

APPLICATION OF PRINCIPAL COMPONENT ANALYSIS (PCA) FOR EVALUATION OF RELEVANT DROUGHT-RELATED PARAMETERS

Ninoslava Tihi¹, Miloš Todorov², Srđan Popov³

Abstract: Drought is a complex natural phenomenon that requires accurate assessment and monitoring to mitigate its impacts on agriculture, water resources, and ecosystems. Principal Component Analysis (PCA) offers a robust statistical method for evaluating multiple drought-related parameters by reducing dimensionality while retaining critical information. This paper explores the application of PCA to identify and assess the most relevant drought-related parameters. By analyzing historical datasets, PCA effectively identifies patterns and correlations among these variables, enabling the extraction of principal components that explain the majority of variance within the dataset. The results provide insights into the dominant factors driving drought conditions and facilitate the creation of a composite drought index for improved monitoring. The results indicate that soil moisture positively correlated with wind speed and precipitation, whereas air temperature showed the most significant negative correlation. PCA analysis is very useful, as it simplifies multi-dimensional drought datasets, enhances parameter prioritization, and supports decision-making processes for drought preparedness and mitigation strategies.

Keywords: Principal component analysis (PCA), urban drought, drought-related parameters, dimensionality reduction, statistical method.

1. INTRODUCTION

Drought is a recurring natural phenomenon that significantly impacts agriculture, water resources, ecosystems, and economies worldwide. Monitoring and evaluating drought-related parameters is crucial for understanding its intensity, duration, and potential impacts [1]. The complexity and variability of this phenomenon, however, require efficient methods to analyze and interpret multiple, often interrelated, environmental variables [2].

Principal Component Analysis (PCA) is a powerful statistical technique that is widely used for dimensionality reduction, feature extraction, and data interpretation [2]. By transforming high-dimensional datasets into a smaller set of uncorrelated variables called principal components, PCA helps in simplifying complex datasets while retaining the majority of the variance. This makes PCA particularly useful in the context of drought analysis, where multiple climatic, hydrological, and soil-related variables must be considered simultaneously to assess drought conditions.

The application of PCA in evaluating drought-related parameters offers several advantages. It enables the identification of key factors driving drought patterns, assists in the reduction of noise from irrelevant variables, and highlights the most influential components that contribute to the onset and severity of drought. By extracting these principal components, decision-makers can gain a clearer, more concise understanding of drought dynamics, facilitating improved forecasting, risk assessment, and resource management strategies.

This study explores the application of PCA to evaluate and prioritize drought-related parameters, such as precipitation, temperature, soil moisture, wind speed, wind direction, and air humidity, demonstrating its utility in managing and mitigating the adverse effects of drought in vulnerable regions. There is no one method that can be used to measure the severity of drought [3]. Through the identification of critical factors influencing drought severity, PCA can enhance the decision-making processes for sustainable water resource management, agricultural planning, and climate resilience efforts.

¹MSc, The Higher Education Technical School of Professional Studies in Novi Sad, Školska 1, e-mail: tihi@vtsns.edu.rs

²PhD, Faculty of Mathematics and Computer Science, Alfa BK University, Belgrade, Serbia, e-mail: milos.todorov@alfa.edu.rs

³PhD, Faculty of Technical Sciences in Novi Sad, University of Novi Sad, Novi Sad, Serbia, e-mail: srdjanpopov@uns.ac.rs

2. THEORETICAL BACKGROUND

The impact of different variables on certain outcomes or situations is analyzed in data analytics. There are three categories of analysis: Univariate, Bivariate, and Multivariate analysis. Multivariate analysis techniques refer to all statistical techniques that are used to analyze multiple variables at once.

Multivariate analysis techniques can be divided into two categories: dependence techniques and interdependence techniques [2]. Dependence methods are usually used to build predictive models where one or more variables are dependent on others. Interdependence methods are used to understand the dataset structure and its patterns within the dataset. These methods are not looking for causal relationships between variables but rather to group them in meaningful ways.

