Analysis of Meteorological Data Across 50 Stations In Minnesota

Isaiah Thompson Ocansey
University Of Texas at El-Paso

April 22, 2024

Abstract

This study delves into the intricate climate dynamics observed across 50 weather stations scattered throughout Minnesota. It places particular emphasis on discerning noteworthy variations in climate metrics between two distinct time frames: October to March (referred to as group I) and April to September (referred to as group II). Through the application of analytical tools such as Principal Component Analysis (PCA) and K-means clustering techniques, applied to datasets sourced from the Global Summary of the Month (GSOM) archives provided by the National Centers for Environmental Information, the research endeavors to uncover latent patterns and seasonal fluctuations embedded within historical weather records. The analysis unearthed statistically significant disparities in key climate factors, encompassing temperature, precipitation, and evaporation rates, between group I and group II time frames. These findings suggest a consistent climatic backdrop that persists throughout the entire year. The implications of these insights extend far beyond the realms of meteorology, resonating deeply within arenas such as agricultural planning, resource management, and strategic initiatives aimed at mitigating the impacts of climate change across Minnesota. Ultimately, the study highlights the remarkable resilience exhibited by the region's climate and advocates for a shift in paradigms regarding agricultural and resource management practices, especially within the specified temporal contexts.

Contents

1 Introduction

1.1 Background

Meteorological conditions profoundly impact various aspects of human life, ranging from agricultural practices and transportation systems to energy usage and public health considerations (Grossman, 2019; Smith et al., 2020). Accurate comprehension and prediction of weather patterns and climate trajectories are pivotal for making well-grounded decisions across these domains (Jones Brown, 2018; Petersen et al., 2019). Nestled in the Upper Midwest region of the United States, Minnesota witnesses a kaleidoscope of weather phenomena year-round, encompassing harsh winters and sultry summers (Johnson Williams, 2017). The state's topographical diversity, characterized by its numerous lakes and forests, contributes to the complexity of its climatic intricacies (Davis Wilson, 2016).

1.2 Objectives

1. The primary objective of this study is to analyze historical weather data obtained from selected weather stations in Minnesota, focusing on parameters such as average monthly temperature, snow, precipitation, and evaporation rates, to gain insights into long-term climatic conditions and trends within the state.

2. An additional aim is to discern trends and patterns in weather variables over time and across diverse geographical regions within Minnesota. Through an examination of temporal and spatial fluctuations in weather phenomena, encompassing temperature shifts and precipitation dynamics, the goal is to elucidate the fundamental factors driving climate variability across the state.

1.3 Data

A 10-year data spanning from July 1, 2013, to July 31, 2013, data was acquired from the Global Summary of the Month (GSOM) archives, which are part of the extensive Global Historical Climatology Network-Daily (GHCN-Daily) dataset, made available by the National Centers for Environmental Information. These GSOM data files compile a thorough monthly overview comprising more than 50 meteorological indicators sourced from weather stations worldwide

1.4 Variables Explained

- **ADPT (Average Dew Point Temperature)**: Monthly average of daily dew point temperatures, available in degrees Celsius or Fahrenheit. Missing data protocols apply based on consecutive missing days.

- **ASLP (Average Sea Level Pressure)**: Monthly average of daily sea level pressures reported in hectopascals or inches of mercury. Missing values are flagged based on specific criteria.

- **ASTP (Average Station Level Pressure)**: Similar to ASLP but taken at the station level.

- **AWBT (Average Wet Bulb Temperature)**: Average of daily wet bulb temperatures. This variable follows the same missing data rules as other temperature measurements.
- **AWND (Average Wind Speed)**: Reported in miles per hour or meters per second, this is the monthly average of daily wind speeds.
- **PRCP (Total Monthly Precipitation)**: Total precipitation received over a month, measured in inches or millimeters.
- **SNOW (Total Monthly Snowfall)**: Total snowfall amount received in a month, also measured in inches or millimeters.
- **TAVG (Average Monthly Temperature)**: The average of the unrounded daily maximum and minimum temperatures divided by two.
- **TMAX (Monthly Maximum Temperature)**: Average of daily maximum temperatures.
- **TMIN (Monthly Minimum Temperature)**: Average of daily minimum temperatures.

Each variable is accompanied by attributes that indicate the quality, source, and presence of any measurement flags or missing data. These attributes ensure the reliability and comprehensiveness of the data used for climatological analysis and research.

subparagraphData Accessibility

The data files are available for public access and can be downloaded from the National Centers for Environmental Information (NCEI) website. Interested parties can retrieve these files to conduct detailed climatic analyses over monthly periods.

subparagraphFile Structure and Data Integrity

Each file within the GSOM dataset is structured with a specific naming convention that includes the GHCN station ID, country code, and station network code. The files contain detailed records for each meteorological element, accompanied by attributes that provide insight into the measurement's integrity. These attributes include measurement flags, quality flags, source codes, and indicators of any data missing or flagged within the dataset.

2 Limitations of the Study

This study, while providing valuable insights into the climate variability across 50 weather stations in Minnesota, encounters several limitations that need to be acknowledged:

2.1 Geographical Coverage

- **Regional Limitations:** The study is confined to Minnesota, and the findings may not be applicable to other regions with different climatic conditions.

2.2 Data Collection Constraints

- **Selection Bias:** The choice of weather stations may not represent all climatic zones of the state evenly, potentially skewing the results.

2.3 Methodological Constraints

- **Statistical Assumptions:** Methods like PCA and k-means clustering assume linearity and normality, which may not hold across complex climatic data sets.

2.4 Temporal Resolution

- **Seasonal Analysis Limitations:** By dividing the year into two broad periods, the study may overlook more subtle intra-seasonal variations that could be significant.

3 Descriptive Analysis

This section of the report delves into the descriptive analysis conducted using R, focusing on historical weather data collected from 50 selected stations across Minnesota.

