МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЗАПОРІЗЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

МАТЕМАТИЧНИЙ ФАКУЛЬТЕТ
Кафедра прикладної математики і механіки

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему: **«ВИКОРИСТАННЯ МЕТОДОЛОГІЇ ТЕОРІЇ ЧАСОВИХ РЯДІВ У ВИВЧЕННІ ТА ОБРОБЦІ БАГАТОРІЧНИХ ЗМІН МЕТЕОРОЛОГІЧНИХ ДАНИХ»**

Виконав(ла): студент(ка) 2 курсу, групи 8.1139
спеціальності 113 прикладна математика
(шифр і назва спеціальності)
освітньої програми прикладна математика
(назва освітньої програми)
М.С. Голуб
(ініціали та прізвище)

Керівник доцент кафедри прикладної математики і механіки, доцент, к.ф.– м.н. Леонтьєва В.В.
(посада, вчене звання, науковий ступінь, прізвище та ініціали)

Рецензент завідувач кафедри фундаментальної математики, доцент, д.т.н. Гребенюк С.М.
(посада, вчене звання, науковий ступінь, прізвище та ініціали)

Запоріжжя
2020

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ЗАПОРІЗЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет математичний

Кафедра прикладної математики і механіки

Рівень вищої освіти магістр

Спеціальність 113 прикладна математика
(шифр і назва)

Освітня програма Прикладна математика

ЗАТВЕРДЖУЮ

Завідувач кафедри прикладної математики і механіки, д.т.н., професор

_____ Грищак В.З.
(підпис)

« ____ » _____

З А В Д А Н Н Я

НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТОВІ(СТУДЕНТЦІ)

Голуб Марії Сергіївні
(прізвище, ім'я та по– батькові)

1. Тема роботи (проекту) Використання методології теорії часових рядів у вивченні та обробці багаторічних змін метеорологічних даних

керівник роботи (проекту) Леонтьєва Вікторія Володимирівна к.ф.– м.н.,доцент
(прізвище, ім'я та по– батькові, науковий ступінь, вчене звання)

затверджені наказом ЗНУ від « 20 » 05 2020 року № 576 – с

2. Строк подання студентом роботи 03.12.2020

3. Вихідні дані до роботи 1. Постановка задачі.
2. Перелік літератури.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)
1. Аналіз об'єкта та предметної області дослідження, огляд сучасного стану проблеми. 2. Методичні особливості застосування теорії часових рядів. 3. Застосування методології теорії часових рядів до аналізу, обробки та прогнозування метеорологічних даних

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)
Презентація

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1			
2			
3			

7. Дата видачі завдання 20.05.2020

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1.	Розробка плану роботи.	21.05.2020	виконано
2.	Збір вихідних даних.	25.06.2020	виконано
3.	Обробка методичних та теоретичних		
	джерел.	28.07.2020	виконано
4.	Розробка першого розділу.	01.09.2020	виконано
5.	Розробка другого розділу.	15.09.2020	виконано
6.	Розробка третього розділу.	16.10.2020	виконано
7.	Оформлення та нормоконтроль		
	кваліфікаційної роботи.	20.11.2020	виконано
8.	Захист кваліфікаційної роботи.	17.12.2020	виконано

Студент _____ М.С. Голуб
(підпис) (ініціали та прізвище)

Керівник роботи _____ В.В. Леонтьєва
(підпис) (ініціали та прізвище)

Нормоконтроль пройдено

Нормоконтролер _____ В.В. Леонтьєва
(підпис) (ініціали та прізвище)

РЕФЕРАТ

Кваліфікаційна робота магістра «Використання методології теорії часових рядів у вивченні та обробці багаторічних змін метеорологічних даних» : 69 с., 18 рис., 24 джерел, 2 додатки.

ЧАСОВИЙ РЯД, ТРЕНД, СЕЗОННА КОМПОНЕНТА, ЦИКЛІЧНА КОМПОНЕНТА, ВИПАДКОВА КОМПОНЕНТА, ПРОГНОЗУВАННЯ, SARIMA, ARIMA .

Об’єкт дослідження – метеорологічні показники міста Запоріжжя.

Мета роботи : провести аналіз ряду метеорологічних даних та, використовуючи методи прогнозування, побудувати прогноз погоди на 5 років.

Метод дослідження – статичний аналіз ряду, прогнозування часових рядів методами ARIMA та SARIMA.

У кваліфікаційній роботі розглядаються методи аналізу та прогнозування часових рядів. Основною задачею роботи є вивчення та обробка багаторічних змін метеорологічних даних, використовуючи методологію теорії часових рядів. Проведено статичний аналіз ряду : виявлення та усунення аномальних рівнів ряду, перевірено на наявність тренду та сезонної компоненти. Побудовано моделі прогнозування ARIMA та SARIMA, зроблено прогноз погоди на 5 років вперед.

SUMMARY

Master’s Qualification Thesis «The use of methodology of time series theory in the study and processing of long– term changes in meteorological data» : 69 pages, 18 figures, 24 references, 2 supplements.

TIME SERIES, TREND, SEASONAL COMPONENT, CYCLIC COMPONENT, RANDOM COMPONENT, FORECASTING, SARIMA, ARIMA.

The object of the study is meteorological indicators of the city.

The aim of the study is to to analyze a number of meteorological meters and, using forecasting methods, to build a weather forecast for 5 years.

The methods of research are static series analysis, time series forecasting by ARIMA and SARIMA methods.

Methods of analysis and forecasting of time series are considered in the qualification work. The main task of the work is to study and process long– term changes in meteorological data, using the methodology of time series theory. Static analysis of the series was performed: detection and elimination of abnormal levels of the series, checked for the presence of trend and seasonal component. ARIMA and SARIMA forecasting models have been built, and the weather forecast for the next 5 years has been made.

ЗМІСТ

ВСТУП

Моделювання і прогнозування явищ і процесів передбачає використання системи статистичних понять, категорій і методів, трактування яких поглиблюється в відповідності до їх статистичних особливостей.

На даний момент в статистичній теорії існує багато різних методів прогнозування інформації. Основна їх частина відноситься до прогнозування часових рядів. Аналіз лише даних спостереження без додаткової інформації, без аналізу впливу зовнішніх сил є особливістю прогнозування часових рядів. Такий аналіз виглядає досить неповним, але досить часто прогнози часових рядів є більш точними.

Основною задачею кваліфікаційної роботи є вивчення та обробка багаторічних змін метеорологічних даних, використовуючи методологію теорії часових рядів.

Кожного року багато шкоди нашій планеті завдають стихійні лиха, та якщо довіряти гідрометеорологічної інформації і адекватно на неї реагувати, то можна запобігти втрат і повністю уникнути людських жертв. Саме тому у своїй роботі я хочу сформувати методики розрахунку прогнозу погоди за допомогою теорії часових рядів та розробити програмне забезпечення, яке буду реалізувати цю методику.

Існують різні методи для прогнозування метеорологічних явищ і їх величин, наприклад, синоптичні, чисельні, статистичні методи, але в повному обсязі жоден метод не забезпечує поки точного прогнозу. Саме тому дослідження в області прогнозування погодних умов є важливими і корисними, а тема – актуальною.

Метою роботи є проведення аналізу ряду метеорологічних даних та, використовуючи методи прогнозування, будування прогнозу погоди на 5 років.

Структурно робота складається з трьох розділів, кожний з яких повністю

розкриває поставленні у роботі завдання.

У першому розділі описуються основні поняття та особливості моделювання багаторічних змін метеорологічних даних та загальні поняття теорії часових рядів. Проводиться аналітичний огляд сучасного стану проблеми.

У другому розділі розглядаються основні підходи проведення аналізу часових рядів, основні методи прогнозування часових рядів, їх сутність, етапи реалізації та умови використання.

У третьому розділі проводиться аналіз вхідного часового ряду. Будуються дві моделі прогнозування ARIMA та SARIMA. За допомогою цих моделей буде зроблений прогноз погоди міста Запоріжжя на 5 років вперед.

1 АНАЛІЗ ОБ'ЄКТА ТА ПРЕДМЕТНОЇ ОБЛАСТІ ДОСЛІДЖЕННЯ. АНАЛІТИЧНИЙ ОГЛЯД СУЧАСНОГО СТАНУ ПРОБЛЕМИ

У першому розділі описуються основні поняття та особливості моделювання багаторічних змін метеорологічних даних та загальні поняття теорії часових рядів. Проводиться аналітичний огляд сучасного стану проблеми.

1.1 Поняття, основні підходи та особливості моделювання багаторічних змін даних

Моделювання і прогнозування явищ і процесів передбачає використання системи статистичних понять, категорій і методів, трактування яких поглиблюється в відповідності до їх статистичних особливостей.

До найважливіших понять і категорій відноситься статистична сукупність, статистична закономірність, закон великих чисел, статистична взаємозв'язок, а також такі філософські категорії як якість і кількість, міра, явище і сутність, одиничне і загальне, випадкове і необхідне.

Статистична закономірність, зумовлює типовий розподіл одиниць статистичної сукупності на деякий момент часу під впливом всієї сукупності факторів [2].

Умовами її прояви є: наявність статистичної сукупності і дію закону великих чисел.

Знаючи статистичну закономірність, можна виявити умови і причини, які породжують її, для того, щоб направляти її дії в заданий «русло», тобто або підтримувати ці умови для її стійкості в часі, або, змінюючи їх, прагнути отримати потрібний результат [1] .

Під статистичною сукупністю (множиною) розуміється безліч одиниць, що володіють масовістю, однорідністю, певною цілісністю, взаємозалежністю станів окремих одиниць і наявністю варіації.

Статистичні сукупності складаються з елементів, одиниць сукупності, які є носієм властивостей досліджуваного явища або процесу.

Ознаки бувають суттєві і не дуже, прямі і непрямі, атрибутивні і кількісні, первинні і вторинні, факторні і результативні.

Класифікація статистичних ознак має важливе значення для побудови статистичних моделей і здійснення прогнозу. Так, при моделюванні в ряді випадків важливо правильно виділити факторні і результативні ознаки. Серед факторних ознак необхідно відбирати лише найсуттєвіші, що визначають основний зміст явищ [2,4].

Закон великих чисел виявляє стійкі пропорції і співвідношення в соціально-економічні явища і процеси і служить основою для їх моделювання, створює можливість управляти ними і передбачати їх розвиток.

Закон великих чисел визначає загальне, істотне в явищах, в їх масі одиниць, завдяки чому відбувається взаємовідношення індивідуальних випадкових відмінностей [5].

Отже, моделювання – відтворення властивостей досліджуваного об’єкта в спеціально побудованій моделі. Для цієї мети використовуються такі статистичні методи як статистичне спостереження, метод угруповань, узагальнюючих показників, кореляційний і регресійний аналіз.

За допомогою статистичного спостереження і соціального експерименту отримують вихідну інформацію для моделювання і прогнозування.

Метод угруповань встановлює наявність та напрямок зв’язку між факторними і результативними ознаками. Для об’єктивних висновків про зв’язок необхідно попередньо визначити кордон, за межами якого вплив групованої ознаки відсутній.

Знаючи статистичну закономірність, можна з тим або іншим ступенем точності передбачити розвиток явища, розкрити сутність і вивчити його структуру.

На основі регресійного і кореляційного аналізу зв'язки отримують свій аналітичний вираз, встановлюється тіснота і напрямок зв'язків між факторними і результативними ознаками. Значимість кореляційних характеристик визначається об'єктивними особливостями досліджуваної сукупності [1,2].

1.2 Основні теоретичні положення теорії часових рядів, використовувані для здійснення дослідження

На даний момент в статистичній теорії існує багато різних методів прогнозування інформації. Основна їх частина відноситься до прогнозування часових рядів. Аналіз лише даних спостереження без додаткової інформації, без аналізу впливу зовнішніх сил є особливістю прогнозування часових рядів. Такий аналіз виглядає досить неповним, але досить 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 [8].

1.2.1 Загальні поняття теорії часових рядів

Часовий ряд – це послідовність спостережень деякої ознаки (випадкової величини) y в послідовні моменти часу. Рівень ряду – це окремі спостереження, які будемо позначати :

$$y_t\ (t = 1,2, \dots, n), \tag{1.1}$$

де n – число рівнів [2, 5].

При досліджені часового ряду y_t в загальному вигляді виділяються декілька складових (адитивна модель) [1] :

$$y_t = T + S + C + E, \text{де } (t = 1,2, \dots, n) \tag{1.2}$$

чи мультиплікативна модель [1] :

$$y_t = T \cdot S \cdot C \cdot E, \tag{1.3}$$

де Т – тренд, описує довготривалу тенденцію зміну ознаки, компонента, яка плавно змінюється (наприклад, ріст населення, економічний розвиток);

S – сезонна компонента, відображає повторюваність процесів на протязі не дуже довготривалого періоду (року, іноді місяця, неділі, наприклад, об’єм продажу товарів);

С – циклічна компонента, відображає повторюваність процесів на протязі дуже довготривалого періоду (наприклад, цикли сонячної активності);

Е – випадкова компонента, відображає вплив випадкових факторів, які не подаються обліку та реєстрації.

Слід звернути увагу на те, що на відміну від E, перші три складові T, S, C є закономірними, невипадковими. Найважливішою класичної завданням при дослідженні економічних часових рядів є виявлення і статистична оцінка основний тенденції розвитку досліджуваного процесу і відхилень від неї [4]. Відзначимо основні етапи аналізу часових рядів :

– графічне представлення та опис поведінки часового ряду;

– виділення та видалення закономірних компонентів часового ряду;

– згладжування і фільтрація;

– дослідження взаємозв’язку між різними часовими рядами;

– прогнозування розвитку досліджуваного процесу на основі часового ряду;

– дослідження випадковою компонентною часового ряду, побудування та перевірка адекватності математичної моделі для її опису.

Серед найбільш поширених методів аналізу часових рядів є кореляційний та спектральний аналіз, моделі авторегресії і ковзної середньої [8].

1.2.2 Особливості побудови часових рядів, види та умови використання часових рядів у практичних дослідженнях

Часові ряди класифікуються за певними рисами. Вони включають в себе два обов’язкові елементи: час і конкретне значення показника, або рівень ряду. Розрізняються часові ряди за такими ознаками :

– за часом – моменті та інтервальні;

– за формою подання рівнів – ряди абсолютних , відносних і середніх величин;

– по відстані між датами або інтервалами часу виділяють повні і неповні тимчасові ряди. Повні ряди мають місце, коли дати реєстрації або закінчення періодів слідують один за одним з рівними інтервалами , неповні – коли принцип рівних інтервалів не дотримується;

– за змістом показників – ряди приватних і агрегованих показників. Приватні показники характеризують досліджуване явище односторонньо, ізольовано. Наприклад, середньодобовий обсяг випуску промислової продукції дає можливість оцінити динаміку промислового виробництва, чисельність громадян, які перебувають на обліку в службі зайнятості; показує ефективність соціальної політики держави; залишки готівки у населення і

вклади населення в банках відображають платоспроможність населення і т.д [17].

