Analysis of Changes in Weather Elements Clustering in Pagar Alam in 2023 and 2024 Using *K*-Means and Average Linkage Clustering Methods

Sri Indra Maiyanti[1], Irmeilyana[2], Putri Nilam Cayo[3], Ngudiantoro[4]

{sri_indra_maiyanti@mipa.unsri.ac.id[1], irmeilyana@unsri.ac.id[2], nilam@unsri.ac.id[3], ngudiantoro@unsri.ac.id[4]}

University of Sriwijaya, Faculty of Mathematics and Natural Sciences, Department of Mathematics Jl. Raya Palembang-Prabumulih Km. 32, Indralaya, South Sumatra Province, Indonesia[1,2,3,4]

Corresponding author: irmeilyana@unsri.ac.id

Abstract. Climate is a combination of weather elements over a long period of time. Global climate change can impact the productivity of agricultural crops, including coffee plantations in Pagar Alam, South Sumatra, Indonesia. The purpose of this study is to cluster the time of occurrence based on weather elements data in Pagar Alam in 2023 and 2024 using K-means and average linkage clustering methods. The data matrix consists of 53 weeks and 15 weather element variables. Based on the biplot and Silhouette Index results of both data matrices, the K values used were 2 and 4. The K-means clustering results on the 2024 data matrix indicate that a cluster of the majority of weeks is characterized by higher minimum temperature, dew point, humidity, precipitation cover, and cloud cover. Meanwhile, in the 2023 data matrix, a cluster of the majority of weeks is characterized by high levels of dew, humidity, and cloud cover. The average linkage clustering results are almost similar to the K-means clustering results. The results of time data clustering can help develop strategies for timing agricultural land management practices as an adaptation measure to weather variability.

Keywords: Average linkage, K-means clustering, Pagar Alam, weather elements, weekly time clusters.

1 Introduction

Climate refers to the long-term average pattern of weather conditions in a particular region. Climate is determined by several weather elements, including solar radiation, temperature, humidity, clouds, rain, evaporation, air pressure, and wind [1]. Weather can impact all aspects of human activity and the environment [2, 3], including coffee plantations [4] and its trade [5]. Global climate change is impacting coffee productivity [6]-[9], including in Indonesia. Coffee productivity is influenced not only by annual climate trends but also by short-term variations in weather conditions during critical phenological and harvesting periods. Such temporal

SICBAS 2025, November 06, Palembang, Indonesia

DOI 10.4108/eai.6-11-2025.2364259

variability directly affects flowering, fruit development, and bean quality [10], [11]. Understanding temporal weather variability during coffee production is therefore essential for developing effective climate adaptation strategies and improving farm management decisions.

Indonesian coffee production declined from 2021 to 2023, most recently by 2.1% in 2023 [12]. South Sumatra, as the largest coffee producer, also saw a decline, including in Pagar Alam. Data from the Directorate General of Plantations and BPS–Statistics Indonesia [13]–[15] show that coffee production in Pagar Alam has fluctuated and generally declined over recent years. Interviews conducted in previous studies [16]–[19] indicated that coffee farmers in Pagar Alam attributed the decline in coffee production primarily to weather variability. Grouping observation periods according to dominant meteorological conditions can provide a general representation of weather characteristics throughout a coffee harvest period. Furthermore, the integration of meteorological data analysis has been widely proven effective in mapping the direct correlation between seasonal shifts and coffee bean quality degradation in similar tropical regions [20].

Several studies have investigated weather variability in Pagar Alam and surrounding coffee-growing regions. Most have focused on drought risk assessment or comparisons of selected weather variables between years. A GIS-based assessment of the drought threat index in Dempo Tengah District, Pagar Alam Municipality, showed that most areas fall into the high-threat zone [21]. However, the analysis was limited to one month of rainfall data and did not provide a multivariate visualization of weather variability. According to the results of mean difference tests reported in [22], the variables of temperature, maximum temperature, minimum temperature, dew point, rainfall coverage, and cloudiness in Dempo Tengah Sub District in 2021 were significantly different from 2022. Similarly, test results for two time periods, from 2022 to 2024, showed significant differences for temperature, dew, humidity, wind conditions, cloud cover, and sea level pressure [23]. Weather conditions during a coffee harvest period are characterized by multiple interrelated meteorological variables that change simultaneously over time. Therefore, a multivariate analytical approach is required to describe the underlying patterns more comprehensively. Among multivariate exploratory methods, biplot analysis provides an initial graphical representation of the relationships among weather variables and observation periods in a reduced-dimensional space, thereby revealing the overall structure of the data. Subsequently, clustering methods classify observation periods with similar meteorological characteristics into homogeneous groups [24], [25].

Clustering observation periods based on multiple meteorological variables can support climate adaptation strategies and improve automated weather prediction systems [26]. However, the study relied on a Kaggle dataset containing only temperature, rainfall, air pressure, and wind speed observations collected in Australia. According to Rinaldi [27], clustering daily data using *K*-means clustering can support weather pattern analysis for disaster mitigation and resource planning purposes. However, the weather elements used are only average temperature and average humidity over a short time span to identify two intuitive and relevant weather patterns (“rainy weather” and “cold, humid weather”).

K-means clustering is a non-hierarchical partitioning method that is sensitive to the selection of initial centroids but effectively partitions observations into clusters with similar characteristics. Meanwhile, hierarchical methods, particularly average linkage and centroid linkage, can produce more stable clusters and output a dendrogram whose structure can be explored [28]. Average linkage defines the distance between two clusters as the average distance between all pairs of observations, with each pair consisting of one observation from each cluster.

Previous comparative studies on meteorological data have demonstrated the usefulness of comparing *K*-means and hierarchical clustering methods for exploring complex weather datasets [29]. Although previous studies have demonstrated the usefulness of clustering techniques for weather analysis, several important gaps remain. Existing studies generally employ relatively short observation periods, limited sets of meteorological variables, publicly available datasets from regions outside Indonesia, or focus solely on daily weather classification. Despite these advances, little attention has been given to clustering weekly weather conditions across an entire coffee harvest year using comprehensive meteorological variables. Moreover, studies comparing *K*-means and average linkage hierarchical clustering for such data across consecutive coffee harvest years in Indonesia remain limited. Therefore, this study aims to cluster weekly observation periods based on meteorological data collected in Pagar Alam, South Sumatra, Indonesia, during 2023 and 2024 using *K*-means and average linkage hierarchical clustering methods, and to compare the resulting clustering structures to identify temporal weather variability between consecutive coffee harvest periods.

The resulting weekly weather clusters and their characteristic weather profiles can support evidence-based decision making in coffee cultivation by assisting farmers in scheduling fertilization, pruning, weed control, rejuvenation, and other management practices according to prevailing weather conditions. In addition, the findings may serve as valuable references for farmers, researchers, and policymakers in developing mitigation and adaptation strategies to address weather variability in coffee-growing regions.