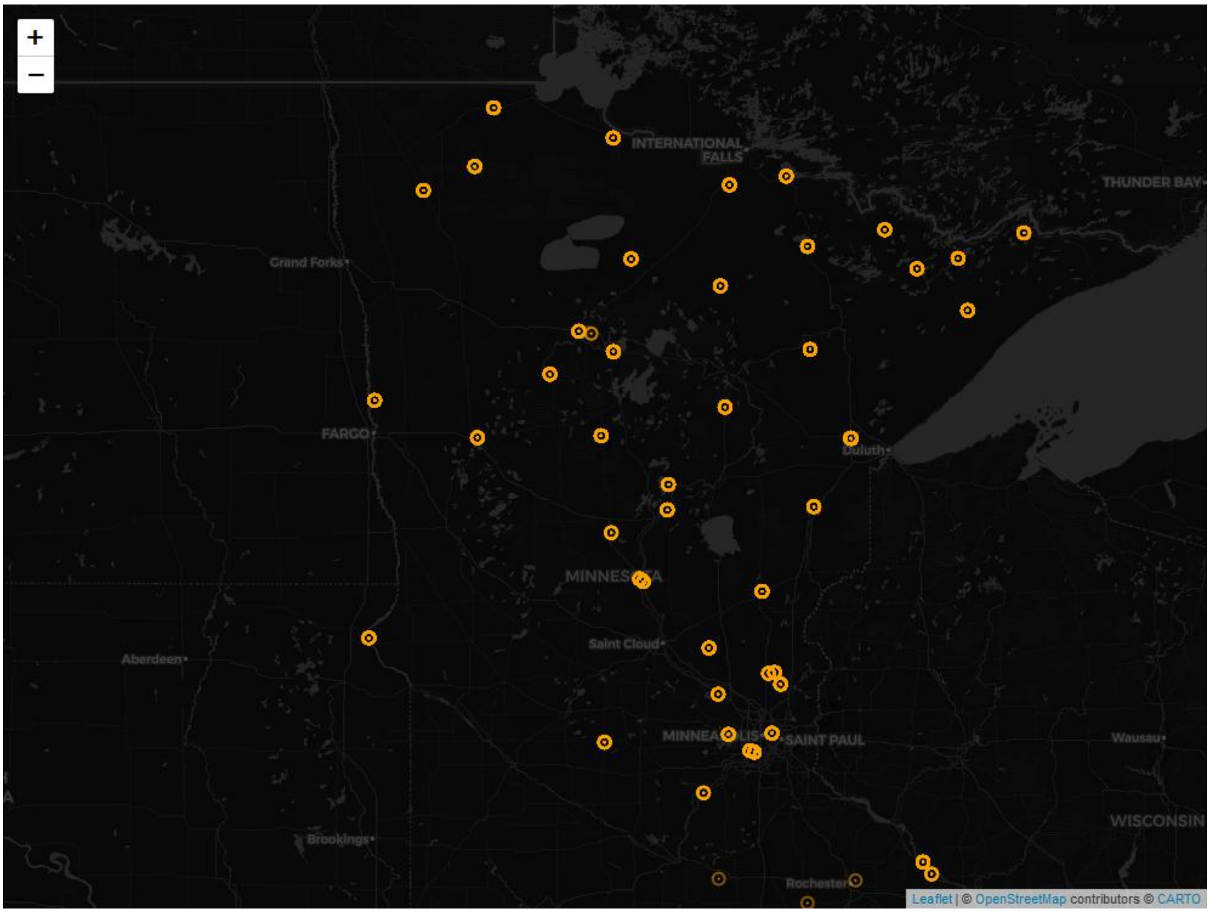


Figure 1: Distribution of 50 Weather Stations Across Minnesota

3.1 Weather Stations

Figure 1 illustrates the spatial arrangement of 50 weather stations throughout Minnesota. Each orange circle denotes the position of a weather station, pinpointed according to its latitude and longitude coordinates. The stations are strategically dispersed across the state, reflecting a well-established network aimed at gathering diverse climatic data from various geographic areas.

3.2 Analysis of Missing Data

Within our dataset, the integrity of the records holds paramount importance to ensure the precision of our meteorological assessment. The subsequent visuals depict the extent of missing data across multiple variables sourced from 50 weather stations.

3.2.1 Missing Data Proportions by Variable

Figure 3 shows the missing data proportion for each variable in a multi-colored bar plot. The distinct colors for each bar make it easier to differentiate between variables, highlighting the extent of missing data for each.

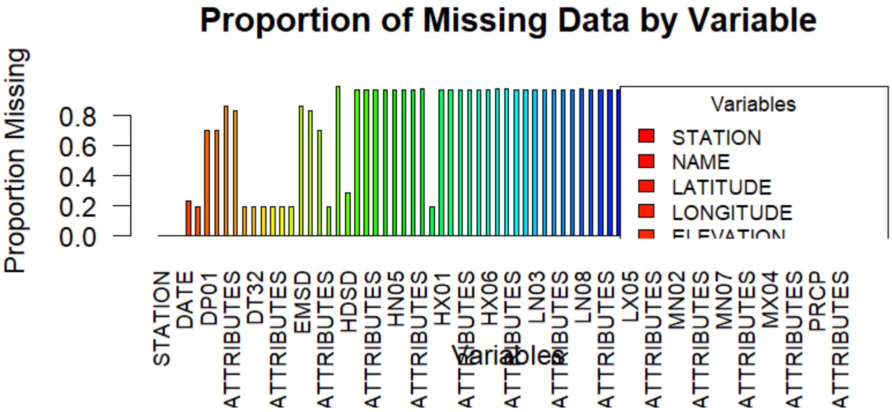


Figure 2: Proportion of Missing Data by Variable

3.2.2 Missing Data Proportions in the Dataset

Figure 3 illustrates the distribution of missing data points across the dataset, revealing that 38.4% of the data is missing with 61.6% present. This substantial proportion warrants careful attention during data preprocessing. Darker bands in the visualization indicate variables with consistently missing data across many observations, hinting at potential issues in data collection for those variables. Conversely, lighter areas indicate a higher presence of data, with some variables exhibiting complete data across all observations. Addressing this missing data appropriately through imputation methods or considering the removal of variables with excessive missingness is crucial to ensure the robustness of subsequent analyses.

After imputation, a re-evaluation of the dataset reveals a reduction in missing data, resulting in a more complete dataset for analysis.

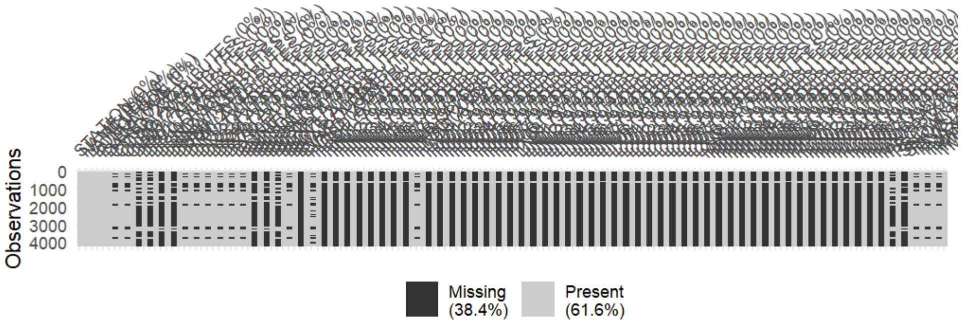


Figure 3: Proportion of Missing Data

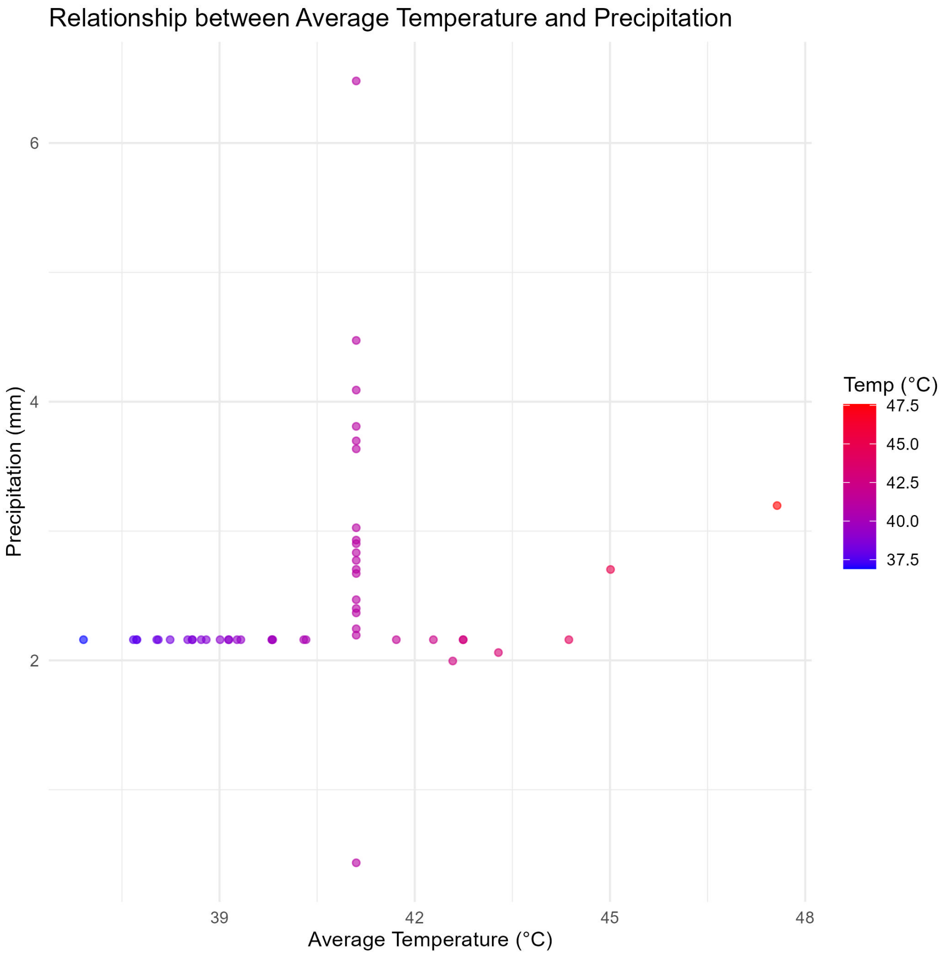
3.2.3 Interpretation of Missing Data Plots

Understanding the dataset's reliability hinges on visualizing missing data. It's evident that several variables exhibit substantial missingness, potentially impacting the dataset's integrity and subsequent analyses. Variables with fewer missing values are deemed more reliable. To address this, median imputation was employed, replacing missing values with the median of each variable, a method less sensitive to outliers than mean imputation.

