Iraqi Journal of Statistical Sciences

http://stats.uomosul.edu.iq

Determining Climate Extremes of Mosul Weather Using Robust Noise Clustering Strategy

Marwan Moysar AL-Hyali[1] , Bashar AL-Talib[2]

[1,2]Department of Statistics and Informatics, College of Computer Science and Mathematics University of Mosul, Mosul, Iraq

Article information

Article history:
Received July 10 ,2024
Revised: December 20,2024
Accepted December 1,2024
Available June 1, 2025

Keywords:
Variables Clustering, Latent Components Noise Cluster, High-Dimensional Data, Local Groups

Correspondence:
Marwan M. Mahmood
marwan.22csp62@student.uomosul.edu.iq

Abstract

The paper aims to analyse the climatic data of the city of Mosul during the summer season from 2013 to 2022, focusing on the maximum temperature variable. Modern methods have been used to detect climate fluctuations that have not been used previously, adapt them to the study data, and explore the general and extreme climates that any of the previous studies have not touched. Variable clustering techniques have been used to discover the latent components according to the local groups model. The "K+1" noise group strategy was used to identify high-noise variables. The researcher proposed a wide format for ordering the data: P > N, which means that the number of variables is greater than the number of observations. The observations represented the school years; the variables were the summer days for three months (June, July, and August). This arrangement proved suitable for the variable aggregation technique of high-dimensional data. The results showed six groups, five of which were almost homogeneous. The five clusters indicate different patterns of maximum temperature increases during the summer. The first cluster highlights heat waves in mid-summer (July and August), while the second cluster focuses on the hot ends of summer (late June and August). The third cluster refers to early and continuous heat waves in June and July, while the fourth cluster reflects persistent heat in late July and August. The fifth cluster shows a variation in temperatures between the beginning and the end of summer. The excluded noise variables represent inconsistent data or outliers that did not belong to any cluster. This contributes to improving the accuracy of climate models. The results highlight characteristic climatic patterns and provide recommendations for strengthening environmental and agricultural planning.

DOI 10.33899/iqjoss.2025.187755, ©Authors, 2025, College of Computer Science and Mathematics University of Mosul.

1. Introduction

Climate change assessments in Iraq, based on forecasts, indicate that temperatures are increasing while rainfall is decreasing. An analysis of rainfall and temperature data from several locations in Iraq shows that rainfall will continue to decline and temperatures will rise in the future. Rainfall occur over shorter periods and be more intense, leading to higher sediment transport rates, reducing storage capacity and agricultural yields (Ali et al., 2024).

The city of Mosul, located in northern Iraq, experiences a dry continental climate. During the summer season, the city is subject to intense heat waves that lead to a significant rise in maximum temperatures. This increase in temperature has a substantial impact on the daily lives of the local population and affects various economic and agricultural activities. Mosul's climate is generally characterized by hot, dry summers and cold, wet winters. The city’s geographic location, far from large bodies of water, makes it susceptible to considerable temperature fluctuations between seasons. Summer, lasting from June to August, is the hottest season. The maximum temperatures during the summer in Mosul are among the highest annual averages, sometimes exceeding [45 (46-49)] degrees Celsius. These temperature levels are influenced by a variety of factors, including (Hassan Ali & Rashad Shaheen, 2013):

1. **Geographic Location**: Mosul is situated in an inland region far from the moderating effects of marine influences, leading to significant temperature increases.
2. **Atmospheric Pressure**: The city is affected by high atmospheric pressure during the summer, contributing to the rise in temperatures.
3. **Dry Winds**: Winds coming from neighboring desert areas increase dryness and elevate temperatures.

The significant rise in maximum temperatures during the summer in Mosul has multiple effects (Al-Hafith et al., 2017):

1. **Public Health**: There is an increase in cases of heat stress and heat strokes, particularly among vulnerable groups such as children and the elderly.
2. **Agriculture**: The high temperatures negatively impact crops that require large amounts of water, increasing irrigation and evaporation challenges.
3. **Infrastructure**: The heavy use of air conditioning puts additional pressure on electricity systems, sometimes leading to power outages. This is what happens during the summer with continuous power outages.

The distinct objective of cluster analysis is to provide a physical classification of weather and climate patterns for several purposes, one of which is to provide a good classification of the fundamental surface climates on Earth, and another is to better understand and predict the occurrence of extreme weather phenomena (storms and floods) and extreme climate phenomena (heat and cold waves) (Straus, 2019).

Clustering of variables is the task of grouping similar variables into different groups. It may be useful in several situations such as dimensionality reduction, feature selection, and detect redundancies (Ghizlane et al., 2021).

The approach discussed here is to cluster variables around latent components. More specifically, the goal is to simultaneously determine K clusters of variables and K latent components such that the variables in each cluster are highly correlated with the corresponding latent component. The solution to this problem is provided by an iterative partitioning algorithm that consists of selecting K initial clusters as a first step. To help practitioners choose the appropriate number of clusters and initial partitioning, a hierarchical clustering approach is proposed. This hierarchical approach shares the same logic as the partitioning algorithm in that both techniques aim to maximize the same criterion. The aim of this research is to identify the extreme days with maximum temperatures in the summer of Mosul city that do not match the prevailing climate according to the specified study years (Vigneau & Qannari, 2003).

2. Methodology

2.1. The Hierarchical Clustering of Variables with Consolidation Method

Let's look at a set of p variables that were seen in n observations. We indicate $X_j = (x_{1j}, x_{2j}, \dots, x_{nj})^t \in \mathbb{R}^n$, the vector of observations for the j^{th} variable. Every variable that has been observed $X_j (j = 1, \dots, p)$ is supposed to be centered. In addition, we may choose to standardize them, or not, to a unit variance. Given a number of clusters, K, the goal of the CLV (clustering of variables around latent components) method is to seek a partition of the observed variables into K groups $(G_1, G_2, \dots, G_K)$ and K latent variables, $(c_1, c_2, \dots, c_K)$ associated with each group respectively, to enhance groups' internal cohesion. A single kind of criterion is taken into, which defines a single kind of group: (I) "local groups" only variables with positive correlations will be included in the same group.

