МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ВАСИЛЯ СТУСА

Сугак Глєб Васильович

Допускається до захисту:
завідувач кафедри
інформаційних технологій
к.т.н., доцент
_____ Т.В. Нескородєва
«___»_____ 2021 р.

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ МЕТЕОРОЛОГІЧНИХ ДАНИХ ДЛЯ ДОСЛІДЖЕННЯ ПОГОДИ

Спеціальність 122 Комп'ютерні науки

Кваліфікаційна (магістерська) робота

Науковий керівник:
Антонов Ю. С., доцент кафедри
інформаційних технологій,
к.ф-м.н., доцент
_____
(підпис)

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

Голова ЕК: _____
(підпис)

Вінниця 2021

АНОТАЦІЯ

Сугак Г. В. Інтелектуальний аналіз метеорологічних даних для дослідження погоди. Спеціальність 122 «Комп’ютерні науки», Освітня програма «Комп'ютерні технології обробки даних (Data Science)». Донецький національний університет імені Василя Стуса, Вінниця, 2021.

У кваліфікаційній роботі досліджено статистичні методи для прогнозу погоди.

Показано недоліки та переваги різних статистичних методів при роботі з метеорологічними даними.

Реалізовано основні статистичні методи для прогнозу погоди.

Ключові слова: прогноз погоди, аналіз даних, метеорологія, Data Science.

Табл. 3. Рис. 22, Бібліограф.: 55 найм.

ANNOTATION

Suhak H. V. Mining meteorological data for weather research. Speciality 122, «Computer Science», Programe «Data Science», Vasyl' Stus Donetsk National University, Vinnytsia, 2021.

In the qualification work statistical methods for weather forecasting are investigated.

The disadvantages and advantages of different statistical methods in working with meteorological data are shown.

The basic statistical methods for weather forecasting are implemented.

Keywords: weather forecast, meteorology, Data Science.

Tabl 3. Fig. 22. Bibliography: 55 items.

ЗМІСТ

ВСТУП

Зміна клімату та поточні зміни погоди стосуються різних секторів, таких як сільське, лісове, міське та регіональне планування, охорона природи, управління водними ресурсами, енергопостачання та туризм. Вплив зміни клімату вже можна спостерігати в багатьох місцях, і він неминуче відчуватиметься в майбутньому [50].

Статистичні методи та процедури оцінки відіграють ключову роль, коли відбувається робота з великими обсягами даних. Ці методи можуть бути різними для питань, пов'язаних з управлінням підземними водами та річками, сильними дощами або економічними наслідками. Крім того, якісні характеристики результатів можна визначити лише за допомогою статистичного аналізу, наприклад тести значущості або стійкості.

Іншою важливою темою є аналіз екстремальних значень, оскільки вони мають сильний вплив на результат аналізу даних при використанні звичайних статистичних методів, але їх важко аналізувати. Для аналізу метеорологічних даних використовується багато різних статистичних методів та процедур, деякі з яких описані у даній роботі.

Актуальність дослідження зумовлена стрімким ростом кількості мікроконтролерів, які застосовуються у тому числі в якості невеликих метеостанцій [51]. З'являється все більше даних, які потребують статистичного аналізу для прогнозування погоди та для планування подальших дій у той чи іншій сфері. Статистичні методи та їх програмна реалізація допомагає аналізувати метеодані та розробляти прогноз погоди, що має можливість застосування у великій кількості сфер життєдіяльності людини, які залежать від погодних умов.

Мета дослідження. Реалізація статистичних методів для аналізу метеоданих та аналіз різних статистичних методів для прогнозу погоди.

Завдання досліждення: Розібратись в сучасних тенденціях прогнозу погоди, алгоритмах та методиках для аналізу метеоданих, дослідити різні метеопоказники, що впливають на один одного, зробити формалізацію задачі, побудувати математичну модель та розробити програмне забезпечення аналізу та прогнозування метеорологічних показників.

У данній магістрескій роботі реалізовані такі **цілі**:

- Ознайомлення зі статистичними методами для аналізу метеоданих
- Порівняння різних методів для прогнозу погоди
- Побудова математичної моделі
- Програмна реалізація статистичних методів аналізу даних

Об'єкт дослідження. Набір метеорологічних даних в Австралії за 2008-2017 роки.

Предмет дослідження дипломної роботи. Статистичні методи для прогнозу погоди.

Методи дослідження: Вивчення існуючих статистичних методів для аналізу метеоданих, створення додатку на Python.

Теоретична та практична значущість дослідження полягає в реалізації статистичних методів для використання їх на наборі метеорологічних даних та підведення висновків щодо їх ефективності у тих чи інших випадках.

Структура роботи. Магістерська робота складається зі вступу, трьох розділів та висновків до них, списку використаних джерел та двох додатків.

У першому розділі магістерської роботи приведено актуальність та сфери застосування прогнозу погоди, сучасний стан розвитку метеорології та огляд основних статистичних методів.

У другому розділі було описано формалізацію задач та математичну модель для розробки програмного забезпечення.

У третьому розділі описано розробку програмної реалізації статистичних методів. Також впроваджено розроблений власноруч метод для прогнозу погоди за метеопоказниками, між якими існує сильна кореляція.

Дипломна робота включає в себе 77 сторінок, 22 рисунків і список літератури з 55 джерел.

РОЗДІЛ 1.
ЗАГАЛЬНІ ВІДОМОСТІ ПРО АНАЛІЗ МЕТЕОРОЛОГІЧНИХ ДАНИХ

Цей розділ містить загальні відомості про метеорологію, її актуальність та методи, які можуть бути застосовані до даних спостережень погоди. Представлені статистичні методи були визнані корисними дослідниками клімату для вивчення різних властивостей кліматичної системи.

Клімат - це парадигма складної системи. Вона має багато змінних, які діють нелінійно в широкому діапазоні просторово-часових шкал. Математичні моделі імітують клімат і його наслідки. Статистичні методи використовують вихідні дані моделі для виведення властивостей кліматичної системи.

1.1 Важливість якісного прогнозування погоди у житті людини

На протязі всієї історії людство потребує якісного прогнозу погоди, але з переходом економіки з аграрного до промислових способів виробництва, з ростом щільності населення у великих містах, питання прогнозу погоди стає ще більш важливим, ніж у минулі часи. Більшість сфер життєдіяльності людини залежить від погодних умов та згодом від прогнозу погоди. Крім того, що прогноз погоди допомагає великій кількості людей обрати одяг для прогулянок, є безліч інших питань, які допомагає вирішити прогноз погоди. Функціонування сільськогосподарського апарату безпосередньо залежить від погоди. Завдяки прогнозу погоди можливо врятувати життя великої кількості людей, попереджуючи громадян про небезпечно високий або низький рівні температури, високу швидкість поривів вітру, блискавки, пожежі та інше. Транспортна логістика також залежить від прогнозу погоди, на суші, воді та у повітрі. Прогнозування погоди здатне попередити про ожеледицю на дорогах, ураган на водному просторі, туман на небі та інші природні явища, які можуть перешкодити пересуванню людини. Енергетична промисловість

також потребує прогнозу погоди, наприклад вітроенергетика та сонячна електроенергетика безпосередньо залежать від погодних умов. При будівництві також бажано використовувати прогноз погоди, особливо при проектуванні великих споруд, це допомагає побудувати більш надійні до природних катаклізмів споруди.

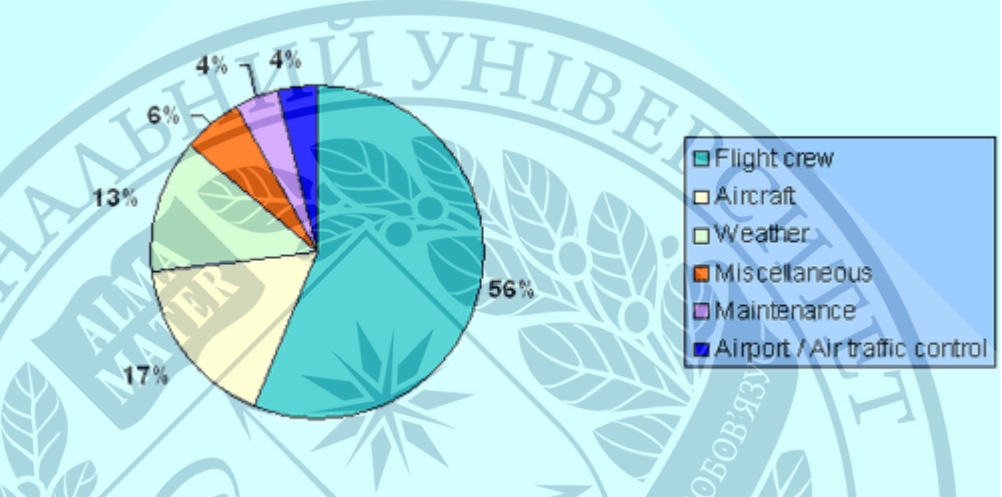

Рисунок 1.1. Основні причини авіакатастроф [7]

Відповідно до статистики основних причин авіакатастроф, яка наведена у вигляді діаграми на рис. 1.1, причини, пов'язані з небезпечними погодними умовами, знаходяться на третьому місці серед основних причин авіакатастроф та складають 13% від усіх причин. NBAA (національна асоціація бізнес авіації) у своїх матеріалах з безпеки та експлуатації заявляє, що більшість аварій, пов'язаних з погодними умовами, є фатальними та нездатність розпізнати погіршення погоди є частою причиною або значним фактором, який сприяє виникненню нещасних випадків [10]. Згідно з правилами льотної експлуатації цивільної авіації метеодані та/або прогноз погоди повинні відповідати вимогам, які призначені для певного льотного апарату. При різних умовах експлуатації для різних повітряних суден існують певні вимоги до погодних умов. Наприклад, для виконання польотів за

допомогою вертольоту над водою поза видимості землі, зведені погодні дані та/або прогнози повинні показувати, що висота нижньої межі хмар вище 180 метрів вдень або 360 метрів вночі [6].

Варто відзначити, що більшість авіакатастроф пов'язані з невеликими повітряними суднами. У 2018 році EASA (європейське агентство авіаційної безпеки) підготувало звіт, який зрівнює аварії, які сталися з легкими повітряними суднами з більш тяжкими [9].

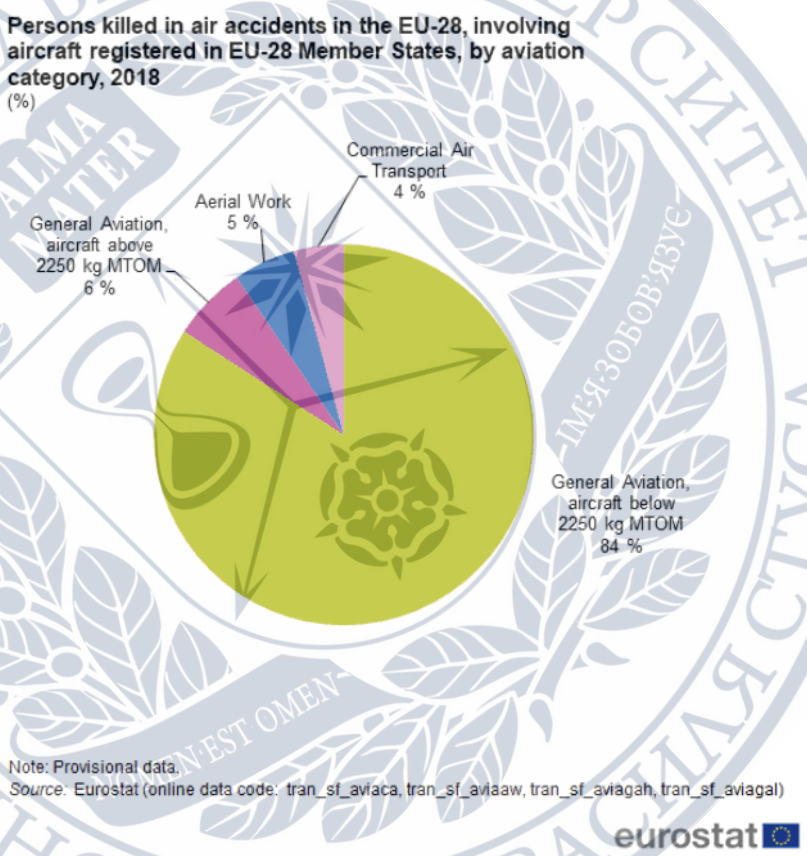

Рисунок 1.2. Статистика летальних випадків у ЄС за 2018, згрупована за категоріями повітряних суден [9]

На рис. 1.2 показано процентне співвідношення летальності між різними категоріями авіації. Легкі повітряні судна відносяться до категорії авіації з максимальною злітною масою до 2250 кг. Ця діаграма показує, що 84% всіх авіакатастроф з летальним випадком, зареєстрованих в ЄС у 2018

році, були аваріями легких повітряних суден. При порівнянні цих даних з аналогічними звітами за 2014-2018 рік, результати не мали великої різниці між собою. На території ЄС за 2018 рік у авіакатастрофах з участю легких повітряних суден загинуло 762 людини, тоді як з участю комерційних літаків — 188 людини. З огляду на те, що легкі повітряні судна перевозять в середньому менше 10 пасажирів, стає очевидним, що авіакатастроф з їх участю було набагато більше, ніж з участю комерційних літаків. Погані погодні умови відіграють значну роль у такій великій кількості летальних випадків серед участі легких повітряних суден. Погодні умови, безумовно, впливають й на тяжку авіацію, але для невеликих повітряних суден ризик польотів при поганих погодних умовах стає значно більшим. Оскільки у легкої авіації значно нижчі межі набору максимальної висоти, таким повітряним суднам дуже тяжко маневрувати в несприятливих погодних умовах. Досвід пілота може стати вирішальним фактором між життям і смертю в таких обставинах. Прогноз туману для висот на яких літають легкі повітряні судна та завчасне інформування пілоту про небезпечні погодні умови можуть зменшити кількість летальних випадків для легкої авіації [9].

Окрім туману, який перешкоджає повітряним суднам вдало приземлитися і злетіти, існує безліч інших погодних умов, які не сприяють успішному польоту. Обмерзання також є розповсюдженою причиною авіакатастроф, особливо у невеликих літаках. Згідно з даними AOPA (асоціація власників і пілотів літаків), за 1990-2000 роки серед всіх авіакатастроф, зареєстрованих у даній асоціації, 12% аварій пов’язані з обмерзанням. Лід на поверхні літака руйнує гладкий потік повітря, збільшуючи опір, зменшуючи при цьому здатність аеродинамічного профілю створювати підйом. Для компенсації створених льодом незручностей підіймається ніс літака, що сприяє подальшому обмерзанню нижній стороні крил та фюзеляжу літака. Лід може накопичуватись на кожній відкритій

лобовій поверхні літака, не тільки на крилах, пропелері та лобовому склі, а й на антенах, вентиляційному отворі, водозабірнику та капоту. Внаслідок чого може з'явитись вібрація, яка зламає антену літака. Також легкі літаки можуть стати настільки обмерзлими, що продовжити політ буде неможливо. Обмерзання може спричинити зупинку двигуна, якщо обмерзнуть важливі деталі двигуна або, у разі двигуна з впорскуванням палива, блокуючи джерело повітря двигуна [8].

Для сільськогосподарського виробництва погодні умови є найважливішим фактором, який впливає на: загальний урожай, кількість шкідників, потреба у воді та добривах, а також на всю сільськогосподарську діяльність, виконувану протягом вегетаційного періоду. Але варто зазначити, що це в першу чергу стосується землеробства під відкритим небом, землеробство у будівлях, теплицях не так сильно залежить від зовнішніх погодних умов. Єдиний спосіб контролювати процес вирощування рослин при випадкових погодних умовах є використовування точних прогнозів погоди. Грунтуючись на прогнозі погоди, фермери можуть планувати час для посіву, захисту, збору врожаю та інших польових робіт, щоб уникнути негативних погодних ефектів і втрат врожаю. Різним рослинам потрібна різна температура, вологість, щоб почати проростання та продовжити рост. У той же час, показники температури та вологості часто використовуються фермерами для прогнозування появи комах-врідників та хвороб у рослин. Для отримання корисної інформації про стан ґрунту та сільськогосподарських культур фермери користуються такими показниками, як: температура повітря і ґрунту, відносна вологість, вологість ґрунту, опади, швидкість/напрямок вітру, евапотранспірація [11]. Для отримання більш точного прогнозу погоди для своєї ділянки фермери можуть використовувати локальні невеликі метеостанції, такі як Arduino. Подібні прилади за допомогою різних датчиків дозволяють виміряти всі, потрібні для фермерів, метеорологічні показники

локально для окремих ділянок, що надає можливість фермерам отримати надійні дані саме для території, яка потрібна фермеру. Зпрогнозувавши кількість опадів для окремої ділянки фермер має змогу уточнити графік зрошення. Фермер може зпланувати оптимальні рівні зберігання кормів і зібраного врожаю, використовуючи інформацію, що стосується вологості. Також прогноз погоди допомагає фермерам підготуватися до екстремальних погодних умов, таких як мороз, град і посуха [12].

Існує багато природних явищ, які можуть завдати шкоду сільськогосподарським культурам. У холодну пору року несприятливі погодні умови можуть завдати шкоду плодовим та лісовим насадженням, озимим культурам, багаторічним травам тощо. Потужний сніговий покрив, сильні морози без снігового покрову, часті й тривалі відлиги завдають найбільшої шкоди, особливо для слаборозвинутих рослин. Але зимуючі культури можуть гинути не тільки з-за низьких температур та обмерзання, а й від вимокання, випрівання, льодової кірки та часткового зрідження посівів. Загибель посівів взимку потребує пересадку озимих культур весною, що потребує додаткового насіння, а також збільшує обсяг польових робіт. Крім того, заморозки, як і суховії, погіршують якість зерна. У той же час протягом вегетаційного періоду також існує безліч природних явищ, які здатні значно зменшити обсяг врожаю. Посухи призводять до висухання ґрунту, суховії призводять до порушення водного балансу рослин, пилові бурі здатні значно пошкодити ґрунтовий покрив, град знищує врожай, сильні вітри і зливи призводять до вилягання посівів, що затрудняє збирання врожаю [13].

