МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ВАСИЛЯ СТУСА
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ І ПРИКЛАДНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

ЮРЧУК ДМИТРО МИКОЛАЙОВИЧ

Допускається до захисту:
в.о. завідувача кафедри
інформаційних технологій,
д-р. техн. наук, професор
_____Наталія ВЕСЕЛОВСЬКА
«____»_____ 2026 р.

РЕАЛІЗАЦІЯ АЛГОРИТМІВ СТАТИСТИЧНОГО НАВЧАННЯ ДЛЯ АНАЛІЗУ МЕТЕОРОЛОГІЧНИХ ДАНИХ

Спеціальність 122 Комп’ютерні науки

Кваліфікаційна (бакалаврська) робота

Керівник:
Надія ПОТАПОВА, доцент кафедри
інформаційних технологій,
к. е. н., доцент

_____

Оцінка: _____ / _____ / _____
(бали за шкалою ЄКТС / за національною шкалою)

Голова ЕК: _____
(підпис)

Вінниця - 2026

АНОТАЦІЯ

Юрчук Д.М. Реалізація алгоритмів статистичного навчання для аналізу метеорологічних даних. Спеціальність 122 «Комп'ютерні науки», освітня програма «Комп'ютерні науки». Донецький національний університет імені Василя Стуса, Вінниця, 2026.

У кваліфікаційній (бакалаврській) роботі досліджено та реалізовано алгоритми статистичного навчання для аналізу метеорологічних даних з відкритих джерел (NOAA, Meteostat). Проведено попередню обробку даних та застосовано кластеризацію методом k-середніх. Виконано кореляційний аналіз, побудовано лінійну регресію для прогнозування середньодобової температури на основі мінімальної, максимальної температури, опадів, швидкості вітру та атмосферного тиску. Розроблено прогноз на майбутні періоди. Програмна реалізація виконана мовою R у середовищі RStudio.

Ключові слова: статистичне навчання, метеорологічні дані, кластеризація, k-середніх, кореляційний аналіз, лінійна регресія, мова R.

55 ст., 13 рис., 7 табл., 1 дод., 42 джерела.

ABSTRACT

Yurchuk D.M. Implementation of Statistical Learning Algorithms for Analysis of Meteorological Data. Specialty 122 «Computer Science», educational program «Computer Science». Vasyl Stus Donetsk National University, Vinnytsia, 2026.

In the qualification (bachelor's) investigates and implements statistical learning algorithms for analyzing meteorological data from open sources (NOAA, Meteostat). The data underwent preprocessing, and k-means clustering was applied. A correlation analysis was performed, and a linear regression model was built to forecast the average daily temperature based on minimum and maximum temperatures, precipitation, wind speed, and atmospheric pressure. A forecast for future periods was developed. The software implementation was carried out in the R language within the RStudio environment.

Keywords: statistical learning, meteorological data, clustering, k-means, correlation analysis, multiple linear regression, R language.

ЗМІСТ

ВСТУП

У сучасних умовах швидкого розвитку інформаційних технологій та зростання обсягів даних особливої актуальності набувають методи аналізу даних та статистичного навчання. Значні обсяги метеорологічних даних, що накопичуються у відкритих міжнародних базах даних, потребують ефективних методів обробки, аналізу та інтерпретації. Використання алгоритмів статистичного навчання дозволяє виявляти закономірності у часових рядах метеорологічних показників, оцінювати їх взаємозв'язки та формувати прогностичні моделі. У зв'язку з цим *актуальним* є дослідження методів статистичного навчання для аналізу метеорологічних даних та реалізація відповідних алгоритмів у вигляді програмної реалізації.

Метою роботи є реалізація алгоритмів статистичного навчання для аналізу та оцінки метеорологічних даних, отриманих із відкритих джерел.

Для досягнення поставленої мети необхідно вирішити такі *завдання*:

1. Дослідити теоретичні основи використання підходів статистичного навчання для аналізу метереологічних даних.
2. Проаналізувати джерела відкритих метеорологічних даних.
3. Сформувати набір даних для проведення аналізу метеорологічних даних з урахуванням підготовки та попередню обробки.
4. Реалізувати алгоритми статистичного навчання для аналізу метеорологічних показників.
5. Розробити програмний продукт мовою програмування R у середовищі RStudio для обробки та аналізу даних.
6. Провести оцінку отриманих результатів.

Об'єктом дослідження є процес використання методик статистичного навчання для аналізу метеорологічних даних.

Предметом дослідження є підходи, методи та алгоритми статистичного навчання, що застосовуються для оцінки та аналізу метеорологічних даних.

У дослідженні *використано методи* статистичного аналізу даних, кореляційний аналіз, методи багатофакторної регресії, методи аналізу часових рядів, а також інструменти статистичного програмування мовою R.

Робота складається зі вступу, трьох розділів, висновків та списку використаних джерел. У першому розділі розглядаються теоретичні основи аналізу даних та статистичного навчання. У другому розділі проводиться аналіз та підготовка метеорологічних даних. Третій розділ присвячений реалізації алгоритмів статистичного навчання мовою програмування R та аналізу отриманих результатів.

Практичне значення роботи полягає у створенні програмного продукту, який дозволяє виконувати автоматизований аналіз метеорологічних даних із відкритих джерел, здійснювати їх обробку, статистичне моделювання та візуалізацію результатів.

Апробація результатів дослідження була представлена у вигляді матеріалів на VII Всеукраїнській науково-практичній конференції здобувачів вищої освіти та молодих вчених «Прикладні інформаційні технології» (м. Вінниця, 2026).

РОЗДІЛ 1
ТЕОРЕТИЧНІ ОСНОВИ ВИКОРИСТАННЯ ПІДХОДІВ СТАТИСТИЧНОГО НАВЧАННЯ В АНАЛІЗІ ДАНИХ

1.1. Поняття та сутність аналізу даних

Аналіз даних – це цілісна, багатоступенева діяльність, що охоплює збір, верифікацію, очищення, перетворення, побудову моделей та тлумачення інформації. Її кінцевою метою виступає формування достовірних і змістовних висновків, розкриття прихованих регулярностей, неочевидних взаємозв'язків та динаміки, а також забезпечення обґрунтованого ухвалення рішень в умовах неповноти даних і невизначеності [1, 7]. У сучасному розумінні аналіз даних виходить за межі простого набору технічних операцій; він трактується як інтегративна дослідницька парадигма, яка об'єднує принципи статистичного висновування, теорії ймовірностей, інформатики та предметного знання [2, 4].

Процес аналізу даних не обмежується формальними обчисленнями, а реалізується у вигляді замкненого циклу дій. Першим кроком є постановка дослідницького завдання та визначення релевантних джерел даних [5]. Далі формується репрезентативна вибірка з генеральної сукупності - від того, наскільки коректно вона відображає властивості всієї сукупності, прямо залежить якість отриманих результатів. Після цього виконується попередня обробка, яка передбачає перевірку даних на достовірність та несуперечність, виправлення систематичних і випадкових помилок вимірювань, ідентифікацію та належне опрацювання аномальних значень, а також заповнення пропущених спостережень із застосуванням статистичних методів – заміни на середнє або медіану, регресійного відновлення, використання методу найближчих сусідів. У виправданих випадках допускається видалення неповних записів [7, 11]. Лише після очищення та стандартизації даних можливе їх перетворення у формат, придатний для аналітичних методів, що нерідко вимагає нормалізації, стандартизації, дискретизації або створення нових похідних ознак.

На практиці аналіз даних реалізується через систему різноспрямованих підходів, які відрізняються за метою та глибиною проникнення в сутність досліджуваних явищ. Описова статистика дає змогу узагальнювати дані через обчислення заходів центральної тенденції, мінливості та характеристик форми розподілу; результати зазвичай подаються в наочній формі, супроводжуючись візуалізаціями [3, 8]. Дослідницький аналіз дани орієнтований на виявлення нових, раніше невідомих закономірностей, формування гіпотез та пошук неочікуваних структур без апріорного нав'язування параметричної моделі, що відповідає індуктивній логіці наукового пізнання [8]. На противагу йому, підтверджувальний аналіз даних використовується для формальної статистичної перевірки апріорно висунутих гіпотез, оцінки значущості виявлених ефектів та верифікації якості побудованих моделей за допомогою критеріїв перевірки гіпотез та довірчих інтервалів [1, 5].

Прогнозна аналітика спрямована на передбачення майбутніх значень показників або неспостережуваних станів, ґрунтуючись на історичних даних. Вона охоплює широкий спектр статистичних моделей та алгоритмів машинного навчання – від класичної лінійної регресії й аналізу часових рядів до складніших непараметричних методів [2, 6]. Важливими аспектами прогнозного аналізу є розрізнення між точковим та інтервальним прогнозуванням, а також проблема компромісу між зміщенням та дисперсією, яка породжує ризик перенавчання моделі на шумових компонентах даних [1, 4]. Окремим напрямом виступають методи інтелектуального аналізу даних, що дозволяють автоматизовано отримувати знання з великих масивів слабкоструктурованих або неструктурованих даних. До них належать кластеризація, пошук асоціативних правил, аналіз послідовностей та виявлення аномалій, які базуються як на принципах класичного статистичного навчання, так і на сучасних підходах штучного інтелекту [3, 7].

Аналіз даних виступає не просто як набір окремих технік, а як цілісна методологія, що органічно поєднує етапи збору, підготовки, перетворення,

моделювання, оцінювання та змістовної інтерпретації інформації. Завдяки своїй універсальності та адаптивності він є фундаментальним інструментом сучасного наукового дослідження, знаходячи застосування в багатьох сферах - від природничих і соціально-економічних наук до інженерної практики, біомедицини, маркетингу та управлінської діяльності. Скрізь, де прийняття обґрунтованих рішень має спиратися на емпіричні докази, а не на інтуїцію, аналіз даних відіграє ключову роль [2, 5, 11].

1.2. Основні принципи та підходи статистичного навчання

Статистичне навчання позиціонується як міждисциплінарний напрям, що поєднує інструментарій математичної статистики, теорію ймовірностей та алгоритмічні засади машинного навчання. Його головна мета - будувати моделі, здатні розпізнавати регулярності в даних, виконувати передбачення та сприяти ухваленню рішень за умов неповноти інформації [2, 4]. Як зауважують Х. Гасті, Р. Тібширані та Дж. Фрідман, принципова відмінність статистичного навчання від класичної статистики полягає у зміщенні акцентів: традиційна статистика зосереджена на висновках про генеральну сукупність, тоді як статистичне навчання робить акцент на передбачувальній точності та здатності моделі до узагальнення на нові, раніше не спостережувані дані [2].

В основі теоретичного підґрунтя статистичного навчання лежить так званий компроміс між зміщенням та дисперсією. Це фундаментальне обмеження визначає граничні можливості будь-якої моделі й не може бути подолане навіть при необмеженому зростанні обсягу вибірки [1, 5]. Зміщення виникає через надмірне спрощення моделі, яке не дає змоги адекватно відобразити істинну залежність між змінними. Моделі з високим зміщенням є мало гнучкими, але стійкими до випадкових коливань. Дисперсія, своєю чергою, відображає чутливість моделі до випадкових флуктуацій у навчальній вибірці; надто гнучка модель може «вивчати» шум, що призводить до різкої зміни прогнозів при незначних варіаціях вхідних даних. Оптимальна модель

досягається в точці рівноваги, де сумарна помилка є мінімальною [2, 6].

На відміну від класичного статистичного аналізу, який часто зосереджується лише на описі наявних даних, статистичне навчання вимагає дотримання низки принципів, що забезпечують здатність моделі до узагальнення. Одним із ключових є поділ даних на навчальну та тестову вибірки, що дозволяє об'єктивно оцінити якість побудованої моделі [1]. Навчальна вибірка слугує для налаштування параметрів, а тестова - для остаточної перевірки на даних, які модель не «бачила» під час навчання. Важливу роль відіграє також принцип узагальнення - здатність моделі зберігати адекватність і передбачувальну силу при застосуванні до нових даних. Цей принцип безпосередньо пов'язаний із компромісом зміщення та дисперсії: підвищення гнучкості моделі покращує її точність на навчальній вибірці, але може погіршити узагальнення [2]. У цьому контексті особливої уваги набуває проблема перенавчання, коли модель надмірно підлаштовується під специфічні особливості навчальної вибірки, включно з її шумовою компонентою, і втрачає здатність до ефективного узагальнення. Перенавчена модель демонструє відмінні показники на навчальних даних, але низьку якість на тестових. Основними методами боротьби з перенавчанням є регуляризація, рання зупинка навчання, збільшення обсягу вибірки, крос-валідація та зменшення кількості ознак [1, 4, 7].

Не менш важливий принцип, що випливає з практики статистичного навчання, - жодна складна модель не компенсує низької якості вхідних даних. Процедура попередньої обробки включає очищення від викидів і явних помилок, нормалізацію або стандартизацію, заповнення пропущених значень, кодування категоріальних змінних та відбір релевантних ознак [7, 8]. Якість підготовки даних безпосередньо впливає на точність, стабільність, збіжність алгоритмів і надійність отриманих моделей.

Оцінювання побудованих моделей виконується за допомогою спеціалізованих метрик, які дозволяють кількісно визначити ефективність

прогнозних залежностей. Вибір показників залежить від типу задачі [1, 5]. Для задач регресії застосовують середньоквадратичну похибку, корінь із середньоквадратичної похибки, середню абсолютну похибку та коефіцієнт детермінації, який вимірює частку дисперсії залежної змінної, пояснену моделлю. Для задач класифікації використовують матрицю неточностей, точність, чутливість, F1-міру та площу під ROC-кривою, що є особливо цінним для незбалансованих вибірок [1, 4].

Залежно від характеру вихідних даних і поставленої мети статистичне навчання реалізується в межах двох фундаментальних парадигм. Навчання з учителем застосовується, коли навчальна вибірка містить як набір вхідних змінних, так і відповідні їм цільові значення. Мета: знайти функціональну залежність, яка з мінімальною помилкою відображає вхідні дані на вихідні. Цей підхід дозволяє розв’язувати задачі регресії та класифікації [1, 2]. Навчання без учителя використовується у випадках, коли цільова змінна відсутня, а аналізуються лише вхідні дані. Його мета - виявлення прихованих внутрішніх структур і закономірностей. Цей напрям охоплює задачі кластеризації, зниження розмірності та пошуку аномалій [7, 9].

Статистичне навчання виступає важливим інструментом сучасного аналізу даних, забезпечуючи теоретично обґрунтований і практично реалізований підхід до побудови ефективних моделей для прогнозування, класифікації, виявлення прихованих закономірностей у складних наборах даних [1-5, 7].

1.3. Методи статистичного аналізу даних та навчання

Методи статистичного аналізу та навчання формують фундамент обробки й інтерпретації даних, даючи змогу виявляти регулярності, оцінювати зв’язки між змінними та створювати прогнозні моделі [1, 7]. Цей інструментарій охоплює широкий спектр підходів, вибір яких визначається типом даних, метою дослідження та необхідною точністю результатів. У сучасній науковій

літературі методи класифікують за кількома ознаками: наявністю цільової змінної, характером залежностей, складністю моделей та сферою застосування [2, 4].

Кореляційний аналіз є одним із базових інструментів, призначених для оцінювання сили та напрямку зв’язку між змінними. Він дозволяє встановити, наскільки зміна однієї величини супроводжується зміною іншої, що робить його важливим етапом попереднього дослідження даних [1, 7]. Залежно від типу даних та характеру розподілу використовуються різні коефіцієнти. Коефіцієнт кореляції Пірсона застосовується для оцінки лінійного зв’язку між двома неперервними змінними за умови нормальності розподілу. Ранговий коефіцієнт Спірмена є доцільним для порядкових змінних або коли розподіл відхиляється від нормального. Коефіцієнт Кендала, також ранговий, вважається більш стійким до викидів [3, 5]. Важливо пам’ятати, що кореляція не свідчить про причинно-наслідковий зв’язок – вона лише фіксує наявність спільної статистичної мінливості.

Регресійні методи займають центральне місце в арсеналі дослідника, оскільки дозволяють моделювати залежності між змінними та виконувати прогнозування. Лінійна регресія є найпростішою формою, що передбачає лінійний зв’язок між однією незалежною змінною та залежною. Багатофакторна лінійна регресія розширює цей підхід, даючи змогу одночасно враховувати вплив кількох незалежних змінних. Це дозволяє оцінювати внесок кожного фактора в результуючу змінну за умови контролю інших [1, 2]. Коефіцієнти регресії інтерпретуються як очікувана зміна залежної змінної при зміні відповідного предиктора на одиницю за незмінності решти факторів. Окрім лінійних моделей, існують нелінійні варіанти: поліноміальна регресія, логістична регресія та узагальнені адитивні моделі [4, 6]. Регресійні моделі є ключовими інструментами статистичного навчання, особливо цінними завдяки своїй інтерпретованості.

