SDSC2102 Statistical Methods and Data Analysis

Semester-B 2024/2025

Group 23

Weather Variables and Infectious Disease Incidence (2015–2024)

Name	Student ID
GAITE Santana Ysabelle Perez	57865485
SHANKAR Vihaan	56690634
MERCER Clayton Risen	56319058
MIRANDA Rhys David	57820443

Table of Contents

1. Project Background

The spread of infectious diseases depends heavily on weather conditions. Research into these connections enables predictions about future disease patterns from environmental changes and identifies key drivers of shifts while selecting suitable models and developing intervention strategies.

The irregular weather patterns in Hong Kong featuring rainfall and temperature fluctuations with changing humidity levels may cause variations in infectious disease occurrence rates. The combination of urban density with its crowded living conditions creates conditions that allow diseases to spread quickly. The knowledge of weather-disease relationships represents a vital component for Hong Kong to prepare its public health system especially when climate change intensifies severe weather conditions.

2. Objectives

The research aims to evaluate Hong Kong's weather information against its infectious disease statistics to discover climate-related risk factors which will guide intervention strategy development. The research will study how Hong Kong's seasonal patterns affect disease distribution while determining peak disease periods after typhoons to schedule public health initiatives optimally. The research project aims to establish predictive models which will help Hong Kong hospitals forecast weather-related increases in infectious diseases including mosquito-borne illnesses that emerge after rainfall.

3. Previous Research

Research evidence demonstrates that climate patterns create conditions which facilitate the spread of infectious diseases. The WHO published a report titled “Using Climate to Predict Infectious Disease Epidemics” in 2005 which established the link between climate patterns and infectious disease spread. Research indicates that sudden weather occurrences including typhoons and heavy rainfall lead to higher rates of infectious diseases, as highlighted by Ebi et al. in their 2021 study “Extreme Weather and Climate Change: Population Health and Health System Implications”.

This project varies from these aforementioned previous research in several key ways. First, our project analyses using the latest data with an extended timeframe, i.e. data from 2015 to 2024, which includes the COVID epidemic period. Second, we focus on newly emerging diseases that have gained prominence in recent years. Finally, our study has a local focus on Hong Kong, addressing the limited existing research on the specific relationships between weather and disease patterns in this unique context.

4. Data Preprocessing

The data integration process combined four separate datasets including notable infectious diseases together with temperature and humidity and rainfall information into monthly-level data. The data transformation required reshaping disease data while normalising weather parameters and establishing a datetime index to support better time-based analysis. The mean temperature replaced maximum and minimum temperatures because it provided an adequate representation of overall trends. The high humidity in Hong Kong throughout the year led to the decision to keep both humidity and rainfall data for a complete understanding of climatic conditions.

5. Hypothesis Testing

5.1. Methodology and Application

In this project, hypothesis testing was employed to investigate the relationship between various **weather variables** (rainfall, humidity, and mean temperature) and **infectious disease incidence**.

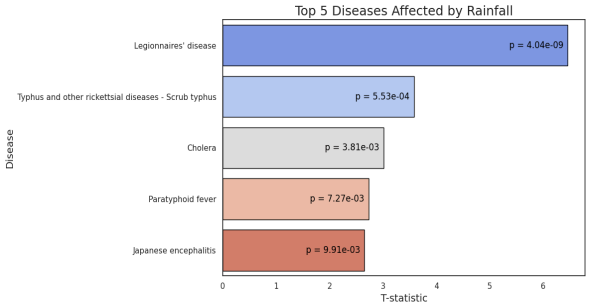
5.1.1. Test Selection and Null and Alternative Hypotheses

We applied the **two-sample independent t-test** (non-equal variances). For each weather variable, the dataset was split into two categories — "High" and "Low" — based on the **median** value of that variable.

- **Null Hypothesis (H_0):** There is **no significant difference** in the average number of disease cases between the high and low weather condition groups.
- **Alternative Hypothesis (H_1):** There **is a significant difference** in the average number of disease cases between the high and low weather condition groups.