2 Methods

This study used secondary data on daily weather elements over a period of one year in Pagar Alam for 2023 and 2024. The data source was the weather website www.visualcrossing.com. The data matrix was compiled used on the weekly average of every weather elements. The steps were as follows:

1. Construct two weather element data matrices $X^* = (x^*_{ij}); i = 1, 2, ..., n; j = 1, 2, ..., p$, representing data for 2023 and 2024.
2. Perform descriptive statistics for each variable.
3. Perform biplot analysis on each data matrix X^* using data standardization with *Z*-scores, data reduction by PCA, and biplot graphs [24], [25], [28], [30]. The biplot representation can be used to determine the number of clusters of observed objects (*K* value).
4. Normalize the data using min-max normalization, then obtain a new data matrix *X*:

$$x_{ij} = \frac{x*_{ij} - x_{j_{min}}}{x_{j_{max}} - x_{j_{min}}} \tag{1}$$

where x_{ij} is the new normalized value of the j-th variable in the i-th data set;
$x*_{ij}$ is the initial value of the j-th in the i-th data set;

$x_{j_{min}}$ is the initial minimum value of the j-th variable;

$x_{j_{max}}$ is the initial maximum value of the j-th variable.

5. Perform K-means clustering algorithm [31]-[34] in every data matrix X.

   5.1 Determine the K centroids, namely centroid c_l; $l = 1, 2, \cdots, K$, by analyzing the biplot results in Step 3, and also using the sillhouette index (SI) to estimate the quality of the formed clusters. The Silhouette Coefficient (SC) is written as:

$$SC = \ max_K \ SI\ (K), \tag{2}$$

   where $SI\ (K) = \ \frac{1}{K}\ \sum_{l=1}^{K} SI_l$ is the Global Silhouette Index (SI); with K is the number of clusters;

   $SI_l = \ \frac{1}{m_l}\ \sum_{i=1}^{m_l} SI_i^l$ is the average SI value for cluster-l;

   $SI_i^l = \ \frac{b_i^l - a_i^l}{max\left\{a_i^l,\ \ b_i^l\right\}}$ is the SI value of object -i in cluster-l;

   $i = 1, 2, \cdots, m_l$; where m_l is number of objects in cluster-l;

   a_i^l is the average distance of object-i to all other objects in a same cluster;

   b_i^l is the average distance of object- i to all other objects in other clusters.

   The selection of the K value is based on the SI graphic output from R Studio version 4.3.1. The Silhouette Index is chosen because of its superior consistency in handling dense, continuous multivariate environmental datasets compared to other internal validation indices [35].

   5.2 Calculates the distance of each object to each centroid c_l; $l = 1, 2, \cdots, K$.

$$d\ (x_i, c_l\ )\ = \sqrt{\sum_{j=1}^{P} \left(x_{ij} - c_{lj}\right)^2}\ , \tag{3}$$

   where $d\ (x_i, c_l\ )$ is the Euclidean distance between object-i and the initial cluster centroid- l; x_{ij} is the value of the j-th variable in the i-th object; c_{lj} is the value of the j-th variable centroid at centroid-l.

   5.3 Calculate the minimum distance from object i to each centroid-l; $l = 1, 2, \cdots, K$.

   5.4 Renew the centroid at cluster- l:

$$c_{lj} = \frac{\sum_{i=1}^{m_l}\ \ x_{ij}}{m_{l.}};\quad i = 1,2,3, ...m_l\ , \tag{4}$$

   where c_{lj} is the centroid of the j-th variable in the l-th cluster; and m_l is the number of objects in cluster-l.

   5.5 Iterates until all centroids converge and are stable.

6. Perform agglomerative hierarchical clustering [28] on n objects from each data matrix X.

   6.1 Start with n clusters, each containing a single entity and $n \times n$ symmetric distance matrix $D = \left(d_{ij}\right)$.

6.2 Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between "most similar" clusters U and V be d_{UV}.

6.3 Merge clusters U and V. Label the newly formed cluster (UV). Update the entries in the distance matrix by (i) deleting the rows and columns corresponding to clusters U and V; (ii) adding a row and column giving the distances between cluster (UV) and the remaining clusters. The distances between (UV) and the other cluster W are determined by

$$d_{(UV)W} = \frac{\sum_i \quad \sum_k \quad d_{ik}}{n_{(UV)}\, n_W}, \tag{5}$$

where d_{ik} is the distance between object i in the cluster (UV) and object k in the cluster W; $n_{(UV)}$ and n_W are the number of items in clusters (UV) and W, respectively.

6.4 Repeat Steps 6.2 and 6.3 a total $N-1$ times. Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

6.5 Visual representation of clustering results through dendograms.

7. Further data analysis.

7.1 Interpretation of the characteristics of each cluster resulting from *K*-means clustering and average linkage in each data matrix 2023 and 2024 based on the values of weather elements.

7.2 interpretation of the characteristics comparison of each cluster based on weather elements in both data matrices.

Data processing and analysis using software RStudio version 4.3.1 , Minitab 19, and SPSS 24.

3 Results and Discussion

The data used is secondary data, limited to the Pagar Alam region of South Sumatra. The objects in the data matrix are weekly periods, from January 1 to December 31 in a year, resulting in 53 weeks, with the 53rd week consisting of only 2 or 3 days. Variables and their notations and units include: Maximum Temperature (Tmax; X_1 ; in oC), Minimum Temperature (Tmin; X_2; in oC), Mean temperature (Temp; X_3; in oC), Dew point (Dew; X_4; in oC), Relative humidity (Hum; X_5; in %), Precipitation (Precip; X_6; in mm), Precipitation Cover (Pcover; X_7; in mm), Wind Gust (Wgust; X_8; in km/h), Wind Speed (WSpeed; X_9; in km/h), Wind Direction (WDir; X_{10}; in $degrees\ from\ North$), Sea level pressure (Seapress; X_{11}; in mb), Cloud cover (CCover; X_{12}; in %), Solar radiation (SolRad; X_{13}; in W/m^2), UV Index (UVI; X_{14}; $between$ 0 and 10), Moonphase (Moophase; X_{15}; $between$ 0 and 10).

The entries of two data matrices used were 15 weather elements data in Pagar Alam in 2023 and 2024. Each data matrix for 2023 and 2024 consists of 53 rows and 15 columns, representing the coordinates of each object (time per week in 1 year) in 15-dimensional space.

The following describes the results of processing the data matrix 2024, and the same process was carried out for the data matrix 2023. Descriptive statistics are presented in **Table 1**. The values of Q_1, Q_2 (median), and Q_3 were used as the segmentation boundaries for the weather elements. A comparison of the boxplots of all variables from both data matrices can be seen in **Figure 1**.

