Sigma J Eng & Nat Sci 36 (4), 2018, 1023-1032

Sigma Journal of Engineering and Natural Sciences
Sigma Mühendislik ve Fen Bilimleri Dergisi

Research Article

ANALYSIS OF THE VIABILITY OF APPLYING THE PRINCIPAL COMPONENTS TECHNIQUE IN MULTIVARIATE DATA FROM TRAFFIC ACCIDENTS

Maria Lígia CHUERUBIM*[1], Irineu Da SILVA[2]

[1]*Faculty of Civil Engineering, Federal University of Uberlândia, Bloco 1Y, Campus, BRAZIL;*
ORCID: 0000-0002-2019-9198
[2]*School of Engineering of São Carlos, Dept. of Transportation Eng., BRAZIL;* ORCID: 0000-0001-5775-6683

Received: 02.07.2018 Revised: 12.10.2018 Accepted: 09.11.2018

ABSTRACT

The risk factors associated with road accident are directly related to the characteristics of the roadway, the vehicle type and the behavior of the driver, among others. For that reason, such traffic elements are intensively investigated and analysed in the field of road safety. Among the techniques and methods developed, those based on statistical analysis have demonstrated a high degree of susceptibility to the problem and have been applied in several studies of traffic accidents. In this perspective, this work presents a theoretical and applied discussion of the technique of Principal Component Analysis (PCA), in the study of road accidents. The main objective is to contribute to the discussion and theoretical foundation of the statistical techniques, used in the multivariate analysis of highway databases, generated by roadway concessionaries. The database used for this study is from the Dom Pedro I Highway, located in the urban area of the Campinas city in Brazil, during the period of four years from 2009 to 2012.
Keywords: Principal Component Analysis (PCA), roadway database, traffic accident.

1. INTRODUCTION

The Principal Component Analysis (PCA) is a multivariate statistical technique based on factor analysis, which is a statistical method used to describe the variability among correlated variables, minimizing possible redundancies and enabling the findings of the principal components of a dataset. The technique was first described by Pearson, in 1901, and later implemented by Hotelling, in 1936. In short, it can be understood as a data reduction method, in which the maximum variance of a dataset can be explained by the classification of eigenvectors associated with the largest eigenvalues of the correlation matrix, allowing the original dataset to be analysed from a small number of independent and orthogonal components.

The application of the PCA technique starts with the computation of eigenvalues and its corresponding eigenvectors, in function of the covariance matrix or the correlation matrix of the variables. An eigenvector is a direction, while an eigenvalue is a number, which expresses how much is the variance in the dataset in that direction. The eigenvector with the highest eigenvalue

* Corresponding Author: e-mail: marialigia@ufu.br, tel: +16997071141

is therefore the principal component. The number of eigenvectors and eigenvalues that exist in a dataset are equal to the number of dimensions the dataset has. The eigenvalues represent how much a factor explains a variable in factor analysis, what can be understood as the correlation between the original variables and the factors, and the key to understand the nature of a particular factor. They are used in deciding how many factors are to be extracted in the overall factor analysis.

According to Hongyu *et al.* (2015), PCA is a multivariate statistical technique that can linearly transform an original set of variables, initially auto correlated, in a substantially smaller set of uncorrelated variables that contains most part of the original dataset, called the principal components, thus reducing the size of the problem under analysis. Geometrically, it can be described as a rotation of existing points in a multidimensional space to express them in a two-dimensional space representing sufficient variability to indicate a pattern to be played.

The main properties of the principal components are that: each principal component is a result of linear combinations of all original variables; they are independent of each other and estimated in order to retain maximum information, in terms of the total variation contained in the dataset (Johnson & Wichern, 1999; Hongyu *et al.*, 2015).

A set of $"n"$ original variables through its linear combinations, generate $"n"$ principal components, whose main property, beyond orthogonality, is that they are obtained in descending order of maximum variance, i.e., the first principal component detains more statistical information than the second principal component, which has more statistical information than the third one and so on. Therefore, although the statistical information presented in the $"n"$ original variables is the same as that of the $"n"$ principal components, in general, it is common to get 90% of the information in no more than 2 or 3 of the first principal components.