На рис. 1.3 простежується вплив погодних умов на приріст урожаю пшениці та кукурудзи в Україні. З графіку видно, що внесок погодних умов досіть відлічається для приросту врожаю пшениці від приросту врожая кукурудзи, що пояснюється декількома факторами. Різні сільськогосподарські культури потребують різних погодних умов. Наприклад, у жовтні та першій

декаді листопада 2018 року був дефіцит опадів, що в результаті стало несприятливими погодними умовами для початку росту та розвитку озимих культур. Але у квітні-травні 2019 року була достатня кількість опадів для початку росту ранніх зернових культур. Крім погодних умов на врожай сільськогосподарських культур впливає багато інших факторів, таких як: родючість ґрунту, збільшення посівних площ під зерновими та зернобобовими, застосування відбірних сортів зернових, використання пестицидів та добрив [14].

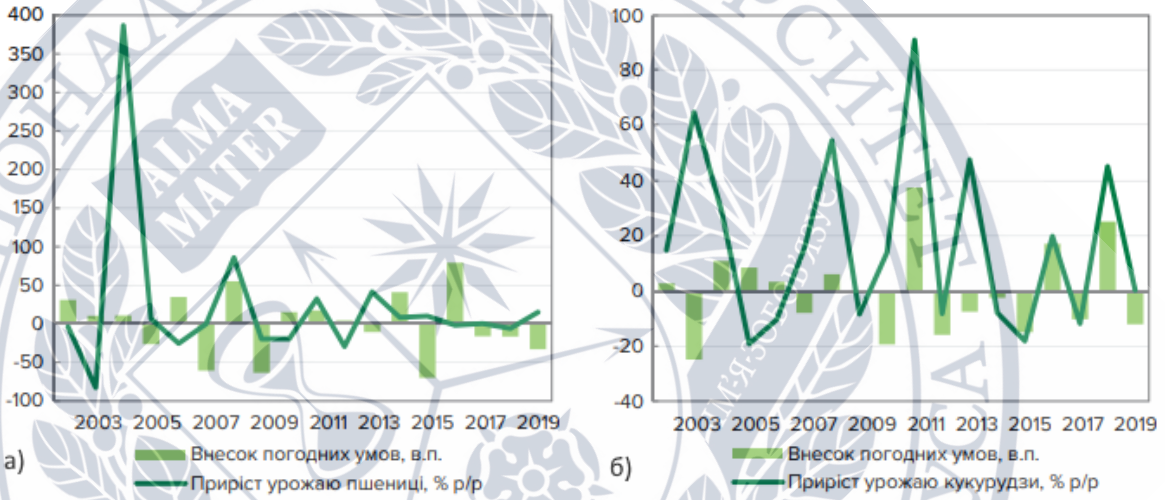

Рисунок 1.3. Внесок погодних умов у приріст обсягів виробництва пшениці (а), кукурудзи (б), % [14]

Для того, щоб запобігти утворення більшості стихійних лих, варто інвестувати у збір метеоданих і їх аналіз для отримання науково обґрунтованих висновків щодо заходів, які потрібно зробити, щоб уникнути жертв серед людей, врожаю, тощо. Наразі саме збір і аналіз приймає першостепове значення для уникнення природних катастроф, йдеться у доповіді ФАО [15]. Для збору метеоданих та їх оцінки з'являється все більше нових методів, таких як збір геопросторової інформації, дистанційне зондування, застосування безпілотників і робототехніки для роботи в надзвичайних ситуаціях, машинне навчання, що дозволяє знизити ризики та

вплив стихійних лих, але це потребує інвестицій, що неможливо для малих підприємств без розвитку державно-приватного партнерства, також доповідає ФАО. Варто зазначити, що крім збору та аналізу метеоданих важливим фактором у спасінні людей при стихійних лихах є своєчасне оголошення та інформування людей про небезпеку.

1.2 Сучасні тенденції та стан розвитку

На сьогоднішній день основною технологією, яку використовує метеорологія, є фізико-математичне моделювання руху атмосферних частин, атмосферних вихрів. Ці дані моделюються за допомогою суперкомп'ютеру [16]. Згідно опублікованому у квітні 2017 року звіту Мінприроди Росії, у 2015 році завдяки роботі Гідрометцентру вдалось зберегти витрат для ліквідації наслідків можливих збитків на суму 35 млрд рублів, а 4,3 млрд з цієї суми для сільського господарства. У тому ж звіті дається оцінка точності прогнозів на одну добу: з 2006 по 2016 рік вона зросла з 94% до 96,5%, а період, прогноз на який збувається у 70% випадків, з 4 до 6 діб [17].

Ізобарична поверхня показує висоту в атмосфері від поверхні Землі, де тиск приймає одне і те ж значення.

Наприклад, карта ізобаричної поверхні 700 гПа (AT700) буде показувати висоту, де тиск повітря досягає даного значення, тобто 700 гПа. Ця висота може десь знижуватися майже до 2.5 км, а десь досягати 3-3.2 км і навіть вище.

Для вільної атмосфери використовують карти висот стандартних изобаричних поверхонь 1000 гПа, 850гПа, 700 гПа, 500 гПа, 300 гПа і т.д .

На рис.1.4 (ковзне середнє за 12 місяців) різних прогностичних центрів у нетропічній частині Північної півкулі по відношенню до об'єктивного аналізу відповідного центру для (а) тиску на рівні моря, (б) висоти

ізобаричної поверхні 500 гПа і (в) модуля вітру на ізобаричної поверхні 250 гПа за період 2008-2014 рр. за даними ведучого центру ВМО з верифікації детерминистичних прогнозів. Помилки прогнозів Гідрометцентру Росії - жовто-коричневі лінії.

З рис. 1.4 видно, що прогнози різних метеоданих з 2008 року замітно поліпшились, особливо для Гідрометцентру Росії, що безумовно свідчить про появу можливості використовувати більш потужні обчислювальні засоби.

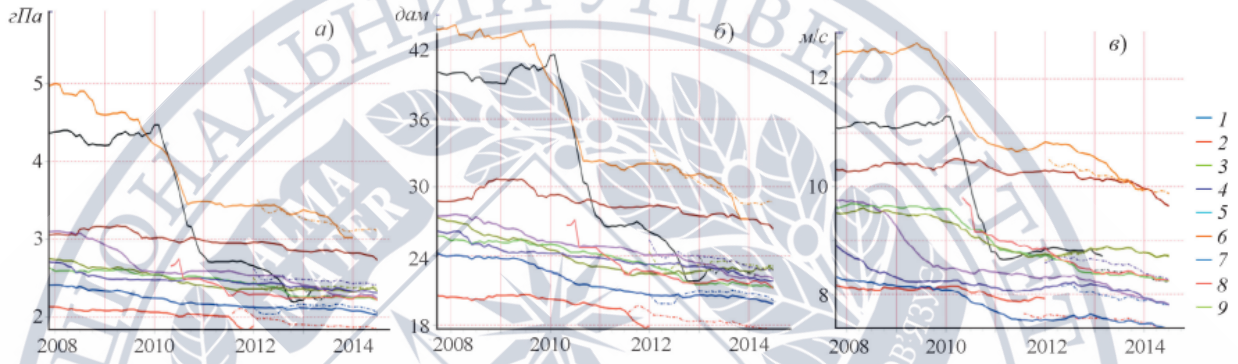

Рисунок 1.4. Середньомісячні середньоквадратичні помилки прогнозів [18].

Існує велика кількість фізико-математичних моделей, які використовують метеорологи, основні з яких: T169L31 - спектральна модель атмосфери ГУ «Гідрометцентр Росії», REGION (ДУ «Гідрометцентр Росії», автор В. М. Лосєв), UKMO - Метеорологічний центр Великобританії, ECMWF - Європейський центр Середньострокових прогнозів погоди, NCEP - Метеорологічний центр США , PLAV - полулагранжева з постійним дозволом (ДУ «Гідрометцентр Росії», автор М.А. Толстих), DWD - метеорологічний центр НГМС Німеччини [20]. Наведемо декілька прикладів динаміки зміни точності прогнозів з різними моделями.

На рис. 1.5 зображена діаграма зміни середньоквадратичної помилки прогнозу атмосферного тиску на рівні моря в залежності від часу прогнозу.

На рис. 1.6 зображена діаграма зміни середньоквадратичної помилки прогнозу температури повітря висоти ізобаричної поверхні 850 гПа в залежності від часу прогнозу.

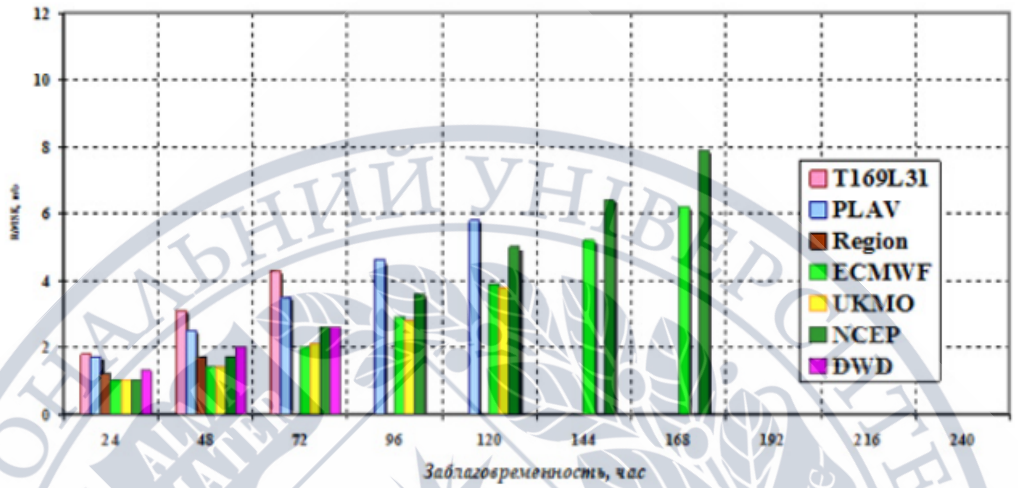

Рисунок 1.5. Діаграма зміни середньоквадратичної помилки прогнозу атмосферного тиску на рівні моря в залежності від часу прогнозу [19]

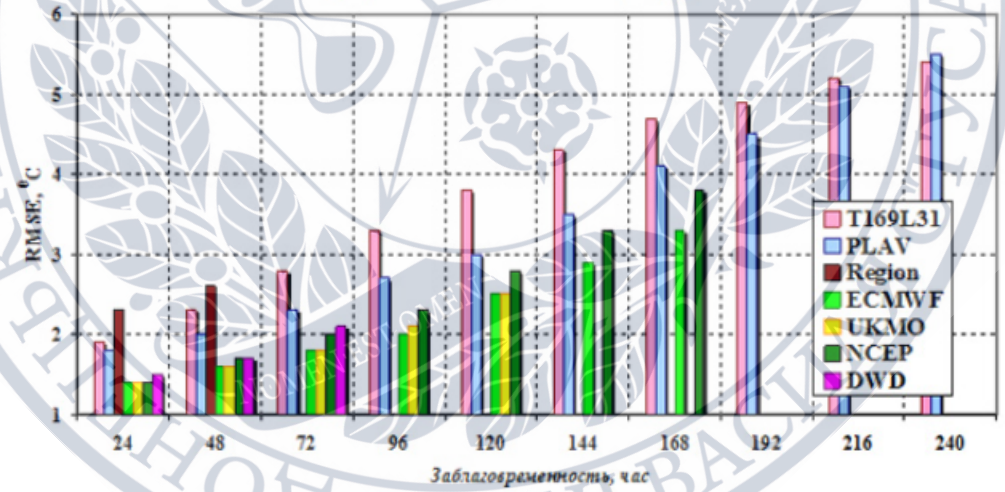

Рисунок 1.6. Діаграма зміни середньоквадратичної помилки прогнозу температури повітря в залежності від часу прогнозу [19]

З діаграм видно, що зі збільшенням терміну прогнозу підвищується середньоквадратична помилка. Найкращими для короткострокових прогнозів з діаграм виявились моделі ECMWF, UKMO. Для більш довгих прогнозів, від

144 до 192 годин, також підходить модель NCEP, для прогнозів від 192 до 264 підходять тільки прогнози від Гідрометцентру Росії, T169L31 та PLAV.

Висновок до розділу 1

Прогноз погоди є актуальним для великої кількості різних сфер життедіяльності людства, тому при прогнозі погоди варто концентруватись на прогнозі специфічних метеопоказників, таких як вологість ґрунту для сільского господарства або вірогідність туману у верхніх слоях атмосфери для авіації, або прогнозувати загальні метеорологічні показники, як температура повітря, тиск, вологість, тощо. Невизначеність та похибки у результатах роботи статистичних методів неминучі. Вони виникають внаслідок мінливості самого клімату, недосконалості моделі клімату внаслідок обмежень у нашому розумінні та в обчислювальної потужності (тим не менш кліматичні моделі належать до найскладніших обчислювальних моделей), відсутності даних вимірювань (необхідних для калібрування клімату моделі) в будь-якому просторі-часі.

Отже, статистичний висновок повинен не лише звітувати про найкращі результати прогнозу (оцінки), а також про похибки та невизначеності результатів.

РОЗДІЛ 2.
ФОРМАЛІЗАЦІЯ ЗАДАЧ ТА ПРОЕКТУВАННЯ МЕТОДІВ ПРОГНОЗУ ПОГОДИ

Метою дипломної роботи є реалізація існуючих методів прогнозу погоди, їх тестування на реальних даних та створення нового методу або нового ансамблю методів.

2.1 Постановка задачі та побудова математичної моделі

Постановка задачі

За допомогою статистичних методів визначити показники, між якими існує лінійна або нелінійна залежність. Зробити прогноз для деяких показників. Підрахувати точність прогнозу. За допомогою діаграм візуалізувати отримані результати та підвести висновки щодо ефективності використання прогнозу. Визначити різницю між роботою різних методів.

Математична модель

Визначимо математичну модель, яку будемо використовувати для прогнозу погоди.

Параметризація моделі

Зробимо параметризацію моделі для більшої зручності використання тих чи інших параметрів.

Визначимо мінімальну температуру, як t(min), максимальну температуру - t(max), температуру у 9 годин ранку - t(9am), температуру у 3 години дня - t(3pm), кількість опадів - r(n), кількість випаровування - e(n), кількість сонячного випромінювання - s(n), напрямок поривів вітру - w_g(dir), швидкість поривів вітру - w_g(sp), напрямок вітру у 9 годин ранку - w(dir9am), напрямок вітру у 3 години дня - w(dir3pm), швидкість вітру у 9

годин ранку - w(sp9am), швидкість вітру у 3 години дня - w(sp3pm), вологість у 9 годин ранку - h(9am), вологість у 3 години дня - h(3pm), тиск у 9 годин ранку - p(9am), тиск у 3 години дня - p(3pm), хмарність у 9 годин ранку - c(9am), хмарність у 3 години дня - c(3pm), випадання дощу у сьогоднішній день - r(tod), випадання дощу у завтрашній день - r(tom), вологість у 9 годин ранку - h(9am), вологість у 3 години дня - h(3pm).

Математична модель

Зробимо формалізації моделі для прогнозу показника температури у 9 годин ранку на основі показника мінімальної температури (2.1). Назви параметрів наведені вище у параметризації моделі.

$$t(9am)_{i+1} = \frac{\sum_{i=1}^{n} (t(9am)_i - t(min)_i)}{n} \quad (2.1)$$

де $t(9am)_{i+1}$ — прогнозуєме значення температури повітря у 9 годин ранку;

$t(9am)_i$ — значення температури повітря у 9 годин ранку за минулий період;

$t(min)_i$ — значення мінімальної температури повітря за минулий період;

n — розмір вибірки.

За формулою 2.1 можливо зробити прогноз не тільки для визначених вище показників, але й для інших показників, між якими існує сильна залежність.

Нижче наведено формулу для прогнозу хмарності у 9 годин ранку на основі показнику температури у 9 годин ранку за допомогою частотного розподілу. Назви параметрів наведені вище у параметризації моделі.

$$c(9am)_{i+1} = \frac{\sum class(t(9am))_i}{n} * 100\% \rightarrow c(9am)_i \quad (2.2)$$

де $c(9am)_{i+1}$ — прогнозуєме значення хмарності у 9 годин ранку;

$t(9am)_i$ — значення температури повітря у 9 годин ранку за минулий період;

$c(9am)_i$ — значення хмарності у 9 годин ранку за минулий період;

class(t(9am))$_i$ — клас хмарності у 9 годин ранку;

n — розмір вибірки.

Формулу 2.2 також можливо використовувати для інших показників між якими є сильна залежність.

Нижче наведено формулу для розрахунку параметрів лінійної регресії за допомогою методу найменших квадратів, формула (2.17), для показника температури у 9 годин ранку. Назви параметрів наведені вище у параметризації моделі.

$$\sum_i e_i^2 = \sum_i (y_i - f_i(t(9am)_i))^2 \to min(t(9am)_i) \quad (2.3)$$

де $t(9am)_i$ – значення температури повітря у 9 годин ранку за минулий період;

$f_i(t(9am)_i)$ – це сукупність змінних набіра показника температури повітря у 9 годин ранку за минулий період;

Нижче наведено формулу для розрахунку параметрів поліноміальної регресії за допомогою методу найменших квадратів, формула (2.17), для показника швидкості вітру у 9 годин ранку. Назви параметрів наведені вище у параметризації моделі.

$$\sum_i e_i^2 = \sum_i (y_i - f_i(t(9am)_i))^2 \to min(t(9am)_i) \quad (2.4)$$

де $t(9am)_i$ – значення температури у 9 годин ранку за минулий період;

$f_i(t(9am)_i)$ – це сукупність змінних набіра показника температури у 9 годин ранку за минулий період.

Нижче наведено формулу для розрахунку параметрів гіперболічної регресії за допомогою методу найменших квадратів, формула (2.17), для показника кількості сонячного випромінювання. Назви параметрів наведені вище у параметризації моделі.

$$\sum_i e_i^2 = \sum_i (y_i - f_i(s(n)_i))^2 \to min(s(n)_i) \quad (2.5)$$

де $s(n)_i$ – значення кількості сонячного випромінювання за минулий період;

$f_i(s(n)_i)$ – це сукупність змінних набіра показника кількості сонячного випромінювання за минулий період.