Prior to imputation, careful consideration was given to each variable's nature. Variables crucial for weather forecasts, like temperature and precipitation, were handled cautiously during imputation to prevent inaccurate predictions stemming from erroneous data replacement.

3.3 Exploratory Data Analysis

3.3.1 Bivariate Analysis



[ht]

Figure 4: Scatter plot illustrating the relationship between total average temperature and precipitation.

The scatter plot depicted in Figure 4 illustrates the correlation between total average temperature and precipitation. Each data point on the plot signifies a pairing of temperature and precipitation values, with color coding denoting the magnitude of temperature. The visualization unveils a broad spectrum of precipitation levels amidst a more confined range of elevated temperatures. Despite this diversity, there's no evident pattern that readily elucidates the intricate interplay between these variables. Noteworthy is the clustering of data points around intermediate temperature values, indicating fluctuating precipitation levels within this span. The absence of a discernible trend suggests that temperature in isolation might not serve as a robust predictor of precipitation outcomes within this dataset.

3.3.2 Multivariate Analysis

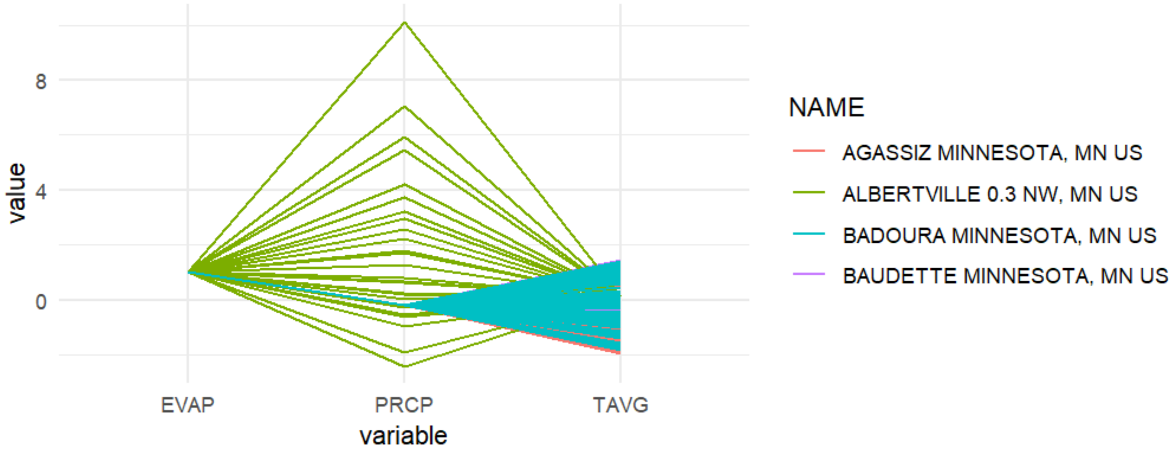


Figure 5: Parallel coordinates plot for evaporation, precipitation, and average temperature across four weather stations.

The parallel coordinates plot shown in Figure 9 shows a graphical representation of the interrelation between evaporation, precipitation, and total average temperature across four distinct weather stations. In this visualization, each axis corresponds to a specific variable, and lines extend across these axes to represent the values of each variable for individual observations. We can observe a high precipitation with quite a moderate total average temperature at Albertville.

3.3.3 Time Series

Figure 6 depicts a time series plot illustrating the three variables: evaporation, precipitation, and total average temperature.

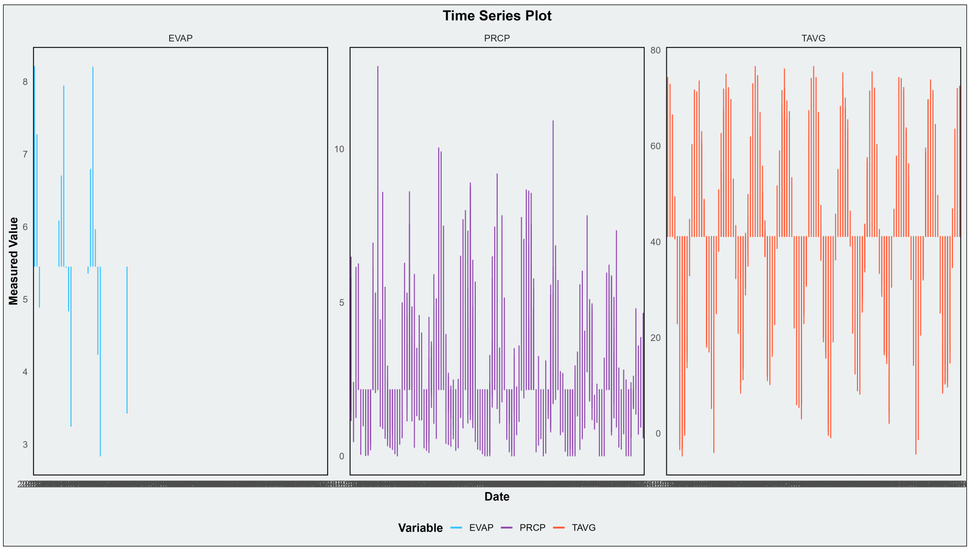


Figure 6: Time Series plot.

4 Hypothesis Testing

This section focuses on testing the hypotheses. This is to determine whether the observed patterns have statistical significance, which could validate the trends identified in the geographical distribution of temperature readings.

4.1 Normality test

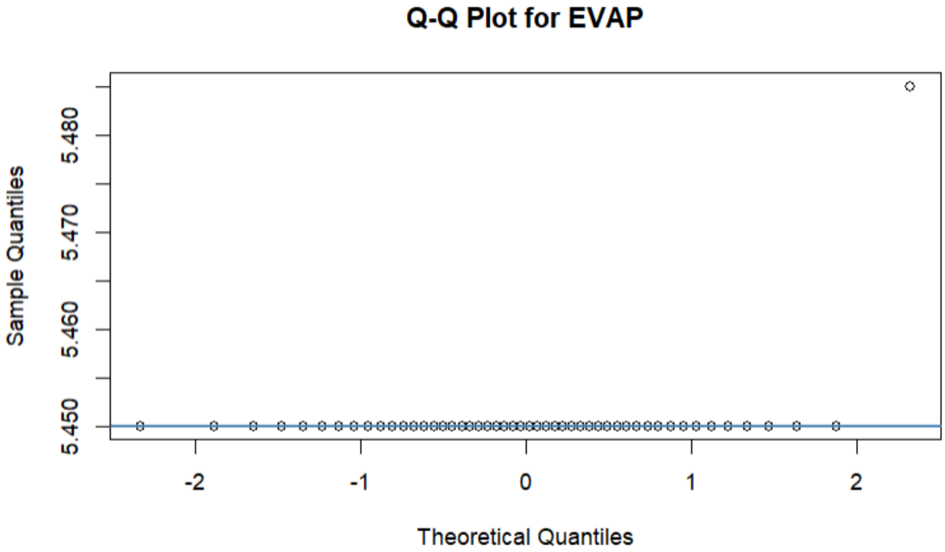


Figure 7: Parallel coordinates plot for evaporation, precipitation, and average temperature across four weather stations.

