МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Харківський національний університет імені В.Н.Каразіна

Факультет математики і інформатики

Кафедра теоретичної та прикладної інформатики

Кваліфікаційна робота

магістр

на тему

«Виявлення статистичних закономірностей між захворюваністю та геоданими обраної місцевості»

Виконав: студент 2 курсу, другого (магістерського) рівня вищої освіти, групи МФ-61
спеціальність 122 Комп’ютерні науки
освітньо-наукова програма «Інформатика»
Шмідов Ярослав Андрійович

Керівник: Меняйлов Євген Сергійович

Рецензент Базілевич Ксенія Олексіївна

Харків – 2023

ЗМІСТ

1. ВСТУП

1.1. Формулювання мети роботи, задач та обґрунтування актуальності теми

У сучасному світі сфера контролю та спостереження за захворюваннями, динамікою їх розвитку та розповсюдження надалі набуває актуальності. Саме тому основна мета цієї роботи - глибше дослідити потенційний зв'язок між захворюваністю на лептоспіроз та геопросторовими даними в Європейському Союзі шляхом вивчення різних кліматичних факторів. Лептоспіроз - це глобальне зоонозне захворювання, яке спричиняє понад 500000 випадків на рік [1]. Захворювання викликається патогенними бактеріями-спірохетами, відомими як *Leptospira interrogans* [2]. Широкий спектр диких і домашніх тварин є потенційними резервуарами для передачі інфекції людині. Зазвичай інфікування людини відбувається через прямий контакт з побічними продуктами життєдіяльності інфікованих тварин, такими як сеча, або опосередковано через контакт із забрудненими *Leptospira* середовищами, зокрема, з джерелами води. Це дослідження має на меті не лише виявити статистично значущі закономірності, тенденції та кореляції між цими змінними, але й дослідити, як різні фактори можуть взаємодіяти та впливати один на одного. Зрештою, цей комплексний аналіз має на меті забезпечити краще розуміння складної динаміки, що лежить в основі поширення лептоспірозу, та може сприяти розробці науково обґрунтованої політики та стратегій втручання у сфері громадського здоров'я.

Лептоспіроз є складним і багатогранним захворюванням, виникненню і поширенню якого сприяють різні фактори. Незважаючи на зростаючу кількість літератури на цю тему, все ще бракує комплексних та

інтегративних досліджень, які б одночасно розглядали різні аспекти захворювання, особливо в контексті Європейського Союзу.

Комплексний підхід цього дослідження, пропонує більш тонке та всебічне розуміння динаміки, що лежить в основі захворюваності на лептоспіроз. Виявивши статистичні закономірності та потенційні взаємодії між цими факторами, дослідження може пролити світло на раніше ігноровані взаємозв'язки та сприяти більш ефективній та цілеспрямованій боротьбі з хворобою.

Крім того, геопросторовий фокус цього дослідження може допомогти заповнити існуючі прогалини в знаннях у сфері просторової епідеміології та контролю за інфекційними захворюваннями. Завдяки застосуванню методів геопросторової візуалізації та використанню геоданих, дослідження може надати більш чітку картину просторового поширення лептоспірозу та сприяти виявленню зон підвищеного ризику. Зрештою, отримані результати можуть бути використані для формування політики громадського здоров'я, розподілу ресурсів і цільових стратегій втручання, спрямованих на пом'якшення впливу лептоспірозу і поліпшення результатів громадського здоров'я в Європейському Союзі.

1.2. Стислий огляд відомих результатів

У результаті аналізу існуючих рішень, був отриманий висновок, що конкретна увага до захворюваності на лептоспіроз в Європейському Союзі є обмеженою, проте кілька досліджень вивчали взаємозв'язок між екологічними, кліматичними та демографічними факторами і поширенням лептоспірозу в різних регіонах світу. Тут ми надаємо короткий огляд відомих результатів на цю тему.

1. Екологічні фактори:
   - Землекористування: Сільськогосподарська діяльність, особливо пов'язана з розведенням худоби та гризунів, була визначена як фактор ризику лептоспірозу. Наявність відповідних середовищ існування для тварин, таких як гризуни, може підвищити ймовірність передачі хвороби людині.
   - Рослинний покрив: Густий рослинний покрив асоціюється з підвищеним ризиком лептоспірозу, оскільки він може створювати сприятливе середовище для гризунів та інших резервуарів інфекції.
   - Близькість до водойм: Передача лептоспірозу є сприятливою на територіях з великою кількістю води, таких як затоплені території, річки та болота, оскільки бактерії можуть довше виживати у вологому середовищі.
2. Кліматичні фактори:
   - Температура: Вищі температури пов'язані зі зростанням захворюваності на лептоспіроз, оскільки бактерії можуть процвітати в теплішому середовищі. Однак надзвичайно високі температури можуть мати протилежний ефект, оскільки вони можуть обмежувати виживання бактерій.
   - Опади: Сильні дощі та повені пов'язані з підвищеним ризиком лептоспірозу, оскільки ці умови сприяють поширенню бактерій у навколишньому середовищі та збільшують ризик контакту людини із забрудненими джерелами води.
   - Вологість: підвищений рівень вологості може сприяти виживанню бактерій у навколишньому середовищі, що потенційно збільшує ризик передачі лептоспірозу.

3. Демографічні та соціально-економічні фактори:
   - щільність населення: висока щільність населення може сприяти підвищенню передачі лептоспірозу;
   - урбанізація: швидка урбанізація може призвести до збільшення контакту людини із забрудненим середовищем, тим самим підвищуючи ризик захворювання на лептоспіроз;
   - доступ до охорони здоров'я: Обмежений доступ до медичної допомоги та низький соціально-економічний статус можуть сприяти підвищенню рівня захворюваності та смертності від лептоспірозу, оскільки постраждалі особи можуть не отримати своєчасну діагностику та лікування.

Проте важливо зазначити, що взаємозв'язок між цими факторами та захворюваністю на лептоспіроз може змінюватися залежно від місцевих умов та конкретних штамів бактерій. Більше того, ефекти взаємодії між різними факторами також можуть відігравати вирішальну роль у формуванні просторового розподілу захворювання. Подальші дослідження, зосереджені на контексті Європейського Союзу, можуть допомогти з'ясувати конкретні фактори та взаємодії, що сприяють виникненню та поширенню лептоспірозу в регіоні.

1.3. Відомості про одержані результати та їх новизна

Це дослідження одне з перших в Україні намагається дослідити зв’язок між геоданими та захворюваністю на лептоспіроз:

- використано формат геоданих GeoTIFF;
- створено унікальний набір даних про захворюваність на лептоспіроз;
- виявлено залежність між ростом захворюваності та кліматичними геоданими;

– отримано прогнози захворюваності на лептоспіроз з урахуванням всіх використаних наборів даних.