Нижче наведено формулу для розрахунку параметрів квадратичної регресії за допомогою методу найменших квадратів, формула (2.17), для показника тиску у 9 годин ранку. Назви параметрів наведені вище у параметризації моделі.

$$\sum_i e_i^2 = \sum_i (y_i - f_i(s(n)_i))^2 \to min(s(n)_i) \quad (2.6)$$

де $p(9am)_i$ – значення тиску у 9 годин ранку за минулий період;

$f_i(p(9am)_i)$ – це сукупність змінних набіра показника тиску у 9 годин ранку за минулий період.

Нижче наведено формулу для розрахунку параметрів показникової регресії за допомогою методу найменших квадратів, формула (2.17), для показника хмарності у 3 години дня. Назви параметрів наведені вище у параметризації моделі.

$$\sum_i e_i^2 = \sum_i (y_i - f_i(c(9am)_i))^2 \to min(c(9am)_i) \quad (2.7)$$

де $c(9am)_i$ – значення хмарності у 3 години дня за минулий період;

$f_i(c(9am)_i)$ – це сукупність змінних набіра показника хмарності у 3 години дня за минулий період.

2.2 Математичний апарат

2.2.1 Кореляційний аналіз

Фундаментальною концепцією для аналізу двовимірних наборів даних (двох змінних) є кореляційний аналіз, який є кількісним показником того, наскільки сильно змінюються обидві змінні [1].

Рисунок 2.1. Види кореляцій [1]

Коефіцієнт кореляції Пірсона

Міра для ступеня лінійного відношення між двома змінними. Використовується для підрахунку кореляції між парою змінних, які знаходяться у відношенні лінійного збільшення/спаді.

Коефіцієнт кореляції Пірсона є стандартною мірою кореляції, сприйнятливий до відхилень.

Результатом коефіцієнту Пірсона є число від –1 до +1, де

+1 - повне позитивне лінійне відношення;

0 - відсутність відношення;

–1 - повне від’ємне лінійне відношення.

Коефіцієнт кореляції Пірсона визначається формулою 2.8 [1]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \underline{x})^2} \bullet \sqrt{\sum_{i=1}^{n}(y_i - \underline{y})^2}}, \quad (2.8)$$

де $\underline{x}$ — середнє арифметичне $x_{(i)}$;

$\underline{y}$ — середнє арифметичне $y_{(i)}$;

n — обсяг вибірки.

Коефіцієнт кореляції рангу Спірмена

Міра статистичної залежності між двома змінними. Використовується для підрахунку кореляції між парою змінних, які знаходяться в монотонно зростаючому/спадаючому відношенні. Коефіцієнт кореляції рангу Спірмена надійний проти присутності викидів; не вимагає нормальної форми розподілу; не вимагає лінійного відношення (тобто підходить також для логарифмічного чи експоненційного відношення).

Розрахунок коефіцієнта кореляції Спірмена рангу (ρ) робиться наступним чином [22]:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n}(R_i - S_i)^2}{n^3 - n}, \quad (2.9)$$

де R_i — ранг спостереження x_i в ряду x;

S_i — ранг спостереження y_i в ряду y;

n — обсяг вибірки.

Дані приводяться у порядку зростання, вихідні значення даних замінюються їх рангами і ранги рядів співвідносяться між собою. Дані повинні знаходитися в порядковій шкалі та бути незалежними (без

автокореляції); повинні бути відсутністніми треті змінні, що впливають на відношення, обсяг вибірки ≥ 30.

Результатом коефіцієнту Спірмена є число від –1 до +1, де:

+1 - повне позитивне монотонне відношення;

0 - відсутність відношення;

–1 - повне від'ємне монотонне відношення.

Коефіцієнт кореляції рангу Кендала

Міра статистичної залежності між двома змінними. Використовується для підрахунку кореляції між парою змінних, які знаходяться в монотонно зростаючому/спадаючому відношенні. Коефіцієнт Кендалла зазвичай приймає дещо менші абсолютні значення ніж міра кореляції Спірмена.

Коефіцієнт кореляції рангу Кендала корисний для малих розмірів вибірки, надійний проти присутності викидів; не вимагає нормальної форми розподілу; не вимагає лінійного відношення (тобто підходить також для логарифмічного чи експоненційного відношення).

Розрахунок коефіцієнта кореляції рангу Кендала робиться наступним чином [23]:

$$\tau = \frac{C-D}{n(n-1)^2}, \quad (2.10)$$

де D — кількість суперечливих пар даних;

C — кількість узгоджених пар даних;

n — обсяг вибірки [23].

Пару даних (x, y) називають узгодженою, якщо ранг зменшується (збільшується) як в x, так і в y (наприклад, ранг x1 > ранг x4 і ранг y1 > ранг y4). Якщо це не так, тоді пара даних називається суперечливою (наприклад, ранг x1 > ранг x4 та ранг y1 < ранг y4). Дані приводяться у порядку

зростання, вихідні значення даних замінюються їх рангами і ранги рядів співвідносяться між собою. Дані повинні знаходитися в порядковій шкалі та бути незалежними (без автокореляції); повинні бути відсутністніми треті змінні, що впливають на відношення, невеликий обсяг вибірки.

Результатом коефіцієнту Кендала є число від –1 до +1, де:

+1 - повне позитивне відношення;

0 - відсутність відношення;

–1 - повне від'ємне відношення.

Статистична кореляція не обов'язково передбачає причинно-наслідковий зв'язок відношення, прикладом такого фальшивого відношення є кореляція між народжуваністю та популяцією лелеки. Чим більший обсяг вибірки, тим більш вірний результат.

2.2.2 Частотний розподіл

Фундаментальна концептуалізація полягає в тому, що невизначені або випадкові компоненти клімату описуються за допомогою розподілу значень, які може приймати змінна клімату. Функція щільності неперервної випадкової величини (густина ймовірності) від змінної X визначає ймовірність знаходження X між деяким значенням x та іншим x + dx, де dx не є нулем; ймовірність задається інтегралом густини ймовірності за інтервал [x; x + dx]. Представлені статистичні методи у розділі 1.3 роблять висновок про густину ймовірності за допомогою гістограм.

Багато статистичних методів роблять припущення щодо густини ймовірності випадкової величини такі як нормальне або припущення Гаусса (дзвоноподібна крива). На практиці припущення можуть бути порушені, а статистичні методи називаються надійними, якщо у випадку порушення вони все одно приносять результати прийнятної точності.

Стосовно моделей клімату точність називається прийнятною, якщо, наприклад, 95% довірчий інтервал має справжнє охоплення лише у 91% через порушення припущення про розподіл, але справжнє охоплення не повинно бути меншим за 78%).

Одновимірний частотний розподіл

За допомогою цього метода можливо зробити:

1. сортування даних вибірки за розміром;
2. поділ на класи з постійною або різною шириною класу (інтервал) та верхньою (правою) або нижньою (лівою) межою класу;
3. підрахунок кількості точок даних, які потрапляють до певного класу;
4. побудова графіків частоти появи за допомогою таблиці частот або гістограм.

Крок (2) потрібно обробляти обережно, оскільки він змінює частотний графік і може призвести до неправильної інтерпретації основної емпіричної функції розподілу. Тому слід звернути увагу на наступні підкроки:

(a) Вибір ширини класу (інтервалу);

(b) Вибір нижньої (лівої) межі класу.

Для розрахунку приблизної оцінки ширини класу (інтервалу) нормально розподілених даних може бути застосована формула SCOTT [42]:

$$h = 3,49\,\sigma/\sqrt[3]{n}\,, \tag{2.11}$$

де σ — стандартне відхилення;

n — кількість елементів.

Цим методом можуть бути оброблені будь-які метричні дані, які можна сортувати за їх значеннями. Гістограма густини ймовірності показує, скільки елементів даних, в абсолютних чи відносних числах, потрапляє до певного класу. Одновимірний частотний розподіл це простий та інтуїтивно зрозумілий метод статистичного аналізу часових рядів. Якщо аналізуються спостережувані часові ряди, потрібно включити певний клас похибок вимірювань, щоб сума частоти зустрічальності дорівнювала одній відповідно до ста відсотків. За допомогою цього методу густину ймовірності можна порівняти між різними джерелами даних (наприклад, спостережувані та змодельовані дані) та обчислити зміну густину ймовірності.

Приклад:

Порівняння частоти появи p[%] середньої швидкості вітру з шириною класу 2м/с (рис. 2.2) та середнім напрямком вітру з шириною класу 30° (рис. 2.3) за результатами моделі (Cosmo-CLM) та спостереження (Німецька служба погоди), поблизу Варнемюнде, період відліку 1971–2000 [3].

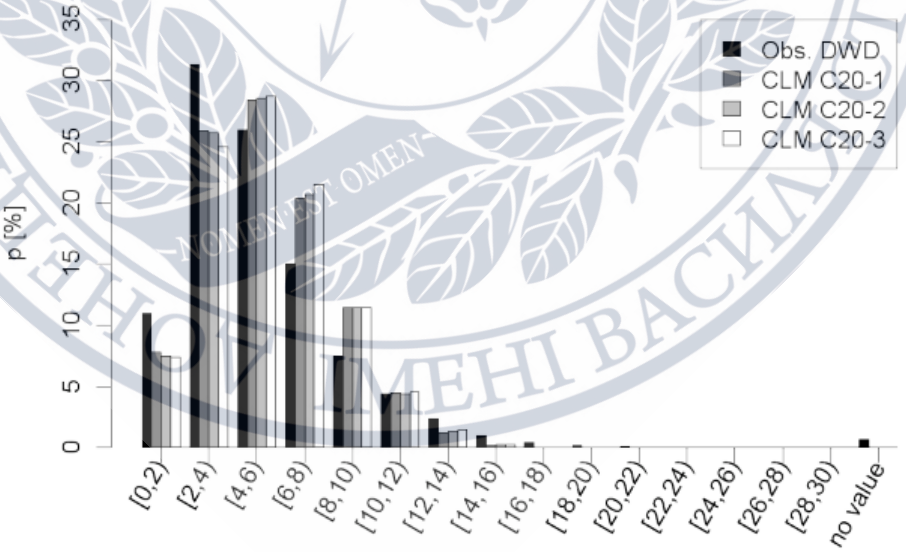

Рисунок 2.2. Модель Cosmo-CLM, середня швидкість вітру [3]

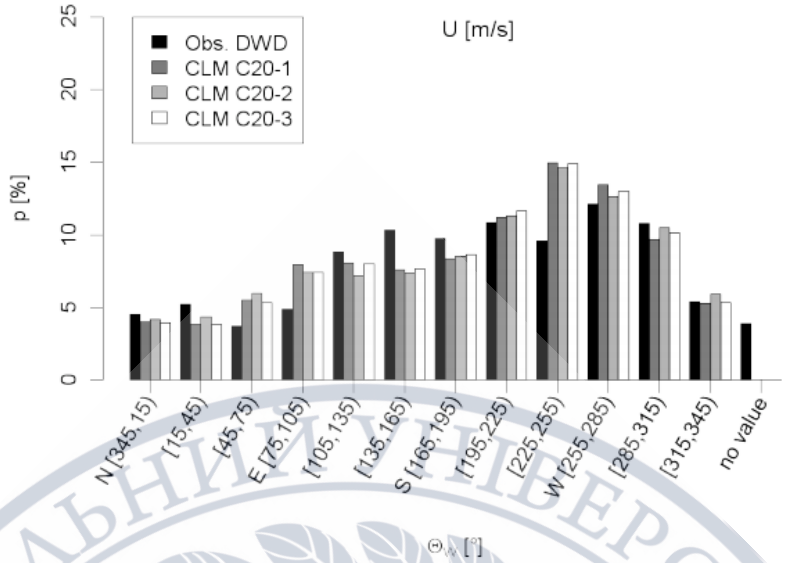

Рисунок 2.3. Модель Cosmo-CLM, середній напрямок вітру [3]

Порівняно з спостереженнями, результати моделі Cosmo-CLM показують менше подій при низькій та високій швидкості вітру, але більше подій для середніх швидкостей (рис. 2.2). Щодо напрямку вітру, є більше подій зі Сходу, Заходу і кілька подій з Півдня, ніж спостерігалося (рис. 2.3).

Відносний частотний розподіл

За допомогою цього метода можливо зробити сортування даних вибірки за розміром, поділ на класи: підрахунок, скільки точок даних потрапляє в клас і нормування (тобто ділення на загальний обсяг вибірки). Застосовується до температури та опадів. Дані повинні бути динамічним рядом з рівними інтервалами. Відносний частотний розподіл це простий та інтуїтивно зрозумілий метод статистичного аналізу часових рядів. Вибір ширини класу впливає на форму частотних графіків. Якщо дані спостереження аналізуються, потрібно подбати про похибку вимірювання (додатковий клас), щоб переконатися, що відносні частоти складають 1 або 100% [24].

Приклади:

1) Порівняння спостережуваних та змодельованих за допомогою CLM частот середньодобової температури станції Dresden-Klotzsche [4].

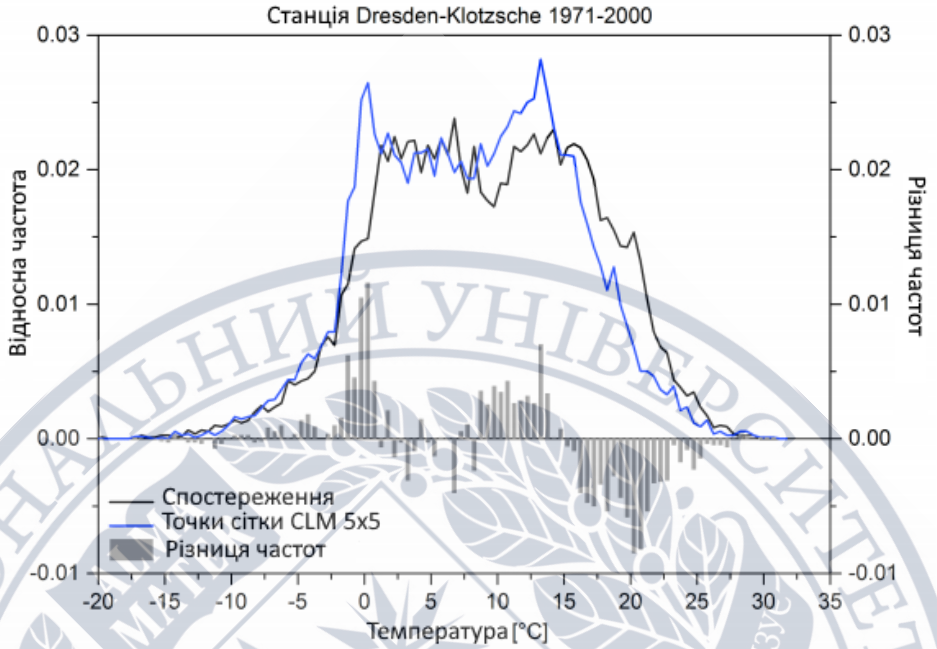

Рисунок 2.4. Приклад роботи відносного частотного розподілу (1) [4]

2) Порівняння спостережуваних та змодельованих за допомогою CLM частот середньодобової температури - з / без просторового усереднення розподілу частоти для станції Dresden-Klotzsche, незалежно від кількості аналізованих сіткових коробок.

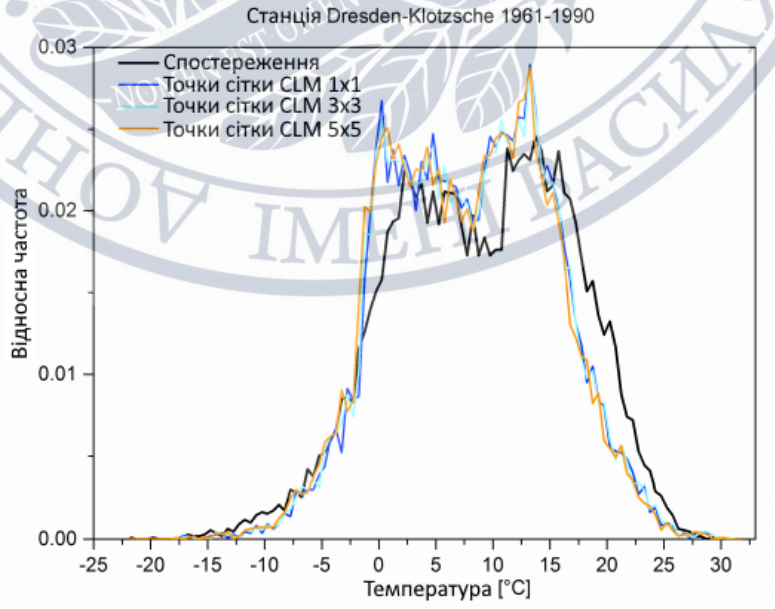

Рисунок 2.5. Приклад роботи відносного частотного розподілу (2) [4]

Двовимірний частотний розподіл

Двовимірний частотний розподіл визначає сутність усіх комбінацій двох змінних X (зі значеннями x_i, $i = 1,\ldots, l$) та Y зі значеннями $y_j, j = 1,\ldots, m$). У збігу з одновимірною частотою, появу комбінації значень (x_i, y_j) може бути подано або в абсолютних числах n (x_i, y_j) та / або відносно обсягу вибірки h (x_i, y_j) = n (x_i, y_j) / n за умов [43]:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} n(x_i, y_i) = n_{ges} \quad (2.12)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} h(x_i, y_i) = 1, \quad (2.13)$$

Усі комбінації значень (x_i, y_j) можна описати за допомогою таблиці 2.1 перехресних класифікацій або у формі гістограми.

