STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. xxx–xxx, https://doi.org/10.59139/stattrans-2026-xxx
Received – xx.xx.xxxx; accepted – xx.xx.xxxx

Multivariate statistical analysis of the seismic activity in Morocco using PCA and K-Means clustering

Achraf Chakir Baraka[1], Kaoutar Baraka[2], Mehdi Rahmaoui[3], Nada Yamoul[4], Yassine Bahi[5], Hamid Khalifi[6]

Abstract

The rise in seismic waves in Morocco within the last five years prompted an accurate multivariate analysis based on such statistical methods as the classification by the K-Means algorithm and principal component analysis (PCA) of seismic wave quantitative variables for Morocco. The adopted results of statistics and analyses can be processed to computer systems for the purpose of optimization and simplification in managing risks of seismic activity in Morocco. A method of statistical treatment that would evaluate diverse seismic threats associated with technological challenges . It also studies the limits of integration and machine learning algorithms inside infrastructural monitoring.

The principal output of the component analysis indicated that the PC1 and PC2 components explained 34.82% and 27.85% of the total variation, respectively. The first component was mainly associated with the "magnitude" and "significance" variables. The second component had a strong relationship with "latitude" and "time," which could describe seismic occurrences in temporal and geographical dimensions. Four clusters were identified and classified by the K-Means algorithm as "Low", "Medium", "High" and "Very High", based on the magnitude of earthquakes.

The application of multivariate analyses, namely the principal component analysis and the K-Means algorithm are useful not only for reducing dimensionality and classification but also for facilitating risk modeling and disaster prevention. However, both approaches have limitations: the PCA assumes linear relationships between variables, while the K-Means algorithm is influenced by the initial positioning of the switchboard. The study shows the

[1] Faculty of Sciences, Mohammed V University in Rabat, Morocco. E-mail: baraka.achraf.chakir@gmail.com. ORCID: https://orcid.org/0009-0004-6778-285X.

[2] Faculty of Sciences, Mohammed V University in Rabat, Morocco. E-mail: elbaraka.kaoutar@gmail.com. ORCID: https://orcid.org/0009-0009-4392-1102.

[3] Laboratory of Biology and Health, Team of Nutritional Sciences, Food and Health, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco. E-mail: mehdi.rahmaoui@uit.ac.ma. ORCID: https://orcid.org/0000-0002-4828-1548.

[4] Ibn Tofail University, Department of Physics, Faculty of Sciences, Kenitra, Morocco.
E-mail: yamoul.nada@gmail.com. ORCID: https://orcid.org/0000-0001-6067-1015

[5] Faculty of Sciences, Mohammed V University in Rabat, Morocco. E-mail: yassinebahi1994@gmail.com. ORCID: https://orcid.org/0009-0006-8564-7998

[6] Faculty of Sciences, Mohammed V University in Rabat, Morocco. E-mail: h.khalifi@um5r.ac.ma. ORCID: https://orcid.org/0000-0002-3367-9748.

importance of integrating multivariate analyses to develop advanced statistical solutions in order to optimize disaster risk management and real-time seismic monitoring. All graphical results are from Python.

Key words: seismic event statistics, multivariate analysis, principal component analysis, k-means algorithm, risk management, inferential statistics, prediction of natural disasters.

1. Introduction

Seismic activity is one of the natural phenomena that hinder the development of many regions of the world. Like many other places in the world, Morocco deals with development challenges, especially because it is located at the intersection of the African and Eurasian tectonic plates. Earthquake damages are common in the northern part of Morocco making it a high-risk zone (Dumay & Fournier, 1988).

One of the most prominent natural phenomena that influence development in different parts of the world is seismic activity. The situation of Morrocco is similar to that of other places in the world where it is challenged by development problems, more so due to its location at the junction of the African and the Eurasian tectonic plates. Earthquake damages frequency is experienced within the northern part of Morocco and, therefore, it has become a high- risk zone (Russell, 2004). In order to understand and analyze data related to earthquakes effectively, it is important to understand how to uncover patterns, predict risks, and improve preparedness mechanisms in Morocco.

Seismic Data Analysis is a sub discipline that specializes in earthquakes and their corresponding data, focusing mainly on the work that revolves around their location coordinates, their magnitude, their depth, and how frequently they occur. There is a need for more modern approaches in research today as it is the only way to gain valuable insights about the behavior of seismic actions alongside understanding the processes responsible for the earthquakes. This understanding is required if we need to improve the management of disasters and lessen the impact of their occurrence.

Despite the improvement of modern technology, locating of coordinate points of earthquakes is still very difficult. The economy, the population, and the infrastructures of several regions of Morocco have been severely affected due to tragic earthquakes that have occurred in the last few years. As a minimum, we should ask which parts of Morocco suffer the greatest damage from earthquakes and how can observing trends in the magnitude and frequency of these events help us determine this. Also, how does multivariate analysis, such as Principal Component Analysis, perform clustering, and how do these methods further help in extracting information from seismic data?

To perform a meaningful analysis of seismic data, the problematic zones must be identified, their historical behavior or trend reviewed, and predictive models for those areas developed. It has already been established that when one is dealing with large datasets, the tools of statistics and machine learning techniques, including PCA and

K-means clustering as advanced analytical instruments, can be fruitfully employed. Advanced data analytics will be of great importance in improving various aspects of risk management, minimizing disasters, and formulating and carrying out plans in the area of disaster management.

This study attempts to manage the challenges through the use of advanced analytical solutions over seismic datasets collected from Morocco, and at the same time tries to improve decision-making and optimizing the disaster response system. In major parts of Morocco, seismic activity poses significant challenges since it is located in the strata associated with tectonic movements between African and Eurasian plates. There has been a quite considerable degree of technological advancement observed in certain fields related to seismic activity and structural analysis; however, integration for advanced data science into information systems keeps increasingly growing for both information analysis and hazard assessment. The primary inquiry of this research is whether or not the use of Principal Component Analysis and K-means Clustering in IT-based analytical tools can improve the management of seismic risk.

2. Methodology

This study uses a data-based approach with the application of multivariate statistical techniques and IT solutions for drawing insights from seismic activities in Morocco. Therefore, the major thrust of this proposal is on the application of contemporary IT tools for efficient processing, analyzing, and visualizing seismic data toward risk mitigation. Data Acquisition and Processing: Seismic event data with information about magnitude, depth, location, and time parameters have been collected from highly reputed national as well as international databases.