2. ОСНОВНА ЧАСТИНА

2.1. Постанова задачі

Головна задача дипломної роботи - це виявлення та аналіз статистично значущих закономірностей, тенденцій та кореляцій між захворюваністю на лептоспіроз та кліматичними геопросторовими даними в межах Європейського Союзу з 2005 по 2020 рік. Зосереджуючись на взаємозв'язку між кліматичними факторами, такими як температура, опади і вологість, ваше дослідження має на меті сприяти кращому розумінню того, як ці фактори впливають на виникнення і поширення лептоспірозу в регіоні, що в кінцевому підсумку сприятиме формуванню політики громадського здоров'я та стратегій втручання.

В рамках даної роботи стоять такі задачі:

- провести комплексний огляд літературних джерел про захворюваність на лептоспіроз, епідеміологічний процес та відібрати пов'язані з ним геопросторові дані в Європейському Союзі;
- отримати та попередньо обробити кліматичні геопросторові дані та дані про захворюваність на лептоспіроз в Європейському Союзі з відповідних джерел;
- розробити та впровадити відповідні статистичні моделі та методи для виявлення, аналізу та прогнозування закономірностей, тенденцій та кореляцій між захворюваністю на лептоспіроз та кліматичними геопросторовими даними;
- оцінити достовірність та надійність отриманих результатів.
- інтерпретувати та контекстуалізувати результати;

– сформулювати рекомендації щодо політики громадського здоров'я, стратегій втручання та майбутніх досліджень у світлі отриманих результатів.

2.2. Розвинутий огляд сучасного стану справ в області

На сьогоднішній момент існує велика кількість досліджень та звітів, що розповідають про розповсюдження лептоспірозу в світі, зокрема і в Європейському Союзі. Серед світових джерел інформації було розглянуто такий перелік:

1. https://www.cdc.gov/ - Center for Disease Control and Prevention
2. https://www.ecdc.europa.eu/ - European Centre for Disease Prevention and Control
3. https://www.who.int/ - World Health Organization
4. https://www.ncbi.nlm.nih.gov/ - National Library of Medicine

Всі вище наведені джерела містять звітність про захворюваність на лептоспіроз в тій, чи іншій частині світу. Здебільшого ці публікації носять декларативний характер та не містять в собі аналізу.

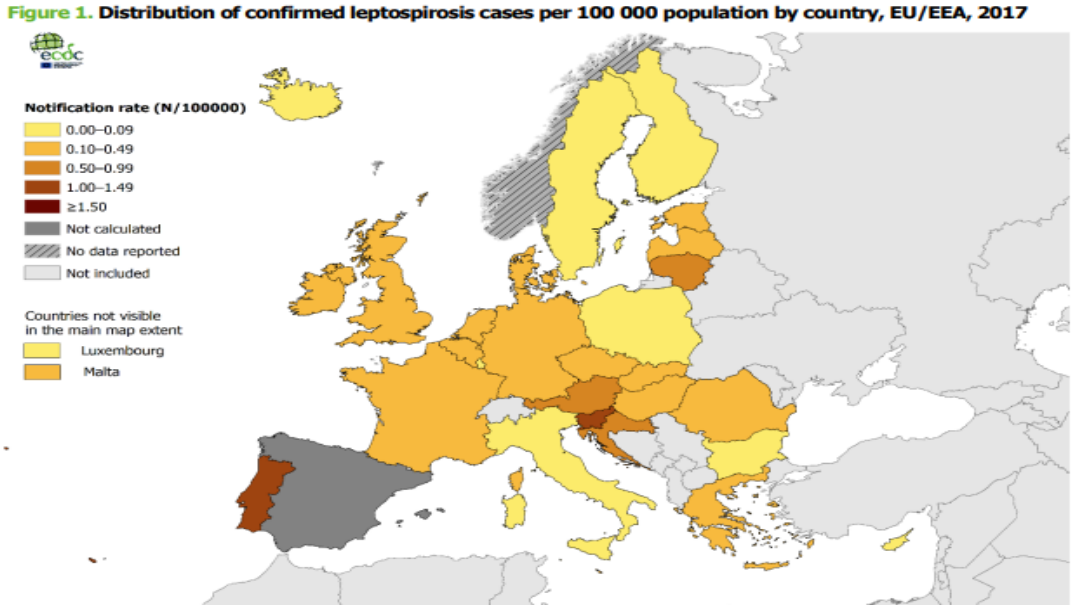

Рис. 1 – Приклад звіту з https://www.ecdc.europa.eu/

Також варто відзначити декілька досліджень, що базуються на порівняльному аналізі статистично значущих закономірностей, тенденцій та кореляцій між захворюваністю на лептоспіроз та кліматичними геопросторовими даними[3][4].

Саме тому для визначення актуальності дослідження було проведено поглиблений аналіз, в результаті чого була складена таблиця 1.

Таблиця 1 - Відмінності між існуючими дослідженнями та проведеним дослідженням

Існуючі дослідження	**Проведене дослідження**
Регіон дослідження обмежений однією країною	Регіон дослідження включає в себе всі країни Європейського Союзу
Проміжок часу, що охоплює дослідження від 1 до 4 років	Проміжок часу, що охоплює дослідження - 16 років
Метою досліджень здебільшого є прогнозування захворюваності	Метою досліджень є не тільки прогнозування захворюваності, а й пошук потенційного зв'язку між захворюваністю на лептоспіроз та геопросторовими даними
Використано один датасет геоданих	Використано декілька датасетів геоданих

У висновку до цієї частини можна сказати, що після проведеного огляду, не було знайдено жодного дослідження, що має аналогічний зміст.

2.3. Матеріали та методи дослідження

2.3.1. Джерела даних

Для роботи над дослідженням були обрані такі дані:

Набір даних, що представляє щорічну кількість випадків лептоспірозу, зареєстрованих у країнах Європейського Союзу з 2005 по 2020 рік. Дані включають в себе показники захворюваності на лептоспіроз з країн Європейського Союзу та країн, що входять в Європейську асоціацію вільної торгівлі, але країни з нульовими показниками захворюваності за всю історію спостережень були виключені з набору даних. Дані надають історичну перспективу виникнення лептоспірозу в різних країнах, що може бути корисним для розуміння тенденцій, виявлення зон підвищеного ризику та аналізу впливу кліматичних геопросторових факторів на поширення хвороби. Потрібно зазначити, що цей датасет був зібраний з різних джерел, немає єдиного походження і є унікальним для нашого дослідження[5][6][7][8][9][10][11][12][13][14][15][16].