У межах навчання з учителем важливе місце належить методам

класифікації, які призначені для віднесення об’єктів до одного з наперед визначених класів. На відміну від регресії, де результатом є неперервне значення, класифікація дає дискретний вихід [1, 4]. До найбільш поширених належать: логістична регресія, метод k-найближчих сусідів, метод опорних векторів, дерева прийняття рішень та їх ансамблі, зокрема випадковий ліс. Дерева рішень цінуються за високу інтерпретабельність, оскільки вони генерують правила у вигляді послідовності питань про значення ознак. Випадковий ліс, який будує велику кількість дерев і усереднює їхні прогнози, значно підвищує точність класифікації за рахунок зменшення дисперсії [2, 7].

Методи кластеризації належать до навчання без учителя і дозволяють групувати об’єкти за схожістю без попередньо заданих міток. Це дає змогу виявляти приховані структури в даних та формувати узагальнені групи спостережень [7, 9]. Найвідомішим є алгоритм k-середніх, який мінімізує внутрішньокластерну суму квадратів відстаней між об’єктами та центрами кластерів. Ієрархічні методи будують деревоподібну структуру вкладених кластерів, дозволяючи досліднику обирати бажаний рівень деталізації. Метод DBSCAN, заснований на щільності, здатний виявляти кластери довільної форми та ефективно працює з викидами. Важливою проблемою кластеризації є визначення оптимальної кількості кластерів; для цього застосовують метод ліктя, силуетний аналіз та статистику розриву [1, 5].

Для дослідження даних, що змінюються в часі, використовують методи аналізу часових рядів. Вони дозволяють вивчати тенденції, сезонні компоненти та випадкові коливання, а також будувати прогнози [3]. Класичним підходом є декомпозиція ряду на трендову, сезонну та залишкову складові. Серед статистичних моделей найбільшого поширення набуло сімейство ARIMA, яке об’єднує авторегресійну компоненту, процедуру інтегрування для досягнення стаціонарності та компоненту ковзного середнього. Модель SARIMA розширює ARIMA шляхом додавання сезонних складових, що робить її придатною для рядів з яскраво вираженою сезонністю. Для прогнозування

також використовуються методи експоненційного згладжування, прості наївні моделі, а в останні роки – методи машинного навчання, адаптовані до часових залежностей [3, 6].

Важливим напрямом є методи зменшення розмірності, які перетворюють вихідний простір великої кількості ознак у простір меншої розмірності з мінімальною втратою інформації. Найбільш відомим є метод головних компонент, що знаходить ортогональні лінійні комбінації вихідних ознак, які пояснюють максимальну дисперсію даних [4, 7]. PCA широко використовується як для візуалізації багатовимірних даних, так і для підготовки даних перед застосуванням інших методів навчання, оскільки дозволяє пом’якшити проблему «прокляття розмірності». Іншими підходами є факторний аналіз, багатовимірне шкалювання та t-SNE [9].

Особливу увагу в контексті статистичного навчання слід приділити методам контролю якості моделей, зокрема крос-валідації. Крос-валідація – це метод оцінювання узагальнювальної здатності моделі шляхом багаторазового розбиття даних на навчальну та тестову вибірки. Найпоширенішими формами є k-блокова крос-валідація, де дані випадково поділяються на k приблизно рівних частин, а також метод залишкового одного спостереження. Крос-валідація дозволяє не лише оцінити якість моделі, але й налаштувати гіперпараметри, що є критичним для запобігання перенавчанню [1, 5, 8].

Застосування комплексу методів – кореляційного аналізу, регресійних і класифікаційних моделей, кластеризації, аналізу часових рядів, зменшення розмірності та крос-валідації – дозволяє проводити глибоке дослідження даних, будувати точні прогнозні моделі та отримувати обґрунтовані результати, які слугують основою для подальшого використання в різних предметних областях [1–9].

У практичній частині даної роботи з розглянутих методів застосовуються кореляційний аналіз, кластеризація методом k-середніх та множинна лінійна регресія, оскільки вони найкраще відповідають структурі даних та поставленим

завданням [1, 2, 7].

1.4 Сфери застосування аналізу даних та статистичного навчання

Аналіз даних та статистичне навчання активно використовуються в різних сферах діяльності, оскільки дають змогу ефективно опрацьовувати великі інформаційні масиви, виявляти приховані закономірності та підтримувати прийняття рішень в умовах невизначеності [1, 7]. Завдяки розвитку інформаційних технологій, зростанню обчислювальних потужностей та накопиченню даних роль цих методів постійно зростає, проникаючи практично в усі галузі людської діяльності. У сучасному світі здатність отримувати знання з даних стає конкурентною перевагою як на рівні окремих організацій, так і на рівні цілих економік [2, 5].

У сфері економіки та бізнесу аналіз даних застосовується для прогнозування попиту, оцінювання ризиків, оптимізації логістичних процесів та підвищення ефективності діяльності підприємств [1]. Маркетингові відділи компаній використовують методи статистичного навчання для сегментації клієнтської бази, персоналізації пропозицій, аналізу споживчої поведінки та прогнозування відтоку клієнтів. Системи рекомендацій, що базуються на аналізі попередніх покупок або переглядів, стали невід’ємною частиною електронної комерції [4]. Управління ланцюгами постачання використовує прогнозні моделі для оптимізації запасів, планування поставок та мінімізації витрат. У сфері управління персоналом аналітика даних дозволяє оцінювати ефективність працівників, прогнозувати плинність кадрів та виявляти фактори, що впливають на продуктивність праці [5].

У фінансовій галузі зазначені підходи застосовуються для аналізу ринкових тенденцій, виявлення аномалій та формування інвестиційних стратегій [2]. Банки використовують методи статистичного навчання для оцінки кредитоспроможності позичальників, що дозволяє мінімізувати ризики неповернення кредитів. Системи виявлення шахрайства аналізують транзакції

в реальному часі, виявляючи підозрілі операції, які відхиляються від типової поведінки клієнта [7]. Алгоритмічна торгівля, заснована на аналізі ринкових даних, дозволяє автоматизовано здійснювати угоди з метою отримання прибутку. Страхові компанії застосовують актуарні розрахунки та прогнозні моделі для оцінки ризиків і встановлення страхових тарифів [1, 4].

У наукових дослідженнях аналіз даних є важливим інструментом для перевірки гіпотез, моделювання складних процесів і отримання нових знань. Сучасна наука, особливо в таких галузях, як астрофізика, геноміка та нейронаука, генерує величезні обсяги даних, які неможливо проаналізувати без автоматизованих методів [3, 6]. У фізиці високих енергій аналіз даних з прискорювачів частинок дозволяє підтверджувати або спростовувати теоретичні моделі будови матерії. У біології методи статистичного навчання застосовуються для аналізу генетичних послідовностей, передбачення структури білків та дослідження еволюційних зв’язків [4]. В екології моделі машинного навчання допомагають прогнозувати зміни в популяціях видів, оцінювати вплив забруднення довкілля та моделювати розповсюдження інвазійних видів [7].

У технічних системах та інженерії аналіз даних використовується для контролю якості, прогнозування відмов та оптимізації роботи складних систем [8]. Промислові підприємства впроваджують системи прогнозованого технічного обслуговування, які на основі даних з датчиків передбачають момент виходу обладнання з ладу, що дозволяє проводити ремонт завчасно та уникати незапланованих простоїв. У сфері енергетики моделі статистичного навчання допомагають прогнозувати навантаження на електромережі, оптимізувати виробництво енергії з відновлюваних джерел та виявляти витоки в трубопроводах [5]. В автомобільній промисловості аналіз даних є ключовою технологією для створення систем допомоги водієві та розробки безпілотних автомобілів. Окреме значення аналіз даних має у сфері охорони здоров’я, де застосовується для діагностики захворювань, аналізу медичних зображень та

прогнозування перебігу хвороб [2, 6]. Системи підтримки прийняття рішень допомагають лікарям інтерпретувати результати діагностичних тестів, обирати оптимальні схеми лікування та оцінювати ймовірність ускладнень. Аналіз медичних зображень – рентгенівських знімків, комп'ютерної томографії, магнітно–резонансної томографії – з використанням методів глибокого навчання дозволяє виявляти патології з точністю, яка в багатьох випадках перевершує людську [4]. У фармацевтиці статистичне навчання застосовується для аналізу ефективності ліків під час клінічних випробувань, виявлення побічних ефектів та прискорення розробки нових препаратів. В епідеміології прогнозні моделі допомагають передбачати поширення інфекційних захворювань, що набуло особливої актуальності в контексті пандемій [3, 7].

У соціальних науках методи статистичного аналізу даних дозволяють досліджувати поведінку людей, аналізувати соціальні процеси та приймати обґрунтовані управлінські рішення [1]. Соціологічні опитування, дані з соціальних мереж та адміністративні бази даних обробляються з використанням методів статистичного навчання для виявлення трендів у громадській думці, моделювання соціальних мереж та аналізу політичних процесів [5]. Кримінологія використовує предиктивне моделювання для прогнозування рівня злочинності та оптимізації патрулювання. У міському плануванні аналіз даних допомагає оптимізувати транспортні потоки, планувати розвиток інфраструктури та покращувати якість життя мешканців [8].

Важливою сферою застосування є також метеорологія та кліматологія, де аналіз даних використовується для обробки великих масивів спостережень, дослідження кліматичних змін та побудови прогнозів погодних умов [3, 13]. Використання методів статистичного навчання дозволяє підвищити точність прогнозування та виявляти складні залежності між різними природними процесами [15]. Сучасні системи короткострокового прогнозування погоди обробляють дані з супутників, радарів та наземних метеостанцій, використовуючи нейронні мережі та ансамблеві методи. Кліматичні моделі, які

застосовуються для оцінки глобальних змін, базуються на аналізі багаторічних рядів спостережень за температурою, атмосферним тиском, вологістю та іншими параметрами [14]. Прогнозування стихійних лих – ураганів, повеней, лісових пожеж – дедалі більше спирається на методи машинного навчання, які враховують комплекс взаємопов'язаних факторів.

У сільському господарстві аналіз даних застосовується в рамках концепції точного землеробства, де дані з супутників, дронів та наземних датчиків використовуються для оптимізації поливу, внесення добрив та захисту рослин [5]. Прогнозування врожайності на основі історичних даних та погодних умов дозволяє аграрним підприємствам планувати свою діяльність та мінімізувати ризики [13]. У сфері освіти аналітика даних допомагає оцінювати успішність учнів та студентів, виявляти фактори, що впливають на академічну успішність, та розробляти персоналізовані навчальні плани [8]. Аналіз даних про взаємодію студентів з навчальними матеріалами в електронних курсах дозволяє вдосконалювати освітній процес та вчасно виявляти студентів, які потребують додаткової підтримки.

Окремої уваги заслуговує застосування методів статистичного навчання у сфері кібербезпеки, де системи виявлення вторгнень аналізують мережевий трафік, виявляючи аномалії, які можуть свідчити про кібератаки [7]. Аналіз логів та поведінки користувачів дозволяє своєчасно виявляти несанкціонований доступ до інформаційних систем та запобігати витокам даних.

Аналіз даних та статистичне навчання є універсальними інструментами, що застосовуються в різних галузях людської діяльності [1–8]. Від бізнесу та фінансів до науки, медицини, освіти, сільського господарства та кібербезпеки – скрізь, де накопичуються дані, методи статистичного навчання дозволяють отримувати нові знання, оптимізувати процеси та приймати обґрунтовані рішення. У міру подальшого зростання обсягів даних та вдосконалення обчислювальних алгоритмів роль цих методів буде лише посилюватися, стаючи невід'ємною частиною сучасного наукового та практичного дискурсу [2, 3, 7].

РОЗДІЛ 2
АНАЛІЗ ТА ПІДГОТОВКА МЕТЕОРОЛОГІЧНИХ ДАНИХ ДЛЯ ПРОВЕДЕННЯ СТАТИСТИЧНОГО НАВЧАННЯ

2.1 Джерела метеорологічних даних та їх характеристика

Для виконання аналізу та побудови моделей у межах цієї роботи було використано метеорологічні дані, отримані з відкритих міжнародних джерел. Основним постачальником інформації виступає Національне управління океанічних та атмосферних досліджень США «National Oceanic and Atmospheric Administration» NOAA [13], яке займається збором, обробкою та поширенням даних про атмосферні, океанічні та кліматичні процеси. Для зручного доступу та структурованого отримання інформації застосовано платформу Meteostat [15], яка надає метеорологічну інформацію у вигляді готових до аналізу наборів даних.

Дослідження ґрунтується на метеорологічних даних, що характеризують погодні умови в місті Вінниця (Україна). Вибір саме цього регіону зумовлений, по–перше, наявністю повних історичних даних за період 2021–2024 роки, а по–друге, можливістю дослідження змін метеорологічних показників у межах конкретної географічної області [14].

Безпосередній доступ до архівів метеорологічних спостережень забезпечується через Національні центри екологічної інформації «National Centers for Environmental Information, NCEI» [14], які є одним із найбільших у світі сховищ кліматичних даних. Цей ресурс надає відкритий доступ до численних наборів даних, що охоплюють багаторічні спостереження з метеорологічних станцій у різних регіонах світу.

У цій роботі використовується набір даних Global Surface Summary of the Day GSOD [13], який містить щоденні усереднені значення метеорологічних показників. Цей набір даних акумулює інформацію, зібрану з наземних метеорологічних станцій, і представлений у табличному форматі, що є зручним

для подальшого аналізу в середовищі програмування R [12].

До складу робочої вибірки було включено такі основні метеорологічні показники:

1. середня температура повітря;
2. мінімальна температура;
3. максимальна температура;
4. атмосферний тиск;
5. швидкість вітру;
6. кількість опадів.

Інші показники, такі як напрямок вітру, висота снігового покриву, пориви вітру та тривалість сонячного сяйва, не використовуються в подальшому аналізі через надмірно велику частку пропущених значень [11, 15].

Кожен запис у вихідному наборі даних відповідає конкретному дню спостереження та містить значення відповідних параметрів. Така структура даних дозволяє формувати часові ряди для аналізу динаміки змін показників у часі та є придатною для застосування методів статистичного аналізу, зокрема кореляційного аналізу, регресійного моделювання та прогнозування [1, 3].

Особливістю метеорологічних даних є наявність сезонних коливань, трендів та випадкових відхилень, що обумовлює необхідність їхньої попередньої обробки [3, 13]. Дані можуть містити пропуски, аномальні значення або похибки вимірювання, які виникають у процесі збору інформації з різних джерел. Тому для забезпечення коректності подальшого аналізу необхідно виконати підготовку даних, яка включає очищення, обробку пропущених значень, нормалізацію показників та перевірку їх узгодженості [7, 8]. Це дозволяє підвищити якість побудованих моделей і забезпечити достовірність отриманих результатів.

Використання набору даних GSOD із відкритих ресурсів NOAA [13] та платформи Meteostat [15] забезпечує надійну та репрезентативну інформаційну основу для проведення статистичного аналізу, моделювання та прогнозування метеорологічних процесів у місті Вінниця за період 2021–2024 роки.

2.2 Формування вибірок метеорологічних даних

Формування вибірки є одним із вирішальних етапів дослідження, оскільки саме на цьому кроці закладаються структура, обсяг і якість інформації, яка використовуватиметься для подальшого аналізу, моделювання та прогнозування [1, 7]. Коректно побудована вибірка безпосередньо впливає на достовірність отриманих результатів і адекватність статистичних моделей.

У цій роботі формування вибірки виконується на основі відкритих метеорологічних даних, отриманих через платформу Meteostat [15], яка агрегує інформацію з глобальних джерел, насамперед NOAA [13]. Використання відкритих даних забезпечує прозорість дослідження та можливість відтворення результатів іншими дослідниками.

Об’єктом формування вибірки є метеорологічні дані, що характеризують погодні умови в місті Вінниця, Україна. Вибір цього регіону зумовлений необхідністю дослідження реальних природних процесів у межах визначеної географічної області, а також наявністю достатньо повних і безперервних рядів спостережень [14].

Часовий інтервал вибірки охоплює період з 1 січня 2021 року по 31 грудня 2024 року включно. Такий діапазон дозволяє врахувати як короткострокові коливання, так і сезонні та міжрічні зміни метеорологічних показників [3]. Наявність чотирьох повних річних циклів є важливою умовою для застосування методів аналізу часових рядів та побудови прогностичних моделей.