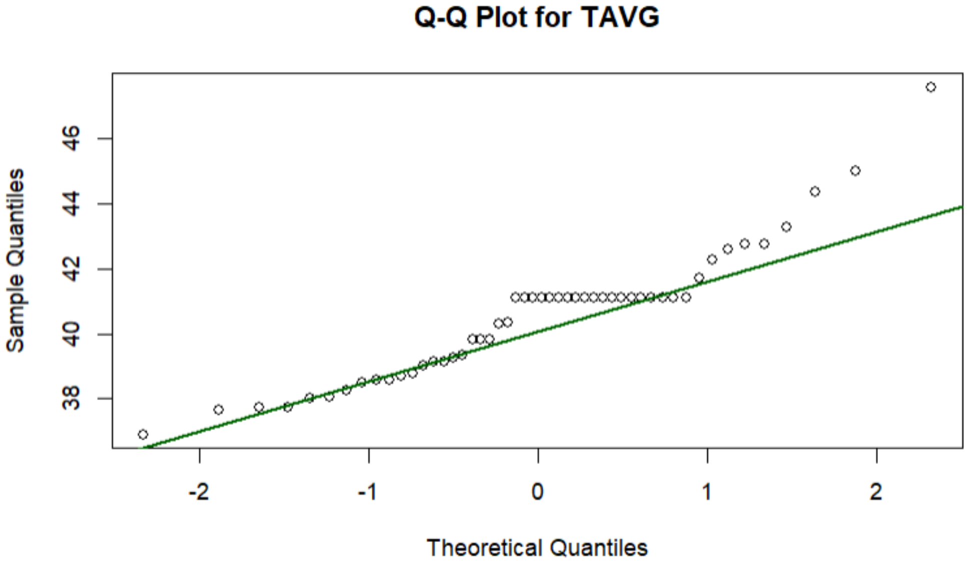


Figure 8: Parallel coordinates plot for evaporation, precipitation, and average temperature across four weather stations.

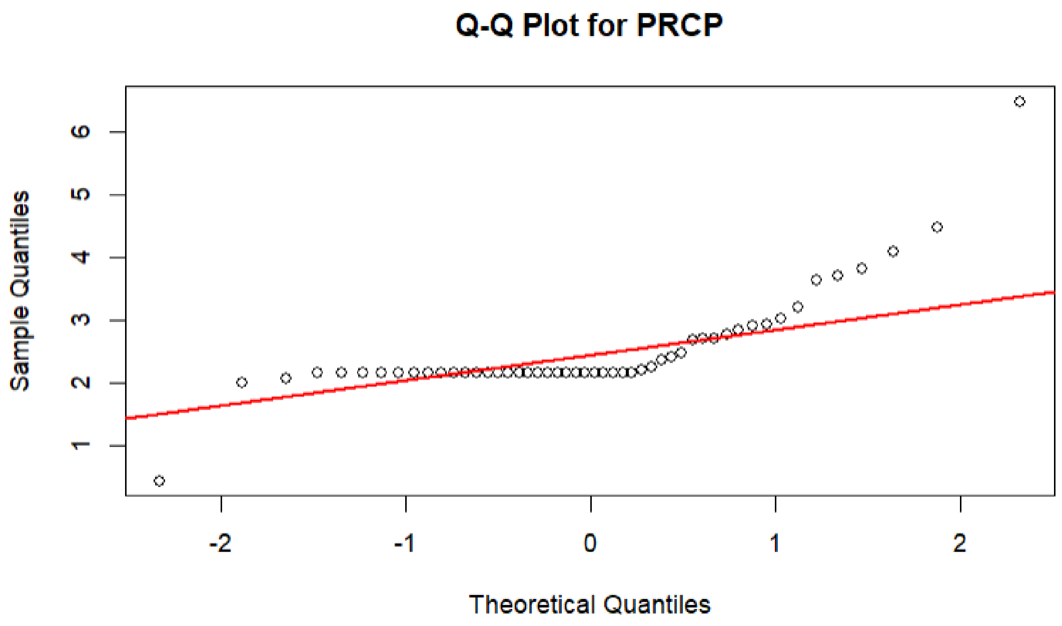


Figure 9: Parallel coordinates plot for evaporation, precipitation, and average temperature across four weather stations.

4.2 Methodology

To evaluate the statistical significance of climate variable disparities between the colder period (October to March) and the warmer season (April to September), we utilize Hotelling's T-squared test, assuming the approximate normality of the three variables. This method is well-suited for analyzing multiple climate metrics concurrently, taking into account their interrelationships.

4.3 Test Setup

The null hypothesis asserts that there is no notable distinction in the average climate metrics (i.e total average temperature, precipitation and evaporation levels) between the two time frames. Conversely, the alternative hypothesis proposes that significant differences exist, potentially signaling shifts in climate or alterations in weather patterns.

4.4 Results and Discussion

Detailed results of the statistical tests will be presented here, including T-squared values, degrees of freedom, and p-values, which will help us determine whether to reject or fail to reject the null hypothesis.

Statistic	Value	Significance
T-squared	6325.4	
p-value	0	significant

Table 1: Hotelling's T-squared test results for climate variables between October-March and April-September.

The findings reveal a statistically significant contrast in climate variables between the two periods, evidenced by a p-value below the conventional alpha threshold of 0.05. Consequently, the null hypothesis is rejected, indicating that the observed disparities in climate variables across the two time frames are indeed statistically meaningful.

5 Data Analysis Techniques

5.1 Principal Component Analysis (PCA)

Principal Component Analysis is used to reduce the dimensionality of the dataset's while retaining the most significant variance, which simplifies the complexity in multidimensional data and highlights patterns.

5.1.1 Original Distribution of Weather Data

Figure 10: Box plot showing the original distribution of variables across various weather stations.

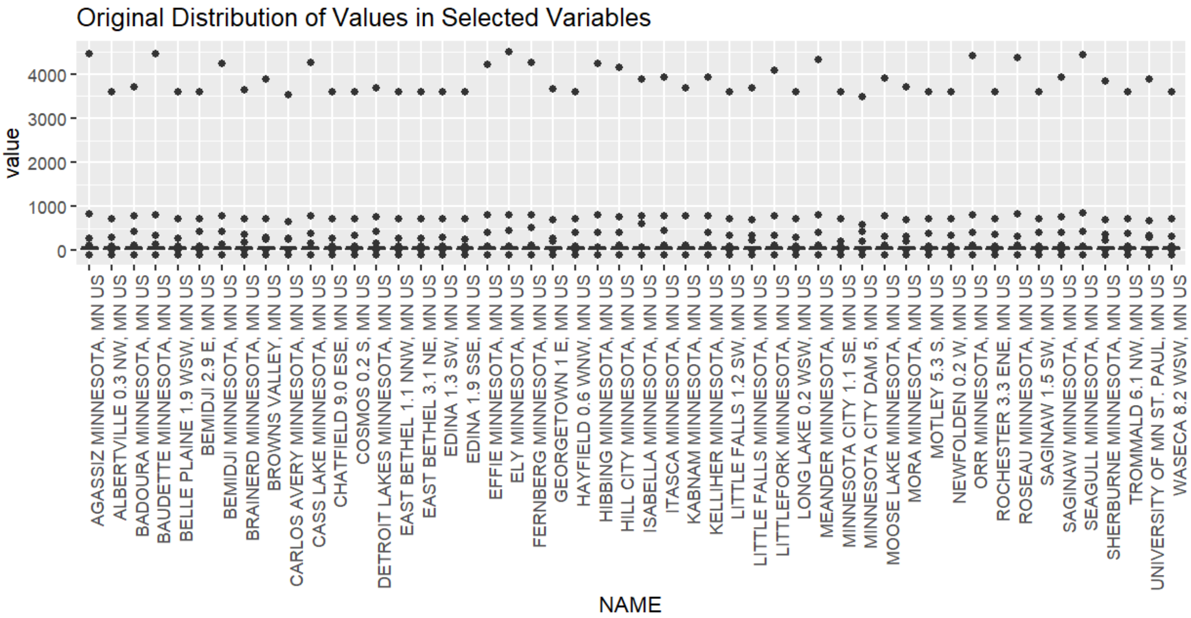


Figure 10 depicts the distribution of variables across various weather stations. Notably, outliers are evident in the data, potentially influencing the means of the variables. Therefore, standardizing the data is essential to mitigate the impact of outliers.