For local groups, the clustering criterion is *S*, described as:

$$S = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} cov\left(X_j, c_k\right), \ \ with\ var(c_k) = 1 \qquad (1)$$

In (1), δ_{kj}=1 if the j^{th} daily variable is a group member G_k, and $\delta_{kj} = 0$ else. $cov(X_j, c_k)$ stands for the covariance between the variable X_j and latent variable c_k and $var(c_k)$ is the variance of c_k.

S is optimized using a partitioning algorithm. Two steps alternate: the assignment step, during which the δ_{kj} are defined given the latent variables c_k, for k = 1, … , K, and the estimation step of the latent variables, given the partition of the variables. Mor precisely, this algorithm is considered as follows (Vigneau, 2016):

1. Initialization step. K clusters are created randomly or chosen at random from the nested partitions that a hierarchical algorithm produces. In the case of random initialization, multiple random starts (let's say 100) are made, and the solution that maximizes the value of the clustering criterion, S (Eq. 1), is chosen. Running an ascendant hierarchical clustering beforehand depending on the taken clustering criterion is a suitable option for a non-random initialization. This method

works step-by-step from a stage when each variable is a group by itself to a stage where they are all combined. The two variable clusters that result in the least drop in the clustering criterion at a particular level are combined. This method of initializing the alternating optimization process is recommended when the number of variables is not too great (less than a few hundred, for example), even if it requires more computational resources to form a hierarchy. This method provides a relevant initial solution.

2. Estimation step. In a cluster $G_k (k = 1, \dots, K)$, the latent variable c_k is defined as:

- For local groups, as the standardized mean variable of the variables in G_k.

3. Assignment step. Each variable $X_j (j = 1, \dots, p)$ is considered in turn. X_j is assigned to the cluster G_k for which its covariance coefficient, for local groups, with c_k is greater than other group latent variables combined. Formally, More,

- For local groups:

$$\delta_{kj} = 1 \; if \; \max_l \{cov\left(X_j, c_1\right)\} = cov\left(X_j, c_k\right) \qquad (2)$$

4. Continue steps 2 and 3 until the partition is stable.

All of the variables in this algorithm are allocated to a single group. The anomalous or noisy variables ought to be eliminated, ideally. To study a changed approach, this question was addressed.

2.2. "K+1" Strategy

This strategy consists in introducing an additional cluster for handling the atypical or noise variables in the clustering. This additional cluster, also named the "noise cluster", can be represented by a prototypical variable that is expected to have the same correlation with all the observed variables $X_j (j = 1, \dots, p)$. The CLV criterion is consequently updated and a fixed parameter, ρ, representing the common correlation coefficient associated with the definition of the "noise cluster" prototype, is introduced. According to the type of groups sought, this consists of maximizing:

For local groups, a new criterion S:

$$S_{new} = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \, cov\left(X_j, c_k\right) + \sum_{j=1}^{p} (1 - \sum_{k} \delta_{kj}) \, \rho \sqrt{var\left(X_j\right)} \qquad (3)$$

With var$(c_k) = 1$. In (3), The final term is the criterion's additional "noise cluster" contribution. Specifically, if the variable X_j is a member of one of the primary groups $G_k (k = 1, \dots, K)$, we have $\delta_{kj} = 1$ for a given k and $(1 - \sum_k \delta_{kj}) = 0$. Conversely, if the variable X_j is not a part of any major groups, $(1 - \sum_k \delta_{kj}) = 1$. The parameter ρ determines how many variables will be included in the "noise cluster". The majority of variables will be allocated to one of the groups G_k rather than the "noise cluster" if ρ is set extremely tiny. A large number of variables will be assigned to the "noise cluster" if ρ is big. A tuning parameter, ρ, must be selected between 0 and 1. Selecting $\rho = 0$ results in the fundamental CLV criterion (Eq.1). All variables will belong to the "noise cluster" when $\rho = 1$.

The same kind of algorithm as the one outlined in section CLV technique can be used to maximize the criterion S_{new}; however, the assignment step's construction differs and is as follows:

Assignment step. Unless this number is very tiny in comparison to the value of the ρ parameter, a variable $X_j$will be assigned to cluster G_k if the covariance between $X_j$and c_k is greater than with other latent variables. Formally speaking, we have:

- For local groups,

$$\begin{cases} \delta_{kj} = 0 \; \forall k \; if \max\left\{max_l\{cov\left(X_j, c_1\right)\}, \rho\sqrt{var\left(X_j\right)}\right\} = \rho\sqrt{var\left(X_j\right)} \\ \delta_{kj} = 1 \qquad if \max\,\{max_l\{cov(X_j, c_1)\}\;, \rho\sqrt{var\left(X_j\right)}\;\} = cov(X_j, c_k) \end{cases} \qquad (4)$$

By the construction of criterion S_{new}(3), there is a correlation coefficient equivalent to the tuning parameter ρ. (Vigneau, 2016).

(i) For each daily variable j, $if \; cov\left(X_j, c_k\right) < \rho\sqrt{var\left(X_j\right)}, for \; all \; k = 1, \dots, K,$ or equivalently, $if \; cor\left(X_j, c_k\right) < \rho \; for \; all \; k = 1, \dots, K,$ where cor() stands for the correlation coefficient, the j^{th} daily variable will be assigned to the "noise cluster"

(ii) Otherwise, daily variable j will be assigned to the group G_k for which $cov(X_j, c_k) > cov(X_j, c_{k'})\ \forall k' \neq k$ and $cov(X_j, c_k) > \rho\sqrt{var(X_j)}$.

The number of daily variables that will be in the "noise cluster" depends on the value of the parameter ρ. Is akin to a positive correlation coefficient and ranges between 0 and 1. If ρ is chosen to be close to 1 then almost all the daily variables are likely to have a correlation coefficient with any latent variable smaller than the value of ρ and the size of the "noise cluster" will be large. Contrariwise, if ρ is close to 0, the "noise cluster" will be almost empty. As a matter of fact, with daily climate data, it may occur that the "noise cluster" is not empty, even if $\rho = 0$, if there are daily variables whose direction of climate extremism is negatively correlated with all the group latent variables $c_k(for\ k = 1, \dots, K)$. (Vigneau et al., 2016).

