Agroclimatic & Crop Modeling

Justin Djagba Ph.D

2025-02-07

Contents

1 Introduction

Climate change poses significant challenges to agricultural productivity, particularly in regions heavily reliant on rain-fed agriculture. Understanding the relationship between climate hazards and crop production is critical for developing adaptation strategies. This report analyzes the impact of climate hazards on sub-national crop production in India using a combination of crop production data and climate indicators.

The study aims to:

- Explore crop production trends (1947-2014)
- Calculate a climate hazard indicator
- Analyze its relationship with crop yields
- Provide insights using statistical modeling and visualization

2 Dataset Overview

2.1 Some needed packages

```
#install.packages("naniar")
#install.packages("ggcorrplot")
#install.packages("magrittr")
#install.packages("factoextra")
```

```
contry_apy <- read.csv("//cloud/project/data 13 models/country_apy_fao_1956-2014.csv")
district_apy <- read.csv("/cloud/project/data 13 models/district_apy_interpolated_1956-2008.csv")
contry_a_f = read.csv('/cloud/project/data 13 models/county_a_fao_1947-2014.csv')
```

2.2 Data Explorations

2.2.1 Check the First Few Rows

```
head(contry_apy)  # Display the first 6 rows
```

```
##   year      crop area production yield   prop_area
## 1 1956    Barley 3418       2815  0.82 0.026638402
## 2 1956   Cassava  247       1792  7.30 0.001925010
## 3 1956  Chickpea 9779       5418  0.55 0.076213263
## 4 1956    Cotton 8050       1684  0.21 0.062738191
## 5 1956 Dry beans 6599       1454  0.22 0.051429729
## 6 1956  Dry peas  919        576  0.63 0.007162285
```

```
tail(contry_apy)  # Display the last 6 rows
```

```
##      year                    crop      area production   yield    prop_area
## 4581 2014 Tobacco, unmanufactured   432.679    720.725  1.6657 0.0021561195
## 4582 2014                Tomatoes   882.030  18735.910 21.2418 0.0043953186
## 4583 2014   Vegetables, fresh nes  2623.000  36838.000 14.0442 0.0130708940
## 4584 2014     Walnuts, with shell    31.000     43.000  1.3871 0.0001544787
## 4585 2014             Watermelons    29.253    411.547 14.0688 0.0001457731
## 4586 2014                   Wheat 30470.000  95850.000  3.1457 0.1518376436
```

```r
library(dplyr)


replace_missing_values <- function(main_df, ref_df, key_columns) {

  merged_df <- main_df %>%
    left_join(ref_df, by = key_columns, suffix = c("", "_ref"))


  cols_to_replace <- setdiff(names(main_df), key_columns)


  for (col in cols_to_replace) {
    ref_col <- paste0(col, "_ref")
    if (ref_col %in% names(merged_df)) {
      merged_df[[col]] <- ifelse(is.na(merged_df[[col]]), merged_df[[ref_col]], merged_df[[col]])
    }
  }


  merged_df <- merged_df %>%
    select(all_of(names(main_df)))

  return(merged_df)
}


contry_apy_clean <- replace_missing_values(contry_apy, district_apy, key_columns = c("year", "crop"))


head(contry_apy_clean)
```

```
##   year   crop area production yield prop_area
## 1 1956 Barley 3418       2815  0.82 0.0266384
## 2 1956 Barley 3418       2815  0.82 0.0266384
## 3 1956 Barley 3418       2815  0.82 0.0266384
## 4 1956 Barley 3418       2815  0.82 0.0266384
## 5 1956 Barley 3418       2815  0.82 0.0266384
## 6 1956 Barley 3418       2815  0.82 0.0266384
```

2.3 Check Dataset Dimensions

```r
dim(contry_apy)  # Returns (number of rows, number of columns)
```

```
## [1] 4586    6
```

- This dataset contains 4586 observations from different crop. It provides information on crop production statistics over time, including year, crop type, area cultivated, production, yield, and proportion of area.

```r
str(contry_apy)  # Structure of the dataset (column names, data types)
```

```
## 'data.frame':    4586 obs. of  6 variables:
##  $ year      : int  1956 1956 1956 1956 1956 1956 1956 1956 1956 1956 ...
##  $ crop      : chr  "Barley" "Cassava" "Chickpea" "Cotton" ...
##  $ area      : num  3418 247 9779 8050 6599 ...
##  $ production: num  2815 1792 5418 1684 1454 ...
##  $ yield     : num  0.82 7.3 0.55 0.21 0.22 0.63 0.78 NA 1.01 0.42 ...
##  $ prop_area : num  0.02664 0.00193 0.07621 0.06274 0.05143 ...
```

```
colnames(contry_apy)  # List of column names
```

```
## [1] "year"       "crop"       "area"       "production" "yield"     
## [6] "prop_area"
```

```
summary(contry_apy)  # Get summary statistics for each column
```

```
##       year          crop                area           production     
##  Min.   :1956   Length:4586        Min.   :    0.2   Min.   :     0.3  
##  1st Qu.:1973   Class :character   1st Qu.:   59.0   1st Qu.:   137.9  
##  Median :1987   Mode  :character   Median :  303.2   Median :   648.4  
##  Mean   :1987                      Mean   : 2289.5   Mean   :  6712.1  
##  3rd Qu.:2001                      3rd Qu.: 1183.4   3rd Qu.:  3114.8  
##  Max.   :2014                      Max.   :45537.4   Max.   :361037.0  
##                                    NA's   :233       NA's   :51        
##      yield           prop_area      
##  Min.   : 0.0382   Min.   :0.00000  
##  1st Qu.: 0.7410   1st Qu.:0.00035  
##  Median : 2.6923   Median :0.00176  
##  Mean   : 6.2754   Mean   :0.01355  
##  3rd Qu.: 9.0709   3rd Qu.:0.00690  
##  Max.   :76.5328   Max.   :0.25858  
##  NA's   :234       NA's   :233      
```

2.3.1 Statistical Analysis of Variables:

2.3.1.1 1. Year

- The dataset spans from **1956 to 2014**, with the **median year being 1987**.
- The data is **relatively evenly distributed** across the years, with:
  - **First quartile (25th percentile)** at **1973**.
  - **Third quartile (75th percentile)** at **2001**.