Формування вибірки здійснюється у вигляді часових рядів із щоденною частотою спостережень. Такий підхід дозволяє детально відстежувати динаміку змін показників та аналізувати їхню поведінку в часі. Вихідна вибірка містить 1461 запис за період 2021–2024 роки, що забезпечує достатній обсяг даних для статистичного аналізу та навчання моделей після проведення процедур очищення.

До складу вибірки включено комплекс основних метеорологічних параметрів, що характеризують стан атмосфери. Наявність кількох взаємопов’язаних змінних дозволяє здійснювати багатофакторний аналіз,

оцінювати вплив окремих факторів на результуючі показники, а також виявляти складні залежності між змінними [1, 2]. Вибір саме цих показників обумовлений їхньою високою інформативністю та широким використанням у задачах аналізу метеорологічних процесів [3, 13]. Температурні показники дають змогу досліджувати тепловий режим, атмосферний тиск і вітер характеризують динаміку повітряних мас, а опади – гідрологічні процеси.

Дані представлено в табличному форматі, що забезпечує зручність їх імпорту в середовище програмування R [12]. Такий формат дозволяє ефективно виконувати подальшу обробку, аналіз та візуалізацію інформації.

У процесі формування вибірки також ураховано необхідність подальшої обробки даних. Зокрема, передбачається можливість наявності пропущених значень, аномалій і похибок вимірювання, що є типовим для реальних метеорологічних спостережень [11, 14]. Це обумовлює необхідність проведення етапу попередньої обробки та очищення даних, який детально розглянуто в підрозділі 2.3.

Сформована вибірка метеорологічних даних є достатньо повною, структурованою та репрезентативною для проведення статистичного аналізу, дослідження взаємозв’язків між показниками та побудови моделей статистичного навчання [1, 7, 15].

2.3 Попередня обробка та очищення даних

Попередня обробка даних є обов’язковим етапом статистичного навчання, оскільки якість вхідних даних безпосередньо визначає достовірність побудованих моделей [7, 10]. У цьому дослідженні всі операції з підготовки метеорологічних даних виконувались у середовищі RStudio з використанням мови програмування R [12].

Етап 1. Завантаження та первинний огляд. Вихідний CSV–файл імпортовано за допомогою функції read.csv(). Набір даних містив щоденні спостереження за період з 01.01.2021 по 31.12.2024, усього 1461 запис та 11 змінних. Змінна date

зберігалась у текстовому форматі, тому її було перетворено на тип Date для коректної роботи з часовими рядами. Фрагмент початкових даних наведено на рис. 2.5.

Етап 2. Аналіз пропущених значень. Для кількісної оцінки пропусків застосовано функцію colSums(is.na(data)). Результати зведено в табл. 2.1, візуалізацію наведено на рис. 2.1. Виявлено, що змінні snow, wdir, wpgt, tsun мають понад 80% пропущених значень, відповідно 89,2%, 100%, 98,4% та 100%. Стовпці wdir та tsun взагалі не містять жодного заповненого значення. Такі змінні визнано непридатними для статистичного аналізу, оскільки будь–яке відновлення даних призвело б до неприпустимих спотворень [11, 14].

Таблиця 2.1 – Кількість пропущених значень за змінними 2021–2024

Змінна	Кількість NA	Частка, %
date	0	0,0
tavg	58	4,0
tmin	124	8,5
tmax	58	4,0
prcp	268	18,3
snow	1303	89,2
wdir	1461	100,0
wspd	282	19,3
wpgt	1438	98,4
pres	282	19,3
tsun	1461	100,0

Візуалізація кількості пропущених значень за змінними наведена на рисунку 2.1.

Етап 3. Видалення малоповних змінних та обробка рядків з пропусками. Відповідно до загальноприйнятої практики [7, 10], змінні з часткою пропусків понад 50% виключаються з подальшого аналізу. Тому з набору даних видалено

стовпці snow, wdir, wpgt, tsun. Після цього застосовано метод повного виключення рядків «na.omit», який є допустимим, оскільки відносна кількість рядків, що містять хоча б один пропуск, становить приблизно 19%, а відсутність значень має випадковий характер [11].

Рис. 2.1 – Візуалізація кількості пропущених значень за змінними

Таке рішення дозволяє уникнути штучного внесення похибок, пов’язаного з імпутацією. Лістинг обробки має наступний вигляд:

```
vars_to_remove <- c("snow", "wdir", "wpgt", "tsun")
data <- data[, !(names(data) %in% vars_to_remove)]
    data_clean <- na.omit(data)
```

Етап 4. Контроль якості очищення та формування робочої вибірки. Після очищення повторно виконано функцію «summary(data)», що підтвердило відсутність пропущених значень. Кількісні зміни обсягу вибірки наведено в таблиці 2.2.

Таблиця 2.2 – Зміна обсягу вибірки після очищення

Показник	До очищення	Після очищення
Кількість змінних	11	7
Кількість спостережень	1461	1445
Частка видалених рядків	–	18,8%

Отримана робоча вибірка містить 1445 повних щоденних спостережень за шістьма метеорологічними показниками: середня температура «tavg», мінімальна температура «tmin», максимальна температура «tmax», кількість опадів «prcp», швидкість вітру «wspd», атмосферний тиск «pres» – а також змінну дати «date». Ці дані є придатними для подальшого кластерного, кореляційного та регресійного аналізу. На рисунку 2.2 подано блок-схему виконаних дій, яка візуалізує послідовність кроків: завантаження даних; перетворення типу дати; підрахунок пропусків; видалення змінних з «>50% NA»; фільтрація за періодом 2021–2024; видалення рядків з будь–якими NA перевірка результату. Така схема відповідає вимогам до оформлення алгоритмів обробки даних у дипломних роботах.

Застосування того чи іншого методу статистичного навчання для певного набору даних є виправданим, якщо цей метод забезпечує низьку помилку на контрольній вибірці. Помилку на контрольній вибірці можна легко обчислити при наявності призначеного для цього контрольного набору даних. На жаль, зазвичай такий набір даних відсутній. У той же час помилку на навчальній вибірці можна легко розрахувати шляхом застосування деякої моделі безпосередньо до спостережень, за якими вона була навчена. За відсутності дуже великої і спеціально створеної контрольної вибірки, безпосередньо підходящої для оцінювання частоти помилок, можна застосувати ряд методів для знаходження подібної оцінки за наявними навчальними даними. Деякі з цих методів оцінюють частоту помилок на контрольній вибірці за рахунок певної математичної поправки, що застосовується до частоти помилок на навчальній вибірці.

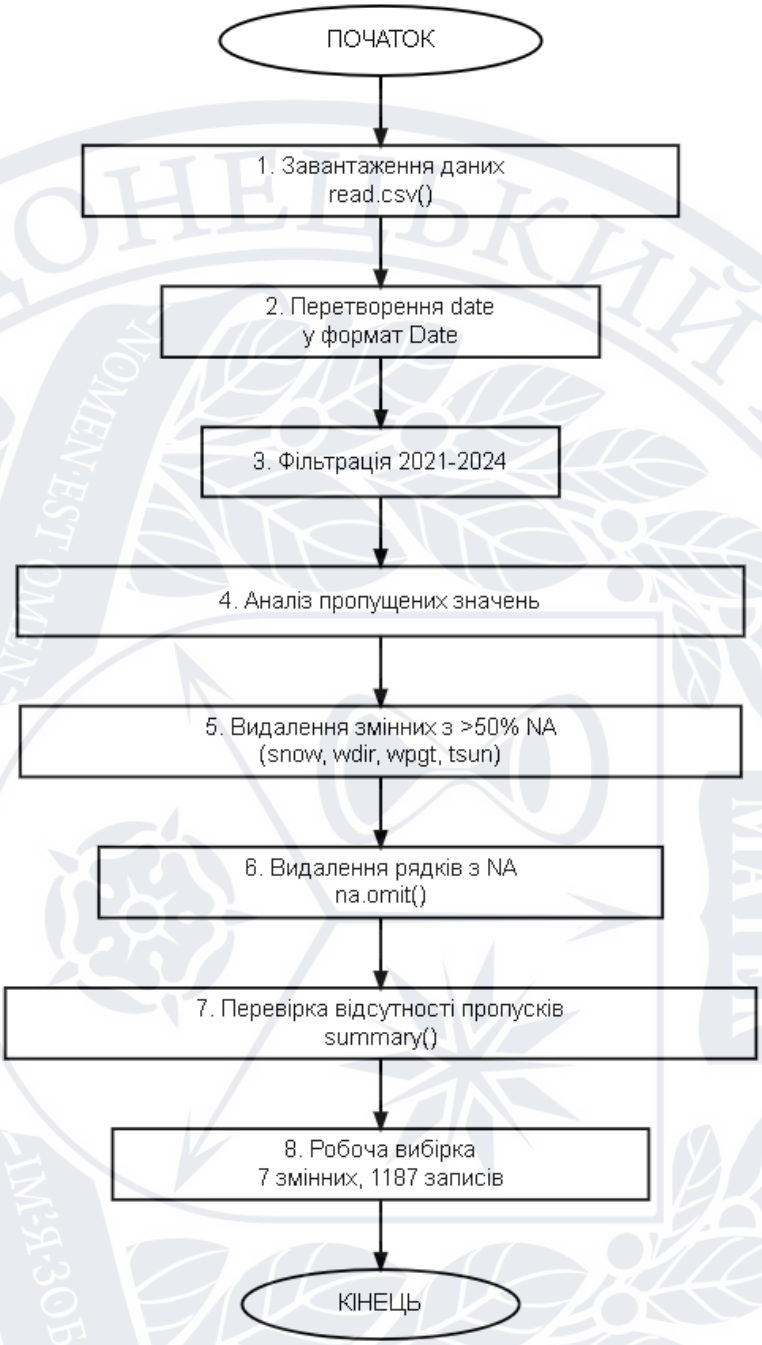

Рис. 2.2 – Блок-схема процедури очищення даних

2.4. Кластеризація метеорологічних даних

Кластеризація належить до методів навчання без учителя і дає змогу виявляти приховані структури в даних шляхом групування об’єктів за ознаками подібності [1, 7]. У цьому дослідженні кластеризація застосовується для виявлення типових погодних умов на основі сукупності метеорологічних показників за період 2021–2024 рр.

Для кластеризації відібрано шість числових змінних, які характеризують погодні умови: середня температура «tavg», мінімальна температура «tmin», максимальна температура «tmax», кількість опадів «prcp», швидкість вітру «wspd» та атмосферний тиск «pres». Оскільки ці показники мають різні одиниці вимірювання, перед застосуванням алгоритму кластеризації виконано їх масштабування методом z–нормалізації «функція scale()», що забезпечує рівноправний вплив кожної змінної на результат групування [2, 9].

Для вибору оптимальної кількості кластерів використано метод «ліктя», «elbow method», який аналізує залежність суми внутрішньокластерних відстаней від кількості кластерів. Результати наведено на рис. 2.3. Як видно з графіка, починаючи з k = 3, зменшення суми внутрішньокластерних відстаней сповільнюється, утворюючи характерний «злам». Тому для подальшого аналізу обрано три кластери.

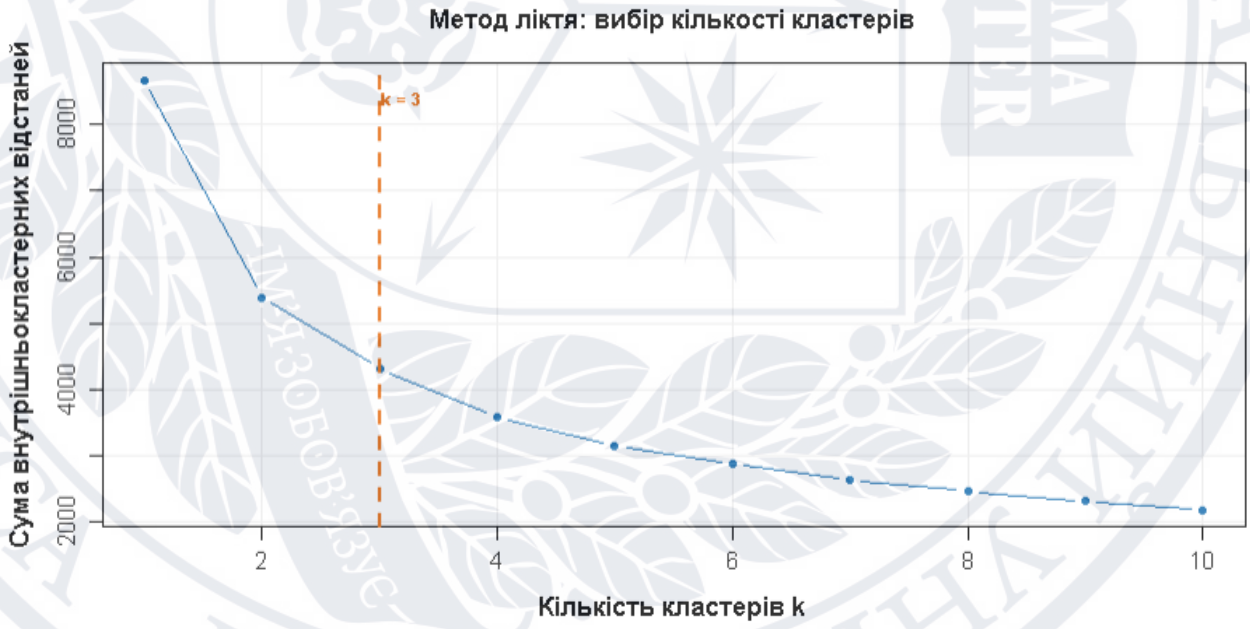

Рис. 2.3 – Графік методу ліктя

Кластеризацію виконано за допомогою алгоритму k-середніх функція «kmeans()» з кількістю центрів, рівною 3, та 25 запусками для забезпечення стійкості результату. У результаті отримано три кластери, які після впорядкування

за середньою температурою ідентифіковано як холодний, перехідний та теплий періоди. Середня температура в кластерах становить: для холодного періоду – 2,1°C, перехідного – 5,1°C, теплого – 18,7°C.

Центри кластерів, середні значення кожної змінної у вихідному масштабі наведено в табл. 2.3.

Таблиця 2.3 – Центри кластерів 2021–2024

Кластер	tavg, °C	tmin, °C	tmax, °C	prcp, мм	wspd, м/с	pres, гПа
Холодний зима	2,14	–1,63	5,87	0,44	11,02	1023,56
Перехідний весна/осінь	5,06	1,80	8,29	3,82	16,39	1009,14
Теплий літо	18,72	12,74	24,95	1,39	10,37	1015,20

На рис. 2.3 показано розподіл спостережень у просторі змінних «середня температура – атмосферний тиск», а на рис. 2.4 – у просторі «середня температура – швидкість вітру». Кольори точок відповідають приналежності до кластерів (синій – холодний період, зелений – перехідний, помаранчевий – теплий).

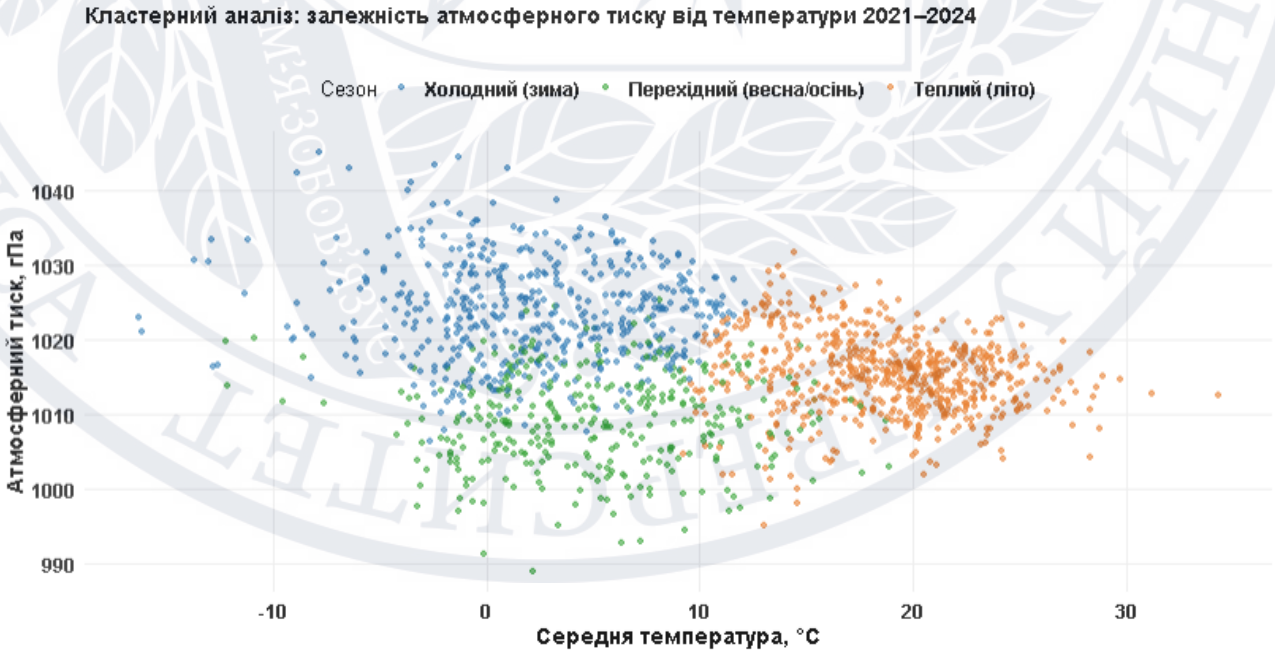

Рис. 2.3 – Кластеризація у просторі «середня температура – атмосферний тиск»

Результати кластеризації у просторі «середня температура – швидкість вітру» наведено на рисунку 2.4.