Bootstrapping on the variables (in column) is performed. Choose the "column" option, if the variables are taken from a population of variables. e.g. when variables are days assessing specific years. Each bootstrapped data matrix is submitted to CLV in order to get partitions from 1 to nmax clusters. For each number of clusters, K, the Rand Index, the adjusted Rand Index, as well as the cohesion and the isolation of the clusters of the observed partition and the bootstrapped partitions are computed. These criteria are used for assessing the stability of the solution in K clusters. Parallel computing is performed for time-saving. The process of bootstrapping variables in columns to enhance cluster analysis. If variables are sampled from a larger population (e.g., days representing specific years), the "column" option is selected. Bootstrapping generates new data matrices, which are then submitted to the CLV algorithm to obtain partitions ranging from 1 to nmax clusters. Stability of the clustering solution is assessed by calculating metrics such as the Rand Index, Adjusted Rand Index, cohesion, and isolation for both the original and bootstrapped partitions. Parallel computing is employed to save time during the process.

.

3. Study Data

The remote sensing center at the University of Mosul was credited with providing climate data from the ministry of agriculture, agricultural meteorology center, and Nineveh governorate - Mosul station with longitude E 43.16 and latitude N 36.33 for the period between 2013 and 2022. Actual data monitored by that station for variable maximum temperatures is an example of 92 different climate variables because of space limitations.

4. Order of Data

In the context of high-dimensional data with a large number of variables, P, and a small number of observations, N. The long format of ordering the data is a common traditional approach to dealing with multivariate data in climatology (N > P). Therefore, a wide-format method for ordering the data for dealing with climate variables has been proposed, which is called high-dimensional data (P > N). Also, the interpretation and visualization of data are less clear with the traditional method compared to the high-dimensional method, which works to build a new structure for the data that is different from the structure of the traditional method, which gives results with more accuracy and clarity. The maximum temperature variable has 92 daily variables throughout the summer (June, July, and August) for ten years and has 10 observations (the years), which are the observed years for that variable. The R program version (4.3.3) and Excel version (2019) were used to extract illustrations and results (Vigneau, 2020).

5. Results

5.1. Review and Summarize Climate Data

After arranging the climate data in a table and examining the cells and ensuring that they are free of empty values, at this stage we will review the climate data and visualize it in an illustrative form that shows us the straight annual trends of the maximum temperature variables in the summer season, as follows:

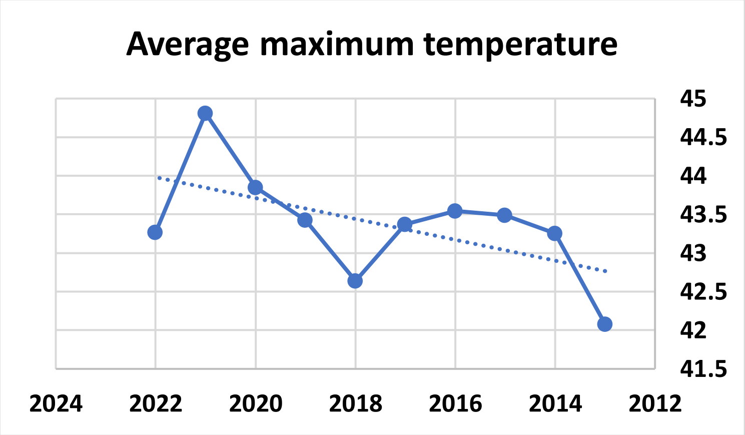

Figure 1: Line chart to display AT Max trends.

Interpretation: The x-axis in figure 1 above represents the years of study, and the y-axis represents the period of average annual maximum temperatures, measured in Celsius C°, for the summer season. It is clear to us from displaying the data form above that the trend of average maximum temperatures is positive.

5.2. Clustering of Variables Around Latent Variables Using the Classical CLV Method

We will take into account the data collected in the Excel sheet. The goal is to divide the days into groups or clusters with similar patterns of years, which is the method of diagnosis of internal climatic climate, so that these clusters are as homogeneous as possible and have a set of underlying variables, each of which is linked to the cluster. These underlying variables make it possible to determine the main directions of climate extremism in the data set (i.e., the most extreme days of these years). The strategy mainly consists of hierarchical cluster analysis followed by a repetitive division algorithm. Both algorithms aim to maximize the same criterion, which reflects the extent of variables in each group with the underlying variable associated with this group, and this is clarified with fees. The CLV method allows the set of variables to be of two different types: (i) directional groups that combine the variables associated with direct and oppositely, and (ii) local groups that merge the changes that are directly related only. Since our goal is to separate days that have different extreme trends for each climate season separately, the situation of local groups will be considered. If dealing with all climatic seasons uniformly, the goal would be to consider the state of directional groups only. And this method aims to determine a simple perfect structure, meaning each variable has one non-zero load for one latent variable. This method is used to reduce the dimensions of data and explain complex problems more easily, unlike the principal components. The components of this method are not perpendicular, and they are not designed to take into account a greater amount of total contrast, but they may be more important in terms of interpretation.

Step 1: We start loading the variable chart from the principal component analysis. The x - axis depends on the first principal component (i.e., the first dimension), and the y-axis is the second principal component (i.e., the second dimension). And we have the following plot:

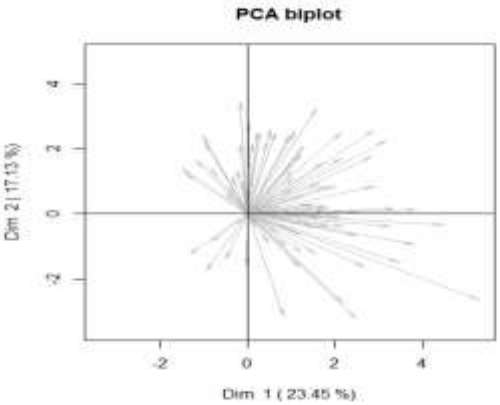

Figure 2: Biplot of the internal Climate mapping for the compote dataset

Interpretation: Since the CLV method is based on two types of variables (directional and local), we will consider the method of local variables that integrates only positively related variables based on the figure above. We note that there are no variables that are significantly negatively or positively correlated in the figure above.