Principal Component Analysis (PCA) is a widely used unsupervised learning technique for dimensionality reduction, pattern recognition, and data exploration. While in supervised learning methods the goal is to predict a class or a number, in unsupervised learning the goal is to look for patterns in the data which is shown in Figure 1. It is used when there is no label or target for each observation in the dataset and it simplifies the complexity of high-dimensional data while retaining its essential structure and variance. In general, PCA is a highly helpful method for representing spatially dependent data [4]. The idea is to reduce the number of variables of a dataset while preserving as much information as possible.

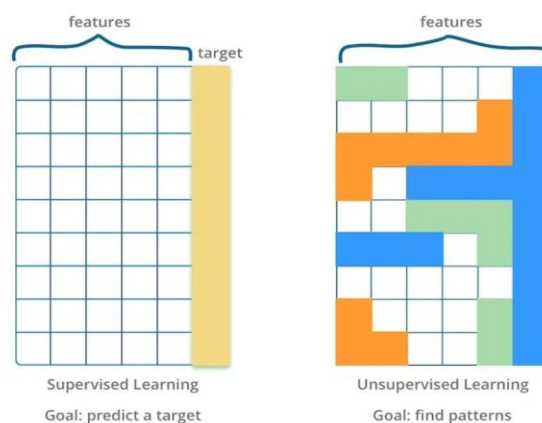


Figure 1 -The goals of supervised vs. unsupervised learning

PCA consists of five main steps (Figure 2) for computing principal components which are:

- Standardization
- Covariance Matrix Calculation
- Eigenvalue and Eigenvector Computation
- Sorting Eigenvectors
- Projection

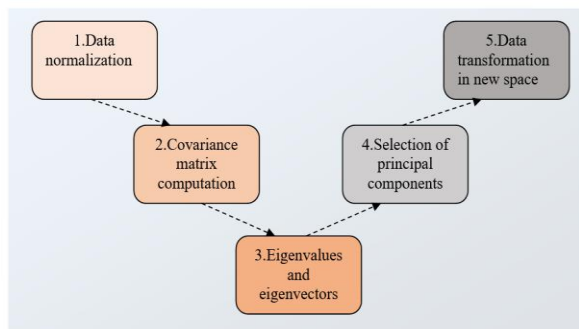


Figure 2 -The five main steps for computing principal components

1.1. Standardization

Since Principal Component Analysis is very sensitive to the scales of the data, it is often necessary to normalize or standardize the data especially when the variables have different units or magnitudes (e.g., temperature in Celsius, precipitation in millimeters). Standardization ensures that each variable contributes equally to the analysis. It is very important to normalize data features especially when they have different scales because normalizing will make all values across features comparable since each data feature is transformed into a common representation. There are several ways to normalize the features usually called feature scaling and one of them is the Z-score Normalization also referred to as Standardization.

To apply Z-score normalization, each data point value needs to be updated by using this formula [5]:

$$Z = (X - \text{mean}) / \text{standard deviation} \quad (1)$$

Where:

Z – standardized vector,

X – vector,

mean – vector's mean,

$\text{standard deviation}$ – vector's standard deviation.

1.2. Covariance matrix calculation

Once the data is standardized, PCA computes the covariance matrix, which measures the relationship between pairs of variables. So generally speaking, the primary objective of PCA is to transform an original set of possibly correlated variables into a smaller set of variables called principal components (PCs). These new variables, linear combinations or mixtures of the original variables, capture the maximum possible variance in the data with fewer dimensions. The covariance matrix shown in Figure 3 is used to identify noise and redundancy in the dataset. PCA analyzes the variance of each feature and the strength of the linear relationship between pairs of features. Variance analysis of each feature will show if a feature is relevant or a pure noise and this is shown in the diagonal cells of the matrix. The strength of linear relationship between pairs of features helps to spot redundant features and it is shown in all non-diagonal values. These off-diagonal elements represent covariance between pairs of variables. Covariance measures how much two variables change together. A positive covariance means that as one variable increases, the other tends to increase as well, and a negative covariance indicates an inverse relationship. So in summary, the covariance matrix highlights how much each variable deviates from its mean (variance) and how two variables co-vary or change together (covariance) [6].