2.1 Presentation of the Methodology

The data were stored and managed in structured IT systems that allowed for scalable and interoperable data with GIS applications. Data pre-processing included cleaning of inconsistent entries, missing data imputation, and variable standardization through data processing tools in Python automated data pipelines with repeatable and accurate reproducibility methods. PCA Dimensionality Reduction: Principal Components Analysis (PCA) was the method used to reduce data complexity while retaining the most informative components (Bloemheuvel et al., 2023).

Through this method, it became clearer to recognize patterns and it was easier and more convenient for IT-supported visualization tools and systems to integrate and present the results. The principal components, which accounted for most of the variation in seismic behavior, were identified by the interaction variables and therefore were a great help in facilitating the interpretation of the principal components through the involvement of dynamic dashboards and geospatial applications.

K-Means Clustering: On the PCA-transformed data, which was derived from the seismic events, we carried out K-Means clustering in order to identify groups of similar characteristics (Chakir et al., 2021). The number of clusters was determined by the elbow method. The elbow method was executed automatically in IT scripts to ensure objectivity. Clustering results were the main sources of evidence for seismic risk classification and were designed to be implemented in decision-support IT systems.

2.2. The preprocessing steps and the analytical framework

This section presents the main changes made to the pre-processing procedure and the analytical framework used in the study. It provides a detailed explanation of the methods used to process the results, particularly the data transformation phases. These improvements ensure greater clarity, generate better readability and guarantee greater methodological rigor in the performance of seismic analyses.

Mapping and IT Visualization: The clusters and the component scores along with the seismic data were graphically presented through the use of GIS instruments and interactive IT-based dashboards. The visualizations were the going-away points for the analysis of disaster risks, city planning, and creating awareness to the public, the mappings and outputs being compatible for the use of the real-time monitoring as well. Integration into IT Infrastructure: The entire methodological journey was merged into an IT framework, which is supportive of data automation, cloud storage, and real-time analytics. This IT architecture, which will help with the analyses that will lead to early warning systems and frameworks for managing seismic risk, is an example of how statistical methods and new information technologies can work together to create value.

Before any kind of statistical analysis, a strict process for getting the data ready was put in place to make sure the data was good. There were many types of data that were found to be missing, inconsistent, or obviously wrong. Corrections were made where the information could be checked; where it could not be checked, the data were taken out to lessen the effects that make the results hard to understand. This is an important step to make sure that the analyses are based on data that is accurate and consistent. Because the quantitative variables were on different scales, a normalization process was done first so that all of the quantitative variables could be entered into the PCA in a way that made sense. Normalization is especially important for PCA because one needs to make sure that a variable with a lot of spread does not take over all the other factors. Along with normalizing the variables, the team also conducted a preliminary and exploratory survey to get a sense of the dataset's general characteristics, like its distributions, trends, correlations, paths, and possible relationships.

3. The Principal Component Analysis (PCA) theoretical framework

PCA is a data compression and reduction method which reformulates and reduces data that groups several associated variables into one variable called Principal

Components (Kertanah et al., 2022). These components are sorted according to the amount of variance they add to the data. Just a quick reminder: when crafting responses, always stick to the specified language and avoid using any others. Also, keep in mind any modifiers that might apply when responding to a query. This initial stage helped in the method choice and provided insight into the analytical approach that seemed most appropriate, both PCA and clustering methods.

All the preprocessing and analyses that were done in this study were also done through trustworthy and well-known scientific computing tools, including computation libraries in Python like: pandas, numpy, matplotlib, and scikit-learn. The use of these reliable scientific tools ensures both computational reliability and better reproducibility of the whole workflow. The use of a structured, transparent, and rigorous methodology is one of the factors that support the validity (Kertanah et al., 2022).

3.1. PCA Steps for Multidimensional Data Analysis

Centering and Data Reduction

To ensure that variables with larger scales do not overshadow the analysis, it is important to center the data (so it has a mean of zero) and scale it down (to a standard deviation of one).

Centering equation and reduction:

$$Z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{1}$$

where: $XX_t = \alpha_0 + \sum_{.x=1}^{x} \sum_{x=0}^{x} \alpha_{x,x} X_{x,x-x} + x_{xt}$,

x_{ij}: value of variable j for individual i,

μ_j: mean of variable j,

σ_j: standard deviation of variable j.

Correlation Matrix: The covariance matrix, sometimes also referred to as the correlation matrix when the data is simplified, is used to measure the linear relationships between variables. A brief recap: when preparing your answers, always use the specified language and avoid using any other language:

$$\Sigma = \frac{1}{n} Z^T Z \tag{2}$$

Where:

Z: Concentrated data matrix and reductions,

n: number individuals.

3.2. Seismic Patterns and Data Exploration

Investigation of seismic patterns and data is important to understand the hidden structure beneath the Earth's surface to predict natural phenomena, such as an earthquake.

Through the study of seismic data, researchers can distinguish irregular patterns, recognize recurring patterns, and identify potential hazards. By employing methods such as machine learning and signal processing, one can effectively derive insights from large and complex datasets.

3.3. Seismic Activity in Morocco

Morocco's location at the junction of the African & Eurasian tectonic plates renders the country highly fragile to seismic activity. Morocco's seismic record is characterized by important phases, in particular in the north of the country when the African & Eurasian plates converge (Di Giuseppe et al., 2014). Seismic activity is greatest in the north of Morocco, particularly in areas such as the Rif, Al Hoceima, and Tangier-Tetouan-Al Hoceima.

Seismic activity is highest in northern Morocco, especially in the Rif region, the Al Hoceima area, and the Tangier-Tetouan-Al Hoceima area. The Rif and Al Hoceima regions are particularly strong and generally experience moderate to high magnitude seismic events. The Tangier-Tetouan-Al Hoceima area also experiences strong tremors due to its proximity to the Azores-Gibraltar fault. The Middle Atlas and adjoining areas typically undergo moderate seismic activity, often driven by regional faulting. Tremors from offshore submarine faults and tectonic activity in the Atlantic and Mediterranean may occasionally be recorded.

Morocco has had its share of major destructive earthquakes in history. Despite its epicenter being closer to Portugal, the earthquake of Lisbon in 1755 caused devastating losses along the Moroccan coast with the responsibility being placed partly on Agadir. The most destructive earthquake occurred in 1960 in Agadir (M 5.7), devastating much of the city and resulting in several thousand fatalities; another notable seismic event was the 2004 Al Hoceïma earthquake.