Розглянемо основні види часових рядів, а сама інтервальні, моментні , стаціонарні та нестаціонарні часові ряди.

Інтервальний часовий ряд – це послідовність, в якій рівень явища відносять до результату, накопиченому або знову зробленому за певний інтервал часу. Інтервальним, наприклад, є часовий ряд показника випуску продукції підприємством за тиждень, місяць або рік, обсяг електроенергії, виробленої за годину, день, місяць і інші.

Моментний часовий ряд – це сукупність значень , які характеризують досліджене явище в конкретний момент часу. Прикладами моментних рядів є послідовності фізичні показники, такі як температура навколишнього повітря, вологість, тиск, виміряні в конкретні моменти часу, і інші [14, 17].

Стаціонарний часовий ряд – це ряд, який не має тренду або циклічної компоненти , кожен рівень цього ряду дорівнює сумі середнього рівня цього ряду та випадкової компоненти.

Нестаціонарний ряд – це ряд, в склад якого входить дві або три компоненти.

Дисперсія та математичне очікування – важливі характеристики часового ряду.

Ряд $y(t)$ строго стаціонарний за умови, що спільний розподіл m спостережень $y(t_1),\ y(t_2), \dots ,\ y(t_{m+1})$ не залежить від зміни часу, іншими словами, збігається з розподілом [14].

Ряд $y(t)$ є слабо стаціонарним за умови, що математичне очікування, дисперсія і коваріація незалежні від часового моменту.

Якщо ж одне з наведених вище умов порушується, то ряд буде нестаціонарним.

При суворої стаціонарності мається на увазі слабка стаціонарність.

Стаціонарність може бути порушена як по математичному очікуванню, так і по дисперсії [2,17].

Часовий ряд $y(t)$ буде стаціонарним по відношенню до детермінованого ряду $f(t)$, в разі, якщо ряд $(y_t - f(t))$ стаціонарний. Коли ряд $y(t)$ стаціонарний по відношенню до деякого детермінованого тренду, то даний ряд відноситься до класу рядів, стаціонарних по відношенню до детермінованого тренду, тобто є TS поруч.

Типові структури, які можна виділити в часі ряду – тренд, сезонна компонента, циклічна компонента. Тоді детермінована складова може бути записана у вигляді [16] :

$$d_i = t_i + s_i + c_i\,, \tag{1.4}$$

де t_i – тренд,

s_i – сезонна компонента,

c_i – циклічна компонента.

Тренд – компонента тимчасового ряд, яка повільно змінюється, описує вплив на часовий ряд довготривало діючих факторів, що викликають плавні і тривалі зміни ряду.

Щоб уявити характер тренда, зазвичай досить поглянути на графік тимчасового ряду. Найбільш популярні моделі для опису тренда [17]:

– проста лінійна модель : $t_i = a + bi$;

– поліноміальна модель : $t_i = a + b_1 i + b_2 i^2 + \cdots + b_n i^n$. У більшості реальних задач ступінь полінома не перевищує 5;

– експоненціальна модель $t_i = \exp(a + bi)$. Використовується у випадках, коли процес характеризується рівномірним збільшенням темпів зростання;

– логістична модель $t_i = a/(1 + be^{-ki})$, де k – константа, що управляє крутизною логістичної функції. Такого типу криві, що мають S – подібну форму, часто називають сигмоїд. Вони добре описують процеси з непостійними темпами зростання.

Багатьом процесам властива повторюваність в часі, причому періодичність таких повторень може змінюватися в дуже широкому діапазоні. Очевидно, що для опису таких періодичних змін, присутніх у тимчасових рядах, тренд непридатний.

Сезонна компонента – складова часового ряду, що описує регулярні зміни його значень в межах деякого періоду і представляє собою послідовність майже повторюваних циклів.

Сезонна компонента може бути прив’язана до певного календарному тимчасового інтервалу: дня, тижня, місяця – або до якої– небудь події, яка прямо не співвідноситься з конкретними календарними інтервалами. Сезонну компоненту змінливих періодом іноді називають плаваючою [14].

Часто часові ряди містять зміни, занадто плавні і помітні для випадкової складової. У той же час такі зміни не можна віднести ні до тренду, оскільки вони не є достатньо протяжними, ні до сезонної компоненті, оскільки вони не є регулярними. Подібні зміни називаються циклічної компонентою часового ряду.

Циклічна компонента часового ряду – інтервали підйому або спаду, які мають різну протяжність, а також різну амплітуду розташованих в них значень.

Вивчення циклічної компоненти часто виявляється корисним для прогнозування, особливо короткострокового [17].

Випадковою компонентою, називається випадковий шум або помилка, яка впливає на часовий ряд нерегулярно.

Випадкова (стохастична) компонента часового ряду – послідовність значень, яка є результатом впливу на досліджуваний процес випадкових чинників. Випадкова складова і її вплив на часовий ряд можуть бути оцінені тільки за допомогою статистичних методів.

Випадкова компонента проявляється як результат впливу набору випадкових факторів на досліджуваний процес і зазвичай виражається в підвищеній мінливості часового ряду, а також у відхиленні значень

детермінованою складової. Результуюче значення часового ряду – це результат взаємодії детермінованої і випадкових складових. Найпростіший вид такої взаємодії – випадок, коли, кожне значення часового ряду можна розглядати як суму (різницю) двох значень, одне з яких обумовлено детермінованою складовою, а інше – випадковою , тобто $x_i = d_i + p_i$ [1].

Існує кілька підходів до аналізу структури часових рядів, що містять сезонні або циклічні коливання.

Найпростіший підхід – розрахунок значень сезонної компоненти методом ковзної середньої і побудова адитивної або мультиплікативної моделі часового ряду. Загальний вигляд адитивної моделі наступний [2] :

$$Y = T + S + C + E \tag{1.5}$$

Ця модель передбачає, що кожен рівень часового ряду може бути представлений як добуток трендової, сезонної і випадкової компонент. Загальний вигляд мультиплікативної моделі виглядає так [2] :

$$Y = T \cdot S \cdot C \cdot E \tag{1.6}$$

Ця модель передбачає, що кожен рівень часового ряду може бути представлений як добуток трендової, сезонної і випадкової компонент. Вибір однієї з двох моделей здійснюється на основі аналізу структури сезонних коливань. Якщо амплітуда коливань приблизно постійна, будують адитивну модель тимчасового ряду, в якій значення сезонної компоненти передбачаються постійними для різних циклів. Якщо амплітуда сезонних коливань зростає або зменшується, будують мультиплікативну модель часового ряду, яка ставить рівні ряду в залежність від значень сезонної компоненти.

Побудова адитивної і мультиплікативної моделей зводиться до розрахунку значень трендової, циклічної і випадкової компонент для кожного рівня ряду [17].

Процес побудови моделі включає в себе наступні кроки.

1. Вирівнювання вихідного ряду методом ковзної середньої.
2. Розрахунок значень сезонної компоненти.
3. Усунення сезонної компоненти з вихідних рівнів ряду і отримання вирівняних даних в адитивної або мультиплікативної моделі.
4. Аналітичне вирівнювання рівнів і розрахунок значень тренду з використанням отриманого рівняння тренда.
5. Розрахунок отриманих за моделлю значень
6. Розрахунок абсолютних і відносних помилок.

Якщо отримані значення помилок не містять автокореляції, ними можна замінити вихідні рівні ряду і надалі використовувати часовий ряд помилок для аналізу взаємозв’язку вихідного ряду і інших часових рядів [20].

1.3 Аналітичний огляд сучасного стану проблеми

На даний момент в нашому сучасному світі екологічні проблеми мають дуже глобальний характер і тому різко підвищується інтерес до вивчення різномасштабних кліматичних процесів. Саме клімат відображає антропогені зміни [10].

Зараз в наш час максимально інтенсивно розвиваються прикладні галузі кліматології – транспортна, авіаційна, будівельна, технічна та інші. Ці галузі мають певні вимоги до кліматичної інформації та методам отримання даних, оскільки вони відрізняються від традиційних розрахунків показників. Багато робіт з оцінками змін клімату у майбутньому було виконано та опубліковано у зв’язку з потеплінням клімату останнього століття та їх негативними наслідками [18, 22].

Клімат можна визначити, як багаторічний режим погоди, опис і вивчення, якого опирається на багаторічні спостереження за метеорологічними величинами. Одна з основних задач сучасної кліматології – обробка багаторічних спостережень. Методи кліматологічної обробки є самостійним розділом кліматології , який засновується на матеріалах, які доставляються кліматологічною обробкою метеоданих [23].

Вихідні ряди спостережень складаються з множини цифр, тому первині дані спостережень попередньо систематизуються. Так як кліматичні дані мають всі основні властивості статистичних сукупностей, при статичній обробці даних можливе застосування багатьох методів варіаційної статистики [10].

Одна з найскладніших завдань фізики атмосфери ,з наукової точки зору є передбачення погоди. Існує багато різних методів прогнозування метеорологічних явищ , наприклад, синоптичні, чисельні, статистичні, але ні один з цих методів не забезпечує точного прогнозу. Саме тому дослідження в області прогнозування погодних умов є важливими і корисними, а тема – актуальною.

2 МЕТОДОЛОГІЧНІ ОСОБЛИВОСТІ ЗАСТОСУВАННЯ ТЕОРІЇ ЧАСОВИХ РЯДІВ. МАТЕМАТИЧНИЙ АПАРАТ ТЕОРІЇ ЧАСОВИХ РЯДІВ

У другому розділі розглядаються основні підходи проведення аналізу часових рядів, основні методи прогнозування часових рядів, їх сутність, етапи реалізації та умови використання

2.1 Основні підходи до проведення аналізу часових рядів

Аналіз часових рядів – сукупність статистичних методів для виявлення складових часового ряду і його прогнозування.

Існують дві основні мети аналізу часових рядів :

– визначення природи ряду;

– прогнозування (передбачення майбутніх значень часового ряду по теперішнім і минулим значенням) [12].

Обидві ці цілі вимагають, щоб модель ряду була ідентифікована і, більш – менш, формально описана. Як тільки модель визначена, ви можете з її допомогою інтерпретувати розглянуті дані. Не звертаючи уваги на глибину розуміння і справедливість теорії, ряд можна екстраполювати на основі знайденої моделі, тобто передбачити його майбутні значення.

Основна мета аналізу часового ряду – побудувати прогноз його значень на майбутні періоди. А основні завдання аналізу часового ряду – зрозуміти, під впливом яких компонент формується значення часового ряду, і побудувати математичну модель для кожної компоненти або їх сукупності. Будь-який часовий ряд можна розкласти на такі складові: тренд, сезонну складову, циклічну складову і випадкову складову. Перші три компоненти утворюють невипадково складову часового ряду. Випадкова складова присутня в будь-

якому часовому ряді. А ось присутність в структурі часового ряду компонент невипадковою складової не обов’язково .

Підходи до моделювання часового ряду можна розділити на два напрямки :

– моделювання невипадковою складової в сукупності;

– розкладання тимчасового ряду на складові компоненти і моделювання значень кожної компоненти окремо [5].

Статистичні методи прогнозування поділяються на алгоритмічні методи і аналітичні методи. До алгоритмічних методів відносять методи простий і зваженою ковзної середньої. До аналітичних методів відносять методи прогнозної екстраполяції на основі кривих росту у вигляді функцій часу. У разі наявності сезонної або циклічної компоненти в часі ряду проводять аналіз періодичних коливань або спектральний аналіз тимчасового ряду.

Часові ряди класифікують на стаціонарні та нестаціонарні. Для аналізу і побудови прогнозу по стаціонарному тимчасового ряду використовують особливі методи: моделі змінного середнього, моделі авторегресії або змішані моделі або моделі про інтегрувати змінного середнього і авторегресії [14].

Існують дві основні мети аналізу часових рядів :

– визначення природи ряду;

– прогнозування (передбачення майбутніх значень часового ряду по теперішнім і минулим значенням).

Обидві ці цілі вимагають, щоб модель ряду була ідентифікована і, більш – менш, формально описана. Як тільки модель визначена, ви можете з її допомогою інтерпретувати розглянуті дані. Не звертаючи уваги на глибину розуміння і справедливість теорії, ряд можна екстраполювати на основі знайденої моделі, тобто передбачити його майбутні значення.

Методи аналізу повинні змінюватися в залежності від характеру досліджуваних процесів, їх специфіки, особливостей і форм прояву.

2.1.1 Попередній аналіз часових рядів

Суть попереднього аналізу часових рядів полягає у виявленні та усуненні аномальних значень рівнів ряду.

Аномальний рівень – це значення часового, яке не відповідає потенційним можливостям досліджуваної системи і яке робить істотний вплив на значення основних характеристик часового ряду, втому числі на відповідну трендову модель. Причинами аномальних спостережень можуть бути помилки технічного порядку або помилки першого роду. Помилки першого роду підлягають виявленню та усуненню.

Для виявлення аномальних значень ряду використовується критерій Ірвіна, згідно з яким аномальної вважається точка Y_t, що відстоїть від попередньої точки Y_{t-1} на величину, більшу середньоквадратичного відхилення [16] :

$$\lambda_i = \frac{|Y_t - Y_{t-1}|}{\sigma}, \tag{2.1}$$

де σ – середньоквадратичне відхилення,

λ_i – критерій Ірвіна

$$\sigma = \sqrt{\frac{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}{n-1}}. \tag{2.2}$$

Точка вважається аномальною, якщо $\lambda_i > \lambda_{\text{таб}}$. Табличні значення $\lambda_{\text{таб}}$ зменшуються зі зростанням довжини ряду [18].

Дуже часто рівні ряду динаміки коливаються, так що тенденція розвитку процесу прихована випадковими відхиленнями. Згладжування часового ряду дозволяє відфільтрувати дрібні випадкові коливання і виявити основну тенденцію зміни досліджуваної величини.

Основною метою згладжування ряду є виділення трендової компоненти процесу. При згладжуванні часового ряду в більшій чи меншій мірі згладжується вплив нерегулярної складової відгуку, так що згладжений ряд фактично виявляється суперпозицією тренда і циклічної (і можливо сезонної) складових процесу, що полегшує їх подальше дослідження. Зазвичай використовується метод ковзного середнього або метод експоненціального згладжування; обидва методи є суб'єктивними щодо вибору параметрів згладжування, але саме в коректному виборі параметрів і проявляється досвід і інтуїція дослідника.

Найпростіший метод згладжування рядів – ковзне середнє. Ідея полягає в тому, що для будь – якого непарного кількості точок послідовності ряду замінювати центральну точку на середнє арифметичне решти точок [16] :

$$s_i = \frac{1}{2k+1} \sum_{j=-k}^{k} x_{i+j} \,, \qquad (2.3)$$

де x_i – вхідний ряд,

s_i – згладжений ряд.