Explained variance represents the proportion of the total variance in the data that is explained by each principal component. It helps in determining how many components are needed to capture the majority of the variance in the data. Maximum possible information is compressed in the first component and then the maximum remaining information is put in the second and so on with each succeeding component being uncorrelated to the others. In the end, the result is a dataset that reveals the most important characteristics of the data.

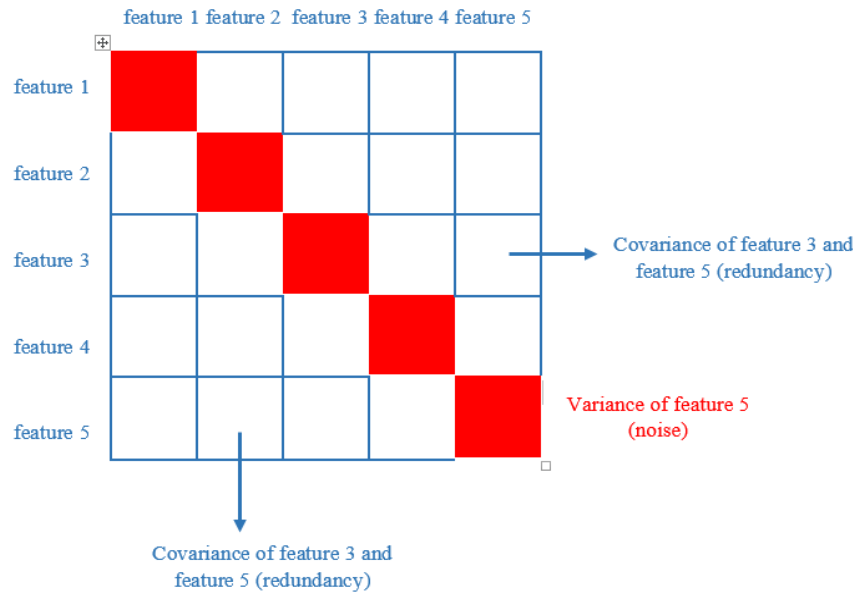


Figure 3 -Interpretation of covariance matrix

1.3. Eigenvalue and eigenvector computation

PCA involves solving the covariance matrix to compute the eigenvalues and eigenvectors. A scalar value that represents the amount of variance that is captured by each principal component is called an eigenvalue. A higher eigenvalue indicates that the principal component accounts for more of the variance in the data. The eigenvalue, represented with λ in the equation for calculation eigenvector, is associated with the eigenvector v .

Mathematically, solving the equation $A * v = \lambda * v$ involves finding the eigenvalue λ and the corresponding eigenvector v . This is equivalent to solving the characteristic equation [6]:

$$\det(A - \lambda I) = 0 \quad (2)$$

Where:

$\det$ –the determinant of the matrix,

A – matrix,

λ – eigenvalue,

I – identity matrix of the same size as A

A vector that defines the direction of the principal components is called an eigenvector. It shows how the original variables contribute to each principal component. In other words, eigenvalues indicate the variance captured by each principal component, while eigenvectors represent the direction of these components in the multidimensional space.

Mathematically, for a square matrix A and a vector v , an eigenvector satisfies the equation [6]:

$$A * v = \lambda * v \quad (3)$$

Where:

A – square matrix (linear transformation),

v – eigenvector,

λ – eigenvalue.

1.4. Sorting eigenvectors

Sorting eigenvectors helps to prioritize those that have the largest eigenvalues because eigenvectors with the largest eigenvalues capture the most variance in the dataset or carry the most significant information about the data, and thus, are selected as the principal components. Sorting starts with computing the eigenvalues ($\lambda_1, \lambda_2 \dots$) and their corresponding eigenvectors ($v_1, v_2 \dots$). Once the eigenvalues are computed, they are sorted in descending order from the largest to smallest because the most significant directions (in terms of variance or contribution) need to be first. After sorting the eigenvalues, corresponding eigenvectors need to be rearranged in the same order so that each eigenvector is associated with its eigenvalue.

