Kurdistan Journal of Applied Research (KJAR)
Print-ISSN: 2411-7684 | Electronic-ISSN: 2411-7706

Website: Kjar.spu.edu.iq | Email: kjar@spu.edu.iq

Application of Principal Component Analysis for Steel Material Components

Kawa Muhammad Jamal Rasheed
Statistics Department
College of Administration and Economics
university of Sulaimani
Sulamaniyah, Iraq
kawa.rashid@univsul.edu.iq

Miran Othman Tofiq
Statistics Department
College of Administration and Economics
university of Sulaimani
Sulamaniyah, Iraq
miran.toffiq@univsul.edu.iq

Article Info

Volume 7 - Issue 2- December 2022

DOI: 10.24017/Science.2022.2.7

Article history:

Received: 22/09/2022
Accepted: 01/12/2022

Keywords:

Principal Component Analysis (PCA) , multivariate, transformation , uncorrelated , Eigenvalues.

ABSTRACT

In this research, we made use of a technique known as principal component analysis (PCA). PCA is an approach to statistics that takes into account several variables that converts a fixed number of correlated variables into a fixed number of orthogonal, uncorrelated axes known as principal components by making use of orthogonal transformation. In other words, PCA transforms correlated variables into uncorrelated axes. The principal component analysis (PCA) method, to put it another way, transforms correlated variables into uncorrelated axes. We used (PCA) technique in order to bring the dimensionality of a data collection down to a more tolerable level that had a wide variety of interconnected variables while yet preserving as much of the natural variation that existed within the data set. Because of this, we were able to examine eleven different steel components. To do this Each independent variable is combined with others to generate a third independent variable. Known as principal components, which are not associated with one another (PC). The order of the principle components is chosen in such a manner that they maintain the vast majority of the variety that exists in each and every one of the several variables. This is accomplished by using a variogram. This is accomplished by reworking the individual variables into a brand new group of variables called principle components, which are not associated with one another (PC). Because this percentage reflects the principal aspect that is best among all 11 principal components, we are able to reach the conclusion that the five principal components that collectively account for approximately sixty-seven percent of the Total variance in all of the data are the best principal components. This allows us to come to the conclusion that the best principal components are the five principal components that collectively account for approximately sixty-seven percent of the variance in all of the data.

1. INTRODUCTION

The rapidly developing discipline of machine learning depends heavily on the use of statistics, which is a branch of mathematics. It requires the examination, interpretation, and presentation of numerical information pertaining to the past, the present, and the future [1]. Statistics is a collection of methods that can be used for the purpose of creating evaluations about a process or population based on an examination of the facts obtained from a sample taken from that population. Statistics is the discipline of science that is concerned with the delineation of effects and the construction of perspectives based on data that is subject to variation. This is the purpose of statistics. The identification of statistical trends is an essential step toward achieving first-rate management and improvement. They provide the main method that is used in order to try, test, and estimate a product, and the facts that are contained in the statistics are employed in order to modify and enhance the product. In a similar vein, statistics are the language that development masterminds, production, procurement, operation, and other functional components of the company utilize to communicate with one another on quality [2]. When performing research in statistics, it is standard practice to revise an explanation of a phenomenon after completing an analysis of data acquired through experiments or observations. This is done in order to account for new information that has been uncovered. During this iterative process of learning, it is possible that new variables may be added, or that previously used variables will be removed. This is necessary in order to draw valid conclusions about the phenomenon under investigation. This body of methodology is known as multivariate analysis and it gets its name from the fact that the data comprises measurements on a number of different variables all at the same time[3].