2.3.1.2 Insight: The dataset covers a **long time period**, allowing for the analysis of **long-term trends** in crop production.

2.3.2 2. Crop

- There are **4,586 observations** across multiple crops.
- The **crop** variable is **categorical**, with each observation representing a specific crop type.

2.3.2.1 Insight: The dataset includes a **diverse range of crops**, enabling analysis of **crop-specific responses to climate hazards**.

2.3.3 3. Area

- The cultivated area (`area`) ranges from **0.2 to 45,537.4** units, with a **mean of 2,289.5** units.
- The **median area** is **303.2** units, indicating that most observations are for **smaller cultivated areas**, while a few **large outliers** skew the mean.

- There are **233 missing values** in this variable.

2.3.3.1 Insight: The **wide range of cultivated areas** suggests **significant variability** in farm sizes and crop types. The presence of **missing values** may require **imputation or exclusion** in further analysis.

2.3.4 4. Production

- Crop production (`production`) ranges from **0.3 to 361,037.0** units, with a **mean of 6,712.1** units.
- The **median production** is **648.4** units, indicating a **right-skewed distribution** with a few **high-production outliers**.
- There are **51 missing values** in this variable.

2.3.4.1 Insight: The **skewness in production data** suggests that a **small number of regions or crops** contribute disproportionately to total production. **Missing values** should be addressed before analysis.

2.3.5 5. Yield

- Crop yield (`yield`) ranges from **0.0382 to 76.5328** units, with a **mean of 6.2754** units.
- The **median yield** is **2.6923** units, indicating a **right-skewed distribution** with some **high-yield outliers**.
- There are **234 missing values** in this variable.

2.3.5.1 Insight: The **variability in yield** reflects **differences in crop types, agricultural practices, and regional conditions**. High-yield **outliers** may represent crops with **exceptional productivity or data errors**.

2.3.6 6. Proportion of Area (`prop_area`)

- The proportion of area (`prop_area`) ranges from **0.00000 to 0.25858**, with a **mean of 0.01355**.
- The **median proportion** is **0.00176**, indicating that most crops **occupy a small fraction** of the total cultivated area, while a **few crops dominate**.
- There are **233 missing values** in this variable.

2.3.6.1 Insight: The **low median proportion** suggests that **most crops are minor contributors** to the total cultivated area, while a **few crops (e.g., rice, wheat) likely dominate**. **Missing values** should be handled appropriately.

2.4 Handle Missing Values

2.4.1 Count Missing Values

```
sum(is.na(contry_apy))  # Total number of missing values
```

```
## [1] 751
```

```
colSums(is.na(contry_apy))  # Missing values per column
```

```
##       year       crop       area production      yield  prop_area
##          0          0        233         51        234        233
```

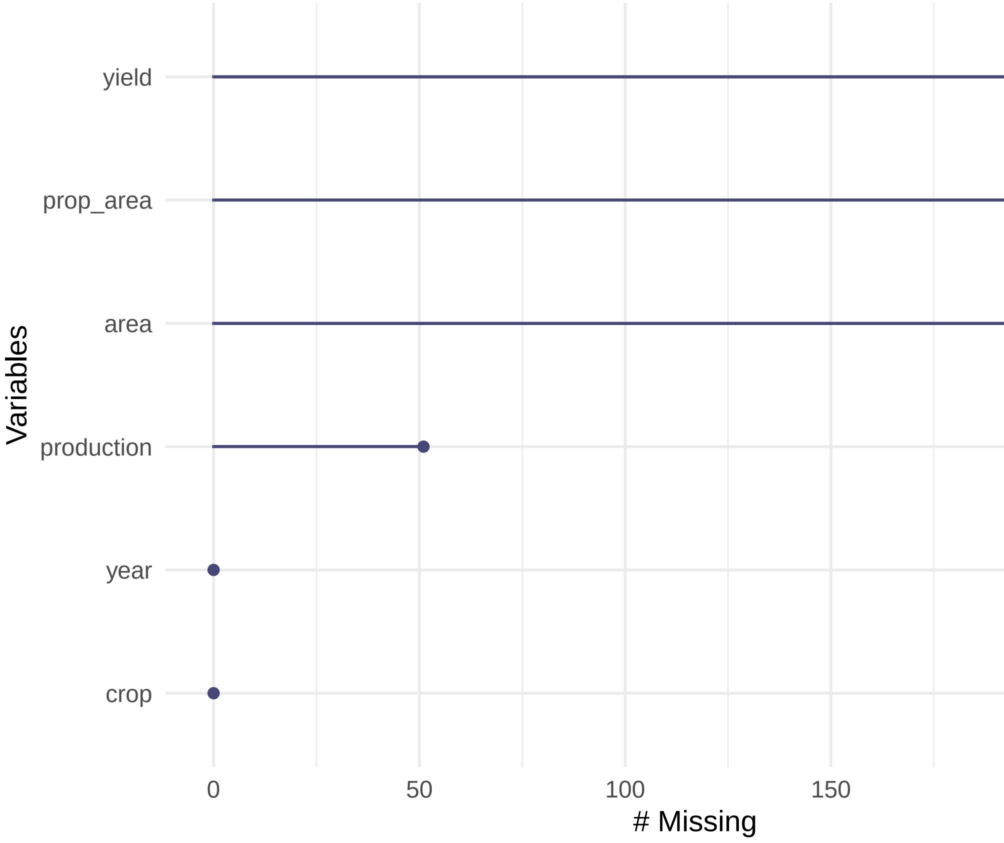
2.4.2 Detect Duplicates

```
sum(duplicated(contry_apy))  # Count duplicate rows
```

```
## [1] 0
```