1.5. Projection

Finally, the original data is projected onto the selected principal components, reducing the data from a higher-dimensional space to a lower-dimensional one. In other words, data are transformed from their original coordinate system into a new coordinate system defined by the principal components or eigenvectors of the covariance matrix. This transformation allows the data to be represented in a way that maximizes the variance along the new axes, making it easier to analyze and visualize high-dimensional data in lower dimensions.

Mathematically, the projection of the data onto the principal components is given by [6]:

$$Y = X * V \quad (4)$$

Where:

X – matrix of original data,

V – matrix of eigenvectors (sorted by eigenvalue),

Y – transformed (projected) data in the new space of principal components.

The new data matrix Y contains the projections of the original data onto the principal components, with each column representing the projection along a particular principal component.

2. RESEARCH METHODOLOGY

2.1. Software selection

For this research, a combination of R programming language and RStudio IDE (version 2021.09.0 Build 351), was used as a software tool. R is a statistical programming language widely used in the academic community for all sorts of statistical computing, data analysis, and data visualization. A comprehensive R archive network called CRAN serves as a central repository for R software packages which are collections of functions, data, and documentation that extend the functionality of the R programming language. CRAN is accessible via its website (<http://cran.r-project.org>) where users can search for packages, download R, and access documentation on R and its packages [7].

2.2. Data selection

The raw data were collected by various sensors from the sensor station located near Kać in the vicinity of Novi Sad from January 2014 to December 2020. Once the data are collected, they are sent to the web portal. From there, data obtained from different sensors can be saved in Excel format. The sensors belong to a type of network called „Proactive Networks“. Those sensors record data each hour so in total there are 288 measurements within 24h. The sensors can be divided into seven categories. The first category represents six sensors that measure only soil moisture (SM1, SM2, SM3, SM4, SM5, SM6). The other categories include a humidity sensor (AH1), wind speed (WS1), wind direction (WD1), air temperature (AT1), precipitation (PP1), and battery power state (BT1). All sensor measures were used for this study, except the BT1 sensor's readings, which were not included because of their insignificance. Since SM3, SM4, SM5, and SM6 readings were obtained from sensors positioned deeper in the soil and their results may have an adverse effect on the analysis, they were also not included in the study [8].

The data set consists of four attributes such as time, device, ID value, and value. Descriptions of all the attributes are given in Table 1 [9].

Table 1 □ Structure of the obtained data

Attribute name	Description
Time	represents the date and time of the measured parameter
Device	represents the device identification number
ID value	represents the sensor identification number
Value	represents the measured value of the parameter

2.3. Data Preparation

After the data have been retrieved from the online portal in Excel format, they are converted and stored in a data frame. A data frame is a two-dimensional, tabular data structure used in R and other programming languages for handling and analyzing data in a structured format for statistical and data science tasks. The preparation process first includes filtering the insignificant parameters and then aggregating only the values of SM1 and SM2 by using the „mean“ function to get the average values. In the end, the data are saved as a data table and arranged in ascending order according to the months and years. After the data were prepared, PCA can be applied [9].

3. RESULTS AND DISCUSSION

The first step in applying PCA is to normalize the data and to ensure that all features are normally distributed. This means that data will follow a standard normal distribution, with a mean equal to zero and a standard deviation equal to one. Normalization in R can be applied using the *scale()* function. Normalized data are shown in Table 2.

Table 2 □ Normalized data

SM1	PP1	AT1	AH1	WS1	WD1
0.6534625	2.1500275	-1.14505313	0.95668922	0.9638429	0.45620579
1.1233465	2.9615747	-0.89623531	0.65690605	1.9776675	0.43843127
1.1403155	0.4487842	-0.37139775	-0.55781894	1.2195045	-0.01094872
0.7856930	0.8515520	0.05727868	-0.33016013	0.2715157	-1.18476687
1.5574475	3.1298956	0.43243547	-0.05144874	-0.1343998	-0.97857234
-0.3313081	-0.7054163	1.04287149	-1.01388171	-0.8323515	-0.92419611

Then, the PCA analysis is conducted using the `princomp()` function, which generates six principal components that explain a percentage of the total variance in the dataset. All six principal components, which also correspond to the number of variables in the data, are shown in Table 3.