The PCA technique also groups dataset on the basis of variation of its characteristics in a population, in other words, based on the similarities, what allows it to be applied in several areas of knowledge such as agronomy, biotechnology, ecology, biology, psychology, medicine, forestry, transportation, among others (Hongyu *et al.*, 2015).

In this context, the objective of this study is to provide a theoretical and applied discussion in the analysis and interpretation of multivariate dataset applied to traffic accidents, with the intention of contributing to the discussion of the choice and use of the PCA technique in analysis and interpretation of the results of such events.

2. THE MATHEMATICS OF PRINCIPAL COMPONENTS

Algebraically, the principal components are linear combinations of p original variables $X_1, X_2, \ldots, X_p$, scalars, with averages and variances. The principal components are determined based on the covariance matrix (S) or on the correlation matrix (R). In general, (R) and (S) are used when all of the original variables have similar scales since PCA is not invariant at scale. In practice, extracting principal components as eigenvectors of (R) is equivalent to calculating the principal components of the original variables, after each one is standardized in order to have unit variance. The standardization must be done when the units of measurement of the observed variables are not the same (Mardia *et al.*, 1979).

It is noteworthy, that rarely there is any correspondence among the principal components obtained from (S) or (R), even if there are considerable differences amid its standard deviations. Therefore, the user must know the characteristic of its variables in the database in order to define which of the two matrices shall be used, once that choice is not an arbitrary process.

The characteristic equation of the matrix (S) or (R) is as follows (Johnson & Wichern, 1999):

$$\det[R-\lambda I]=0 \text{ or } |R-\lambda I|=0 \tag{1}$$

Where (I) is the identity matrix and λ are the eigenvalues. The matrix (R) is described as:

$$R=\begin{bmatrix} 1 & r(X_1,X_2) & r(X_1,X_3) & \cdots & r(X_1,X_p) \\ r(X_2,X_1) & 1 & r(X_2,X_3) & \dots & r(X_2,X_p) \\ r(X_3,X_1) & r(X_3,X_2) & 1 & \dots & r(X_3,X_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(X_p,X_1) & r(X_p,X_2) & r(X_p,X_3) & \cdots & 1 \end{bmatrix} \tag{2}$$

Matrix (R) has full rank if its rank is equal to p, or in other words, if it presents the maximum number of rows or columns, linearly independent. Equation (1) will have p roots, called eigenvalues or characteristic roots of the matrix (R). Thus, considering $\lambda_1,\lambda_2,\lambda_3,\cdots,\lambda_p$ the roots of Equation (1), we have:

$$\lambda_1>\lambda_2>\lambda_3>\cdots>\lambda_p \tag{3}$$

For each eigenvalue λ_i there will be an eigenvector $\tilde{a}_l$ of the type:

$$\sum_{j=1}^{p} a_{ij}^2=1 \text{ with } \tilde{a}_l=\begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ip} \end{bmatrix} \tag{4}$$

The eigenvectors $\tilde{a}_l$ are standardized, i.e., the sum of the squares of the coefficients are equal to one and mutually orthogonal. For that reason, they have the following properties:

$$\sum_{j=1}^{p} a_{ij}\cdot a_{kj}=1 \quad \text{with} \quad (\tilde{a}_l^{'}\cdot\tilde{a}_l)=1, \tag{5}$$

with $\tilde{a}_l$ ranging from 1 to p and $\tilde{a}_l^{'}$ the transposed standardized eigenvector and

$$\sum_{j=1}^{p} a_{ij}\cdot a_{kj}=0 \text{ with } (\tilde{a}_l^{'}\cdot\tilde{a}_k)=0 \text{ for } i\neq k \tag{6}$$

where i and k ranging from 1 to p.