Table 1. Statistics descriptive data matrix 2024 and data matrix 2023

Variable	Mean	StDev	Minimum	Q_1	Median	Q_3	Maximum
Tmax_2024	27.0	1.0	24.7	26.2	26.9	27.6	29.6
Tmax_2023	26.8	1.3	24.3	26.2	26.6	27.2	30.4
Tmin_2024	19.7	0.6	18.2	19.3	19.7	20.1	20.9
Tmin_2023	19.4	0.6	18.1	19.0	19.4	19.8	20.7
Temp_2024	22.8	0.6	21.5	22.3	22.7	23.0	24.2
Temp_2023	22.5	0.7	20.8	22.1	22.4	22.8	24.3
Dew_2024	20.0	1.0	17.5	19.3	20.2	20.8	21.6
Dew_2023	19.5	1.1	16.6	19.0	19.8	20.3	21.1
Hum_2024	85.7	5.9	68.2	82.6	87.7	90.1	92.6
Hum_2023	84.7	6.4	69.0	82.4	87.0	89.1	93.2
Precip_2024	8.9	6.4	0.2	4.4	7.7	13.2	25.1
Precip_2023	6.8	4.9	0.3	2.8	5.5	10.0	19.7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
MooPhase_2024	0.5	0.2	0.1	0.3	0.5	0.7	0.8
MooPhase_2023	0.5	0.2	0.1	0.3	0.5	0.7	0.9

Based on **Table 1**, mean, standard deviation, minimum, Q_1, Q_2, Q_3, and maximum values of variables on both data matrices are almost the same and not too different. Even some values are the same with rounding 1 decimal. Based on the data value scale in Figure 1, the means of several variables are quite different, namely Tmin, Temp, Dew, Wgust, and WSpeed. The means of the variables in the data matrix 2024 are relatively higher than the means of the data matrix 2023, except for WDir, Seapress, SolRad, and UVI. Meanwhile, the means for MooPhase are relatively similar. A significant difference in the variation of variable values occurs in the Precip variable.

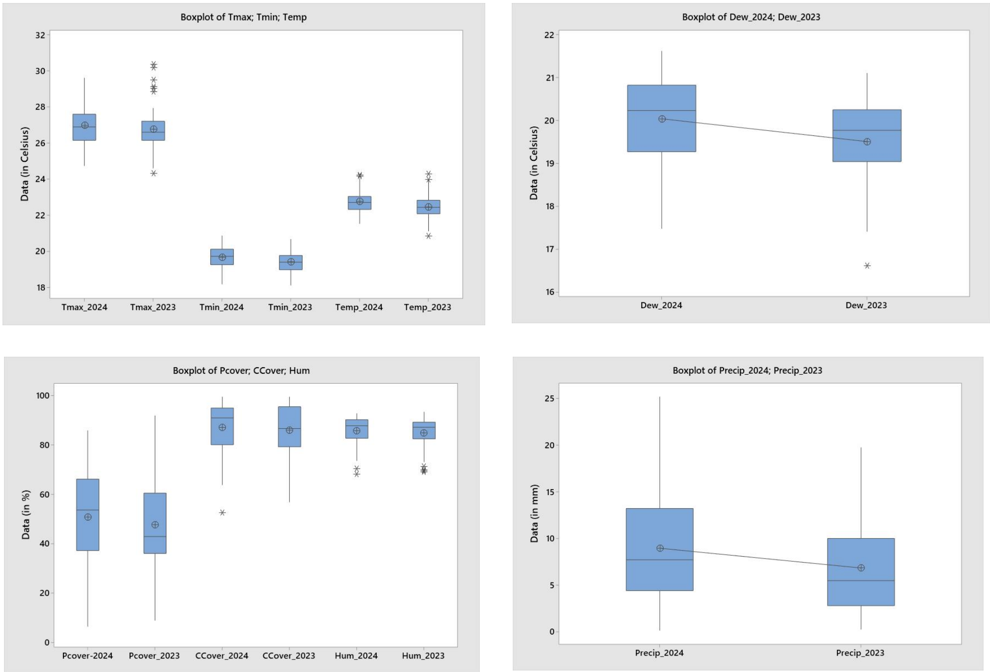

Fig. 1. Boxplot of Tmax, Tmin, Temp (on upper left side), Dew (on upper right side), Pcover, CCover, Hum (on bottom left side), Precip (on bottom right side) variables in data matrix 2024 and 2023

The coordinates of each object in the data matrix can be represented in the form of a biplot. Furthermore, biplots can also represent the correlation between variables and the relative relationships between objects and variables. Biplots are obtained by first performing PCA (Principal Component Analysis) on the standardized data matrix. The variance and loading plots of the variables, as well as the component scores for each object, are obtained from the correlation matrix. The biplots of both data matrices can be seen in **Figure 2** and **Figure 3**.

Based on the biplot of data matrix 2024, the Dim1 (PC1; 45.5%) and Dim2 (PC2; 18.3%) axes explain 63.8% of the data variation. The red arrows represent the variable vectors, while the blue dots represent the coordinates of objects or observations. The variables Tmax, UVI, and SolRad are highly correlated with each other and have a large positive contribution to Dim1. It means that objects on the right side of the graph tend to have higher Tmax (maximum temperature), solar radiation, and UV index. The variables Hum, PCover, and Dew also contribute significantly to PC1 but have a large negative correlation with Tmax, SolRad, and UVI. The variables WSpeed and Wgust have a more dominant influence on PC2 and tend to have very low correlations with the variables that contribute significantly to PC1.

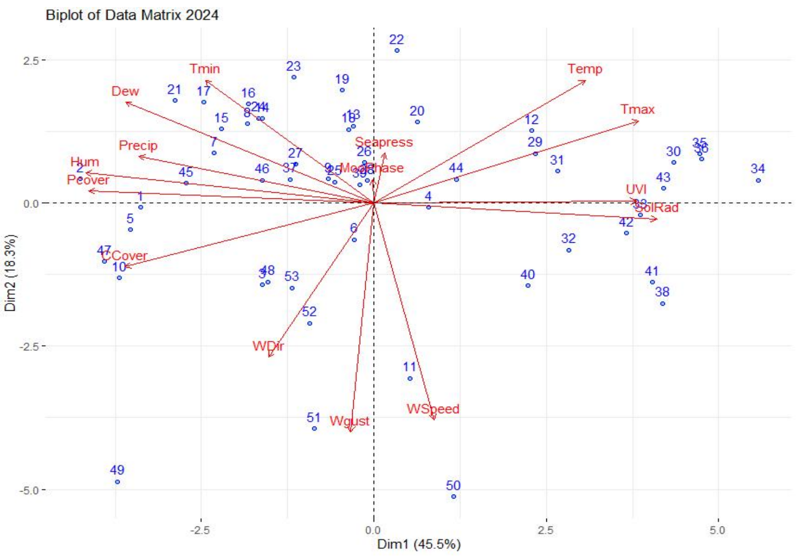

Fig. 2. Biplot of data matrix 2024

In general, observations (blue dots) adjacent to the direction of the arrow of a variable have higher values for that variable. For example, Week 30, 33, 34, 35, 36, 42, and 43 have higher Tmax, SolRad, and UVI values than other weeks, indicating that the conditions of these observed objects are closely related to high temperatures and UV radiation. Week 1, 2, 5, 7, 37, 45, and 46 tend to have higher Hum, PCover, Dew, and Precip values than other weeks.