Principal Component Analysis (PCA), which is one of the methods that are used in multivariate analysis, and it is statistics approach that takes into account several variables that is the process of changing a group of correlated variables into a group of orthogonal, uncorrelated axes often referred to principal components [4]. PCA is one of the techniques that are used in multivariate analysis and one of the methods that is used in multivariate analysis. Moreover, Principal component analysis is a statistical approach that aims to decrease the number of interrelated variables that are present in a data set while retaining as much of the set's intrinsic variability as is practically possible. Changing to a new group of uncorrelated variables, which are known as principle components, is one way to achieve this goal principal component (PC), which can then be ordered in such a way that the principal components hold the majority of the variant that is determined in all of the unique variables. This can be done by switching to a new set of uncorrelated variables known as principal components (PC). Utilizing a principal component analysis to accomplish this goal is one option [5]

2. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is one of the methods that are used in multivariate analysis, PCA is a statistical approach that takes into account several variables, PCA is the process of changing a group of correlated variables into a group of orthogonal that are needed to analyze a set of variables [4]. This method is known as the "principal component" analysis. The key compounds, often referred to as axes, are the variables that are formed as a result of this transformation process. The main objective of principal component analysis (PCA) is to reduce the size of a dataset that contains a high number of different variables that are correlated to one another while still maintaining as much variety as is practically possible. This is accomplished through the use of an iterative process known as principal component extraction (PCEE). This is accomplished by converting the data into a new set of variables, each of which has no correlation with the others in the set [5] One of the first and best-known methods for multivariate analysis is principal component analysis. It resolves the issue of how well a low-dimensional interpolation subspace is likely to fit to a collection of data points that are located in a high-dimensional space. [7]. Additionally, it is one of the multivariate analysis approaches that is used by the greatest number of people. As shown in Figure (1) and Figure(2) [2] the directions of maximum variability are represented by the axes of a new

coordinate system that was generated by rotating the axes of the first system. This new system was produced by combining the two previous systems to fit a low-dimensional interpolation subspace to a set of data points in a high-dimensional space [6]. It is also one of the most widely used multivariate analysis methods. As demonstrated in Figure (1) and Figure(2) [2] the directions of highest variability are represented by the axes of a new coordinate system that was produced by rotating the axes of the first system.

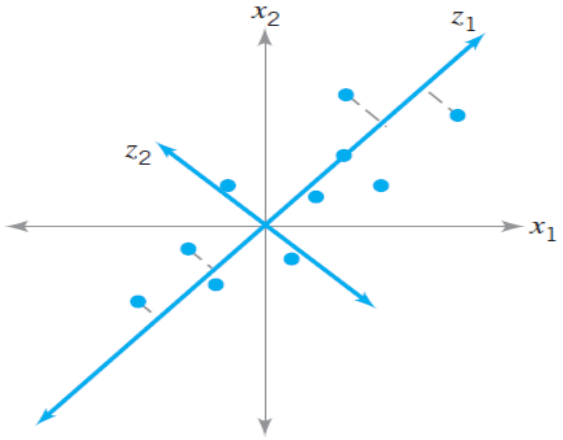

Figure 1: Principal components for p =2 process variables.

The majority of the differences between the two primary variables are explained by the first principal component, as seen in Figure 1.

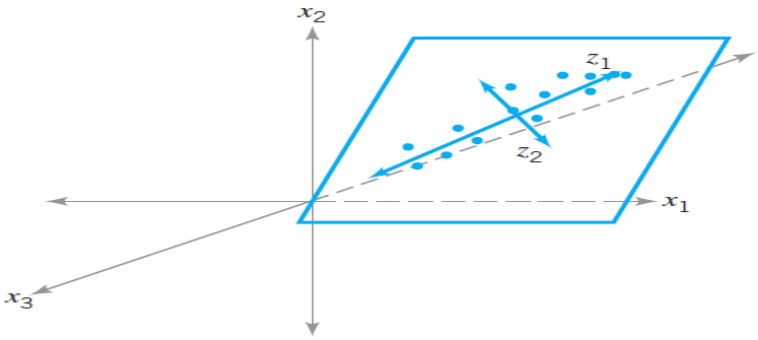

Figure 2: Principal components for p =3 process variables

In Figure (2), we can see an illustration of the three initial process variables. Due to the fact that the majority of the "motion" or variability in these two variables is contained inside a line, just two main components have been utilized to characterize them.

The scree plot is yet another method that may be used in order to ascertain the suitable quantity of components. The eigen values are plotted against the component number first in the resulting graph. The scree plot illustrates the significantly reduced rate at which variance is explained by additional principal components, and this scree plot is shown in Figure (3) [6]. Eigen values are constrained to minimize in a monotonically decreasing fashion from the first principal component to the last principal component.