Метод ковзного середнього має певні недоліки :

– ковзне середнє неефективно в обчисленні. Для кожної точки середнє необхідно переобчислювати заново. Ми не можемо перевикористати результат, обчислений для попередньої точки;

– ковзне середнє не можна продовжити на перші і останні точки ряду. Це може викликати проблему, якщо нас цікавлять саме ці точки;

– ковзне середнє не визначене за межами ряду, і як наслідок, не може використовуватися для прогнозування.

Більш просунутий метод згладжування – експоненціальне згладжування, також іноді зване методом Хольта-Уінтерс (Holt-Winters) в честь імен його творців [18].

Існує декілька варіантів цього методу:

– одинарне згладжування для рядів, у яких немає тренда і сезонності;

– подвійне згладжування для рядів, у яких є тренд, але немає сезонності;

– потрійне згладжування для рядів, у яких є і тренд, і сезонність.

Метод експоненціального згладжування обчислює значення згладженого ряду шляхом оновлення значень, розрахованих на попередньому кроці, використовуючи інформацію з поточного кроку. Інформація з попереднього і поточного кроків береться з різними вагами, якими можна управляти.

У найпростішому варіанті одинарного згладжування співвідношення таке [16] :

$$s_i = \alpha x_i + (1-\alpha)s_{i-1}, \text{ де } 0 \le \alpha \le 1 \qquad (2.4)$$

Коли ця формула застосовується рекурсивно, то кожне нове згладжене значення (яке є також прогнозом) обчислюється як зважене середнє поточного спостереження і згладженого ряду. Очевидно, результат згладжування залежить від параметра α (альфа). Якщо α дорівнює 1, то попередні спостереження повністю ігноруються. Якщо α дорівнює 0, то ігноруються поточні спостереження. Значення α між 0, 1 дають проміжні результати.

Якщо в даних є тренд, просте згладжування буде «відставати» від нього (або доведеться брати значення α близькими до 1, але тоді згладжування буде недостатнім). Потрібно використовувати подвійне експоненціальне згладжування.

Подвійне згладжування використовує вже два рівняння – одне рівняння оцінює тренд як різницю між поточним і попереднім згладженим значеннями, потім згладжує тренд простим згладжуванням. Друге рівняння виконує згладжування як в разі простого варіанту, але в другому доданку використовується сума попереднього згладженого значення і тренду.

Потрійне згладжування включає ще один компонент – сезонність, і використовує ще одне рівняння. При цьому розрізняються два варіанти

сезонного компонента – адитивний і мультиплікативний. У першому випадку амплітуда сезонного компонента постійна і з часом не залежить від базової амплітуди ряду. У другому випадку амплітуда змінюється разом зі зміною базової амплітуди ряду. З ростом ряду амплітуда сезонних коливань збільшується.

2.1.2 Аналіз ряду на наявність тренду

Для визначення наявності тренда часового ряду використовуються метод перевірки різниць середніх рівнів і метод Фостера-Стьюарта [8].

Метод перевірки різниць середніх рівнів.

Реалізація цього методу складається з чотирьох етапів.

На першому етапі вихідний часовий ряд $y_1, y_2, y_3, \dots, y_n$ розбивається на дві приблизно рівні за кількістю рівнів частини: в першій частині n_1 перших рівнів вихідного ряду, у другій – n_2 інших рівнів $(n_1 + n_2 = n)$ [16].

На другому етапі для кожної з цих частин обчислюються середні значення і дисперсії [16] :

$$\overline{y_1} = \frac{\sum_{t=1}^{n_1} y_t}{n_1}; \tag{2.5}$$

$$\sigma_1^2 = \frac{\sum_{t=1}^{n_1} (y_t - \overline{y_1})^2}{n_1 - 1}; \tag{2.6}$$

$$\overline{y_2} = \frac{\sum_{t=n_1+1}^{n} y_t}{n_2}; \tag{2.7}$$

$$\sigma_2^2 = \frac{\sum_{t=n_1+1}^{n_1} (y_t - \overline{y_2})^2}{n_2 - 1}. \tag{2.8}$$

Третій етап полягає в перевірці рівності (однорідності) дисперсій обох частин ряду за допомогою F – критерію Фішера, яка заснована на порівнянні розрахункового значення цього критерію [18] :

$$F = \begin{cases} \frac{\sigma_1^2}{\sigma_2^2}, \text{якщо } \sigma_1^2 > \sigma_2^2 \\ \frac{\sigma_2^2}{\sigma_1^2}, \text{якщо } \sigma_1^2 < \sigma_2^2 \end{cases} \quad (2.9)$$

Якщо отримане значення F менше табличного $F_{\text{табл}}$, то гіпотеза про однорідність дисперсії приймається і переходять до наступного етапу розрахунку. Якщо F більше або дорівнює табличному значенню $F_{\text{табл}}$, то гіпотеза про однорідність дисперсій відхиляється і метод не дає відповіді на питання про наявність чи відсутність тренда.

Табличне значення $F_{\text{табл}}$ залежить від рівня значущості і довжини порівнюваних рядів.

Остаточна перевірка гіпотези про відсутність тренда проводиться з використанням t– критерію Стьюдента, який обчислюється за формулою [18] :

$$t = \frac{|\overline{y_1} - \overline{y_2}|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad (2.10)$$

де σ – середньоквадратичне відхилення різниці середніх [16] :

$$\sigma = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}. \quad (2.11)$$

Якщо розрахункове значення t менше табличного значення $t_{\text{табл}}$, то гіпотеза приймається, тобто тренда немає, в іншому випадку тренд є. Для визначення табличного значення число ступенів свободи приймається рівним $n_1 + n_2 - 2$.

Метод Фостера– Стьюарта.

Цей метод має великі можливості і дасть більш надійні результати. Крім тренда самого ряду (як кажуть, тренда в середньому), він дозволяє встановити наявність тренда дисперсії часового ряду: якщо тренду дисперсії немає, то

розкид рівнів ряду постійний; якщо дисперсія збільшується, то ряд "розгойдується" і т.д [14].

Реалізація методу містить чотири етапи.

На першому етапі проводиться порівняння кожного рівня вихідного часового ряду, починаючи з другого рівня, з усіма попередніми, при цьому визначаються дві числові послідовності [16] :

$$k_t = \begin{cases} 1, \text{якщо } y_t \text{ більше всіх попередніх рівнів ;} \\ 0 \text{ в іншому випадку ;} \end{cases} \quad (2.12)$$

$$l_t = \begin{cases} 1, \text{якщо } y_t \text{ менше всіх попередніх рівнів ;} \\ 0 \text{ в іншому випадку.} \end{cases} \quad (2.13)$$

де $t = 2, 3, 4, \dots, n$.

Обчислимо величини s і d, що характеризують зміна часового ряду і дисперсії [14] :

$$s = \sum_{t=2}^{n}(k_t + l_t); \quad (2.14)$$

$$d = \sum_{t=2}^{n}(k_t - l_t). \quad (2.15)$$

Величина s характеризує зміну часового ряду, вона може приймати значення від 0 (коли всі рівні ряду рівні) до $n - 1$ (ряд монотонний). Величина d характеризує зміну дисперсії часового ряду і змінюється від $-(n-1)$ (коли ряд монотонно убуває) до $(n - 1)$ (коли ряд монотонно зростає). Ці величини є випадковими з математичним очікуванням μ для значення s і 0 для значення d.

Перевіримо гіпотези про випадковість відхилення величини s від її математичного очікування μ і про випадковість відхилення величини d від нуля за допомогою критерію Стьюдента для середньої і для дисперсії [14, 16] :

$$t_s = \frac{|s-\mu|}{\sigma_1}; \quad (2.16)$$

$$\sigma_1 = \sqrt{2\ln n - 3{,}4253}; \quad (2.17)$$

$$t_d = \frac{|d-0|}{\sigma_2}; \quad (2.18)$$

$$\sigma_2 = \sqrt{2\ln n - 0{,}8456}. \quad (2.19)$$

де μ – математичне очікування величини s для випадкового часового ряду;

σ_1 – середньоквадратичне відхилення s для випадкового часового ряду;

σ_2 – середньоквадратичне відхилення d для випадкового часового ряду.

Отримані значення t_s, t_d необхідно порівняти з табличними значеннями критерію Стьюдента $t_{\text{табл}}$. якщо $t_{\text{табл}}$ більше розрахункового значення, то відповідний тренд відсутній: тобто, якщо $t_s > t_{\text{табл}}$, а $t_d > t_{\text{табл}}$, то тренд ряду є, а тренда дисперсії немає [14].

2.1.3 Аналіз ряду на наявність сезонності

При аналізі коливання динамічних рядів поряд з виділенням випадкових коливань, виникає завдання вивчення періодичних коливань. Як правило, вивчення періодичних (сезонних) коливань необхідно з метою виключення їх впливу на загальну динаміку для виявлення чистої (випадкової) коливання [8].

До сезонних відносять всі явища, які виявляють в своєму розвитку чітко виражену закономірність річних змін, тобто більш-менш стійко повторюються з року в рік коливання рівнів. Часто ці коливання можуть бути не пов’язані зі зміною пір року.

Багато часових рядів мають яскраво виражені сезонні компоненти, що повторюються з певною періодичністю. Ця періодичність має місце щороку.

Якщо в аналізованій часовій послідовності спостерігаються стійкі відхилення від тенденції (в більшу або в меншу сторону), то можна припустити наявність у ряду динаміки деяких (одного або декількох) коливальних процесів [16].

Це особливо помітно, коли досліджувані явища мають сезонний характер, зростання або спадання рівнів повторюється регулярно з інтервалом в один рік [1].

Завдання, які необхідно вирішити в ході дослідження сезонності :

– виявити наявність сезонності;
– чисельне висловити сезонні коливання;
– виділити фактори, що викликають сезонні коливання;
– оцінити наслідки сезонних коливань;
– провести математичне моделювання сезонності.

Для вимірювання сезонних коливань статистикою запропоновані різні методи. Найбільш прості і часто вживані з них :

– метод абсолютних різниць;
– метод відносних різниць;
– побудова індексів сезонності.

Перші два способи припускають знаходження різниць фактичних рівнів і рівнів, знайдених при виявленні основної тенденції розвитку (тренда) [16].

Застосовуючи спосіб абсолютних різниць, оперують безпосередньо розмірами цих різниць, а при використанні методу відносних різниць, визначають ставлення абсолютних розмірів зазначених різниць до вирівняні рівню. При виявленні основної тенденції використовують або метод ковзної середньої, або аналітичне вирівнювання. У деяких випадках в стаціонарних рядах можна користуватися різницею фактичних рівнів і середнім місячним рівнем за рік. Використання даних за кілька років пов’язано з тією обставиною, що в відхиленнях по окремих роках сезонні коливання

змішуються з випадковими. Щоб елімінувати випадкові коливання, беруть середні відхилення за кілька років .

Для виділення сезонної хвилі треба визначити середній рівень за кожен місяць по 3 – 5 – річним даними і загальну середню за весь аналізований період.

Загальна середня виходить розподілом суми рівнів за все три– п’ять років на 36 або 60 (загальне число місяців). Потім визначається абсолютне відхилення середніх місячних показників від загальної середньої [8].

Метод абсолютних різниць полягає в розрахунку місячних середніх і загальної середньої з подальшим їх порівнянням [16] :

$$\Delta_{\text{сез}} = \overline{y_t} - \overline{y_c}\ , \qquad (2.20)$$

де $\overline{y_t}$ – середній місячний рівень показника за три та більше років,

$\overline{y_c}$ – середньомісячне значення показника за всі роки.

Якщо сезонність оцінюється по даним за 3 роки (36 місяців), якщо за п’ять років (60 місяців) [16] :

$$\overline{y_c} = \frac{\sum y_i}{36}, \qquad (2.21)$$

де y_i – значення рівня динамічного ряду.

Величина і знак значень абсолютних відхилень визначають наявність сезонності.

Як показник, що характеризує сезонну нерівномірність, використовується показник відносного відхилення.

Метод відносних різниць є розвитком методу абсолютних різниць. Для знаходження відносних різниць абсолютні відхилення ділять на загальну середню і виражають у відсотках. За величиною і знакам значень відносних відхилень можна судити про величину і силі впливу сезонного фактору [17].

$$\Delta_{\text{отн}} = \frac{\overline{y_t} - \overline{y_c}}{\overline{y_c}}, \qquad (2.22)$$

Замість відносних різниць за кожен місяць може бути обчислений індекс сезонності, який розраховується як відношення середнього рівня відповідного місяця до загальної середньої. Індекс сезонності розраховується [16] :

$$I_{\text{сез}} = \frac{\overline{y_t}}{\overline{y_c}}, \qquad (2.23)$$

де y_t – середній рівень показника відповідного місяця за три та більше років,

$y_{\text{с}}$ – середньомісячне (по року) значення показника за всі роки (загальне середнє).

Розраховані значення індексу сезонності порівнюються із значенням 100%. Якщо індекс сезонності перевищує 100% – це свідчить про вплив сезонного фактора в бік збільшення рівнів динамічного ряду і навпаки. Розрахунок індексу сезонності по даній формулі не враховує наявність тренда. Виділення сезонної хвилі можна виконати на основі побудови аналітичної моделі прояви сезонних коливань. Побудова аналітичної моделі виявляє основний закон коливання даного часового ряду в зв'язку з переходом від місяця до місяця і дає лише середню характеристику річних коливань [17].

2.2 Основні методи прогнозування часових рядів: сутність, етапи реалізації та умови використання

Прогнозування на основі часових рядів – один із самих популярних підходів до прогнозування розвитку економічних процесів, об'ємів торгових операцій, об'ємів виробництва та накопичення продукції на складах,

оцінювання альтернативних економічних стратегій, формування бюджетів підприємств та держави, прогнозування та менеджмент економічних і фінансових ризиків та інше. Загалом методи прогнозування можна розділити на три широкі класи [14] :

а) прогнозування на основі суджень, тобто, прогнозування, що ґрунтується на суб'єктивних судженнях (оцінках), інтуїції, поглиблених знаннях конкретної області та іншій інформації, що має відношення до прогнозованого процесу – так зване передбачення;

б) методи прогнозування на основі використання часового ряду однієї змінної, тобто, на основі авторегресії, авторегресії з ковзним середнім (ACF) та AACF плюс модель тренду;

в) методи прогнозування на основі використання часових рядів декількох змінних.

Можна по-різному ставити задачу прогнозування в залежності від рівня прийняття рішення та конкретної поставленої задачі управління чи контролю. Прогнозування може стосуватись таких складових процесу [16] :

– детермінованого тренду, як індикатора довгострокових змін процесу;

– випадкового (нерегулярного) тренду, як показника коротко – та середньострокових змін;

– короткострокових змін, тобто, прогнозування коливань (відхилень), що накладаються на тренд;

– сезонних ефектів;

– приростів (швидкості) зміни процесу, які визначаються першими різницями;

– дисперсії або стандартного відхилення, як міри розсіювання процесу;

– якісних змінних (за допомогою нечітких множин, мереж Байєса і т. ін.);

– комбінацій вказаних елементів процесів.