Thus, being $\tilde{a}_l$ the corresponding eigenvector of the eigenvalue λ_i, the $i-th$ principal component is obtained by the following linear combination:

$$Y_i=a_{i1}X_1+a_{i2}X_2+\cdots+a_{ip}X_p \tag{7}$$

or, $Y=AX$. (8)

with variance S_Y^2 maximized and conditioned to $\sum_{i,j=1}^{p} a_{ij}^2=1$.

Where,

$Y_1,Y_2,\cdots,Y_p$ = principal components not correlated

a_{ij} = coefficients of the linear combination the weight of the $j-th$ variable with the $i-th$ - principal component, represented in the matrix A of order p

Y_p = are designated as scores of the principal components

Geometrically, the linear combinations represent a selection of a new coordinate system, obtained by rotation of the original system, with axis $X_1, X_2, \ldots, X_p$. The new axis $Y_1, Y_2, \cdots, Y_p$ represent the directions with maximum variability, allowing a simpler interpretation of the structure of the covariance matrix. As a result, the principal components possess the following properties (Tran, 2008; Johnson & Wichern, 1999):

The variance of the principal component Y_i is equal to eigenvalue λ_i:

$$\hat{\sigma}(Y_i) = \lambda_i \tag{9}$$

The first component is the one with the largest variance among the possible combinations, and so on:

$$\hat{\sigma}(Y_1) > \hat{\sigma}(Y_2) > \cdots > \hat{\sigma}(Y_p) > \tag{10}$$

The total variance of the original variables is equal to the sum of the eigenvalues, that is equal to the total variance of the principal components:

$$\sum \hat{\sigma}(X_i) > \sum \lambda_i > \cdots > \sum \hat{\sigma}(Y_i) \tag{11}$$

The principal components are uncorrelated to each other:

$$\hat{C}orr(Y_i, Y_j) = 0 \tag{12}$$

2.1. Contribution of Each Principal Component and Number of Components to Retain

The contribution C_i of each principal component Y_i is obtained by dividing the variance of Y_i by total variance, expressed in percentage. The value of C_i makes it possible to evaluate how many components are required to perform the analysis, ensuring the maximum variance among the variables. In most of the application areas, the number of components used are the ones that accumulate 70% or more of the total variance, as indicated below (Solanas *et al.*, 2011):

$$\frac{\hat{\sigma}(Y_1) + \cdots + \hat{\sigma}(Y_p)}{\sum_{i=1}^{p} \hat{\sigma}(Y_i)} \cdot 100 \geq 70\% \text{ where } \; k < p \tag{13}$$

where k is the sum of the first k eigenvalues, that is, it represents the proportion of information retained in the reduction of p to k dimensions.

2.2. Interpretation and Scores of Each Principal Component

This step is performed by identifying the degree of influence that each X_j has on the component Y_i, expressed by the correlation between each X_j and Y_i, that is being interpreted. The correlation between X_j and Y_1 is expressed by:

$$\hat{C}orr(X_j, Y_1) = r_{X_j} \cdot r_{Y_1} = a_{1j} \cdot \frac{\sqrt{\hat{\sigma}(Y_1)}}{\sqrt{\hat{\sigma}(X_j)}} = \sqrt{\lambda_1} \cdot \frac{a_{1j}}{\sqrt{\hat{\sigma}(X_j)}} \tag{14}$$

In order to analyse the influence of $X_1, X_2, \ldots, X_p$ over Y_1, it is necessary to check the weight or loading of each variable over the component Y_1, calculated as follows:

$$w_1 = \frac{a_{11}}{\sqrt{\hat{\sigma}(X_1)}}, w_2 = \frac{a_{12}}{\sqrt{\hat{\sigma}(X_2)}}, \cdots, w_p = \frac{a_{1n}}{\sqrt{\hat{\sigma}(X_p)}}, \tag{15}$$

where w_1 is the weight of X_1 (Johnson & Wichern, 1999).