Each component explains a percentage of the total variance in the dataset. In the Cumulative proportion section of the table, the first principal component explains almost 43% of the total variance, while the second, third, and fourth components explain nearly 61%, 78.73%, and 89.25% of the total variance respectively. This implies that almost half of the data in the set of six variables can be represented by just the first principal component. So the cumulative proportion of the first four principal components explains nearly 89,25% of the total variance. This means that the first four principal components can accurately represent the data.

Table 3 □Principal components of the data

Importance of components	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6
Standard deviation	1.5922063	1.0469163	1.0182074	0.7897297	0.70856587	0.36733085
Proportion of variance	0.4276107	0.1848732	0.1748729	0.1051978	0.08468576	0.02275961
Cumulative proportion	0.4276107	0.6124839	0.7873568	0.8925546	0.97724039	1.00000000

Next, the loading matrix will be calculated to explore how components relate to each column using the loadings of each principal component. The loading matrix shows that the first principal component has high positive values for soil moisture (SM1), precipitation (PP1), air humidity (AH1), and wind speed (WS1), but relatively negative values for air temperature (AT1) and wind direction (WD1). The second principal component has high negative values for air humidity (AH1) and high positive values for wind speed (WS1) and wind direction (WD1). The third principal component has high positive values for soil moisture (SM1), precipitation (PP1), and air temperature (AT1), while values for air humidity (AH1) and wind speed (WS1) are negative. The last component has relatively high values for wind speed (WS1) and high negative values for air humidity (AH1) and wind direction (WD1). The loading matrix of all four principal components is shown in Table 4.

Table 4 □Loading matrix of the first four principal components

	Comp. 1	Comp. 2	Comp. 3	Comp. 4
SM1	0.4543221	0.09903135	0.31989025	0.2711357
PP1	0.2229617	-0.05322383	0.85379274	-0.1939551
AT1	-0.5492468	-0.12460436	0.34317959	0.1126082
AH1	0.4572305	-0.34398801	-0.18339655	-0.6509583
WS1	0.4275757	0.52804842	-0.11183405	0.3535814
WD1	-0.2243362	0.75807050	0.06924449	-0.5722060

Analysis of the loading matrix showed the relationship between each of the first four principal components and the attributes in the data. However, it might not be visually appealing. There are certain standard visualization techniques such as scree plots which can visualize the importance of each principal component and can be used to determine the number of principal components to retain. The scree plot, shown in Figure 4, can be generated using the `fviz_eig()` function from the *factoextra* package. This plot shows the eigenvalues in a downward curve, from the highest to the lowest. The

first four components can be considered to be the most significant since they contain almost 89% of the total information of the data.