Відповідно до того, які складові процесу необхідно прогнозувати, ставиться задача побудови математичної, ймовірнісної або логічної моделі, що має меті забезпечити високу якість прогнозу на заданому горизонті [5].

Методи прогнозування часових рядів:

– ковзне середнє;
– зважене ковзне середнє;
– експоненціальне згладжування;
– авторегресія ковзного середнього (ARMA);
– інтегрована модель ковзного середнього (ARIMA);
– сезонна інтегрована модель ковзного середнього (SARIMA);
– екстраполяція;
– лінійне прогнозування;
– оцінка тренду;
– крива зростання (статичні дані);
– нейромережеві моделі.

Далі розглянемо детально методи які були використані в цій роботі.

2.2.1 Авторегресійна інтегрована модель ковзного середнього ARIMA

ARIMA – інтегрована модель авторегресії ковзного середнього – модель і методологія аналізу часових рядів. ARIMA – розширення моделей ARMA для нестаціонарних часових рядів, які можна зробити стаціонарними. Модель $ARIMA(p, d, q)$ означає, що різниці часового ряду порядку d підкорюються моделі $ARMA(p, q)$ [16].

Модель авторегресії порядку описується як [11] :

$$AR(p): y_t = c + \varphi 1 y_{(t-1)} + \varphi 2 y_{(t-2)} + \cdots + \varphi p y_{(y-p)} + \varepsilon_t, \quad (2.24)$$

і показує залежність значення нинішнього періоду від минулих значень p періодів.

Модель змінного середнього порядку q описується як [16] :

$$MA(q): y_t = c + \varepsilon_t + \theta_1 \varepsilon_{(t-1)} + \theta_2 \varepsilon_{(t-2)} + \cdots + \theta_q \varepsilon_{(t-q)}, \qquad (2.25)$$

і показує залежність значення нинішнього періоду від помилок передбачення попередніх q періодів.

Модель авторегресії з інтеграцією і ковзаючим середнім порядків (p, d, q) є сумою AR (p) і MA (q) моделей і може бути представлена у вигляді [11] :

$$ARIMA(p,d,q): \left(1 - \varphi_1 L - \cdots - \varphi_p L^p\right)((1-L)^d y_t - \mu) = \\ = \left(1 + \theta_1 L + \cdots + \theta_q L^q\right)\varepsilon_t \,, \qquad (2.26)$$

де d – кількість диференціювання вихідного часового ряду до досягнення його стаціонарності;

L – величина лагу.

Підхід ARIMA до часових рядах полягає в тому, що в першу чергу оцінюється стаціонарність ряду. Різними тестами виявляються наявність поодиноких коренів і порядок інтегрованості тимчасового ряду (зазвичай обмежуються першим або другим порядком). Далі при необхідності (якщо порядок інтегрованості більше нуля) ряд перетворюється взяттям різниці відповідного порядку і вже для перетвореної моделі будується деяка ARMA – модель, оскільки передбачається, що отриманий процес є стаціонарним, на відміну від вихідного нестаціонарного процесу (інтегрованого процесу порядку d) [1].

Побудову $ARIMA(p,d,q)$ моделі часового ряду складається з таких етапів :

Перший крок. Необхідно отримати стаціонарний ряд. При тестуванні вихідних даних на стаціонарність насамперед використовується візуальний аналіз графіка. Наприклад, вже на цьому етапі можна виявити яскраво виражену трендову складову [18].

Також в методиці Бокса-Дженкінса рекомендується проводити аналіз ACF (PACF). Швидке спадання значень вибіркової ACF є простим критерієм стаціонарності (аналогічна поведінка повинна демонструвати і PACF).

Часто на цьому етапі використовуються статистичні тести на наявність одиничного кореня (тест Дікі-Фуллера, розширений тест Дікі-Фуллера).

Для переходу до стаціонарного ряду традиційно застосовують оператор взяття послідовних різниць (процедуру дискретного диференціювання). Швидке загасання АКФ буде свідчити про те, що необхідна для стаціонарності ряду ступінь різниці досягнута.

Другий крок. Після отримання стаціонарного ряду досліджується характер поведінки вибіркових ACF і PACF, висуваються гіпотези про значення параметрів p (порядок авторегресії) і q (порядок змінного середнього).

При цьому слід мати на увазі, що вибіркові кореляційні функції можуть не демонструвати детального подібності з теоретичними. Тому для ідентифікації моделі можуть використовуватися головні риси ACF, при розбіжності більш тонких деталей, в результаті формується базовий набір, що включає 1-2 або навіть більше число моделей.

Третій крок. Після здійснення ідентифікації моделей необхідно оцінити їх параметри. В сучасних економетричних пакетах прикладних програм використовуються різні підходи (метод найменших квадратів, нелінійний МНК, метод максимальної правдоподібності). Всі ці оцінки при великих обсягах вибірок асимптотично еквівалентні.

На наступному, четвертому етапі для перевірки кожної пробної моделі на адекватність аналізується ряд її залишків. У адекватної моделі залишки повинні

бути схожими на білий шум, тобто їх вибіркові автокореляції не повинні істотно відрізнятися від нуля.

При перевірці значимості коефіцієнтів ACF використовуються два підходи :

– перевірка значимості кожного коефіцієнта автокореляції окремо;

– перевірка значимості множини коефіцієнтів автокореляції як групи.

Крім того, при побудові моделі ARIMA необхідно перевірити значимість коефіцієнтів (по t-критерієм). При цьому модель не повинна містити зайвих параметрів, тобто зменшення числа параметрів буде сприяти появі значимої автокореляції залишків.

Якщо в результаті перевірки кілька моделей виявляються адекватними вихідними даними, то при остаточному виборі слід врахувати дві вимоги :

– підвищення точності (якість підгонки моделі);

– зменшення числа параметрів моделі.

На заключному етапі за допомогою моделі, обраної на четвертому кроці, можна будувати точковий та інтервальний прогноз на L кроків вперед [16].

2.2.2 Сезонна авторегресійна інтегрована модель ковзного середнього (SARIMA)

Сезонна авторегресійна інтегрована модель ковзного середнього, SARIMA або Seasonal ARIMA, є розширенням ARIMA, яке явно підтримує одновимірні дані часових рядів з сезонною компонентою [8].

Він додає три нових гіперпараметра для вказівки на авторегресію (AR), різницю (I) і ковзне середнє (MA) для сезонної складової ряду, а також додатковий параметр для періоду сезонності.

Сезонна модель ARIMA формується шляхом включення додаткових сезонних термінів в ARIMA. Сезонна частина моделі складається з термінів,

які дуже схожі на несезонні компоненти моделі, але включають зворотні зрушення сезонного періоду [16].

Налаштування SARIMA вимагає вибору гіперпараметрів як для трендових, так і для сезонних елементів ряду.

Є три елементи тренда, які вимагають налаштування [21].

Вони такі ж, як модель ARIMA; зокрема:

– p: Порядок авторегресії тренда.

– d: Порядок зміни тренда.

– q: Тренд ковзної середньої.

Є чотири сезонних елемента, які не є частиною ARIMA, які повинні бути налаштовані:

– P: Сезонний порядок авторегресії.

– D: Порядок сезонних різниць.

– Q: Сезонний порядок ковзних середніх.

– m: Кількість тимчасових кроків за один сезонний період.

Разом позначення для моделі SARIMA задається як [16]:

$$ARIMA(p,d,q)(P,D,Q)m. \qquad (2.27)$$

Параметри P, D, Q , m дозволяють врахувати циклічні коливання процесу.

Сезонна модель ARIMA використовує різницю з запізненням, що дорівнює кількості сезонів, для усунення адитивних сезонних ефектів. Як і у випадку з різницею в запізненні для видалення тренду, відмінність в запізненні вводить термін ковзне середнє. Сезонна модель ARIMA включає умови авторегресії і ковзного середнього з затримкою [21].

Алгоритм побудови моделі SARIMA [14] :

– перевіряємо стаціонарність: якщо часовий ряд має компонент тренда або сезонності, він повинен бути стаціонарним, перш ніж почати використовувати ARIMA для прогнозування;

– різниця: якщо часовий ряд не є стаціонарним, він повинен бути стаціонарним за допомогою диференціювання;

– фільтрація перевірочного зразка: використовується для перевірки точності нашої моделі;

– вибір термінів AR і MA: використовується ACF і PACF, щоб вирішити, чи слід вносити термін AR, термін MA або обидва;

– побудування моделі: будується модель і встановлюється кількість періодів для прогнозу N (залежить від потреб);

– перевірка моделі: порівнюються прогнозні значення з фактичними даними в перевірочному зразку [21].

3 ЗАСТОСУВАННЯ МЕТОДОЛОГІЇ ТЕОРІЇ ЧАСОВИХ РЯДІВ ДО АНАЛІЗУ, ОБРОБКИ ТА ПРОГНОЗУВАННЯ БАГАТОРІЧНИХ ЗМІН МЕТЕОРОЛОГІЧНИХ ДАНИХ

У третьому розділі проводиться аналіз вхідного часового ряду. Буде побудовано дві моделі прогнозування ARIMA та SARIMA. За допомогою цих моделей зробимо прогноз погоди міста Запоріжжя на 5 років вперед. Проводиться аналіз, яка з моделей підходить більше для прогнозування метеоданих .

3.1 Вхідні дані для проведення дослідження. Побудова вихідного часового ряду метеорологічних даних

Часовий ряд – послідовні виміри , впорядковані в невипадкові моменти часу. В нашому випадку часовий ряд представлений у вигляді метеорологічних даних.

Для побудови часового ряду метеорологічних даних, побудови моделей прогнозування та проведення прогнозу було реалізовано програму на мові програмування Python.

Для того, щоб провести прогнозування ряду було взято середні значення температури міста Запоріжжя по місяцям за минулі 10 років в період з 2009 – 2019 роки. Дані були взяті з архіву метеоданих міста Запоріжжя [24].

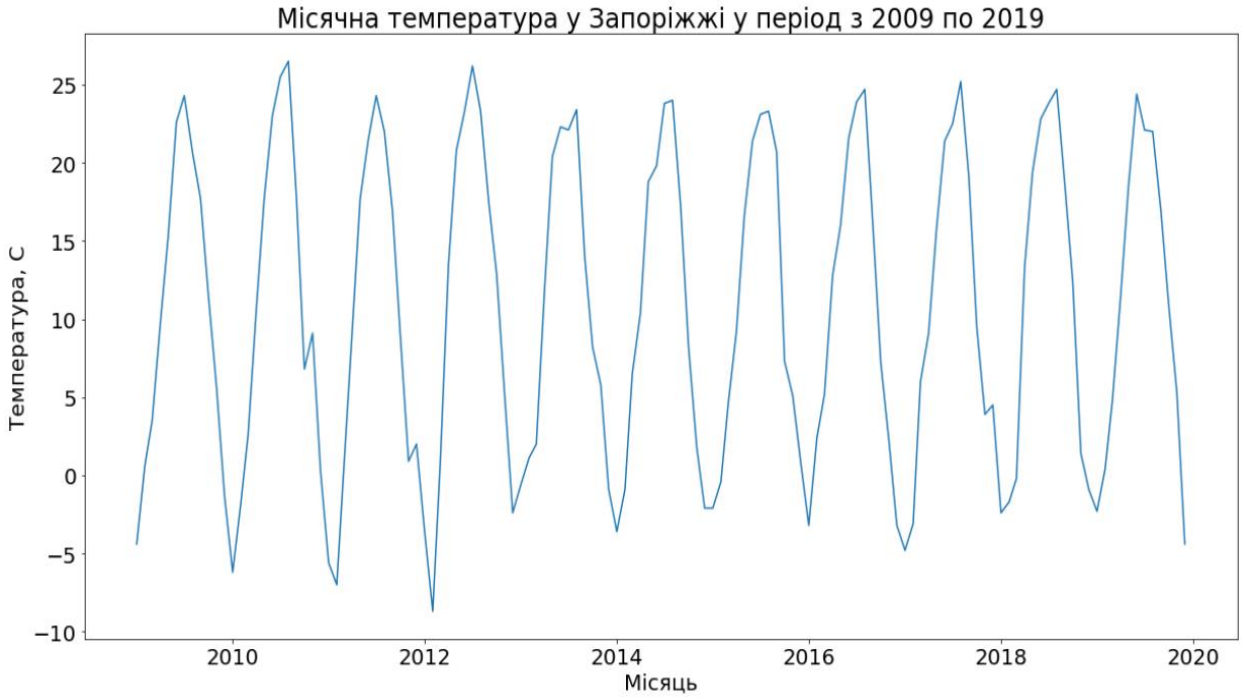

Рисунок 3.1 – Графік місячної температури міста Запоріжжя у період з 2009 по 2019 роки

3.2 Проведення аналізу побудованого часового ряду метеорологічних даних

Існують дві основні мети аналізу часових рядів: визначення природи ряду і прогнозування (передбачення майбутніх значень часового ряду по теперішнім і минулим значенням). Обидві ці цілі вимагають, щоб модель ряду була ідентифікована і, більш-менш, формально описана. Як тільки модель визначена, ви можете з її допомогою інтерпретувати дані, які розглядаються.

На основні побудованого часового ряду метеоданих проводиться аналіз.

Спочатку обчислюється базова модель (прогноз на день на значеннях попереднього дня) для визначення середньоквадратичної похибки (RMSE) даного часового ряду (див. рис. 3.2).

```
# обчислюємо базову модель(прогноз на день на знеченні попереднього дня)
from sklearn.metrics import mean_squared_error
from math import sqrt
# завантажуємо дані
series = read_csv('dataset.csv', header=None, index_col=0, parse_dates=True,
                  squeeze=True)
# готуємо дані
X = series.values
X = X.astype('float32')
train_size = int(len(X) * 0.50)
train, test = X[0:train_size], X[train_size:]
# покрокова перевірка
history = [x for x in train]
predictions = list()
for i in range(len(test)):
    # прогноз
    yhat = history[-1]
    predictions.append(yhat)
    # спостереженні
    obs = test[i]
    history.append(obs)
    print('>Predicted=%.3f, Expected=%.3f' % (yhat, obs))
# висновок - RMSE=корень середнє квадратичного відхилення
rmse = sqrt(mean_squared_error(test, predictions))
print('RMSE: %.3f' % rmse)
```

Рисунок 3.2 – Обчислення базової моделі

Та за отриманими даними обчислення, приходимо до висновку, що $RMSE = 5.622$. Це не велике значення похибки і ми можемо далі продовжувати наш аналіз.

Потім визначаються мінімальні та максимальні значення, медіана та середньоквадратичне відхилення.(див. рис. 3.3).

```
count    132.000000
mean      10.584091
std       10.037813
min       -8.700000
25%        1.050000
50%       10.050000
75%       20.700000
max       26.500000
Name: 1, dtype: float64
```

Рисунок 3.3 – Аналіз побудованого часового ряду

Далі проводиться перевірка ряду на щільність, для того, щоб переконатися, що не має пропущених значень та подивитися, як ці значення розподіленні (див. рис. 3.4).

