Project Title: TEMPORAL ANALYSIS OF CLIMATIC INFLUENCES ON FOREST FIRE PATTERNS IN PENINSULAR MALAYSIA USING STATISTICAL METHOD

Prepared by: GRACE LING KIAN HWAI, MCS231026, GraceLKH

# Chapter 3 Methodology

## 3.1 Introduction

This chapter highlighted the structured approach taken to analyse the temporal patterns of forest fires in Peninsular Malaysia and their climatic influences using statistical methods. The chapter begin with stating the research design for this study, following by explaining the data science project life cycle which describes data sources, data collection methods and data pre-processing steps. In the next session, it continued with describing the data analysis, defining evaluation metrics and lastly comparing the methods used in previous studies with those proposed in this research and identified the research gap.

## 3.2 Research Design

The study's research design is observational, which is suitable for understanding and analysing the existing patterns and relationships between climatic variables and forest fire occurrences. This design enables the gathering of information from natural settings without manipulating variables, ensuring that the findings are reflective of real-world scenarios. The choice of an observational design is justified by the need to analyse historical data to identify trends and correlations, which is essential for developing predictive models and management strategies.

## 3.3 Data Science Project Life Cycle

The Data Science Project Life Cycle (Figure 1) is a systematic approach that guides the progression of a data-driven project from initial problem definition to deployment and maintenance of the solution. This cycle encompasses a series of interconnected stages designed to ensure thorough understanding, accurate analysis, and effective implementation of data-centric solutions. Every step in the process builds on the one before it, leading to a cohesive workflow that improves the accuracy and usefulness of the outcomes. The data science project life cycle followed in this session includes the first three stages: problem formulation, data collection and data pre-processing. The other stages such as data exploration will be discussed in Chapter 4, model building will be presented in Chapter 5, and the last two stages: deployment and monitoring and maintenance will be outlined in Chapter 7.

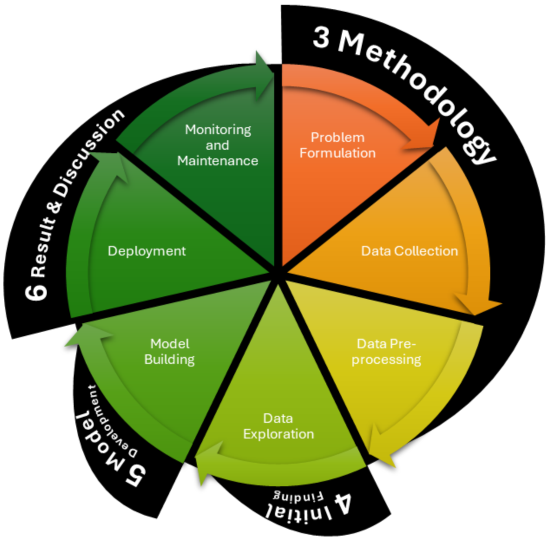

Figure 3.2: Data Science Project Life Cycle

### 3.2.1 Problem Formulation

Forest fires in Peninsular Malaysia pose significant risks to ecosystems, biodiversity, and human settlements. Understanding the temporal patterns of these fires and their climatic influences is crucial for developing effective fire management strategies and early warning systems.

### 3.2.2 Data Source and Data Collection

Data from multiple sources was gathered for this research to ensure a thorough examination of the factors that impact forest fire trends:

- Google Earth Engine: The dataset on forest fires in Peninsular Malaysia is obtained from multiple-source remote sensing data through google earth engine. This platform integrates data from various satellites and provides high-resolution imagery, enabling detailed analysis of fire hotspots and burned areas.
- IMF: The climate change dataset is obtained from IMF Climate Change Dashboard. This dataset includes comprehensive climate variables such as temperature, precipitation, humidity, and other relevant indicators, offering valuable insights into the climatic factors influencing forest fires.

### 3.2.3 Data Pre-Processing

The data pre-processing steps are critical to ensuring the accuracy and reliability of the analysis. These steps involve several key activities: data cleaning, data transformation, and feature engineering. Figure 3.2.3 illustrates the process.