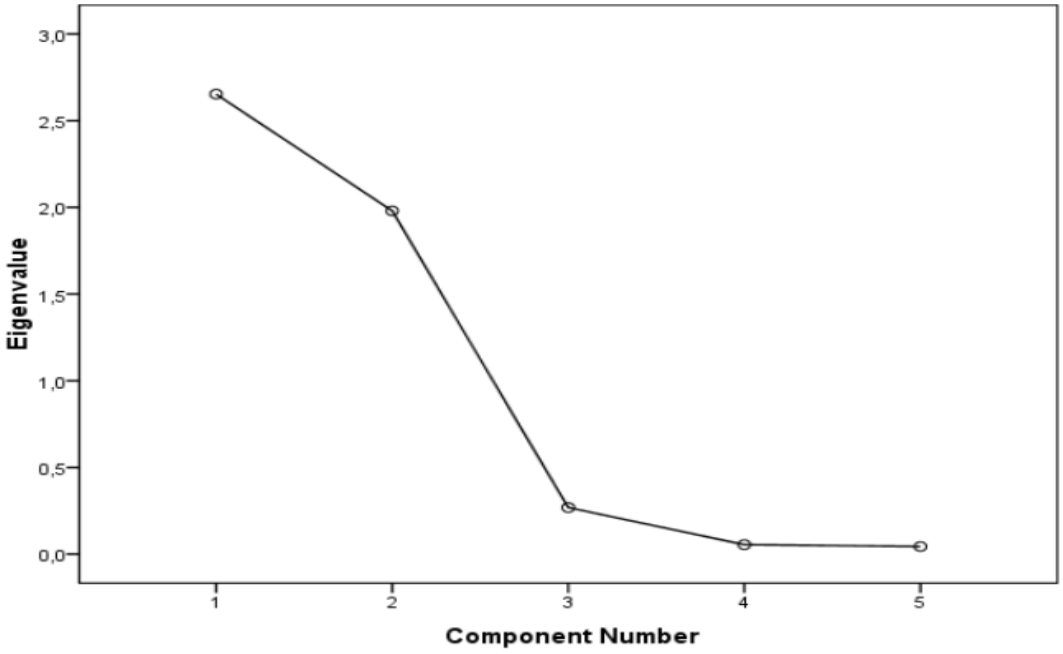

Figure 3: shows scree plot for the Eigen values of components

We can explain the principal component Analysis by the following equations [8]:

$$Z_i = a_{i1}x_1 + a_{i2}x_2 + ... + a_{ip}x_p \quad i = 1, 2, ..., p \ (1)$$

the above equation can be express it by matrices approach

$$\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \ (2)$$

the principal component variables $Z_1, Z_2, \cdots, Z_p$ *is the value of* vector Z and The random variables $x_1, x_2, \ldots x_p$ *is the value of* vector x and the eigenvalues be $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$. Then the constants a_{ij} are simply the elements of the i^{th} eigenvector associated with the eigenvalue λ_i [9].

Then we calculate the correlation matrix or the covariance matrix for the explanatory variables as it shows in the eq(3) [8]:

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \ (3)$$

r_{ij} is simple correlation between two variables, when i=1,2,3,….,ρj= 1,2,3,….,ρ

Then we can find the value of the characteristic roots(λ_i) by solving the characteristic equation of the correlation matrix, whichis :

$$|R-\lambda_i| = \begin{vmatrix} 1-\lambda & r_{12} & \cdots & r_{1p} \\ r_{21} & 1-\lambda & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1-\lambda \end{vmatrix} = 0 \quad (4)$$

Where:

$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p$

The first characteristic vector corresponding to the first characteristic value of the following equation [10]:

$$(R-\lambda_1)\underline{a_1} = 0 \quad (5)$$

The values of the elements of this characteristic vector are chosen as long as the following condition is satisfied [8]

$$\underline{a_1'}\underline{a_1} = 1$$

Thus, the first principal component is in the following form:

$$Z_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p \quad (6)$$

Then we can find the second characteristic vector corresponding to the second characteristic value from the following equation[10]:

$$(R-\lambda_2)\underline{a_2} = 0 \quad (7)$$

Thus, the second principal component is in the following form [9] :

$$Z_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p \quad (8)$$

The values of the elements of this characteristic vector are chosen as long as the following condition is satisfied [8]

$\underline{a'_2}\,\underline{a_2} = 1$, $\underline{a'_1}\,\underline{a_2} = 0$

It is a condition that the components (Z_1, Z_2) are orthogonal.