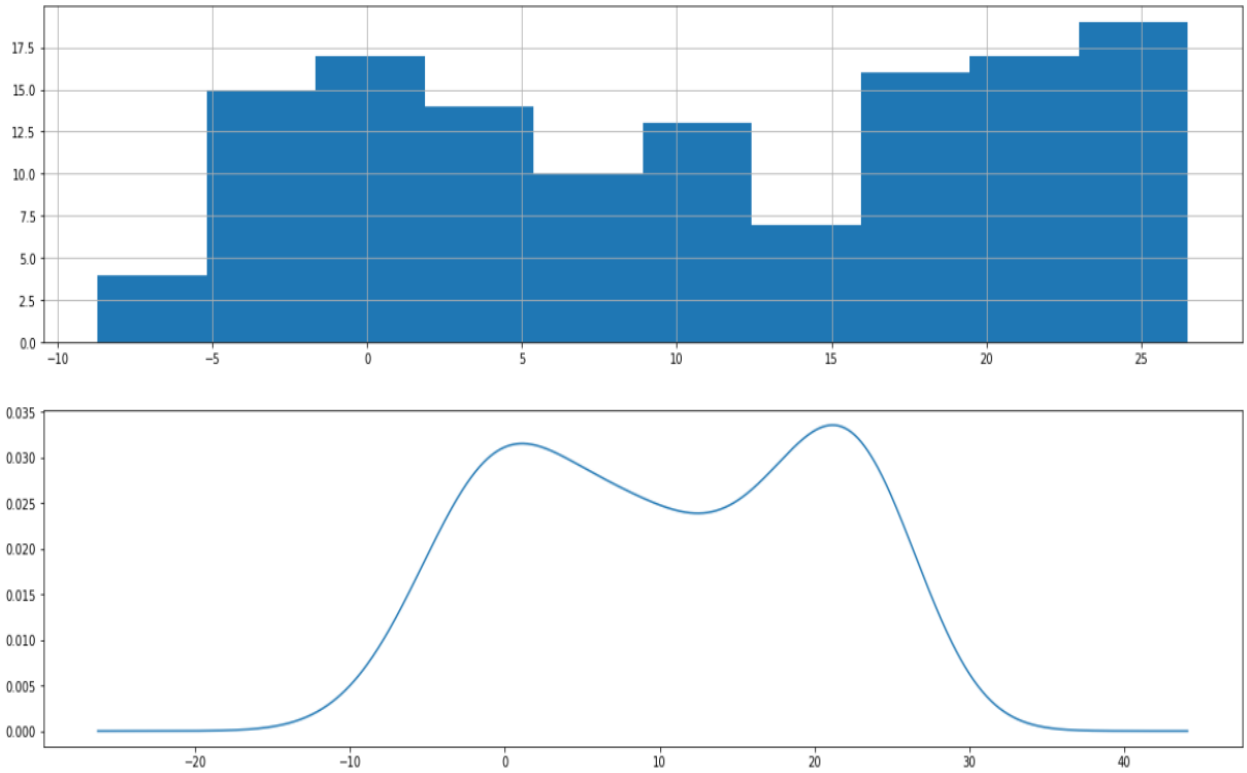

Рисунок 3.4 – Графік щільності ряду

Подивившись на графік (див. рис. 3.4), можна переконатися, що з даним рядом все добре та найбільш актуальні значення знаходяться на відмітці 0 та 25, і це є логічним.

Останнім кроком аналізу є розгляд розмаху часового ряду метеоданих по рокам (див. рис. 3.5).

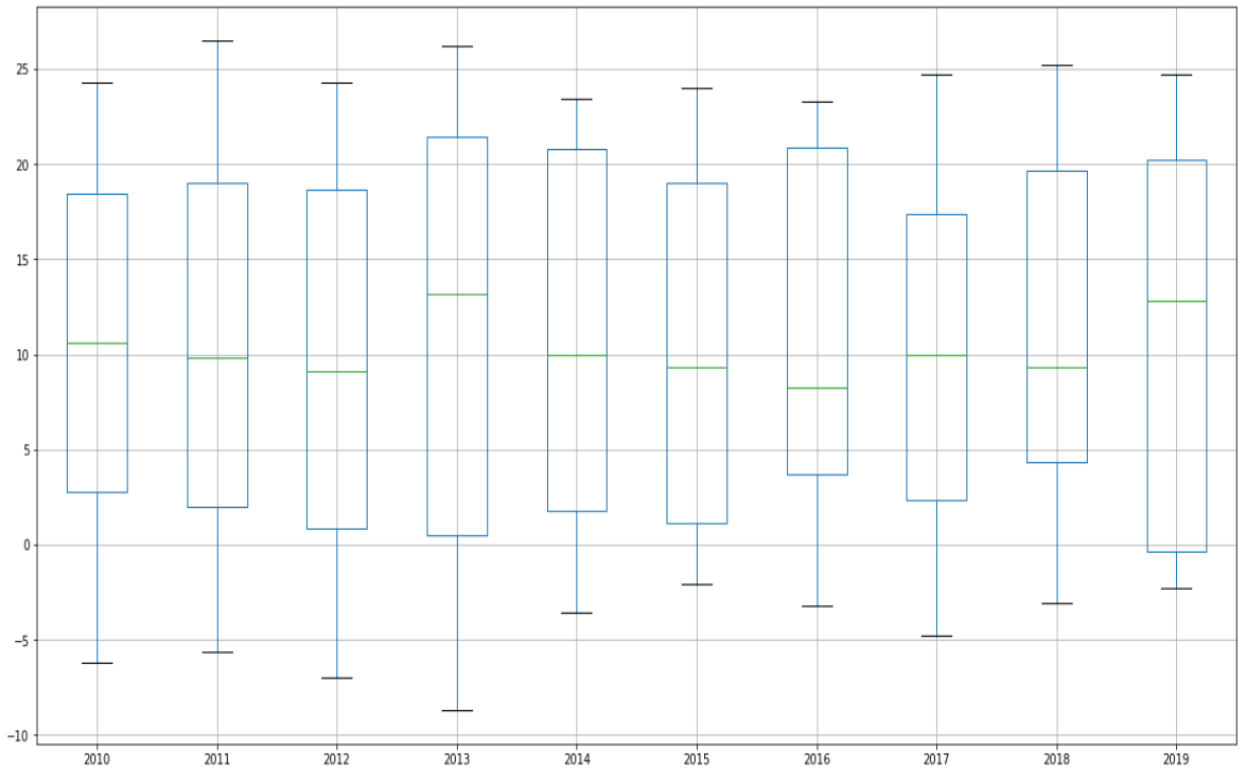

Рисунок 3.5 – Діаграма розмаху ряду по рокам

На діаграмі (див. рис. 3.5) можна побачити, що ми маємо 10 розмахів, що дорівнює 10 рокам, які розглядаються.

Провівши аналіз базової моделі ряду, ми можемо переходити до моделювання та прогнозування часового ряду метеоданих методами ARIMA та SARIMA.

3.3 Здійснення прогнозування за побудованим часовим рядом метеорологічних даних із використанням досліджених методів теорії часових рядів

Для побудови моделей прогнозування та проведення прогнозу, методами ARIMA та SARIMA, було реалізовано програму на мові програмування Python.

Для того, щоб провести прогнозування ряду було взято середні значення температури міста Запоріжжя по місяцям за минулі 10 років в період з 2009 – 2019 роки. Дані були взяті з архіву метеоданих міста Запоріжжя.

3.3.1 Метод прогнозування ARIMA

Для моделювання використовується модель ARIMA. Дана модель має загальний вигляд: $ARIMA(p, q, d)$. Для того, щоб побудувати модель нам потрібно знати її порядок , який складається з трьох параметрів : p – порядок компоненти AR; d – порядок інтегрування ряду; q – порядок компоненти MA.

Параметр d є і він дорівнює 0, тому що параметр d – це скільки разів нам треба інтегрувати ряд, щоб він став стаціонарним, а наш ряд – стаціонарний, тому, що стаціонарність означає, що значення залежать від часу.

Отже залишилося визначити p і q. Для їх визначення нам треба розглянути автокореляційну (ACF) і частково автокореляційну (PACF) функції для ряду перших різниць (див. рис. 3.6).

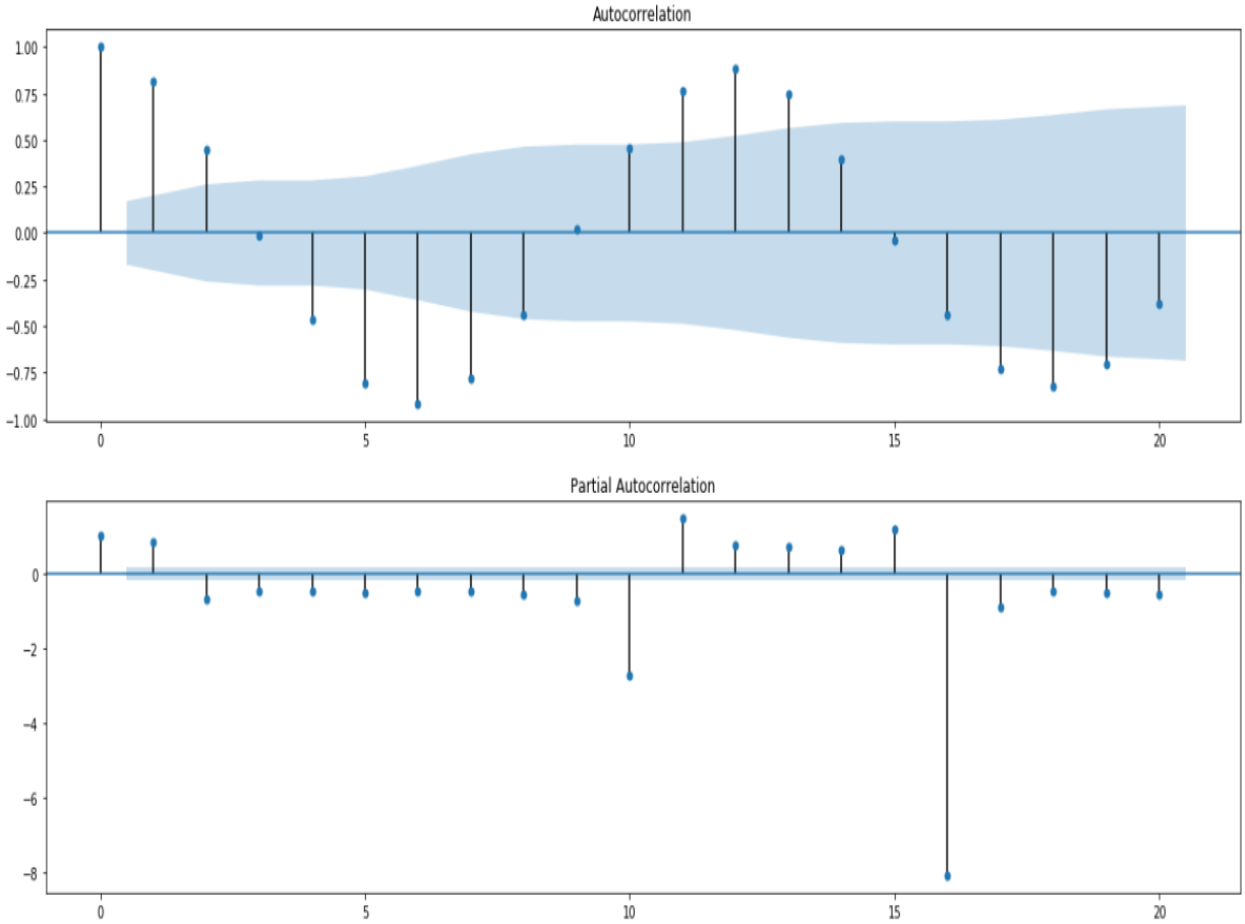

Рисунок 3.6 – Графік ACF та PACF

Розглядаючи графік (див. рис. 3.6) ACF можна зробити висновок, що p = 4, тому що на ньому 4 лаги сильно відмінний від нуля. За графіком PACF можна побачити, що q = 1, тому що на ньому 1 лаг сильно відмінний від нуля.

Отже зараз маємо параметри ARIMA(4, 0, 1).

Будуємо модель ARIMA за даними параметрами (див. рис. 3.7).

```
# створюємо ARIMA-модель на основі знайдених вище параметрів
from sklearn.metrics import mean_squared_error
from statsmodels.tsa.arima.model import ARIMA
from math import sqrt


# завантажуємо дані
series = read_csv('dataset.csv', header=None, index_col=0, parse_dates=True,
                  squeeze=True)
# готуємо дані
X = series.values
X = X.astype('float32')
train_size = int(len(X) * 0.50)
train, test = X[0:train_size], X[train_size:]
# покрокове прогнозування
history = [x for x in train]
predictions = list()
for i in range(len(test)):
    # прогноз
    model = ARIMA(history, order=(4,0,1))
    model_fit = model.fit()
    yhat = model_fit.forecast()[0]
    predictions.append(yhat)
    # спостеження
    obs = test[i]
    history.append(obs)
    print('>Прогноз=%.3f, Спостереження=%.3f' % (yhat, obs))
# висновок RMSE
rmse = sqrt(mean_squared_error(test, predictions))
print('RMSE: %.3f' % rmse)
```

Рисунок 3.7 – Побудова моделі ARIMA

```
>Прогноз=20.679, Спостереження=23.800
>Прогноз=22.382, Спостереження=24.000
>Прогноз=16.697, Спостереження=17.200
>Прогноз=9.968, Спостереження=8.200
>Прогноз=3.095, Спостереження=1.800
>Прогноз=-1.906, Спостереження=-2.100
>Прогноз=-3.513, Спостереження=-2.100
>Прогноз=0.042, Спостереження=-0.400
>Прогноз=4.943, Спостереження=4.600
>Прогноз=12.065, Спостереження=9.200
>Прогноз=16.602, Спостереження=16.600
>Прогноз=22.246, Спостереження=21.400
>Прогноз=22.893, Спостереження=23.100
>Прогноз=21.807, Спостереження=23.300
>Прогноз=18.361, Спостереження=20.700
>Прогноз=12.613, Спостереження=7.300
>Прогноз=1.605, Спостереження=5.100
>Прогноз=1.679, Спостереження=0.900
>Прогноз=-3.989, Спостереження=-3.200
>Прогноз=-1.217, Спостереження=2.400
>Прогноз=6.028, Спостереження=5.200
>Прогноз=9.170, Спостереження=12.800
>Прогноз=9.170, Спостереження=12.800
>Прогноз=18.071, Спостереження=16.000
>Прогноз=19.232, Спостереження=21.600
>Прогноз=22.420, Спостереження=23.900
>Прогноз=19.803, Спостереження=24.700
>Прогноз=17.532, Спостереження=16.000
>Прогноз=8.734, Спостереження=7.200
>Прогноз=2.536, Спостереження=2.200
>Прогноз=-2.251, Спостереження=-3.200
>Прогноз=-4.274, Спостереження=-4.800
>Прогноз=-0.659, Спостереження=-3.100
>Прогноз=4.275, Спостереження=6.000
>Прогноз=13.636, Спостереження=9.100
>Прогноз=17.362, Спостереження=15.800
>Прогноз=23.881, Спостереження=21.400
>Прогноз=24.005, Спостереження=22.500
>Прогноз=21.987, Спостереження=25.200
>Прогноз=20.189, Спостереження=19.100
>Прогноз=11.709, Спостереження=9.500
>Прогноз=4.654, Спостереження=3.900
>Прогноз=-0.583, Спостереження=4.500
>Прогноз=-0.659, Спостереження=-2.400
>Прогноз=-0.659, Спостереження=-2.400
>Прогноз=-2.374, Спостереження=-1.700
>Прогноз=4.121, Спостереження=-0.200
>Прогноз=6.345, Спостереження=13.300
>Прогноз=17.534, Спостереження=19.400
>Прогноз=20.803, Спостереження=22.800
>Прогноз=24.452, Спостереження=23.800
>Прогноз=21.393, Спостереження=24.700
>Прогноз=17.440, Спостереження=18.600
>Прогноз=9.635, Спостереження=12.400
>Прогноз=4.316, Спостереження=1.400
>Прогноз=-3.803, Спостереження=-0.900
>Прогноз=-3.316, Спостереження=-2.300
>Прогноз=-2.565, Спостереження=0.400
>Прогноз=4.286, Спостереження=4.800
>Прогноз=10.511, Спостереження=11.400
>Прогноз=17.146, Спостереження=18.500
>Прогноз=21.794, Спостереження=24.400
>Прогноз=23.942, Спостереження=22.100
>Прогноз=19.758, Спостереження=22.000
>Прогноз=16.608, Спостереження=17.000
>Прогноз=8.797, Спостереження=10.900
>Прогноз=3.852, Спостереження=5.300
>Прогноз=-0.296, Спостереження=-4.400
RMSE: 2.453
```

Рисунок 3.8 – ARIMA модель на основі знайдених параметрів

Середньоквадратична похибка – 2.453, це дуже гарний результат.