No duplicated row.

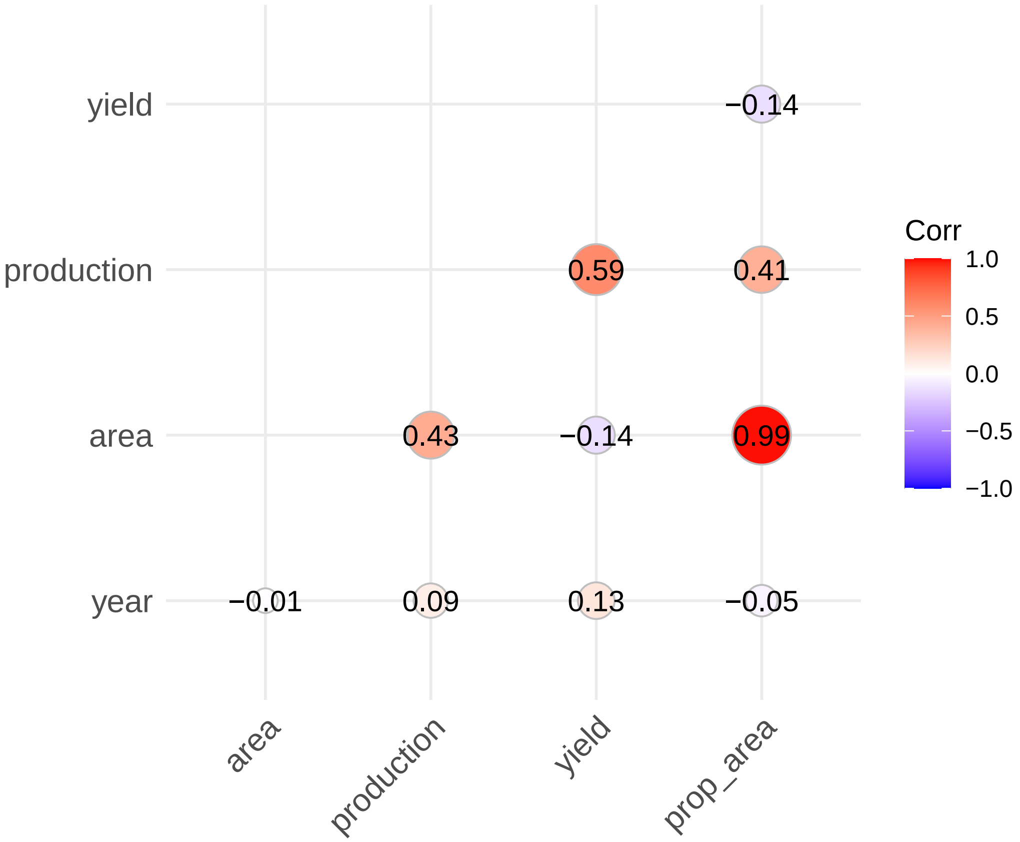
```
library(ggplot2)
library(naniar)

gg_miss_var(contry_apy)  # Plot missing values per column
```



2.4.2.1 Visualize Missing Data

Some data are not available. We have to take it in account in ours next step by inputing or dropping off some of them based on quantity of i,formation we are about to lose. ## Analysis of the Correlation Graph

```
library(ggcorrplot)  # Install if needed

# Compute correlation matrix
cor_matrix <- cor(contry_apy[, sapply(contry_apy, is.numeric)], use = "complete.obs")

# Plot with ggcorrplot
ggcorrplot(cor_matrix, method = "circle", type = "lower", lab = TRUE)
```



2.4.2.2 Key Observations:

2.4.2.2.1 1. Strongest Positive Correlation:

- **Area & Proportion of Area (0.99, Red Circle)**
  - Nearly perfect positive correlation, meaning that as the **cultivated area** increases, the **proportion of total cultivated area** occupied by a specific crop also increases.

2.4.2.3 2. Moderate Positive Correlations:

- **Production & Area (0.43, Red Circle)**
  - Larger cultivated areas tend to produce more crops, which is expected.
- **Production & Yield (0.59, Red Circle)**
  - Higher yields are moderately associated with **higher production**.

2.4.2.4 3. Weak/Negligible Correlations:

- **Year & All Variables (~ -0.01 to 0.13, Small Circles)**
  - Year does not significantly influence area, production, or yield trends.
- **Yield & Area (-0.14, Light Purple Circle)**
  - Weak negative correlation, indicating that **yield does not strongly depend on cultivated area**.