For the third component is in the following form [2] :

$$Z_3 = a_{31}x_1 + a_{32}x_2 + \ldots + a_{3p}x_p \quad (9)$$

The values of the elements of this characteristic vector are chosen as long as the following condition is satisfied [9]

$\underline{a'_3}\,\underline{a_3} = 1$, $\underline{a'_1}\,\underline{a_3} = 0$, $\underline{a'_2}\,\underline{a_3} = 0$

It is a condition that the third component (Z_3) is orthogonal withthe other two component (Z_1, Z_2) .
We will continue like this until we complete all the main components one after the other, provided that each new one is orthogonal to all the previous ones.

And the variance and covariance of Z_i is explained as followed

$$Var(x_i) = \sigma_{ii}\ i = 1, 2, \ldots., p$$
$$Var(Z_i) = \lambda_i\ i = 1, 2, \ldots., p$$
$$\mathrm{Cov}(Z_i, Z_j) = 0$$

The ratio, which is described as follows, will provide us the percentage of the original data's variability that can be explained by the ith principal component. This may be found by using the following formula: [2] :

$$\frac{Var(Z_i)}{\sum Var(Z_i)} = \frac{\lambda_i}{\sum \lambda_i} \quad (10)$$

Therefore, one can easily see how much variability is explained by retaining just a few of the principal components simply by computing the sum of the eigen values for those components and comparing that total to the sum of alleigen values.

3. DESCRIPTION OF THE DATA

The data was collected in Sulaimani Steel Company, and it consisted of 400 observations and 11 variables include the following materials as it is shown in table(1) and table(2), and it produces a high quality steel rebar in many sizes10mm, 12mm, 16mm, 20mm, 25mmwhich 10mm is equal to 0.62 kg per meter, 12mm is equal to 0.89 kg per meter,16mm is equal to 1.58 kg per meter,20mm is equal to 2.47 kg per meter,25mm is equal to 3.86 kg per meter, Sulaimani Steel Company was established in 2012 to create and manufacture various types of construction steel in accordance with standards established in the United States and the United Kingdom. To minimize the dimensionality of a data collection that consists of a high number of variables that are correlated to one another, while preserving as much of the variance that is already there in the data set as is feasible, we use a method of principal component analysis (PCA), and we use this method to reduce the dimensionality of a data set that was constructed by Sulaimani Steel Company and began production in 2015 by using R-programming to analyze the data. This method is used to reduce the dimensionality of a data set by using eq(10) and data from table(2) as is shown in table(3).

Table 1: Steel Materials Components

Component	Full Name	unit	Summary of steel components	component values limit
C	Carbon	gram	Carbon, which is one of the most important chemical elements in steel. An increase in carbon content yields a material with lower ductility and higher strength	0.27-0.35
Si	silicon	gram	Silicon, is one of the principal deoxidizers for structural steel	0.2-0.7
Mn	Manganese	gram	Manganese, is probably the second most important alloying element after carbon on steel	1.1-1.3
P	Phosphorus	gram	Phosphorus, is generally considered to be an undesirable impurity in steels, Phosphorus is often added with sulphur to improve	0.01-0.05

			machinability	
S	Sulphur	gram	Sulphur, is present in raw materials used in iron making	0.04-0.08
Cr	Chromium	gram	Chromium, are the most commonly found residuals in steel, Chromium is present in small amounts and is used in combination with copper and nickel to increase the material's resistance	0.1-0.4
Ni	Nickel	gram	Nickel, is added to steels to increase hardenability, It is frequently used to improve toughness at low temperature	0.06-0.3
Cu	Copper	gram	Copper, it helps paint bond the steel. It also has a small impact on hardenability	0.23-0.8
Al	Aluminum	gram	Aluminum, is used primarily as a deoxidizing agent in steelmaking, combining with oxygen in the steel to form aluminum oxides which can float out in the slag	0.002-0.007
Mo	Molybdenum	gram	Molybdenum, is used to increase the strength of boiler and pressure vessel steels at typical boiler operating temperatures of 400°C	0.01-0.03
V	Vanadium	gram	Vanadium, are strengthening elements that are added to steel singly or in combination. In very small quantities they can have a very significant effect	0.001-0.009