Залишається лише завантажити модель та зробити прогноз на п’ять років.

```
[-8.95563485  3.63865901  6.64597528 13.65604155 18.50850105 25.77761259
 24.09083256 17.12664138 14.32533385  7.15196141  1.89225198 -4.59365061
 -7.71192339  1.90328713  6.72620227 15.77008873 21.74475189 25.1305288
 25.38888917 25.06123372 11.22341824 -0.59262276  9.52738642 -3.17826244
 -8.01366313 -6.16919785  6.35234211 15.30649492 22.51519903 22.96158193
 24.27966598 19.10334632 12.46082041  3.52157405 -3.65156256  2.89445131
 -4.88465897 -9.75418227  7.26141896 21.51628753 24.74506666 23.4467023
 25.81368791 19.76846639 12.47818028  9.05598044  0.34029115 -5.99360859
  1.32808958  3.51513777  3.78301738 17.5804436  25.01225032 22.33089974
 20.33658799 22.28241133  7.62312702  3.98160136  4.3905548  -3.72125609]
```

Рисунок 3.9 – Отриманні значення прогнозу погоди на 2020-2024 роки

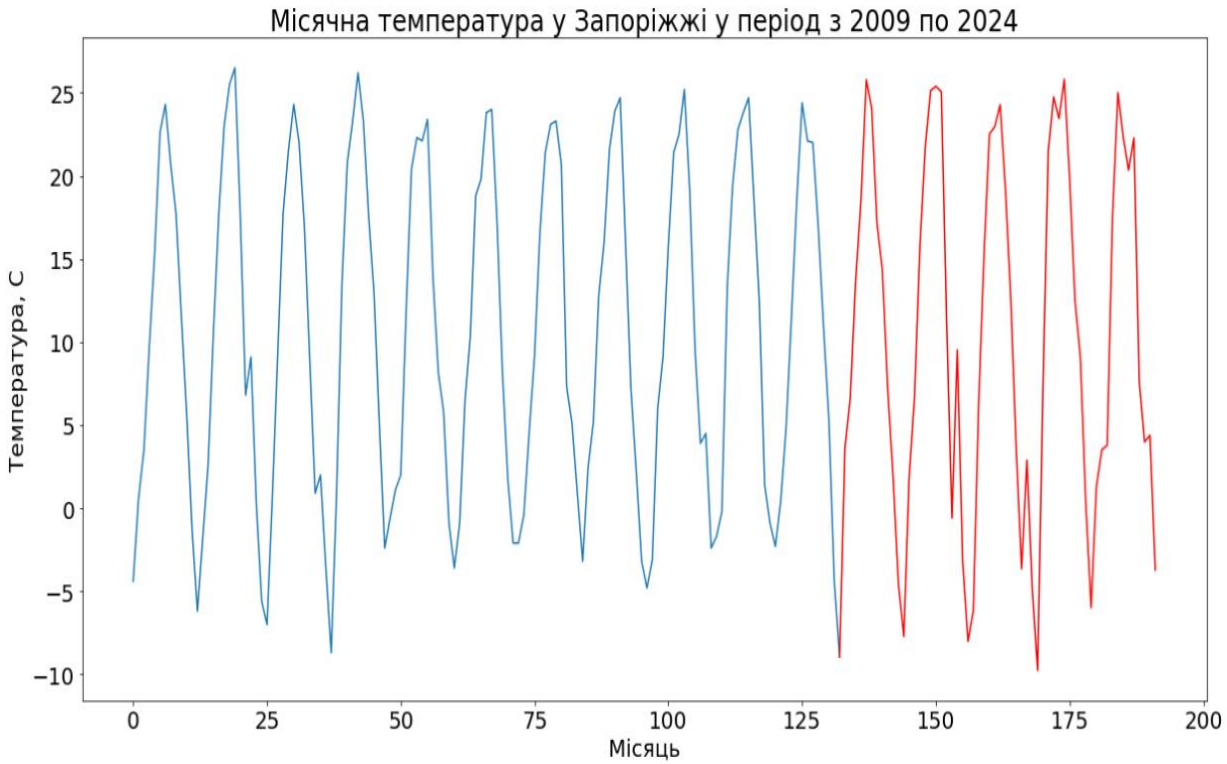

Рисунок 3.10 – Графік отриманих значень прогнозу

На графіку (див. рис. 3.10) зображена місячна температура у місті Запоріжжя у період з 2009 по 2024 рік. Синім виділено наші вхідні дані, а червоним – прогноз , який було проведено.

3.3.2 Метод прогнозування SARIMA

Для моделювання використовується модель SARIMA. Дана модель має загальний вигляд $ARIMA(p,q,d)(P,Q,D)m$. В цій моделі пармаетри означають: p – порядок компоненти AR; d – порядок інтегрування ряду; q – порядок компоненти MA; P – порядок сезонної компоненти SAR; Q – порядок сезонної компоненти SMA; D – порядок інтегрування сезонної компоненти; m – розмірність сезонності, в даному випадку місяць.

Для початку розглянемо такі властивості ряду, як тренд, сезонність та залишки (див. рис. 3.11).

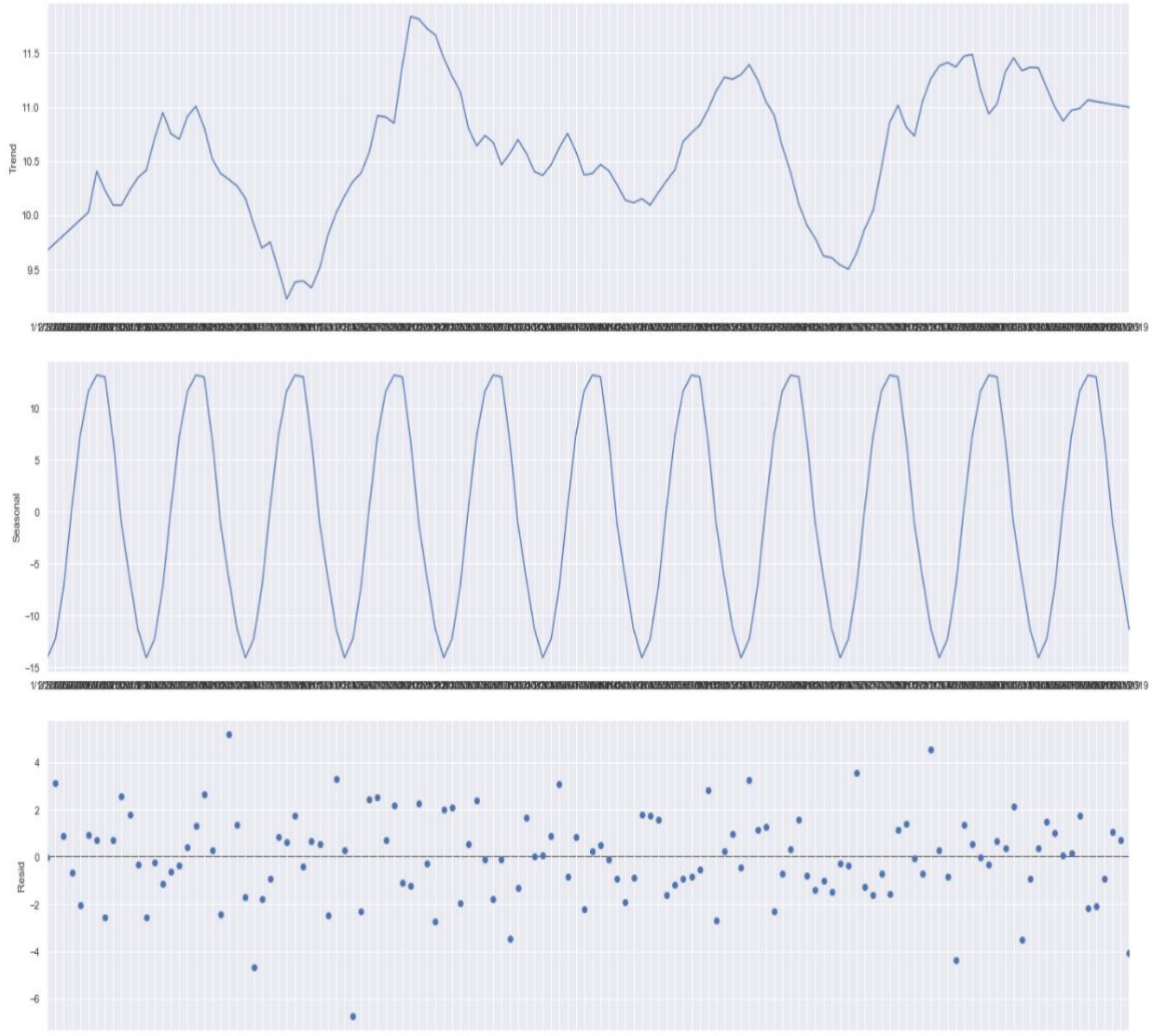

Рисунок 3.11 – Графік тренду, сезонності та залишків

На даному графіку (див. рис. 3.11) досить добре показано сезонність ряду.

Далі, як і в попередньому методі розглянемо не сезонні компоненти, але вже для SARIMA.

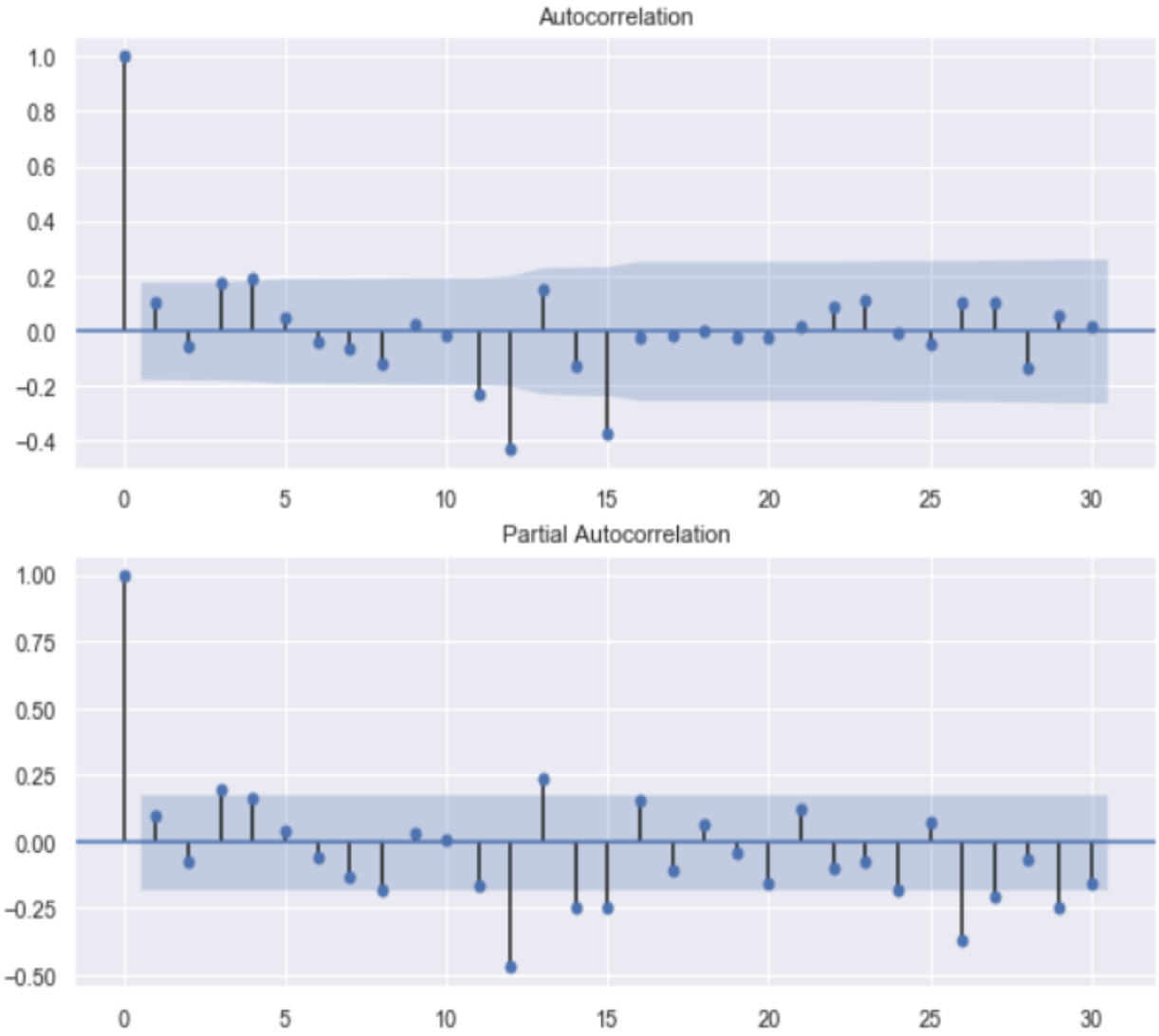

Рисунок 3.12 – Графік ACF та PACF для не сезонних параметрів

Розглядаючи графік (див. рис. 3.12) ACF можна зробити висновок, що p = 1, тому що на ній 1 лаги сильно відмінних від нуля. За графіком PACF можна побачити, що q = 1, тому що на ній 1 лаг сильно відмінний від нуля. Параметр d є і він дорівнює 0, тому що параметр d – це скільки разів нам треба

інтегрувати ряд, щоб він став стаціонарним, а наш ряд – стаціонарний, тому, що стаціонарність означає, що значення залежать від часу.