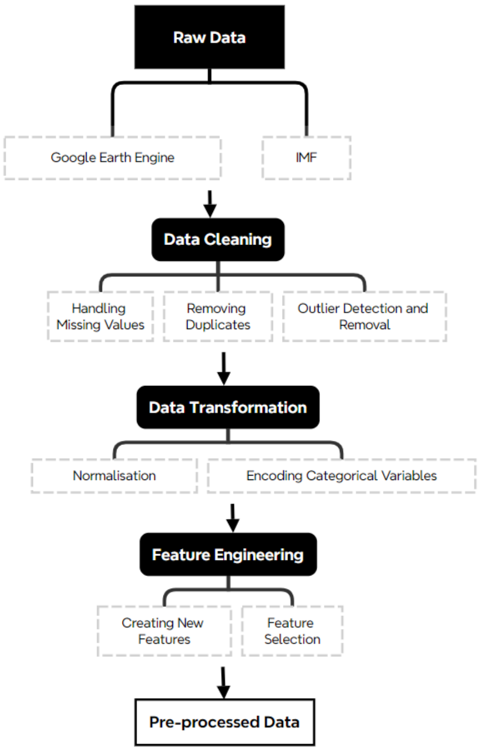

Figure 3.2.3: Data Pre-Processing Flow

Step 1: Data Cleaning

Data cleansing is the initial phase in the preprocessing stage, aimed at enhancing the dataset's quality by removing or correcting inaccuracies and inconsistencies. This involves removing duplicate records to avoid redundancy and eliminating irrelevant data points that do not contribute to the analysis. Handling missing values is another essential part of data cleaning, which can be done through imputation methods, such as filling in empty data entries with the mean, median, or mode, or by removing records with missing data if the proportion is insignificant. Additionally, filtering out noise and outliers is crucial to prevent skewed analysis results. Statistical methods such as z-scores or IQR (Interquartile Range) can be utilised in detecting and addressing the outliers.

Step 2: Data Transformation

Following the cleaning process, data transformation ensures that the data is in a format suitable for analysis. Normalizing and standardizing the data is a critical step to ensure consistency across the dataset, especially when dealing with variables that operate on different scales. This procedure consists of adjusting numerical data to a uniform scale without altering variations in the values' ranges. Converting categorical variables into numerical formats is also essential for statistical analysis, which can be achieved by using methods such as label encoding or one-hot encoding. Aggregating data to appropriate temporal and spatial scales ensures that the analysis captures the relevant patterns and trends. For instance, weather data might be aggregated on a daily or monthly basis to match the temporal resolution of fire incident data.

Step 3: Feature Engineering

Feature engineering involves generating new features using available data to improve the efficacy of the models. This involves leveraging domain knowledge to generate features that capture critical aspects of the data. For example, calculating fire weather indices such as the Fire Weather Index (FWI) or the Keetch-Byram Drought Index (KBDI) can provide insights into fire risk levels. Analysing time-related characteristics, such as the specific day or time of year, helps in understanding the seasonal patterns of forest fires. Additionally, generating interaction terms between climatic variables (e.g., interaction between temperature and humidity) can help in capturing the combined effects of multiple factors on fire occurrences.

These initial processing tasks together guarantee that the dataset is ready to use for future analysis steps, enhancing the quality and reliability of the insights derived from the statistical models.

## 3.3 Data Analysis

The data analysis involves employing advanced statistical methods to process and analyse the collected data. Time series analysis and regression models are used to quantify the relationship between climatic factors and fire patterns, offering a robust analytical framework. These techniques are suitable for handling the high-resolution temporal data and identifying significant trends and correlations.

## 3.4 Evaluation Metrics

The research's effectiveness is measured using defined evaluation metrics. These metrics help determine whether the findings support the research question and objectives. Examples of evaluation metrics include:

- Prediction Accuracy: The accuracy of the models in predicting fire occurrences based on climatic variables.
- Error Rates: Methods like mean absolute error (MAE) and root mean square error (RMSE) are utilized for evaluating the model's performance.
- Model Robustness: The ability of the models to generalize to new data and avoid overfitting, assessed through cross-validation techniques.

### 3.5 Comparison of Methods

#### 3.5.1 Previous Studies

Recent advancements in remote sensing and machine learning have significantly enhanced the detection and monitoring of forest fires. Various studies have employed satellite imagery, drones, and ground-based sensors to detect and analyse fire patterns. In 2019, Sabani et al. utilized MODIS satellite data to detect forest fire hotspots in Southeast Asia, demonstrating the effectiveness of remote sensing in fire detection. In 2021, Saruni Dwiasnati and Yudo Devianto applied machine learning algorithms to classify fire-prone areas using historical fire data and climatic variables. A study by Ghali and Akhloufi in 2023 integrated drone-based imagery with deep learning models to accurately map burned areas and predict fire spread. In 2023, Sudiana et al. experiment with five detection methods and found CNN-RF hybrid method achieved 97% accuracy in fire detection. Finally, a theoretical framework proposed by Meng et al. in 2024 emphasized the importance of integrating cutting-edge fire detection and suppression technologies to reduce the likelihood and impact of forest fires.