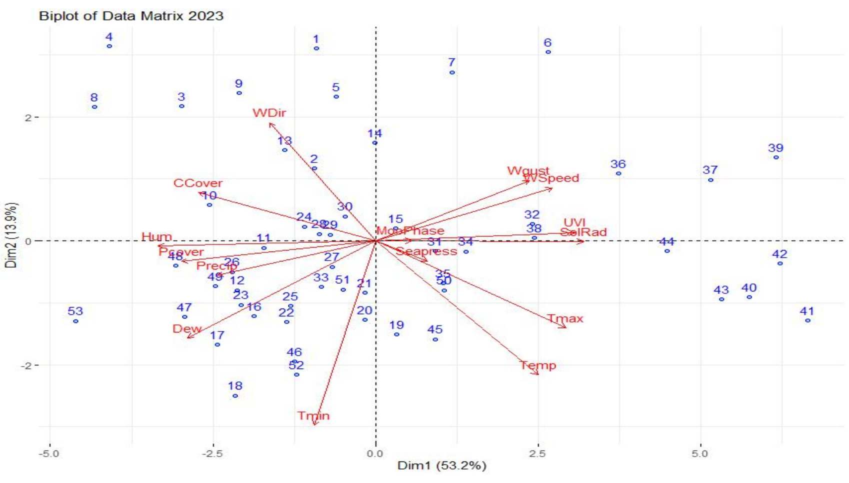

Fig. 3. Biplot of data matrix 2023

The biplot of data matrix 2023 shows that the two main dimensions (Dim1 = 53.2% and Dim2= 13.9%) explain 67.1% of the data variation. The UVI and SolRad variables contribute predominantly to PC1. Similarly, the Hum, PCover, Precip, and CCover variables also contribute predominantly to PC1. These two groups of variables are highly negatively correlated. The minimum temperature (Tmin) variable contributes highest to PC2. Tmin variable has a very low correlation with the two groups of variables that contribute highest to PC1. Tmax and Temp have a high correlation. Wgust and WSpeed are also highly correlated. Observing the relative position of objects in the variable vector reveals the characteristics of object clustering. For example, Weeks 31, 32, 34, 39, 40, 42, 43, 41, and 44 have higher UVI and SolRad values than other weeks. Next, cluster analysis was performed to analyze the object and variable clustering.

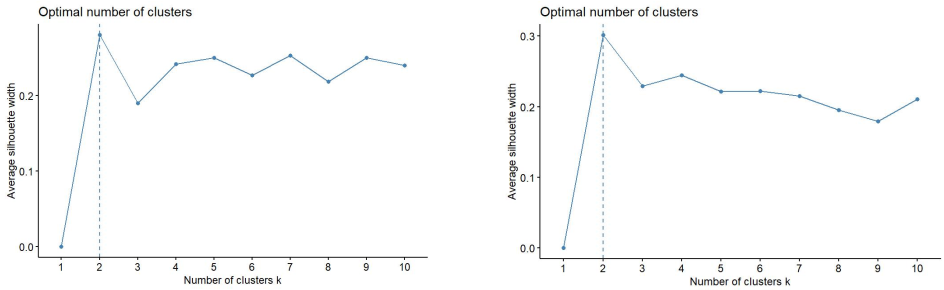

Fig. 4. Silhouette Index values for each *K* value in data matrix 2024 (on left side) and data matrix 2023 (on right side)

The use of non-hierarchical cluster analysis, namely *K*-means clustering, is carried out by first normalizing each entry of the data matrix, using equation (1). Determination of the *K* value in *K*-means clustering is based on the Silhouette Index (SI) using equation (2) and also by using the biplot of each data matrix. Determination of the SI value uses the help of R studio version 4.3.1. The optimal SI value is obtained at $K = 2$. Furthermore, based on the SI value and biplot, clustering is carried out at $K = 2$ and $K = 4$. **Figure 4** is the SI of each data matrix.

K-means clustering begins by determining the initial centroids (centers) for the *K* clusters. For example, for the data matrix 2024, the initial cluster centroids for K =2 are Week 49 and Week 37. For K=4, the initial cluster centroids are Week 50, Week 7, Week 14, and Week 34. These initial cluster centroids can be seen in **Table 2**.

The *F* tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among objects in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal. The significance level (alpha-value) used is 5%, so if the probability value of the *F* test result (noted as *p*-value or Sig.) is greater than 0.05, then the test result is considered statistically insignificant, thus failing to reject the hypothesis that there is insufficient evidence to state a significant difference in the "variable characteristics" between clusters. So, it can be said that the variable is not significant in characterizing the differences between clusters.

Table 2. Initial cluster centroids for *K*=2 and *K*=4 in the data matrix 2024

Variable	*K*=2 Cluster		*K*=4 Cluster			
	1	2	1	2	3	4
Tmax	0.00	0.93	0.38	0.29	0.46	0.93
Tmin	0.42	0.54	0.37	0.85	0.77	0.54
Temp	0.00	0.93	0.39	0.27	0.54	0.93
Dew	0.55	0.09	0.39	0.73	0.90	0.09
Hum	0.90	0.09	0.55	0.90	0.90	0.09
Precip	0.44	0.03	0.17	0.38	0.82	0.03
Pcover	0.91	0.22	0.41	0.82	0.76	0.22
Wgust	0.92	0.30	1.00	0.22	0.15	0.30
WSpeed	0.76	0.42	1.00	0.26	0.13	0.42
WDir	0.90	0.25	0.91	0.00	0.60	0.25
Seapress	0.10	0.54	0.00	0.79	0.29	0.54
CCover	0.98	0.29	0.86	0.87	0.82	0.29
SolRad	0.01	0.84	0.65	0.14	0.50	0.84
UVI	0.00	0.83	0.73	0.13	0.60	0.83
MooPhase	0.03	0.64	0.37	0.09	0.99	0.64

Final cluster centroids and its distance, observed significance levels, and the number of objects in each cluster can be seen in **Table 3**. Next, the grouping process is carried out using equation (3) and equation (4). *K*-means clustering was carried out with the help of SPSS. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is 0.000. For *K*=2, the current iteration is 6. The minimum distance between initial centers is 2.569. Meanwhile, for *K*=4, the current iteration is 4. The minimum distance between initial centers is 1.457.