5.2. Results

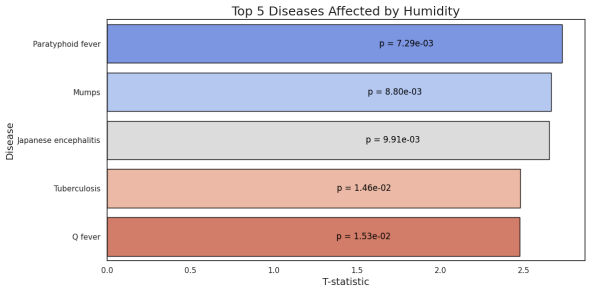
5.2.1. Rainfall



The strongest relationship is observed in **Legionnaires' disease**, which demonstrates the highest t-statistic and a highly significant p-value. This is consistent as Legionella bacteria thrive in moist environments such as water systems and plumbing.

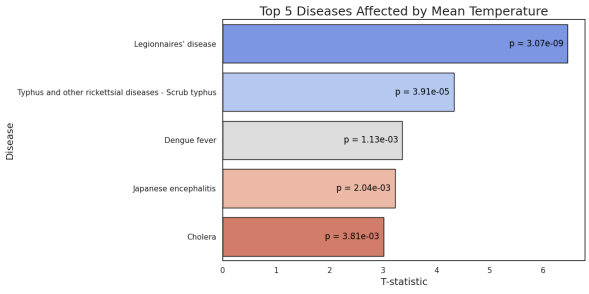
Following this, **Scrub typhus**, **Cholera** and **Paratyphoid fever** rank second, third and fourth, likely reflecting contamination of water sources after rainfall. Lastly, **Japanese encephalitis**, a mosquito-borne infection, displays moderate associations.

5.2.2. Humidity



Paratyphoid fever shows the strongest relationship, indicated by a strong p-value. This suggests that increased humidity may create favorable conditions for the transmission of this disease. **Mumps** exhibits a notable association with humidity. Elevated humidity levels can facilitate the spread of respiratory pathogens, increasing the transmission risk in populated areas.

5.2.3. Temperature



Legionnaires' disease has the highest t-statistic and a highly significant p-value aligning with the biological fact that Legionella bacteria thrive in warmer conditions. Following this, **Scrub typhus** and **Dengue fever** demonstrate notable associations with temperature. Both diseases are influenced by warmer temperatures that enhance the activity of their respective vectors (mites and mosquitoes).

6. Logistic Regression

6.1. Methodology

Logistic Regression is a statistical modeling technique used to predict a binary outcome. In our project we wanted to consider whether a disease occurred in a given month (1) or not (0)—based on one or more predictor variables. The probability is modeled as follows:

$$P(Y = 1) = \frac{1}{1 + e^{\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn}}$$

For each disease with sufficient variation, a logistic regression model was trained using the weather variables: mean temperature, humidity, and rainfall. Binary targets were created per disease to denote whether at least one case occurred in a month. Class balance was maintained using stratified sampling where possible.

6.2. Results

6.2.1. Rainfall

Legionnaires' disease showed the strongest positive relation with rainfall (↑0.99), underscoring its link to water system contamination after typhoons or acid rain. *Scarlet fever* (↑0.15) and *Melioidosis* (↑0.09) also increased with rainfall, highlighting the role of waterborne and flood-associated pathways in disease emergence. However, some diseases exhibited reduced incidence during rainy periods. *Tetanus and Mumps* declined substantially (↓0.62) and (↓0.20), possibly due to lower exposure to outdoor and agricultural environments during wet weather. *Novel influenza A (H7)* again showed a negative correlation (↓0.22), reinforcing its association with drier, cooler conditions.