Таблиця 2.1 — Перехресна класифікація відносної частоти виникнення [43]

Змінна X	Змінна Y					
	y_1	…	y_j	…	y_m	Одновимірна частота появи змінної X
x_1	h_{11}	…	h_{1j}	…	h_{1m}	$h_{1.}$
…	..	…	…	…	…	…
x_i	h_{i1}	…	h_{ij}	…	h_{im}	$h_{j.}$
…	…	…	…	…	…	…
x_l	h_{l1}	…	h_{lj}	…	h_{lm}	$h_{l.}$
Одновимірна частота появи змінної Y	$h_{.1}$	…	$h_{.j}$	…	$h_{.m}$	h..=1

Цей метод ефективно використовується для аналізу декількох параметрів, наприклад показники морського стану (висота та напрямок хвилі) або метеорологічні показники (швидкість та напрямок вітру). Цим методом можуть бути оброблені метричні дані або дані з номінальною порядковою шкалою. Таблиця перехресних класифікацій або гістограма інформує про абсолютну або відносну кількість комбінацій змінних X та Y (наприклад, висота та напрямок хвилі). Більш того, одновимірна частота появи змінних X та Y може бути описана в останньому рядку та стовпці таблиці. Це більш складний метод статистичного аналізу часових рядів. Якщо аналізуються спостережувані часові ряди, потрібно включити певний клас похибок вимірювань, щоб сума частоти появи дорівнювала одній на сто відсотків.

За допомогою цього методу частоту зустрічальності можна порівняти між різними джерелами даних (наприклад, спостережувані та змодельовані дані) та обчислити зміну частоти появи.

Приклад:

Гістограма частоти зустрічальності спостережуваної швидкості та напрямку вітру (Німецька метеорологічна служба) поблизу Warnemünde та за контрольний період 1971-2000. Ширина класу швидкості вітру становить 1 м/с та напрямку 10°. Зазвичай ширина класу складає 22.5°, що ділить всі напрямки вітру на 16 класі, але у цьому випадку усього 36 класів, що може пояснюватись більш точними засобами збору метеоданих для визначення напрямку вітру. Зазвичай, приладом для визначення напрямку вітру слугує флюгер. Кількість подій класифікується кольорами від кольорового поля праворуч [3].

З рис. 2.6 видно, що сильних вітрових подій зі швидкістю вітру більше 20 м/с (що відповідає 9 класу Бофорта) мало. Для цих подій основний напрямок вітру коливається від південного заходу до північного заходу.

Більшість подій із середньою швидкістю вітру (наприклад, 5-20 м/с) також надходять із західних напрямків. Більше того, з рисунка видно, що чим вище швидкість вітру, тим менша кількість подій і напрямок вітру повертається у напрямі північного заходу. На відміну від подій з низькою швидкістю вітру (наприклад, 0-5 м/с), які, очевидно, спостерігались у більшості випадків, переважно відбуваються з південно-східних до південно-західних напрямків.

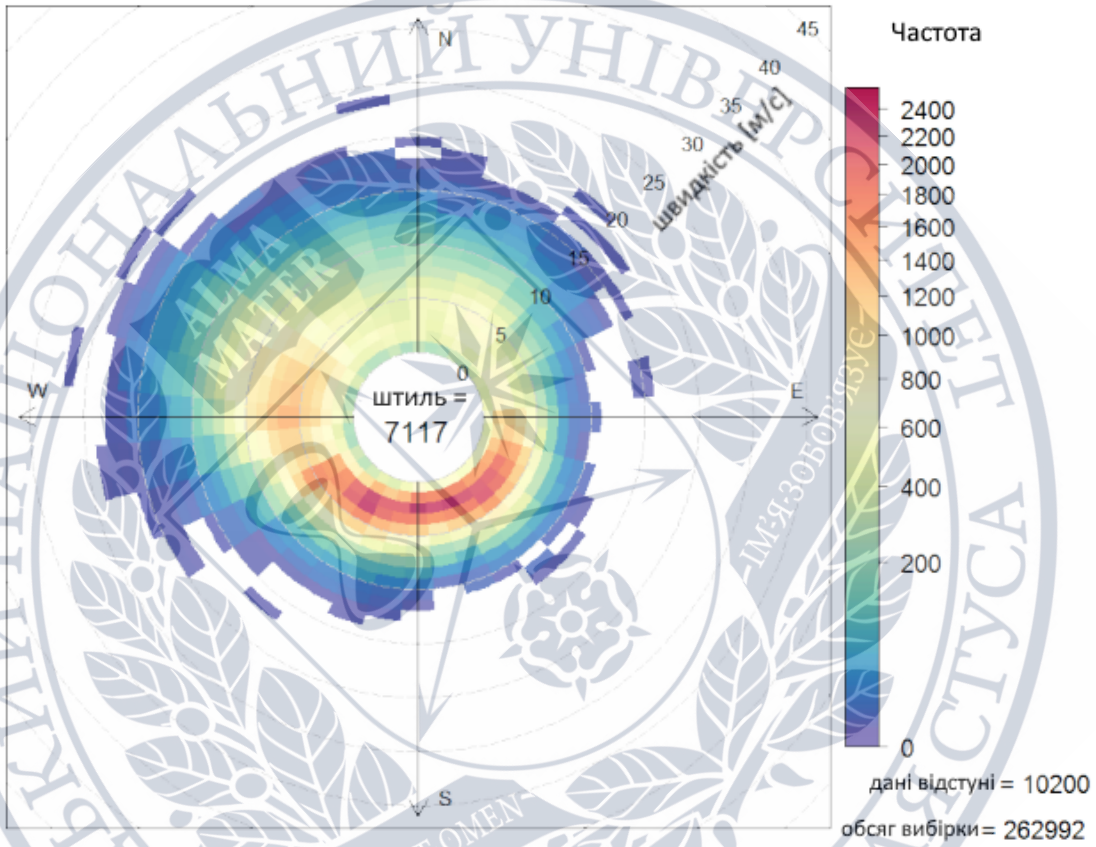

Рисунок 2.6. Приклад роботи двовимірного частотного розподілу [3]

2.2.3 Аналіз часових рядів

«Зміна клімату» відноситься до часу, і аналіз змодельованих або спостережуваних часових рядів, таких як глобальна температура поверхні повітря протягом останнього тисячоліття, є важливим полем для аналізу клімату.

Одна з найперших робіт статистичного аналізу часових рядів досліджувала “передбачуваний 26-денний період метеорологічних явищ” [26].

Можливий погляд на кліматичні зміни - це залежна від часу випадкова величина, яка складається з тенденції, викидів/екстремумів та мінливості/шуму [25]; цей структурний підхід також є основою представленого методу середнього значення.

Завдання аналізу полягає у використанні даних для оцінки параметрів, що описують тенденцію, мінливість та інші компоненти. Оцінка тенденцій, тобто кількісна зміна кліматичних змін, є пріоритетним. Це відображається різноманітністю представлених методів: лінійна регресія, де застосовується проста параметрична модель тенденції; додаток поточного середнього значення, де не слід приймати параметричну форму тренду; та ковзна середня, яка є надійним аналогом поточного середнього значення; також гнучкий аналіз тенденцій, який є непараметричним методом. Непараметричну регресію також називають згладжуванням. Іншими інструментами вилучення низькочастотних тенденцій є числова фільтрація частот або порівняння різних часових зрізів. Окрім тенденції, мінливість кліматичної змінної може бути розкладена на різні класи.

Поточне середнє значення

Метод поточного середнього значення визначає оцінку тенденцій за допомогою розрахунку арифметичних середніх часово послідовних точок даних. Використовується для, наприклад, визначення частоти переливу, розміру переливу; тимчасової роздільної здатності. Дані повинні бути динамічним рядом без розривів. Метод зменшує мінливість і дозволяє аналізувати тенденції. Метод поточного середнього значення дозволяє зробити простий і швидкий аналіз часових рядів [37].

Ковзна середня

Метод ковзної середньої визначає надійну непараметричну оцінку тренду [38]. Надійна означає, що на метод не впливає наявність крайнощів. Використовується для не параметричної оцінки тренду. Дані повинні бути однорідними та репрезентативними. Ковзна середня є одним з статистичних методів, який згладжує коливання, які спостерігаються в даних, щоб отримати уявлення про наявність тренду.

Ковзна середня може бути в основному трьох типів:

- Проста ковзна середня [39];
- Зважена ковзна середня [40];
- Експоненціальна змінна середня [41].

Проста ковзна середня (англ. Simple Moving Average – SMA) – є одними з найбільш простих і популярних індикаторів в статистичному аналізі. SMA є звичайним середнім арифметичним від значень за певний період. SMA відноситься до класу індикаторів, які слідують за трендом, воно допомагає визначити початок нової тенденції і її завершення, за його кутом нахилу можна визначити силу (швидкість руху). Іноді ковзну середню називають лінією тренда.

Формула простої ковзної середньої [39]:

$$SMA = \frac{\sum_{i=1}^{n} P_i}{n}, \tag{2.14}$$

де P_i – значення з вибірки;

n – основний параметр, довжина згладжування або період SMA (кількість значень, що входять у розрахунок ковзного).

SMA являє собою якийсь показник рівноваги значень за певний період, чим коротше SMA, тим за менший період береться рівновага. Усереднюючи значення, вона завжди слідує за головною тенденцією ринку, фільтруючи дрібні коливання. Чим менший параметр SMA (коротке ковзне середнє), тим швидше воно визначає нову тенденцію, але й одночасно робить більше помилкових коливань, і навпаки чим більший параметр (довге ковзне середнє), тим повільніше визначається новий тренд, але надходить менше помилкових коливань [2].

Чисельна фільтрація: високий, низький та смуговий фільтри

Методи чисельної фільтрації представляє дані різних періодів у часових рядах. Використовується для гідрологічних часових рядів (опади, випаровування) річних, місячних, добових та годинних значень. Дані повинні бути повним (без розривів) рівновіддаленим часовим рядом. У результаті метод формує відфільтрований часовий ряд, який показує короткострокові/довгострокові зміни початкового часового ряду. Фільтр низьких частот показує кращі результати, ніж метод поточного середнього значення на довгострокових періодах [44].

Лінійна регресія

Лінійна регресія описує лінійну залежність однієї змінної у від іншої незалежної змінної x.

Формула лінійної регресії [45]:

$$y = c + ax, \tag{2.15}$$

де y – значення ознаки по лінії регресії, тобто теоретичні значення,

a – кутовий коефіцієнт регресії;

x – значення ознаки-фактору (предиктора);

c – вільний член, константа.

Для знаходження лінійної регресії використовується метод найменших квадратів. Метод найменших квадратів у даному випадку використовується для рішення системи лінійних рівнянь [32].

За умови, що x – набір n невідомих змінних (параметрів), (2.16) – це сукупність змінних цього набіра змінних [32].

$$f_i(x) i = 1,\dots, m,\ m > n \tag{2.16}$$

Формула методу найменших квадратів [32]:

$$\sum_i e_i^2 = \sum_i (y_i - f_i(x))^2 \to min(x) \tag{2.17}$$

де x – набір n невідомих змінних (параметрів).

Точність лінійної регресії визначається за формулою середньої помилки апроксимації [46]:

$$\underline{A} = \frac{\Sigma \left| y_i - y_x \right| : y_i}{n} 100\% \tag{2.18}$$

При оцінці найменших квадратів сума квадратів помилок (тобто різниця між значенням даних та значенням функції регресії), також звана залишковою дисперсією, зводиться до мінімуму. Це означає, що лінійна регресія найкраще відповідає емпірично визначеним (або виміряним) значенням у. Сума квадратів у вертикальному (у) напрямку менша за суми в будь-якому іншому напрямку. Використовується для змінних, які лінійно залежать від інших, неперервних змінних. Дані повинні бути у лінійному відношенні між залежними та незалежними змінними. Незалежні та нормально розподілені залишки з постійною дисперсією. Порушення цих припущень може призвести до помилкових результатів та висновків [27].

Порівняння різних відрізків часу щодо середнього значення, мінливості та/або розподілу. Різниця між часовими зрізами - порівняння „майбутніх зрізів часу” від модельованих проекцій (наприклад, 2021–2050, 2071–2100) із зрізами часу з контрольного періоду (наприклад, 1961–1990); останній період використовує змодельовані або спостережувані значення або розподіли. Порівняння різних відрізків часу щодо середнього значення, мінливості та/або розподілу. Використовується для різних кліматичних змінних, таких як опади, температура, швидкість вітру тощо, а також похідні показники, такі як кліматологічні порогові дні. Порівняні часові зрізи повинні складати однаковий проміжок, і вони повинні бути достатньо довгими для статистичного опису клімату (бажано принаймні 30 років). У результаті метод виявляє сигнали щодо зміни клімату. Оскільки розглядаються відносні сигнали змін (що стосуються змодельованих даних, а не спостережуваних), тенденції, що виникають внаслідок різних моделей з різними систематичними помилками, стають порівнянними.

Структурний аналіз часових рядів, метод максимальної ймовірності

Функції щільності ймовірності Гаусса (нормальної) [47], Гумбеля [48] та Вейбулла [49] описуються за допомогою 2 залежних від часу параметрів (середнє та стандартне відхилення). Використовується для таких параметрів, як середньомісячна температура, середньомісячна загальна кількість опадів, дані повинні бути довгими часовими рядами без розривів тривалістю щонайменше 100 років. У результаті отримуються тенденції середнього та стандартного відхилення. Застосування методу вимагає опису значень часових рядів за допомогою функції щільності ймовірності (тест Колмогорова – Смірнова) [28].

Нелінійна регресія

Нелінійна регресія - це спосіб знаходження нелінійної моделі взаємозв'язку між залежною змінною та набором незалежних змінних. На відміну від традиційної лінійної регресії, яка обмежена оцінкою лінійних моделей, нелінійна регресія може оцінювати моделі з довільними взаємозв'язками між незалежними та залежними змінними. Це досягається за допомогою ітераційних алгоритмів оцінки, хоча така процедура не є обов'язковою для простих поліноміальних моделей виду, яку можна оцінювати з використанням традиційних методів, таких як процедура лінійної регресії [52].

Розрізняють дві групи нелінійних регресійних моделей: моделі, нелінійні щодо включених в аналіз пояснюючих змінних, але лінійні за параметрами, що оцінюються; моделі нелінійні за параметрами, що оцінюються [53].

До першої групи належать, наприклад, такі види функцій [53]:

$$y = a + b \cdot x + c \cdot x^2 + \varepsilon \quad (2.19)$$

- поліном 2-го ступеня;

$$y = a + b \cdot x + c \cdot x^2 + d \cdot x^3 + \varepsilon \quad (2.20)$$

- поліном 3-го ступеня;

$$y = a + \frac{b}{x} + \varepsilon \quad (2.21)$$

- гіпербола.

До другої групи належать [53]:

$$y = a \cdot x^b \cdot \varepsilon \quad (2.22)$$

- степенева;

$$y = a \cdot b^x \cdot \varepsilon \quad (2.23)$$

- показникова;

$$y = a + b \ln x + e \quad (2.24)$$

- лінійно логаріфмічна (полулогаріфмічна) та ін. види функцій.

У нелінійній регресії, статистичні моделі пов'язує вектор незалежних змінних і пов'язані з ним залежні змінні, що спостерігаються. Функція нелінійна в компонентах вектора параметрів, але в іншому лінійна.

Деякі функції, такі як експоненційні або логарифмічні, можуть бути перетворені таким чином, щоб вони були лінійними. При такому перетворенні можна виконати стандартну лінійну регресію, але слід застосовувати її з обережністю.

Як правило, відсутній вираз закритої форми для параметрів найкращої відповідності, оскільки існує лінійна регресія. Зазвичай алгоритми чисельної оптимізації застосовуються для визначення параметрів найкращого підбору. На відміну від лінійної регресії, можливо багато локальних мінімумів оптимізованої функції, і навіть глобальний мінімум може створювати зміщення. На практиці оціночні значення параметрів використовуються у поєднанні з алгоритмами оптимізації, щоб спробувати знайти глобальний мінімум суми квадратів.

В основі оцінки точності нелінійних регресій лежить те, що модель може бути апроксимована лінійною функцією.

Найкращою залежністю часто вважається та, яка мінімізує суму квадратів. Однак у випадках, коли залежна змінна немає постійної дисперсії, можна мінімізувати зважену суму квадратів. В ідеалі, кожна вага повинна дорівнювати дисперсії спостережень, але ваги можуть бути повторно обчислені для кожної ітерації [54].

2.2.4 Дисперсійний аналіз

Розкладання мінливості змінної залежно від різних класів. З цією метою досліджувана змінна поділяється на різні класи (залежно від обраних факторів). Наскільки фактори впливають на мінливість змінної, можливо

побачити в середніх показниках класів, які пов'язані з факторами; ці середні значення відрізняються між собою. Також можливо визначити, чи пояснюється мінливість відомими впливами (факторами) чи іншими, поки невідомими впливами. Існує кілька підгалузей дисперсійного аналізу. Загальним для всіх є те, що для кожної групи факторів розраховується тестова статистика, яка кількісно визначає відношення пояснених до незрозумілих дисперсій. Статистика тесту розподілена за двома ступенями свободи. Щоб прийняти рішення серед вибраних нульових гіпотез (які, як правило, стверджують, що між підгрупами факторів не існує відхилення), використовується статистична статистика для порівняння її з таким розподілом і отримується ймовірність P, щоб знайти коефіцієнт відхилення принаймні настільки високий, яким спостерігається коефіцієнт. Якщо P дуже малий (менше рівня значущості), ця нульова гіпотеза відхиляється. Використовується для змінних, які залежать від інших змінних, які розкладаються на фактори. Дисперсійний аналіз описує цю залежність. Кожна група факторів повинна мати подібний розмір. Досліджувані змінні повинні бути у нормальній формі розподілу з постійною дисперсією. Метод визначає, чи існують суттєві відхилення між окремими групами змінних, які можна розкласти на фактори [29].

Приклад:

У статті “Оцінка атмосферного тиску”, Krüger O, von Storch H [30] використовується “двосторонній дисперсійний аналіз”, який дозволяє оцінити взаємодію двох факторів. База даних містить кореляційні зв’язки між геострофічними вітрами (обчисленими з просторово-часових полів тиску повітря за допомогою тріангуляції) та при поверхневими вітрами, які були перетворені на приблизну нормальну форму розподілу. Автори вивчали вплив поверхні (суші чи океану) та розміру трикутників (великого, середнього чи малого) на кореляцію між обома змінними вітру та з'ясовували, чи існує

взаємодія між поверхнею факторів та розмірами. Якщо ці два фактори не залежать один від одного, їх відповідні впливи на аналізовану змінну вітру також не залежать один від одного.