If the objective of the analysis is to obtain indexes, the process ends in the Equation (14). However, if the objective is to compare or group the individuals, the analysis extends to the calculation of scores (values) of each principal component. After reducing p to k dimensions, the k principal components will be the new individuals and all analysis are made using the scores of these components (Johnson & Wichern, 1999).

3. MATERIALS AND METHODS

The database used in this article was obtained from the Rota das Bandeiras Concessionaire. It describes the history of accidents in the urban section of the Campinas city, Brazil, along the kilometre 125 to 145 of Dom Pedro I Highway (SP-065). It considers the period of four years from 2009 to 2012, when the planned infrastructure improvements were not yet implemented. The Dom Pedro I Highway has an extension of 145.5 km and it connects the Paraíba Vale to the metropolitan region of Campinas. Its layout begins at the city of Jacareí, at the intersection with the Carvalho Pinto Highway (SP-070), and finishes at the intersection with the Anhanguera Highway. All statistical analysis of this article was made by computational routines implemented in the software SPSS - Statistical Package for Social Science for Windows.

In order to illustrate the application of the PCA technique, a database of 14 variables has been studied. They are the: date of occurrences; type of data acquisition; time of the occurrence; mileage (meters); number of vehicles involved in the accidents; number of victims not informed; number of uninjured, mild, moderate, severe and fatal victims; and, latitude and longitude of each accident. The amount of data corresponds to 5,744 samples for each of the variables.

4. RESULTS AND DISCUSSION

Applying PCA technique resulted in to five principal components, explaining 72.35% of the total variance of the original database, slightly above the minimum in the literature which is 70% (Solanas *et al,* 2011). Thus, an acceptance threshold of 0.3 was defined for the commonalities which is the total number of variances (correlations) of each variable, i.e., which can be extracted from the original variable. The higher the commonalities, the greater will be the capacity of explanation of the variable by factor, which culminates in the discarding of the variables which present values below threshold.

Table 1 shows the values obtained for commonalities of each variable analysed. Note that the original database was reduced to nine variables.

Table 1. Commonalities.

Variables	Extraction
N° de Vehicles	0.715
Uninjured victims	0.699
Mild victims	0.829
Severe victims	0.554
Latitude	1.000
Longitude	1.000
Data of Occurrence	0.736
Fatal victims	0.541
Moderate victims	0.437

Table 1 also shows that all analysed variables have communality values higher than 0.3. The variables such as, serious, fatal and moderate victims were those that had the lowest communality values, which shows that these factors are not 100% suitable to explain the original variable. In contrast, the other variables presented values as expected, which indicate the possibility of extracting information from the principal components.

For studies of this type, it is also recommended to apply the Kaiser-Meyer-Olkin (KMO) and Bartlett´s Sphericity tests, which indicate the degree of susceptibility or adjustment of the database for factor analysis, i.e., whose confidence level can be expected from the database after applying PCA technique. The KMO method presents normalized values (in the range from 0 to 1.0) and shows the proportion of variation due to common factors. To interpret the results, values near to 1.0 indicate that the factor analysis method is perfectly suitable for this kind of database analysis. On the other hand, values lower than 0.6 (adopted as limen), indicate the inadequacy of the method. In this study, the result was 0.499, what indicate the inadequacy of using the PCA technique for this kind of analysis.

Bartlett´s sphericity test examines the null hypothesis, that an original correlation matrix is an identity matrix. If the variables are perfectly correlated, only one factor is sufficient. In this case, the correlation matrix came out the same as the identity matrix. Hence, the PCA technique cannot be applied, since it can only be applied if the null hypothesis is rejected.

In order to measure the overall relation between the variables, we first compute the determinant of the correlation matrix - $|R|$. Under null hypothesis H_0, $|R|=1$; if the variables are highly correlated, $|R| \approx 0$. The test of significance is performed based on the following formula:

$$\chi^2 = -\left(n-1-\frac{2p+5}{6}\right)\cdot \ln|R| \tag{16}$$

Where,

n = number of events

p = number of variables

Under H_0, it follows a χ^2 distribution with a $[p\cdot(p-1)/2]$ degree of freedom. We reject the null hypothesis at a level of significance of 5% if P-value (probability) is less than 0.05. In this case, we can perform efficiently a PCA technique on the related database. In this paper, Bartlett´s test resulted in a value of χ^2 equal to 67,387.243 with 36 degrees of freedom, what gives Sig = 0.00, which indicates that the null hypothesis must be rejected.