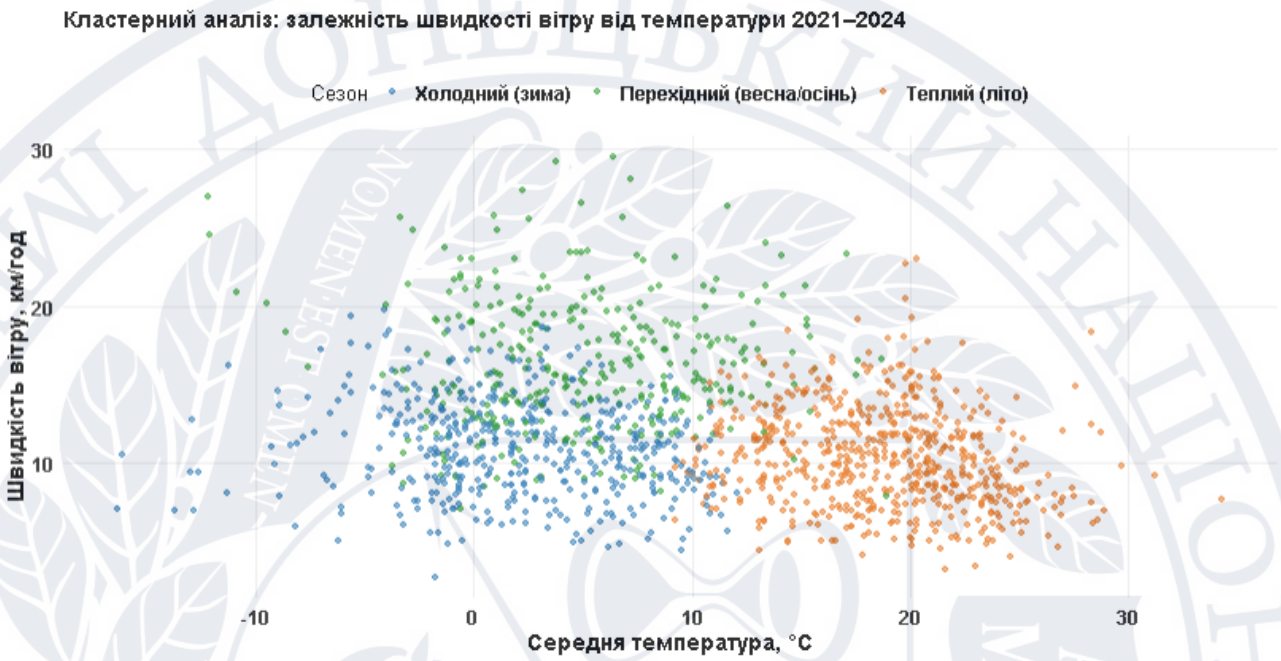

Рис. 2.4 – Кластеризація у просторі «середня температура – швидкість вітру»

Аналіз отриманих результатів показав, що:

1. Кластер 1 «холодний період». Характеризується найнижчими температурами середня tavg ≈ 2,1°C, найвищим атмосферним тиском 1023,6 гПа та помірною швидкістю вітру 11,0 м/с. Опади в цьому кластері мінімальні 0,4 мм. Цей кластер відповідає зимовим та ранньовесняним умовам.

2. Кластер 2 «перехідний період». Має помірні температури середня tavg ≈ 5,1°C, найнижчий атмосферний тиск 1009,1 гПа та найвищу швидкість вітру 16,4 м/с, а також найбільшу кількість опадів 3,8 мм. Відповідає весняним та осіннім місяцям, а також окремим дням з нестабільною погодою.

3. Кластер 3 «теплий період». Об’єднує дні з високими температурами середня tavg ≈ 18,7°C, високими екстремальними температурами tmin 12,7°C, tmax 25,0°C, дещо нижчим тиском 1015,2 гПа та найнижчою швидкістю вітру 10,4 м/с. Цей кластер відповідає літньому сезону.

Проведена кластеризація підтвердила наявність природної сезонної

структури в метеорологічних даних та дозволила кількісно охарактеризувати кожен тип погодних умов. Отримані результати будуть використані на наступних етапах дослідження для побудови регресійних моделей прогнозування температури [1, 2, 7].

2.5. Первинний статистичний аналіз даних

Після виконання очищення даних та кластеризації було проведено первинний статистичний аналіз, метою якого є узагальнення основних характеристик досліджуваних показників, виявлення особливостей їх розподілу та попередня оцінка динаміки змін у часі. Цей етап є важливим для розуміння структури даних перед побудовою регресійних моделей [1, 7].

Для кожної з шести метеорологічних змінних: середня; мінімальна та максимальна температура; кількість опадів; швидкість вітру; атмосферний тиск. Обчислено основні статистичні показники: мінімум, максимум, середнє арифметичне, медіану, а також перший та третій квартилі. Результати наведено в таблиці 2.4.

Таблиця 2.4 – Описові статистики метеорологічних показників

Показник	tavg, °C	tmin, °C	tmax, °C	prcp, мм	wspd, м/с	pres, гПа
Мінімум	–16,3	–22,9	–13,7	0,0	2,7	988,8
1–й квартиль	2,3	–0,9	5,1	0,0	9,0	1011,4
Медіана	9,4	4,9	14,2	0,1	11,6	1016,3
Середнє	9,94	5,34	14,62	1,64	12,0	1016,6
3–й квартиль	17,9	12,3	24,3	1,5	14,3	1021,9
Максимум	34,3	23,4	39,2	33,2	29,5	1045,0

Аналіз табл. 2.4 показує, що середня температура коливається від –16,3°C до 34,3°C, що відповідає континентальному клімату регіону. Мінімальна температура опускається до –22,9°C, а максимальна сягає 39,2°C. Кількість опадів є сильно варіабельною: більшість днів мають незначні опади, медіана 0,1 мм, однак трапляються дні з екстремальними значеннями до 33,2 мм. Швидкість вітру в середньому становить 12,0 м/с, з максимумом 29,5 м/с. Атмосферний тиск змінюється в межах від 988,8 до 1045,0 гПа, що є типовим для помірних широт [3, 13]. Для візуальної оцінки форми розподілу кожної змінної побудовано гістограми (рис. 2.5–2.8).

Розподіл середньої температури наведено на рисунку 2.5.

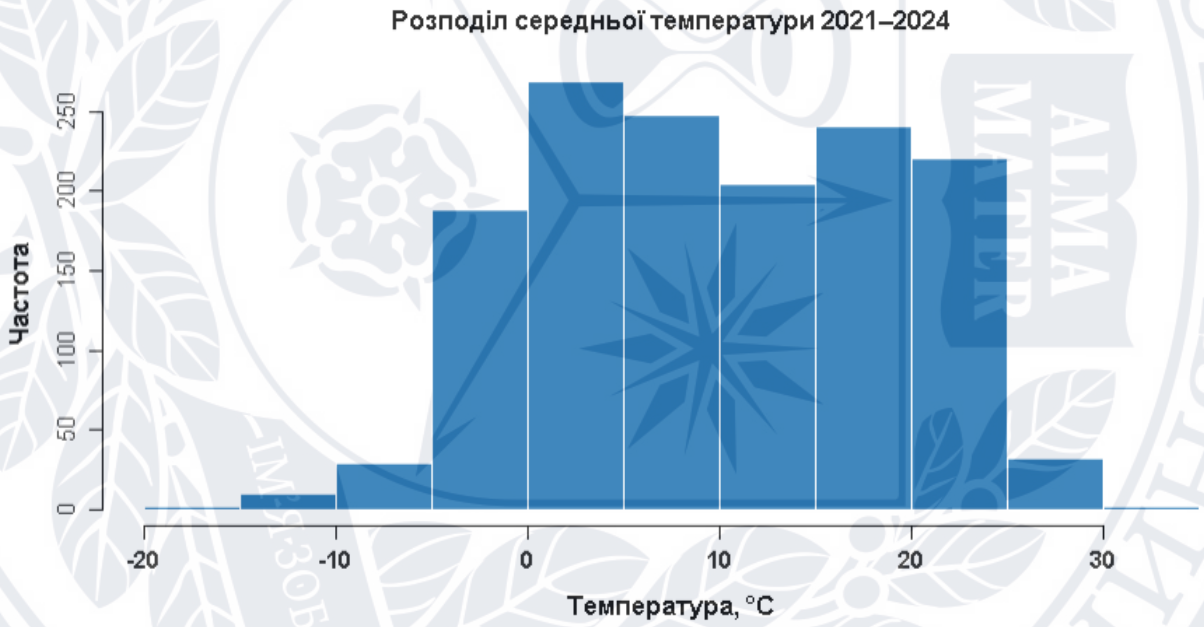

Рис. 2.5 – Розподіл середньої температури

Гістограма (рис. 2.5) демонструє симетричну форму, близьку до нормального розподілу, з одним чітко вираженим піком у діапазоні 5–15°C. Це свідчить про відсутність суттєвих аномалій та придатність даних для параметричних методів аналізу.

Розподіл опадів наведено на рисунку 2.6.

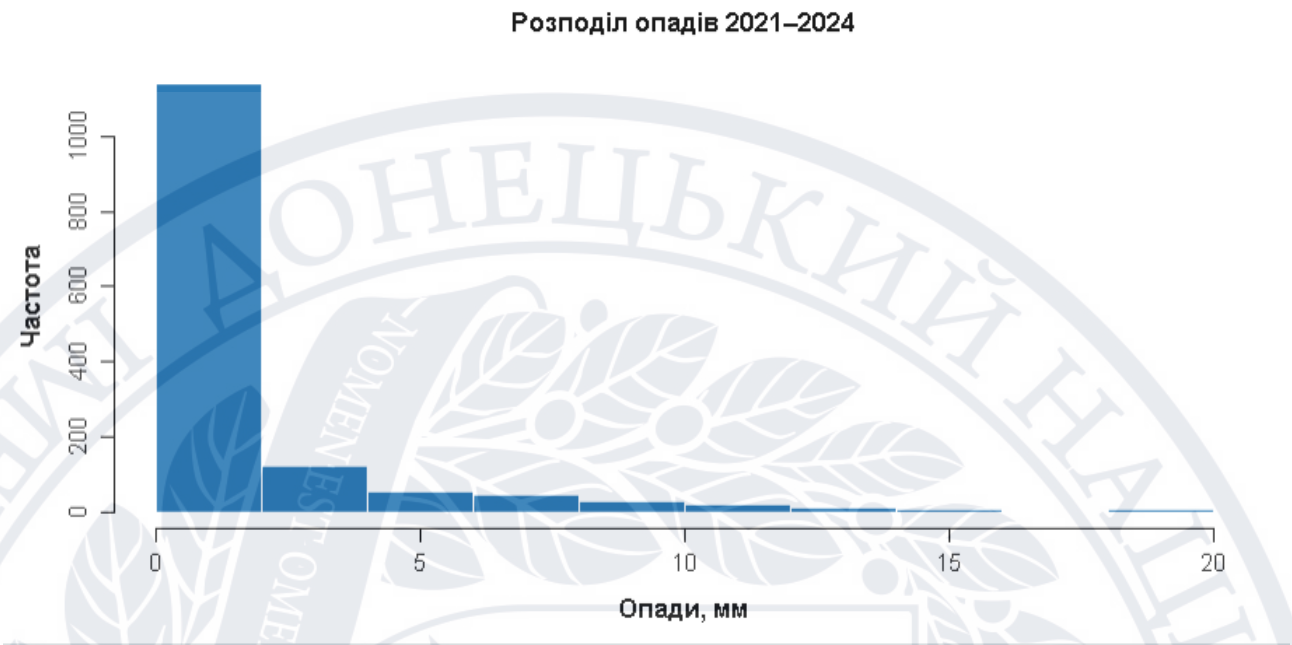

Рис. 2.6 – Розподіл кількості опадів

Розподіл опадів (рис. 2.6) є різко асиметричним , правостороння асиметрія з довгим «хвостом» у бік великих значень. Переважна більшість днів має нульові або незначні опади, що є типовим для континентального клімату.

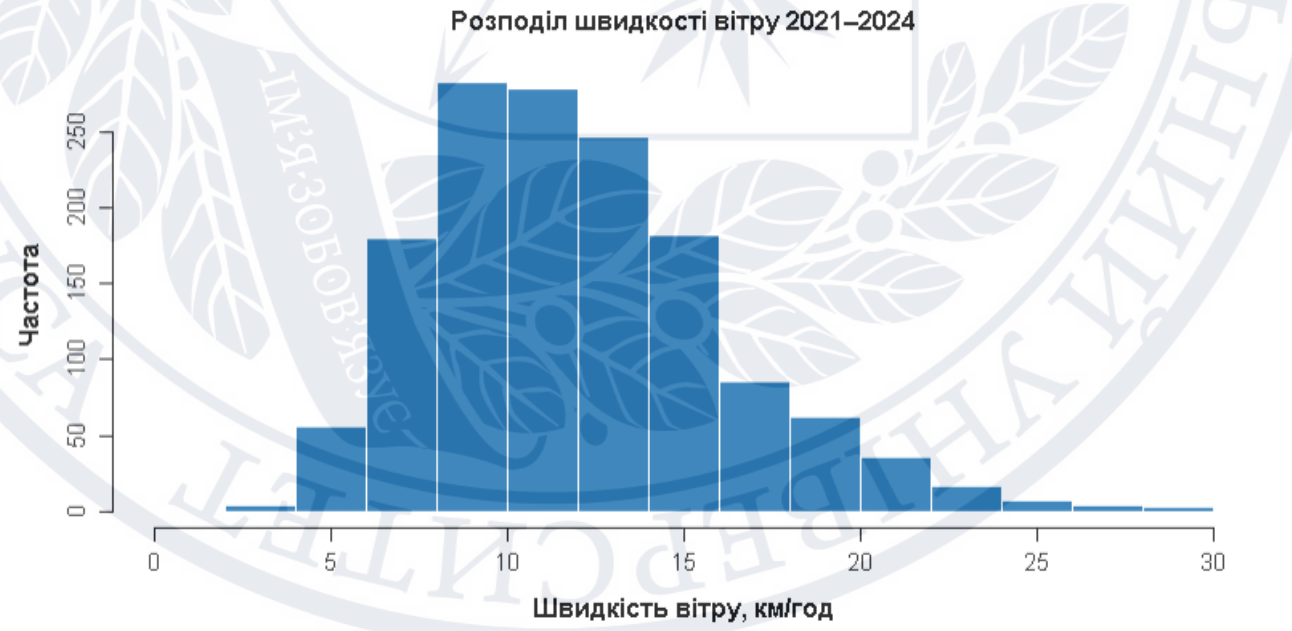

Рис. 2.7 – Розподіл швидкості вітру

Швидкість вітру (рис. 2.7) має одновершинний розподіл з деякою правосторонньою асиметрією, що пояснюється наявністю окремих днів зі штормовими вітрами, понад 25 м/с.