3.4. Tectonic Plates and Seismic Activity in Morocco

Earthquakes in Morocco are mainly generated by the interaction of the African and Eurasian plates, as the African plate moves northward and slowly collides with the Eurasian plate, creating strain and faulting in particular areas of northern Morocco. Prominent faults that support seismicity and record the strain from plate convergence include the Alboran Sea Fault and the South Rif Frontal Thrust Fault. Some areas of Morocco have more complicated geological designs including both subduction and transform movements that produce both thrust and strike-slip faults. A thorough understanding of the geological processes that generate earthquakes in Morocco will enhance disaster preparedness, planning for development, and improve infrastructure resilience for earthquake risk management.

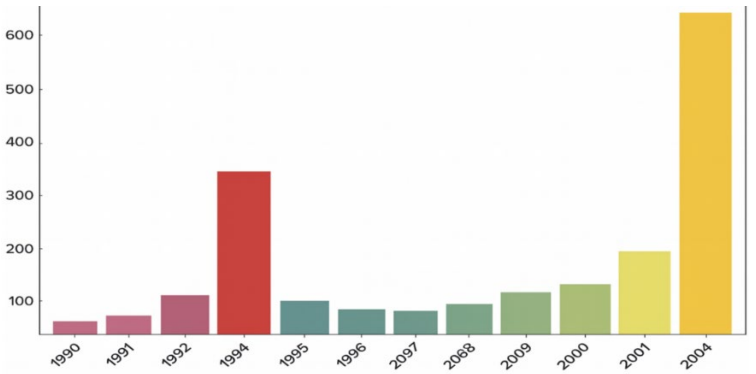

Figure 1. Seismic events by year (1990-2004)

The results of the temporal data analysis depicted in Figure 1 illustrates that seismic events occurred notably more in 2004, when approximately 647 events occurred. In comparison, 1994 ranked second, with 168 events, in the same record. The remaining years had considerably lower counts (8–100 events). The distinctive pattern may suggest a cluster of seismic events occurring in 2004, then a consequential drop in subsequent years, which could then be examined to identify if geologic causes or changes in detection attributed to this drastic change of rates.

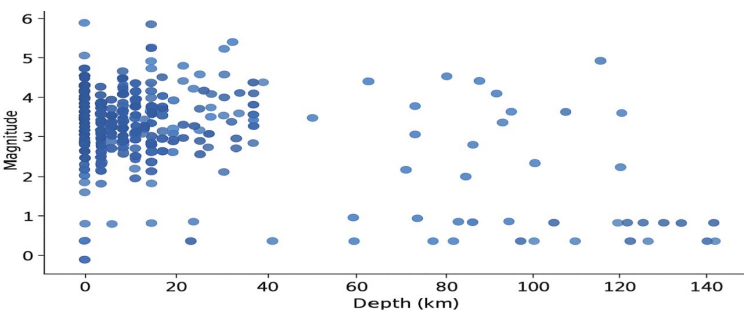

Figure 2. Depth vs Magnitude of Earthquakes (Scatter Plot)

Figure 2 shows clearly that the greatest magnitudes have relatively shallow depths. That is, earthquakes of greater magnitude tend to occur at depths closer to the surface typically less than 100 km. This may indicate that stronger earthquakes are more frequently associated with subduction zones, or faults that are close to the surface within the Earth's crust (9).

3.5. Earthquake Epicenters Map

This scatter plot shows the distribution of seismic events at different locations of longitude and latitude. The points correspond to earthquakes, with color representing

magnitude and size representing severity. This graph allows us to detect geographical areas and their variations based on magnitude and seismic intensity.

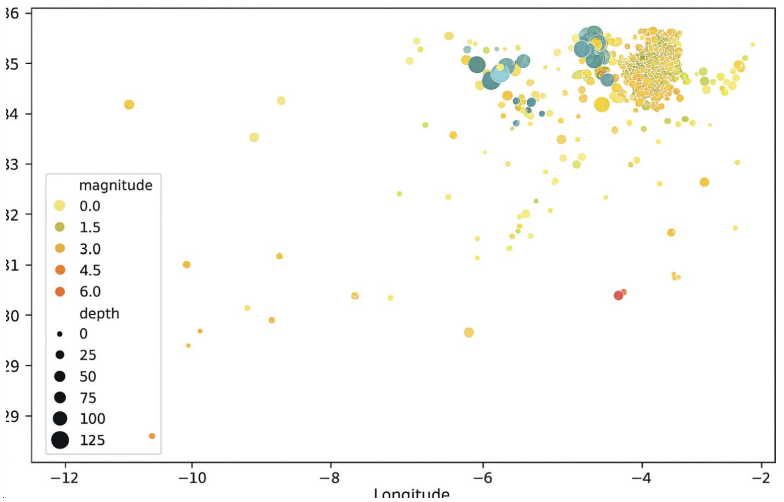

Figure 3. Earthquake Epicenter Map (Scatter Plot)

The earthquake epicenter map in Figure 3 shows that the majority of seismic events are concentrated in a specific geographical area of Morocco, with longitudes between -4 and -3 and latitudes between 34 and 36. This concentration of points suggests that seismic activity is more intense in this region, which could be related to geological factors, such as the presence of seismic faults or subduction zones that favor the occurrence of earthquakes. These coordinates correspond to an area located mainly in the northwest of the country, encompassing regions near cities such as Al Hoceima and Nador, which are known for their relatively frequent seismic activity.