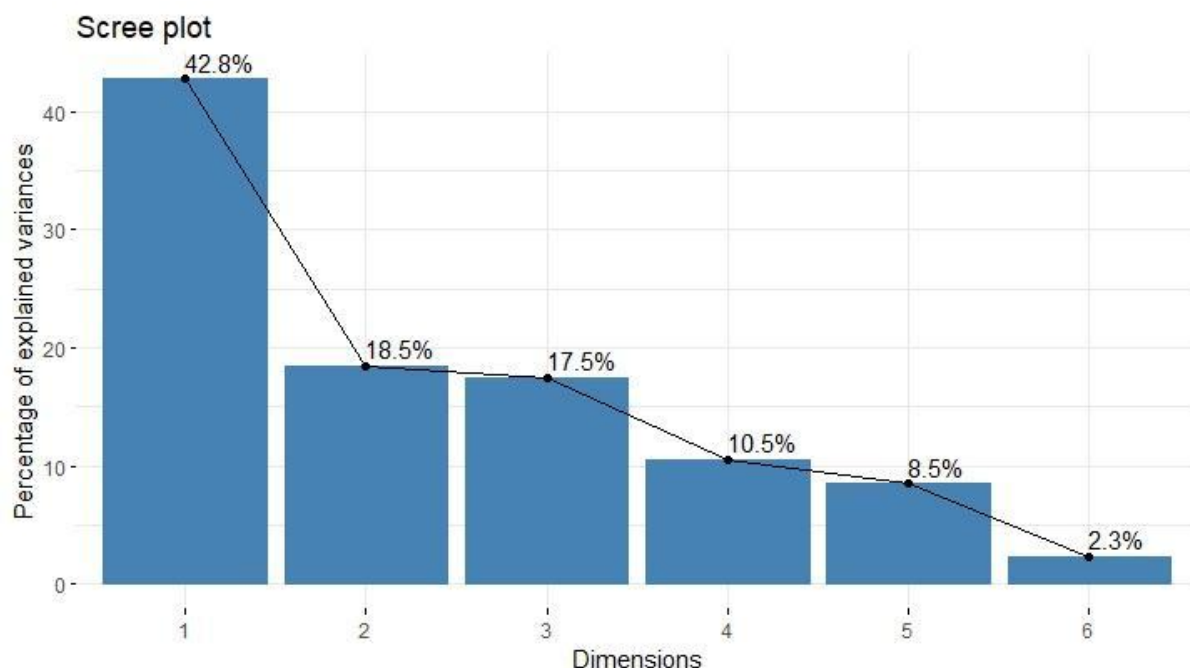


Figure 4 –Scree plot of the components

The biplot can be used to visualize the similarities and dissimilarities between samples and the impact of each attribute on each of the principal components. The biplot is shown in Figure 5.

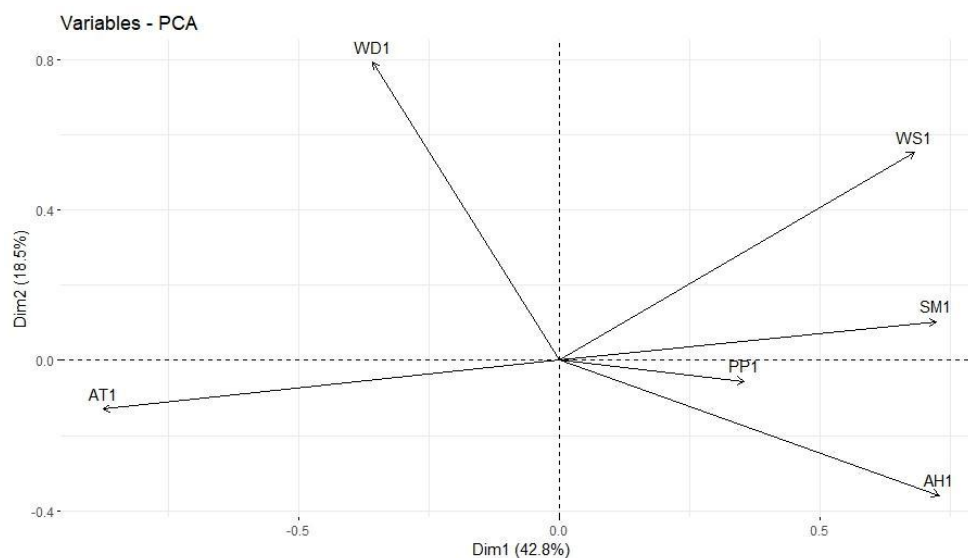


Figure 5 –Biplot of the variables concerning the principal components

Three main conclusions can be observed from the biplot. First, variables such as WS1 and SM1 that are grouped are positively correlated to each other. They have the highest values in the loading matrix concerning the first principal component. Then the higher the distance between the variable and the origin, the better represented that variable is. Since WS1 and SM1 have higher magnitudes, it means they are well represented. Finally, variables that are negatively correlated are displayed on the opposite sides of the biplot's origin. Variable AT1 is negatively correlated with WS1 and SM1.

The third visualization determines how much each variable is represented in a given component. Such a quality of representation is called the Cos2 and corresponds to the square cosine and it is computed using the *fviz_cos2()* function. A low value means that the variable is not well represented by that component while a high value means a good representation of the variable. Figure 6 shows the square cosine value for each variable with respect to the first two principal components.