В якості геопросторових даних були обрані GeoTIFF датасети. GeoTIFF - це загальнодоступний стандарт метаданих, який дозволяє вбудовувати інформацію про географічну прив'язку у файл TIFF. Потенційна додаткова інформація включає картографічну проекцію, системи координат, еліпсоїди, точки відліку та все інше, необхідне для встановлення точної просторової прив'язки файлу. Були використані такі датасети:

ERA5-Land daily: Air temperature at 2 meter above surface (2000 - 2020) (Температура повітря на висоті 2 метри над поверхнею)[17]. Це набір даних повторного аналізу, що надає послідовне уявлення про еволюцію наземних змінних протягом кількох десятиліть з підвищеною

роздільною здатністю. ERA5-Land був створений шляхом повторного аналізу наземного кліматичного компоненту ECMWF(Європейський центр середньострокових прогнозів погоди) ERA5. Це аналіз поєднує модельні дані зі спостереженнями з усього світу в глобально повний і узгоджений набір даних з використанням законів фізики. Повторний аналіз дає дані, які сягають на кількох десятиліть назад у часі, забезпечуючи точний опис температури минулого. Доступні середньодобові, мінімальні та максимальні значення температури повітря (2 м). Всі значення масштабовані до цілих чисел.

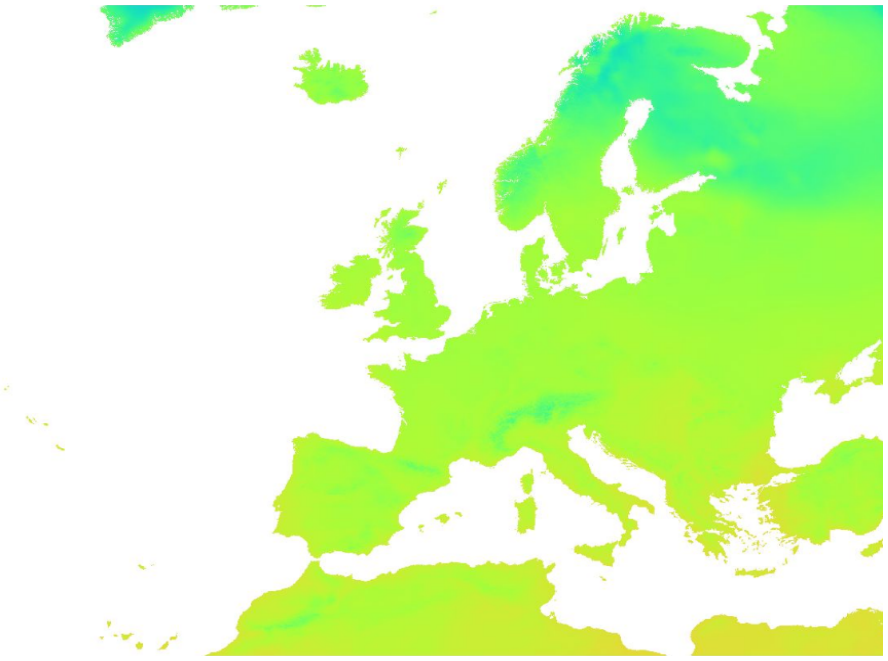

Рис. 2 – ERA5-Land daily: Air temperature at 2 meter above surface (2000 - 2020).

Monthly time series of spatially enhanced relative humidity for Europe at 1000 m resolution (2000 - 2021) derived from ERA5-Land data[18]. Це набір даних, що включає в себе щомісячні часові ряди просторово розширеної відносної вологості для Європи з роздільною здатністю 1000 м (2000 - 2021), отримані з даних ERA5-Land. Всі значення масштабовані до цілих чисел.

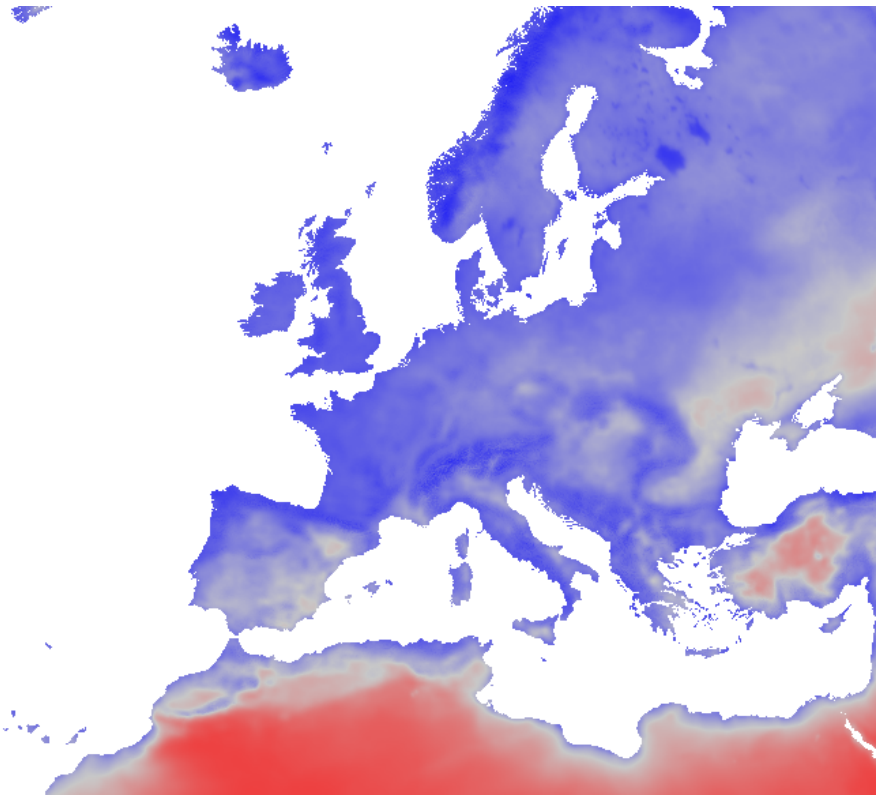

Рис. 3 – Monthly time series of spatially enhanced relative humidity for Europe at 1000 m resolution (2000 - 2021) derived from ERA5-Land data.

Global SnowPack — MODIS[19][20]. Цей набір даних показує середню тривалість залягання снігового покриву починаючи з 2001 гідрологічного року. Гідрологічний рік починається метеорологічною осінню (1 жовтня попереднього року в північній півкулі або 1 березня базового року в південній півкулі) і закінчується метеорологічним літом (північна півкуля: 31 серпня базового року; південна півкуля: 28/29 лютого поточного року. Тут враховується весь рік, а також ранній сезон (до середини зими) та пізній сезон (з середини зими). "Global SnowPack" отримано на основі щоденних оперативних даних MODIS(Візуалізаційний спектрорадіометр з помірною роздільною здатністю) про сніговий покрив за кожен день, починаючи з лютого 2000 року. Прогалини в даних, пов'язані з полярною ніччю та хмарністю, заповнюються в декілька етапів

обробки, що забезпечує унікальний глобальний набір даних, який характеризується високою точністю, просторовою роздільною здатністю 500 метрів та постійним розширенням у майбутньому.

