Chapter 6

Principal Component Analysis: An Overview and Applications in Multivariate Engineering Problems

Chapter details

Chapter DOI:

https://doi.org/10.4322/978-65-86503-88-3.c06

Chapter suggested citation / reference style:

Almeida, Fabricio A., et al. (2022). "Principal Component Analysis: An Overview and Applications in Multivariate Engineering Problems". *In* Jorge, Ariosto B., et al. (Eds.) *Uncertainty Modeling: Fundamental Concepts and Models*, Vol. III, UnB, Brasilia, DF, Brazil, pp. 172–194. Book series in Discrete Models, Inverse Methods, & Uncertainty Modeling in Structural Integrity.

P.S.: DOI may be included at the end of citation, for completeness.

Book details

Book: Uncertainty Modeling: Fundamental Concepts and Models

Edited by: Jorge, Ariosto B., Anflor, Carla T. M., Gomes, Guilherme F., & Carneiro, Sergio H. S.

Volume III of Book Series in:

Discrete Models, Inverse Methods, & Uncertainty Modeling in Structural Integrity

Published by: UnB City: Brasilia, DF, Brazil Year: 2022

DOI: https://doi.org/10.4322/978-65-86503-88-3

Principal Component Analysis: An Overview and Applications in Multivariate Engineering Problems

Fabricio A. Almeida[1*], Guilherme F. Gomes[2], Pedro P. Balestrassi[3], Gabriela Belinato[4], and Pedro A. R. C. Rosa[5]

[1]Institute of Electrical Systems and Energy, Federal University of Itajubá (UNIFEI), Itajubá, Brazil – fabricio-almeida@unifei.edu.br
[2]Mechanical Engineering Institute, Federal University of Itajubá (UNIFEI), Itajubá, Brazil – guilhermefergom@unifei.edu.br
[3]Institute of Electrical Systems and Energy, Federal University of Itajubá (UNIFEI), Itajubá, Brazil – pedro@unifei.edu.br
[4]Federal Institute of South Minas Gerais (IFSULDEMINAS), Brazil – gabrielabelinato@ifsuldeminas.edu.br
[5]Department of Mechanical Engineering, Technician Superior Institute of Lisbon, University of Lisbon (ULisboa), Portugal – pedro.rosa@tecnico.ulisboa.pt

*Corresponding author

Abstract

This chapter presents an overview of principal component analysis (PCA), introducing and presenting the steps for using this powerful technique for data processing. In addition, three different examples are described, applying PCA to different multivariate engineering problems: in the manufacturing process using laser beam machining; power quality indices and in turbofan engine degradation data.

Keywords: *principal component analysis*; *multivariate statistics*; *laser beam machining; power quality indices; turbofan engine degradation*.

1 Contextualization

Most datasets usually have datas with multiple characteristics that can be analyzed. For example, in a standard machining process, which features characteristics such as average roughness (R_a), cutting tool wear and material removal rate (MRR). Such characteristics present a relationship with each other, which can be statistically verified through their variance-covariance structure. The need to understand the relationships between several variables of a correlated nature makes multivariate analysis an intrinsically complex subject (Johnson, R.A., Wichern, 2007).

When analyzing a set of characteristics, using univariate strategies (which deal with only one variable at a time) can bring unsatisfactory or even inadequate results. This can happen, because the multicollinearity existing in the set would be neglected (Almeida et al., 2020). Thus, it must be necessary to verify the relationship between the characteristics (which usually present correlation between them), requiring the use of multivariate strategies, promoting more informative and robust evaluations (Ferreira, 2018).

Among the commonly used strategies, principal component analysis (PCA) stands out, which was introduced by Pearson (Pearson, 1901) and later attributed differently by Hotelling (Hotelling, 1933). PCA is characterized as an exploratory multivariate technique that models correlated data from the variance-covariance structure (Ferreira, 2018). In addition, this technique allows the reduction of the dimensionality of the dataset (Gaudêncio et al., 2019; Jolliffe, 2010), finding a linear combination of uncorrelated variables that adequately explains the original variables, with the least possible loss of information (Mardia et al., 1995). In this way, the principal components can be obtained through a diagonalization, specifically, of defined semipositive symmetric matrices (Ferreira, 2018). The use of this technique can be found in many studies with different applications, such as: (Bounoua & Bakdi, 2021; Mahmoudi et al., 2021; Nhu et al., 2020; Song & Li, 2021; Yu et al., 2020).