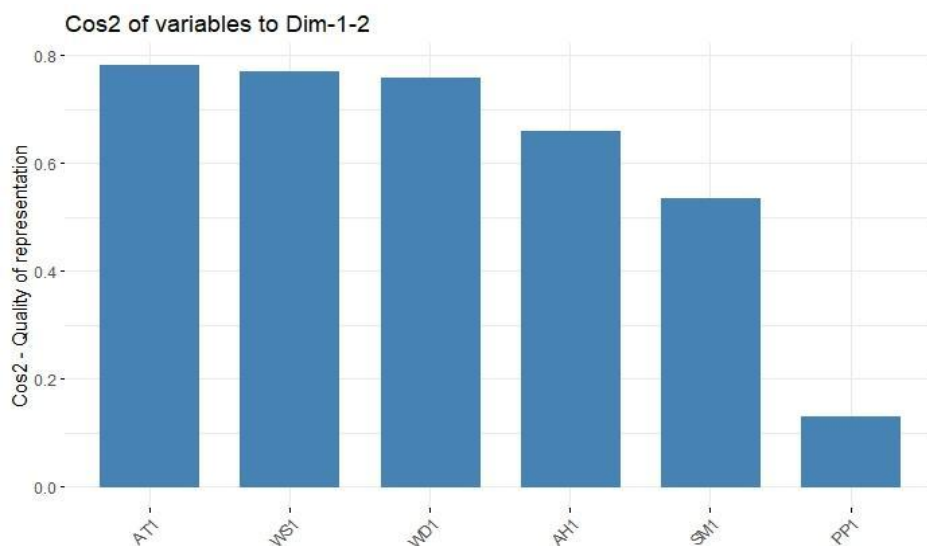


Figure 6 –Contribution of the variables to principal components

A covariance (correlation) matrix can be created using the R function *cor()*. The correlation matrix of the data with its coefficients is shown in Table 4.

Table 5 Correlation matrix

	SM1	PP1	AT1	AH1	WS1	WD1
SM1	1.00	0.35	-0.48	0.34	0.41	-0.17
PP1	0.35	1.00	-0.04	0.18	0.15	-0.08
AT1	-0.48	-0.04	1.00	-0.66	-0.66	0.20
AH1	0.34	0.18	-0.66	1.00	0.20	-0.33
WS1	0.41	0.15	-0.66	0.20	1.00	0.01
WD1	-0.17	-0.08	0.20	-0.33	0.01	1.00

The function *geom_tile()* from the *ggplot2* package is used to visualize the correlation matrix. Negative correlations are in purple and positive correlations are in red. To identify hidden patterns the matrix needs to be reordered according to the correlation coefficients. Figure 7 shows the correlation matrix with added correlation coefficients visualized with the heatmap. It can be concluded that for example soil moisture and air humidity are in negative correlation with air temperature and wind direction while in positive correlation with precipitation and air humidity. Wind speed has in high negative correlation with the air temperature but in positive correlation with the precipitation, air humidity, and soil moisture.