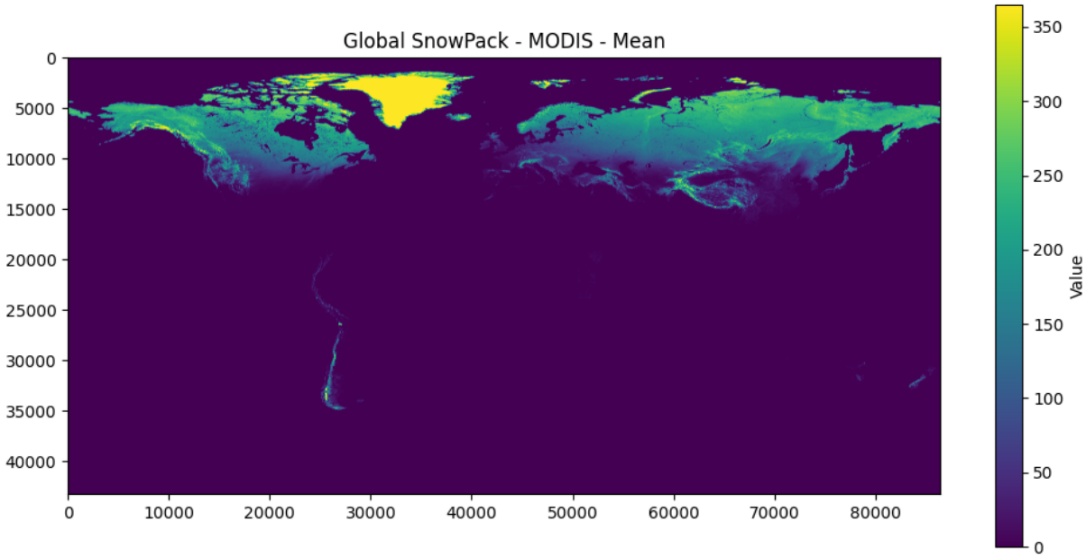

Рис. 4 – Global SnowPack — MODIS

Через те, що GeoTIFF датасети містили більший об’єм даних ніж було потрібно для дослідження, з них були отримані та використані тільки ті дані, що підпадають під необхідні географічні і часові межі. На основі обраних наборів даних було створено єдиний датасет.

2.3.2. Методи аналізу та прогнозування

Для вивчення закономірностей і тенденцій захворюваності на лептоспіроз було використано поєднання описової статистики, візуального аналізу та кореляційного аналізу.

Описова статистика: Були обчислені зведені статистичні показники для даних про захворюваність, такі як середнє значення, медіана, стандартне відхилення та діапазон. Ці статистичні дані допомагають зрозуміти основну тенденцію, дисперсію та розподіл даних.

Візуальний аналіз: Були створені графіки часових рядів та бокс-графіки для візуалізації даних про захворюваність на лептоспіроз у часі. Графіки часових рядів дозволяють спостерігати тенденції, сезонність і нерівномірність даних, тоді як секторні діаграми дають уявлення про розподіл і потенційні відхилення від норми.

Кореляційний аналіз: Були досліджені потенційні рушійні сили або фактори, які можуть впливати на структуру захворюваності, вивчаючи взаємозв'язок між даними про захворюваність на лептоспіроз та іншими змінними, такими як температура та опади. Були обчислені коефіцієнти кореляції Пірсона, Спірмена та Кендала.

Коефіцієнт кореляції Пірсона: Вимірює лінійний зв'язок між двома безперервними змінними. Він варіюється від -1 (ідеальний негативний лінійний зв'язок) до 1 (ідеальний позитивний лінійний зв'язок). Значення 0 вказує на відсутність лінійного зв'язку. Кореляція Пірсона чутлива до викидів і передбачає нормальний розподіл даних. Коефіцієнт кореляції Пірсона розраховують за формулою:

$$r_{xy} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m} (x_i - \bar{x})^2 \sum_{i=1}^{m} (y_i - \bar{y})^2}} = \frac{cov(x,y)}{\sqrt{s_x^2 s_y^2}},$$

де - $\bar{x}$, $\bar{y}$ вибіркові середні, x^m, y^m, s^2_x, s^2_y — вибіркові дисперсії, $r_{xy} \in [-1,1]$.

Коефіцієнт рангової кореляції Спірмена: Вимірює монотонний зв'язок між двома змінними на основі їхніх рангів. Це непараметричний тест, тобто він не передбачає нормального розподілу даних. Рангова кореляція Спірмена менш чутлива до викидів порівняно з кореляцією Пірсона. Коефіцієнт кореляції Спірмена розраховують за формулою(для вибірки обсягу n множини X_i, Y_i перетворюються в ряди x_i, y_i):

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

Коефіцієнт рангової кореляції Кендала: Також вимірює монотонний зв'язок між двома змінними на основі їхніх рангів. Це ще один непараметричний тест, менш чутливий до викидів. Тау Кендалла є обчислювально більш ефективним для великих наборів даних, але зазвичай має нижчу статистичну потужність порівняно з ранговою кореляцією Спірмена. Коефіцієнт кореляції Кендала розраховують за формулою:

$$\tau = \frac{s_1 - s_2}{\frac{1}{2} n(n-1)},$$

де s_1 - кількість узгоджених пар, s_2 - кількість неузгоджених пар.

Для прогнозування рівня захворюваності на лептоспіроз була використана модель ARIMA (Авторегресійне інтегроване ковзне середнє). Модель ARIMA - це популярний метод прогнозування лінійних часових рядів, який поєднує компоненти авторегресії (AR), диференціювання (I) та ковзного середнього (MA). Модель визначається трьома параметрами (p, d, q), де:

- p - порядок авторегресійного компонента (AR);
- d - ступінь диференціювання (I), що застосовується до часового ряду для того, щоб зробити його стаціонарним;
- q - порядок компоненти ковзного середнього (MA).

Для визначення оптимальних значень параметрів ARIMA моделі був поведений сітковий пошук з діапазоном потенційних значень для p, d та q. Були оцінені моделі за допомогою інформаційного критерію Акаіке (AIC)

або байєсівського інформаційного критерію (BIC), підбираючи комбінацію параметрів, яка мінімізує обраний критерій.

Після підбору оптимальних параметрів ARIMA-моделі були згенеровані прогнози щодо майбутніх рівнів захворюваності на лептоспіроз.

Поєднання методів аналізу даних з ARIMA моделлю, було спричинено прагненням виявити фактори, що впливають на рівень захворюваності на лептоспіроз, і створити прогнози щодо майбутніх тенденцій захворюваності.