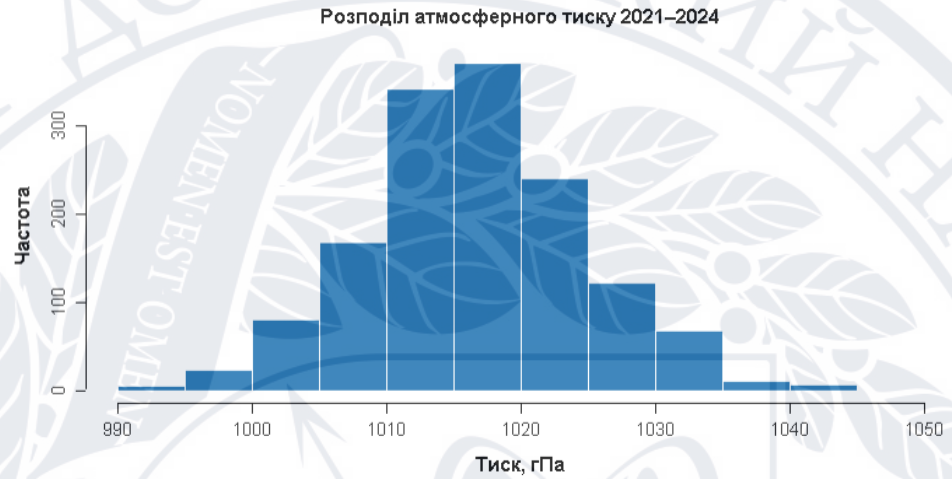

Рис. 2.8 – Розподіл атмосферного тиску

Атмосферний тиск (рис. 2.8) демонструє симетричний, практично нормальний розподіл із концентрацією значень навколо середнього ≈1016,6 гПа, що відповідає відносно стаціонарним умовам.

Для дослідження зміни температури в часі побудовано графік часового ряду (рис. 2.9). Дані охоплюють період з 1 січня 2021 року по 31 грудня 2024 року.

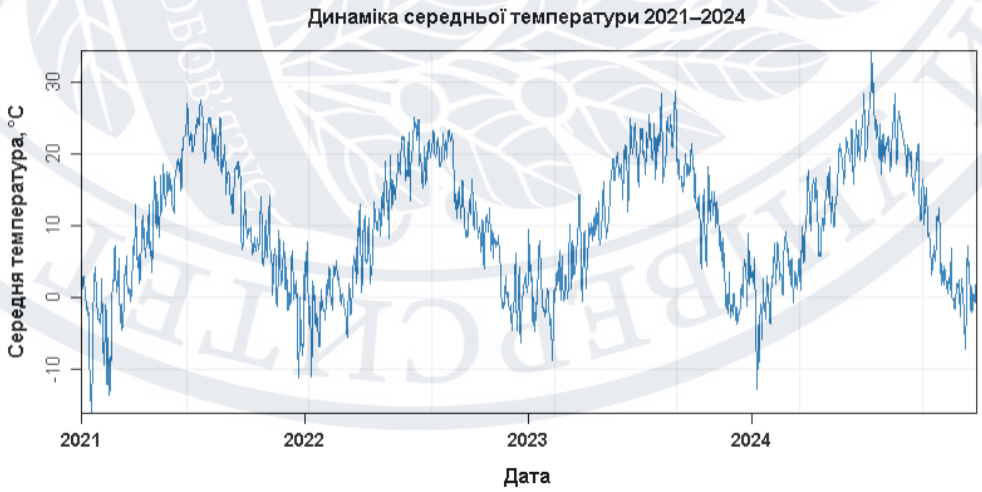

Рис. 2.9 – Динаміка середньої температури

На рис. 2.9 чітко видно сезонні коливання температури з максимумами в літні місяці червень–серпень та мінімумами в зимові грудень–лютий. Амплітуда коливань становить приблизно 50°C від –16°C до +34°C. Ряд є повним за всі роки 2021–2024 і демонструє стійку річну періодичність, що підтверджує репрезентативність вибірки для подальшого аналізу [3].

Первинний статистичний аналіз дозволив охарактеризувати основні властивості досліджуваних метеорологічних показників, виявити особливості їх розподілів та підтвердити наявність чітко вираженої сезонної компоненти в часовому ряді температури. Отримані результати є основою для подальшого кореляційного та регресійного аналізу [1, 2].

2.6 Кореляційний аналіз метеорологічних показників

Кореляційний аналіз є методом статистичного навчання, який дозволяє кількісно оцінити силу та напрямок лінійного зв’язку між досліджуваними змінними. На відміну від регресійного аналізу, кореляція не передбачає причинно–наслідкових відношень, але дає змогу виявити наявність спільної мінливості показників, що є важливим етапом перед побудовою прогнозних моделей [7, 10].

Для оцінки взаємозв’язків між метеорологічними показниками використано коефіцієнт кореляції Пірсона, який застосовується для неперервних змінних, розподіл яких наближається до нормального. Аналіз проведено для шести змінних: середня температура tavg, мінімальна температура tmin, максимальна температура tmax, кількість опадів prcp, швидкість вітру wspd, атмосферний тиск pres. Розрахунки виконано на основі робочої вибірки, що містить 1445 щоденних спостережень за період 2021–2024 рр.

Результати обчислень зведено в кореляційну матрицю (табл. 2.5). Для наочного представлення зв’язків побудовано графічну візуалізацію (рис. 2.16), де колір комірки відповідає силі кореляції, сині відтінки – додатна кореляція, червоні – від’ємна, а числові значення вказані безпосередньо.

Таблиця 2.5 – Кореляційна матриця Пірсона

	tavg	tmin	tmax	prcp	wspd	pres
tavg	1,000	0,966	0,982	0,021	–0,278	–0,236
tmin	0,966	1,000	0,915	0,107	–0,219	–0,280
tmax	0,982	0,915	1,000	–0,029	–0,307	–0,190
prcp	0,021	0,107	–0,029	1,000	0,145	–0,318
wspd	–0,278	–0,219	–0,307	0,145	1,000	–0,237
pres	–0,236	–0,280	–0,190	–0,318	–0,237	1,000

Аналіз отриманих результатів показав, що:

1. Сильні позитивні зв’язки. Коефіцієнти кореляції між «tavg» та «tmax» 0,982, а також між «tavg» та «tmin» 0,966 є дуже високими. Це закономірно, оскільки всі три показники характеризують одну фізичну величину – температуру повітря. Висока кореляція свідчить про те, що зміни мінімальної та максимальної температури добре пояснюють варіацію середньої температури.

2. Слабкі або практично нульові зв’язки. Кількість опадів prcp має коефіцієнт кореляції з температурою, близький до нуля 0,021 з «tavg», –0,029 з «tmax», що вказує на відсутність лінійної залежності між опадами та температурним режимом. Це відповідає реальним кліматичним умовам, де опади визначаються іншими факторами циркуляція атмосфери, вологість.

3. Помірні обернені зв’язки. Атмосферний тиск «pres» демонструє слабку негативну кореляцію з температурою від –0,190 до –0,280 та найбільш помітну з опадами –0,318. Це узгоджується з фізичними уявленнями: зниження тиску часто супроводжується циклональною діяльністю, що приносить опади та зниження температури.

4. Зв’язки швидкості вітру. Швидкість вітру «wspd» має слабкі негативні коефіцієнти з температурою від –0,219 до –0,307 та тиском –0,237, що може бути пов’язане з посиленням вітру під час проходження атмосферних фронтів. Найсильніший негативний зв’язок вітру спостерігається з максимальною

температурою –0,307.

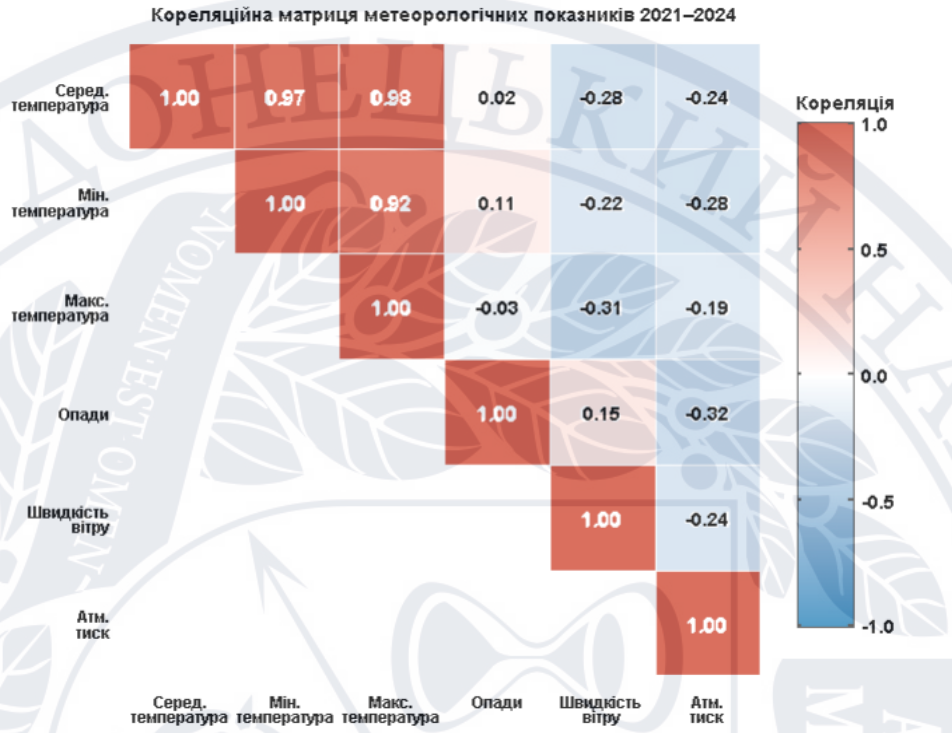

Рис. 2.16 – Візуалізація кореляційної матриці

Результати кореляційного аналізу підтверджують доцільність включення всіх п’яти факторів «tmin», «tmax», «prcp», «wspd», «pres» до регресійної моделі для прогнозування середньої температури tavg. Високі кореляції між температурними показниками не створюють проблеми мультиколінеарності, оскільки tmin та tmax є окремими вимірами, що несуть різну інформацію. Слабкі зв’язки з опадами та вітром свідчать про їх відносну незалежність, що також є корисним для побудови прогнозної моделі [1, 2, 7].

РОЗДІЛ 3
РЕАЛІЗАЦІЯ АЛГОРИТМІВ СТАТИСТИЧНОГО НАВЧАННЯ

3.1. Архітектура та алгоритм обробки метеорологічних даних

Розроблений програмний продукт являє собою сукупність модулів, реалізованих мовою програмування R у середовищі RStudio, які забезпечують виконання повного циклу статистичного аналізу метеорологічних даних: від імпорту первинних даних до оцінки якості прогнозної моделі. Архітектура продукту ґрунтується на принципі послідовної обробки інформації з чітким розділенням на логічні модулі (рис. 3.1).

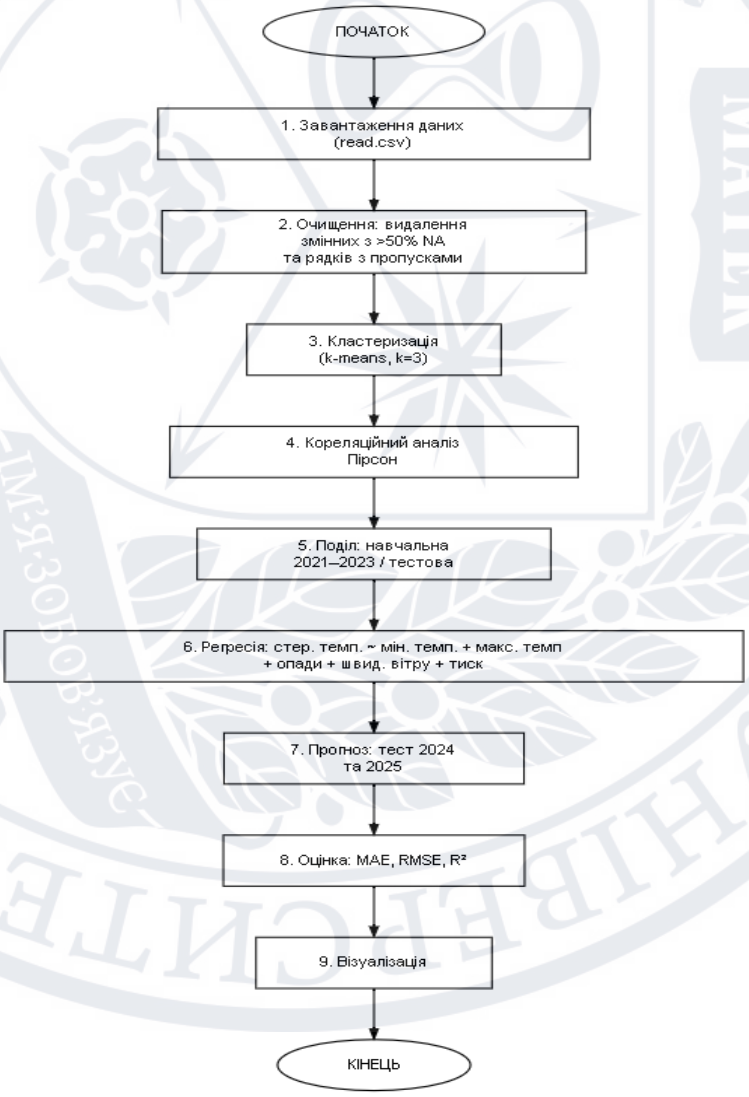

Рис. 3.1 – Архітектура та алгоритм обробки метеорологічних даних

Розроблений програмний продукт є сукупністю модулів на мові R у середовищі RStudio, які реалізують повний цикл статистичного аналізу метеорологічних даних – від імпорту первинних даних до оцінювання якості прогнозної моделі та побудови прогнозу. Архітектура побудована на засаді послідовної обробки інформації з чітким розділенням на логічні блоки (рис. 3.1).

Модуль завантаження та первинної обробки забезпечує імпорт даних з CSV–файлу через «read.csv()», перетворення колонки «date» у тип «Date» з урахуванням формату часу, а також первинний огляд структури набору даних за допомогою «head()», «str()» і «summary()».

Модуль очищення виконує фільтрацію рядків за період 2021–2024 рр., аналіз пропущених значень функцією «colSums(is.na()», видалення змінних із часткою пропусків понад 50% «snow», «wdir», «wpgt», «tsun» і, нарешті, видалення рядків з будь–якими пропусками через «na.omit()». Результатом є робоча вибірка з 1445 спостережень за сімома змінними: «date», «tavg», «tmin», «tmax», «prcp», «wspd», «pres».

Модуль кластеризації застосовує алгоритм k-середніх «kmeans()» для виявлення типових погодних умов. Він включає масштабування даних «scale()», визначення оптимальної кількості кластерів методом «ліктя» з побудовою графіка залежності внутрішньокластерної суми квадратів від k та візуалізацію результатів у вигляді діаграм розсіювання в координатах «температура – тиск» і «температура – швидкість вітру».

Модуль кореляційного аналізу обчислює коефіцієнти кореляції Пірсона «cor()» між усіма метеорологічними показниками та візуалізує отриману матрицю у вигляді кольорової теплової карти з числовими мітками, використовуючи пакет «ggplot2».

Модуль регресійного моделювання будує множинну лінійну регресію «lm()» для прогнозування середньої температури tavg на основі змінних tmin, tmax, prcp, wspd, pres. Перед цим дані розділяються на навчальну вибірку 2021–2023 рр., 1092 спостереження та тестову 2024 р., 353 спостереження. Модуль

виконує прогнозування (predict()) на тестовій вибірці, обчислює метрики якості (MAE, MSE, RMSE, R^2), а також будує прогноз на 2025 рік, використовуючи середньомісячні значення предикторів.

Модуль візуалізації та інтерпретації забезпечує побудову гістограм розподілів, графіків часових рядів, діаграм розсіювання для кластерів, кореляційної матриці та порівняльних графіків фактичних і прогнозованих значень – зокрема суміщеного графіка для навчальної, тестової вибірок і прогнозу на 2025 рік.

Алгоритм реалізує таку послідовність кроків (загальна блок–схема наведена на рис. 3.1):

1. Завантаження даних – зчитування CSV–файлу, перетворення «date», перевірка структури.

2. Очищення даних – фільтрація 2021–2024, видалення змінних з >50% NA та рядків із будь–якими пропусками; формування робочої вибірки 1445 записів, 7 змінних.

3. Кластеризація – масштабування, визначення оптимального k = 3 методом ліктя, застосування k–means, інтерпретація кластерів: холодний, перехідний, теплий періоди з обчисленням центрів.

4. Кореляційний аналіз – обчислення кореляційної матриці, візуалізація та аналіз зв’язків.

5. Регресійне моделювання – поділ на навчальну 2021–2023 та тестову 2024 вибірки, побудова множинної лінійної регресії, отримання коефіцієнтів.

6. Прогнозування – обчислення прогнозних значень для тестової вибірки та для 2025 року на основі середньомісячних значень предикторів.

7. Оцінка якості – розрахунок метрик MAE, MSE, RMSE, R^2, порівняння фактичних і прогнозованих значень.