Table 2: Data Table

Date	C	Si	Mn	p	⋯	⋯	⋯	V
2/1/2022	0.356	0.321	1.21	0.0335	⋯	⋯	⋯	0.005
2/1/2022	0.285	0.388	1.18	0.0272	⋯	⋯	⋯	0.0036
2/1/2022	0.298	0.385	1.21	0.0287	⋯	⋯	⋯	0.0042
2/1/2022	0.281	0.324	1.03	0.0313	⋯	⋯	⋯	0.0043
3/1/2022	0.301	0.267	1.17	0.0347	⋯	⋯	⋯	0.0039
3/1/2022	0.295	0.357	1.21	0.039	⋯	⋯	⋯	0.0034
⋮	⋮	⋮	⋮	⋮	⋯	⋯	⋯	⋮
⋮	⋮	⋮	⋮	⋮	⋯	⋯	⋯	⋮
⋮	⋮	⋮	⋮	⋮	⋯	⋯	⋯	⋮
30/3/2022	0.281	0.316	1.09	0.0385	⋯	⋯	⋯	0.0035

Table 3: Principal Component Table

Component Number	**Standard deviation**	**Proportion of variance**	**Cumulative Proportion**
1	1.5566	0.2203	0.2203
2	1.2418	0.1402	0.3605
3	1.1299	0.1161	0.4765
4	1.0697	0.1040	0.5805
5	1.01730	0.09408	0.67462
6	0.96378	0.08444	0.75907
7	0.88644	0.07143	0.83050
8	0.72805	0.04819	0.87869
9	0.68779	0.04301	0.92169
10	0.66420	0.04011	0.96180
11	0.6482	0.0382	1.0000

In table(3) it is Shown that each principal component has proportion of variance, andwe can say that 5 principal component from 11 principal component is good to use because 5 principal component together is represent about 67 percent of variance to all the data which is best principal component of all the eleven principal component , so we can it is shown that these 5 principal component is the best principal component

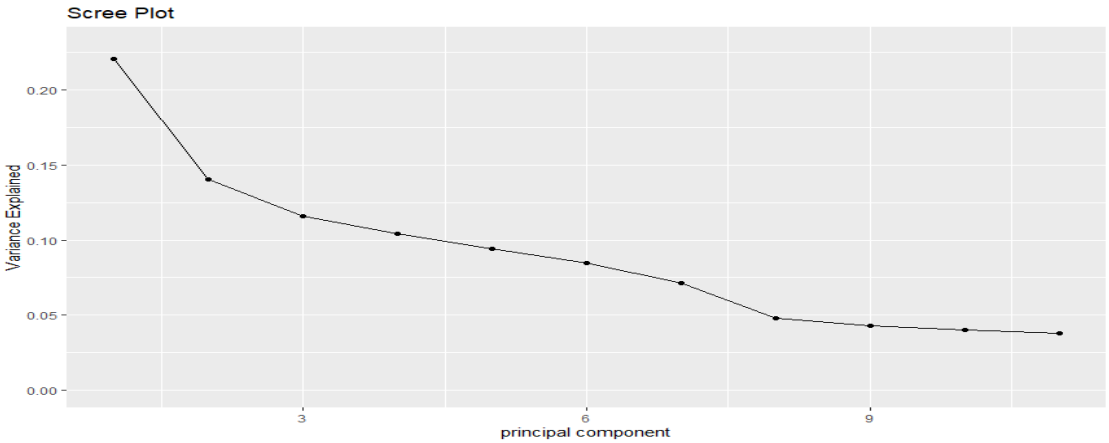

Figure 4: Scree Plot

In Figure (4), a Scree plot of the eigenvalue and a decreasing order of the percentage of variance explained by each component is shown.
In table (4) it shows the value of eigenvalue of each component in the principal component , and from table(4) we can find which component is more significance and better than the other component by using the eq(6),eq(8),eq(9) and data from table(1), and that by compare the value of eigenvalue of the different component in the same principal component or compare component in different principal component, and like that we can find the best and more significance component.