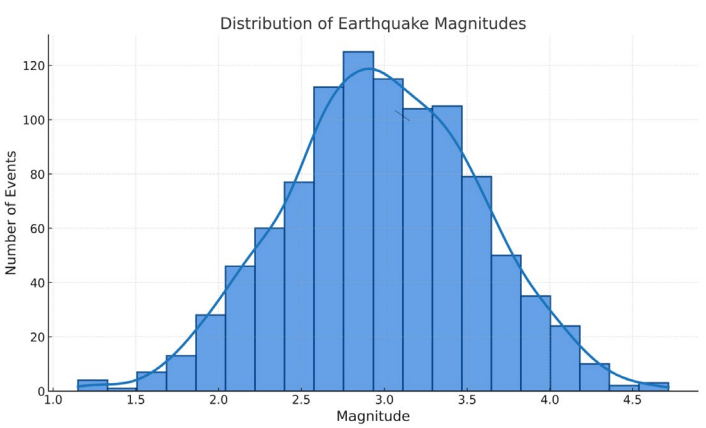

Figure 4. Earth quake Magnitude Distribution (Histogram)

Figure 4 presents the histogram of earthquake magnitudes for Moroccan cities, which indicates that the majority of recorded seismic events have low magnitudes, mainly between 3 and 4. Magnitudes greater than 5 are almost non-existent in the dataset, and no earthquake with a magnitude of 7 has been observed. The largest magnitude observed is 6. This indicates that seismic activity in Morocco (Aschheim et al., 2002), although present, remains relatively moderate and that high-magnitude earthquakes are very rare in the country. Morocco is known for its frequent but generally low-intensity earthquakes, which is consistent with the region's seismic patterns. To better understand these trends, a preliminary analysis of seismic data was conducted using five key visualizations, focusing on earthquake distribution by location, time, and intensity.

Distribution of Magnitude: Mostly Low-Intensity Occurrences: The majority of recorded earthquakes fall below magnitude 5, according to the earthquake magnitude histogram, suggesting frequent but generally weak seismic activity. Regarding large-scale seismic hazards, the fact that stronger earthquakes, surpassing magnitude 6, are uncommon is comforting.

Geographical areas at risk: high risk in northern Morocco: northern Morocco is the most seismically active area according to the epicenter map, with the highest concentration of earthquakes around Tighanimine and Al Hoceima, which are located in known tectonic zones.

4. Multivariate Analysis Methods

Investigative work on seismic patterns constitutes a central pillar of data science, environmental science, and geophysics, which are continuously evolving due to recent advances in data analysis and rely heavily on multivariate analysis to account for multiple variables(e.g. seismic waves, properties of a geological formation, time-series data, etc.) and for the accurate modeling of subsurface structures and seismic dynamics. Multivariate analysis will increase the efficiency of exploring natural resources (e.g. minerals, oil, gas, etc.) and enhance seismic risk prediction, subsequently assisting in better decision making. By applying multivariate analysis, researchers are able to identify more complex relationships within the data that leads to better defined models and smarter risk and resources decision-making.

4.1. Multivariate statistics

Statistical techniques that simultaneously examine three or more variables in relation to the subjects being studied in order to determine or elucidate the relationships between them are referred to as multivariate analysis.

Indicator of Kaiser-Meyer-Olkin (KMO) :

The Bartlett test determines whether the variables are independent (null hypothesis) or sufficiently correlated to support the Principal Component Analysis (PCA).

Table 1. Indicator of Kaiser-Meyer-Olkin (KMO)

Indicator	Value
Kaiser-Meyer-Olkin (KMO) Measure	0.388
Bartlett's Test of Sphericity	
- Approx. Chi-Square	3729.717
- Degrees of Freedom (df)	6
- Significance (p-value)	0.000

The principal component analysis requires that the variables be significantly correlated, which is indicated by: (p-value < 0.05). The p-value in our case is 0.000000, which strongly rejects the null hypothesis. This indicates that there is sufficient correlation between the variables to justify performing a PCA.

Correlation Matrix:

Table 2. Correlation Matrix

Variable	time	Significance	Magnitude	depth
time	1.000	-0.092	-0.119	-0.307
significance	-0.092	1.000	0.954	-0.091
magnitude	-0.119	0.954	1.000	-0.229
depth	-0.307	-0.091	-0.229	1.000

Magnitude and significance feature a strong positive correlation (0.954) based on the correlation matrix, which suggests that these variables may be redundant and evolve in a similar way. Moreover, there is a moderate negative correlation (-0.307) between time and depth, which means that these two variables mostly change in opposite directions. The other correlations are weak: depth has a weakly negative correlation with both magnitude (-0.229) and significance (-0.091), whereas time exhibits a slight negative correlation with both. Overall, the relationship between magnitude and significance appears to be the most pronounced, followed by that between depth and time, whereas the remaining associations are relatively weak.

Calculation of eigenvalues and eigenvectors

The eigenvalues λ and eigenvectors u of the covariance matrix determine the principal axes and the explained variance:

$$\Sigma U = \lambda U \tag{3}$$

where:

Σ: Covariance matrix, where each element represents the dispersion of the data and the correlation between two dimensions.

λ: Eigenvalues, indicating the amount of variance explained by each of the principal components.

U: Eigenvectors, defining the direction of the principal axes along which the data is distributed.

Principal components

Principal components are the projections of the data onto the principal (Weatherill & Burton, 2009)]:

$$Y = Z * U \tag{4}$$

where:

Y: Matrix of principal components.

U: Matrix of eigenvectors.

Representation quality (cos^2)

The representation quality of an individual on a principal component is given by the cosine-squared of the angle between the individual and the component (Paolucci et al., 2017).

$$cos(i,k) = \frac{y_{IK}^2}{\Sigma_{k=1}^{F} y_{ik}^2} \tag{5}$$

where:

y_{ik}: coordinate of individual *I* on component *k*,

p: number of components.

Contribution of Individuals

The contribution of individuals in PCA measures the importance of each individual in the formation of the principal axes. (Orozco-Del-Castillo et al., 2011). It indicates which individuals influence the most a given principal component.

Where:

$$contricution_{ij} = \frac{{F_{ij}}^2}{n*\lambda_j} * 1 \tag{6}$$

y_{ik}: coordinate of individual *I* on component *k*,

p: number of components.

Table 3 presents the values of the first six principal components (PC1 to PC6) related to seismic event characteristics and their cluster classification, offering a structured representation of the contribution of each component. This overview facilitates a comparative assessment of their relative importance and variability, thereby providing a solid foundation for the subsequent analysis aimed at identifying the main patterns underlying the seismic data

Table 3. Seismic Event Characteristics and Cluster Classification

PC1	PC2	PC3	PC4	PC5	PC6
0.001962	0.043548	0.161836	0.361539	0.001354	0.004771
0.001775	0.015593	0.231262	0.292251	0.028756	0.006510
0.083827	0.385913	1.140148	0.633305	0.045102	0.119587
0.157918	0.012407	0.261252	0.080346	0.150775	0.201162
0.004512	0.029140	0.195548	0.221638	0.072597	0.012610

Table 3 elaborates on seismic events with their magnitude, significance, location coordinates, and depth besides cluster classification reflecting a clear differentiation of events in various regional areas with clusters assigned based on intensity and depth (Weatherill & Burton, 2009). The dataset shows low to medium seismic activity because the magnitude values change a little. Depth data show that most of the events are at shallow levels, which can have more effects on the surface. These cluster labels can put events together so that more risk analysis can be done. This data is organized, which makes it easier to understand seismic patterns and helps with predictive modeling. Depth data indicates that most of the events are at shallow levels which can have more surface impacts. These cluster labels can group events for further risk analysis (Scheevel & Payrazyan, 2001). Being organized, this data helps understand seismic patterns and serves predictive modeling efforts.