2.4.2.5 4. Color Interpretation:

- **Red Circles (Positive Correlation):** Strong relationships where **both variables increase together**.
- **Blue/Purple Circles (Negative Correlation):** Weak inverse relationships.

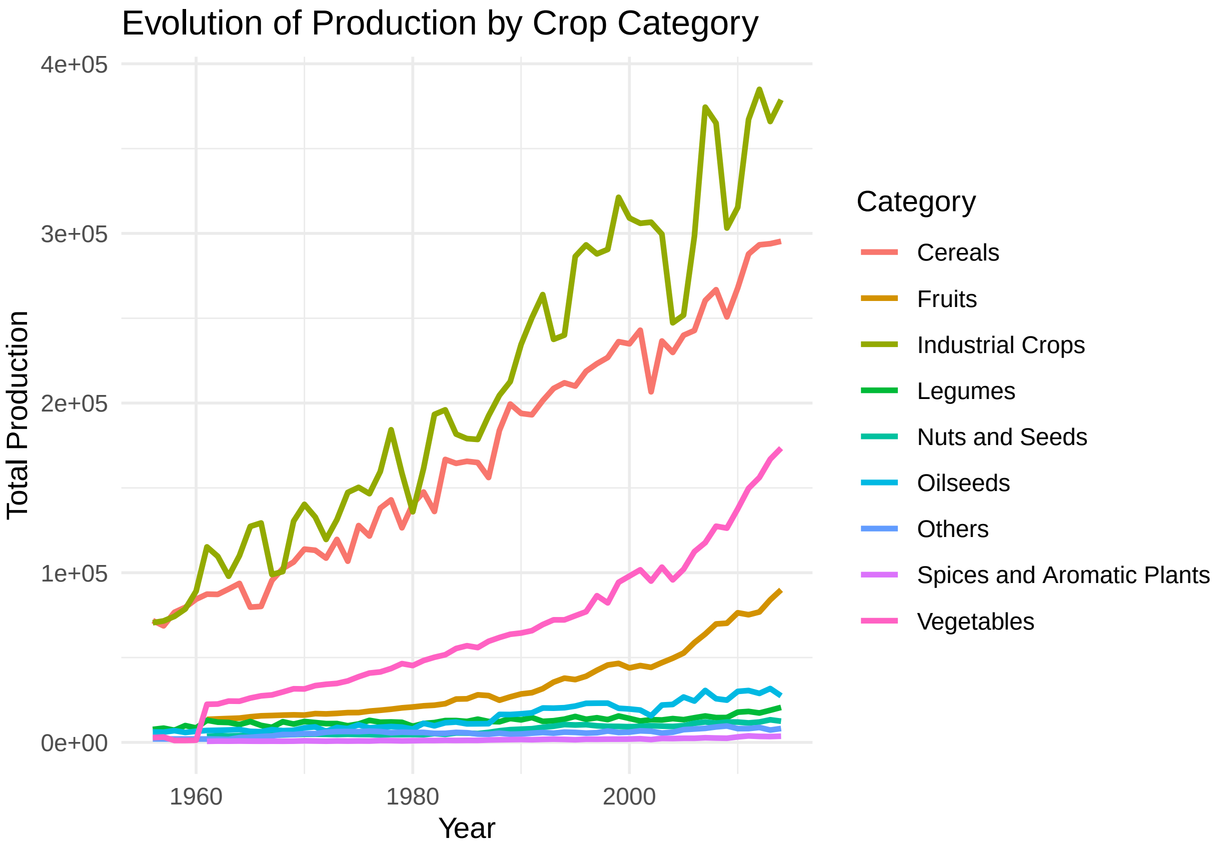
2.4.2.6 Conclusion:

- The most important relationship is between **Area and Proportion of Area** (almost perfect correlation).
- **Production is moderately dependent on both Area and Yield**, but not strongly.
- **Year has minimal impact on other variables**, suggesting **stability in trends over time**.

2.5 Distribution of Crops by Cultivated Area

```
# Load the packages
library(tidyverse)
library(ggplot2)

# Add a "category" column to the data
data <- contry_apy %>%
  mutate(category = case_when(
    # Cereals
    crop %in% c("Barley", "Maize", "Millet", "Rice", "Sorghum", "Wheat") ~ "Cereals",

    # Legumes
    crop %in% c("Chickpea", "Dry beans", "Dry peas", "Lentils", "Pigeonpea", "Beans, dry", "Beans, gree

    # Oilseeds
    crop %in% c("Groundnut", "Linseed", "Rape and mustard", "Sesamum", "Soybean", "Sunflower", "Safflow

    # Fruits
    crop %in% c("Apples", "Apricots", "Bananas", "Grapes", "Mangoes, mangosteens, guavas", "Oranges", "

    # Vegetables
    crop %in% c("Potato", "Sweet potatoes", "Tomatoes", "Onions", "Garlic", "Carrots and turnips", "Cuc

    # Industrial Crops
    crop %in% c("Cotton", "Jute", "Sugarcane", "Tobacco, unmanufactured", "Rubber, natural", "Tea", "Co

    # Spices and Aromatic Plants
    crop %in% c("Chillies and peppers, dry", "Chillies and peppers, green", "Ginger", "Pepper (piper sp

    # Nuts and Seeds
    crop %in% c("Cashew nuts, with shell", "Walnuts, with shell", "Areca nuts", "Coconuts") ~ "Nuts and

    # Others
    TRUE ~ "Others"
  ))

# Aggregate data by category and year
category_data <- data %>%
  group_by(year, category) %>%
  summarise(total_production = sum(production, na.rm = TRUE),
            total_area = sum(area, na.rm = TRUE),
            mean_yield = mean(yield, na.rm = TRUE))

# Plot the evolution of production by category
```

```
ggplot(category_data, aes(x = year, y = total_production, color = category)) +
  geom_line(size = 1) +
  labs(title = "Evolution of Production by Crop Category",
       x = "Year",
       y = "Total Production",
       color = "Category") +
  theme_minimal()
```



2.5.0.1 Key Observations:

2.5.0.1.1 Dominance of Cereals:

- **Cereals (e.g., rice, wheat, maize)** are likely the dominant category in terms of total production.
- This reflects their importance as staple foods and their widespread cultivation across regions.
- Insight: Cereals play a critical role in food security, and their production trends are a key indicator of agricultural performance.