Step 2: We draw the "Delta Plot," which depicts the decrease in the criterion *S* through different numbers of summer days after the additional integration phase that follows the hierarchy CLV. As in the following form:

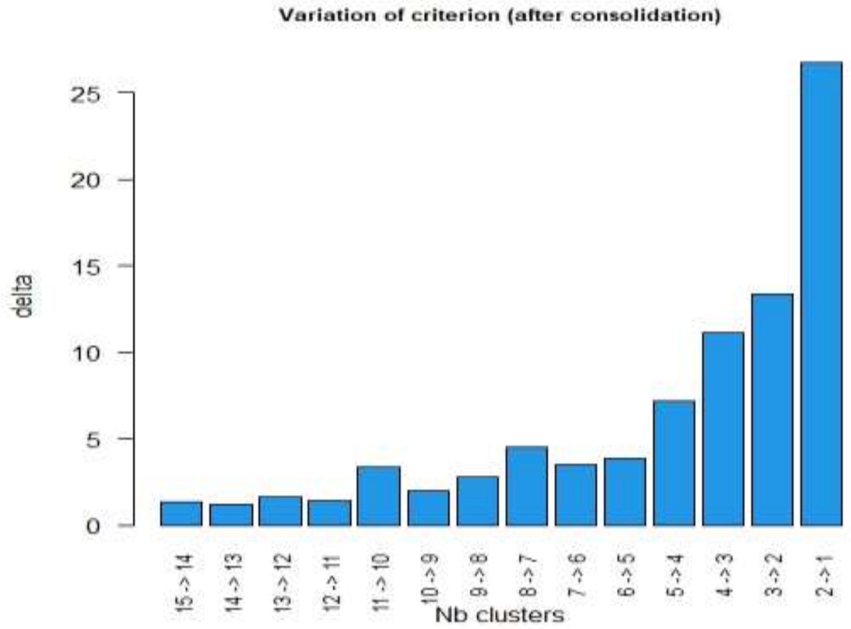

Figure 3: Evolution of the aggregation criterion

Interpretation: The delta division scheme allows a clearer decision in favor of 5 sectors of climatic days (because the decrease in *S* is relatively pronounced when moving from 5 to 4 sectors). The x-axis represents the number of divisions for sectors (nmax= 15), and the y-axis represents the delta standard for maximum temperatures. This graph of the difference in hashing criterion between splitting into k clusters and splitting into k-1 clusters after merging is useful for determining the number of clusters to keep.

Step 3: is to describe the groups of variables (the days) in a two-dimensional space obtained by principal component analysis, as shown in the following figure:

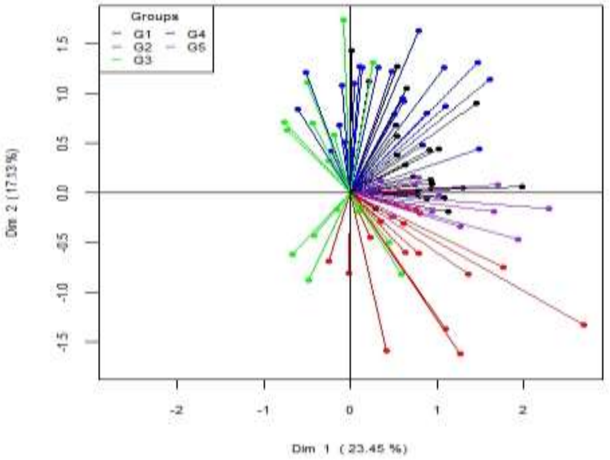

Figure 4: Group membership to be divided into five clusters

Interpretation: G1 (blue), G2 (red), G3 (green), G4 (black), and G5 (violet) indicate the five groups identified based on the analysis. Each group represents a group of variables that are similar in behavior and trends within the dimensions described. Dim 1 and Dim 2 are the basic dimensions that have been used to represent variables in two-dimensional space. The percentage attached to the dimensions (23.45% and 17.13%) indicates the amount of variance in the data that is explained by each dimension.

Points clustered close together indicate maximum temperatures that behave similarly over the specified periods. Points farther from the center of the 2D space indicate larger differences in values, which means they may be outliers or have different properties.

- G1 (blue): Points in this group are clustered close together, indicating relatively low variation in temperature extremes.
- G2 (red): Shows greater divergence, which may indicate greater variation in temperature extremes within this group.
- G3 (green) and G4 (black): These groups show greater divergence and dispersion, which may indicate greater variation in temperature extremes in the different time periods.
- G5 (violet): This group shows greater divergence, indicating large differences in temperature extremes.

These analyses will help in better understanding climate patterns in the city of Mosul and provide useful data for various purposes. The distribution of days across groups reflects daily and weekly variations in maximum temperatures, which helps in understanding patterns.

Step 4: In addition to the drawing above, we will draw an intuitive graph of cluster similarity based on principal component analysis, as follows:

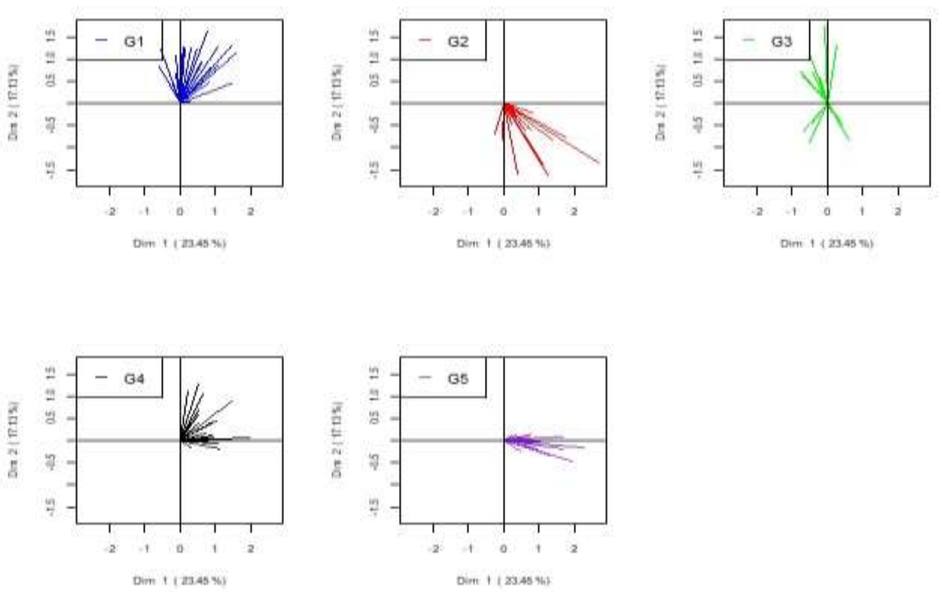

Figure 5: Sectional loading diagrams for each cluster

Interpretation: The x-axis displays loadings related to the first principal component; the y-axis displays loadings related to the second principal component. The first principal component explains approximately 23% of the variance in portion benefits among all daily variables for the highest summer temperatures, while the second principal component explains an additional 17%. We see that he expressed all the daily variables of the summer. Therefore, visual perception is considered simple because it does not explain 60% of the variance in years of schooling. When comparing the clusters of daily variables, it becomes clear that the first part (blue) is, by trend, closer to the fourth part (black) compared to the second part (red), because clusters 1 and 4 indicate the same direction, while the second cluster (red) indicates a different direction. Vectors pointing in opposite directions represent days whose temperatures vary greatly. Vectors perpendicular to each other represent days that have different temperatures with respect to the levels of other daily variables.