2.3.3. Використані технології та обґрунтування вибраного інструментарію

2.3.3.1. Rasterio

Rasterio - це популярна бібліотека Python, яка надає простий і зрозумілий API для роботи з геопросторовими растровими даними. Бібліотека побудована на основі широко використовуваної GDAL (Бібліотека абстракції геопросторових даних) і пропонує більш зручний інтерфейс для читання, запису та маніпулювання растровими даними. Вона широко використовується в таких додатках, як дистанційне зондування, ГІС і моделювання навколишнього середовища.

Ключові особливості та причини вибору Rasterio:

– Читання та запис растрових даних: Rasterio може читати і записувати широкий спектр растрових форматів, включаючи такі популярні, як GeoTIFF, JPEG і PNG;

- Система координат і прив'язка до місцевості: Rasterio підтримує обробку та перетворення растрових даних між різними системами координат (CRS);
- Растрові обчислення: Бібліотека пропонує зручний спосіб виконання растрових обчислень, таких як арифметичні операції, маскування та повторна вибірка;
- Віконне читання і запис: Ця функція дозволяє читати і записувати лише частину растрового набору даних, що корисно при роботі з великими файлами або коли обробку потрібно виконувати частинами;
- Інтеграція з NumPy: Rasterio використовує масиви NumPy для представлення растрових даних, що дозволяє легко інтегруватися з іншими науковими бібліотеками та інструментами Python;
- Деформація: Rasterio надає функції для відтворення та передискретизації растрових даних.

2.3.3.2 Pandas

Pandas - це бібліотека Python з відкритим вихідним кодом, яка надає гнучкі, високопродуктивні структури даних та інструменти для маніпулювання та аналізу даних. Вона побудована на основі бібліотеки NumPy і призначена для того, щоб зробити роботу зі структурованими даними (наприклад, електронними таблицями та таблицями SQL) легкою та ефективною. Pandas широко використовується в науці про дані, аналізі даних і машинному навчанні.

Ключові особливості та причини вибору Pandas:

- Структури даних: Pandas надає дві основні структури даних - Series та DataFrame. Серія являє собою одновимірний маркований масив,

тоді як DataFrame - двовимірну марковану структуру даних зі стовпчиками потенційно різних типів;

- Імпорт/експорт даних: Pandas може читати і записувати дані з різних форматів, включаючи CSV, Excel, JSON, SQL та інші;
- Очищення даних: Pandas надає різні інструменти для очищення та попередньої обробки даних, такі як обробка відсутніх даних, видалення дублікатів та фільтрація рядків або стовпців;
- Маніпулювання даними: Бібліотека пропонує широкий спектр функцій для маніпулювання даними, включаючи злиття, приєднання, зміну форми та обертання даних;
- Аналіз даних: Pandas надає повний набір функцій для описової статистики, агрегації та групових операцій, що робить його придатним для різних завдань аналізу даних;
- Функціональність часових рядів: Pandas має вбудовану підтримку для обробки та аналізу даних часових рядів, таких як повторна вибірка, обробка часових поясів та операції з рухомим вікном.

2.4. Описання та обґрунтування алгоритмів та результатів дослідження

Розпочати роботу було вирішено з описового аналізу даних про захворюваність на лептоспіроз у різних країнах Європи з 2005 по 2020 рік. Дані представлені у двох таблицях: таблиця 2 - зведені статистичні дані та таблиця 3 - додаткова описова статистика.

Таблиця 2 - Зведені статистичні дані

Параметр	**Значення**
Середнє значення	18.61
Медіана	10
Стандартне відхилення	42.10
Мінімальне значення	0
Максимальне значення	479

У таблиці 2 наведено зведені статистичні дані щодо захворюваності на лептоспіроз у всіх європейських країнах і за всіма роками. Середнє значення 18,61 випадків на рік вказує на те, що в середньому в кожній країні щорічно реєструється близько 19 випадків лептоспірозу. Медіанне значення 10 випадків на рік свідчить про те, що половина точок даних є нижчими за це значення, а половина - вищими за нього. Стандартне відхилення 42,10 означає, що точки даних розкидані, причому в деяких країнах спостерігається відносно велика кількість випадків, а в інших - дуже мале або зовсім немає. Мінімальне значення 0 випадків вказує на те, що в деяких країнах не було зареєстровано жодного випадку лептоспірозу за певний рік, тоді як максимальне значення 479 випадків вказує на найбільшу кількість випадків, зареєстрованих за один рік у країні.

Таблиця 3 - Додаткова описова статистика

Параметр	**Значення**
Дисперсія	1772.42
1 квартиль(Q1)	4
2 квартиль(Q2)	32
Міжквартильний розмах (IQR)	28

У таблиці 3 наведено додаткові описові статистичні дані, які дають змогу глибше зрозуміти розподіл і поширення даних. Дисперсія 1772,42 вимірює середнє квадратичне відхилення від середнього значення, що вказує на значний ступінь варіабельності кількості випадків між країнами та роками. Значення 1-го квартиля (Q1), що дорівнює 4 випадкам, свідчить про те, що 25% точок даних мають 4 або менше зареєстрованих випадків. Значення 3-го квартилю (Q3), що дорівнює 32 випадкам, вказує на те, що 75% точок даних мають 32 або менше зареєстрованих випадків. Міжквартильний розмах (IQR) у 28 випадків представляє розкид середніх 50% даних, що свідчить про значну різницю в кількості випадків між країнами.

Аналогічні таблиці були отримані і для інших наборів даних.

Таблиця 4 - Зведені статистичні дані для Global SnowPack - MODIS

Параметр	**Значення**
Середнє значення	22.201873657131664
Медіана	20.730146290491117
Стандартне відхилення	11.751090966201932
Мінімальне значення	1.0444096133751306
Максимальне значення	61.182079414838036

Таблиця 5 - Зведені статистичні дані для Monthly time series of spatially enhanced relative humidity for Europe at 1000 m resolution (2000 - 2021)

Параметр	**Значення**
Середнє значення	581.4423516013762
Медіана	577.426231060606
Стандартне відхилення	61.23274838455749
Мінімальне значення	448.86003787878786
Максимальне значення	752.786268939394

Таблиця 6 - Зведені статистичні дані для ERA5-Land daily: Air temperature at 2 meter above surface (2000 - 2020)

Параметр	**Значення**
Середнє значення	112.95452160139948
Медіана	113.60992366412214
Стандартне відхилення	28.0926381841473
Мінімальне значення	3.5319222761970854
Максимальне значення	180.94788341429563

Наступним кроком є візуалізація наявних даних. На рис. 5 і рис. 6 можна побачити загальні діаграми, що відображають дані про захворюваність на лептоспіроз для кожної країни Європейського союзу у вигляді лінійної діаграми, що ілюструє динаміку захворюваності з часом.