Дані були відсортовані за 6 класами (“поверхня”, 2; “розмір”, 3) з однаковим розміром класу. Застосований метод дисперсійного аналізу використовує три нульові гіпотези, з яких дві стосуються прямих ефектів, а одна - інтерактивних.

Перша нульова гіпотеза, H0, стверджує, що немає різниці середніх значень (трансформованих кореляцій), яка пов’язана з фактором “поверхня”. Відповідна альтернативна гіпотеза H1 стверджує, що існує різниця.
Аналогічно, H0 та H1 будувались на основі коефіцієнта "розмір". Нульова гіпотеза H0 щодо інтерактивних ефектів стверджує, що ефекти поверхні та розміру не залежать один від одного і впливають один на одного. Альтернативна гіпотеза, H1, стверджує, що існує взаємодія та залежність. ANOVA виявив для тесту щодо недіючих факторів співвідношення між поясненою та незрозумілою дисперсією приблизно 0,7. Розподіл F (з 2 і 690 градусами свободи) дає значення P приблизно 0,45 для спостереження коефіцієнта дисперсії 0,7 або вище. Це велике значення P (більше, ніж зазвичай використовуються рівні значущості, скажімо, 0,05 або 0,01) означає, що H0 (“відсутність взаємодії”) не можна відхилити. Фактори «поверхня» та «розмір» виявляються в цьому аналізі незалежними один від одного. Крім того, автори встановили, що обидва інші тести призводять до відхилення нульових гіпотез: фактори "поверхня" та "розмір" мають значний вплив на кореляцію швидкості вітру.

2.2.5 Корекція упередженості

У контексті кліматичного моделювання упередження означає систематичне відхилення змінної кліматичної моделі від спостережуваного

аналога. Наприклад, кількість опадів протягом часового інтервалу після 1950 року, з ряду регіональних кліматичних моделей є вказівки, що для регіону Центральної Європи – Скандинавії упередження є позитивним, тобто кліматичні моделі систематично завищують кількість опадів (Goodess et al (2009). Зазвичай вважається, що упередженість пов'язана з неадекватними формулюваннями моделей, корінням яких є наші неповні знання про кліматичні процеси та обмежені можливості наших комп'ютерів.

Одним із засобів вирішення цієї дискомфортної ситуації, очевидно, є побудова кращих кліматичних моделей; хоча це робиться постійно моделюючими групами, це процес, який вимагає часу на розробку. Іншим, "швидким і брудним" засобом є виправлення результатів кліматичної моделі таким чином, щоб упередженість зникала. Успіх корекції зміщення кліматичної моделі критично залежить від (1) відповідного стохастичного опису форми зміщення (наприклад, адитивного/мультиплікативного або постійного/залежного від часу) та (2) наявності точних та чітко вирішених (у просторі і часі) даних спостережень. Іншим критичним моментом є небезпека невідповідності між змінними кліматичної моделі, які коригуються із зміщенням, та іншими змінними, які не є такими; розглянемо, наприклад, залежність між температурою повітря та типом опадів (дощ, сніг). Область корекції упередженості є досить новою для кліматичного моделювання, і потрібно очікувати значно нових подій у майбутньому.

Вибір методу корекції упередженості, який зараз використовують багато моделей кліматичних систем, - це квантильне відображення, де встановлюється залежність між функцією розподілу змодельованої змінної та функцією розподілу спостережуваної змінної. Цей метод може добре працювати, коли метою аналізу є середні кліматичні стани, але метод може працювати менш ефективно, коли аналізуються екстремальні кліматичні стани.

2.2.6 Квантильне відображення з функцією передачі

За допомогою методу здійснюється статистична корекція систематичних відхилень даних кліматичної моделі від спостережуваних кліматичних даних у минулому.

Використовується для аналізу таких кліматичних показників, як опади, температура приземного повітря, глобальної освітленість. Дані спостережень повинні бути достатньо якісними, з добовою роздільною здатністю та досить високою просторовою роздільною здатністю. Крім того, ряд спостережень повинен бути достатньо довгим, щоб гарантувати, що очевидна упередженість моделі не є результатом короткочасної мінливості. Сигнал зміни клімату на даних, скоригованих із зміщенням, може відхилятися від сигналу на нескоригованих даних; незрозуміло, який із двох сигналів зміни клімату є ближчим до реальності.

Застосовуючи дані про клімат, скориговані цим методом, потрібно врахувати:

(1) "внутрішня модель" узгодженості різних змінних клімату може бути втрачена внаслідок корекції упередженості;

(2) сигнал зміни клімату може сам піддаватися змінам через корекцію зміщення;

(3) дані спостережень і сам метод схильні до невизначеності (і в даний час аналізуються в кліматичних дослідженнях) [31].

Приклад з столичного району Гамбурга: річні значення (рис. 2.7) та сезонний цикл (рис. 2.8) опадів, змодельованих за допомогою REMO в 3 реалізаціях клімат-контролю, 1961–2000 рр., Та виправлені з урахуванням зміщення опадів 1-ї реалізації кліматичного контролю (bc_C20_1) з використанням даних спостереження DWD_REGNIE [5].

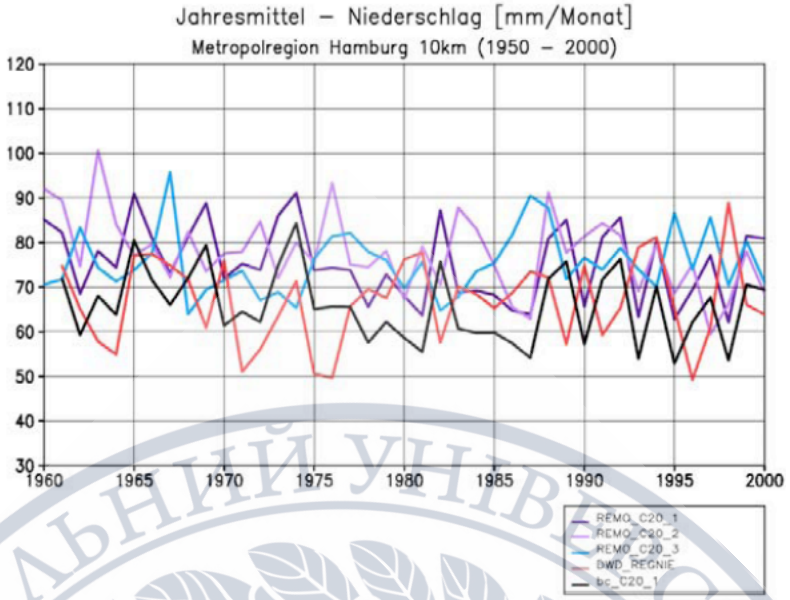

Рисунок 2.7. Приклад роботи методу квантильного відображення (1) [5]

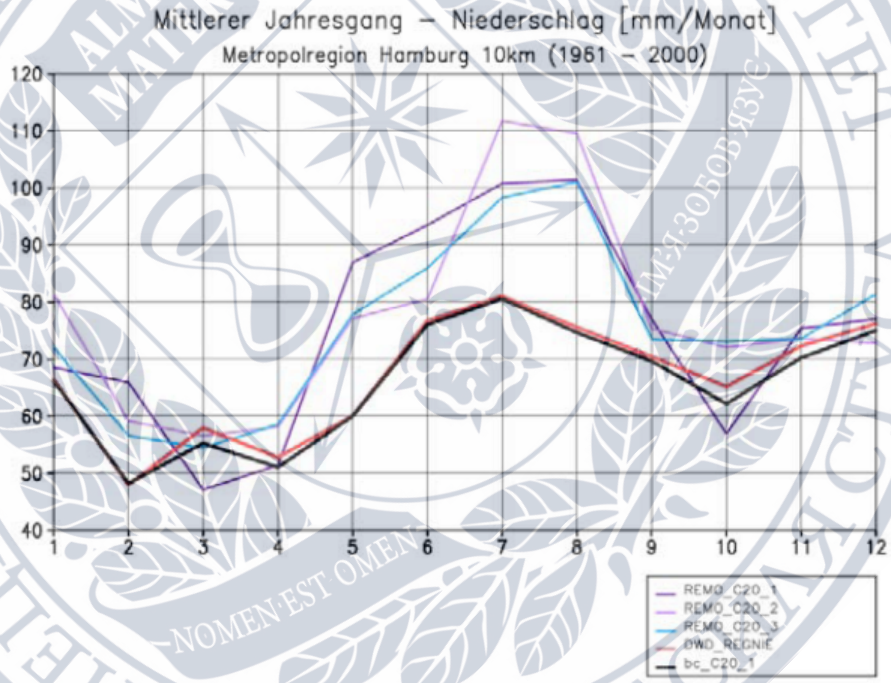

Рисунок 2.8. Приклад роботи методу квантильного відображення (2) [5]

Висновок до розділу 2

У даному розділі було визначено параметризацію математичних моделей, математичні моделі різних статистичних методів для деяких метеорологічних показників.

Також у даному розділі були розглянуті основні статистичні методи, які використовуються для прогнозу погоди та при прогнозуванні клімату, стихійних лих.

РОЗДІЛ 3.
РЕАЛІЗАЦІЯ МЕТОДІВ АНАЛІЗУ МЕТЕОРОЛОГІЧНИХ ДАНИХ

У цьому розділі застосовуються методи аналізу даних до даних спостережень погоди. У якості даних спостережень погоди береться набір метеорологічних даних в Австралії [55].

3.1. Виявлення лінійних залежностей за допомогою кореляційного аналізу

Кореляційний аналіз визначає міру статистичної залежності між двома змінними.

За допомогою стандартної формули коефіцієнту кореляції Пірсона було виведено таблицю з коефіцієнтами кореляції всіх показників до всіх показників. Коефіцієнти кореляції для показників, які не приймали числове значення, у додатку А, таблиці 3.1 позначені як False. Назви параметрів наведені у параметризації моделі у розділі 2.

Коефіцієнти кореляції, більші за 0.7 між різними показниками наведено у додатку А за таблицею 3.2.

Розглянемо першу кореляцію між показниками мінімальної температури повітря та температури у 9 годин ранку, яка складає 0.95. Ця кореляція означає, що мінімальна температура сильно корелює з температурою в 9 годин ранку.

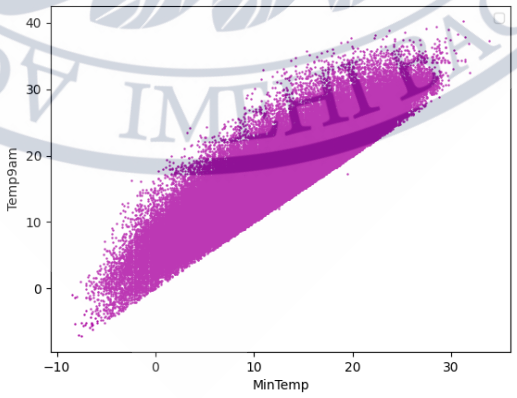

Рисунок 3.1 Точкова діаграма для показників температури в 9 годин ранку та мінімальної температури

З діаграми видно, що у показників температури в 9 годин ранку та мінімальної температури є сильна лінійна позитивна кореляція, що пояснюється низькими температурами відносно доби у регіонах Австралії.

Спробуємо спрогнозувати температуру повітря в 9 годин ранку на основі показника мінімальної температури. Для цього застосуємо математичну модель (2.1).

Представимо деякі фрагменти коду, які прогнозують дану модель. Весь код написаний мовою програмування Python у додатку Б, лістингу 3.1.

Після здійснення прогнозу температури у 9 годин ранку за математичною моделью (2.1) на основі показника мінімальної температури беручи до уваги весь набір даних середня похибка склала 1.89.

Виведемо отримані результати у вигляді точкової діаграми у часовому розрізі:

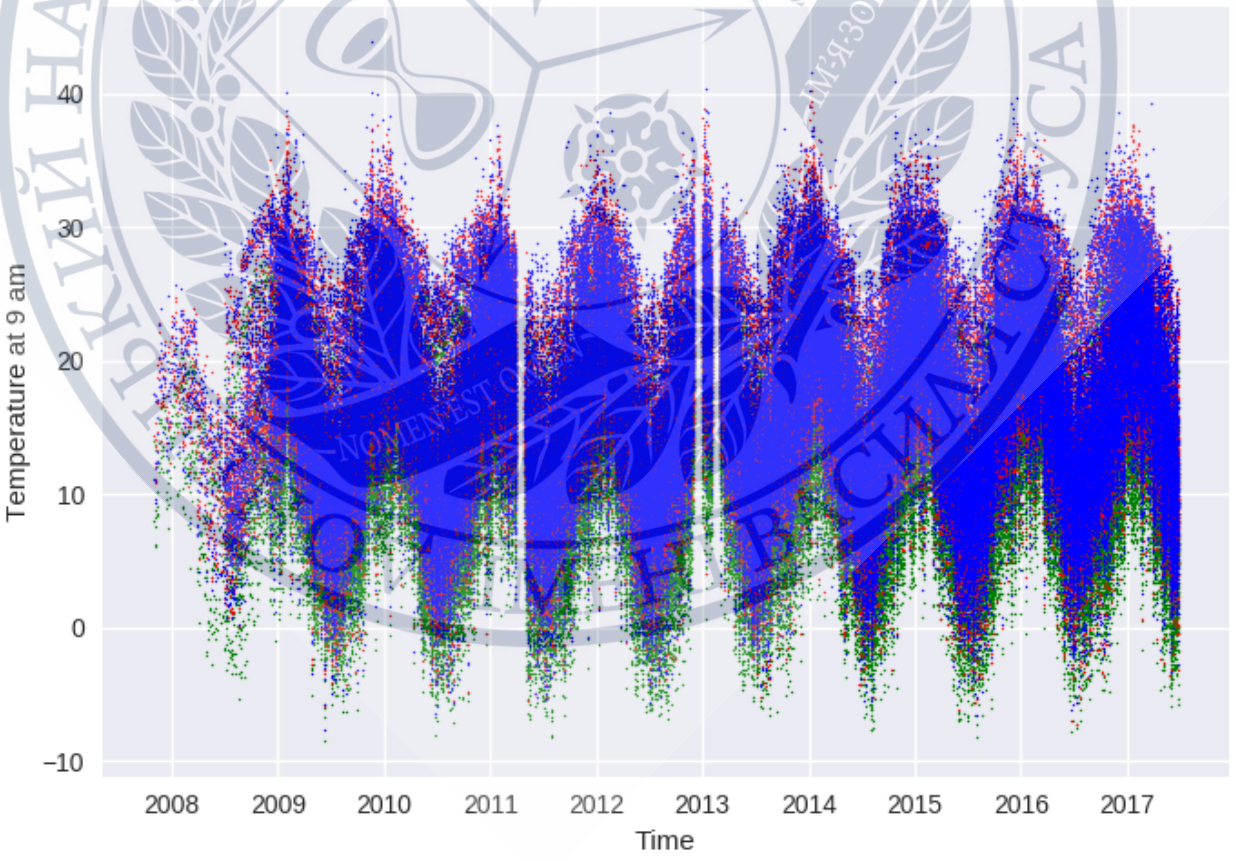

Рисунок 3.2 Точкова діаграма для показників температури в 9 годин ранку та мінімальної температури з прогнозом погоди на весь набір даних

На рис. 3.2 показники мінімальної температури позначені зеленим кольором, температура у 9 годин ранку - червоним кольором та прогнозовані значення - синім кольором. Прогнозовані дані майже збігаються з фактичними та похибка зі значенням у 1.89 градусів підходить для прогнозу погоди для пересічних громадян.

Спробуємо локалізувати дані за одним містом та зробити прогнози погоди, беручи до уваги тільки ці дані. Місто, за яким будуть фільтруватись дані має назву Albury. Після здійснення прогнозу з локалізацією середня похибка склала 13.06, що є кращим, ніж у попередньому прогнозі більше ніж на 3 градуси.

Виведемо отримані результати у вигляді точкової діаграми у часовому розрізі:

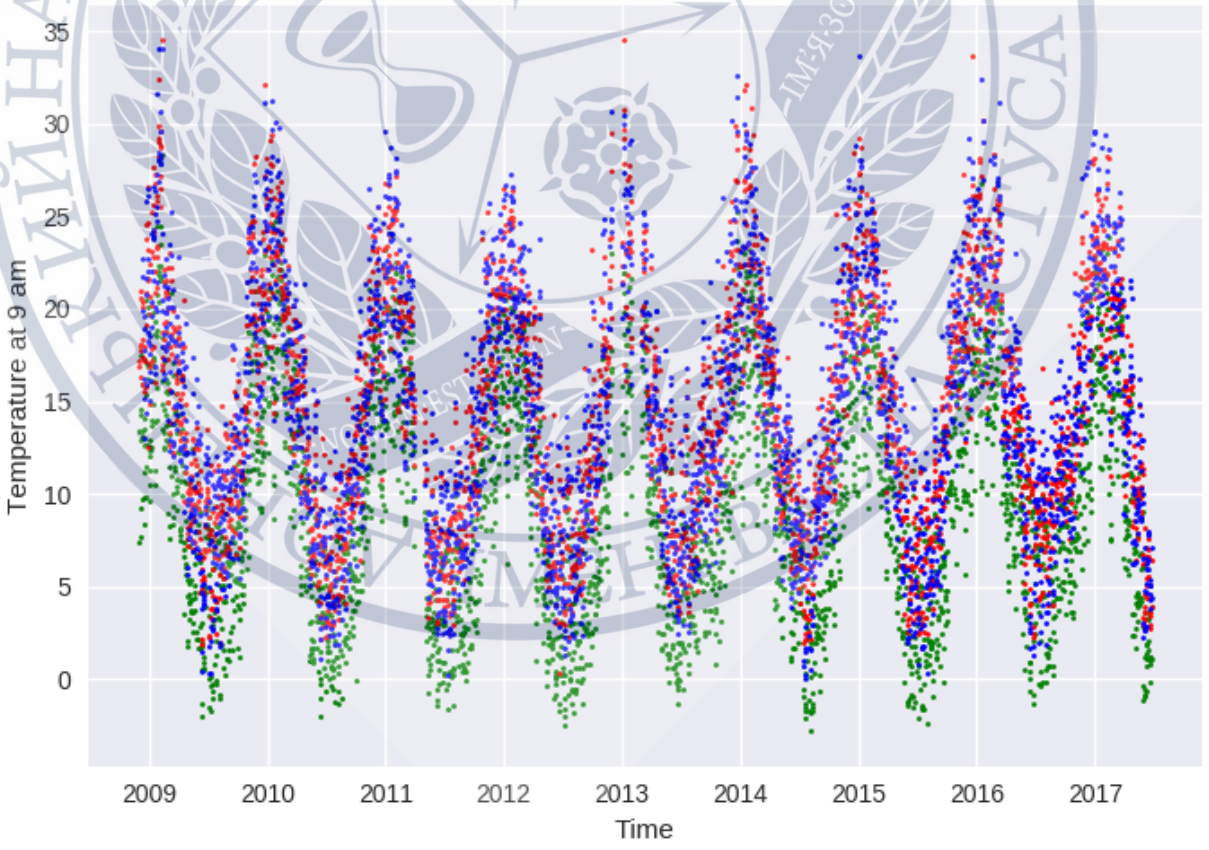

Рисунок 3.3 Точкова діаграма для показників температури в 9 годин ранку та мінімальної температури з прогнозом погоди для міста Albury

На рис. 3.3 показники позначені тими ж самими кольорами, як і на рис. 3.2. З цього рисунку ще краще видно, що прогнозовані дані дуже часто збігаються з фактичними, але не у 100%, про що свідчить похибка у 1.65 градусів, що трохи краще, ніж у минулому прогнозі.