Step 5: is to include a table of the latent components associated with the five clusters resulting from the CLV. We see each latent component (from principal components analysis) associated with a cluster from the CLV method as follows:

Table 1: Latent components associated with clusters

```
          Comp1       Comp2      Comp3       Comp4      Comp5
2013  -2.6048913 -1.15778125 -0.9675938 -0.84309091 -0.4378333
2014   1.1046739 -2.50403125  1.5674062  0.31145455 -1.8711667
2015   0.9494565  1.68034375 -2.4738438  0.20100000 -0.1905000
2016  -1.6514130  1.32159375  0.8720937  1.42463636 -0.8365000
2017   1.0270652 -0.41184375 -0.4532188  0.25622727 -1.0308333
2018  -0.2970652 -0.90528125 -0.3057188 -1.41990909 -0.6925000
2019  -0.1505435  1.74221875  1.9249062 -2.12809091 -0.2345000
2020   1.6659783 -0.06903125 -0.9857188 -0.31309091  1.9541667
2021   0.4407609  0.85846875  0.9467812  2.46259091  2.5955000
2022  -0.4840217 -0.55465625 -0.1250938  0.04827273  0.7441667
```

Interpretation:

1.**Component 1 (Comp1)**:

○ Negative values indicate years with lower-than-average temperatures.
○ Positive values indicate years with higher-than-average temperatures.

2.**Component 2 (Comp2)**:

○ Negative values indicate years with greater temperature variations.
○ Positive values indicate years with smaller temperature variations.

3.**Component 3 (Comp3)**:

○ Negative values indicate years with stable temperatures.
○ Positive values indicate years with fluctuating temperatures.

4.**Component 4 (Comp4)**:

○ Negative values indicate years with lower deviations in temperatures.
○ Positive values indicate years with higher deviations in temperatures.

5.**Component 5 (Comp5)**:

○ Negative values indicate years with temperatures closer to the average.
○ Positive values indicate years with temperatures far from the average.

Interpretation of Key Points:

- **Year**:

○ Lower-than-average temperatures (Comp1 = -2.6048913).
○ Smaller temperature variations (Comp2 = -1.15778125).
○ Stable temperatures (Comp3 = -0.9675938).

- **Year**:

○ Higher-than-average temperatures (Comp1 = 0.4407609).
○ Smaller temperature variations (Comp2 = 0.85846875).
○ Relatively stable temperatures (Comp3 = 0.9467812).
○ Large deviations in temperatures (Comp4 = 2.46259091).
○ Temperatures far from the average (Comp5 = 2.5955000).

Changes over the years: Significant changes can be observed in the pivot values over the years. For example, the year 2020 shows high values for Comp1 and Comp5, indicating a large variation in temperature extremes.

Anomalous behavior: 2019 shows anomalous behavior in Comp3 and Comp4, where the values are very high and very negative, respectively. This indicates that there was an abnormal weather event or exceptional circumstances that year.

Principal Component Analysis provides insight into the different patterns of maximum temperatures during the summer season in Mosul over the ten-year period. The five components help in understanding the variations, stability, and deviations in temperatures during this time frame.

Step 6: We explain the stability of the clusters according to the results of the CLV method, using the bootstrap method through the rand index, the adjusted rand index, the cohesion or interconnectedness within the clusters, and the isolation between the clusters. The following figure shows this:

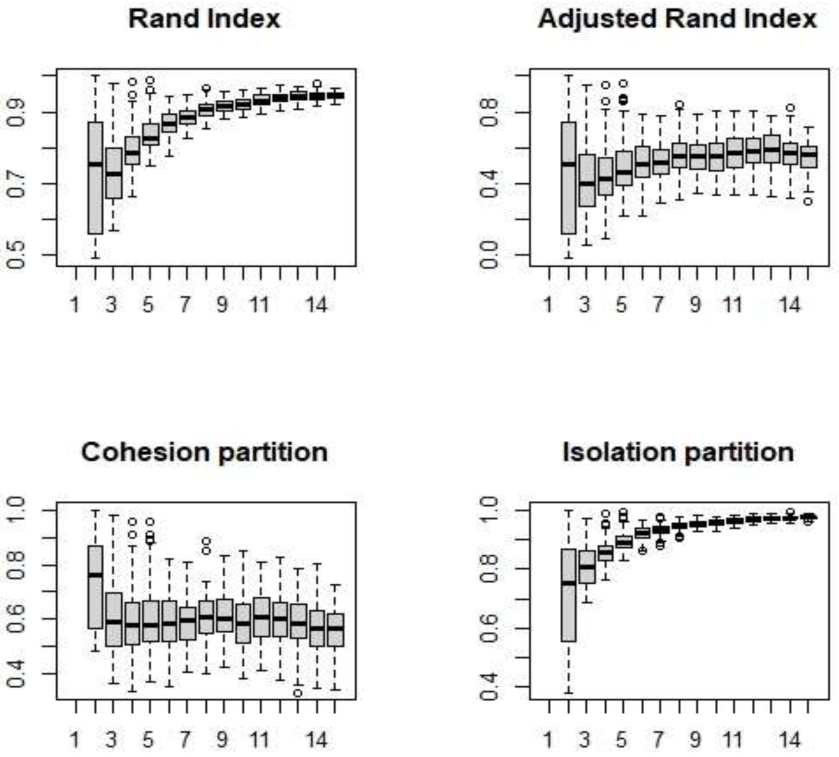

Figure 6: Bootstrap to evaluate the stability of CLV results

Interpretation: The bootstrap method indicated that the rand index result for clustering number 15 is 0.96616883, suggesting a high similarity between the clustering result and the original data categories. This indicates that the identified clusters align well with the actual clusters in the data. The adjusted rand index value was 0.67705475 for clustering number 15, suggesting accurate clustering after correcting for the effect of chance. The cohesion value was 0.6601942 in the final clustering, which measures how tightly data points are grouped within a cluster, indicating high intra-cluster similarity. The separation value was 0.9832985 in the final clustering, measuring how distinct the clusters are from each other. Given these metrics:

- Rand Index and Adjusted Rand Index values close to 1 indicate a high level of clustering accuracy.
- Cohesion values close to 1 indicate strong intra-cluster cohesion.
- Separation values close to 1 indicate effective inter-cluster separation.