5.1.2 Distribution of Weather Data after Standardization

Figure 11: Box plot showing the original distribution of variables across various weather stations.

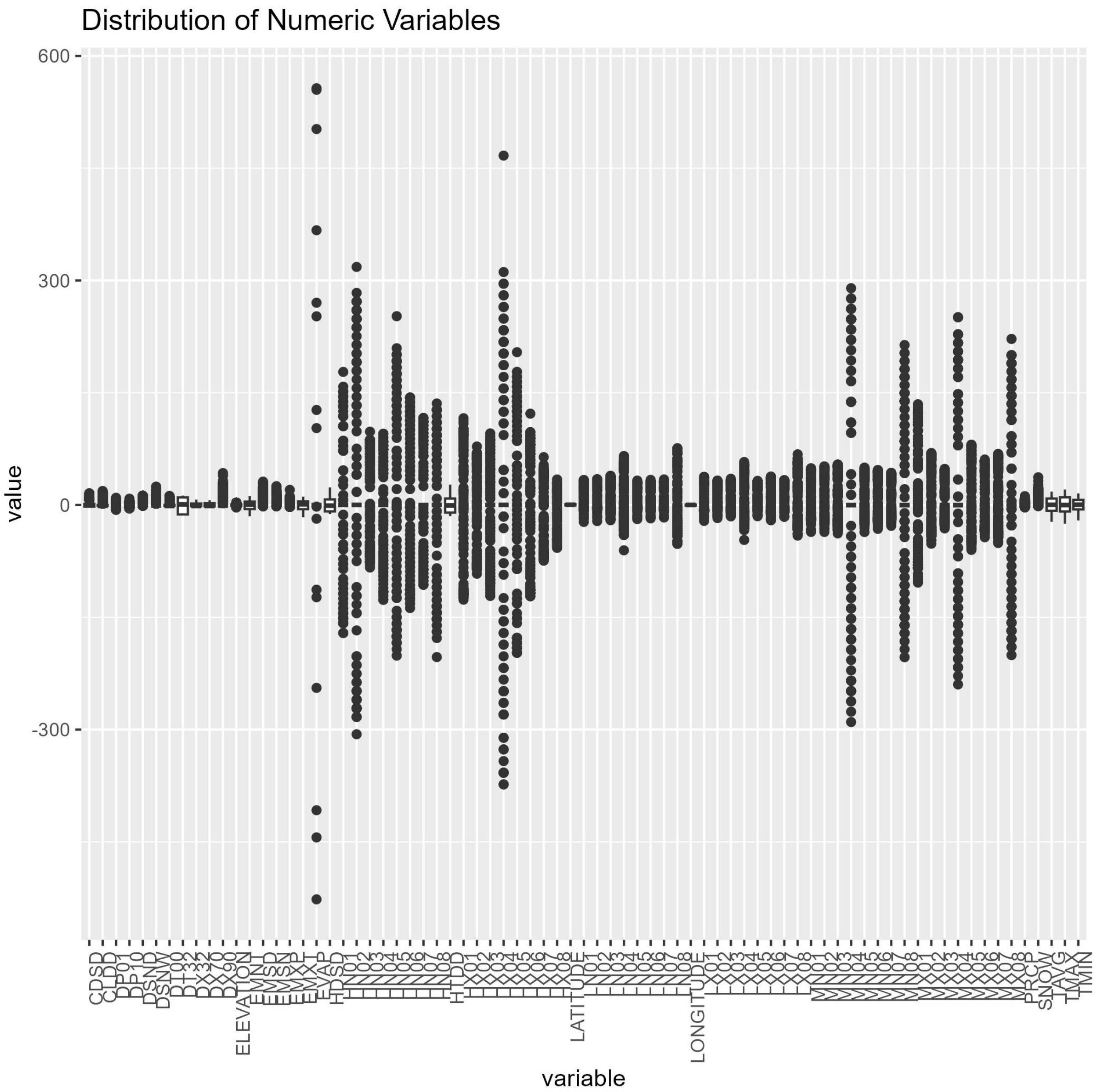


Figure 11 illustrates the distribution of variables across various weather stations after standardization. It is evident that the variability in the standardized distribution is noticeably milder compared to the original distribution.

5.1.3 Biplot

Figure 12: Biplot of the Variables

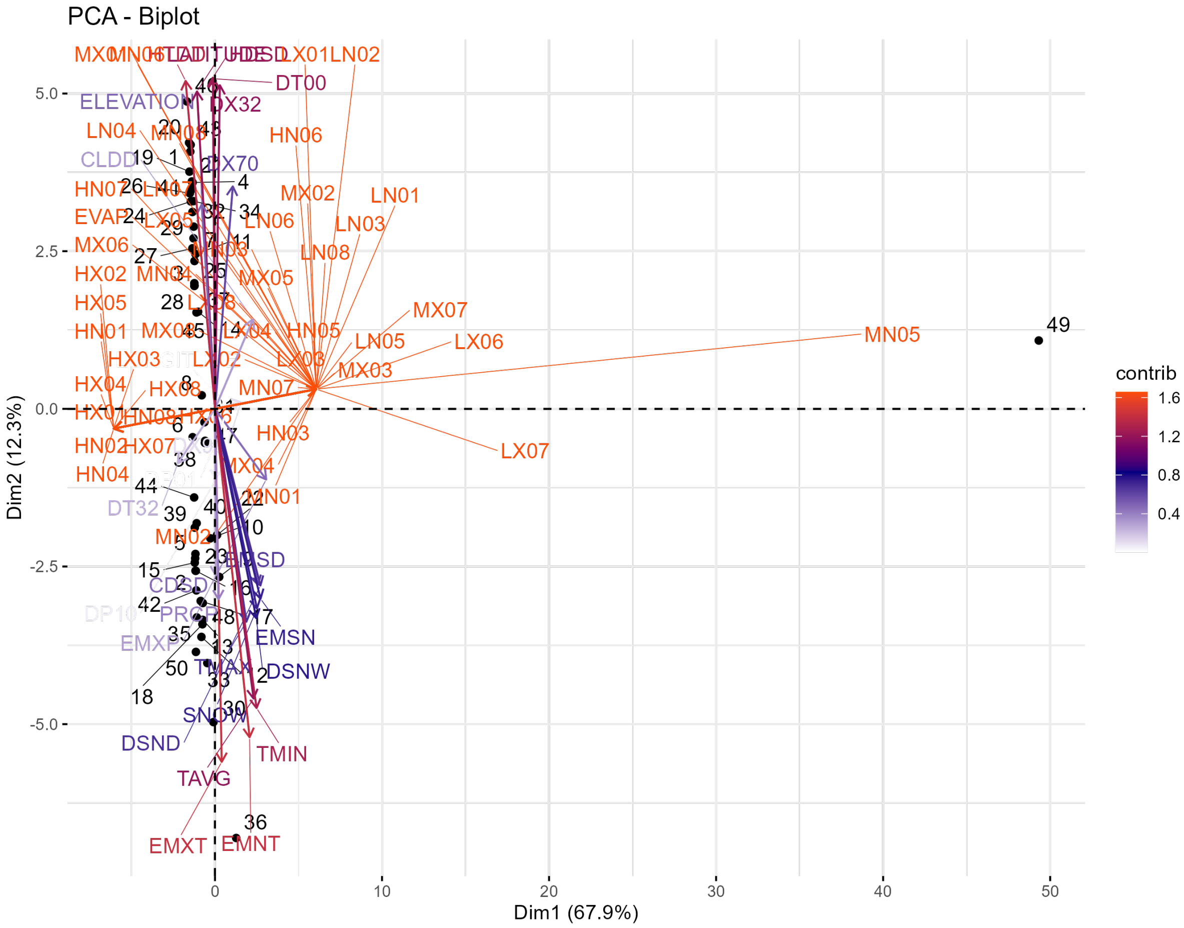


Figure 12 displays the projection of the variables onto a two-dimensional space. It is noticeable that variables such as LX07 and MN05 EVAP contribute significantly to the overall pattern observed in the plot.