Спробуємо додати фільтр екстремальних частот, не беручи до уваги показники температури в 9 годин ранку, значення яких більші за 22 та менші за 5. Після здійснення прогнозу з локалізацією середня похибка склала 1.68, що є трохи гіршим за похибку у попередньому прогнозі, що свідчить про те, що додавання фільтрів не завжди добре впливає на прогноз.

3.2. Поділ метеоданих на класи за допомогою частотного розподілу

Цим методом можуть бути відсортовані та поділені на класи будь-які метричні дані за їх значеннями.

Одновимірний частотний розподіл

Проаналізуємо методом одновимірного частотного розподілу показники температури у 9 годин ранку та напрямок вітру у 9 годин ранку.

За допомогою формули SCOTT (2.4) поділимо показник температури на класи та виведемо отримані дані у вигляді діаграми.

Фрагмент коду, які використовувались для сортування даних представлені у додатку Б, лістингу 3.2.

Так як у попередньому підрозділі з проведенням локалізації була отримана більша точність прогнозу, відсортуємо дані за містом Albury.

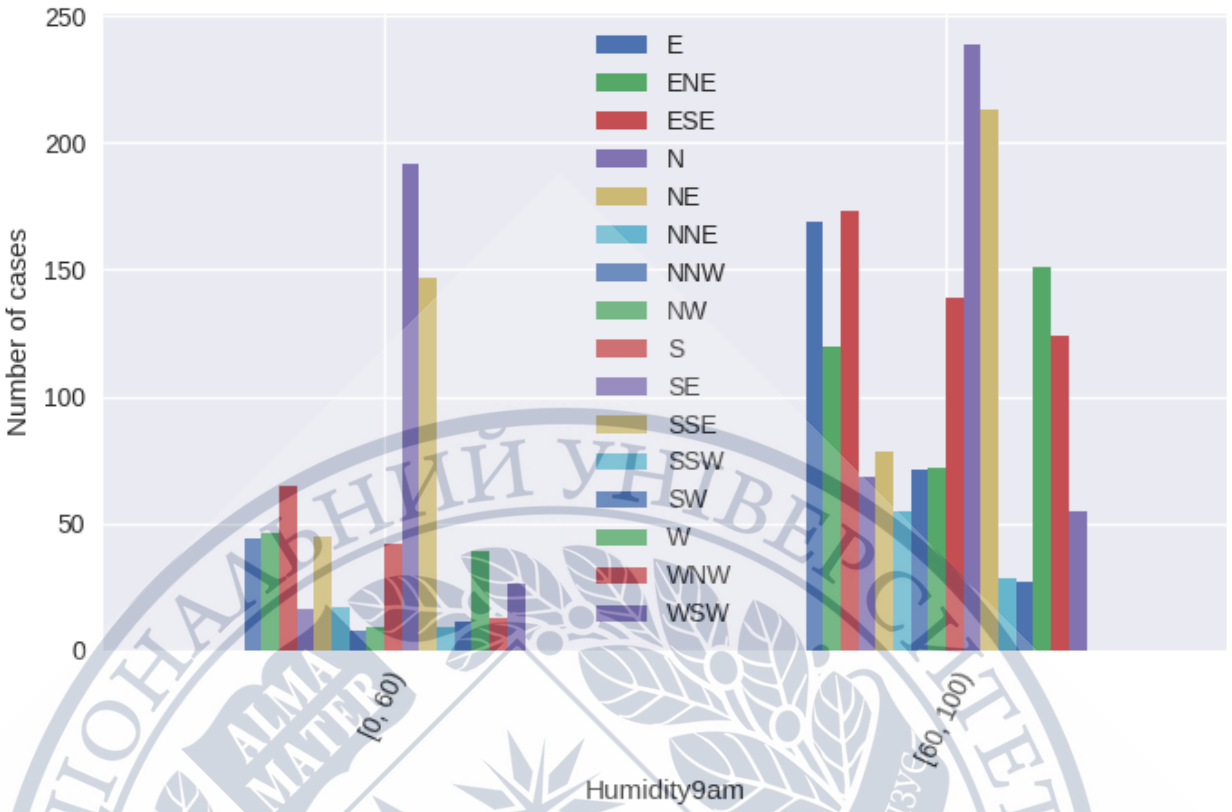

Рисунок 3.4 Гістограма для показників вологи повітря у 9 годин ранку та напрямку вітру в 9 годин ранку

За допомогою гістограми на рис. 3.4, яку було побудовано за допомогою методу частотного розподілу, було просорторовано показник вологи у 9 годин ранку за показниками напрямку вітру в 9 годин ранку. Різними кольорами позначено швидкість вітру. З гістограми видно, що у більшості випадків вологість повітря не менша за 60%. Також за вологістю меншою за 60%, напрямок вітру приймає північний, північно-східний напрямки. Інші напрямки вітру розподілені майже рівномірно.

Спробуємо зробити прогноз напрямку вітру за показником вологи повітря за допомогою методу частотного розподілу за мат. моделью (2.1). Вибірку також локалізуємо за містом Albury.

Представимо деякі фрагменти коду, які прогнозують дану модель у додатку Б, лістингу 3.3. Весь код також написаний мовою програмування

Python. Поділ на класи відбувався також за формулою SCOTT (2.4) та написаного вище коду, який її реалізує.

Після апробації даного коду напрямок вітру був спрогнозований вірно лише у 260 з 3040 випадків, точність складає 8.55%.

Невелика точність одномірного розподілу обумовлена тим, що метеорологічні дані розподілені досить рівномірно, а при прогнозі напрямків вітру класів розбиття занадто багато для отримання якісного прогнозу. Отже, одновимірний розподіл краще підходить для візуалізації того, як розподіляються дані, поділені за класами та прогнозування у ансамблях з іншими методами у якості корректировки інших прогнозів, але використання даного методу, як єдиного методу прогнозування не забезпечує необхідної точності прогнозу.

Двовимірний частотний розподіл

За допомогою кругової гістограми можливо проаналізувати двовимірний частотний розподіл.

Гістограма частоти зустрічальності спостережуваної швидкості та напрямку вітру у регіоні Австралії та за контрольний період 2008-2017. Ширина класу швидкості вітру становить 1 м/с та напрямку 22.5°. Кількість подій класифікується кольорами від кольорового поля праворуч.

Представимо деякі фрагменти коду, які прогнозують дану модель у додатку Б, лістингу 3.4. Весь код також написаний мовою програмування Python.

На рисунку нижче представлена реалізована гістограма двовимірного частотного розподілу за показниками швидкості та напрямку вітру.

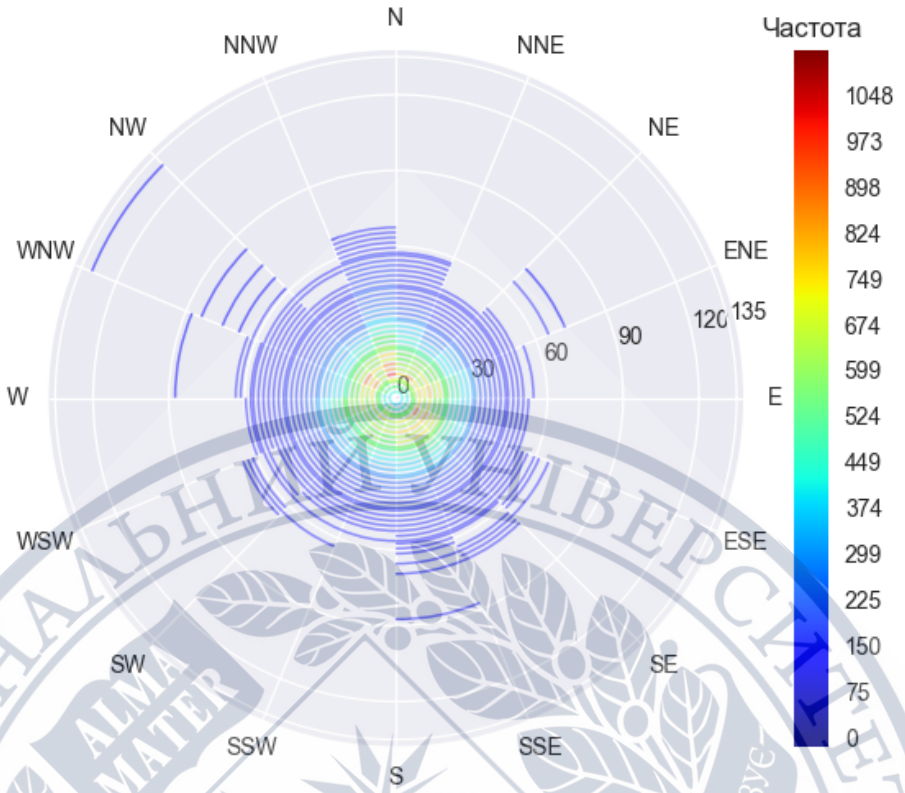

Рисунок 3.5. Кругова гістограма для показників швидкості вітру у 9 годин ранку та напрямку вітру в 9 годин ранку

З рисунка 3.5 видно, що сильних вітрових подій зі швидкістю вітру більше 30 м/с (що відповідає 11 класу Бофорта) мало. Для цих подій напрямок вітру коливається розподілений рівномірно. Подій із середньою швидкістю вітру (наприклад, 5-20 м/с) більше надходять із північно-західних напрямків, ніж х решти інших напрямків. Більше того, з рисунка видно, що чим вище швидкість вітру, тим менша кількість подій і напрямок вітру повертається у північному напрямі. На відміну від подій з низькою швидкістю вітру (наприклад, 5-20 м/с), які, очевидно, спостерігались у більшості випадків, переважно відбуваються події з південно-східних, північно-західних напрямків.

Кругова гістограма добре підходить для аналізу показників, які залежать від простору, таких як напрям вітра, вона зрозуміло пояснює залежність показників один від одного, але для порівняння інших показників

за методом частотного розподілу більше підходить звичайна гістограма, тому що за допомоги числової шкали можливо більш точніше оцінити кількість випадків, ніж за допомогою кольорової шкали.

3.3. Виявлення лінійних та нелінійних залежностей за допомогою регресійного аналізу

Для прогнозу різних показників погоди у майбутньому використовуються показники, отримані у минулому часі та аналізуються для отримання прогнозу погоди. У результаті можливо отримати тенденцій у зміні клімату, наявність сезонних і циклічних компонент.

Лінійна регресія

У статистиці важливим є визначення співвідношення між двома випадковими змінними. Це дає можливість робити прогнози щодо однієї змінної щодо інших. Регресійний аналіз, як і кореляція, застосовуються у прогнозі погоди.

Регресійний аналіз – кількісне уявлення зв’язку або залежності між залежною змінною і незалежної/незалежними змінними.

Для знаходження лінійної регресії використовується метод найменших квадратів для рішення систем рівнянь, формула (2.17). Метод найменших квадратів дозволяє отримати такі оцінки параметрів, при яких сума квадратів відхилень фактичних значень результативної ознаки у від теоретичних y(x) мінімальна.

Представимо деякі фрагменти коду, які прогнозують дану модель у додатку Б, лістингу коду 3.5. Весь код також написаний мовою програмування Python.

На рисунку нижче представлена реалізована діаграма розсіювання за показником температури у 9 годин ранку з візуалізацією лінійної регресії.

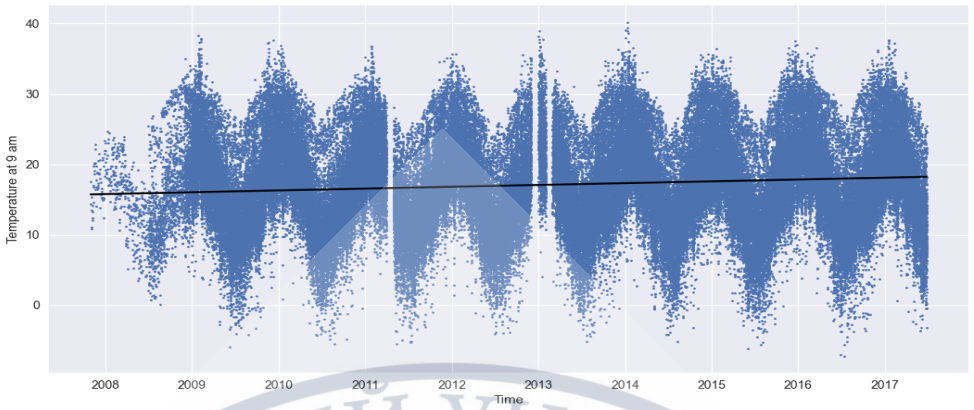

Рисунок 3.6. Діаграма розсіювання показника температури в 9 годин ранку у розрізі часу та лінія регресії

На рис. 3.6 зображено показник температури в 9 годин ранку за 2008-2017 роки у Австралії. На рисунку видно сезонні коливання температури, при чому саме у початку року показники температури набувають максимальних значень, що свідчить про те, що у Австралії літній сезон діє у грудні-лютні.

Також на рисунку зображена лінія тренду, яка свідчить про стабільні сезонні зміни показника температури, але приблизно 2 градуси Цельсія, на які у середньому збільшився показник температури за менше, ніж у 10 років свідчить про те, що існує тренд потепління, що підтверджує теорію про глобальне потепління [33]. Згідно отриманій лінійній регресії середні значення показників температури в Австралії постійні та кожного року приймають майже однакові значення та можливо зробити припущення, що отримані значення показників температури будуть надалі повторюватись у майбутньому.

Розраховуємо середню помилку апроксимації за формулою (2.18). Значення середньої помилки апроксимації до 15% свідчить про добре підібрану модель рівняння.

Представимо деякі фрагменти коду, які прогнозують дану модель. Весь код також написаний мовою програмування Python.

Лістинг коду 3.6:

Розрахунок помилки апроксимації:

```
expression1s = []

i = 0
for y in values:
   if y != 0:
      expression1s.append(abs(y-regressionXs[i])/y)

approximationError = sum(expression1s)/len(values)*100
```

Середня помилка апроксимації склала 41.75%, що свідчить про те, що показники температури від сезону до сезону коливаються більш менш рівномірно та відсутність сильного тренду для показників температури у Австралії.

Нелінійна регресія

Нелінійна регресія – регресійна модель залежності результативної змінної від однієї або декількох змінних, що пояснюють, що виражається у вигляді нелінійної функції [34].

Нелінійні регресії бувають різними. Роздивимось поліноміальну регресію. Для знаходження поліноміальної регресії використовується функція np.poly1d [35] від бібліотеки Numpy[36], математична модель - (2.4).

Представимо деякі фрагменти коду, які прогнозують дану модель. Весь код також написаний мовою програмування Python.

Лістинг коду 3.7:

Розрахунок нелінійної регресії:

```
p5 = np.poly1d(np.polyfit(x,y,2))
p5x = p5(x)
```

Візуалізація даних:

```
plt.plot(x, p5x, 'k-', color='red')
plt.scatter(values1,values2, c = '#0800a3', s=5)
plt.xlabel("Minimum Temperature")
plt.ylabel("Temperature at 9 am")
plt.show()
```

За допомогою вище приведеного коду було розраховано поліноміальну регресію для показників температури в 9 годин ранку та мінімальної температури в 9 годин ранку за містом Albury.

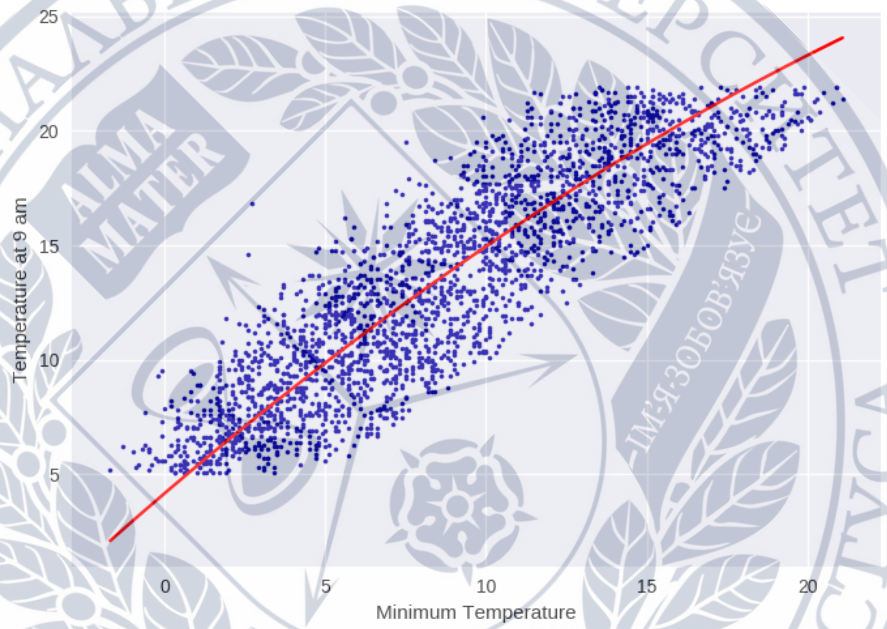

Рисунок 3.7 Точкова діаграма для показників температури в 9 годин ранку та мінімальної температури з поліноміальною регресією 2-го порядку

На рис. 3.7 спостережується нелінійна регресія за показниками температури, видно, що показники пов’язані між собою та мають нелінійну залежність.