Table 2 shows the total variance explained by the latent root criterion from 9 original variables, resulting in 5 components with eigenvalues greater than or equal to 1, that are considered significant. These components result in a cumulative variance of 72.35%, slightly above the recommended minimum of 70%, wherein, the first component showed the highest variance (22.542%).

Table 2. Total variance explained.

	Initial Eigenvalues			Rotation Sum of Squared Loading		
Component	Total	% variance	% cumulative	Total	% variance	% cumulative
1	2.029	22.542	22.542	2.029	22.542	22.542
2	1.386	15.398	37.940	1.386	15.398	37.940
3	1.068	11.872	49.812	1.068	11.872	49.812
4	1.021	11.346	61.158	1.021	11.346	61.158
5	1.007	11.192	72.350	1.007	11.192	72.350
6	0.955	10.611	82.962			
7	0.933	10.366	93.328			
8	0.601	6.672	100.000			
9	4.812E-6	5.347E-5	100.000			

Another option to verify the usefulness of the PCA technique is through a screen plot, which allows to determine the appropriate number of principal components. Screen plot shows the eigenvalues versus the number of principal components. It always displays a downward curve. The point where the slope of curve is clearly levelling out (the elbow) indicates the number of the principal components that should be generated by the analysis.

The Screen Plot presented in Figure 1, shows first five components to be above the line of the eigenvalue one. In relation to the selection of the number of components as a function of the "elbow", one should select only the factors located before the "elbow" (Catell, 1966) and observe the value of the "elbow". If the eigenvalue corresponding to the "elbow" is high, it should be included in the analysis, otherwise it should be neglected (Catell, 1977).

Thus, the selection of the quantity of the principal components was based on the truncation criterion of Kaiser that considers the most significant eigenvalues the ones with values higher than 1.0. In this study, value of 0.4 was considered as the minimum acceptable value of the variable to be included as the principal component. The variables presented factor loading significantly higher, ranging between 0.50 and 0.90.

In Table 3 it is possible to verify which variables have more influence on each component.

Table 3. Principal Component matrix.

	Component				
Variables	1	2	3	4	5
N° of Vehicles	0.215	0.753	0.137	0.172	0.232
Uninjured victims	0.189	0.807	0.107	0.028	0.008
Mild victims	-0.033	-0.202	-0.260	0.443	0.723
Severe victims	-0.009	-0.182	0.565	0.170	0.416
Latitude	0.986	-0.162	-0.038	0.007	-0.022
Longitude	0.986	-0.163	-0.038	0.007	-0.022
Date of Occurrence	-0.027	-0.014	0.088	0.725	-0.449
Fatal victims	0.045	-0.096	0.614	-0.377	0.106
Moderate victims	-0.009	-0.177	0.514	0.313	-0.208