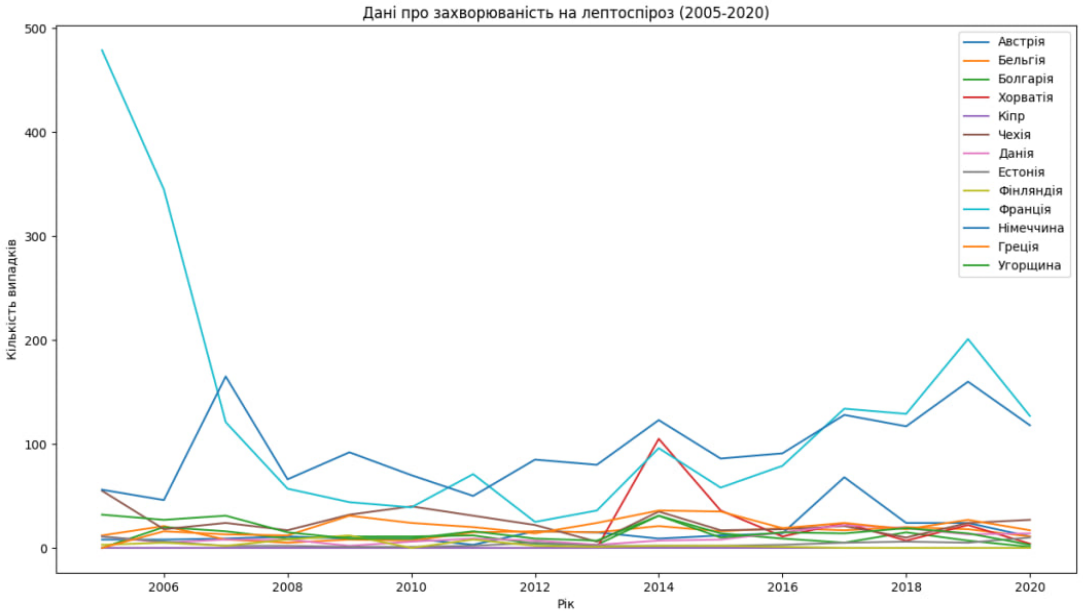

Рис. 5 – Загальна діаграма про захворюваність на лептоспіроз для кожної країни Європейського союзу (Частина 1)

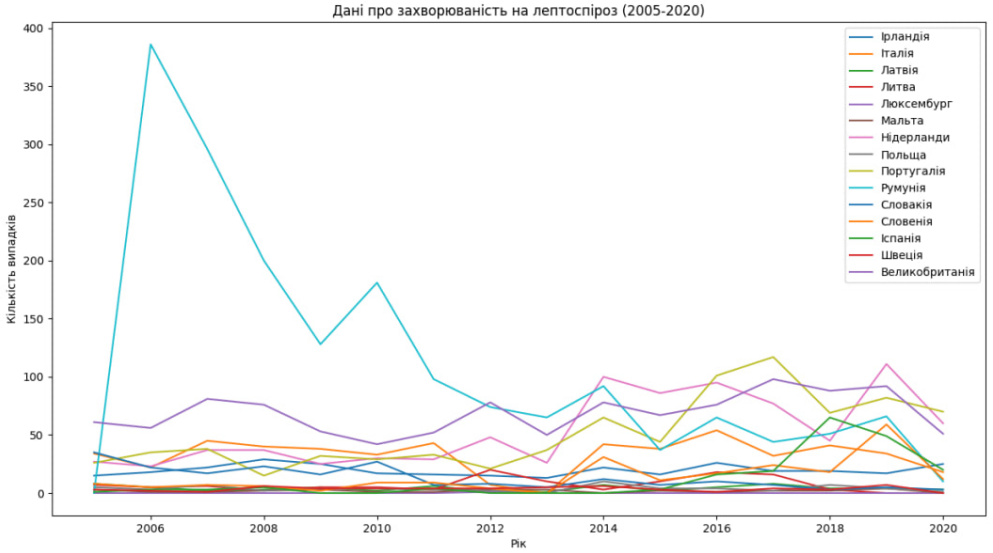

Рис. 6 – загальна діаграма про захворюваність на лептоспіроз для кожної країни Європейського союзу (Частина 2)

Така візуалізація допомагає швидко визначити зміни в рівнях захворюваності та виявити будь-які аномалії або значущі тенденції. З іншого боку, графік середніх значень на рис. 7 показує середній рівень захворюваності в усіх країнах за кожен рік. Цей графік надає ширшу перспективу, дозволяючи оцінити загальну тенденцію захворюваності на лептоспіроз протягом досліджуваного періоду. На ньому можна помітити значне збільшення кількості випадків з 2014 по 2019 роки.

Рис. 7 – графік середніх значень захворюваності на лептоспіроз для кожної країни Європейського союзу

Аналогічні графіки були побудовані і для GeoTIFF наборів даних. Особливу увагу варто звернути на графіки на рис. 8 та рис. 9.

Рис. 8 – графік середніх та медіанних значень вологості в ЄС (2005-2020)

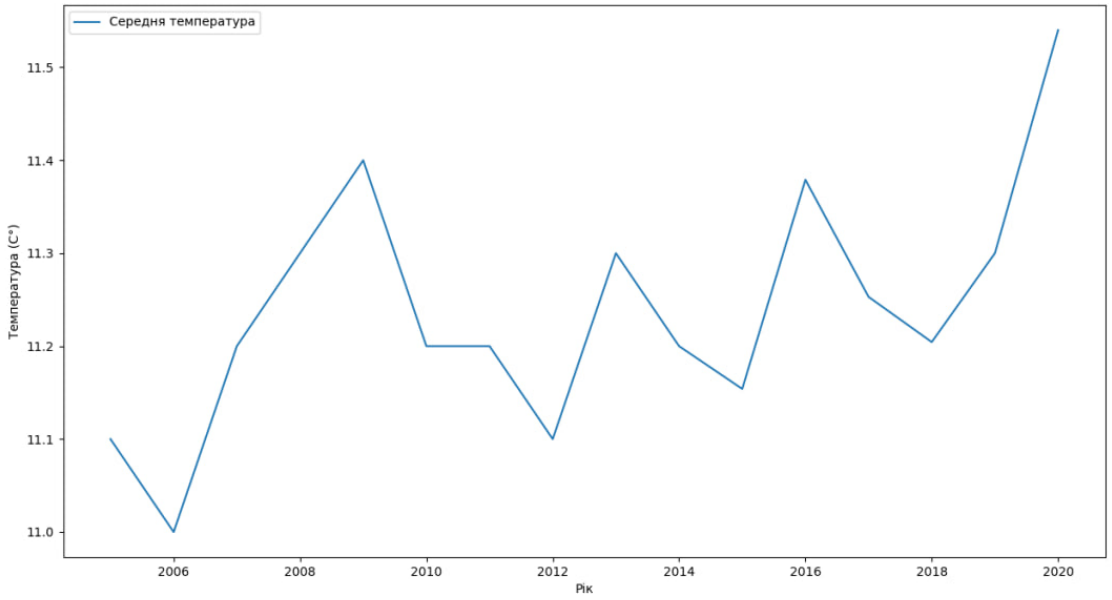

Рис. 9 – графік середніх значень температури в ЄС (2005-2020)

На графіках на рис. 8 та рис. 9 можна побачити чіткий тренд на збільшення значень в проміжку з 2014 по 2019 роки.