2.5.0.2 Growth in Oilseeds and Industrial Crops:

- **Oilseeds (e.g., soybean, sunflower) and Industrial Crops (e.g., cotton, sugarcane)** may show significant growth in production over time.
- This growth is likely driven by increasing demand for edible oils, biofuels, and industrial raw materials.

-Insight: The rise in oilseeds and industrial crops reflects shifting market demands and economic priorities.

2.5.0.3 Variability in Fruits and Vegetables:

- **Fruits and Vegetables** may exhibit more variability in production compared to other categories.

-This variability can be attributed to factors such as seasonal fluctuations, climate conditions, and market demand.

-Insight: Fruits and vegetables are sensitive to environmental and market changes, making them more vulnerable to disruptions.

2.5.0.4 Long-Term Trends:

- The visualization likely shows long-term trends in production, such as:

-Steady growth in Cereals and Oilseeds.

-Fluctuations in Fruits and Vegetables.

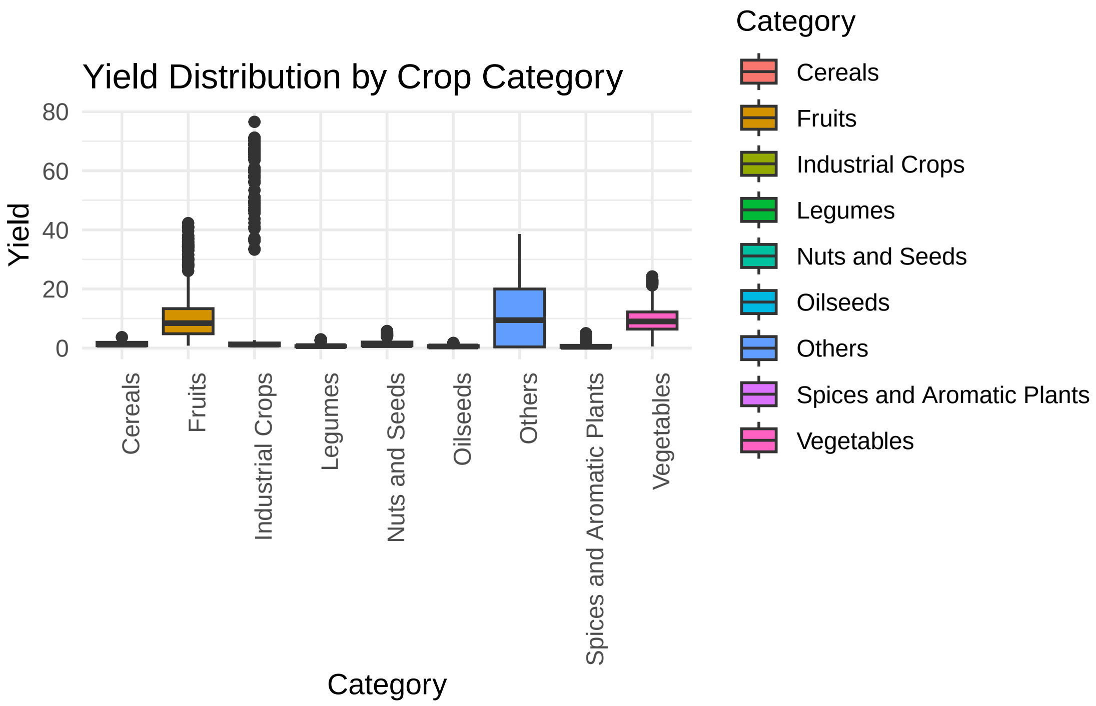
-Stable or gradual growth in Spices and Nuts.

-Insight: Long-term trends highlight the impact of technological advancements, policy changes, and climate conditions on agricultural production.

2.5.0.5 Conclusion: The visualization provides a comprehensive overview of production trends by crop category over time. It highlights the dominance of Cereals and Oilseeds, as well as the variability in Fruits and Vegetables. Further analysis can deepen our understanding of the impact of climate hazards and inform strategies for sustainable agricultural development.