Table 4: Eigenvectors Table

	Component1	Component2	Component3	Component4	Component5
C%	0.408883	-0.0993626	-0.437013	-0.0202483	0.286033
Si%	-0.168723	0.346819	-0.250404	0.55044	0.0498693
Mn%	-0.201626	0.366136	0.438713	0.345959	0.150446
P%	0.147707	-0.185755	0.563905	-0.0686612	0.485272
S%	0.206358	-0.421512	0.135044	0.170354	0.0741282
Cr%	0.489184	0.141954	-0.204965	0.165749	-0.108225
Ni%	0.400782	0.0852444	0.20896	0.248565	-0.450941
Cu%	0.163317	-0.454088	0.175425	0.264759	-0.285164
Al%	0.0851444	0.227719	0.0350968	-0.595938	-0.135863
Mo%	0.327956	0.413618	0.315889	-0.157669	-0.249362
V%	0.397372	0.258962	-0.0125293	0.061426	0.524458

It was found that the first and second axes of the main component explained 22 and 14 percent of the total variance, respectively, when the variables were projected onto the factor axis. This was revealed when the variables were projected onto the factor axis. After doing so, we came to this realization after projecting the variables onto the factor axis. The finding of this was made possible as a direct consequence of projecting the variables onto the factor axis. The projection led to this discovery. Figure 5 displays the Pearson correlation coefficients in addition to some basic visuals that are associated with the correlations the relationships that already exist between the principal components and the ambient aspects. These relationships

are there because the primary factors are responsible for them. and the environmental elements are interrelated. There is a relationship between the basic components and the ambient aspects, which explains why these connections may be found. These graphical representations have some kind of relationship to the data that was presented in the figures that came before them. They are shown on a circle that, for the sake of this particular graphical depiction, has been dubbed "the correlation circle." [Circle] The axis of the circle is composed of the two axes that are used in the construction of the factor that is used in the construction of the factor. If the strength of the the relationship that can be shown to exist between a variable and the factor that has an effect on it is high, then the strength of the It is expected that there would be a high level of correlation between the related variable and the factor axis. This is due to the fact that what decides whether or not the strength of the degree to which the strength of the correlation that already exists between the variable and the factor is strong is the degree to which the correlation that already exists between the variable and the factor is already existent. As a result of this, it is feasible to determine the variables that are connected to a certain factor; doing so provides information on the variables that are capable of explaining the aforementioned factor. As a consequence of this, Additionally, it is feasible to determine the variables that are connected to a certain factor. Each and every one of the values for nickel(Ni), molybdenum (Mo), vanadium (V), chromium (Cr), and aluminum (Al) go up whenever the value of the value of the first significant component goes down. This continues to be the case even if the values of these components are reduced. In addition, the value of these components grows as a direct consequence of an increase in the value of the first principal component. It is possible that the existence of the first principal component might most credibly explain the wide variation in the concentrations of manganese (Mn) and silicon (Si). These values go up in a way that is exactly proportional to the degree to which the first main component is lessened. The degree to which it is diminished determines the magnitude of the increase. The values of Copper (Cu), Sulphur (S), Phosphorus (P), and Carbon (C) rise in parallel with the expansion of the second principal component. The second principle component may be the best candidate for describing the variability of the elements Copper (Cu), Sulphur (S), and Phosphorus (P); the values of these elements rise in parallel with the expansion of the second principal component.