На основі зібраних даних переходимо до результатів кореляційного аналізу.

Таблиця 7 - Кореляційна матриця

	Середня вологість	Середня температура	Загальна кількість випадків
Середня вологість	1.0	-0.324396	0.382656
Середня температура	-0.324396	1.0	-0.145007
Загальна кількість випадків	0.382656	-0.145007	1.0

Середня вологість та загальна кількість випадків: Позитивна кореляція 0,38 свідчить про те, що зі збільшенням середньої вологості повітря може зростати загальна кількість випадків лептоспірозу. Ця кореляція не дуже сильна, але все ж таки вказує на можливий зв'язок.

Середня температура та загальна кількість випадків: Негативна кореляція -0,14 свідчить про те, що зі зростанням середньої температури може спостерігатися незначне зменшення загальної кількості випадків лептоспірозу. Однак ця кореляція є слабкою і не може бути значущою.

Середня вологість та середня температура: Негативна кореляція -0,32 свідчить про те, що зі збільшенням середньої вологості повітря середня температура може знизитися. Ця кореляція не дуже сильна, але все ж таки вказує на можливий зв'язок.

Також згідно з обраними методами дослідження були розраховані коефіцієнти кореляції Пірсона, Спірмена та Кендала.

Таблиця 8 - Кореляційна матриця Пірсона

	Середня вологість	Середня температура	Загальна кількість випадків
Середня вологість	1.0	-0.324396	0.382656
Середня температура	-0.324396	1.0	-0.145007
Загальна кількість випадків	0.382656	-0.145007	1.0

Таблиця 9 - Матриця рангової кореляції Спірмена

	Середня вологість	Середня температура	Загальна кількість випадків
Середня вологість	1.0	-0.265362	0.398876
Середня температура	-0.265362	1.0	-0.084764
Загальна кількість випадків	0.398876	-0.084764	1.0

Таблиця 10 - Матриця рангової кореляції Кендала

	Середня вологість	Середня температура	Загальна кількість випадків
Середня вологість	1.0	-0.241926	0.276170
Середня температура	-0.241926	1.0	-0.052223
Загальна кількість випадків	0.276170	-0.052223	1.0

На основі кореляційних матриць для коефіцієнтів кореляції Пірсона, Спірмена та Кендала можна зробити наступні висновки:

- Існує позитивна кореляція між середньою вологістю повітря та загальною кількістю випадків лептоспірозу. Цей зв'язок є найсильнішим у матриці рангової кореляції Спірмена (rho = 0,398876), за ним слідує матриця кореляції Пірсона (r = 0,382656) та матриця рангової кореляції Кендала (tau = 0,276170). Це свідчить про те, що зі збільшенням середньої вологості повітря кількість випадків лептоспірозу також має тенденцію до зростання.
- Між середньою температурою та середньою вологістю існує негативний кореляційний зв'язок. Цей зв'язок є найсильнішим у матриці кореляції Пірсона (r = -0,324396), за ним слідують матриця рангової кореляції Кендала (tau = -0,241926) та матриця рангової кореляції Спірмена (rho = -0,265362). Це вказує на те, що зі зростанням середньої температури середня вологість має тенденцію до зниження.
- Зв'язок між середньою температурою та загальною кількістю випадків лептоспірозу також є негативним. Це означає, що з підвищенням середньої температури кількість випадків лептоспірозу має тенденцію до зниження. Однак кореляція є відносно слабкою за всіма трьома кореляційними матрицями (r Пірсона = -0,145007, rho Спірмена = -0,084764 і tau Кендала = -0,052223), що свідчить про те, що цей зв'язок може бути не таким сильним, як зв'язок між середньою вологістю повітря і загальною кількістю випадків лептоспірозу.

Ці результати свідчать про те, що вологість може бути більш важливим фактором, що впливає на кількість випадків лептоспірозу, ніж температура.

Також цьому аналізі була використана модель ARIMA (авторегресійне інтегроване ковзне середнє) для прогнозування захворюваності на лептоспіроз у різних європейських країнах. Наданий набір даних для побудови моделі містив агрегований та об’єднаний масив всіх датасетів, що використовувались під час дослідження. Отримані прогнози з таблиці 11, можуть бути корисними для органів охорони здоров'я, щоб передбачити та підготуватися до майбутніх випадків лептоспірозу. Варто зазначити, що прогноз зроблений на 2021 рік через те, що даних про поширення лептоспірозу на цей рік ще немає, адже ECDC(Європейський центр з профілактики та контролю захворювань) публікує такі дані з затримкою в два роки.

Таблиця 11 - Прогноз захворюваності на лептоспіроз на 2021 рік

Країна	**Кількість випадків**	**Країна**	**Кількість випадків**	**Країна**	**Кількість випадків**
Австрія	8	Німеччина	130	Нідерланди	98
Бельгія	11	Греція	18	Польща	5
Болгарія	10	Угорщина	3	Португалія	38
Хорватія	16	Ісландія	0	Румунія	32
Кіпр	0	Ірландія	35	Словакія	4
Чехія	20	Італія	17	Словенія	47
Данія	14	Латвія	4	Іспанія	22
Естонія	12	Литва	4	Швеція	3
Фінляндія	1	Люксембург	0	Великобританія	51
Франція	211	Мальта	2		

Щоб оцінити точність ARIMA моделі, важливо виконати позавибіркову перевірку, порівнявши її прогнози з фактичними спостережуваними значеннями, тому був отриманий прогноз на 2020 рік, дані про який є в наборі даних, що був використаний під час дослідження. В результаті цього була отримана таблиця 12.

Таблиця 12 - Порівняння прогнозованої захворюваності з актуальними результатами на 2020 рік

Країна	**Прогнозована кількість випадків**	**Реальна кількість випадків**
Австрія	9	11
Бельгія	8	11
Болгарія	0	1
Хорватія	0	4
Кіпр	0	0
Чехія	23	27
Данія	17	14
Естонія	8	10
Фінляндія	0	0
Франція	135	127
Німеччина	106	118
Греція	21	17
Угорщина	15	3
Ісландія	0	0
Ірландія	19	25
Італія	23	18

Латвія	3	3
Литва	2	0
Люксембург	0	0
Мальта	0	0
Нідерланди	63	60
Польща	4	1
Португалія	71	70
Румунія	9	10
Словакія	3	3
Словенія	15	12
Іспанія	25	20
Швеція	0	0
Великобританія	59	51

ВИСНОВКИ

В межах дипломної роботи було досягнуто наступних результатів:

1. Був проведений комплексний огляд літератури про лептоспіроз, його епідеміологію та пов'язані з ним геодані. Та виявлено, що аналогічні дослідження, в географічних і часових межах, відсутні.
2. Кліматичні геодані та дані про захворюваність на лептоспіроз були отримати та попередньо оброблені.
3. Були розроблені та впроваджені відповідні статистичні моделі та методи для виявлення та аналізу закономірностей, а саме було пораховано зведені статистичні дані та додаткову описову статистику для наявних датасетів. Також було побудовано діаграму про загальну захворюваність на лептоспіроз для кожної країни Європейського Союзу та графік середніх значень захворюваності на лептоспіроз для кожної країни ЄС, графік середніх та медіанних значень вологості в ЄС, графік середніх значень температури в ЄС.