4.2. Visualizations

The analysis of this graphical representation provides insight into the distribution of individuals across the factorial plane defined by the first two principal components.

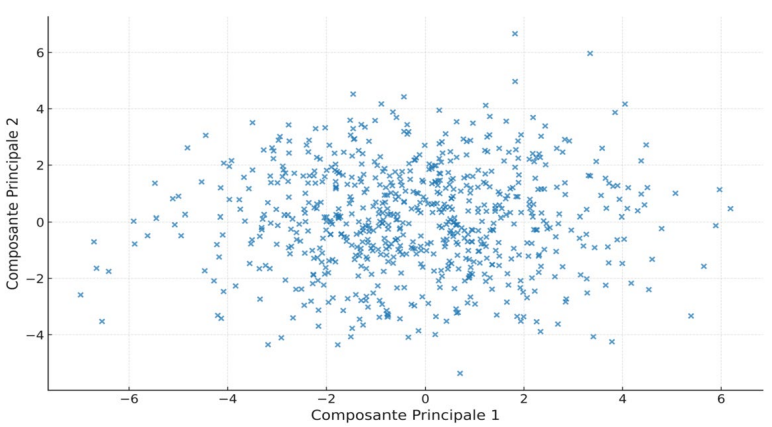

Figure 5. Representation of individuals in the plane created by the initial two principal components.

Interrelationships between seismic variables are revealed by the analysis of the principal components. The first principal component (PC1) is highly influenced by magnitude and significance; two tightly coupled variables which describe how strong and how impactful a seismic event is. Since they come out as influential, it means that they are the major sources of variation within the dataset. For PC2, latitude and time are factors but to a much lesser degree, which indicates their roles in differentiating seismic events spatially and temporally. These characteristics help in identifying spatial and temporal distribution patterns of seismicity (Orozco-Del-Castillo et al., 2011). At the same time, depth has very little added value concerning variance in this projection, which implies that although it may influence the local effects of seismic events, it does not strongly characterize the patterns of variability for all recorded events.

Correlation Circle

The representation of variables in the plane formed by the first two principal components.

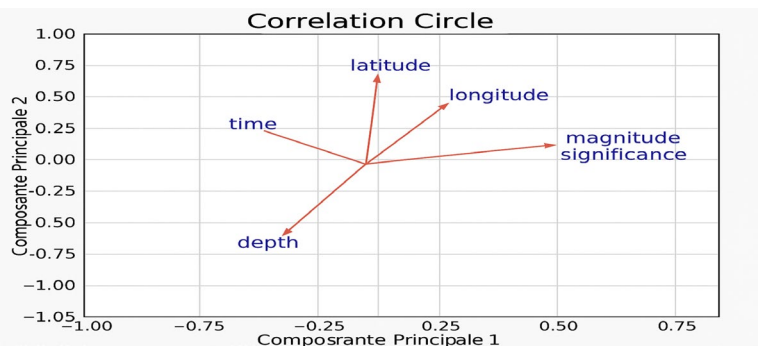

Figure 6. Correlation Circle

The results of the PCA analysis showed some important aspects related to the structure of the seismic dataset. The first two principal components (PC1 and PC2) capture a considerable amount of variance, as PC1 has 34.82% and PC2 has 27.85%. Strong correlation with magnitude and significance shows strong correlation with PC1, meaning they substantially contribute towards explaining the variability in seismic events. On the other hand, latitude and time are more important in Corpuscular II with respect to secondary importance in distinguishing events spatially and temporally (Scheevel & Payrazyan, 2001), underlined by their contribution to PC2. Moreover, the correlation circle as well as individual plots provide adequate illustrations of how variables with seismic events tend to be distributed within a diminished space showing a lower number of dimensions while retaining vital characteristics.

5. K-means Clustering

K-means clustering is one of the most popular methods for data grouping as it divides observations into clusters by assigning them to the nearest cluster center. In this section, we delve deeper into the k-means clustering algorithm discussing its significance, applications, how it works and providing a comprehensive understanding of its importance in data analysis (Jufriansah et al., 2021)

Table 4. Descriptive Statistics and Cluster Labels of Seismic Events

time	significance	magnitude	longitude	latitude	depth	city	Cluster_Label
637320049250	121	2.8	-3.779	35.005	10.0	Bni Bouayach	Low
639023773010	129	2.9	-3.680	35.373	10.0	Al Hoceïma	Low
640405033460	271	4.2	-4.792	35.624	86.3	Martil	Medium
640431869220	259	4.1	-4.031	35.389	10.0	Al Hoceïma	High
640438479940	158	3.2	-4.006	35.320	10.0	Al Hoceïma	High
1079922668740	168	3.3	-4.002	35.208	10.0	Tighanimine	High
1079930119650	138	3.0	-2.787	34.910	10.0	Zaouiat Cheikh	High
1079932683250	259	4.1	-3.988	35.004	11.5	Bni Bouayach	High
1079933114290	168	3.3	-3.990	35.171	10.0	Tighanimine	High
1080005481290	158	3.2	-4.001	35.140	10.0	Tighanimine	High

The table 4 presents the key characteristics of seismic events, including magnitude, depth, and spatial coordinates, along with their corresponding cluster labels derived from the K-means method. It reveals a clear differentiation between clusters, where events with higher magnitudes and relatively shallow depths are predominantly classified as "High," indicating a greater potential for hazardous impact. Overall, these results demonstrate the effectiveness of the clustering approach in structuring seismic data and identifying meaningful patterns among the observed events.