6.2.2. Humidity

Humidity particularly causes respiratory and vector-borne illnesses. Increased humidity was linked to higher incidence of *Q fever* (↑0.28) and *Novel influenza A (H7)* (↑0.26), potentially due to the better environment created for aerosolized pathogens in the air. Other diseases such as *Leprosy*, *Measles*, and *Mumps* also rose under humid conditions. *COVID-19* showed a modest positive association (↑0.14), contrary to some earlier assumptions about its seasonal behavior.On the other side of the bridge, several diseases like *invasive pneumococcal disease* (↓0.25) and *Legionnaires' disease* (↓0.47) both declined in humid environments, supporting the notion that drier air may facilitate respiratory spread. Additionally, moderate humidity-related reductions were observed for *Melioidosis*, *Scarlet fever*, and *Typhoid fever*, suggesting a potential dampening effect of high humidity on certain pathogens.

6.2.3. Temperature

Japanese Encephalitis demonstrated the strongest temperature-related rise (coefficient: ↑0.48), peaking during hot seasons. *Mumps* also exhibited a substantial increase (↑0.47), likely due to seasonal dynamics and overlap with school calendars. *Streptococcus suis*, *Typhoid fever*, *Malaria*, *Measles*, and *Mpox* followed similar upward trends, suggesting that rises in temperatures may create favorable conditions for transmission. On the other hand, certain diseases declined with warmer temperatures. *Novel influenza A (H7)* showed a moderate reduction (↓0.33), and *Whooping cough* (↓0.24), indicating increased risk during cooler seasons. These patterns align with the well-established seasonality of many respiratory and vector-borne infections.

6.3. Classification Report Analysis

Average Accuracy: 0.76
Average F1 Score: 0.40
Average ROC-AUC: 0.53
Best Model: Legionnaires' disease with Accuracy: 1.00
Worst Model: Measles with Accuracy: 0.38

7. Decision Tree Classification

To explore non‑linear interactions between weather and Legionnaires' disease incidence, we trained a decision‑tree classifier using mean temperature and total monthly rainfall to predict whether case counts fall at or below the median ("Low") or above it ("High").

7.1. Model Development

- **Data split:** Stratified train–test split (30% train, 70% test), yielding 36 training samples and 84 test samples.
- **Features:**
  - *Mean_Temp* (°C)
  - *Rainfall* (mm)
- **Target:** `Legionnaires_Category ∈ {Low, High}` based on median monthly cases.

7.2. Performance

On held‑out data, the tree achieved an accuracy of 61%. This moderate score reflects the limited feature set and sample size but nevertheless captures key environmental thresholds.

7.3. Tree Structure and Interpretation

The figure below shows the fitted tree. At the root (Gini = 0.50), **Mean_Temp ≤ 19.26 °C** cleanly separates cooler months, mostly "Low" incidence, from warmer ones. For warmer months, **Rainfall** splits further: low‑rainfall months tend to remain "Low," while high‑rainfall months overwhelmingly classify as "High." Additional splits on temperature and rainfall refine these groups into pure terminal nodes (Gini = 0), illustrating specific climate regimes most conducive to Legionella outbreaks.