Therefore, based on the very strong values across all metrics in the final clustering, the cluster analysis can be interpreted as successful, indicating well-defined clusters in the data. This result is positive, reflecting strong and accurate clustering (Vigneau et al., 2022).

5.3. Clustering of Variables Around Latent Variables Using the CLV_Kmeans Method

This procedure takes less time when the number of variables to be merged is large. The number of clusters must be specified before execution, which is one of the characteristics of this procedure. This method features two strategies for detecting outlier cases and provides solutions for merging partitions and allocating influential cases, automatically setting them aside (which are observations that significantly differ from others in one or more characteristics).

The first strategy is the kplusone (K+1) approach, which sets aside atypical (or unusual) variables into a noise cluster. The concept of this approach is to allocate variable j to cluster G_k when the correlation between daily variables x_j and the latent component centers c_k is high and positive. This method assumes a lower threshold (for equation processing error); if variable j (equation processing error) = cor (x_j, c_k) fails to exceed this threshold for any center of variable partition, then this variable will be allocated to the noise cluster. The choice of the threshold (equation processing error) can be arbitrary, but often 0.3 is used as a constant.

The second strategy is the sparselv (sparse LV) approach, which involves assigning a zero loading. The soft-thresholding algorithm is adapted for sparse principal component analysis. The thresholding procedure depends on the parameter ρ\rho as in the (K+1) strategy, and this approach has been explained previously and will not be used in this thesis.
The CLV_kmeans() function will be used instead of the CLV() function. This function eliminates hierarchical clustering and instead uses random starts *nstart* in a kmeans-like algorithm to search for the range with the highest resulting value targeted for the *S* criterion. Thus, outlier detection and the merging of partitioning solutions will be combined into a single function. The work in this thesis will be limited to the first strategy only.
Step 1: The number of clusters is 5, as previously determined in the third step and according to the CLV() function. In this step, we will determine the correlation coefficient rho according to the figure below:

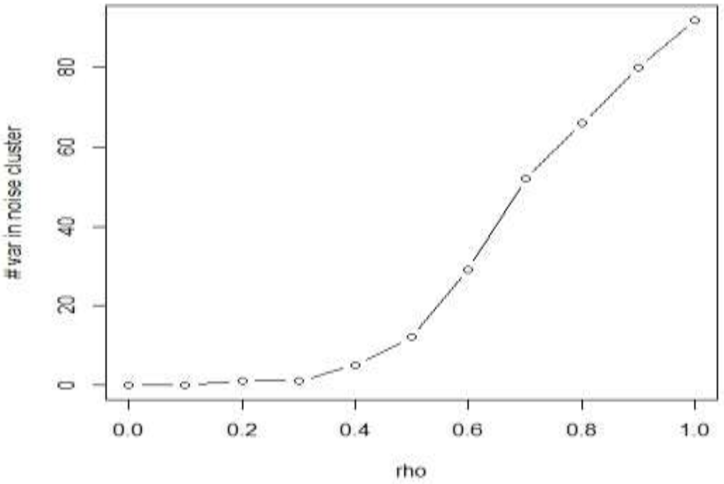

Figure 7: The number of variables to be assigned to the noise cluster
Interpretation: The clustering result showed that the value of the correlation coefficient is (0.4) and the number of variables in the noise cluster is (5). The x-axis shows the value of the correlation coefficient, and the y-axis shows the number of variables that are the focus of the study.
Step 2: In this step, we will insert a figure showing the internal assignment of variables to the highest temperature data for the summer of 2013-2022, with five clusters of daily variables numbered according to each cluster (black lines), and the unnumbered variables appear aside (gray lines) and are called the noise cluster. We can also see that some variables are better explained by the loading plot (long vectors) than others (short vectors). According to the figure below:

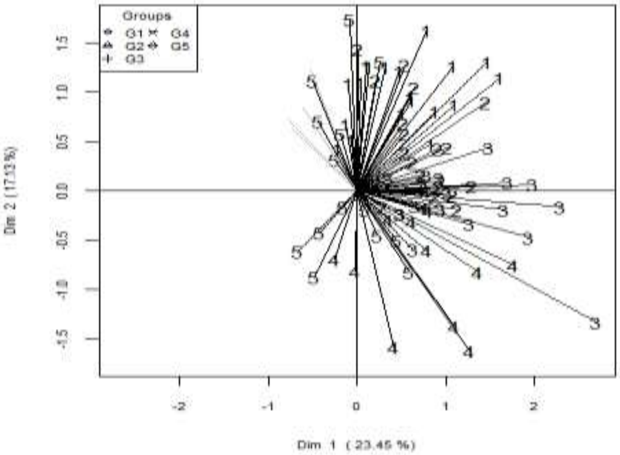

Figure 8: Internal assignment to the K+1 strategy while setting aside the noisy variables

Step 3: We will include a figure that shows the similarity of the clusters based on principal component analysis. The figure below shows this:

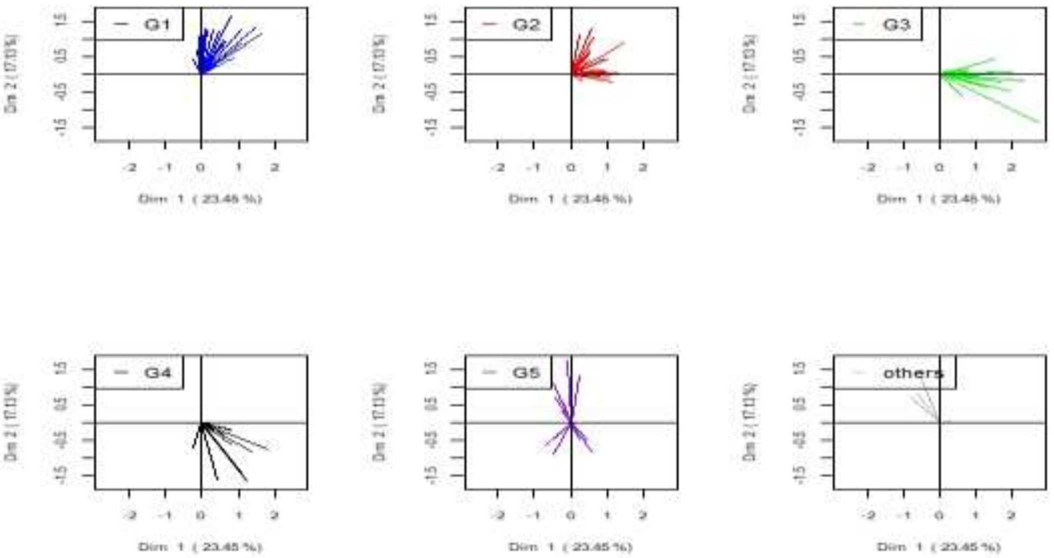

Figure 9: Partition loading plot for each variable using the first two principal components

Interpretation: The figure shows five clusters. Each cluster is plotted using the first two principal components, and the sixth cluster (others) represents recessive variables, which is the noise cluster. It becomes clear that cluster 1 is closer by direction to cluster 2 than to cluster 3, because clusters 1 and 2, by direction, point in the same direction. The last plot in the output is for the variables in the noise group. As we can see, the vectors for the daily variables differ from those for the other clusters and, in general, are not well explained by PCA (Vigneau et al., 2015).

Step 4: In this step, we will include the noise cluster, that is, the variables that have been set aside, as in the table below:

Table 2: Noise variables for the maximum temperature variable

```
var_set_aside    ID
X6_Jul_AT_Max    36
X23_Jul_AT_Max   53
X16_Aug_AT_Max   77
X17_Aug_AT_Max   78
X18_Aug_AT_Max   79
```

Interpretation: The days that do not follow the patterns identified by the other groups in the analysis are noisy variables. These days show a significant difference from the usual patterns of maximum temperatures in the selected groups. These may represent days of unusual weather conditions or extremes, such as heat waves or unusual storms. These days may not fit the usual seasonal patterns of the rest of the data, perhaps due to certain local effects such as changes in vegetation, human activity, or local topographic effects, and this could indicate that these days do not follow the general temperature pattern of the remaining days in each year. We can consider them climate fluctuations in the summer season for the years 2013–2022, for the city of Mosul**.**

Table 3: Daily noise variables data

Years	**X6_Jul**	**X23_Jul**	**X16_Aug**	**X17_Aug**	**X18_Aug**
2013	43.71	44.4	44.68	45.91	43.13
2014	45.87	44.4	46.89	46.63	47.29
2015	42.33	45.83	42.28	43.48	44.8
2016	41.82	44.27	43.28	42.2	42.3
2017	44.1	45.115	44.585	45.055	46.045
2018	44.22	42.37	44.96	45.87	43.33

2019	43.39	44.27	44.38	41.74	43.13
2020	41.81	45.64	44.51	45.16	41.97
2021	43.61	43.45	43.74	44.32	43.99
2022	44.79	43.26	43.64	43.72	45.45

Interpretation: The results presented represent the maximum temperature values on specific days for the summer season, and we will analyze them in detail for each group:

- **X6_Jul:** These values indicate the maximum temperatures on a specific day. It can be seen that temperatures range between 41.81 and 45.87, with an approximate average of about 43.77.
- **X23_Jul:** These values indicate maximum temperatures on another day. Temperatures range between 42.37 and 45.83, with an approximate average of around 44.39.
- **X16_Aug:** These values indicate maximum temperatures on another day. Temperatures range between 42.28 and 46.89, with an approximate average of around 44.38.
- **X17_Aug:** These values indicate maximum temperatures on another day. Temperatures range between 41.74 and 46.63, with an approximate average of around 44.34.
- **X18_Aug:** These values indicate maximum temperatures on another day. Temperatures range between 41.97 and 47.29, with an approximate average of around 44.15.

Statistical analysis: The approximate average of all studied values is about 44.2 degrees Celsius, which reflects relatively high and homogeneous maximum temperatures.

Analysis: Analyzing these points may help identify and explain unusual climate phenomena, which helps in understanding extreme climate changes and how they affect the local environment. These days can indicate the effects of climate change on local patterns, helping to understand how the climate may change in the future and how to adapt to these changes.

Planning: This information can be used to plan to better manage natural and agricultural resources by understanding how unusual climatic conditions affect resources.

Conclusion

1) By using variable clustering, the number of climate variables can be reduced to a smaller number of latent components that explain most of the variance in the data. This makes the analysis simpler and less complicated.

2) Determining the methodology for directional or local variables in the method of clustering variables around the latent components remains unclear until the direction of the variables is known, which is the presentation of the loadings from principal components analysis.

3) Measures for evaluating the stability of the CLV results according to the bootstrap method for the rand index, the adjusted rand index, the connectivity and isolation of the clusters showed high accuracy and quality and indicated the success of the clustering process.

4) The number of clusters according to the rho criterion (adjustment parameter) is determined by clustering the variables using the kmeans method. The “kplusone” strategy to identify noise clusters and the “sparselv” strategy to identify sparse variables (neutralized in this study) clearly show anomalous climate fluctuations.

5) The years 2013 and 2016 show prominent negative latent components, indicating lower than average extreme temperatures in those years.

6) 2020 and 2021 show prominent positive latent components, especially in the fifth component, indicating years with higher than usual extreme temperatures, which may reflect strong heatwaves.

7) The first underlying component reflects a clear pattern where certain years (such as 2014, 2017, and 2020) have higher maximum temperatures than other years, indicating the likelihood of climatic impacts or specific environmental conditions during those periods.

8) The second latent component shows variation in the years 2015 and 2019 with high positive values, which may indicate unexpected climatic effects or shifts in weather conditions during those years.

9) Some years, like 2018 and 2022, show fewer volatile components compared to other years, which may indicate more stable climatic conditions during those periods.

10) There is a general trend towards increased temperature fluctuations in recent years (2020 and 2021) compared to the beginning of the studied period (2013-2016), which may reflect a gradual climate change leading to an overall rise in maximum temperatures.