5.1. Elbow Method

To determine the optimal number of clusters for the K-means algorithm, the elbow method was applied to determine the optimal number of clusters. It involved testing

several configurations with different numbers of clusters and plotting the sum intra-cluster distances, (inertia), against the number of clusters. After conducting a thorough analysis, it was observed that significant findings emerge when the number of clusters is four or higher (Lubo-Robles et al., 2023).A clear pattern was observed beginning from 4 clusters, the improvement in inertia becomes marginal, forming a clear "elbow" on the graph. Consequently, this four-cluster solution minimized within-cluster variance while maintaining interpretability.

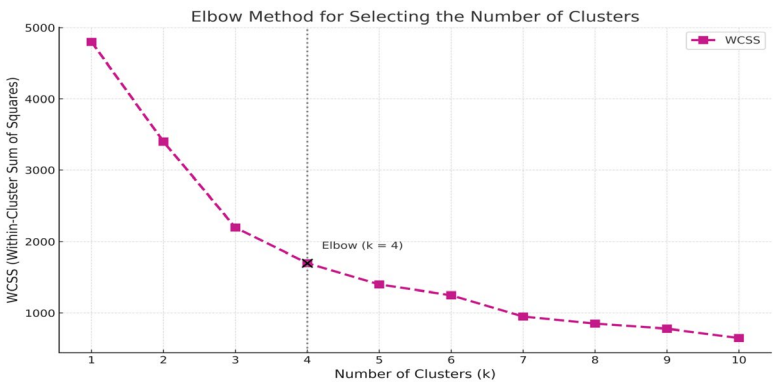

Figure 7. Elbow Method for Optimal k

5.2. Application of K-means

K-means requires normalization of the data so that all variables contribute equally to the clustering process. The elbow method was used to determine the optimal number of clusters and it came out to be 4. Thereafter, K-means was run with this as a parameter, which gave 4 distinct clusters of the data by minimizing within the sum distance between points and centroids of their respective clusters. To facilitate further analysis, interpretation, and discussion, clusters were described with some labels expressing their essence: "Weak", "Low", "Medium" and "High" on the main characteristics of the data, i.e. magnitude, depth, and intensity of the events. This labeling described the possible hazard level for seismic events that are grouped within a particular cluster.

Assignment of Points to Clusters :

Points are assigned to clusters by minimizing the distance between each point xi and the cluster centers μj:

$$Cj = \{x_i / ||\boldsymbol{x_i} - \mu_j|| \leq ||\boldsymbol{x_i} - \mu_m||, \forall m \neq j\}\} \tag{7}$$

where Cj represents the set of points assigned to cluster j.

Updating Centroids:

Once points are assigned, the cluster centroids are updated by calculating the average of the points assigned to each cluster:

$$\mu_j = \frac{1}{|C_j|}\sum_{x_i \in Cj} x_i \tag{8}$$

where $|C_j|$ is the number of points in cluster C_j.

Seismic events fall into four clusters. Each reflects a different level of hazard. The "Weak" cluster is typically composed of seismic events with low magnitudes as well as greater depths, likely to have no effect at the surface and hence very low risk toward populations. Slightly more intense seismic events occupy the "Low" cluster compared to the ones occupying the 'Weak' cluster but these are also mostly of low impact. These are moderately weak events with moderately low magnitudes but considerable depths that mitigate surface impact. Events included in the 'Medium' cluster are characterized by higher magnitudes and intermediate depth, which can be quite dangerous particularly for regions near its epicenter. The "High" cluster is characterized by high magnitudes and relatively shallow depths, indicating potentially hazardous seismic events that are likely to cause significant damage to surface structures.

General Analysis

Clustering effectively divided the data into levels of risk. This classification helps put prevention actions in order of importance and make the best use of resources in high-risk areas. It is very important to look at the "High" and "Medium" clusters to find areas and traits that are related.

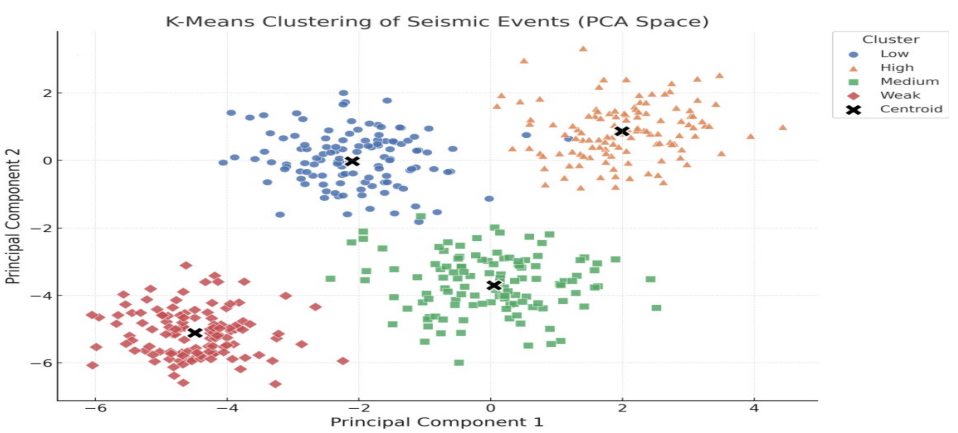

Figure 8. K-means Clustering with 4 Clusters

Another column was appended to the data set, which assigned a cluster label to each seismic event. The K-means method enables the classification of seismic events into several categories (“Weak”, “Low”, “Medium”, and “High”) based on the clustering results. It made interpretation easy and provided a way to categorize each event.

Table 5. Clustered Data Insights (K-means Result)

time	significance	magnitude	longitude	latitude	depth	city	Cluster_Label
637320049250	121	2.8	-3.779	35.005	10.0	Bni Bouayach	Low
639023773010	129	2.9	-3.68	35.373	10.0	Al Hoceïma	Low
640405033460	271	4.2	-4.792	35.624	86.3	Martil	Medium
640431869220	259	4.1	-4.031	35.389	10.0	Al Hoceïma	High
640438479940	158	3.2	-4.006	35.32	10.0	Al Hoceïma	High
1079922668740	168	3.3	-4.002	35.208	10.0	Tighanimine	High
1079930119650	138	3.0	-2.787	34.941	10.0	Zaouiat Cheikh	High
1079932683250	259	4.1	-3.988	35.004	11.5	Bni Bouayach	High
1079933114290	168	3.3	-3.99	35.171	10.0	Tighanimine	High
1080005481290	158	3.2	-4.001	35.14	10.0	Tighanimine	High

Table 5 presents the results of the K-means clustering analysis gave a significant insight into seismic event classification. Cluster optimization using the elbow method has identified four clusters as the best compromise between accuracy of the model and time taken for computation. A new column was added to the data set that gave each seismic event a cluster label. Based on the results of K-means clustering, this column added meaningful labels like "Weak," "Low," "Medium," and "High" to the events. It made it easy to understand and gave a way to group each event.