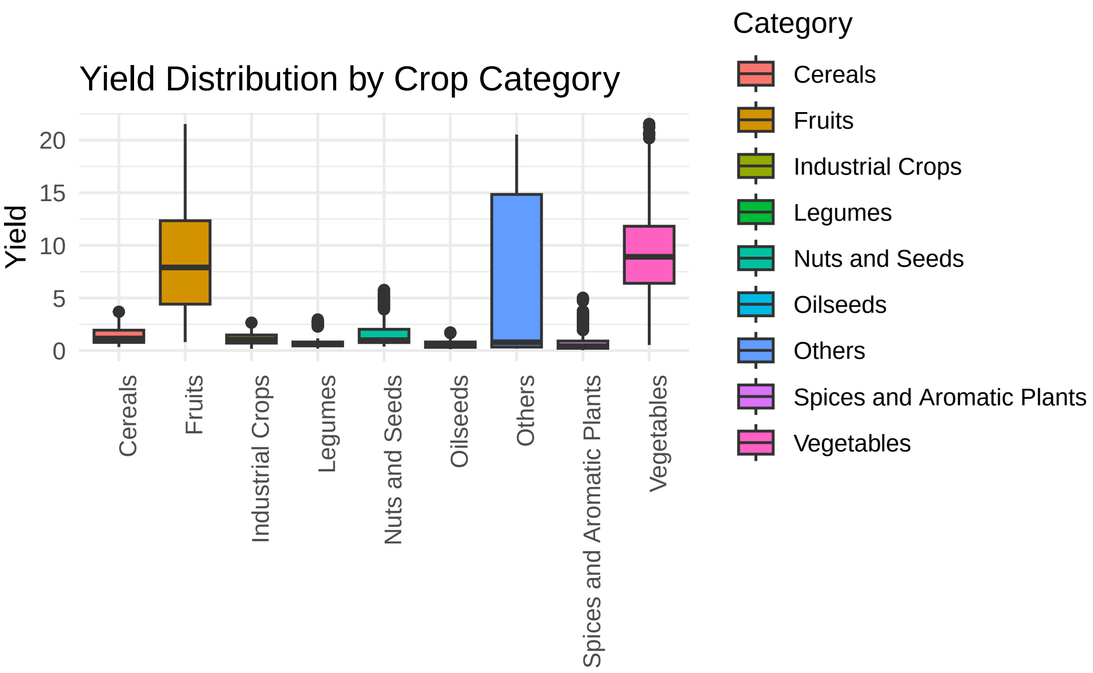
2.6 Analysis of Yield Distribution by Crop Category

```r
# Plot the distribution of yield by category
ggplot(data, aes(x = category, y = yield, fill = category)) +
  geom_boxplot() +
  labs(title = "Yield Distribution by Crop Category",
       x = "Category",
       y = "Yield",
       fill = "Category") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +  # Rotate x-axis labels for better readabil
  theme(plot.margin = unit(c(1, 1, 2, 1), "cm"))  # Adjust margin for better spacing
```



We should remove the outliers data from the dataset to make accurate analysis using the boxplots.

```
# Calculate IQR for yield
Q1 <- quantile(data$yield, 0.25, na.rm = TRUE)
Q3 <- quantile(data$yield, 0.75, na.rm = TRUE)
IQR_value <- Q3 - Q1

# Define the lower and upper bounds for outliers
lower_bound <- Q1 - 1.5 * IQR_value
upper_bound <- Q3 + 1.5 * IQR_value

# Filter out outliers
data_no_outliers <- data %>%
  filter(yield >= lower_bound & yield <= upper_bound)

# Plot the distribution of yield by category without outliers
ggplot(data_no_outliers, aes(x = category, y = yield, fill = category)) +
  geom_boxplot() +
  labs(title = "Yield Distribution by Crop Category",
       x = "Category",
       y = "Yield",
       fill = "Category") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  theme(plot.margin = unit(c(1, 1, 2, 1), "cm"))  # Adjust margin for better spacing
```



2.6.0.1 1. Dominance of High-Yield Categories

- **Fruits and Vegetables** are likely to have the **highest yields** among crop categories.
- These crops (e.g., **tomatoes, bananas, potatoes**) benefit from **intensive cultivation practices, high inputs** (e.g., irrigation, fertilizers), and **shorter growing cycles**.
- **Cereals** (e.g., rice, wheat) and **Oilseeds** (e.g., soybean, sunflower) may show **moderate yields** due to their **larger cultivation areas** and **dependence on rainfall** in some regions.

Insight:
High-yield crops like **fruits and vegetables** are critical for **diversified diets and high economic returns**, but they may require **significant resource inputs and management**.

2.6.0.2 2. Moderate Yields in Cereals and Oilseeds

- **Cereals and Oilseeds** have **moderate yields** compared to fruits and vegetables.
- These crops are **staple foods** with widespread cultivation, but their yields may be limited by factors such as **soil quality, water availability, and longer growing cycles**.
- **Industrial Crops** (e.g., sugarcane, cotton) may also show **moderate yields** due to their **economic importance** and **intensive cultivation methods**.

Insight:
Cereals and oilseeds remain **critical for food security and economic stability**, but their moderate yields highlight the need for **improved agricultural practices and climate resilience**.

2.6.0.3 3. Lower Yields in Legumes, Nuts, and Spices

- **Legumes** (e.g., chickpeas, lentils), **Nuts**, and **Spices** tend to have **lower yields**.
- These crops are often grown in **marginal lands with fewer inputs** (e.g., fertilizers, irrigation) and are **more susceptible to climate hazards** such as droughts and pests.

Insight:
Legumes and spices are **important for crop rotation, soil health, and niche markets**, but their productivity could improve with **better agronomic practices and targeted support**.

2.6.0.4 4. Variability in Yield Distribution

- Yield distributions **within each category vary significantly**.
- Some varieties of **Fruits and Vegetables** (e.g., tomatoes, grapes) may have **exceptionally high yields**, while others (e.g., leafy greens, root vegetables) may show **moderate yields**.
- Similarly, **Cereals and Oilseeds** may exhibit **wide yield variations** depending on the crop type, region, and management practices.