Based on the previous discussion, this chapter will present an overview of the PCA strategy, indicating how the application should be carried out and interpreted, from the previous analysis of the data (before the application of the PCA). In addition, the steps to be considered by using this technique in the applied in datasets with multiple characteristics will be discussed. Finally, three different examples will be explored using problems in several areas of engineering, such as the manufacturing process, power quality indices and heath monitoring of aeronautical engine.

2 Principal Component Analysis

As previously inferred, PCA is characterized by being a multivariate technique widely used to interpret and reduce extensive and correlated data (Wang & Chien, 2010). Thus, the first step to consider before using the PCA is to verify the significance of the data correlation structure. This analysis can be done through correlation tests such as Pearson's, in which it is possible to verify (through the p-value of the test) if the data present a significant correlation. In addition, it is possible to verify whether the characteristics are directly or inversely correlated. Correlation between the characteristics can be verified according to Equation (1):

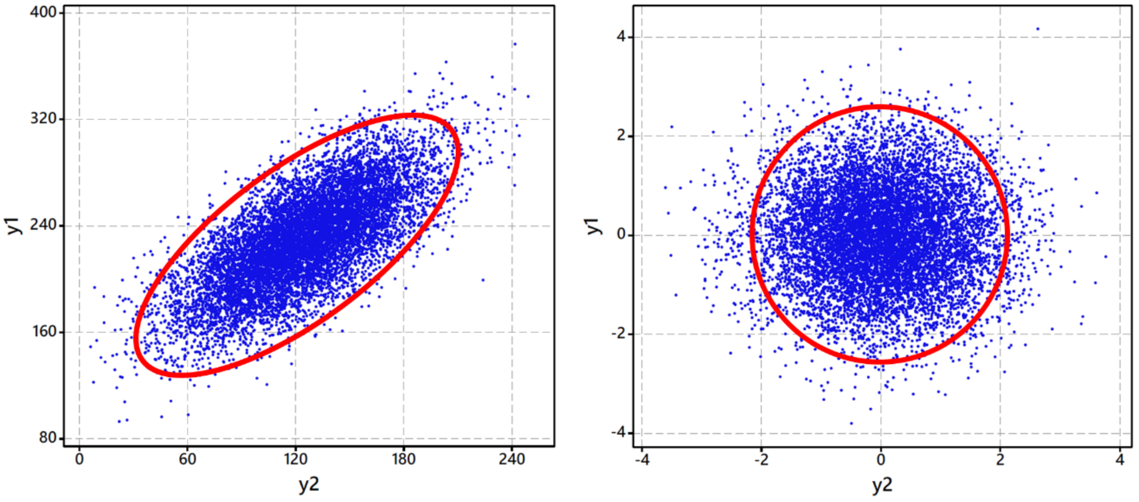
$$Cor_{y_i y_j} = \frac{CoVar_{y_i y_j}}{\sqrt{Var_{y_i} Var_{y_j}}} \qquad \forall i = 1, 2, ..., q; \quad j = 1, 2, ..., q \tag{1}$$

Where:

Var_{y_i} and Var_{y_j} are i^{th} and j^{th} variance;

$CoVar_{y_i y_j}$ represents the covariance between the characteristics.

Datasets with a significant level of correlation usually present an ellipsoidal geometric structure, while variables without correlation, that is, independent, present a spherical structure. Figure 1 illustrates both behaviors: no significant correlation and significant correlation level.



(a) (b)

Figure 1: Data behavior with correlation: (a) r = +0.705; (b) r = 0 (Almeida, 2021).

If the data set does not show significant correlation, there is no need to use a multivariate strategy. However, when dealing with variables from the same source, there is usually a significant correlation between the analyzed characteristics. In this way, one can proceed with the application of the PCA strategy.