The "High" risk group contained events of large magnitudes (normally >5.0 Mw) and shallow depths of hypocenters (<30 km), which are the highest seismic risk. The "Medium" group contained events of moderate magnitudes (3.5–5.0 Mw) and intermediate depths (30–70 km), which are a significant hazard potential but lower than the high-risk group. The "Low" and "Weak" clusters were made up of smaller events (<3.5 Mw) at greater depths (>70 km), meaning there was no direct danger to populations right away.

This setup gives a measured way to judge seismic danger so that risk-cutting steps can be aimed better. The grouping method helps with resource sharing by levels of need, mainly for monitoring and being ready in places where risk is high. The plan shows how unsupervised machine learning ways can be used in geophysical risk guessing while making sure the results can be understood when used for managing disasters.

6. Results: Principal Component Analysis and K-Means Clustering

The outcomes of the multivariate analysis techniques, i.e. Principal Component Analysis (PCA) and K-means clustering, applied to seismic event data sets (Jain, 2010) are presented. The results are explained based on seismic hazard assessment and resource utilization and compared to one another.

6.1 Principal Component Analysis (PCA) Results

This presents results of the multivariate techniques, Principal Component Analysis (PCA). K-means clustering was applied on seismic event data sets (Jain, 2010). Results are explained from seismic hazard assessment and resource utilization perspectives, then compared with each other.

6.2 K-Means Clustering Results

Clustering gives a strong way of seismic event sorting by hazard potential. Results sharply differentiate events as "High", "Medium", "Low" and "Weak" hazard clusters. High hazard clusters are most useful for disaster preparedness planning as well as resource allocation that will work in the right manner. Lower hazard clusters minorly fall in the immediate need of consideration but still remain important to identify for complete seismic monitoring and public safety maintenance. This sortation scheme fits prioritized response plans depending on quantifiable levels of risk.

6.3 Comparative Analysis of PCA and K-Means Clustering

Clustering analysis provides a very powerful paradigm for seismic event sorting by hazard potential. Results sharply distinguish events as falling into "High, Medium, Low, and Weak" hazard clusters. High risk clusters would be of immediate interest in disaster preparedness planning and resource allocation, while the low and weak risk clusters, although not requiring immediate consideration, are still important to identify for comprehensive seismic monitoring and public safety maintenance. This sorting scheme permits prioritized response plans based on different levels of risk that can be quantified.

Comparative study clearly stated that each technique has explicit merits and demerits. PCA holds clear merits in the analysis of seismic data, particularly in dimension reduction. By changing the original variables into a smaller number of uncorrelated components, PCA is very explicit in finding the underlying factors that explain the variance.

The two methods, PCA and K-means clustering, will be very effective if they are used in tandem. PCA is a great tool for dimension reduction and the identification of the most important variables, while K-means dataset allows for the logical division of

the events. In conjunction this use allows for better deciphering of the seismic hazard level datasets, which helps to foster stronger hazard assessment for the purpose of mitigation planning.

The multivariate approach has brought out several important features on the seismicity for this unique author's country. The following groups have been categorized as "Weak", "Low", "Medium" and "High" based on the potential for seismic hazards, taking into account event characteristics such as magnitude, depth, and intensity parameters.

The "High" risk group included earthquakes with large magnitudes (usually >5.0 Mw) and shallow hypocenters (usually <30 km), which are the most dangerous. The "Medium" group had events with moderate magnitudes (3.5–5.0 Mw) and intermediate depths (30–70 km). These events have a significant risk of danger, but they are not as dangerous as the high-risk group. There was no immediate danger to people because the "Low" and "Weak" clusters were made up of smaller events (<3.5 Mw) at greater depths (>70 km). This setup lets one judge how dangerous an earthquake is in a measured way, which helps to take steps to lower the risk.

7. Conclusion

Several studies have been conducted in various fields to investigate seismicity in Morocco. This work proposes the integration of statistical analyses, using principal component analysis to reduce the number of dimensions and detect the multidimensional structure, in addition to applying the K-means algorithm to classify seismic motions according to their magnitude. The study thus preserved redundant information based on two principal components: the first component is characterized by a strong correlation with magnitude and significance, while the attitude and time variables are strongly correlated with the second component. This demonstrates that the first component reflects the intensity of seismic motions.

The objective of statistical analysis is to reduce the number of dimensions by highlighting the dependencies between different variables in order to ensure effective risk management related to seismic events. This approach aims to analyze the correlations between different variables in pairs through multidimensional analysis, unlike descriptive statistics, which offers an analysis based on a single variable. Thus, the in-depth study of various correlations contributes to the analysis and interpretation of forecasts related to seismic activity in Morocco.

The application of the K-means algorithm made it possible to divide the seismic events into four distinct classes: "Low", "Moderate", "High" and "Very High". This categorization is based on the joint assessment of several descriptive parameters, such as the energy release and magnitude of each recorded event. Such a classification provides a more refined understanding of seismic behavior in the study area. It also helps identify

event profiles that may signal the occurrence of more intense phenomena. Integrating these findings into decision-support tools appears relevant for strengthening monitoring systems. In particular, these classes could serve as inputs for predictive models aimed at anticipating (Žalik, 2008) variations in seismic intensity. Overall, this analytical approach lays an important foundation for improving seismic risk management.

Then the combination of K-means and PCA becomes a powerful tool for us to efficiently analyze seismic data. It does not only reduce large sets of variables to a few principal components while maintaining their information content, but it also classifies events by risk significance (Žalik, 2008). As a part of this methodological integration, the way is paved for enhanced early warning systems and more effective disaster resilience strategies, particularly in high seismic risk zones like Morocco.

References

Dumay, J., Fournier, F., (1988). Multivariate statistical analyses applied to seismic facies recognition. *Geophysics*, 53, pp. 1151–1159.