Розраховуємо середню помилку апроксимації за формулою (2.18).

Представимо деякі фрагменти коду, які прогнозують дану модель. Весь код також написаний мовою програмування Python.

Лістинг коду 3.8:

Розрахунок помилки апроксимації:

```
expression1s = []

i = 0
for val in y:
    if y != 0:
        expression1s.append(abs(val-p5x[i])/val)

approximationError = sum(expression1s)/len(y)*100
```

Середня помилка апроксимації склала 56.57%, що свідчить про те, що хоча між показниками температури є нелінійна залежність, вони мають стохастичний характер, який важко описати за допомогою поліноміальної функції. Після прогнозу показника температури у 9 годин ранку за допомогою даної регресії середня відхилення прогнозуємого значення від фактичного при прогнозі на один день складало 0.00903 значення, на два дні - 0.01227, на сім днів - 0.0284.

Спробуємо розрахувати поліноміальну регресію з порядком 5.

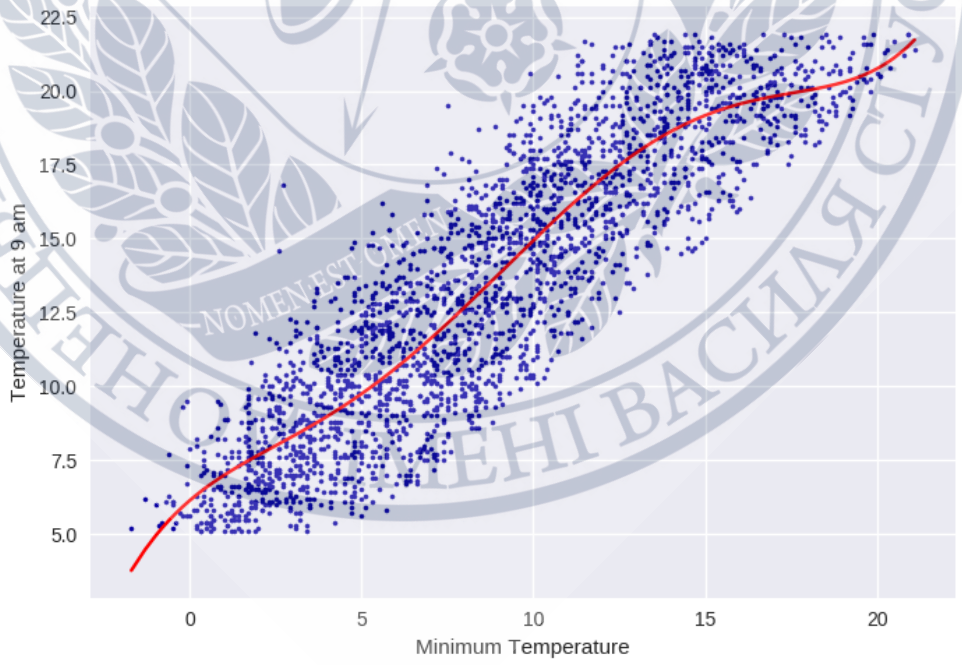

Рисунок 3.8 Точкова діаграма для показників температури в 9 годин ранку та мінімальної температури з поліноміальною регресією 5-го порядку

З рис. 3.8 видно, що поліноміальна регресія 5-го порядку більш чутлива до коливань, що дозволяє більш точно скласти прогноз. Помилка апроксимація склала 60.8%, що більше, ніж з функцією апроксимації 2-го порядку. Це пояснюється тим, що з функцією більшого порядку модель стає більш чутлива до відхилень та охоплює дані з більшим відхиленням для більшості даних. Після прогнозу показника температури у 9 годин ранку за допомогою даної регресії середня відхилення прогнозуємого значення від фактичного при прогнозі на один день складало 0.00894 значення, на два дні - 0.0121, на сім днів - 0.02788.

Спробуємо розрахувати поліноміальну регресію з порядком 20.

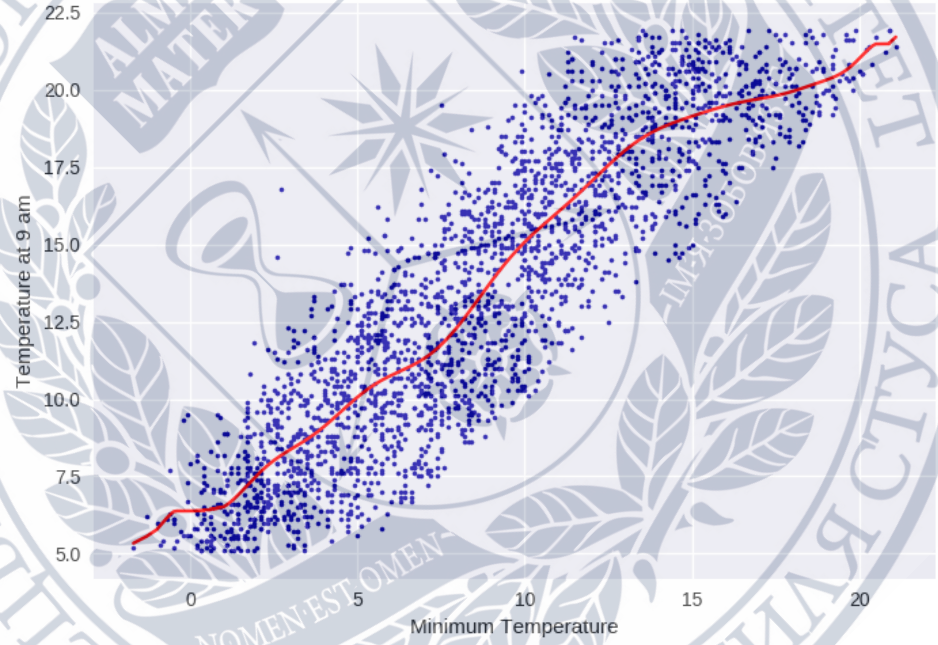

Рисунок 3.7 Точкова діаграма для показників температури в 9 годин ранку та мінімальної температури з поліноміальною регресією 20-го порядку

З рис. 3.8 видно, що поліноміальна регресія 20-го порядку ще більш чутлива до коливань, що дозволяє більш точно скласти прогноз. Помилка апроксимація склала 61.73%, що трохи більше, ніж з функцією апроксимації 5-го порядку. Результати поліноміальної функції 20-го порядку не сильно відрізняються за результати, отримані за допомогою функції 5-го порядку, що

свідчить, про те, що чутливість поліноміальної функції має певний максимум, до якого вона наближається з підвищенням порядку функції. Після прогнозу показника температури у 9 годин ранку за допомогою даної регресії середня відхилення прогнозуємого значення від фактичного при прогнозі на один день складало 0.00899 значення, на два дні - 0.01219, на сім днів - 0.02819.

За результатами, отриманими за допомогою поліноміальної регресії, можливо зробити висновок, що нелінійний регресійний аналіз більш чутливий до змін у даних, в тому числі метеорологічних даних. Та за допомогою нелінійної регресії можливо більш точно визначити тенденції змін у динаміці даних та зробити більш точний прогноз.

Висновок до розділу 3

Як було визначено у першому розділі, при використанні статистичних методів похибки неминучі, результат роботи статистичних методів є завжди правильним лише з якоюсь долею вірогідності, що пояснюється не тільки недосконалостями статистичних методів, але й внаслідок мінливості самого клімату.

Статистичні методи надають однобоку інформацію яка часто потребує інтерпретації зі сторони статистиків. Також для повного використання потенційних можливостей статистичних методів необхідно враховувати різні метеопоказники з різних сусідніх ділянок на різних висотах атмосфери, що потребує у розвертані великої кількості нових метеостанцій, а також потребує великих обчислювальних потужностей.

Велика частина аналізу метеоданих складається з інтерпретації отриманого результату. Завдяки цьому можливо знайти тенденції, які впливають глобально на клімат, локації, які потребують більшої уваги з-за великої небезпеки стихійного лиха, та можливі загрози, пов’язані з динамікою змін деяких показників.

ВИСНОВКИ

В даній магістерській роботі було розглянуто важливість та актуальність якісного прогнозування погоди у житті людини, сучасний стан аналізу метеоданих та прогнозу погоди, існуючі статистичні методи для прогнозу погоди. Також було протестовано основні статистичні методи на реальних даних з австралійського континенту.

Варто відмітити, що актуальність прогнозу погоди з часом не спадає, особливо у регіонах з великою кількістю стихійних лих, таких як, Японія. Прогноз погоди рятує життя мільйонів пересічних громадян завдяки попередженням стихійних лих. Також аналіз метеоданих використовується у великій кількості інших сфер, серед яких найбільше у авіації, флоту, сільському господарстві, енергетичному секторі, будівництві.

На даний момент сфера аналізу метеоданих швидко розвивається, використовуються та розробляються різні гідродинамічні моделі, які беруть до уваги якомога більше факторів, які впливають на стан погодних умов. Прогноз погоди став дуже точним у порівнянні з минулим сторіччям, для прогнозів на один день точність стала складати до 96%, що пояснюється появою можливості використовувати більш потужні обчислювальні пристрої. На даний час прогноз погоди може складатися максимум на 14 днів, тому що на стан атмосфери впливає велика кількість випадкових факторів, які неможливо спрогнозувати. Збільшення точності навіть на 0.01% є результатом збільшості обчислювальних потужностей у декілька разів. У даний час метеорологи ставлять задачу не зробити більш точним прогноз погоди, а зробити більш точним прогнозування небезпечних природних явищ.

Тестування статистичних методів підтвердило їх недоліки у точності. При отримані результатів тестування методів було виявлено важливість правильної інтерпретації їх результатів.

СПИСОК ВИКОРИСТАНИХ ПОСИЛАНЬ

1. Кендалл М., Стьюарт А. Статистические выводы и связи. 1973. 389 с.

2. Ya-lun Chou. Statistical Analysis. Holt International, 1975. section 17.9

3. Statistical methods for the analysis of simulated and observed climate data. Climate Service Center, 2013. pages 22-26

4. Andrea S Schaller, Johannes Franke, Christian Bernhofer. Climate dynamics: temporal development of the occurrence frequency of heavy precipitation in Saxony. Germany, 2020. pages 6-7

5. Robert Schoetter. Can local adaptation measures compensate for regional climate change in Hamburg Metropolitan Region? 2013. pages 111-112

6. Про затвердження Авіаційних правил України «Технічні вимоги та адміністративні процедури щодо льотної експлуатації в цивільній авіації»: Наказ В.о. Голови Державіаслужби від 05.07.2018 № 682
URL: https://zakon.rada.gov.ua/laws/show/z1109-18

7. Commercial aviation accidents statistics - Causes
URL: https://www.1001crash.com/index-page-statistique-lg-2-numpage-4.html

8. Aircraft Owners and Pilots Association. "Aircraft Icing". 2007. pages 2-4
URL:
https://web.archive.org/web/20070202074833/http://www.aopa.org/asf/publications/sa11.pdf

9. Why Do So Many Light Airplanes Crash? What's the Cause of Most Small Airplane Crashes?
URL:
https://www.highskyflying.com/why-do-so-many-light-airplanes-crash-whats-the-cause-of-most-small-airplane-crashes/

10. Weather Concerns for General Aviation
URL: https://flightsafety.org/asw-article/weather-concerns-for-general-aviation/

11. Importance of Weather Monitoring in Farm Production
URL:
https://blog.agrivi.com/post/importance-of-weather-monitoring-in-farm-production
12. Weather Forecasting for the Farmer. 2020
URL:
https://www.agritechtomorrow.com/article/2020/02/weather-forecasting-for-the-farmer/11981
13. МІНІСТЕРСТВО АГРАРНОЇ ПОЛІТИКИ УКРАЇНИ, НЕБЕЗПЕЧНІ ДЛЯ СІЛЬСЬКОГО ГОСПОДАРСТВА МЕТЕОРОЛОГІЧНІ ЯВИЩА, С. 1-15
URL:
http://www.tsatu.edu.ua/rosl/wp-content/uploads/sites/20/ahrometeorolohija-sr-nebezpechn-javyshcha.pdf
14. Ганна Яценко. ВПЛИВ ПОГОДНИХ УМОВ НА ЕКОНОМІЧНУ ДІЯЛЬНІСТЬ В УКРАЇНИ. 2020. С. 2-6
URL: https://journal.bank.gov.ua/uploads/articles/249_3_Yatsenko_Ukr.pdf
15. FAO. The impact of disasters and crises on agriculture and food security: 2021. Rome, 2021. pages 32-36
URL: https://doi.org/10.4060/cb3673en
16. Точность прогноза на завтра — 96%. 2015
URL:
https://polymus.ru/ru/museum/news/prognoz-pogody-na-zavtra-tochen-v-96-sluchaev/
17. Что плохо в российской метеорологии?. 2017
URL: http://www.sib-science.info/ru/news/prognoziruyut-pogodu-v-30082017
18. М. А. Толстых, Ж. Ф. Желен , Е. М. Володин. Разработка многомасштабной версии глобальной модели атмосферы ПЛАВ. 2015. С. 3-6
URL: https://core.ac.uk/download/pdf/287442079.pdf

19. Методический кабинет Гидрометцентра России. Краткосрочные и среднесрочные прогнозы полей метеорологических величин на основе гидродинамических моделей циркуляции атмосферы. 2011
URL: http://method.meteorf.ru/estimate/average/apr11/apr11.html
20. INTERGOVERNMENTAL OCEANOGRAPHIC COMMISSION. WMO ATMOSPHERIC RESEARCH AND ENVIRONMENT PROGRAMME WMO/IOC/ICSU WORLD CLIMATE RESEARCH PROGRAMME. 2005
21. Метеословарь — глоссарий метеорологических терминов. Поверхность изобарическая
URL: https://pogoda.by/glossary/?nd=14&id=143
22. Симушкин С.В. Многомерный статистический анализ. – Казань.: Издательство КГУ, 2006. С. 59
23. Кобзарь А. И. Прикладная математическая статистика. — М.: Физматлит, 2006. С. 624
24. Lothar Sachs. Statistische Methoden. 1982. pages 46-48
25. Mudelsee M. Climate Time Series Analysis: Classical Statistical and Bootstrap Methods. Springer. Dordrecht, 2010. page 474
26. Schuster A. On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. Terrestrial Magnetism, 1898. pages 13-41
27. Дж. Себер. Линейный регрессионный анализ. — М.: Мир, 1980. С. 456
28. Колмогоров А. Н. Основные понятия теории вероятностей. — 2-е изд. — Москва : Наука, 1974. С. 41
29. Шеффе Г. Дисперсионный анализ, пер. с англ. — М., 1963. С. 626
30. Krueger, O & J-S von Storch. (2011). A simple empirical model for decadal climate prediction. Journal of Climate, 2-3, doi: 10.1175/2010JCLI3726.1

31. Mudelsee, M., D. Chirila, T. Deutschländer, C. Döring, J.O. Haerter, S. Hagemann, H. Hoffmann, D. Jacob, P. Krahé, G. Lohmann, C. Moseley, E. Nilson, O. Panferov, T. Rath, B. Tinz. 2010: Climate Model Bias Correction und die Deutsche Anpassungsstrategie. Mitteilungen der Deutschen Meteorologischen Gesellschaft 03/2010, pages 2-7

32. Лоусон Ч., Хенсон Р. Численное решение задач методом наименьших квадратов. — М.: Наука, 1986. С. 10

33. А. М. Заморока. Антарктичне потепління триває // Станіславівський Натураліст

URL: http://www.naturalist.if.ua/?p=3840

34. Центр статистического анализа. Нелинейный регрессионный анализ

URL:

https://www.statmethods.ru/statistics-metody/nelinejnyj-regressionnyj-analiz/

35. Numpy - Function numpy.poly1d

URL: https://numpy.org/doc/stable/reference/generated/numpy.poly1d.html

36. Numpy - What is NumPy?

URL: https://numpy.org/doc/stable/user/whatisnumpy.html

37. Franziska Hufsky, Léon Kuchenbecker, Katharina Jahn, Jens Stoye. Swiftly Computing Center Strings. 2011, page 2-3

URL:

https://www.researchgate.net/publication/51062225_Swiftly_Computing_Center_Strings

38. В.Є. Бахрушин. МЕТОДИ АНАЛІЗУ ДАНИХ. Запоріжжя, Класичний приватний університет, 2011. С. 10

URL:

http://web.kpi.kharkov.ua/auts/wp-content/uploads/sites/67/2017/02/DAMAP_Ivashko_posobie2.pdf

39. Puchong Praekhaow. Determination of Trading Points using the Moving Average Methods. Conference: INTERNATIONAL CONFERENCE FOR A SUSTATIONAL GREATER MEKONG SUBREGION, King Mongkut's University of Technology-Thonburi, June 2010. page 1
URL: https://www.researchgate.net/publication/233988919_Determination_of_Trading_Points_using_the_Moving_Average_Methods
40. Marcus B. Perry. The Weighted Moving Average Technique. University of Alabama. June 2010. page 1
URL: http://dx.doi.org/10.1002/9780470400531.eorms0964
41. FRANK KLINKER. EXPONENTIAL MOVING AVERAGE VERSUS MOVING EXPONENTIAL AVERAGE FRANK KLINKER. Dortmund, Germany, 2011. page 5
URL: https://arxiv.org/pdf/2001.04237.pdf
42. Scott DW. On optimal and data-based histograms. 1979. pages 605-610
43. Von der Lippe. Deskriptive Statistik. Friedrich-Schiller-Universität Jena, 1993. pages 2-12
URL: http://www.von-der-lippe.org/dokumente/buch/buch07.pdf
44. C. D. Schönwiese, Gebrüder Bornträger. Praktische Statistik für Meteorologen und Geowissenschaftler. Berlin, Stuttgart, 1985. page 163
45. GEOFF NICHOLLS. BS1A APPLIED STATISTICS - LECTURES 1-16. UNIVERSITY Of Oxford, Department of statistics, 2012. page 2
URL: http://www.stats.ox.ac.uk/~nicholls/bs1a/lecturenotes1-16.pdf
46. Гамалий В. Ф., Дмитришин Б. В., доцент Сотников В.С. ЭКОНОМИКО-МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИ. Кировоградский национальный технический университет, Украина, 2014. С. 27