#### 3.5.2 Current Study

This study focuses on several enhancements using statistical methods to analyse forest fire patterns in Peninsular Malaysia. First, it utilizes high-resolution temporal data to capture more detailed patterns and trends, providing a finer granularity of analysis. Second, the research utilizes sophisticated statistical techniques such as regression models and time series analysis, to quantify the relationship between climatic factors and fire patterns, offering a robust analytical framework. Lastly, the research implements real-time validation techniques to ensure the robustness and reliability of the models, addressing potential issues of overfitting and ensuring the models' applicability in real-world scenarios. By building on the strengths of previous studies and introducing these novel elements, this research aims to provide a deeper and more accurate understanding of the climatic influences on forest fire patterns in the region.

### 3.5.3 Summarisation and Research Gap

| Aspect | Previous Research | Current Research | Research Gap |
|---|---|---|---|
| **Data Collection** | Satellite imagery, weather stations | Google Earth Engine, IMF Climate Change Dashboard | Need for high-resolution temporal data |
| **Methodology** | Statistical analysis, machine learning | Advanced statistical methods | Integration of advanced AI techniques |
| **Geographic Focus** | Southeast Asia | Peninsular Malaysia | Need for region-specific studies |
| **Validation** | Separate dataset | Cross-validation with real-time data | Implementation of real-time validation techniques |
| **Applications** | Fire detection, prediction | Fire detection, prediction, management | Development of comprehensive fire management systems |

Table 3.5.3 Summarisation of The Comparison Between Previous and Current Methods and The Research Gap.

The current research's aim is to add to the existing knowledge by pinpointing these gaps in research by addressing the need for high-resolution temporal data, integrating advanced statistical techniques, focusing on region-specific studies, implementing real-time validation techniques, and developing comprehensive fire management systems.

## 3.6 Summary

This methodology chapter outlines the structured approach used to analyse the temporal patterns of forest fires in Peninsular Malaysia and their climatic influences using statistical methods. The chapter covers the full data science project life cycle, which includes problem definition, data collection from Google Earth Engine and IMF Climate Change Dashboard, data pre-processing involving cleaning, transformation, and feature engineering, and rigorous model building and validation. The study builds on previous research from 2019 to 2024, incorporating high-resolution temporal data, advanced statistical techniques, and real-time validation to improve the precision and dependability of the analysis. By identifying and addressing research gaps, this study aims to provide valuable insights and contribute to effective fire management strategies in the region.

## References

Ghali, R., & Akhloufi, M. A. (2023). Deep Learning Approaches for Wildland Fires Using Satellite Remote Sensing Data: Detection, Mapping, and Prediction. *Fire*, *6*(5), 192. https://doi.org/10.3390/fire6050192

Meng, L., O'Hehir, J., Gao, J., Peters, S., & Hay, A. (2024). A theoretical framework for improved fire suppression by linking management models with smart early fire detection and suppression technologies. *Journal of Forestry Research*, *35*(1), 86. https://doi.org/10.1007/s11676-024-01737-3

Sabani, W., Rahmadewi, D. P., Rahmi, K. I. N., Priyatna, M., & Kurniawan, E. (2019). Utilization of MODIS data to analyze the forest/land fires frequency and distribution (case study: Central Kalimantan Province). *IOP Conference Series: Earth and Environmental Science*, *243*, 012032. https://doi.org/10.1088/1755-1315/243/1/012032

Saruni Dwiasnati & Yudo Devianto. (2021). Classification of forest fire areas using machine learning algorithm. *World Journal of Advanced Engineering Technology and Sciences*, *3*(1), 008–015. https://doi.org/10.30574/wjaets.2021.3.1.0048

Sudiana, D., Lestari, A. I., Riyanto, I., Rizkinia, M., Arief, R., Prabuwono, A. S., & Sri Sumantyo, J. T. (2023). A Hybrid Convolutional Neural Network and Random Forest for Burned Area Identification with Optical and Synthetic Aperture Radar (SAR) Data. *Remote Sensing*, *15*(3), 728. https://doi.org/10.3390/rs15030728