The results of the two-cluster clustering show that the distance between cluster 1 and cluster 2 is 1.191. At *K*=2, 36 weeks in cluster 1 are predominantly characterized by the highest Tmin, Dew, Hum, Precip, Pcover, and CCover values. Seventeen weeks in cluster 2 are predominantly characterized by higher Tmax, Temp, SolRad, and UVI values. Meanwhile, at *K*=4, members of clusters 2 and 3, each with 17 weeks, have high Tmin, Dew, Hum, Pcover, and CCover. An additional characteristic of cluster 3 is its highest MooPhase. Cluster 4 at *K*=4 shares the same characteristics as Cluster 2 at *K*=2, namely the highest Tmax, Temp, SolRad, and UVI values. At *K*=4, a small portion of weeks (there are 5 weeks) in Cluster 1 are characterized by high Hum, Wgust, WSpeed, WDir, and CCover. According to Table 3, at *K*=2, Wgust, WSpeed, WDir, Seapress, and MooPhase variables do not significantly influence the dissimilarity between clusters. Meanwhile, at *K*=4, only Seapress has a non-significant effect.

Furthermore, the same method was used for the data matrix 2023, resulting in the summary as in **Table 4**. For *K*=2, the current iteration is 2. The minimum distance between initial centers is 2.768. Meanwhile, for *K*=4, the current iteration is 4. The minimum distance between initial centers is 1.457.

Table 3. Final cluster centroids, observed significance levels, and the number of objects in each cluster on data matrix 2024

Variable	*K*=2 Final cluster centroid		Sig.	*K*=4 Final cluster centroid				Sig.
	1	2		1	2	3	4	
Tmax	0.37	*0.68*	0.00	0.26	0.38	0.41	*0.71*	0.00
Tmin	*0.60*	0.42	0.02	0.40	*0.65*	*0.59*	0.41	0.02
Temp	0.39	*0.61*	0.00	0.27	0.42	0.42	0.64	0.001
Dew	*0.76*	0.33	0.00	0.55	*0.79*	*0.73*	0.30	0.00
Hum	*0.86*	0.42	0.00	*0.75*	*0.88*	*0.82*	0.39	0.00
Precip	*0.45*	0.15	0.00	0.21	0.51	0.46	0.09	0.00
Pcover	*0.69*	0.29	0.00	0.56	*0.71*	*0.67*	0.24	0.00
Wgust	0.25	0.26	0.89	*0.77*	0.12	0.27	0.21	0.00
WSpeed	0.23	0.34	0.09	*0.72*	0.12	0.24	0.30	0.00
WDir	0.60	0.49	0.18	*0.93*	0.50	0.64	0.42	0.001
Seapress	0.43	0.44	0.92	0.32	0.42	0.48	0.43	0.59
CCover	*0.85*	0.51	0.00	*0.92*	*0.83*	*0.82*	0.46	0.00
SolRad	0.33	*0.76*	0.00	0.42	0.31	0.37	*0.78*	0.00
UVI	0.43	*0.77*	0.00	0.47	0.41	0.48	*0.80*	0.00
MooPhase	0.52	0.49	0.75	0.32	0.29	*0.80*	0.49	0.00
Number of objects	36	17		5	17	17	14	

Note: A significance level > 0.05 indicates that the variable has no significant effect on determining dissimilarity between clusters. Numbers in italics indicate the value of the variable that predominantly characterizes the cluster.

The results of grouping 2 clusters in the data matrix 2023, it was found that the distance between cluster 1 and cluster 2 was 1.185. At $K = 2$, there are 36 weeks in cluster 1 dominantly characterized by higher Dew, Hum, Precip, Pcover, WDir, and CCover values. There are 17 weeks in cluster 2 dominantly characterized by higher Tmax, Temp, WGust, WSpeed, SolRad, and UVI. Meanwhile, at $K = 4$, cluster 1 which has 21 weeks, does not have dominant weather element characteristics. Cluster 3 which consists of 19 weeks, is characterized by high Tmin, Dew, Hum, and CCover. While cluster 4 which only consists of 5 weeks, is characterized by high Hum, Precip, PCover, WDir, and CCover. Cluster 2 which consists of 8 weeks, has the same characteristics as Cluster 2 at K=2, namely the highest values for Tmax, Temp, WGust, WSpeed, SolRad, and UVI.

Based on **Table 4**, at K=2, the variables of Tmin, Seapress, and MooPhase do not significantly influence dissimilarity between clusters. Meanwhile, at K=4, Seapress and MooPhase do not significantly influence dissimilarity between clusters.

The results of K-means clustering by week within a year at K=2 for 2024 and 2023 can be summarized in **Table 5**. Variables that do not significantly differentiate characteristics between clusters were not considered in the cluster comparison for those two years.

Table 4. Final cluster centroids, observed significance levels, and the number of objects in each cluster on data matrix 2023

Variable	*K*=2 Final cluster centroid		Sig.	*K*=4 Final cluster centroid				Sig.
	1	2		1	2	3	4	
Tmax	0.31	0.61	0.00	0.38	0.79	0.35	0.12	0.00
Tmin	0.51	0.47	0.52	0.43	0.53	0.63	0.25	0.002
Temp	0.39	0.63	0.00	0.44	0.78	0.45	0.14	0.00
Dew	0.76	0.39	0.00	0.63	0.21	0.83	0.71	0.00
Hum	0.80	0.34	0.00	0.65	0.11	0.81	0.92	0.00
Precip	0.43	0.14	0.00	0.20	0.11	0.49	0.68	0.00
Pcover	0.57	0.25	0.00	0.37	0.16	0.61	0.82	0.00
Wgust	0.26	0.61	0.00	0.39	0.75	0.20	0.38	0.00
WSpeed	0.20	0.55	0.00	0.33	0.69	0.16	0.26	0.00
WDir	0.53	0.31	0.01	0.49	0.20	0.42	0.88	0.00
Seapress	0.48	0.54	0.37	0.52	0.58	0.45	0.51	0.55
CCover	0.79	0.47	0.00	0.66	0.31	0.80	0.99	0.00
SolRad	0.39	0.77	0.00	0.55	0.89	0.40	0.19	0.00
UVI	0.38	0.79	0.00	0.53	0.93	0.40	0.16	0.00
MooPhase	0.46	0.57	0.21	0.57	0.58	0.39	0.42	0.17
Number of objects	36	17		21	8	19	5	

Table 5. Members of the two clusters and their characteristic variables

Variable	Data matrix 2024 Cluster		Data matrix 2023 Cluster	
	1	2	1	2
Tmax		√		√
Tmin	√		*	
Temp		√		√
Dew	√		√	
Hum	√		√	
Precip	√		√	
Pcover	√		√	
Wgust	*			√
WSpeed	*			√
WDir	*		√	
CCover	√		√	
SolRad		√		√
UVI		√		√
Number of Weeks	36	17	36	17
Week	Member of Cluster 1	Member of Cluster 2	Member of Cluster 1	Member of Cluster 2
1 to 4	1 2 3 4		1 2 3	4
5 to 8	6 7 8	5	5 6	7 8
9 to 12	9 10 11	12	9 10 11	12
13 to 16	13 15 16	14	13 14 15 16	

Variable	Data matrix 2024								Data matrix 2023							
	Cluster								Cluster							
	1				2				1				2			
17 to 20	17	18	19	20					17	18	19	20				
21 to 24	21	22	23	24					21	22	23	24				
25 to 28	25	26	27	28					25	26	27	28				
29 to 32	29					30	31	32					29	30	31	32
33 to 36			35	36	33	34						36	33	34	35	
37 to 40	37	38	39	40					37	38	39	40				
41 to 44	41		43	44		42			41		43	44		42		
45 to 48					45	46	47	48	45	46					47	48
49 to 52		50			49		51	52	49	50					51	52
53					53								53			

Note: A symbol "√" indicates the dominant variable characterizing the cluster.
A symbol * indicates that the variable is not significant in determining cluster dissimilarity.