   В межах кореляційного аналізу були отримані коефіцієнти кореляції Пірсона, Спірмена та Кендала. В наслідок чого було виявлено, що існує додатній зв'язок між середньою вологістю повітря та кількістю випадків лептоспірозу. Найсильніше цей зв'язок спостерігається у ранговій кореляції Спірмена (rho = 0,398876), потім йде кореляція Пірсона (r = 0,382656), а наостанок рангова кореляція Кендала (tau = 0,276170). Ці дані вказують на те, що зі збільшенням середньої вологості повітря кількість випадків лептоспірозу має схильність до зростання. За допомогою моделі ARIMA були отримані прогнози з низьким коефіцієнтом помилки на 2020 та 2021 роки.

4. Результати були інтерпретовані та контекстуалізовані.
5. Були сформовані рекомендації щодо політики громадського здоров'я та майбутніх досліджень у світлі отриманих результатів та прогнозів.

4. СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. TR Fraga, AS Barbosa, L Isaac 2011. Leptospirosis: Aspects of innate immunity, immunopathogenesis and immune evasion from the complement system. 73(5) 408-419.
2. M DebMandal, S Mandal, NK Pal 2011. Serologic evidence of human leptospirosis in and around Kolkata, India: A clinico–epidemiological study. 4(12) 1001-1006.
3. SUMI, A., TELAN, E., CHAGAN-YASUTAN, H., PIOLO, M., HATTORI, T., & KOBAYASHI, N. (2017). Effect of temperature, relative humidity and rainfall on dengue fever and leptospirosis infections in Manila, the Philippines. *Epidemiology & Infection, 145*(1), 78-86.
4. Sudarat Chadsuthi, Charin Modchang, Yongwimon Lenbury, Sopon Iamsirithaworn, Wannapong Triampo, Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time–series and ARIMAX analyses, Asian Pacific Journal of Tropical Medicine, 5(7), 2012, 539-546
5. European Centre for Disease Prevention and Control. Annual epidemiological report 2014 – food- and waterborne diseases and zoonoses. Stockholm: ECDC; 2014.
6. Suggested citation: European Centre for Disease Prevention and Control. Annual Epidemiological Report 2016 – Leptospirosis. Stockholm: ECDC; 2016.
7. European Centre for Disease Prevention and Control: Annual Epidemiological Report on Communicable Diseases in Europe 2008. Stockholm, 2008.
8. European Centre for Disease Prevention and Control: Annual Epidemiological Report on Communicable Diseases in Europe 2007. Stockholm, 2007.

9. European Centre for Disease Prevention and Control. Annual Epidemiological Report on Communicable Diseases in Europe 2009. Stockholm: ECDC; 2009.
10. European Centre for Disease Prevention and Control. Leptospirosis. In: ECDC. Annual epidemiological report for 2015. Stockholm: ECDC; 2018.
11. European Centre for Disease Prevention and Control. Leptospirosis. In: ECDC. Annual epidemiological report for 2016. Stockholm: ECDC; 2021.
12. Suggested citation: European Centre for Disease Prevention and Control. Leptospirosis. In: ECDC. Annual Epidemiological Report for 2017. Stockholm: ECDC; 2022.
13. European Centre for Disease Prevention and Control. Leptospirosis. In: ECDC. Annual Epidemiological Report for 2018. Stockholm: ECDC; 2022.
14. European Centre for Disease Prevention and Control. Leptospirosis. In: ECDC. Annual Epidemiological Report for 2019. Stockholm: ECDC; 2022.
15. Suggested citation: European Centre for Disease Prevention and Control. Leptospirosis. In: ECDC. Annual Epidemiological Report for 2020. Stockholm: ECDC; 2022.
16. Common animal-associated infections (England and Wales): Health Protection Report. 15(20) 2021.
17. ERA5-Land daily: Air temperature at 2 meter above surface (2000 - 2020)
18. Monthly time series of spatially enhanced relative humidity for Europe at 1000 m resolution (2000 - 2021) derived from ERA5-Land data.

19. Dietz, A.J., Kuenzer, C., Dech, S., 2015. Global SnowPack: a new set of snow cover parameters for studying status and dynamics of the planetary snow cover extent. Remote Sensing Letters 6, 844-853.
20. Dietz, A.J., Kuenzer, C., Conrad, C., 2013. Snow-cover variability in central Asia between 2000 and 2011 derived from improved MODIS daily snow-cover products. International Journal of Remote Sensing 34, 3879-3902.

АНОТАЦІЯ

Основною ціллю цієї дипломної роботи було дослідження потенційного зв'язку між захворюваністю на лептоспіроз та геопросторовими даними в Європейському Союзі шляхом вивчення різних кліматичних факторів. У ході роботи було розглянуто та проаналізовано існуючі роботи на цю тему, розроблено та впроваджено відповідні статистичні моделі та методи для виявлення, аналізу та прогнозування закономірностей, тенденцій та кореляцій між захворюваністю на лептоспіроз та кліматичними геопросторовими даними . Було використано GeoTIFF датасети та набір даних про захворюваність на лептоспіроз. У якості мови програмування була обрана мова програмування Python, дослідження проводилися за допомогою бібліотек Pandas та Rasterio.

У ході дипломної роботи була знайдена та досліджена кореляція між геоданими та захворюваністю на лептоспіроз і наданий прогноз на декілька років.

ANNOTATION

The main objective of this thesis was to investigate the potential relationship between the incidence of leptospirosis and geospatial data in the European Union by examining various climatic factors. As part of the thesis, existing work on this topic was reviewed and analyzed, and appropriate statistical models and methods were developed and implemented to identify, analyze and predict patterns, trends and correlations between leptospirosis incidence and climate geospatial data. GeoTIFF datasets and a dataset on the incidence of leptospirosis were used. Python was chosen as the programming language, and the research was conducted using the Pandas and Rasterio libraries.

In the course of the thesis, the correlation between geodata and the incidence of leptospirosis was found and studied, and a forecast for several years was provided.