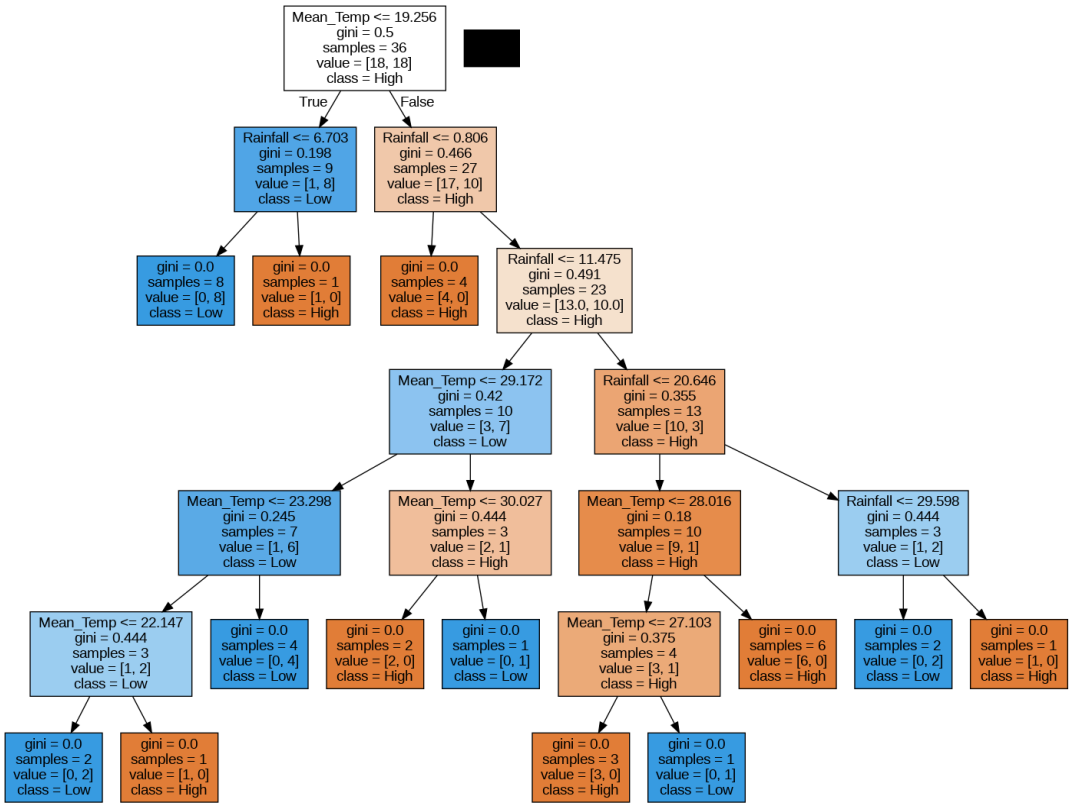


Figure: Decision tree model predicting monthly Legionnaires' disease incidence category, based on mean temperature and rainfall.

8. Conclusion

This study investigated the relationship between weather conditions—rainfall, humidity, and temperature—and the incidence of infectious diseases using hypothesis testing, logistic regression, and decision tree modeling. We found that environmental elements are reliable indicators of disease patterns because this research established their significant relationship with infectious diseases.

Statistical analysis revealed significant associations between environmental factors and disease occurrence, with Legionnaires' disease showing the strongest and most consistent links, particularly to rainfall and temperature. Other diseases, including Paratyphoid fever, Scrub typhus, and Dengue, also demonstrated weather-sensitive patterns. Logistic regression quantified these effects, showing both positive and negative associations depending on disease type and climate condition. While some diseases increased with warmer, wetter conditions, others declined, particularly those influenced by reduced outdoor exposure or respiratory seasonality. Model performance varied, with an average accuracy of 0.76 and high predictive success for Legionnaires' disease. The decision tree model, though moderate in accuracy (61%), offered clear thresholds for high-risk periods based on temperature and rainfall, reinforcing the utility of interpretable models in public health surveillance.

Weather-pattern predictive models provide hospitals and healthcare systems with an essential tool to anticipate disease outbreaks. Through these models healthcare facilities can predict rising patient numbers which enables them to improve resource management and emergency readiness. Hospitals can improve their crisis management by scheduling their staff and distributing medical resources more effectively. Timely intervention planning becomes possible through recognizing these high-risk periods. Protective measures become most effective for vulnerable populations when implemented during high-risk periods. The research demonstrates that climate and weather data plays a vital role in developing an advanced public health response system and we hope officials in Hong Kong can help us do that.

9. References

- Extreme weather and climate change: Population health and health system implications. (n.d.). PMC Home. https://pmc.ncbi.nlm.nih.gov/articles/PMC9013542/
- Kuhn, K., Campbell-Lendrum, D., Haines, A., & Cox, J. (2005). Using climate to predict infectious disease epidemics. IRIS Home. https://iris.who.int/bitstream/handle/10665/43379/9241593865.pdf?sequence=1

10. Appendix

The datasets used and the code file can be found at the following link: SDSC2102 Project