Наступним кроком вже розглянемо сезонні параметри (див. рис. 3.13).

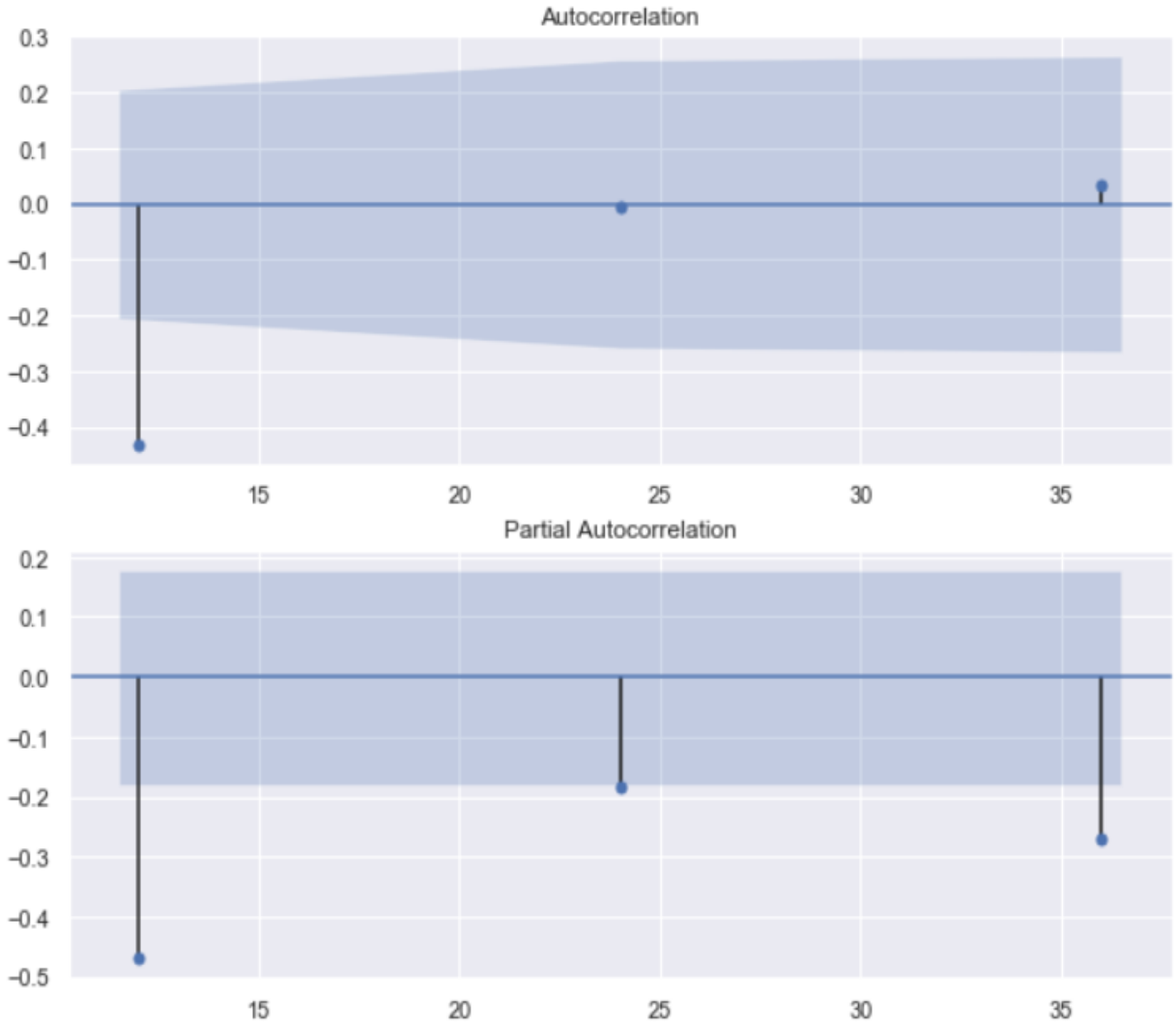

Рисунок 3.13 – Графік ACF та PACF для сезонних параметрів

Розглядаючи графік (див. рис. 3.13) ACF можна зробити висновок, що P = 1, тому що на ньому 1 лаг сильно відмінний від нуля. За графіком PACF можна побачити, що Q = 3, тому що на ньому 3 лаг сильно відмінний від нуля. Параметр D = 1.

Отже зараз дана модель має параметри ARIMA(1,0,1)(1,1,3)12.

Будуємо модель SARIMA за даними параметрами (див. рис. 3.14, 3.15).

```
# будуємо модель
model = SARIMAX(train,order=(1,0,1),seasonal_order=(1,1,3,12),trend='n')
results = model.fit()
```

Рисунок 3.14 – Побудова моделі SARIMA

Dep. Variable:	temp	**No. Observations:**	132
Model:	SARIMAX(1, 0, 1)x(1, 1, [1, 2, 3], 12)	**Log Likelihood**	-269.436
Date:	Sun, 13 Dec 2020	**AIC**	552.872
Time:	20:25:11	**BIC**	572.384
Sample:	01-01-2009	**HQIC**	560.796
	- 12-01-2019		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.2709	0.455	-0.596	0.551	-1.162	0.620
ma.L1	0.4747	0.416	1.141	0.254	-0.341	1.290
ar.S.L12	-0.6323	1.839	-0.344	0.731	-4.237	2.972
ma.S.L12	-0.3379	19.477	-0.017	0.986	-38.512	37.836
ma.S.L24	-0.6799	13.837	-0.049	0.961	-27.801	26.441
ma.S.L36	0.0201	0.317	0.064	0.949	-0.601	0.641
sigma2	4.1132	82.081	0.050	0.960	-156.762	164.988

Ljung-Box (L1) (Q):	0.06	**Jarque-Bera (JB):**	0.41
Prob(Q):	0.81	**Prob(JB):**	0.81
Heteroskedasticity (H):	0.76	**Skew:**	-0.10
Prob(H) (two-sided):	0.39	**Kurtosis:**	3.21

Рисунок 3.15 – Модель SARIMA

Далі обчислюємо середньоквадратичну похибку (RMSE) (див. рис. 3.16).

```
# обчислюємо RMSE
print(np.mean(np.abs(results.resid)))
```

```
2.6448937510273076
```

Рисунок 3.16 – Середньоквадратична похибка RMSE для моделі SARIMA

Останній крок, проведення прогнозування даного ряду на п’ять років уперед.

2020-01-01	-4.343888	2021-08-01	23.692764	2023-05-01	17.918141
2020-02-01	-1.527916	2021-09-01	17.479317	2023-06-01	22.161905
2020-03-01	3.476407	2021-10-01	9.442120	2023-07-01	23.813682
2020-04-01	10.820016	2021-11-01	4.051704	2023-08-01	23.640005
2020-05-01	17.809775	2021-12-01	-0.712217	2023-09-01	17.486840
2020-06-01	22.221728	2022-01-01	-3.553674	2023-10-01	9.435210
2020-07-01	23.656510	2022-02-01	-1.704560	2023-11-01	4.138970
2020-08-01	23.517976	2022-03-01	3.432948	2023-12-01	-0.704437
2020-09-01	17.524162	2022-04-01	10.920554	2024-01-01	-3.576827
2020-10-01	9.367739	2022-05-01	17.872095	2024-02-01	-1.713379
2020-11-01	4.387057	2022-06-01	22.228743	2024-03-01	3.371435
2020-12-01	-0.571485	2022-07-01	23.702631	2024-04-01	10.964906
2021-01-01	-3.629732	2022-08-01	23.549278	2024-05-01	17.889027
2021-02-01	-1.748865	2022-09-01	17.499776	2024-06-01	22.204166
2021-03-01	3.170102	2022-10-01	9.423326	2024-07-01	23.743465
2021-04-01	11.110847	2022-11-01	4.289028	2024-08-01	23.582639
2021-05-01	17.945046	2022-12-01	-0.691059	2024-09-01	17.495019
2021-06-01	22.123001	2023-01-01	-3.616639	2024-10-01	9.427696
2021-07-01	23.878274	2023-02-01	-1.728544	2024-11-01	4.233850
2021-08-01	23.692764	2023-03-01	3.265661	2024-12-01	-0.695978
		2023-04-01	11.041170		

Рисунок 3.17 – Значеня прогнозу метеоданих у період з 2020-2024 роки

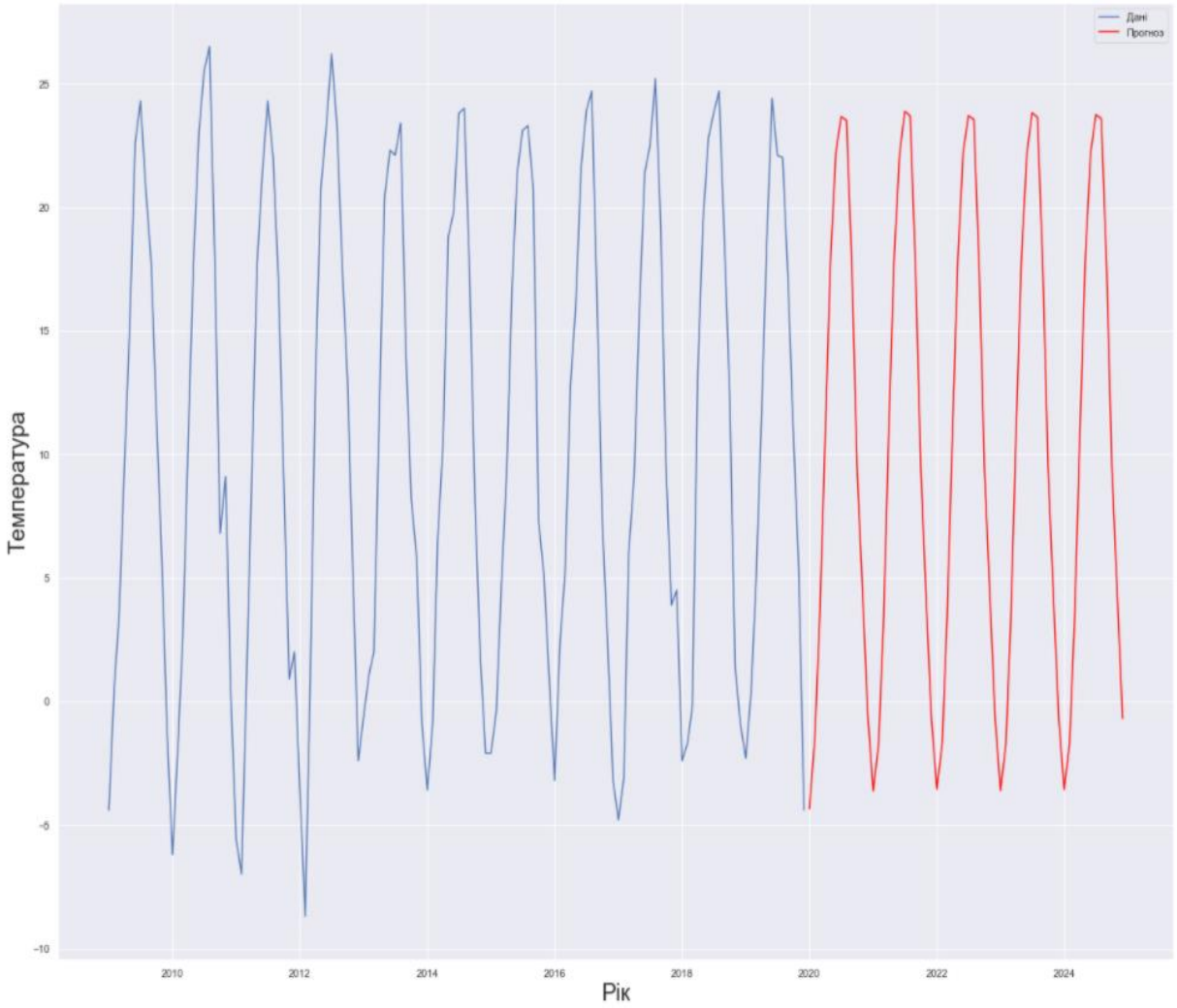

Рисунок 3.18 – Графік отриманих значень прогнозу

На графіку (див. рис. 3.18) зображена місячна температура у місті Запоріжжя у період з 2009 по 2024 рік. Синім виділено наші вхідні дані, а червоним – прогноз , який було проведено.

ВИСНОВКИ

У кваліфікаційній роботі було розглянуто методи аналізу та прогнозування часових рядів. Виконано основну задачу роботи, а саме виконано обробку метеорологічних даних , використовуючи методологію теорії часових рядів. Проведено статичний аналіз ряду : виявлено аномальні рівні ряду, перевірено на наявність тренду та сезонної компоненти. Побудовано моделі прогнозування ARIMA та SARIMA, зроблено прогноз погоди на 5 років вперед. Проаналізовано , яка модель прогнозування підходить більше для прогнозування метеоданих. Модель SARIMA буде більш доречною для прогнозування погоди.

Структурно робота складається з трьох розділів, кожний з яких повністю розкриває поставленні у роботі завдання.

У першому розділі було описано основні поняття та особливості моделювання багаторічних змін метеорологічних даних та загальні поняття теорії часових рядів. Проведено аналітичний огляд сучасного стану проблеми.

У другому розділі було розглянуто основні підходи проведення аналізу часових рядів, основні методи прогнозування часових рядів, їх сутність, етапи реалізації та умови використання.

У третьому розділі було проведено аналіз вхідного часового ряду. Побудовано дві моделі прогнозування ARIMA та SARIMA. За допомогою цих моделей було зроблено прогноз погоди міста Запоріжжя на 5 років вперед. Проаналізовано яка з моделей підходить більше для прогнозування метеоданих .

ПЕРЕЛІК ПОСИЛАНЬ

1. Афанасьев В. Н. , Юзбашев М. М. Анализ временных рядов и прогнозирования. Москва : Финансы и статистика, 2012. 320 с.

2. Бокс Дж., Дженкинс Г.М. Анализ временных рядов. Прогноз и управление / [пер. з англ. В. Ф. Писаренко]. Москва : Мир, 1974. 406 с.

3. Важнова Н. А., Верещагин М. А. О многолетней динамике приземного термического режима на территории приволжского федерального округа (ПФО) во второй половине XX и начале XXI века. *Вестник Удмуртского университета.* Биология. Науки о Земле. 2014. Вип. 1. С. 112-120.

4. Елисеева И. И. Эконометрика : учебник для магистров. Москва : Юрайт, 2014. 453 с.