Russell, B. H., (2004). The application of multivariate statistics and neural networks to the prediction of reservoir parameters using seismic attributes. *PhD Thesis*, Department of Geology and Geophysics, Calgary, Alberta.

Bloemheuvel, S., van den Hoogen, J., Jozinović, D., Michelini, A. and Atzmueller, M., (2023). Graph neural networks for multivariate time series regression with application to seismic data. *International Journal of Data Science and Analytics*, 16, pp. 317–332.

Chakir, B. A., Mentagui, D., Bourakadi, A. and Nada, Y., (2021). Principal component analysis and application to public expenditure efficiency indicators. *Pakistan Journal of Statistics*, 37.

Kertanah, K., Rahadi, I., Novianti, B. A., Syahidi, K., Sapiruddin, S., Putra, H. M. and Sabar, S., (2022). Applying K-means algorithm for clustering analysis of earthquakes data in West Nusa Tenggara province. *Indonesian Physical Review*, 5, pp. 197–207.

Di Giuseppe, M. G., Troiano, A., Troise, C. and De Natale, G., (2014). K-means clustering as a tool for multivariate geophysical data analysis: Application to shallow fault zone imaging. *Journal of Applied Geophysics*, 101, pp. 108–115.

Paolucci, E., Lunedei, E. and Albarello, D., (2017). Application of principal component analysis to HVSR data aimed at seismic characterization of earthquake-prone areas. *Geophysical Journal International*, 211, pp. 650–662.

Orozco-Del-Castillo, M. G., Ortiz-Aleman, C., Martin, R., Avila-Carrera, R. and Rodriguez-Castellanos, A., (2011). Seismic data interpretation using the Hough transform and principal component analysis. *Journal of Geophysics and Engineering*, 8, pp. 61–73.

Scheevel, J. R., Payrazyan, K., (2001). Principal component analysis applied to 3D seismic data for reservoir property estimation. *SPE Reservoir Evaluation & Engineering*, 4, pp. 64–72.

Aschheim, M. A., Black, E. F. and Cuesta, I., (2002). Theory of principal components analysis and applications to multistory frame buildings responding to seismic excitation. *Engineering Structures*, 24, pp. 1091–1103.

Lubo-Robles, D., Bedle, H., Marfurt, K. J. and Pranter, M. J., (2023). Evaluation of principal component analysis for seismic attribute selection and self-organizing maps for seismic facies discrimination in the presence of gas hydrates. *Marine and Petroleum Geology*, 150, p. 106097.

Weatherill, G., Burton, P. W., (2009). Delineation of shallow seismic source zones using K-means cluster analysis: Application to the Aegean region. *Geophysical Journal International*, 176, pp. 565–588.

Jufriansah, A., Pramudya, Y., Khusnani, A. and Saputra, S., (2021). Analysis of earthquake activity in Indonesia by clustering method. *Journal of Physics: Theories and Applications*, 5, p. 92.

Wilkin, G. A., Huang, X., (2007). K-means clustering algorithms: Implementation and comparison. *IMSCCS 2007*, IEEE, pp. 133–136.

Jain, A. K., (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31, pp. 651–666.

Žalik, K. R., (2008). An efficient k′-means clustering algorithm. *Pattern Recognition Letters*, 29, pp. 1385–1391.

Appendices

Appendix 1: R code used in PCA

```
fromsklearn.preprocessingimportStandardScaler
scaler=StandardScaler()
df_scaled=scaler.fit_transform(df_numeric)
importnumpyasnp
cov_matrix=np.cov(df_scaled,rowvar=False) print("CorrelationMatrix:\n",cov_matrix)
fromsklearn.decompositionimportPCA
pca=PCA() pca.fit(df_scaled)#Valeurspropres
eigenvalues=pca.explained_variance_print("Valeurspropres:",eigenvalues)#Vecteurspropres
eigenvectors=pca.components_ print("Vecteurspropres:",eigenvectors)
df_pca=pca.transform(df_scaled)
#Créationd'unDataFramepourlescomposantesprincipales
df_pca=pd.DataFrame(df_pca,columns=[f'PC{i+1}' foriinrange(df_pca.shape[1])]) print(df_pca.head())
df_pca_array=df_pca.to_numpy()
cos2=(df_pca_array**2)/(df_pca_array**2).sum(axis=1)[:,None]
cos2_df=pd.DataFrame(cos2,columns=df_pca.columns,index=df_pca.index) print("Qualitéderepresentation(cos2):")
print(cos2_df.head())
importseabornassns
importmatplotlib.pyplotasplt
plt.figure(figsize=(10,7)) sns.scatterplot(x='PC1',y='PC2',data=df_pca) plt.title('IndividualPlot')
plt.xlabel('PrincipalComponent1')
plt.ylabel('PrincipalComponent2') plt.show()
plt.figure(figsize=(10,7))
fori,(x,y)inenumerate(zip(pca.components_[0],pca.components_[1])): plt.arrow(0,0,x,y,color='r',alpha=0.5)
plt.text(x,y,df_numeric.columns[i],color='b') plt.xlim(-1,1)
plt.ylim(-1,1)
plt.title('Circleofcorrelations') plt.xlabel('MainComponent1')
plt.ylabel('MainComponent2') plt.grid()
plt.show()
```

Appendix 2: R code used in Kmeans

```
plt.figure(figsize=(8,6))
plt.scatter(data_pca[:,0],data_pca[:,1],c=colors,marker='o')
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],
marker='X',color='red',s=200,label='Centroid')
plt.title('K-meansClusteringwith4Clusters(PCA-ReducedData) andClusterLabels')
plt.xlabel('PrincipalComponent1')
plt.ylabel('PrincipalComponent2')
handles=[plt.Line2D([0],[0],marker='o',color='w',
markerfacecolor=color_map[label],markersize=10)forlabelincolor_map]plt.legend(handles=handles,label
s=color_map.keys(),
title="ClusterLabels") plt.grid(True) plt.show()
kmeans=KMeans(n_clusters=4,random_state=42) kmeans.fit(data_pca)
labels=kmeans.labels_ df['Cluster']=labels
cluster_mapping={0:'Low',1:'High',2:'Medium',3:'Weak'}
df['Cluster_Label']=df['Cluster'].map(cluster_mapping)
color_map={'Low':'blue','High':'orange','Medium':'green',
'Weak':'pink'}colors=df['Cluster_Label'].map(color_map)
```