Insight:
Yield variability highlights the importance of adopting **best practices** such as **improved seeds, irrigation, and pest management** to maximize productivity.

2.6.0.5 5. Regional and Crop-Specific Patterns

- The **yield distribution** may reflect **regional patterns**, such as:
  - **High yields** of **Fruits and Vegetables** in regions with **favorable climates** (e.g., **Maharashtra for grapes, Punjab for potatoes**).
  - **Moderate yields** of **Cereals and Oilseeds** in **irrigated regions** (e.g., **Punjab, Haryana**).
  - **Lower yields** of **Legumes and Spices** in **rain-fed or marginal areas**.

Insight:
Regional variations in yields underscore the need for **location-specific agricultural strategies and investments**.

2.6.0.6 6. Implications for Climate Resilience

- **Crops with higher yields** (e.g., fruits, vegetables) may still face risks from **climate hazards** such as **extreme temperatures, pests, and water scarcity**.
- **Lower-yield crops** (e.g., legumes, spices) are particularly **vulnerable to climate hazards** and may require **targeted support** to enhance resilience.

Insight:
Enhancing **climate resilience** through **drought-tolerant varieties, water-efficient irrigation, and integrated pest management** is critical for **sustaining yields** across all crop categories.

2.6.0.7 7. Recommendations for Further Analysis

- **Crop-Specific Yield Trends:**
  - Analyze **yield trends** for individual crops (e.g., tomatoes, potatoes, rice, wheat).
- **Regional Yield Analysis:**
  - Identify areas with the **highest and lowest yields** for specific crops.
- **Impact of Agricultural Practices:**
  - Investigate the role of **irrigation, fertilizers, and crop varieties** in determining yields.
- **Climate Hazard Correlation:**
  - Assess the impact of **droughts, floods, and climate hazards** on productivity, particularly for **high-yield crops** like fruits and vegetables.

2.6.0.8 Conclusion The **visualization of yield distribution by crop category** provides **valuable insights** into agricultural productivity and its variability.
- It highlights the **dominance** of **Fruits and Vegetables** in terms of yields.
- It also reveals the **challenges faced by Legumes, Nuts, and Spices**.

Further analysis can help identify strategies to:
Enhance yields
Improve resilience
Support sustainable agricultural development

3 Climate Hazard Indicators

Assessing the impact of climate on crop production involves monitoring various climate indicators. For instance, temperature and precipitation are traditional climate indicators used to assess agricultural impacts. Their effects can be complex and region-specific. The increased temperatures accelerate evapotranspiration, which leads to water stress. The changes in precipitation patterns result in droughts or floods, both adversely affecting crop production (Toté et al. 2015; Dembélé and Zwart 2016).

3.1 Calculation of a Climate Hazard Indicator: Precipitation Extremes

To calculate a climate hazard indicator such as **Precipitation Extremes** (e.g., heavy rainfall, droughts), the following steps are required:

1. **Data Collection**:
   - Collect historical rainfall data (e.g., monthly precipitation totals) and flood data from freely available datasets such as:
     - NOAA
     - CHIRPS
     - FEWS NET
     - IMD (for India, the study country) (KC et al. 2015).
2. **Risk Level Assignment**:
   - Assign a risk level for each hazard variable based on its historical data:
     - Low
     - Medium
     - High
     - Very High
3. **Weighting Hazard Variables**:
   - Weight the hazard variables based on their relative importance.
4. **Climate Hazard Indicator (CHI) Calculation**:
   - Calculate the Climate Hazard Indicator (CHI) using the formula:

     $$CHI = \sum_{i=1}^{n} (W_i \times H_i)$$

     Where:
     - W_i is the weight assigned to each hazard variable i,
     - H_i is the risk level (1 to 4) for that hazard variable.
5. **Interpretation of Results**:
   - A higher CHI score indicates a greater level of climate hazard for a region. Regions can be categorized based on their CHI scores:
     - **Low risk**: $CHI < 10$
     - **Moderate risk**: $10 \leq CHI < 20$
     - **High risk**: $CHI \geq 20$

4 Modeling and Analysis

4.1 Problem Definition

• Perform statistical analysis to understand the relationship between the climate hazard indicator(s) and crop production at the sub-national level.

The goal is to predict crop yield based on:

- **Climate variables**: Consecutive dry days (CDD), cumulative rainfall, temperature extremes, etc.
- **Crop-specific variables**: Crop type, cultivated area, region, etc.

4.2 Methodology and Simulation Models for Yield Prediction Based on Climate and Agricultural Variables

Before diving into the methodology and models, it's important to note that **all data must be properly prepared**. This includes cleaning, handling missing values, aggregating climate data (e.g., monthly averages), and creating derived variables such as rainfall anomalies, temperature extremes, and drought indices. Once the data is ready, we can proceed with the analysis and modeling.

4.3 Methodology

4.3.1 Data Collection and Preparation

4.3.1.1 Data Sources:

- **Agricultural Data**: Collect data on area under cultivation, production, and yield from reliable sources such as government agricultural departments or FAO.
- **Climate Data**: Gather historical data on minimum temperature (MinTemp), maximum temperature (MaxTemp), and rainfall from meteorological departments or platforms like NOAA, CHIRPS, or IMD.
- **Extreme Events Data**: Obtain data on heavy rainfall and droughts from disaster management agencies or satellite-based platforms.