In **Table 5**, the variables Seapress and MooPhase are the variables that do not significantly determine dissimilarity between the two clusters in both data matrices. Cluster 1, which comprises the majority of Weeks (there are 36 weeks, or approximately 68% of the weeks in a year) in both data matrices, is predominantly characterized by higher weather elements, namely Dew, Hum, Precip, Pcover, and CCover. However, in the data matrix 2024, the majority of these weeks are characterized by higher Tmin. While in the data matrix 2023, the majority of these weeks are also characterized by high WDir (wind direction blowing away from the North).

On the data matrix 2024, almost all weeks from January to July, September, and October have the characteristics of Cluster 1. Meanwhile, weeks in August, November, and December are characterized by Cluster 2, with higher Tmax, Temp, SolRad, and UVI values.

Meanwhile, in the data matrix 2023, weeks in early January, early February, early March, April to mid-July, early September, almost all weeks in October, early November, and early December are characterized by Cluster 1. Weeks in end of January, end of February, end of July, all weeks in August, end of November, and end of December are characterized by Cluster 2, with higher Tmax, Temp, SolRad, UVI, Wgust, and WSpeed values.

The same analysis was also performed for K=4. Therefore, based on the results of K-means clustering on both data matrices, Wgust, WSpeed, and WDir in 2023 were more characteristic of weather conditions than in 2024. Furthermore, Tmin did not significantly differentiate the clusters.

Next, by using equation (5) in the average linkage algorithm, a clustering in the form of a dendrogram was obtained in **Figure 5**. The data matrix used for object clustering (Weeks) was the normalized data matrix. Meanwhile, for variable clustering, the data matrix used is the standardized result of the original data matrix. The number of clusters in the resulting dendogram was based on the level of similarity in the amalgamation steps.

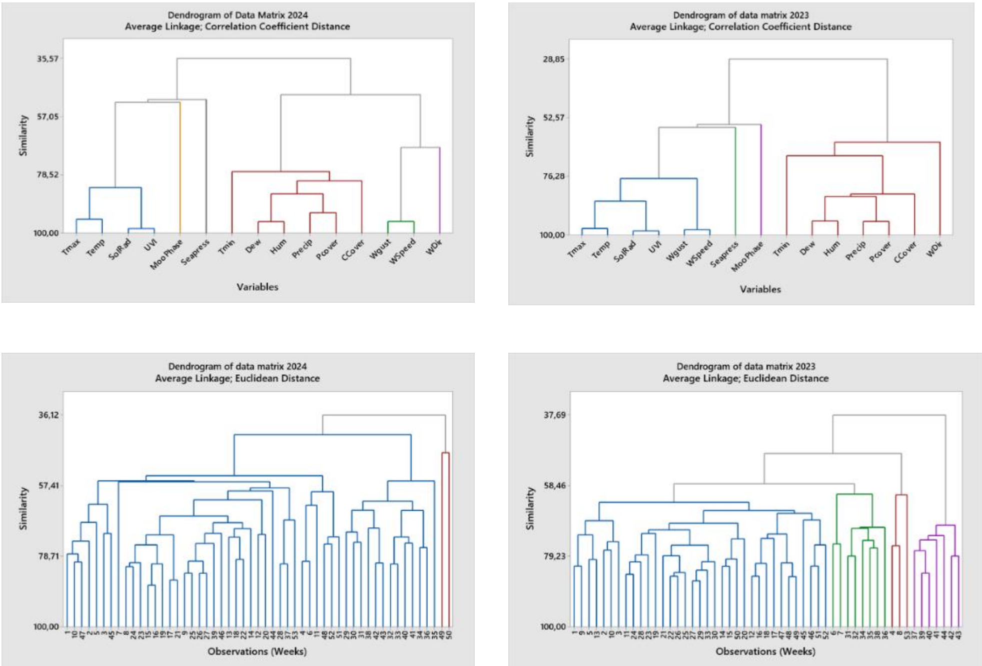

Fig. 5. Dendogram of variables clusters in data matrix 2024 (on upper left side) and data matrix 2023 (on upper right side), dendrogram of observation clusters in data matrix 2024 (on bottom left side) and data matrix 2023 (on bottom right side)

The results of *K*-means clustering can be compared with the results of hierarchical cluster analysis, namely average linkage. In the dendrogram of the data matrix 2024, there are clusters of highly correlated variables with a similarity level of 77%. These are the cluster consisting of Tmax, Temp, SolRad, and UVI; the cluster consisting of Tmin, Dew, Hum, Precip, PCover, and CCover; and the cluster consisting of Wgust and WSpeed.

In the amalgamation steps, four clusters were obtained in Step 49, with a similarity level of 54.3%. Therefore, all weeks, except Week 49 and Week 50, have relatively similar weather element characteristics. In Step 51, two clusters were obtained with a similarity level of 41.9%. According to **Table 6**, 38 weeks (in Cluster 1) have high Tmin, Dew, Hum, and CCover characteristics. Thirteen weeks (in Cluster 2) were characterized by the highest Tmax, SolRad, and UVI. Week 49 (in Cluster 3) and Week 50 (in Cluster 4) had extreme weather elements, namely: higher Hum, PCover, Wgust, WSpeed, WDir, CCover, and UVI values.

Meanwhile, in the dendrogram of the data matrix 2023 revealed clusters of highly correlated variables with a similarity level of 77%. These were the cluster consisting of Tmax, Temp, Wgust, WSpeed, SolRad, and UVI; and the cluster consisting of Dew, Hum, Precip, PCover, and CCover.

In the amalgamation steps of the data matrix 2023, four clusters were obtained at Step 49, with a similarity level of 60.9%. Seven weeks had weather element characteristics that were very different from the other 46 weeks. In Step 51, two clusters were identified with a similarity level of 48.9%. According to **Table 6**, 35 weeks (in Cluster 1) have high Dew, Hum, and CCover characteristics. Seven weeks in Cluster 4 (where Cluster 4 shares a similarity level with Cluster 2 at 48.9%) have the highest Tmax, Temp, WGust, WSpeed, SolRad, and UVI characteristics. Weeks 4, 8, and 53 (in Cluster 3) have the highest extreme weather elements: Dew, Hum, Precip, PCover, WDir, and CCover. Meanwhile, Cluster 3, consisting of 8 weeks, has no dominant characteristics.