5.1.4 Cumultative Variance Plot based on PCA

Figure 13: Biplot of the Variables

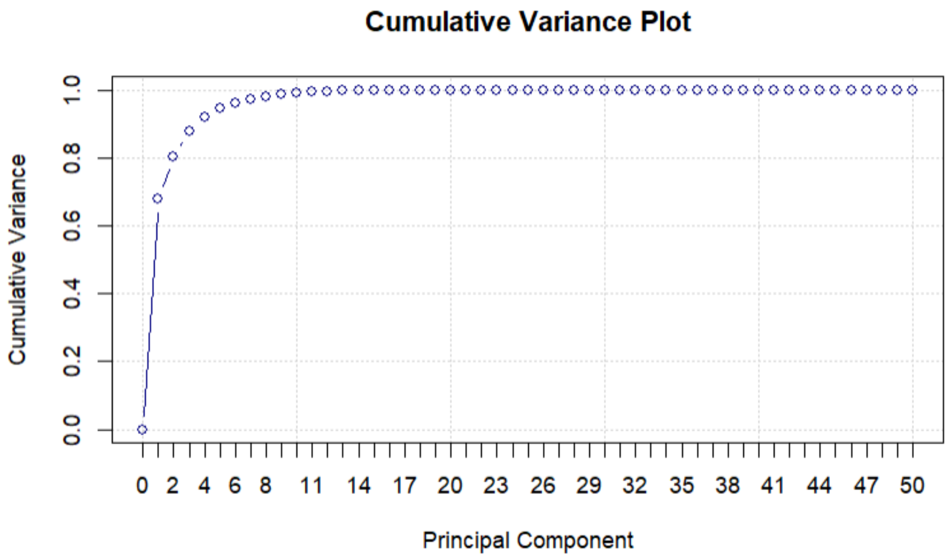


Figure 13 illustrates the cumulative variance plot of the variables. It is evident that approximately three principal components account for around 80% of the variance in the dataset.

5.1.5 PCA Based K-Means Clustering

Figure 14: K-Means Clustering

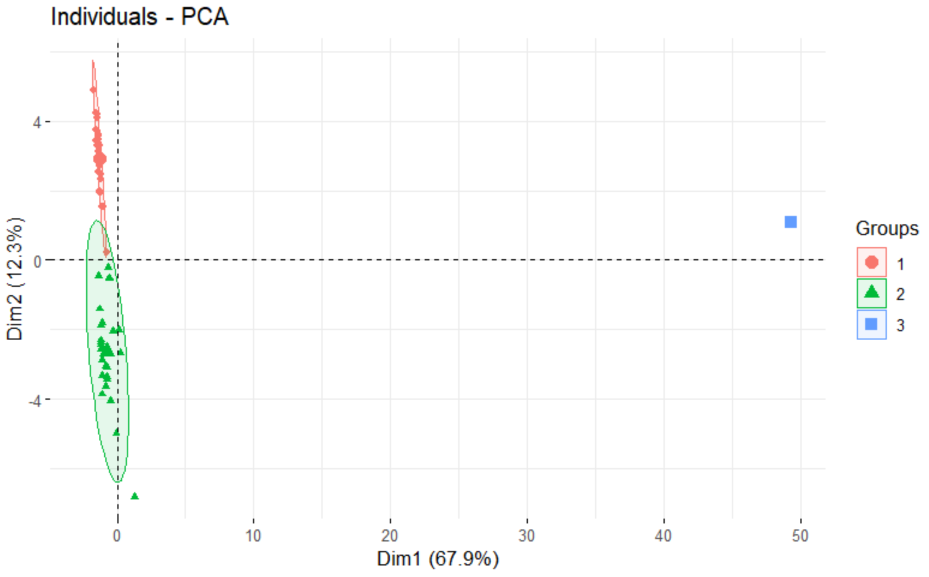


Figure 14 presents the k-means clustering results based on the first three principal components.

Figure 15: PCA biplots with k-means clustering results for different numbers of clusters (K = 3, 4, 5, and 6).

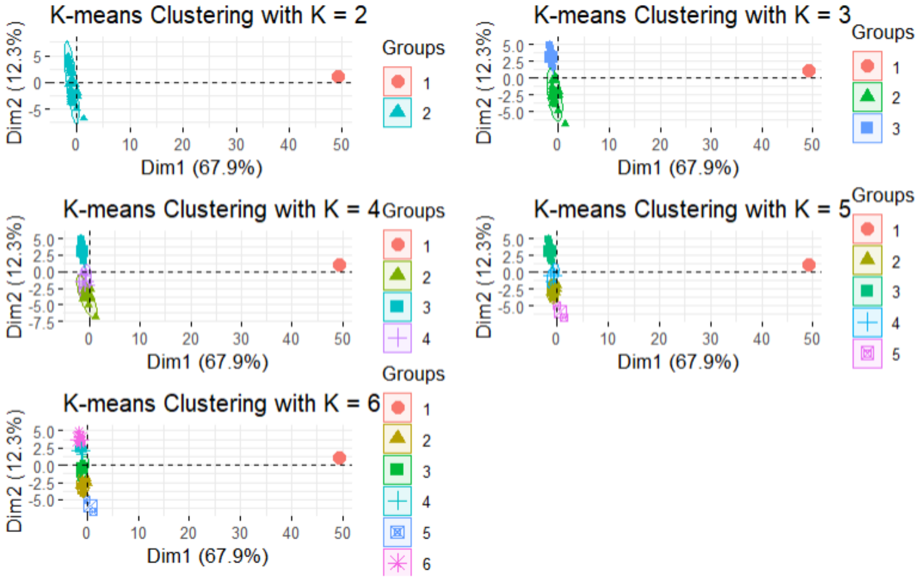
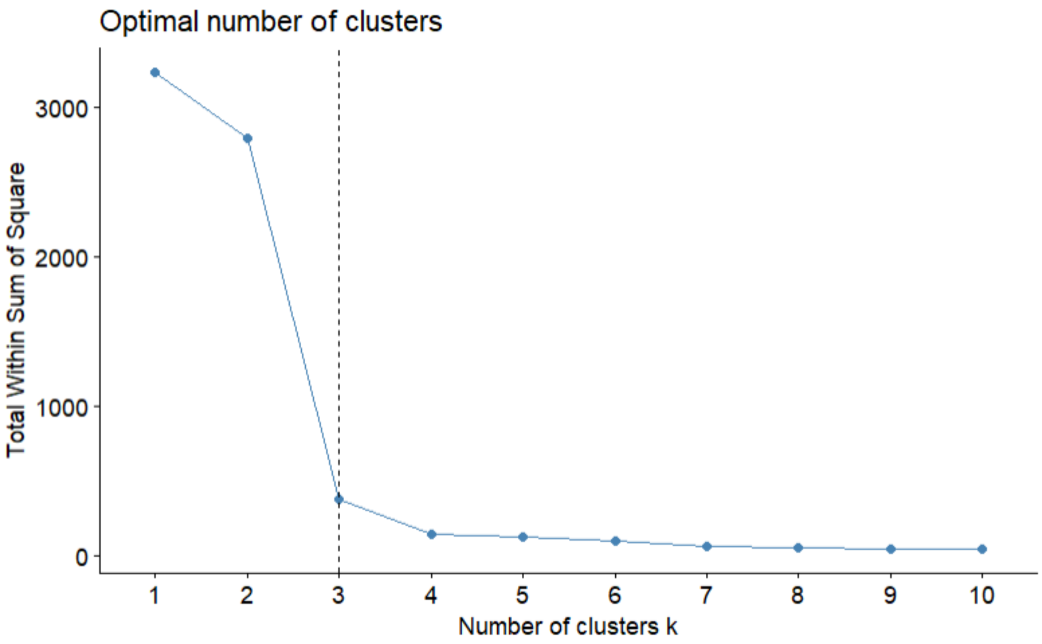