8. Візуалізація та інтерпретація – побудова всіх графіків, формулювання висновків щодо адекватності моделі.

Усі кроки реалізовано у вигляді єдиного R–скрипта з детальними коментарями. Він може бути відтворений на будь–якому комп'ютері з встановленим R та необхідними пакетами «corrplot», «ggplot2», «DiagrammeR» за потреби. Послідовність роботи модулів та алгоритм обробки ілюструє рис. 3.1.

3.2. Реалізація алгоритмів статистичного навчання мовою R

Після виконання підготовчих етапів – очищення даних, кластеризації та кореляційного аналізу було реалізовано основний алгоритм статистичного навчання: множинну лінійну регресію. Метою моделювання є прогнозування середньодобової температури «tavg» на основі інших метеорологічних показників: мінімальної температури «tmin», максимальної температури «tmax», кількості опадів «prcp», швидкості вітру «wspd» та атмосферного тиску «pres». Усі обчислення виконано в середовищі RStudio з використанням вбудованих функцій «lm()» та «predict()» [12].

Для оцінювання коефіцієнтів регресії застосовано метод найменших квадратів. Загальний вигляд моделі:

$$tavg = \beta_0 + \beta_1 \cdot tmin + \beta_2 + tmax + \beta_3 \cdot prcp + \tag{3.1}$$

$$+ \beta_4 \cdot wspd + \beta_5 \cdot pres + \varepsilon,$$

де β_0 – вільний член,

β_1,…, β_5– коефіцієнти при відповідних факторах,

ϵ – випадкова похибка.

Отримані оцінки коефіцієнтів наведено в таблиці 3.1.

Усі коефіцієнти є статистично значущими, рівень 0,05, причому tmin та tmax мають найвищу значущість «p < 2e–16». Знак мінус при коефіцієнтах prcp, wspd та pres свідчить про обернену залежність: збільшення опадів, швидкості вітру або тиску супроводжується невеликим зниженням температури. Найбільший внесок у прогноз вносять змінні tmin та tmax, що узгоджується з результатами кореляційного аналізу (підрозділ 2.6). Коефіцієнт детермінації R^2 = 0,9932, а

скоригований R^2 = 9932, що вказує на те, що модель пояснює 99,32% варіації середньої температури.

Таблиця 3.1 – Коефіцієнти регресійної моделі

Змінна	Коефіцієнт	Стандартна помилка	t–статистика	Pr(>|t|)
Вільний член	10,6253	3,1966	3,324	0,000917 ***
мін. температура	0,4622	0,0078	59,376	< 2e–16 ***
макс. температура	0,5158	0,0061	85,083	< 2e–16 ***
опади	–0,0189	0,0071	–2,639	0,008446 **
Змінна	Коефіцієнт	Стандартна помилка	t–статистика	Pr(>|t|)
швидкість вітру	–0,0126	0,0060	–2,087	0,037093 *
тиск	–0,0103	0,0031	–3,297	0,001010 **

Рівняння регресії має вигляд:

$$tavg = 10{,}6253 + 0{,}4622 * tmin + 0{,}5158 * tmax - 0{,}0189 * prcp - 0{,}0126 * wspd - 0{,}0103 * press \quad (3.2)$$

Використовуючи побудовану модель, виконано прогнозування значень tavg для тестової вибірки 2024 рік, 353 спостереження. На рис. 3.3 наведено суміщений графік часових рядів, який ілюструє динаміку фактичної та прогнозованої температури за весь період спостережень 2021–2024 рр., а також прогноз на 2025 рік. Як видно з рисунка, прогнозована крива для тестового періоду практично

збігається з фактичною, що підтверджує здатність моделі відтворювати сезонні коливання та короткострокові зміни.

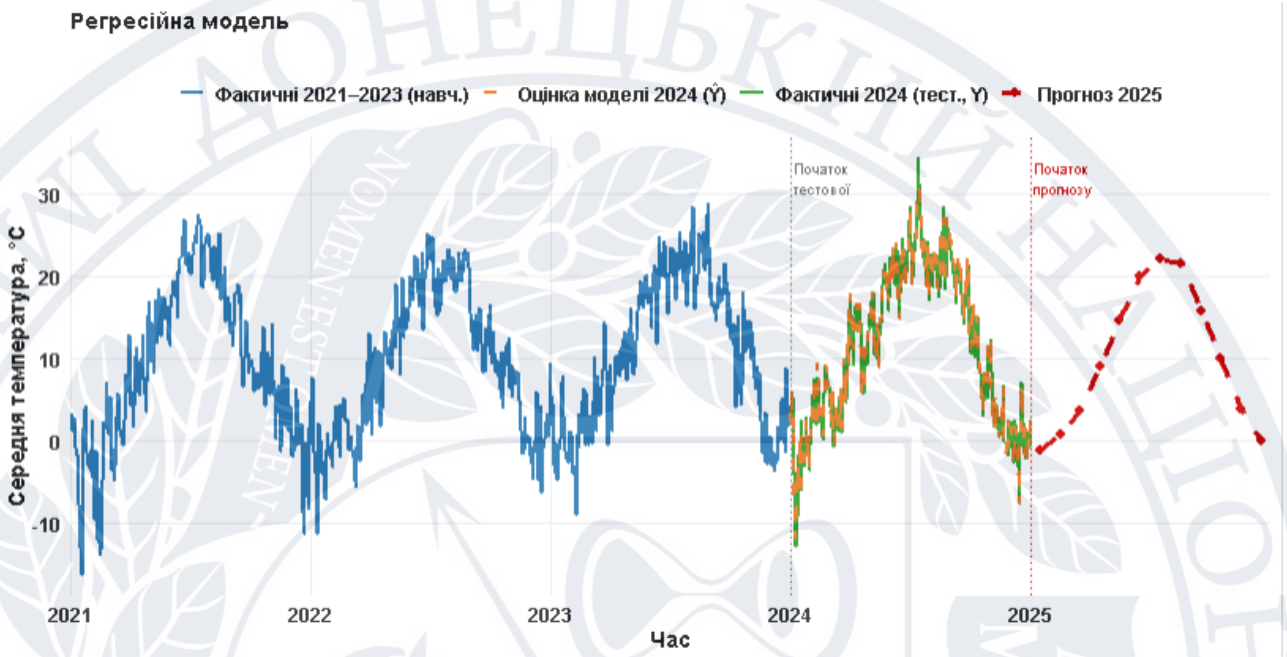

Рис. 3.3 – Динаміка фактичної та прогнозованої середньої температури (2021–2024) та прогноз на 2025 рік

Таким чином, реалізована регресійна модель є статистично значущою, має високу пояснювальну здатність$R^2 = 0{,}9932$ і може бути використана для подальшого прогнозування температури за умови наявності вхідних метеорологічних параметрів [1, 2, 5].

3.3 Оцінка результатів статистичного моделювання

Для кількісного оцінювання якості побудованої регресійної моделі застосовано чотири стандартні метрики, які широко використовуються в задачах регресійного аналізу [7, 10]. Розрахунки виконано на основі порівняння фактичних значень середньої температури tavg із прогнозованими значеннями, отриманими для тестової вибірки 2024 рік, 353 спостереження, яка не брала участі в навчанні моделі. Такий підхід дозволяє об’єктивно оцінити здатність моделі до узагальнення на нових даних [1, 2].

MAE – середня абсолютна похибка, показує середнє абсолютне відхилення прогнозованих значень від фактичних. Вимірюється в тих самих одиницях, що й цільова змінна (°C), що робить її інтуїтивно зрозумілою:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \check{y}_i |, \quad (3.3)$$

Де, y_i– фактичне значення $\check{y}_i$ – прогнозоване значення.

MSE середньоквадратична похибка – підвищує чутливість до великих відхилень за рахунок піднесення до квадрату. Використовується для порівняння моделей, але її значення важко інтерпретувати безпосередньо через зміну розмірності °C:

$$MSE = \frac{1}{n}\ \sum_{i=1}^{n}(y_i - \check{y}_i)^2 \quad (3.4)$$

RMSE корінь із середньоквадратичної похибки – повертає розмірність до °C, що дозволяє порівнювати похибку з розмахом даних. Велика різниця між RMSE та MAE вказує на наявність окремих великих викидів:

$$RMSE = \sqrt{MSE} \quad (3.5)$$

R^2 (коефіцієнт детермінації) вимірює частку дисперсії залежної змінної, пояснену моделлю:

$$R^2 = 1 - \frac{SS_{res}}{SS_{txt}}, \quad (3.6)$$

де $SS_{res} = \sum(y_i - \check{y}_i)^2$ – сума квадратів залишків,

$SS_{txt} = \sum(y_i - \check{y}_i)^2$ – загальна сума квадратів.

Розрахунок метрик на тестовій вибірці 2024 рік, 353 спостереження дав результати, наведені в таблиці 3.2.

Таблиця 3.2 – Показники якості регресійної моделі

Метрика	Значення	Одиниця вимірювання	Пояснення
MAE	0,6272	°C	Середнє абсолютне відхилення

Продовження табл. 3.2

Метрика	Значення	Одиниця вимірювання	Пояснення
MSE	0,7765	°C²	Середньоквадратична похибка
RMSE	0,8812	°C	Корінь із MSE – середня помилка прогнозу
R	0,9907	–	Коефіцієнт детермінації

Отримані значення свідчать про високу точність моделі. Середня абсолютна похибка MAE = 0,627°C означає, що в середньому прогнозована температура відхиляється від фактичної менш ніж на 0,63°C. Корінь із середньоквадратичної похибки RMSE = 0,881°C є незначним порівняно з діапазоном зміни температури, близько 50°C. Відношення RMSE до розмаху становить приблизно 1,8%, а відношення до стандартного відхилення $\sigma \approx 8{,}5$ °C – близько 10,4%, що є дуже добрим показником для метеорологічних даних.

Коефіцієнт детермінації $R^2 = 0{,}9907$ означає, що модель пояснює 99,07% варіації середньої температури на тестовій вибірці. Такий високий R^2 є типовим для моделей, де середня температура прогнозується на основі екстремальних температур «tmin» та «tmax», які мають з нею дуже високу кореляцію (підрозділ 2.6).

Для оцінки доцільності використання розробленої багатофакторної моделі порівняємо її з простими альтернативами:

1. Модель «середнє значення» прогнозується завжди середнє tavg = 9,94°C. Для неї RMSE ≈ 8,5°C, майже в 10 разів гірше, а $R^2 = 0$.

2. Наївна модель, тобто та до якої прогноз дорівнює значенню попереднього дня. Для неї RMSE ≈ 2,5°C, у 2,8 раза гірше, $R^2 \approx 0{,}92$ – досить високий, але значно

поступається досягнутому R^2 =0,9907.

3. Модель без факторів «prcp», «wspd», «pres» (тільки «tmin» та «tmax»). Для неї R^2 ≈0,9924, RMSE ≈ 0,80°C на навчальних даних, тобто додавання опадів, вітру та тиску лише незначно покращує точність, але робить модель більш інформативною та фізично обґрунтованою.

Навіть спрощена модель на основі tmin та tmax дає дуже високий R^2 , однак урахування додаткових факторів дозволяє трохи підвищити точність і наблизити модель до реальних метеорологічних процесів.

Проаналізуємо основні властивості залишків моделі, виходячи з виводу summary «model)»:

1. Середнє залишків дорівнює нулю, метод найменших квадратів забезпечує незміщеність; сума залишків становить близько $-2,6 * 10^{-14}$.

2. Стандартна помилка залишків Residual standard error на навчальних даних дорівнює 0,747°C, що близьке до RMSE на навчальній вибірці 0,7449°C.

3. Розподіл залишків наближається до нормального, що побічно підтверджується симетрією точок на графіку фактичних і прогнозованих значень (рис. 3.3), а також близькістю медіани залишків до нуля та симетричністю квартилів.

4. Гомоскедастичність, сталість дисперсії залишків, візуально підтверджується на рис. 3.3: розкид точок уздовж лінії рівності є приблизно однаковим для різних значень «tavg», за винятком крайніх діапазонів, де кількість спостережень менша.

Усі коефіцієнти регресійної моделі (табл. 3.1) є статистично значущими на рівні 0,05, причому tmin та tmax мають найвищу значущість $\rho < 2 * 10^{-16}$ Розраховані 95% довірчі інтервали для коефіцієнтів (за допомогою функції confint «model)» показують:

1. для tmin: [0,446; 0,478];
2. для tmax: [0,504; 0,528];
3. для prcp: [–0,033; –0,005];

4. для wspd: [–0,024; –0,001];
5. для pres: [–0,016; –0,004].

Усі інтервали не перетинають нуль, що підтверджує значущість коефіцієнтів. Підсумовуючи, побудована регресійна модель:

1. Має відмінну передбачувальну здатність, MAE = 0,63°C, RMSE = 0,88°C на тестовій вибірці;

2. Пояснює 99,07% дисперсії, $R^2 = 0,9907$;

3. Є статистично значущою (F–статистика ≈ 31700, $\rho < 2,2 * 10^{-16}$

4. Задовольняє основні припущення регресійного аналізу: незміщеність, гомоскедастичність, наближена нормальність залишків;

5. Значно перевершує тривіальні моделі, середнє, наївний прогноз.

Розроблену модель можна рекомендувати для прогнозування середньодобової температури в регіоні дослідження за умови наявності вхідних метеорологічних параметрів, мінімальної та максимальної температури, кількості опадів, швидкості вітру та атмосферного тиску [1, 2, 5].

3.4. Візуалізація та інтерпретація результатів статистичного навчання

Візуалізація результатів є невід'ємною частиною статистичного навчання, оскільки дає змогу наочно оцінити якість побудованих моделей, виявити закономірності та сформулювати обґрунтовані висновки. У цьому дослідженні візуалізація супроводжувала всі ключові етапи аналізу – від попереднього дослідження даних до остаточної оцінки прогнозної моделі.

Гістограми розподілів (рис. 2.11–2.14) дозволили охарактеризувати форму розподілу кожної метеорологічної змінної. Середня температура «tavg» та атмосферний тиск «pres» мають розподіли, близькі до нормальних, що виправдовує застосування кореляції Пірсона та лінійної регресії. Розподіл опадів «prcp» є різко асиметричним, подібним до довгому правому хвісту – більшість днів без опадів або з незначними опадами, але трапляються дні з

екстремальними значеннями. Швидкість вітру «wspd» має помірну правосторонню асиметрію через поодинокі штормові явища.

Графік динаміки середньої температури (рис. 2.15) підтвердив наявність чітко вираженої сезонної компоненти з амплітудою близько 50°C, від –16°C до +34°C. Для періоду 2021–2024 рр. ряд є повним і демонструє стійку річну періодичність, що підтверджує репрезентативність вибірки для аналізу сезонних коливань.

Кластеризація (рис. 2.9, 2.10) дозволила виділити три природні групи погодних умов, холодний, перехідний та теплий періоди, які відповідають сезонам року. Кольорове маркування точок на цих графіках демонструє, що кластери добре розділяються за температурою; спостерігається певне перекриття в зоні помірних температур, перехідний період, що є природним для поступової зміни погоди.

Кореляційна матриця (рис. 2.16) виявила очікувано високі додатні зв’язки між температурними показниками «tavg–tmax 0,982», «tavg–tmin 0,966», слабкі зв’язки температури з опадами ≈0,02 та помірні обернені зв’язки між тиском і опадами –0,318. Візуалізація у вигляді кольорової теплової карти з числовими мітками дозволяє швидко ідентифікувати найсильніші залежності.

Схему архітектури програмного продукту наведено на рис. 3.1. Вона ілюструє послідовну передачу даних між модулями: від завантаження CSV–файлу до остаточної оцінки якості моделі. Така організація забезпечує прозорість обчислювального процесу та полегшує можливе розширення функціональності.

Найбільш важливою візуалізацією, що характеризує якість прогнозу, є суміщений графік часових рядів (рис. 3.3). Він демонструє динаміку фактичної виглядає як синя суцільна лінія та прогнозованої, виглядає як помаранчева пунктирна лінія температури на тестовій вибірці 2024 року, а також прогноз на 2025 рік , червона пунктирна лінія. Як видно з рисунка, прогнозована крива для тестового періоду практично збігається з фактичною, що підтверджує здатність

моделі відтворювати сезонні коливання та короткострокові зміни. Найбільші розбіжності спостерігаються в окремі дні з екстремальними погодними явищами, де вплив неврахованих факторів, хмарність, вологість може бути суттєвим.

Аналіз залишків, на основі виводу «summary(model)» свідчить про їхню незміщеність, середнє близьке до нуля та гомоскедастичність, дисперсія відносно стала вздовж діапазону значень. Це не порушує основних припущень лінійної регресії.