Table 6. Cluster centroids resulting from clustering using average linkage

Variable	Data matrix 2024						Data matrix 2023					
	K=2 Centroid		*K*=4 Cluster centroid				*K*=2 Centroid		*K*=4 Cluster centroid			
	1	2	1	2	3	4	1	2	1	2	3	4
Tmax	0.48	0.19	0.39	0.72	0.00	0.38	0.34	0.82	0.34	0.14	0.45	0.82
Tmin	0.55	0.39	0.60	0.38	0.42	0.37	0.49	0.58	0.56	0.22	0.28	0.58
Temp	0.47	0.19	0.42	0.63	0.00	0.39	0.42	*0.81*	0.42	0.17	0.47	0.81
Dew	0.62	0.47	*0.74*	0.27	0.55	0.39	0.71	0.19	*0.76*	*0.75*	0.48	0.19
Hum	*0.72*	*0.73*	*0.84*	0.37	0.90	0.55	0.74	0.08	*0.77*	*0.95*	0.50	0.08
Precip	0.35	0.30	0.45	0.09	0.43	0.17	0.37	0.12	0.39	0.77	0.13	0.12
Pcover	0.55	0.66	0.67	0.22	0.91	0.41	0.51	0.17	0.53	0.95	0.28	0.17
Wgust	0.23	*0.96*	0.23	0.22	0.92	1.00	0.31	0.78	0.26	0.41	0.51	*0.78*
WSpeed	0.24	*0.88*	0.21	0.32	0.76	1.00	0.25	0.72	0.20	0.27	0.47	*0.72*
WDir	0.55	*0.90*	0.60	0.41	0.90	0.91	0.50	0.20	0.49	*0.92*	0.36	0.20
Seapress	0.44	0.05	0.45	0.43	0.10	0.00	0.49	0.60	0.49	0.37	0.50	0.60
CCover	0.73	*0.92*	*0.83*	0.46	0.98	0.86	0.75	0.29	*0.77*	*0.99*	0.57	0.29
SolRad	0.47	0.33	0.36	0.78	0.01	0.65	0.45	*0.90*	0.44	0.10	0.66	*0.90*
UVI	0.55	0.37	0.46	0.78	0.00	0.73	0.45	*0.94*	0.43	0.02	0.68	*0.94*
MooPhase	0.52	0.20	0.53	0.50	0.03	0.37	0.49	0.55	0.49	0.28	0.55	0.55
Number of Weeks	51	2	38	13	1	1	46	7	35	3	8	7

Note: Numbers in italics indicate the value of the variable that predominantly characterizes the cluster.

Overall, based on the average linkage results, the majority of Weeks in both data matrices have similar characteristics, namely high Dew, Hum, and CCover. In the 2023 data, the weather elements WGust, WSpeed, and WDir are more characteristic of cluster dissimilarity than in the 2024 data.

The optimal K value obtained in K-means clustering is 2. The minimum distance between two clusters in the 2024 and 2023 data matrices is 2.569 and 2.768, respectively. In both data matrices, the minimum distance between two clusters is higher than the minimum distance between four clusters. Meanwhile, in the average linkage results, the similarity level between two clusters in the 2024 and 2023 data matrices is 41.9% and 48.9%, respectively. The similarity level in two clusters is lower than the similarity level in four clusters. The results of the two clustering techniques give different results, namely there are differences in several dominant weather elements that characterize each cluster. This may be due to the random initial centroid selection in K-means clustering. Meanwhile, in average linkage, clustering time objects through amalgamation steps.

There are several things that can be obtained from the results of comparing the K-means clustering and the average linkage results for the weather elements, namely:

(i) In the data matrix 2024, Precip and PCover dominantly characterize the cluster of the K-means clustering results. Conversely, this is not the case with the average linkage results.
(ii) In the data matrix 2023, the K-means clustering results show that one cluster in almost all weeks lacks the weather elements that are dominant in determining their dissimilarity. However, in the average linkage results, this is only true for a small number of weeks.
(iii) K-means clustering results on the data matrix 2023, the weather elements that determine the dissimilarity between clusters tend to be the same.

The tendency of differences in weather elements that dominate a weekly time span within a 1-year period can provide information for coffee farmers regarding the time span from flowering to coffee fruit being ready for harvest, the length of the harvest period, and the arrangement of plant maintenance schedules, such as fertilization, weed control, pruning, and rejuvenation. This effective plant maintenance schedule activity is a form of mitigation against weather variability, thereby minimizing the impact of extreme weather on plant productivity, including coffee.

4 Conclusion

Comparing the similarity levels of the average linkage results for the data matrices 2024 and 2023, the weather elements across weeks in 2023 were more stable than across weeks in 2024. In the 2023 data, WGust, WSpeed, and WDir better characterize the dissimilarity of weather element clusters between weeks. K-means clustering results for the data matrix 2024 indicate that clusters of most weeks are characterized by higher minimum temperature, dew point, humidity, precipitation, and cloud cover. Meanwhile, in the data matrix 2023, clusters from most weeks are characterized by higher dew point, humidity, and cloud cover. The average linkage clustering results also show that clusters from most weeks are predominantly characterized by the same weather elements as the K-means clustering results.

Differences in weather elements that dominate a weekly time span within a year indicate weather variability in Pagaralam. If the condition of these weather elements is linked to coffee cultivation activities, the results of this clustering are useful in developing a strategic schedule for plant maintenance activities and also provide information for estimating coffee production, harvest time, and the duration of the harvesting process.

Because the weather elements characterizing clusters obtained from different cluster analysis methods can differ, further research should compare the clustering results from normalized data matrices with those from standardized data matrices. Furthermore, it is also necessary to compare the two data matrices for the two years using Procrustes analysis.

References

[1] Susilo, B.: Mengenal iklim dan cuaca di Indonesia. Yogyakarta: Diva Press (2021)

[2] Hamami, F. H. and Dahlan, I. A.: Klasifikasi cuaca Provinsi DKI Jakarta menggunakan Algoritma Random Forest dengan Teknik Oversampling. Teknoinfo. Vol. 16, No. 1, pp. 87-92 (2022)

[3] Chusyairi, A.: Clustering data cuaca ekstrim Indonesia dengan K-Means dan entropi. Journal of Informatics and Communications Technology. Vol. 5, No. 1, pp. 1–10 (2023)

[4] Guntoro, G., Lisnawati, L., Tantawi, A. R., Safni, I.: Analisis pengaruh cuaca terhadap produksi kopi di Tapanuli Utara, Sumatera Utara. Seminar Nasional Kedaulatan Pertanian 2024. Vol. 1, No. 1, pp. 401-408, Yogyakarta (2024)

[5] Saputro, E. N. and Nugraha, F. S.: Prediksi penjualan kopi berdasarkan cuaca menggunakan Association Rule dan Algoritma FP Growth. Jurnal Ilmiah Media Sisfo. Vol. 17, No. 1, pp. 1–8 (2023)