URL:
http://dspace.kntu.kr.ua/jspui/bitstream/123456789/7935/1/Ekonometryka.pdf
47. Вентцель Е. С. Теория вероятностей. — 10-е изд., стереотипное.. — М.: Высш. шк., 2006. С. 116
48. Лебедев Алексей Викторович. Неклассические задачи стохастической теории экстремумов: дис. ... д-ра ф.-м. наук: 01.01.05. Москва, 2015. 191 с.
URL: http://mech.math.msu.su/~snark/files/diss/0087diss.pdf
49. Papoulis, Pillai. Probability, Random Variables, and Stochastic Processes. 4th Edition. 2002. page 89
URL:
http://ce.sharif.edu/courses/97-98/1/ce181-1/resources/root/Text_Books_References/Papoulis_Pillai_Probability_RandomVariables_and_Stochastic_Processes-4th_Edition_2002.pdf
50. Міністерство захисту довкілля та природних ресурсів України. Як змінюється клімат в Україні. 2020
URL: https://mepr.gov.ua/news/35246.html
51. Історія та майбутнє інтернету речей. 2019.
URL: https://www.itransition.com/blog/iot-history/
52. IBM. SPSS Statistics. Нелинейная регрессия
URL:
https://www.ibm.com/docs/ru/spss-statistics/SaaS?topic=regression-nonlinear
53. Виды нелинейных моделей. Линеаризация моделей
URL:
https://thelib.info/matematika/884252-vidy-nelinejnyh-modelej-linearizaciya-modelej/
54. Померанцев Алексей Леонидович. Методы нелинейного регрессионного анализа для моделирования кинетики химических и физических процессов. Институт химической физики им Н.Н. Семенова, Москва, 2003. С. 21-26

URL: https://chph.chemometrics.ru/papers/thesis_alp.pdf

55. Kaggle. Dataset Rain in Australia - Predict next-day rain in Australia

URL: https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

ДОДАТОК А

Таблиця 3.1 — Таблиця з коефіцієнтами кореляції

Назва параметру	t(min)	t(max)	r(n)	e(n)	s(n)	w_g(dir)	w_g(sp)	w(dir 9am)	w(dir 3pm)
t(min)	1	0.77	0.11	0.42	0.16	False	0.17	False	False
t(max)	0.77	1	-0.07	0.46	0.31	False	0.12	False	False
r(n)	0.11	-0.77	1	-0.06	-0.12	False	0.11	False	False
e(n)	0.42	0.46	-0.06	1	0.71	False	0.19	False	False
s(n)	0.17	0.31	-0.12	0.71	0.71	False	0.19	False	False
w_g(dir)	False	False	False	False	False	False	False	False	False
w_g(sp)	0.17	0.12	0.11	0.19	0.08	False	1	False	False
w(dir9am)	False	False	False	False	False	False	False	False	False
w(dir3pm)	False	False	False	False	False	False	False	False	False
w(sp9am)	0.19	0.02	0.08	0.19	0.12	False	0.66	False	False
w(sp3pm)	0.19	0.07	0.05	0.15	0.16	False	0.78	False	False
h(9am)	-0.16	-0.4	0.22	-0.39	-0.32	False	-0.2	False	False
h(3pm)	0.02	-0.45	0.24	-0.31	-0.32	False	0	False	False
p(9am)	0.49	0.47	0	0.78	1	False	0.7	False	False
p(3pm)	0.48	0.46	0	0.78	1	False	0.71	False	False
c(9am)	0.2	-0.11	0.17	0.21	0.09	False	0.01	False	False
c(3pm)	0.15	-0.09	0.15	0.23	0.12	False	0.09	False	False
t(9am)	0.95	0.94	0.01	0.46	0.26	False	0.16	False	False
t(3pm)	0.72	1	-0.08	0.45	0.37	False	0.16	False	False
r(tod)	False	False	False	False	False	False	False	False	False
r(tom)	False	False	False	False	False	False	False	False	False

Продовження таблиці 3.1

Назва параметру	p(9am)	p(3pm)	c(9am)	c(3pm)	t(9am)	t(3pm)	r(tod)	r(tom)
t(min)	0.48	0.48	0.19	0.15	0.94	0.71	False	False
t(max)	0.47	0.46	-0.11	-0.1	0.94	1	False	False
r(n)	0	0	0.18	0.15	0.01	-0.08	False	False
e(n)	0.78	0.78	0.21	0.23	0.46	0.45	False	False
s(n)	1	1	0.09	0.12	0.26	0.37	False	False
w_g(dir)	False	False	False	False	False	False	False	False
w_g(sp)	0.7	0.71	0.01	0.09	0.16	0.16	False	False
w(dir9am)	False	False	False	False	False	False	False	False
w(dir3pm)	False	False	False	False	False	False	False	False
w(sp9am)	0.42	0.41	0.12	0.14	0.14	0.02	False	False
w(sp3pm)	0.64	0.67	0.1	0.13	0.18	0.12	False	False
h(9am)	0.16	0.12	0.32	0.22	-0.34	-0.39	False	False
h(3pm)	0.16	0.2	0.35	0.41	-0.19	-0.36	False	False
p(9am)	1	1	0.63	0.7	0.53	0.6	False	False
p(3pm)	1	1	0.63	0.71	0.5	0.63	False	False
c(9am)	0.63	0.63	1	1	0.03	-0.13	False	False
c(3pm)	0.7	0.71	1	1	0.04	-0.06	False	False
t(9am)	0.53	0.5	0.03	0.04	1	0.89	False	False
t(3pm)	0.6	0.63	-0.13	-0.06	0.89	1	False	False
r(tod)	False	False	False	False	False	False	False	False
r(tom)	False	False	False	False	False	False	False	False

Таблиця 3.2 — Таблиця з парами коефіцієнтів кореляції, більшими за 0.7

Показник №1	Показник №2	Кореляція

t(min)	t(9am)	0.95
t(min)	t(max)	0.77
t(min)	t(3pm)	0.72
t(max)	t(3pm)	1
t(max)	t(9am)	0.94
e(n)	p(9am)	0.79
e(n)	p(3pm)	0.79
e(n)	s(n)	0.71
s(n)	p(9am)	1
s(n)	p(3pm)	1
w_g(sp)	w(sp3pm)	0.79
w_g(sp)	p(9am)	0.71
w_g(sp)	p(3pm)	0.71
h(9am)	h(3pm)	0.75
p(9am)	p(3pm)	1
p(9am)	c(3pm)	0.71
p(3pm)	c(3pm)	0.72
c(9am)	c(3pm)	1
t(9am)	t(3pm)	0.89

ДОДАТОК Б

Лістинг коду 3.1:

Ініціалізація та заповнення масивів даними:

```
dates = []
values1 = []
values2 = []
i = -1
for loc in datasetWeather[datasetWeather.keys()[1]]:
    i = i+1
dates.append(datetime.strptime(datasetWeather[datasetWeather.keys()[0]][
i], '%Y-%m-%d'))
    values1.append(datasetWeather[column1][i])
    values2.append(datasetWeather[column2][i])

forecastValues = []
period = 10
for i in range(period):
    forecastValues.append(None)
```

Здійснення прогнозу:

```
for v in range(len(values1)):
    if v < period:
        continue
    meanForLastDays = 0
    for i in range(period):
        if is_number(values2[v-i-1]) and is_number(values1[v-i-1]):
            meanForLastDays += values2[v-i-1]-values1[v-i-1]
    meanForLastDays = meanForLastDays/period

    forV = values1[v]+meanForLastDays
    forecastValues.append(forV)
```

Підрахунок похибки:

```
mean_error = 0
i = 0
for value in values2:
    f = forecastValues[i]
    if is_number(f) and is_number(value):
```

```
        error = abs(f - value)
        mean_error = mean_error + error
    i = i + 1
mean_error = mean_error/len(values2)
```

Лістинг коду 3.2:

```
minY = 0
maxY = 0
sumY = 0
for j in y:
    if(j < minY):
        minY = j
    if(j > maxY):
        maxY = j
    if is_number(j):
        sumY = sumY + j
meanY = sumY/len(y)
variance = 0
for j in y:
    if is_number(j):
        variance = variance + (j-meanY)**2
variance = math.sqrt(variance/len(y))
interval = 3.49*variance/(len(y)**(1/3))
```

Лістинг коду 3.3:

Ініціалізація та заповнення масивів даними:

```
classesX = ['N', 'NNE', 'NE', 'ENE', 'E', 'ESE', 'SE', 'SSE', 'S', 'SSW', 'SW',
'WSW', 'W', 'WNW', 'NW', 'NNW']
intervals = []
i = 0
while 1:
    a = minY+(i*interval)
    b = a+interval
    if b >= maxY:
        b = maxY
    intervals.append([a, b])
    if b >= maxY:
        break
    i = i + 1
```

```
arrayValues = {}
for i in classesX:
    arrayValues[i] = []
    for j in range(len(intervals)):
        arrayValues[i].append(0)
```

Аналіз частоти розподілу даних у інтервали:

```
m = 0
for j in x:
    if(pd.isna(j) == 0 and pd.isna(y[m]) == 0):
        l = 0
        for i in intervals:
            if(l == len(intervals)-1):
                if(y[m] >= i[0]):
                    break
            if(y[m] >= i[0] and y[m] < i[1]):
                break
            l = l+1
        arrayValues[j][l] = arrayValues[j][l]+1
    m = m+1
```

Підготовка масивів з діапазонами та частотою попадання даних у них:

```
indexes = []
indexes2 = []
for k in intervals:
    digit1 = int_r(k[0]*100)/100
    digit2 = int_r(k[1]*100)/100
    indexes.append("[" + str(digit1) + ", " + str(digit2) + ")")
    indexes2.append(str(digit1) + "-" + str(digit2))

probabilitiesOld = {}
probabilities = {}
for index in indexes2:
    probabilitiesOld[index] = {}
    probabilities[index] = {}

sumsOfDiapazones = 0

for obj in arrayValues:
    for val in arrayValues[obj]:
        sumsOfDiapazones = sumsOfDiapazones + val
for obj in arrayValues:
```

```
    i = 0
    for val in arrayValues[obj]:
        probabilitiesOld[indexes2[i]][str((float(val)/sumsOfDiapazones))] = obj
        i = i+1

arrayProp = []
for p in probabilitiesOld:
    for pp in probabilitiesOld[p]:
        arrayProp.append(float(pp))
arrayProp.sort()

sumP = 0
for p in arrayProp:

    for po in probabilitiesOld:
        if str(p) in probabilitiesOld[po]:
                        probabilities[po][str(sumP) + '-' + str(sumP+p)] =
probabilitiesOld[po][str(p)]
            sumP = sumP + p

intervalsOriginal = []
intervals = []
for p in probabilities:
    intervalsOriginal.append(p.split('-'))
for interv in intervalsOriginal:
    intervals_arr = []
    for inter in interv:
        intervals_arr.append(inter)
    intervals.append(intervals_arr)
```

Реалізація прогнозу напрямку вітру та підрахунок похибки:

```
precisionOfForecast_sum = 0
i = 0
for valY in y:
    for interval in intervals:
        if valY > float(interval[0]) and valY <= float(interval[1]):
            arr = probabilities[str(interval[0])+'-'+str(interval[1])]
            arrOriginal = []
            for a in arr:
                arrOriginal.append(a.split('-'))
            forecastValue = ''
            while forecastValue == '':
                randomNumber = random.uniform(0, 1)
```

```
            for a in arrOriginal:
                        if randomNumber > float(a[0]) and randomNumber <=
float(a[1]):
                    forecastValue = arr[a[0]+'-'+a[1]]
          if forecastValue == x[i]:
             precisionOfForecast_sum = precisionOfForecast_sum + 1
   i = i +1
```

Лістинг коду 3.4:

Ініціалізація та заповнення масивів даними:

```
fig = plt.figure()
polar_ax = fig.add_subplot(1, 1, 1, projection="polar")

polar_ax.bar(0, 1, 0.4, 0, align='edge', color='#a0f', alpha=0.5)
polar_ax.bar(0.4, 1, 0.4, 0, align='edge', color='#a9f', alpha=0.5)

dirs = ['E', 'ENE', 'NE', 'NNE', 'N', 'NNW', 'NW', 'WNW', 'W', 'WSW', 'SW',
'SSW', 'S', 'SSE', 'SE', 'ESE']
# Fiddle with labels and limits
polar_ax.set_xticks([0, np.pi/8, 2*np.pi/8, 3*np.pi/8, np.pi/2, 5*np.pi/8,
3*np.pi/4, 7*np.pi/8, np.pi, 9*np.pi/8, 5*np.pi/4, 11*np.pi/8, (3*np.pi)/2,
13*np.pi/8, 14*np.pi/8, 15*np.pi/8])
polar_ax.set_xticklabels(dirs)

valuesFrequency = []
winterdata = {}
for i in dirs:
   winterdata[i] = {}
```

Поділ масивів за класами та підрахунок частоти розподілу даних:

```
i = 0
for j in datasetWeather[datasetWeather.keys()[11]]:
   dir = datasetWeather[datasetWeather.keys()[9]][i]
   if(pd.isna(j) == 0 and pd.isna(dir) == 0):
      speed = math.ceil(j)

      if speed in winterdata[dir]:
         winterdata[dir][speed] = winterdata[dir][speed]+1
      else:
         winterdata[dir][speed] = 1
```

```
    i = i+1
```

Візуалізація даних:

```
colors = []
for color in range(256):
    colors.append('#%02x%02x%02x' % (0, color, 255))
for color in range(254, -1, -1):
    colors.append('#%02x%02x%02x' % (0, 255, color))
for color in range(1, 256, 1):
    colors.append('#%02x%02x%02x' % (color, 255, 0))
for color in range(254, -1, -1):
    colors.append('#%02x%02x%02x' % (255, color, 0))

for i in dirs:
    for j in winterdata[i].keys():
        valuesFrequency.append(winterdata[i][j])
maxFrequency = max(valuesFrequency)

color_step = len(colors)/maxFrequency

for i in dirs:
    for j in winterdata[i].keys():
        indexDir = 0
        k = 0
        for dirr in dirs:
            if dirr == i:
                indexDir = k
            k = k+1
        angle = indexDir*0.3925
        color_index = int_r(winterdata[i][j]*color_step)-1
                    polar_ax.bar(angle,  1,  0.3925,  j,  align='edge',
color=colors[color_index], alpha=0.5)

polar_ax.set_rticks([0, 30, 60, 90, 120, 135])
polar_ax.set_rlabel_position(12.5)
polar_ax.grid(True)

x = np.linspace(0, 5, 100)
N = maxFrequency
cmap = plt.get_cmap('jet',N)

norm = mpl.colors.Normalize(vmin=0,vmax=N)
```

```
sm = plt.cm.ScalarMappable(cmap=cmap, norm=norm)
clb = plt.colorbar(sm, ticks=np.linspace(0,N,16),
            boundaries=np.arange(-14,N,1))
clb.ax.set_title('Частота')

plt.show()
```

Лістинг коду 3.5:

Ініціалізація та заповнення масивів даними:

```
datasetWeather['dates'] =
matplotlib.dates.date2num(datasetWeather[datasetWeather.keys()[0]].tolist(
))
datasetWeatherLocale = datasetWeather.sort_values('dates')

dates = []
values = []

i = 0
for d in datasetWeatherLocale[datasetWeather.keys()[0]]:
    if is_number(datasetWeatherLocale[datasetWeather.keys()[19]][i]):
        dates.append(d)
        values.append(datasetWeatherLocale[datasetWeather.keys()[19]][i])
    i = i + 1

dates = matplotlib.dates.date2num(dates)
indexes = list(range(1, len(values)+1))

indexesX2 = []
indexesXvalues = []
```

Підрахунок параметрів лінійної регресії за методом найменших квадратів:

```
sumValues = sum(values)
sumIndexes = sum(indexes)
sumIndexesX2 = sum(indexesX2)
sumIndexesXvalues = sum(indexesXvalues)

M1 = np.array([[sumIndexesX2, sumIndexes], [sumIndexes, len(values)]])
v1 = np.array([sumIndexesXvalues, sumValues])
result = np.linalg.solve(M1, v1).tolist()

def regressionFunction(x):
```

```
    return result[0]*x+result[1]

parameters = [[dates[dates.tolist().index(min(dates))], dates[len(dates)-1]],
[regressionFunction(0), regressionFunction(len(dates)-1)]]
```

Візуалізація лінійної регресії:

```
plt.plot_date(dates, values, ms=2)
plt.plot(parameters[0], parameters[1], 'k-')
plt.xlabel("Time")
plt.ylabel("Temperature at 9 am")
plt.show()
```

ДЕКЛАРАЦІЯ АКАДЕМІЧНОЇ ДОБРОЧЕСНОСТІ

Сугак Глєб Васильович
Прізвище, ім’я, по батькові

Факультет інформаційних і прикладних технологій
Факультет

122 «Комп'ютерні науки»
Шифр і назва спеціальності

Комп'ютерні технології обробки даних (Data Science)
Освітня програма

ДЕКЛАРАЦІЯ АКАДЕМІЧНОЇ ДОБРОЧЕСНОСТІ

Усвідомлюючи свою відповідальність за надання неправдивої інформації, стверджую, що подана кваліфікаційна (магістерська) робота на тему: «Інтелектуальний аналіз метеорологічних даних для дослідження погоди» є написаною мною особисто.

Одночасно заявляю, що ця робота:

- не передавалась іншим особам і подається до захисту вперше;
- не порушує авторських та суміжних прав, закріплених статтями 21-25 Закону України «Про авторське право та суміжні права»;
- не отримувались іншими особами, а також дані та інформація не отримувались у недозволений спосіб.

Я усвідомлюю, що у разі порушення цього порядку моя кваліфікаційна (магістерська) робота буде відхилена без права її захисту, або під час захисту за неї буде поставлена оцінка «незадовільно».

дата	підпис