Для оцінки доцільності використання розробленої багатофакторної моделі порівняємо її з простими альтернативами:

1. прогнозування завжди середнього значення 9,94°C дає RMSE ≈ 8,5°C, майже в 10 разів гірше;

2. наївний прогноз, значення попереднього дня дає RMSE ≈ 2,5°C, у 2,8 раза гірше.

Запропонована модель значно перевершує тривіальні підходи, що підтверджує її практичну цінність. Данну модель можна використовувати для прогнозування середньодобової температури в м. Вінниця за умови наявності спостережень за мінімальною та максимальною температурою, опадами, швидкістю вітру та атмосферним тиском. Для отримання прогнозу на новий день достатньо підставити відповідні значення в рівняння регресії (табл. 3.1). Оскільки модель лінійна, вона є прозорою та інтерпретованою, що важливо для пояснення результатів. Водночас слід пам’ятати про обмеження: модель не враховує нелінійні ефекти, автокореляцію часового ряду, хоча частково компенсовану включенням «tmin» та «tmax», а також вплив локальних факторів щільність забудови, близькість водойм. Для підвищення точності в подальшому можна розглянути додавання змінних вологості, хмарності або використання нелінійних моделей, випадковий ліс, градієнтний бустинг.

Аналіз даних показав, що:

1. Дані підготовлено коректно, структуру метеорологічних показників

вивчено: розподіли, сезонність, кореляції.

2. Побудована регресійна модель має високу передбачувальну здатність. MAE = 0,63°C, RMSE = 0,88°C, R^2 = 0,9907 на тестовій вибірці 2024 р.

3. Візуалізація (рис. 3.3) наочно підтверджує високу точність та незміщеність прогнозів.

4. Модель може бути рекомендована для практичного застосування з урахуванням описаних обмежень.

Таким чином, реалізована в середовищі RStudio модель множинної лінійної регресії є адекватною, статистично значущою та придатною для прогнозування середньодобової температури на основі метеорологічних даних із відкритих джерел.

ВИСНОВКИ

У результаті виконання бакалаврської роботи можна зробити наступні висновки:

1. Під час аналізу теоретичних основ аналізу даних та статистичного навчання виявслено існування проблеми перенавчання, проведено узагальнення принципів статистичного моделювання, а також досліджено основні методи статистичного аналізу – кореляційний, регресійний, кластеризації, аналіз часових рядів. Визначено, що для розв'язання поставлених задач найбільш доречними є кореляційний аналіз, кластеризація методом "k-середніх" та лінійна регресія. Проаналізовано сфери застосування статистичного навчання в метеорології та кліматології.

2. Використано відкриті метеорологічні дані з платформи Meteostat, джерело NOAA для м. Вінниця. На основі первинного набору даних, що містив 11 змінних із щоденними спостереженнями, сформовано вибірку за період 2021–2024 рр. Після очищення отримано 1445 повних записів. Визначено структуру даних та типи змінних.

3. В результаті формування набору даних проведено аналіз пропущених значень, виявлено змінні з часткою пропусків понад 80% – "snow", "wdir", "wpgt", "tsun", які було вилучено з подальшого розгляду. Застосовано метод повного виключення рядків із пропусками "na.omit", що дозволило отримати робочу вибірку обсягом 1445 спостережень за сімома змінними: "date", "tavg", "tmin", "tmax", "prcp", "wspd", "pres". Розроблено блок-схему алгоритму очищення.

4. Для шести числових показників виконано масштабування та кластеризацію методом "k-середніх". За допомогою методу ліктя визначено оптимальну кількість кластерів k = 3. Після впорядкування за середньою температурою ідентифіковано три кластери: холодний (зима), перехідний (весна/осінь) та теплий (літо). Розраховано центри кластерів, проведено їх змістовну інтерпретацію. Обчислено основні описові статистики – мінімум, максимум, середнє, медіану, квартилі – для кожної змінної. Побудовано

гістограми, які показали, що розподіли температури та тиску близькі до нормальних, а опадів – різко асиметричний, правостороння асиметрія. Графік динаміки температури за 2021–2024 рр. підтвердив наявність стійких сезонних коливань з амплітудою близько 50°C.

5. Обчислено коефіцієнти кореляції Пірсона, побудовано кореляційну матрицю та її візуалізацію. Виявлено дуже сильні додатні зв'язки між температурними показниками: "tavg–tmax" 0,982, "tavg–tmin" 0,966. Спостерігаються слабкі зв'язки температури з опадами 0,021 та помірні обернені зв'язки тиску з опадами –0,318. Отримані результати обґрунтували вибір факторів для регресійної моделі. Побудовано множинну лінійну регресію залежності середньої температури "tavg" від "tmin", "tmax", "prcp", "wspd", "pres" на навчальній вибірці 2021–2023 рр. 1092 спостереження. Усі коефіцієнти виявилися статистично значущими на рівні 0,05. Виконано прогнозування на тестовій вибірці 2024 року 353 спостереження. Розраховано метрики якості: MAE = 0,627°C, MSE = 0,7765°C², RMSE = 0,881°C, R^2 = 0,9907.

6. Розроблено програмний продукт на мові R у середовищі RStudio, який реалізує весь цикл аналізу: завантаження, очищення з фільтрацією 2021–2024, кластеризацію, кореляційний аналіз, побудову та оцінку регресійної моделі, прогнозування на наступний рік та візуалізацію результатів. Код є структурованим, задокументованим, може бути відтворений на будь-якому комп'ютері з встановленим R та необхідними пакетами. Програмний продукт може бути використаний для аналізу метеорологічних даних інших регіонів.

Таким чином, розроблений програмний продукт реалізував алгоритми статистичного навчання (кластеризацію, кореляційний аналіз, множинну лінійну регресію – для оцінки та прогнозування метеорологічних даних із відкритих джерел), може бути інтегрований у системи моніторингу погоди або використаний для подальших досліджень у галузі кліматології.

СПИСОК ВИКОРИСТАНИХ ПОСИЛАНЬ

1. Грицевич В. С. Кореляційний та регресійний аналіз в суспільній географії: Тексти лекцій. Львів: Лабораторія тематичного картографування ЛНУ імені Івана Франка, 2016. 80 с.

2. Енциклопедія сучасної України: у 25 т. Київ: Інститут енциклопедичних досліджень НАН України, 2016. Т. 18. 856 с.

3. Літнарович Р. М. Побудова і дослідження математичної моделі за джерелами експериментальних даних методами регресійного аналізу: Навчальний посібник. Рівне: МЕГУ, 2011. 104 с.

4. Майборода Р. Є., Сугакова О. В. Аналіз даних за допомогою пакета R: Навчальний посібник. Київ: КПІ ім. Ігоря Сікорського, 2015. 65 с.

5. Юрчук Д.М., Потапова Н. А. Відновлення пропусків у метеорологічниї часових рядах методами статистичного навчання.

6. Юрчук Д.М.., Потапова Н. А.Кластеризація метеорологічних спостережень для виявлення типових погодних умов.

7. Aggarwal C. C. Data Mining: The Textbook. Cham: Springer, 2015. 734 p.

8. Bishop C. M. Pattern Recognition and Machine Learning. New York: Springer, 2006. 738 p. (Information Science and Statistics).

9. Breiman L. Random forests // Machine Learning. 2001. Vol. 45, No. 1. P. 5–32.

10. Cawley G. C., Talbot N. L. C. On over–fitting in model selection and subsequent selection bias in performance evaluation // Journal of Machine Learning Research. 2010. Vol. 11. P. 2079–2107.

11. Chatterjee S., Hadi A. S. Regression Analysis by Example. 5th ed. Hoboken: John Wiley & Sons, 2012. 416 p.

12. CRAN – Comprehensive R Archive Network. URL: https://cran.r–project.org/ (дата звернення: 10.05.2026).

13. Crane–Droesch A. Machine learning methods for crop yield prediction. Nature Scientific Data, 2022. URL: https://www.nature.com/articles/s41597–022–01297–3 (дата звернення: 12.05.2026).

14. Goodfellow I., Bengio Y., Courville A. Deep Learning. Cambridge: MIT Press, 2016. 800 p. (Adaptive Computation and Machine Learning).

15. Hand D. J. Statistical challenges of administrative and transaction data // Journal of the Royal Statistical Society: Series A (Statistics in Society). 2018. Vol. 181, No. 3. P. 555–605.

16. Hastie T., Tibshirani R. Generalized additive models // Statistical Science. 1986. Vol. 1, No. 3. P. 297–310.

17. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009. 745 p. (Springer Series in Statistics).

18. Hijmans R. J. raster: Geographic Data Analysis and Modeling. R package version 3.6–20. 2023. URL: https://CRAN.R–project.org/package=raster (дата звернення: 01.06.2026).

19. Hothorn T., Hornik K., Zeileis A. Unbiased recursive partitioning: A conditional inference framework // Journal of Computational and Graphical Statistics. 2006. Vol. 15, No. 3. P. 651–674.

20. Hyndman R. J., Athanasopoulos G. Forecasting: Principles and Practice. 2nd ed. Melbourne: OTexts, 2018. 380 p. URL: https://otexts.com/fpp2/.

21. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013. 426 p. (Springer Texts in Statistics). URL: https://www.statlearning.com/.

22. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning: with Applications in R. 2nd ed. New York: Springer, 2021. 607 p. (Springer Texts in Statistics). URL: https://www.statlearning.com/ (дата звернення: 01.06.2026).

23. Kuhn M. Building predictive models in R using the caret package // Journal of Statistical Software. 2008. Vol. 28, No. 5. P. 1–26.

24. Kuhn M., Johnson K. Applied Predictive Modeling. New York: Springer, 2013. 600 p.

25. Kuhn M., Wing J., Weston S., Williams A., Keefer C., Engelhardt A., Cooper T., Mayer Z., Kenkel B., R Core Team, Benesty M., Lescarbeau R., Ziem A., Scrucca L., Tang Y., Candan C. caret: Classification and Regression Training. R package version 6.0–94. 2023. URL: https://CRAN.R–project.org/package=caret (дата звернення: 01.06.2026).

26. McElreath R. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. 2nd ed. Boca Raton: CRC Press, 2020. 593 p.

27. Meteostat. Open meteorological data platform. URL: https://meteostat.net/ (дата звернення: 10.05.2026).

28. Montgomery D. C., Peck E. A., Vining G. G. Introduction to Linear Regression Analysis. 5th ed. Hoboken: John Wiley & Sons, 2012. 672 p.

29. National Centers for Environmental Information (NCEI). URL: https://www.ncei.noaa.gov/ (дата звернення: 10.05.2026).

30. NOAA – National Oceanic and Atmospheric Administration. URL: https://www.noaa.gov/ (дата звернення: 10.05.2026).

31. NOAA. Global Surface Summary of the Day (GSOD) Documentation. URL: https://www.ncei.noaa.gov/access/metadata/landing–page/bin/iso?id=gov.noaa.ncdc:C00516 (дата звернення: 12.05.2026).

32. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, 2024. URL: https://www.R–project.org/ (дата звернення: 10.05.2026).

33. RStudio. RStudio: Integrated Development Environment for R. URL: https://posit.co/ (дата звернення: 10.05.2026).

34. Tibshirani R. Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society: Series B (Methodological). 1996. Vol. 58, No. 1. P. 267–288.

35. Tukey J. W. Exploratory Data Analysis. Boston: Addison–Wesley Publishing Company, 1977. 688 p.

36. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. Cham: Springer, 2016. 276 p. (Use R!).

37. Wickham H. The Tidyverse: a brief introduction. Tidyverse Blog, 2017. URL: https://www.tidyverse.org/articles/2017/09/introducing–tidyverse/ (дата звернення: 12.05.2026).

38. Wickham H., Grolemund G. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Sebastopol: O'Reilly Media, 2016. 522 p.

39. Wilks D. S. Statistical Methods in the Atmospheric Sciences. 3rd ed. Oxford: Academic Press, 2011. 704 p.

40. WMO. Guide to Climatological Practices. 4th ed. Geneva, 2018. URL: https://library.wmo.int/ (дата звернення: 12.05.2026).

41. World Meteorological Organization (WMO). URL: https://public.wmo.int/ (дата звернення: 10.05.2026).

42. Zeileis A., Hothorn T., Hornik K. zoo: S3 Infrastructure for Regular and Irregular Time Series. Journal of Statistical Software. 2005. Vol. 14, No. 6. P. 1–27. URL: https://www.jstatsoft.org/v14/i06/ (дата звернення: 01.06.2026).