5. Канторович Г.Г. Анализ временных рядов. М. : *Экономический журнал ВШЭ,* 2002. Вип. 4. С. 498-523.

6. Капитанова О. В. Прогнозирование социально-экономических процессов : Учебно-методическое пособие. Нижний Новгород : Нижегородской госуниверситет, 2016. 74 с.

7. Кизбикенов К. О. Прогнозирование и временные ряды : учебное пособие. Барнаул : АртГПУ, 2017. 113 с.

8. Лоскутов А. Ю. Анализ временных рядов. Курс лекций. Москва : Физический факультет МГУ, 2013. 113 с.

9. Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов. Москва : Финансы и статистика, 2003. 416 с.

10. Михалат С. Г., Мингалёв Д. Э., Евдокимов С. И. Использование анализа временных рядов в изучении многолетних температурных изменений. Псков : ПсковГУ, «ЛОГОС Плюс», 2014. 368 с.

11. Модель ARIMA - полное руководство по прогнозированию временных рядов в Python. 2019. URL : https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

12. Мхитарян В. С. Анализ данных : учебник для академического бакалавриата. Москва : Юрайт, 2017. 490 с.

13. Петросянц М.А. Прогноз погоды: состояние и ближайшие задачи. *Метеорология и гидрология*, 1981. Вип. 6. С. 12-31.

14. Пузаченко Ю. Г. Математические методы в экологических и географических исследованиях: Учебное пособие для студентов вузов. Москва : Центр «Академия», 2004. 416 с.

15. Снитковский А.И. К прогнозу температуры воздуха. *Метеорология и гидрология*, 1980. Вип. 12. С. 98.

16. Татаренко С. И. Методы и модели анализа временных рядов : метод. указания к лаб. работам. Тамбов : Тамбовский государственний технический университет, 2008. 32 с.

17. Чуева И. Характеристики прогнозируемых рядов. 2011. URL : https://www.mbureau.ru/blog/harakteristiki-prognoziruemyh-vremennyh-ryadov

18. Шерстюков Б. Г., Салугашвили Р. С. Новые тенденции в изменениях климата Северного полушария Земли в последнее десятилетие. Труды ГУ ВНИИГМИ-МЦД, 2010. Вип. 175. С. 43-51.

19. Шугунов Л.Ж. Исследование и анализ среднегодовой температуры на основе методов спектрального анализа и классической декомпозиции. Изв. вузов. Сев-Кав. регион. Естест. науки. Приложение, 2006. Вип. 1. С. 83-88.

20. Юрченко М.Є. Прогнозування та аналіз часових рядів. Методичні вказівки до практичних занять та самостійної роботи студентів. Чернігів: ЧНТУ, 2018. 88 с.

21. Brownlee J. A Gentle Introduction to SARIMA for Time Series Forecasting in Python. 2018.

22. Crowley T. J. Causes of climate change over the past 1000 years. 2000.

23. Mahmut Firat, Fatih Dikbas, A. Cem Koc, Mahmud Gungor. Analysis of temperature series: estimation of missing data and homogeneity test. Meteorological Applications. 2012.

24. Вхідні дані для проведення практичного завдання. Архів погоди м. Запоріжжя. URL : https://rp5.ua/%D0%90%D1%80%D1%85%D0%B8%D0%B2_%D0%BF%D0%BE%D0%B3%D0%BE%D0%B4%D1%8B_%D0%B2_%D0%97%D0%B0%D0%BF%D0%BE%D1%80%D0%BE%D0%B6%D1%8C%D0%B5_(%D0%B0%D1%8D%D1%80%D0%BE%D0%BF%D0%BE%D1%80%D1%82) .

Додаток А

Реалізація програмного продукту для побудови моделі ARIMA та прогнозування

```
# підключаємо дані
from pandas import read_csv
series = read_csv('data.csv', header=0, index_col=0, parse_dates=True, squeeze=True)
split_point = len(series) - 60
dataset, validation = series[0:split_point], series[split_point:]
print('Dataset %d, Validation %d' % (len(dataset), len(validation)))
dataset.to_csv('dataset.csv', header=False)
validation.to_csv('validation.csv', header=False)
# обчислюємо базову модель(прогноз на день на знеченні попереднього дня)
from sklearn.metrics import mean_squared_error
from math import sqrt
# завантажуємо дані
series = read_csv('dataset.csv', header=None, index_col=0, parse_dates=True, squeeze=True)
# готуємо дані
X = series.values
X = X.astype('float32')
train_size = int(len(X) * 0.50)
train, test = X[0:train_size], X[train_size:]
# покрокова пеервірка
history = [x for x in train]
predictions = list()
for i in range(len(test)):
```

```
        # прогноз
        yhat = history[-1]
        predictions.append(yhat)
        # спостереженні
        obs = test[i]
        history.append(obs)
        print('>Predicted=%.3f, Expected=%.3f' % (yhat, obs))
        # висновок - RMSE=корень середнє квадратичного відхилення
        rmse = sqrt(mean_squared_error(test, predictions))
        print('RMSE: %.3f' % rmse)
        # аналіз часового ряду
        series = read_csv('dataset.csv', header=None, index_col=0, parse_dates=True,
squeeze=True)
        print(series.describe())
        # графік ряду
        from matplotlib import pyplot
        series = read_csv('dataset.csv', header=None, index_col=0, parse_dates=True,
squeeze=True)
        pyplot.figure(figsize=(20,10))
        pyplot.rcParams.update({'font.size': 20})
        pyplot.plot(series)
        pyplot.xlabel('Місяць')
        pyplot.ylabel('Температура, С')
        pyplot.title('Місячна температура у Запоріжжі у період з 2009 по 2019')
        pyplot.show()
        # графік щільності ряду
        series = read_csv('dataset.csv', header=None, index_col=0, parse_dates=True,
squeeze=True)
        pyplot.figure(figsize=(20,10))
        pyplot.figure(1)
```

```
pyplot.subplot(211)
series.hist()
pyplot.subplot(212)
series.plot(kind='kde')
pyplot.show()
# Діаграма розмаху ряду по рокам
from pandas import DataFrame
from pandas import Grouper
series = read_csv('dataset.csv', header=None, index_col=0, parse_dates=True, squeeze=True)
groups = series.groupby(Grouper(freq='12M'))
decades = DataFrame()
for name, group in groups:
if len(group.values) is 12:
decades[name.year] = group.values
pyplot.figure(figsize=(20,10))
decades.boxplot()
pyplot.show()
# Графіки кореляції та автокореляції(для визначення p та q)
from statsmodels.graphics.tsaplots import plot_acf
from statsmodels.graphics.tsaplots import plot_pacf
series = read_csv('dataset.csv', header=None, index_col=0, parse_dates=True, squeeze=True)
pyplot.figure(figsize=(20,10))
pyplot.subplot(211)
plot_acf(series, lags=20, ax=pyplot.gca())
pyplot.subplot(212)
plot_pacf(series, lags=20, ax=pyplot.gca())
pyplot.show()
# створюємо ARIMA-модель на основі знайдених вище параметрів
```

```
from sklearn.metrics import mean_squared_error
from statsmodels.tsa.arima.model import ARIMA
from math import sqrt
# завантажуємо дані
series = read_csv('dataset.csv', header=None, index_col=0, parse_dates=True, squeeze=True)
# готуємо дані
X = series.values
X = X.astype('float32')
train_size = int(len(X) * 0.50)
train, test = X[0:train_size], X[train_size:]
# покрокове прогнозування
history = [x for x in train]
predictions = list()
for i in range(len(test)):
# прогноз
model = ARIMA(history, order=(4,0,1))
model_fit = model.fit()
yhat = model_fit.forecast()[0]
predictions.append(yhat)
# спостереження
obs = test[i]
history.append(obs)
print('>Прогноз=%.3f, Спостереження=%.3f' % (yhat, obs))
# висновок RMSE
rmse = sqrt(mean_squared_error(test, predictions))
print('RMSE: %.3f' % rmse)
# зберігаємо побудовану модель
import numpy
```

```
def __getnewargs__(self):
return ((self.endog),(self.k_lags, self.k_diff, self.k_ma))
ARIMA.__getnewargs__ = __getnewargs__
series = read_csv('dataset.csv', header=None, index_col=0, parse_dates=True, squeeze=True)
X = series.values
X = X.astype('float32')
# модель
model = ARIMA(X, order=(4,0,1))
model_fit = model.fit(trend='nc', disp=0)
# похибка прогнозування
bias = 1.37
5751
# зберігаємо модель
model_fit.save('model.pkl')
numpy.save('model_bias.npy', [bias])
# завантажуємо модель та робимо прогноз на 5 років
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.arima_model import ARIMAResults
# дані
dataset = read_csv('dataset.csv', header=None, index_col=0, parse_dates=True, squeeze=True)
X = dataset.values.astype('float32')
history = [x for x in X]
validation = read_csv('validation.csv', header=None, index_col=0, parse_dates=True,squeeze=True)
y = validation.values.astype('float32')
# модель
model_fit = ARIMAResults.load('model.pkl')
```

```
bias = numpy.load('model_bias.npy')
# перший прогноз
predictions = list()
yhat = bias + float(model_fit.forecast()[0])
predictions.append(yhat)
history.append(y[0])
#print('>Predicted=%.3f, Expected=%.3f' % (yhat, y[0]))
# подальший прогноз
for i in range(1, len(y)):
# predict
model = ARIMA(history, order=(3,0,0))
model_fit = model.fit(trend='nc', disp=0)
yhat = bias + float(model_fit.forecast()[0])
predictions.append(yhat)
# observation
obs = y[i]
history.append(obs)
#print('>Predicted=%.3f, Expected=%.3f' % (yhat, obs))
# висновки
rmse = sqrt(mean_squared_error(y, predictions))
print('RMSE: %.3f' % rmse)
#графік
#pyplot.figure(figsize=(20,10))
#pyplot.plot(y)
#pyplot.plot(predictions, color='red')
#pyplot.show()
list = []
for l in predictions:
list.append(float(l))
forecast = np.array(list)
```

```
res =  np.concatenate([X, forecast])
print(forecast)
ax = np.linspace(0,191, num=192)
ay = res
x0 = 132
plt.figure(figsize=(20,10))
plt.rcParams.update({'font.size': 20})
plt.plot(ax[:x0+1], ay[:x0+1],label='Спостереження' )
plt.plot(ax[x0:], ay[x0:], color="red", label='Прогноз')
plt.xlabel('Місяць')
plt.ylabel('Температура, С')
plt.title('Місячна температура у Запоріжжі у період з 2009 по 2024')
plt.show()
```

Додаток Б

Реалізація програмного продукту для побудови моделі SARIMA та прогнозування

```
import numpy as np
import matplotlib.pyplot as plt
import pathlib
import os
import seaborn as sns
import pandas as pd
from datetime import datetime
import statsmodels.api as sm
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_acf,plot_pacf
from statsmodels.tsa.statespace.sarimax import SARIMAX
sns.set()
data = pd.read_csv('dataset.csv')
data.head()
data = data.set_index('month',drop=True)
fig, ax = plt.subplots()
data['temp'].plot(figsize=(15,12),ax=ax)
plt.rcParams.update({'font.size': 20})
ax.set_xlabel('Місяць')
ax.set_ylabel('Температура, С')
ax.set_title('Місячна температура у Запоріжжі у період з 2009 по 2019')
plt.show()
def plotBoxNdendity(data,col=None):
if col in data.columns:
plt.figure(figsize=(18,8))
```

```
plt.rcParams.update({'font.size': 20})
ax1 = plt.subplot(121)
data.boxplot(col,ax=ax1)
ax1.set_ylabel('Рівні температури', fontsize=10)
ax2 = plt.subplot(122)
data[col].plot(ax=ax2,legend=True,kind='density')
ax2.set_ylabel('Розподілення температури', fontsize=10)
else:
print("Column not in the data")
plotBoxNdendity(data,'temp')
train = data[:'12/1/2019']
test = data['12/1/2019':]
# візуалізуємо такі властивость ряду як тренд, сезонність та залишкі
def decomposeNplot(data):
decomposition = sm.tsa.seasonal_decompose(data)
plt.figure(figsize=(15,16))
ax1 = plt.subplot(411)
decomposition.observed.plot(ax=ax1)
ax1.set_ylabel('Дані')
ax2 = plt.subplot(412)
decomposition.trend.plot(ax=ax2)
ax2.set_ylabel('Тренд')
ax3 = plt.subplot(413)
decomposition.seasonal.plot(ax=ax3)
ax3.set_ylabel('Сезонність')
ax4 = plt.subplot(414)
decomposition.resid.plot(ax=ax4)
ax4.set_ylabel('Залишки')
return decomposition
from statsmodels.tsa.seasonal import seasonal_decompose
```

```
#адитивна декомпозиція
result_add = seasonal_decompose(data.temp, model='additive', extrapolate_trend='freq', period=12)
# графік
plt.rcParams.update({'figure.figsize': (20,20)})
result_add.plot().suptitle('Аддітивна декомпозиція', fontsize=16)
plt.show()
# перевірка на стаціонарність(розширений Дікі-Фулер)
results = adfuller(train.diff(12).dropna())
results
# не сезонні параметри для SARIMAX
plt.figure(figsize=(10,8))
plt.rcParams.update({'font.size': 20})
ax1 = plt.subplot(211)
acf = plot_acf(train.diff(12).dropna(),lags=30,ax=ax1)
ax2 = plt.subplot(212)
pacf = plot_pacf(train.diff(12).dropna(),lags=30,ax=ax2)
# сезонні порядки для SARIMAX  за допомогою ACF & PACF частковими лагами
lags = [12*i for i in range(1,4)]
plt.figure(figsize=(10,8))
ax1 = plt.subplot(211)
acf = plot_acf(train.diff(12).dropna(),lags=lags,ax=ax1)
ax2 = plt.subplot(212)
pacf = plot_pacf(train.diff(12).dropna(),lags=lags,ax=ax2)
# будуємо модель
model = SARIMAX(train,order=(1,0,1),seasonal_order=(1,1,3,12),trend='n')
results = model.fit()
results.summary()
# обчислюємо RMSE
```

```
print(np.mean(np.abs(results.resid)))
# діагностика моделі
diagnostics = results.plot_diagnostics(figsize=(20,20))
# робимо прогноз за допомогою побудованної моделі
forecast = results.get_forecast(steps=60)
predictedmean = forecast.predicted_mean
bounds = forecast.conf_int()
lower_limit = bounds.iloc[:,0]
upper_limit = bounds.iloc[:,1]
print(predictedmean)
# будуємо графік прогнозу
plt.figure(figsize=(22,18))
plt.plot(pd.to_datetime(train.index), train, label='Дані')
plt.plot(predictedmean.index, predictedmean, color='red', label='Прогноз')
plt.xlabel('Рік',fontsize=24)
plt.ylabel('Температура',fontsize=24)
ax.set_title('Місячна температура у Запоріжжі у період з 2009 по 2024')
plt.legend()
plt.show()
```