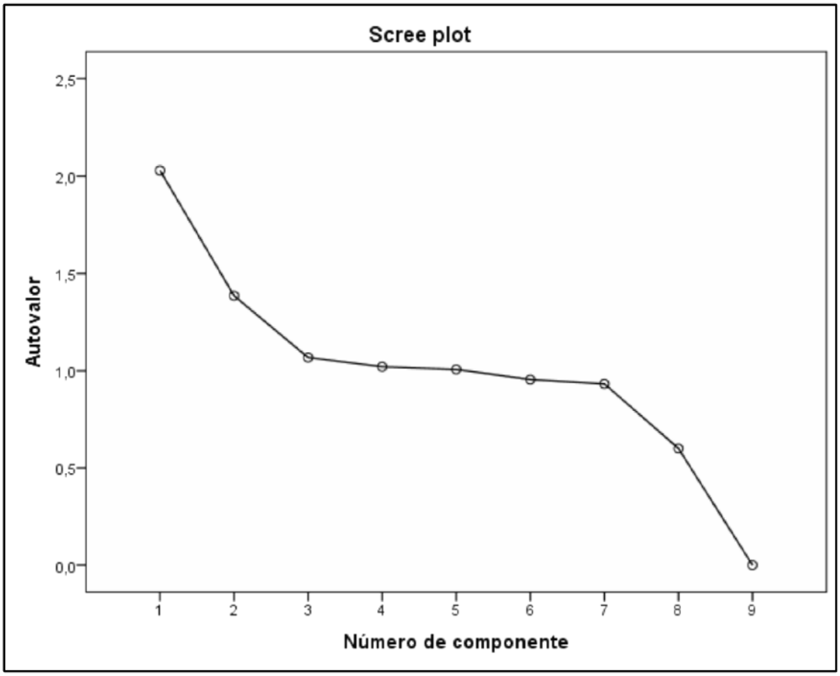

Figure 1. Scree Plot.

Considering the threshold of 0.40, the variables with the highest factor loadings in each component were:

C_1 - the latitude and longitude variables with factor loadings equal to 0.986;

C_2 - the variables such as, numbers of vehicles with factor loadings equal to 0.753 and, uninjured victims with factor loadings equal to (0.807);

C_3 - the variables such as: severe victims with factor loadings equal to 0.565, fatal victims with factor loadings equal to 0.614 and moderate victims with factor loadings equal to 0.514;

C_4 - the variable of date of occurrence with factor loadings equal to 0.725;

C_5 - the variable of mild victims with factor loadings equal to 0.723.

Table 4 presents the principal components extraction applying the axis rotation with Quartimax (a) and Varimax (b), both with normalization of Kaiser, and with the objective of minimizing possible redundancies between the variables represented by each component. In the rotation by the Quartimax method, the weights are elevated to a reduced number of components and near to zero for the remaining. In the Varimax method, some weights are significant and the rest are close to zero.

Table 4. Matrix of rotation.

Variables	Quartimax (a) Component					Varimax (b) Component				
	1	2	3	4	5	1	2	3	4	5
N° Vehicles	0.048	0.838	0.033	0.101	-0.002	0.048	0.838	0.033	0.099	-0.003
Uninjured victim	0.020	0.819	-0.060	-0.154	-0.022	0.020	0.819	-0.059	-0.155	-0.024
Mild victims	0.000	-0.050	-0.014	0.907	-0.053	0.001	-0.049	-0.014	0.907	-0.054
Severe victims	-0.014	0.014	0.665	0.311	0.120	-0.013	0.014	0.664	0.312	0.126
Latitude	0.999	0.039	0.008	-0.001	-0.003	0.999	0.039	0.008	-0.002	-0.003
Longitude	0.999	0.039	0.008	-0.001	-0.003	0.999	0.039	0.009	-0.002	-0.003
Data of Occurrence	-0.010	0.026	-0.179	0.022	0.839	-0.010	0.027	-0.187	0.023	0.837
Fatal victims	0.027	-0.020	0.668	-0.256	-0.168	0.027	-0.021	0.670	-0.256	-0.161
Moderate Victims	0.008	-0.075	0.383	-0.089	0.526	0.008	-0.075	0.378	-0.088	0.529

In the Quartimax rotation, Table 4 (a), an increase in the factor loadings is observed in C_1 for both the latitude and longitude variables, both equals to 0.999. For C_2, same behaviour was identified, where the variable number of vehicles is equal to 0.838 and uninjured victims is equal to 0.819. For C_3, the variable of moderated victims (0.383) presented a factor loading lower than expected and, therefore, it was deleted from C_3. Then, for C_3 only variables considered were the ones that had a considerable increase in its load factor, which are: severe victims (0.665) and fatal victims (0.668). For C_4, there was an increase in loading factor for the mild victims, but the same was not observed for the variable severe victims that reached the unsatisfactory index of 0.311. This variable was therefore excluded from this component. For C_5, the variables that had the highest factor loadings were the date of occurrence (0.839) and moderated victims (0.526).