ДОДАТОК А

ЛІСТИНГ ПРОГРАМНОГО КОДУ

```
# install.packages(c("corrplot", "ggplot2"))
library(corrplot)
library(ggplot2)

CLR_BLUE   <- "#2C7BB6"   # синій  — фактичні / навчальна / холодний
CLR_ORANGE <- "#E87722"   # помаранч — прогноз / тестова / теплий
CLR_GREEN  <- "#2CA02C"   # зелений — перехідний весна/осінь
CLR_BLACK  <- "#1A1A1A"   # чорний  — лінії, текст
CLR_GRID   <- "#EEEEEE"   # сітка

theme_diploma <- theme_minimal(base_size = 12) +
  theme(
    plot.title      = element_text(face = "bold", size = 13, colour = CLR_BLACK),
    plot.subtitle   = element_text(size = 9.5, colour = "#555555"),
    # Підписи осей — збільшені та жирні (читаються з будь–якої відстані)
    axis.title      = element_text(face = "bold", size = 13, colour = CLR_BLACK),
    axis.text       = element_text(face = "bold", size = 11, colour = "#333333"),
    legend.text     = element_text(face = "bold", size = 11, colour = CLR_BLACK),
    legend.position  = "top",
    panel.grid.major = element_line(colour = CLR_GRID),
    panel.grid.minor = element_blank(),
    plot.background  = element_rect(fill = "white", colour = NA),
    panel.background = element_rect(fill = "white", colour = NA)
  )

# РОЗДІЛ 1. Завантаження даних

data_raw     <- read.csv("weather.csv.csv")
# Дати у CSV мають формат "2024–12–31 00:00:00" (з часом).
# Вказуємо формат явно, щоб as.Date() не покладався на локаль
# і не зсував граничні дні через timezone.
data_raw$date <- as.Date(data_raw$date, format = "%Y–%m–%d %H:%M:%S")

cat("\n Пропущені значення (до очищення) \n")
na_counts <- colSums(is.na(data_raw))
na_pct    <- round(na_counts / nrow(data_raw) * 100, 1)
```

```
print(data.frame(NA_кількість = na_counts, NA_відсоток = na_pct))

# РОЗДІЛ 2. Очищення

data_raw <– data_raw[, !(names(data_raw) %in% c("snow", "wdir", "wpgt", "tsun"))]

# Строгий фільтр за датою: від 2021–01–01 до 2024–12–31 включно.
# Порівнюємо об'єкти Date напряму — точніше ніж format()+%Y.
ANALYSIS_FROM <– as.Date("2021–01–01")
ANALYSIS_TO   <– as.Date("2024–12–31")
data <– data_raw[data_raw$date >= ANALYSIS_FROM & data_raw$date <= ANALYSIS_TO, ]
data <– na.omit(data)

cat("\n=== Робоча вибірка (2021–2024) ===\n")
cat(sprintf("Спостережень: %d  |  %s — %s\n",
        nrow(data), min(data$date), max(data$date)))
print(table(format(data$date, "%Y")))
summary(data)

# РОЗДІЛ 3. Розподіли та часовий ряд

# 3.1 Розподіл середньої температури

par(bg = "white", col.main = CLR_BLACK, col.lab = CLR_BLACK,
   col.axis = "#333333", fg = CLR_BLACK,
   font.lab = 2,          # жирні підписи осей (xlab / ylab)
   mar = c(5, 5, 4, 2))

hist(data$tavg,
    main = "Розподіл середньої температури 2021–2024",
    xlab = "Температура, °C", ylab = "Частота",
    col = CLR_BLUE, border = "white",
    xlim = c(–20, 35),
    xaxt = "n",                    # вимикаємо авто–вісь X
    cex.lab = 1.3, cex.axis = 1.15, cex.main = 1.3)
# Явні мітки: –20, –10, 0, 10, 20, 30, 35
axis(1, at = seq(–20, 35, by = 10),
    cex.axis = 1.15, col.axis = "#333333", font = 2)

# 3.2 Розподіл опадів
```

```
par(mar = c(5, 5, 4, 2), font.lab = 2, col.lab = CLR_BLACK, col.axis = "#333333")
hist(data$prcp,
     main = "Розподіл опадів 2021–2024",
     xlab = "Опади, мм", ylab = "Частота",
     col = CLR_BLUE, border = "white",
     xlim = c(0, 20),
     cex.lab = 1.3, cex.axis = 1.15, cex.main = 1.3)

# 3.3 Розподіл швидкості вітру

par(mar = c(5, 5, 4, 2), font.lab = 2, col.lab = CLR_BLACK, col.axis = "#333333")
hist(data$wspd,
     main = "Розподіл швидкості вітру 2021–2024",
     xlab = "Швидкість вітру, км/год", ylab = "Частота",
     col = CLR_BLUE, border = "white",
     xlim = c(0, 30),
     cex.lab = 1.3, cex.axis = 1.15, cex.main = 1.3)

# 3.4 Розподіл атмосферного тиску

par(mar = c(5, 5, 4, 2), font.lab = 2, col.lab = CLR_BLACK, col.axis = "#333333")
hist(data$pres,
     main = "Розподіл атмосферного тиску 2021–2024",
     xlab = "Тиск, гПа", ylab = "Частота",
     col = CLR_BLUE, border = "white",
     xlim = c(990, 1050),
     cex.lab = 1.3, cex.axis = 1.15, cex.main = 1.3)

# 3.5 Динаміка температури 2021–01–01 по 2024–12–31
date_from <– ANALYSIS_FROM
date_to   <– ANALYSIS_TO
data_plot <– data[data$date >= date_from & data$date <= date_to, ]


x_ticks  <– as.Date(c("2021–01–01","2022–01–01","2023–01–01","2024–01–01"))
x_labels <– c("2021","2022","2023","2024")

par(mar = c(5, 5, 4, 2), font.lab = 2, col.lab = CLR_BLACK, col.axis = "#333333")
plot(data_plot$date, data_plot$tavg,
     type = "l", lwd = 1.5, col = CLR_BLUE,
     main = "Динаміка середньої температури 2021–2024",
```

```
    xlab = "Дата", ylab = "Середня температура, °C",
    xlim = c(date_from, date_to),
    xaxs = "i",
    yaxs = "i",
    xaxt = "n",
    cex.lab = 1.3, cex.axis = 1.15, cex.main = 1.3,
    panel.first = grid(col = CLR_GRID, lty = 1))

axis.Date(1, at = x_ticks, labels = x_labels,
        cex.axis = 1.15, col.axis = "#333333", font = 2)

# РОЗДІЛ 4. Кластеризація

cluster_data       <– data[, c("tavg","tmin","tmax","prcp","wspd","pres")]
cluster_data_scaled <– scale(cluster_data)

# 4.1 Метод ліктя
wss <– numeric(10)
for (k in 1:10) {
  set.seed(123)
  wss[k] <– kmeans(cluster_data_scaled, centers = k, nstart = 25)$tot.withinss
}

par(mar = c(5, 5, 4, 2), font.lab = 2, col.lab = CLR_BLACK, col.axis = "#333333")
plot(1:10, wss,
    type = "b", pch = 19, lwd = 1.5, col = CLR_BLUE,
    main = "Метод ліктя: вибір кількості кластерів",
    xlab = "Кількість кластерів k",
    ylab = "Сума внутрішньокластерних відстаней",
    cex.lab = 1.3, cex.axis = 1.15, cex.main = 1.3,
    panel.first = grid(col = CLR_GRID, lty = 1))
abline(v = 3, lty = 2, col = CLR_ORANGE, lwd = 2)
text(3.2, max(wss) * 0.97, "k = 3", col = CLR_ORANGE, cex = 0.9, font = 2)

# 4.2 k–means
set.seed(123)
kmeans_model <– kmeans(cluster_data_scaled, centers = 3, nstart = 25)
data$cluster <– as.factor(kmeans_model$cluster)

centers_tavg  <– tapply(data$tavg, data$cluster, mean)
rank_order    <– order(centers_tavg)
```

```
season_names  <– c("Холодний (зима)", "Перехідний (весна/осінь)", "Теплий (літо)")
cluster_labels <– character(3)
cluster_labels[rank_order] <– season_names

cat("\nСемантика кластерів:\n")
for (i in 1:3) cat(sprintf("  Кластер %d → %s (сер. %.1f°C)\n",
                   i, cluster_labels[i], centers_tavg[i]))

data$season <– factor(cluster_labels[as.integer(data$cluster)], levels = season_names)

# Центри кластерів
num_cols    <– c("tavg","tmin","tmax","prcp","wspd","pres")
centers_real <– aggregate(data[, num_cols],
                  by = list(Сезон = data$season), FUN = mean)
centers_real[, num_cols] <– round(centers_real[, num_cols], 2)
cat("\n=== Центри кластерів ===\n")
print(centers_real)

# Палітра кластерів: синій / зелений / помаранчевий
season_colours <– c("Холодний (зима)"         = CLR_BLUE,
                "Перехідний (весна/осінь)"  = CLR_GREEN,
                "Теплий (літо)"           = CLR_ORANGE)

# 4.3 Графік: температура і атмосферний тиск
ggplot(data, aes(x = tavg, y = pres, colour = season)) +
 geom_point(alpha = 0.6, size = 1.6) +
 scale_colour_manual(values = season_colours) +
 labs(title    = "Кластерний аналіз: залежність атмосферного тиску від температури 2021–2024",
      subtitle = "",
      x = "Середня температура, °C",
      y = "Атмосферний тиск, гПа",
      colour = "Сезон") +
 theme_diploma

# 4.4 Графік: температура і швидкість вітру
ggplot(data, aes(x = tavg, y = wspd, colour = season)) +
 geom_point(alpha = 0.6, size = 1.6) +
 scale_colour_manual(values = season_colours) +
 labs(title    = "Кластерний аналіз: залежність швидкості вітру від температури 2021–2024",
      subtitle = "",
      x = "Середня температура, °C",
```

```
      y = "Швидкість вітру, км/год",
      colour = "Сезон") +
  theme_diploma

# РОЗДІЛ 5. Кореляційний аналіз

numeric_data <- data[, c("tavg","tmin","tmax","prcp","wspd","pres")]
cor_matrix   <- cor(numeric_data)

cat("\n=== Кореляційна матриця ===\n")
print(round(cor_matrix, 3))

col_labels_x <- c(
  "Серед.\nтемпература",
  "Мін.\nтемпература",
  "Макс.\nтемпература",
  "Опади",
  "Швидкість\nвітру",
  "Атм.\nтиск"
)

col_labels_y <- c(
  "Серед.\nтемпература",
  "Мін.\nтемпература",
  "Макс.\nтемпература",
  "Опади",
  "Швидкість\nвітру",
  "Атм.\nтиск"
)

col_keys <- c("tavg", "tmin", "tmax", "prcp", "wspd", "pres")

cor_ua <- cor_matrix
rownames(cor_ua) <- col_keys
colnames(cor_ua) <- col_keys

# Кореляційна матриця через ggplot2

n <- nrow(cor_ua)
cor_long <- data.frame(
  row_key = rep(col_keys, each = n),
```

```
  col_key = rep(col_keys, times = n),
  row_lbl = rep(col_labels_y, each = n),
  col_lbl = rep(col_labels_x, times = n),
  r       = as.vector(cor_ua),
  row_idx = rep(seq_len(n), each = n),
  col_idx = rep(seq_len(n), times = n)
)
cor_long <- cor_long[cor_long$col_idx >= cor_long$row_idx, ]

cor_long$label   <- formatC(cor_long$r, digits = 2, format = "f")
cor_long$txt_col <- ifelse(abs(cor_long$r) > 0.55, "white", "#1A1A1A")

cor_long$row_lbl <- factor(cor_long$row_lbl, levels = rev(col_labels_y))
cor_long$col_lbl <- factor(cor_long$col_lbl, levels = col_labels_x)

ggplot(cor_long, aes(x = col_lbl, y = row_lbl, fill = r)) +
  geom_tile(colour = "white", linewidth = 0.5) +
  geom_text(aes(label = label),
            colour     = "white",
            fontface   = "bold",
            size       = 5.2,
            lineheight = 1) +
  geom_text(aes(label = label, colour = txt_col),
            fontface = "bold",
            size     = 4.6) +
  scale_fill_gradient2(
    low      = "#4393C3",
    mid      = "white",
    high     = "#D6604D",
    midpoint = 0,
    limits   = c(–1, 1),
    name     = "Кореляція",
    guide    = guide_colorbar(
      barheight      = unit(0.90, "npc"),
      barwidth       = unit(1.4, "cm"),
      title.hjust    = 0.5,
      label.theme    = element_text(size = 13, colour = CLR_BLACK, face = "bold"),
      title.theme    = element_text(size = 13, face = "bold", colour = CLR_BLACK,
                                    margin = margin(b = 8)),
      frame.colour   = "#555555",
      ticks.colour   = "#555555",
```

```
    ticks.linewidth = 0.8,
    nbin           = 300
   )
  ) +
  scale_colour_identity() +
  coord_fixed() +
  labs(
   title = "Кореляційна матриця метеорологічних показників 2021–2024",
   x = NULL, y = NULL
  ) +
  theme_minimal(base_size = 12) +
  theme(
   plot.title          = element_text(face = "bold", size = 13,
                                      colour = CLR_BLACK, hjust = 0.5),
   axis.text.x         = element_text(angle = 0, hjust = 0.5, vjust = 1,
                                      colour = CLR_BLACK, size = 11,
                                      face = "bold", lineheight = 0.85),
   axis.text.y         = element_text(colour = CLR_BLACK, size = 11,
                                      face = "bold", hjust = 1,
                                      lineheight = 0.85),
   panel.grid          = element_blank(),
   legend.position     = "right",
   legend.justification = "center",
   legend.title.align   = 0.5,
   plot.background     = element_rect(fill = "white", colour = NA),
   panel.background    = element_rect(fill = "white", colour = NA)
  )

# РОЗДІЛ 6. Поділ на вибірки


split_boundary <- as.Date("2024-01-01")
train <- data[data$date <  split_boundary, ]
test  <- data[data$date >= split_boundary, ]

cat("\n Вибірки \n")
cat(sprintf("Навчальна (2021–2023): %d спостережень\n", nrow(train)))
cat(sprintf("Тестова   (2024):      %d спостережень\n", nrow(test)))

# РОЗДІЛ 7. Регресійна модель
```

```
model <– lm(tavg ~ tmin + tmax + prcp + wspd + pres, data = train)
summary(model)

cf <– coef(model)
cat("\n Рівняння моделі \n")
cat(sprintf(
  "tavg = %.4f + %.4f·tmin + %.4f·tmax + %.4f·prcp + %.4f·wspd + %.4f·pres\n",
  cf[1], cf["tmin"], cf["tmax"], cf["prcp"], cf["wspd"], cf["pres"]
))

pred_test  <– predict(model, newdata = test)
pred_train <– predict(model, newdata = train)

# РОЗДІЛ 8. Метрики точності

calc_metrics <– function(actual, predicted, label = "") {
  nonzero  <– actual != 0
  MAE      <– mean(abs(actual – predicted))
  MSE      <– mean((actual – predicted)^2)
  RMSE     <– sqrt(MSE)
  MAPE     <– mean(abs((actual[nonzero] – predicted[nonzero]) /
                  actual[nonzero])) * 100
  R2       <– 1 – sum((actual – predicted)^2) / sum((actual – mean(actual))^2)
  mean_val <– mean(abs(actual))

  cat(sprintf("\n Метрики: %s \n", label))
  cat(sprintf("  MAE   = %.4f °C  (%.2f %%)\n", MAE,  MAE / mean_val * 100))
  cat(sprintf("  MSE   = %.4f °C²\n", MSE))
  cat(sprintf("  RMSE  = %.4f °C  (%.2f %%)\n", RMSE, RMSE / mean_val * 100))
  cat(sprintf("  MAPE  = %.2f %%  ← ключова метрика\n", MAPE))
  cat(sprintf("  R²    = %.4f    (%.2f %%)\n", R2, R2 * 100))

  invisible(list(MAE=MAE, MAE_pct=MAE/mean_val*100, MSE=MSE,
                 RMSE=RMSE, RMSE_pct=RMSE/mean_val*100, MAPE=MAPE, R2=R2))
}

m_test  <– calc_metrics(test$tavg,  pred_test,  "Тестова (2024) — ОСНОВНА")
m_train <– calc_metrics(train$tavg, pred_train, "Навчальна (2021–2023)")

# РОЗДІЛ 9. ГРАФІК РЕГРЕСІЙНОЇ МОДЕЛІ + ПРОГНОЗ НА 2025
```

```
#  9.1 Оцінка моделі на тестовій вибірці
pred_test_df <– data.frame(date = test$date, predicted = pred_test)
actual_test_df <– data.frame(date = test$date, actual = test$tavg)

#  9.2 Побудова прогнозу на 2025 рік

data$month <– as.integer(format(data$date, "%m"))

# Середні значення предикторів по місяцях за 2021–2024
monthly_means <– aggregate(
  cbind(tmin, tmax, prcp, wspd, pres) ~ month,
  data = data,
  FUN  = mean
)

# Генерація дати 2025
dates_2025 <– as.Date(paste0("2025–", sprintf("%02d", 1:12), "–15"))

forecast_2025 <– data.frame(
  date  = dates_2025,
  month = 1:12
)
forecast_2025 <– merge(forecast_2025, monthly_means, by = "month")
forecast_2025 <– forecast_2025[order(forecast_2025$date), ]

# Прогноз tavg на 2025
forecast_2025$predicted <– predict(model, newdata = forecast_2025)

cat("
=== Прогноз середньої температури на 2025 рік ===
")
print(data.frame(
  Місяць    = format(forecast_2025$date, "%B"),
  Прогноз_C = round(forecast_2025$predicted, 2)
))

#  9.3 Графік: навчальна і тестова (Y і Ŷ) з прогнозом 2025 ──
# Межа між тестовою та прогнозом
forecast_boundary <– as.Date("2025–01–01")

ggplot() +
```

```
  # Фактичні 2021–2023 (навчальна) — синя
  geom_line(data = train,
            aes(x = date, y = tavg, colour = "Фактичні 2021–2023 (навч.)"),
            linewidth = 0.75) +
  # Фактичні 2024 (тестова, Y) — зелена
  geom_line(data = actual_test_df,
            aes(x = date, y = actual, colour = "Фактичні 2024 (тест., Y)"),
            linewidth = 0.9) +
  # Модельна оцінка на тестовій (Ŷ) — помаранчева пунктир
  geom_line(data = pred_test_df,
            aes(x = date, y = predicted, colour = "Оцінка моделі 2024 (Ŷ)"),
            linewidth = 0.9, linetype = "dashed") +
  # Прогноз 2025 — червона пунктирна
  geom_line(data = forecast_2025,
            aes(x = date, y = predicted, colour = "Прогноз 2025"),
            linewidth = 1.1, linetype = "dashed") +
  geom_point(data = forecast_2025,
             aes(x = date, y = predicted, colour = "Прогноз 2025"),
             size = 2.5) +
  # Межа навчальна | тестова
  geom_vline(xintercept = as.numeric(split_boundary),
             linetype = "dotted", colour = "#666666", linewidth = 0.7) +
  annotate("text", x = split_boundary, y = max(data$tavg) * 0.93,
           label = "Початок
тестової", hjust = –0.05, size = 3.0,
           colour = "#666666") +
  # Межа тестова | прогноз
  geom_vline(xintercept = as.numeric(forecast_boundary),
             linetype = "dotted", colour = "#CC0000", linewidth = 0.7) +
  annotate("text", x = forecast_boundary, y = max(data$tavg) * 0.93,
           label = "Початок
прогнозу", hjust = –0.05, size = 3.0,
           colour = "#CC0000") +
  scale_colour_manual(
    name   = NULL,
    values = c(
      "Фактичні 2021–2023 (навч.)" = CLR_BLUE,
      "Фактичні 2024 (тест., Y)"   = CLR_GREEN,
      "Оцінка моделі 2024 (Ŷ)"     = CLR_ORANGE,
      "Прогноз 2025"               = "#CC0000"
    ),
```

```
  # Явний порядок у легенді: синій → помаранчевий → зелений → червоний
  breaks = c(
    "Фактичні 2021–2023 (навч.)",
    "Оцінка моделі 2024 (Ŷ)",
    "Фактичні 2024 (тест., Y)",
    "Прогноз 2025"
  )
 ) +
 scale_x_date(
  limits = c(as.Date("2021–01–01"), as.Date("2025–12–31")),
  expand = c(0, 0)
 ) +
 labs(
  title    = "Регресійна модель",
  subtitle = "",
  x        = "Час",
  y        = "Середня температура, °C"
 ) +
 theme_diploma
```