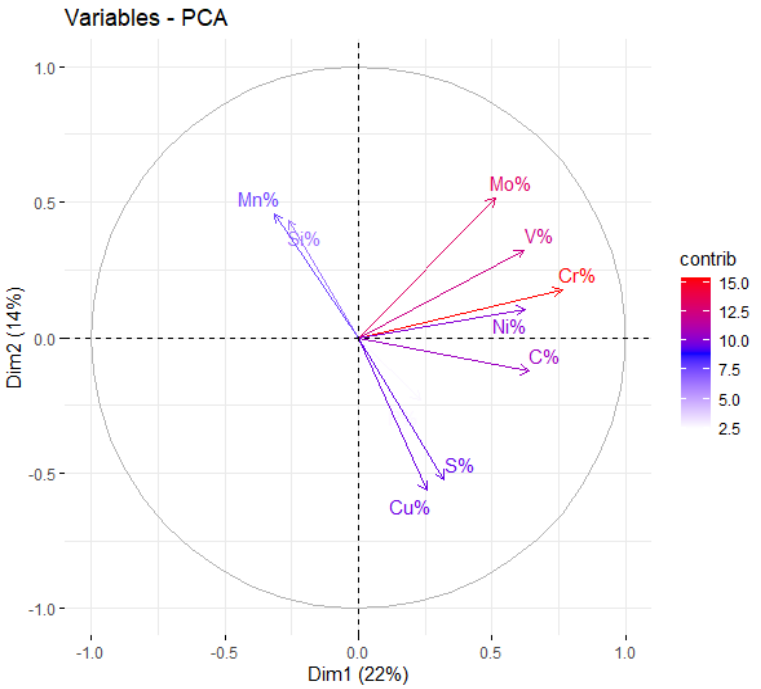

Figure 5: the correlation circle of first and second principal component

.

4. CONCLUSION

We chose to apply principal component analysis (PCA), which is a multivariate approach, in order to carry out an investigation of the steel components that were put to use at Sulai. In addition to this, we made it a top priority to maintain the natural variation that was already a part of the data set to the greatest extent possible by utilizing the technique of principal component analysis, which enabled us to do so. We did this in order to ensure that the data set was as accurate as possible. Because the five principal components collectively account for approximately sixty-seven percent of the variance in all of the data, we came to the conclusion that five of the eleven principal components are superior to the other six principal components. This led us to the conclusion that five of the eleven principal components are superior to the other six principal components. As a result of this, we came to the realization that five of the eleven principal components had advantages over the others. This conclusion may be reached as a result of the fact that the five principal components are responsible for approximately sixty-seven percent of the variation that is present in all of the data. Consequently, this conclusion led to the result that the variation is present in all of the data. Because of this, we got to the insight that five of the eleven main components are superior than the other six major components. The other two principal components, on the other hand, are inferior. Therefore, we demonstrated that these five principal components are the more effect than others principal components, and we arrived at the conclusion from the primary and the second principal components that the first component has the highest variation value include the following materials Chromium (Cr), Nickel (Ni), carbon(C) and Vanadium (V), the second component has the second highest variation value include the following materials Molybdenum (Mo) ,Manganese (Mn) and Silicon (Si), as a result of the information that has been supplied to us.

REFERENCES

[1] T.T. Allen, 2010,"Introduction to Engineering Statistics " and Lean Sigma Statistical Quality Control and Design of Experiments and Systems Second Edition.
[2] D.C. MONTGOMERY 'Introduction to Statistical Quality Control" , seven editions , Arizona State University Printed in the United States of America. 10 9 8 7 6 5 4 3 2 1 ,2013
[3] T. Cleff,Applied Statistics and Multivariate Data Analysis for Business and Economics,Springer, 2019.
[4] P. Sanguansat, Principal component analysis - multidisciplinary applications, Rijeka: InTech, 2012.
[5] I. T. Jolliffe, Principal component analysis,Springer, 2002.
[6] R. Vidal, Y. Ma, and S. Sastry, Generalized Principal Component Analysis,New York, NY: Springer, 2016.
[7] P. Sanguansat, Principal Component Analysis, Intech, 2012.
[8] B. F. J. Manly and J. A. N. Alberto, Multivariate Statistical Methods,Chapman & Hall/CRC, 2017.
[9] R. Johnson and D. Wichern, Applied Multivariate Statistical Analysis,New Jersey: Pearson, 2014.
[10] A. C. Rencher, Methods of multivariate analysis,John Wiley & Sons, Incorporated, 2002.