In the Varimax rotation, Table 4 (b), the results are strictly similar to the Quartimax method, introducing slight variations to the factor loadings for the components that were excluded in the Quartimax method. The Varimax rotation grouped variables identically to the Quartimax rotation. After the application of both methods, one can conclude that the latitude and longitude variables can be grouped as C_1 and named as geographic location of the accident. Due to the high correlation shown by the variables such as number of vehicles and uninjured victims, the C_2 can be named as vehicular use and number of uninjured victims. By grouping the variables like: severe, fatal and moderated victims, for C_3, it can be named as degree of accidentally of the accidents. The C_4, can then be named as the number of uninjured victims and the component C_5 as the date of occurrence of the accident.

5. CONCLUSION

Based on the results obtained, the PCA technique showed effectiveness on the database reduction and multicollinearity elimination among the variables, contributing to the removal of five variables that showed a low variability or redundancy because they were correlated with more than one principal component. Thereby, results have shown that it is possible to work with fewer variables to explain the total variation of the database, and in this way optimizing the future analysis without significant loss of information.

One of the goals of the PCA use, in this case, was achieved since a relatively small number of components could be extracted (C_1 , C_2 , C_3 , C_4 and C_5) with the capability to explain the variability in the original database (72.35%). It is important to notice that the technique has not shown effectiveness in the adjustment of the database considering the KMO test, although it has been accepted by the Sphericity test, indicating an ambiguity on the analysed tests.

We made the PCA technique possible to be applied for the related database, without significant loss of information. It also allowed to group the variables in an array of components for categorizing data on statistical order of significance, what, of course, facilitates the study of this type of information.

It should be noted that the PCA technique is indicated for reducing the dimensionality of the database or for simplifying a problem, as long as the data are without any experimental or statistical delineation. Otherwise, other techniques should be applied, such as canonical analysis, track analysis, and correlation network analysis. In addition, the PCA technique can be applied for visualizing the structure and distribution of data in the two-dimensional or three-dimensional space.

REFERENCES

[1] Abdi, H.; Williams, L. J. *Principal component analysis*, Wiley Interdisciplinary Reviews: Computational Statistics, 2, 2010.]

[2] Cattell, R. B. *The data box: Its ordering of total resources in terms of possible relational systems*. In R. B. Cattell (Ed.), Handbook of multivariate experimental psychology (pp. 67–128). Chicago, IL: Rand-McNally, 1966.

[3] Cattell, R. B. *The grammar of science and the evolution of personality theory*. In R. B, 1977.]

[4] Hotelling, H. *Simplifield calculation of principal components*. Psychometrika, Williamsburg, v.1, p.27-35, 1936.]

[5] Hongyu, K.; Sandanielo, V. L. M.; Oliveira Junior, G. J. *Principal Component Analysis: theory, interpretations and applications*. E&S - Engineering and Science, Volume 1, Edição 5:1, 2015.]

[6] Johnson, R. A.; Wichern, D. W. *Applied multivariate statistical analysis*. 4th ed. Upper Saddle River, New Jersey: Prentice-Hall, 1999, 815 p.]

[7] Mardia, K. V.; Kent, J. T.; Bibby, J. M. 1979. *Multivariate Analysis*. New York: Academic Press.

[8] Pearson, K. 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* [S.l.: s.n.] 2 (6): 559–572.

[9] Solanas, A.; Manolov, R.; Leiva, D. Retaining principal components for discrete variables. *Anuario de Psicología, Facultat de Psicologia*, Universitat de Barcelona, vol. 41, nº 1-3, 33-50, 2011.

[10] Wilks, D. S. *Statistical Methods in the Atmospheric Sciences*, 2nd Ed. v. 91 p. 463-507, 2006.