Recommendations

1) We recommend enhancing climate monitoring systems in Mosul and developing accurate forecasting models for changes in maximum temperatures, given the significant fluctuations observed. This can help predict extreme heat waves and take necessary measures to mitigate their effects.
2) Strategies for adaptation must be developed specifically to cope with extreme heat waves, such as those observed in 2020 and 2021. These strategies include providing public cooling areas, raising awareness among residents about the importance of staying hydrated, and enhancing assistance for the most vulnerable groups, such as the elderly and children.
3) With the trend towards rising temperatures, urban planners must consider designing cities in a way that reduces the impact of heat, such as increasing green spaces, using building materials that minimize heat retention, and developing infrastructure to provide shade and natural ventilation.
4) The local community should be made aware of the impacts of climate change and ways to adapt to it. Awareness campaigns can include how to act during extreme heat waves, the importance of proper cooling, and the wise use of water resources.
5) Given the likelihood of increasing climate fluctuations in the future, it is recommended to develop comprehensive emergency plans to address emergencies caused by extreme heat waves, such as power outages or water shortages.
6) High correlation values within the clusters indicate the stability of the climatic pattern within each cluster. This means that the days within each cluster share similar climatic characteristics.

Conflict of interest

The author has no conflict of interest.

References

1. Al-Hafith, O., Satish, B. K., Bradbury, S., & de Wilde, P. (2017). The impact of courtyard compact urban fabric on its shading: Case study of Mosul city, Iraq. *Energy Procedia*, *122*, 889–894.
2. Ali, S. H., Qubaa, A. R., & Al-Khayat, A. B. M. (2024). Climate Change and its Potential Impacts on Iraqi Environment: Overview. *IOP Conference Series: Earth and Environmental Science*, *1300*(1), 12010. https://doi.org/10.1088/1755-1315/1300/1/012010
3. Ghizlane, E.-Z., Wafae, S., & Abdelkrim, B. (2021). Features Clustering Around Latent Variables for High Dimensional Data. *E3S Web of Conferences*, *297*, 1070.
4. Hassan Ali, T., & Rashad Shaheen, B. (2013). Urban micro-climate in the City of Mosul, Iraq (The Effect of Urban Space Characters on Air Temperature) _ENG. *Al-Rafidain Engineering Journal (AREJ)*, *21*(2), 90–97.
5. Straus, D. M. (2019). Clustering Techniques in Climate Analysis. In *Oxford Research Encyclopedia of Climate Science*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228620.013.711
6. Vigneau, E. (2016). Dimensionality reduction by clustering of variables while setting aside atypical variables. *Electronic Journal of Applied Statistical Analysis*, *9*(1), 134–153.
7. Vigneau, E. (2020). Clustering of variables for enhanced interpretability of predictive models. *ArXiv Preprint ArXiv:2008.07924.*
8. Vigneau, E., Chen, M., Cariou, V., & Vigneau, M. E. (2022). *Package 'ClustVarLV.'*
9. Vigneau, E., Chen, M., & Qannari, E. M. (2015). ClustVarLV: an R package for the clustering of variables around latent variables. *R Journal*, *7*(2).
10. Vigneau, E., & Qannari, E. M. (2003). Clustering of variables around latent components. *Communications in Statistics-Simulation and Computation*, *32*(4), 1131–1150.
11. Vigneau, E., Qannari, E. M., Navez, B., & Cottet, V. (2016). Segmentation of consumers in preference studies while setting aside atypical or irrelevant consumers. *Food Quality and Preference*, *47*, 54–63.

تحديد المناخ المتطرف لطقس الموصل باستخدام استراتيجية حصينة لعنقدة الضوضاء

مروان ميسر محمود[1] ، بشار عبد العزيز الطالب[2]

قسم الإحصاء والمعلوماتية ، كلية علوم الحاسبات والرياضيات ، جامعة الموصل ، موصل ، العراق.

الخلاصة: يهدف هذا البحث إلى تحليل البيانات المناخية لمدينة الموصل خلال موسم الصيف من عام 2013 إلى 2022، مع التركيز على متغير درجات الحرارة القصوى. تم استخدام الأساليب الحديثة للكشف عن التقلبات المناخية التي لم يتم استخدامها سابقاً وتكييفها مع بيانات الدراسة واستكشاف المناخ العام والمتطرف الذي لم تتطرق إليه أي من الدراسات السابقة. تم استخدام تقنيات عنقدة المتغيرات لاكتشاف المكونات الكامنة وفق نموذج المتغيرات الموضعية. تم استخدام استراتيجية عنقود الضوضاء "K+1" لتحديد متغيرات عالية الضوضاء. واقترح الباحث الشكل العرضي لترتيب البيانات: P > N، مما يعني أن عدد المتغيرات أكبر من عدد الملاحظات. وتمثل الملاحظات سنوات الدراسة المرصودة، وكانت المتغيرات هي أيام الصيف لمدة ثلاثة أشهر (يونيو، يوليو، أغسطس). أثبت هذا الترتيب أنه مناسب لتقنية عنقدة المتغيرات للبيانات عالية الأبعاد. وأظهرت النتائج ست عناقيد، خمس منها كانت متجانسة تقريبا. تشير العناقيد الخمسة إلى أنماط مختلفة من ارتفاع درجات الحرارة القصوى خلال الصيف. العنقود الأول يبرز موجات الحرارة في منتصف الصيف (يوليو وأغسطس)، بينما يركز الثاني على نهايات الصيف الحارة (أواخر يونيو وأغسطس). العنقود الثالث يشير إلى موجات حر مبكرة ومتواصلة في يونيو ويوليو، والرابع يعكس حرارة مستمرة في أواخر يوليو وأغسطس. أما العنقود الخامس فيظهر تباينًا في درجات الحرارة بين بداية ونهاية الصيف. المتغيرات الضوضاء المستبعدة تمثل بيانات غير متسقة أو حالات استثنائية لم تنضم لأي عنقود. وهذا يساهم في تحسين دقة النماذج المناخية. وتسلط النتائج الضوء على الأنماط المناخية المميزة وتقدم توصيات لتعزيز التخطيط البيئي والزراعي.

الكلمات المفتاحية: عنقدة المتغيرات ،المكونات الكامنة ،عنقود الضوضاء ،البيانات عالية الابعاد ، المجموعات الموضعية.