[6] Mbwambo, S. G., Mourice, S. K., Tarimo, A. J. P.: The impacts of current climate variability on coffee production in the Northern and Southern Highlands of Tanzania. Journal of Agricultural Science. Vol. 14, No. 3, pp. 78–99 (2022)

[7] Merga, W. and Alemayehu, D.: Effects of climate change on global Arabica Coffee (Coffea arabica L) production. Greener J. Plant Breed. Crop Sci. Vol. 7, No. 1, pp. 23–30 (2019)

[8] Rahn, E., Vaastd, P., Läderachb, P., Astenc, P. van, Jassognec, L., Ghazoul, J.: Exploring adaptation strategies of coffee production to climate change using a process-based model. Ecol. Modell. Vol. 371, pp. 76–89 (2018)

[9] Moreira, T. R., Silva, S. F. da, Silva, N. B. da, Santos, G. M. A. D. A. dos, Santos, A. R. dos.: Global warming and the effects of climate change on coffee production. In Food Engineering Series: Quality Determinants In Coffee Production. 1st ed., Switzerland: Springer Cham, pp. 65-100 (2021)

[10] A. C. W. Craparo, P. J. A. Van Asten, P. Läderach, L. T. P. Jassogne, and S. W. Grab, "Coffea arabica yields decline in Tanzania due to climate change: Global implications," *Agricultural and Forest Meteorology*, vol. 207, pp. 1–10, Jul. 2015, doi: 10.1016/j.agrformet.2015.03.005.

[11] F. M. DaMatta, C. P. Ronchi, M. Maestri, and R. S. Barros, "Ecophysiology of coffee growth and production," *Brazilian Journal of Plant Physiology*, vol. 19, no. 4, pp. 485–510, 2007, doi: 10.1590/S1677-04202007000400014.

[12] Badan Pusat Statistik. Statistik kopi Indonesia 2023. Vol. 8. Jakarta: BPS (2024)

[13] BPS. Kota Pagar Alam dalam Angka 2023. BPS Kota Pagar Alam (2023)

[14] BPS, Kota Pagar Alam dalam Angka 2024. BPS Kota Pagar Alam (2024)

[15] Direktorat Jenderal Perkebunan. Statistik Perkebunan 2023-2025. Jilid 1. Jakarta: Kementerian Pertanian (2024)

[16] Ngudiantoro, Irmeilyana, Samsuri, M. N.: Binary logistic regression modeling on net income of Pagar Alam coffee farmers. International Journal of Applied Sciences and Smart Technologies. Vol. 2, No. 2, pp. 137–156 (2020)
[17] Irmeilyana, Ngudiantoro, Samsuri, M. N., Suprihatin, B.: Logistic regression model on land productivity of Pagar Alam coffee farming. In Journal of Physics: Conference Series, IOP Publishing Ltd. (2021)
[18] Irmeilyana, Ngudiantoro, Maiyanti, S. I.: Regression model on Pagaralam coffee farmers' income with the influence of the use of reductant variable. Barekeng. Vol. 16, No. 2, pp. 409–420, (2022)
[19] Irmeilyana, Ngudiantoro, Maiyanti, S. I.: Application of two groups analysis and cluster analysis on comparison of characteristics of Pagaralam coffee farmers categories. Barekeng. Vol. 17, No. 1, pp. 101-112 (2023)
[20] Wati, S. E. and Sukmono, A.: Analysis of the Relationship between Climate Variability and Coffee Productivity using Remote Sensing and Meteorological Data. Jurnal Agroekoteknologi Indonesia. Vol. 10, No. 2, pp. 115-125 (2022)
[21] Pasmah, K, Dhiniati, F., Azizah, B.: Penilaian Indeks Ancaman Kekeringan di Kecamatan Dempo Tengah Kota Pagaralam menggunakan Gis. Bering's. Vol. 09, No. 01, pp. 23–30 (2022)
[22] Irmeilyana, Suratama, B., Maiyanti, S. I., Ngudiantoro, Suprihatin, B, Desiani, A.: Deskripsi dan uji hipotesis data elemen-elemen cuaca menggunakan uji selisih mean dan uji rasio variansi pada Tahun 2021-2022 di dua kecamatan penghasil kopi. Jurnal Penelitian Sains. Vol. 27, No. 2, pp. 1–11 (2025)
[23] Irmeilyana, Suprihatin, B., Desiani, A., Ngudiantoro, Maiyanti, S. I.: Graphical representation and two groups analysis on data matrix of Robusta green cherries production in two harvest. Barekeng. Vol. 19, No. 2, pp. 1279–1294 (2025)
[24] Johnson, R. A. and Wichern, D. W.: Applied multivariate statistical analysis. 6th ed. New Jersey: Prentice-Hall, Inc. (2007).
[25] Husson, F., Lê, S., Pagês, J.: Exploratory multivariate analysis by example using R. 2th ed. London: Chapman & Hall/CRC Press (2017)
[26] Siregar, A. M.: Klasifikasi untuk prediksi cuaca menggunakan esemble learning. PETIR. Vol. 13, No. 2, pp. 138–147 (2020)
[27] Rinaldi, I.: Penggunaan algoritma clustering K-Means untuk mengelompokkan pola cuaca. Journal of Science and Social Research. Vol. 8, No. 1, pp. 343–348 (2025)
[28] Gabriel, K. R.: Goodness-of-fit of biplots and correspondence analysis. Biometrika. Vol. 89, No. 2, pp. 423-436 (2002)
[29] Prasetyo, B. and Wiguna, A.: Comparative Study of Non-Hierarchical and Agglomerative Clustering for Regional Weather Pattern Identification. Indonesian Journal of Computing and Modeling. Vol. 7, No. 1, pp. 45-53 (2024)
[30] Greenacre, M.: Biplots in practice. Madrid: BBVA Foundation. (2010)
[31] Thakare, Y. S. and Bagal, S. B.: Performance evaluation of K-means clustering algorithm with various distance metrics. International Journal of Computer Applications. Vol. 110, No. 11, pp. 975–987 (2015)
[32] Desgraupes, B.: Clustering indices (Package clusterCrit for R). University Paris Quest (2017)
[33] Kingrani, S. K., Levene, M., Zhang, D.: Estimating the number of clusters using diversity. Artificial Intelligence Research. Vol. 7, No. 1, pp. 15–22 (2017)

[34] Muningsih, E., Maryani, I., Handayani, V. R.: Penerapan metode K-Means dan optimasi jumlah cluster dengan Index Davies Bouldin untuk clustering propinsi berdasarkan potensi desa. Jurnal Sains dan Manajemen. Vol. 9, No. 1, pp. 95–100 (2021)
[35] Nugroho, H. and Lestari, D.: Evaluation of Cluster Validity Indices for High-Dimensional Environmental and Climate Datasets. Journal of Data Science and Environmental Statistics. Vol. 15, No. 3, pp. 201-210 (2023)