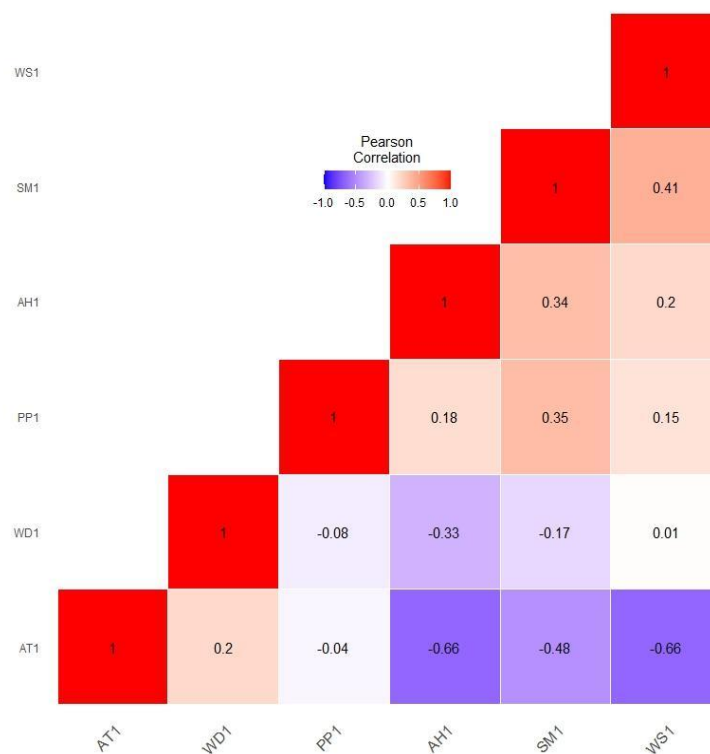


Figure 7 –Correlation matrix with added coefficients

4. CONCLUSION

A reliable and effective technique for lowering dimensionality and locating important factors that aid in drought characterization is the use of Principal Component Analysis (PCA) to assess pertinent drought-related data. With PCA, researchers may preserve the majority of the original variance while condensing large datasets into a smaller number of main components. This process highlights the interrelationships among various parameters, such as precipitation, temperature, and soil moisture, which are critical for understanding drought dynamics.

PCA improves monitoring systems, helps prioritize mitigation methods, and makes it easier to create more accurate and interpretable drought indices by separating the most important components. Additionally, PCA is a useful tool for stakeholders and policymakers entrusted with controlling drought risks due to its ability to detect patterns and trends across a variety of spatial and temporal scales.

Research shows that soil moisture and air humidity negatively correlate with air temperature and wind direction, while positively with precipitation and air humidity. Wind speed positively correlates with precipitation, air humidity, and soil moisture but negatively correlates with air temperature.

In conclusion, the integration of PCA into drought assessment frameworks provides a powerful mechanism for improving the accuracy of drought evaluations, advancing scientific understanding, and supporting evidence-based decision-making. Future studies could explore the combination of PCA with machine learning techniques to further refine drought prediction and management strategies.

5. REFERENCES

- [1] Mishra, A. K.; Singh, V. P.: *A review of drought concepts*, Journal of Hydrology 391(1–2): 202–216, 2010, doi: 10.1016/j.jhydrol.2010.07.012.
- [2] Hair, J. F.: *Multivariate data analysis*, Eighth edition. Cengage, Andover, Hampshire: 2019.
- [3] Keyantash, J.; Dracup, J. A.: *The Quantification of Drought: An Evaluation of Drought Indices*, Bull. Amer. Meteor. Soc. 83(8), 1167–1180, 2002, doi: 10.1175/1520-0477-83.8.1167.
- [4] Tadić, L.; Brleković, T.; Potočki, K.; Leko-Kos, M.: *Application of principal component analysis to drought indicators of three representative Croatian regions*, Elektronički časopis građevinskog fakulteta Osijek: 41–55, 2021, doi: 10.13167/2021.22.4.
- [5] Jolliffe, I. T.; Cadima, J.: *Principal component analysis: a review and recent developments*, Phil. Trans. R. Soc. A. 374(2065): 20150202, 2016, doi: 10.1098/rsta.2015.0202.
- [6] *Principal Component Analysis*. in Springer Series in Statistics. New York: Springer-Verlag, 2002. doi: 10.1007/b98835.
- [7] Tihi, N.; Popov, S.; Bondžić, J.; Dujović, M.: *Visualization of Big Data as Urban Drought Monitoring Support in Smart Cities*, Fresenius environmental bulletin 30(01): 719–723, 2021.
- [8] Tihi, N.; Popov, S.: *Selection of the Best ARIMA models for Urban Drought Prediction*, Fresenius Environmental Bulletin 32(06): 2564–2572, 2023.
- [9] Tihi, N.; Popov, S.: *A Comparison of Arima and Random Forest Time Series Models for Urban Drought Prediction*, in *Proceedings of the International Scientific Conference - Sinteza 2024*, Beograd, Serbia: Singidunum University, 2024, pp. 51–56. doi: 10.15308/Sinteza-2024-51-56.