Figure 15 illustrates the clusterings for k=2,3,4,5 and 6

Each cluster arrangement provides a unique insight into the organization of data, offering valuable perspectives for discerning the underlying patterns within climate datasets.

5.1.6 Scree Plot

Figure 16: K-Means Clustering



The scree plot in Figure 16 indicates a sharp decline in variance contribution, suggesting that the first few components capture the majority of information.

5.2 Heatmap of Clustered Data: Selected Variables

Figure 17: K-Means Clustering

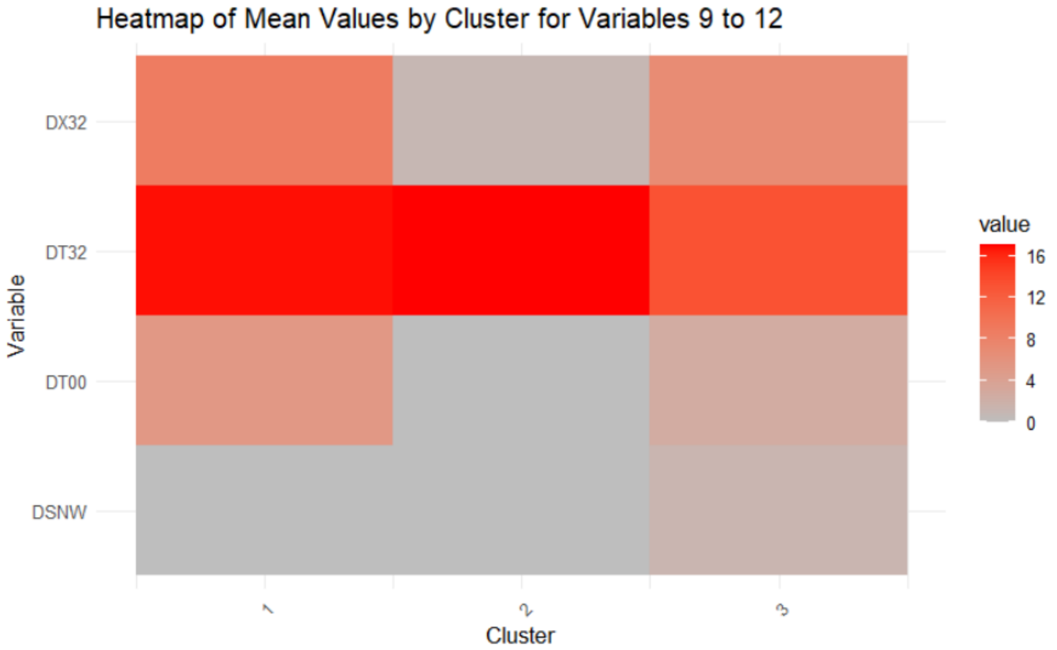


Figure 18: K-Means Clustering

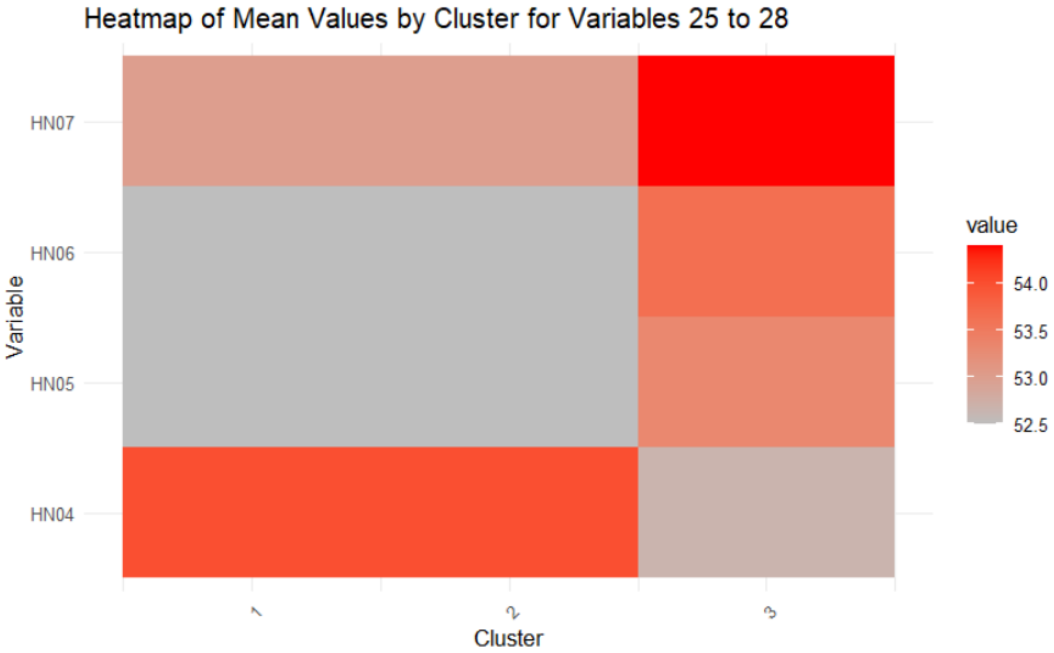


Figure 19: K-Means Clustering

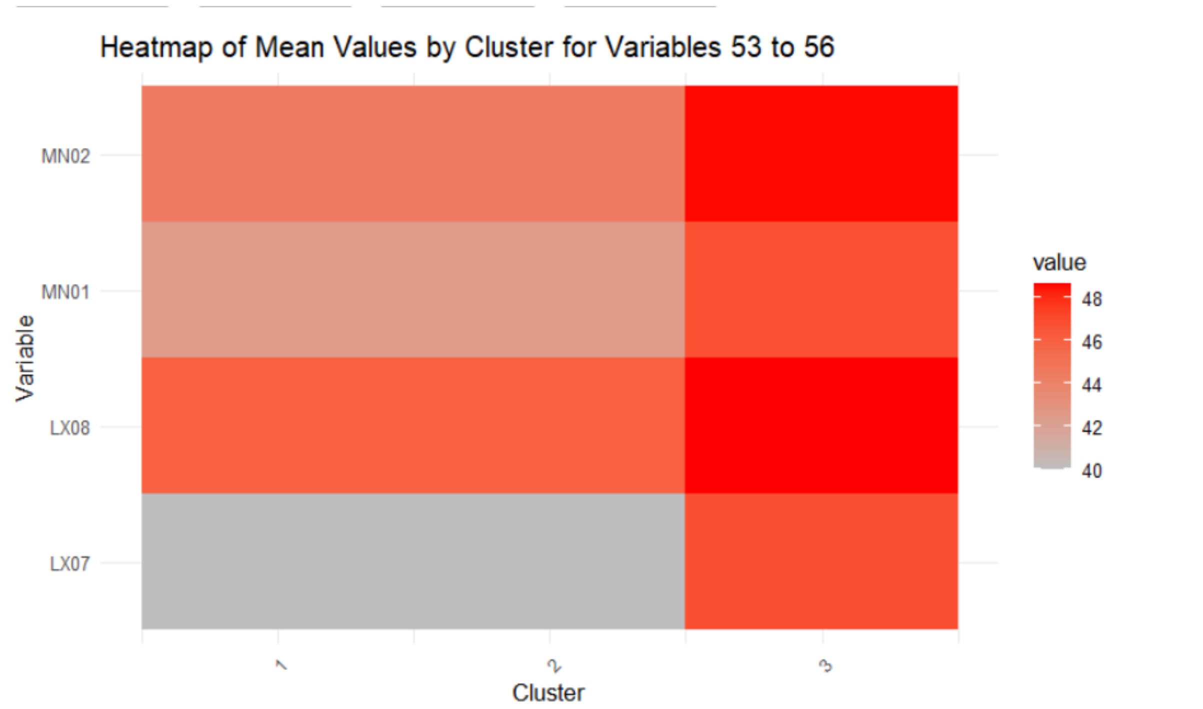
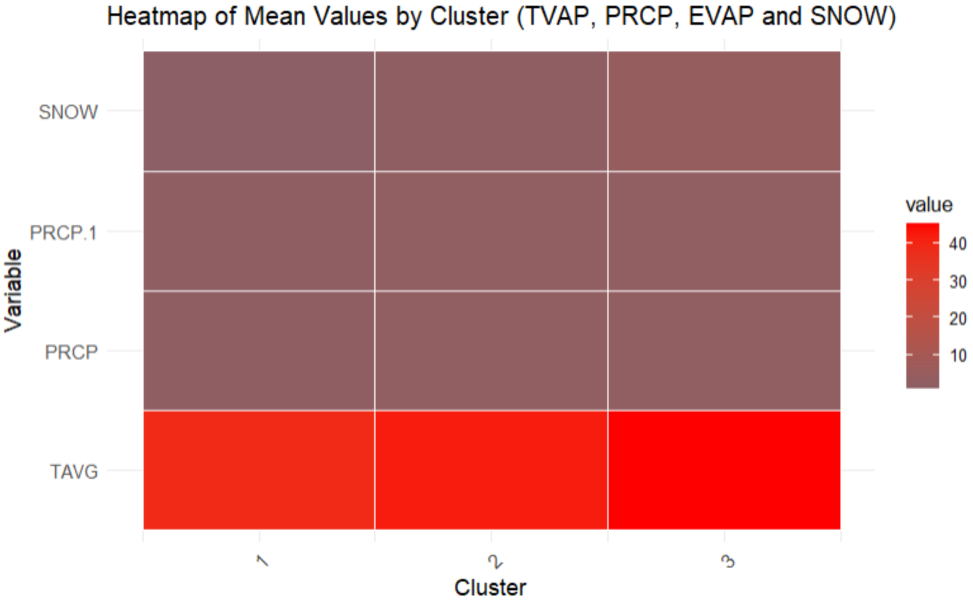


Figure 20, Figure 18, and Figure 19 showcase specific variables delineated by the K-means clustering

5.3 Heatmap of Selected Variables cluster Variables

Figure 20: K-Means Clustering



Total average temperature, precipitation, Evaporation and Snow were selected to see their patterns.

5.4 Map of 5 Weather Stations with highest Total Average Temperature

Figure 21: K-Means Clustering

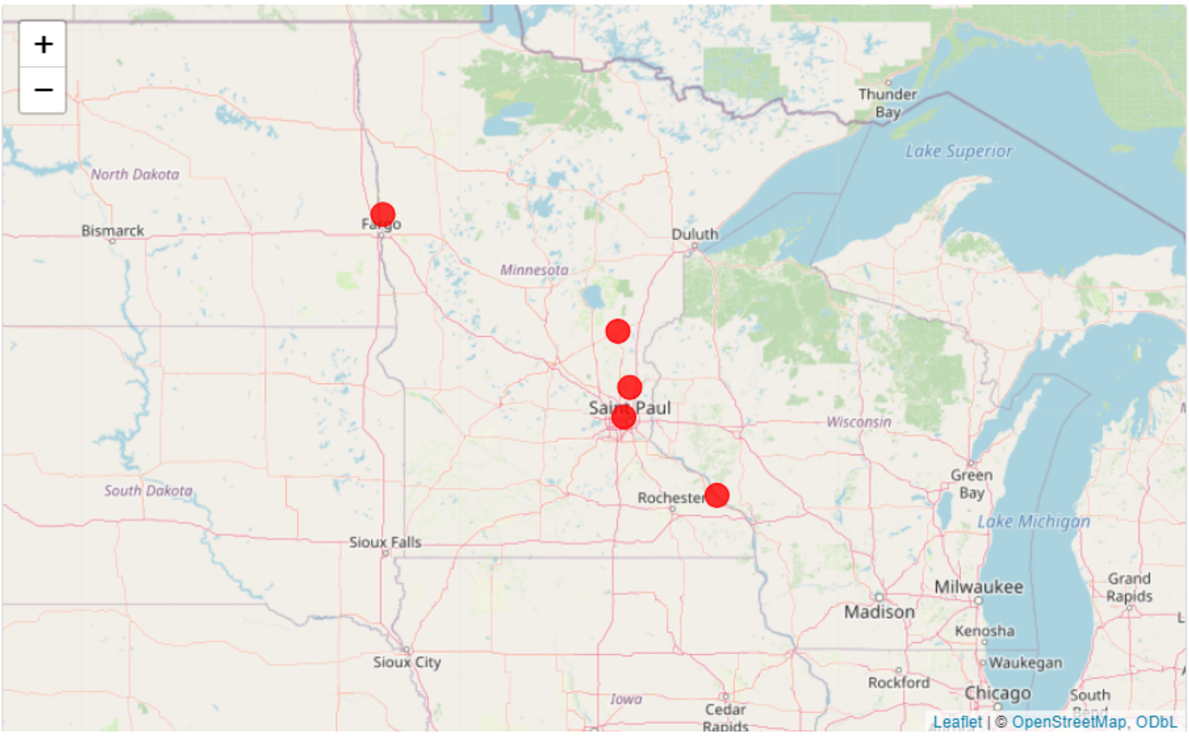


Figure 21 illustrates the map displaying five Weather Stations in Minnesota with highest total average temperature.

5.5 Conclusion

The statistical analyses, including Hotelling's T-squared test, revealed significant disparities in climate variables between the two periods under investigation. This finding aligns with our initial hypothesis, indicating temporal variations in Minnesota's climate. Furthermore, cluster analysis identified evaporation, total average temperature, and snow as pivotal factors influencing the overall climatic conditions in Minnesota. Notably, weather stations situated at Carlos Avery, Mora, Georgetown, University of Minnesota St. Paul, and Minnesota City Dams exhibited the highest total average temperatures across the state. Specifically, Minnesota City Dam recorded 47.867°F, University of Minnesota St. Paul recorded 45.008°F, Georgetown recorded 43.28°F, Mora recorded 42.74°F, and Carlos Avery recorded 44.36°F, all in Fahrenheit.

References

[1] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis* (7th ed.). Prentice Hall.

[2] Krzanowski, W.J. (2000). *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press.

[3] Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall.

[4] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[5] Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). Springer.

[6] NOAA National Centers for Environmental Information (NCEI). Available at: https://www.ncei.noaa.gov/

[7] Scikit-Learn Documentation. Available at: https://scikit-learn.org/stable/

[8] Journal of Climate. American Meteorological Society.

[9] Statistics in Medicine. Wiley.