4.3.1.2 Data Preparation:

- Clean the data to handle missing values, outliers, and inconsistencies.
- Aggregate climate data to a suitable temporal scale (e.g., monthly averages for temperature, total monthly rainfall).
- Create derived variables such as:
  - **Rainfall Anomalies**: Deviation from long-term averages.
  - **Temperature Extremes**: Number of days exceeding a threshold (e.g., days with MaxTemp > 35°C).
  - **Drought Index**: Use standardized precipitation index (SPI) or other drought indices.

4.4 Model Selection

Several models can be used to simulate yield based on the variables. Below, we focus on **Linear Regression**, **Random Forest**, and **Support Vector Machines (SVM)**.

4.4.1 1. Linear Regression

- **Description**: Simple and interpretable. Models linear relationships between yield and predictors.

- **Equation**:

$$Yield = \beta_0 + \beta_1 \cdot Area + \beta_2 \cdot MinTemp + \beta_3 \cdot MaxTemp + \beta_4 \cdot Rainfall + \epsilon$$

  Where:
  - β_0 is the intercept,
  - $\beta_1, \beta_2, \beta_3, \beta_4$ are coefficients,
  - ϵ is the error term.

4.4.2 2. Random Forest

- **Description**: Handles non-linear relationships and interactions. Robust to outliers and missing data.
- **Equation**:

$$Yield = f(Area, Production, MinTemp, MaxTemp, Rainfall)$$

  Where f is a non-linear function learned by the random forest.

4.4.3 3. Support Vector Machines (SVM)

- **Description**: Effective for high-dimensional data. Can model complex relationships using kernel functions.
- **Equation**:

$$Yield = f_{SVM}(Area, Production, MinTemp, MaxTemp, Rainfall)$$

  Where f_{SVM} is a function learned by the SVM, often using a kernel (e.g., linear, radial basis function).

4.4.4 Summary of Models

Model	Strengths	Weaknesses	Use Case
Linear Regression	Simple, interpretable	Assumes linearity	Linear relationships
Random Forest	Handles non-linearities, robust	Less interpretable	Complex, non-linear relationships
SVM	Effective for high-dimensional data	Computationally expensive	High-dimensional data

5 Unified Model Calibration Methodology

Model calibration involves tuning model parameters and validating performance to ensure accurate predictions. Below is a unified methodology for calibrating **Linear Regression**, **Random Forest**, and **Support Vector Machines (SVM)**.

5.1 Methodology

5.1.1 1. Data Preparation

- **Split Data**: Divide the dataset into training (80%) and testing (20%) sets.
- **Normalize Data**: Scale numerical variables (e.g., using `scale()` in R) to ensure consistent performance across models.

5.1.2 2. Thematic mapping and Model Training

- **Thematic yield** maps per crop and per decade introduction, and interpretation.
- **Linear Regression**: Fit the model using the training data and estimate coefficients.
- **Random Forest**: Train the model using the training data and tune hyperparameters (`ntree`, `mtry`).
- **SVM**: Train the model using the training data and tune hyperparameters (`C`, `gamma`, kernel type).

5.1.3 3. Hyperparameter Tuning

- Use **cross-validation** or **grid search** to find the optimal hyperparameters for Random Forest and SVM (Hastie et al., 2009).
- For Linear Regression, no hyperparameter tuning is required, but feature selection can improve performance. (Challinor et al., 2018; Djagba et al;,2018)

5.1.4 4. Model Validation

- Evaluate the model on the testing set using metrics like **RMSE**, **MAE**, and **R^2**.(Chai et al;,2014)
- Compare the performance of different models to select the best one.

5.1.5 5. Final Model Selection

- Choose the model with the best performance on the testing set.
- Retrain the selected model on the full dataset (training + testing) for deployment.

5.2 R Code for Unified Methodology

5.2.1 Load Libraries

```
#library(caret)       # For model training and tuning
#library(ranàdomForest) # For Random Forest
#library(e1071)       # For SVM
```

---

6 References

1. **Challinor AJ, Müller C, Asseng S, et al (2018)**. Improving the use of crop models for risk assessment and climate change adaptation. *Agricultural Systems* 159:296–306. https://doi.org/10.1016/j.agsy.2017.07.010

2. **Dembélé M, Zwart SJ (2016)**. Evaluation and comparison of satellite-based rainfall products in Burkina Faso, West Africa. *International Journal of Remote Sensing* 37:3995–4014. https://doi.org/10.1080/01431161.2016.1207258

3. **Djagba JF, Sintondji LO, Kouyaté AM, et al (2018)**. Predictors determining the potential of inland valleys for rice production development in West Africa. *Applied Geography* 96:86–97. https://doi.org/10.1016/j.apgeog.2018.05.003

4. **KC B, Shepherd JM, Gaither CJ (2015)**. Climate change vulnerability assessment in Georgia. *Applied Geography* 62:62–74. https://doi.org/10.1016/j.apgeog.2015.04.007

5. **Mao X, Hu A, Wu M, et al (2023)**. Evaluation of Water Inrush Hazard in Karst Tunnel Based on Improved Non-Linear Attribute Variable Weight Recognition Model. *Applied Sciences (Switzerland)* 13:. https://doi.org/10.3390/app13085026

6. **Toté C, Patricio D, Boogaard H, et al (2015)**. Evaluation of satellite rainfall estimates for drought and flood monitoring in Mozambique. *Remote Sensing* 7:1758–1776. https://doi.org/10.3390/rs70201758

7. **Willmott, C. J., & Matsuura, K. (2005)**. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Climate Research*, 30(1), 79–82.

8. **Chai, T., & Draxler, R. R. (2014)**. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? *Geoscientific Model Development*, 7(3), 1247–1250.