It is known that PCA minimizes the dimensionality of the original variables, in order to absorb significant elements in the principal axis, while maintaining the error variation in the secondary axes. Consequently, PCA stands out for being very widespread in the literature to reduce the computational effort in analyzes involving large and correlated data sets. This strategy makes use of an orthogonal conversion to transform the observations into a set of variables that are not correlated with each other (Almeida, 2021). Thus, one of the necessary parameterizations for the application of the PCA is to define the ideal number of principal components (PC_i) that will be considered in the model. This choice is not arbitrary and can be defined through specific criteria.

One of the guidelines used when defining the amount of PC_i is the Kaiser criterion (Johnson, R.A., Wichern, 2007). This criterion indicates that: the more correlation there is between the variables, the smaller the number of components needed to represent the observations. Thus, the components need to explain at least 80% of the accumulated variance. For example, on a suitable dataset with seven correlated characteristics: if the first principal component (PC_1) has an explanation of 63.4% and the second principal component (PC_2) has a 19.9% explanation, it is known that two components are enough to adequately explain the original data set. Thus, there is a dimensionality reduction of 71.43% of the original dataset.

In addition to the percentage of explanation, another criterion commonly used to define the number of principal components is based on their eigenvalue. Thus, if the eigenvalue associated with the component is greater than or equal to 1 ($\lambda \geq 1$), this principal component will be associated in the model. Figure 2 shows the "Scree Plot" graph, which exemplifies the behavior of the eigenvalue in data from a welding process (adapted from (Almeida, 2017).

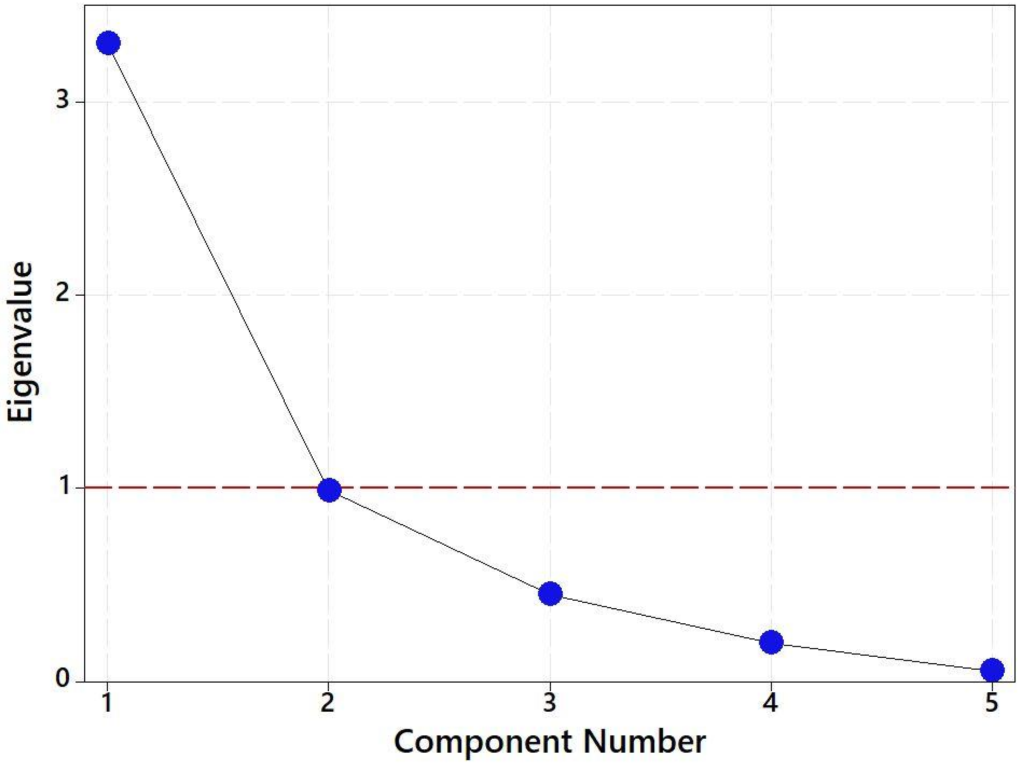
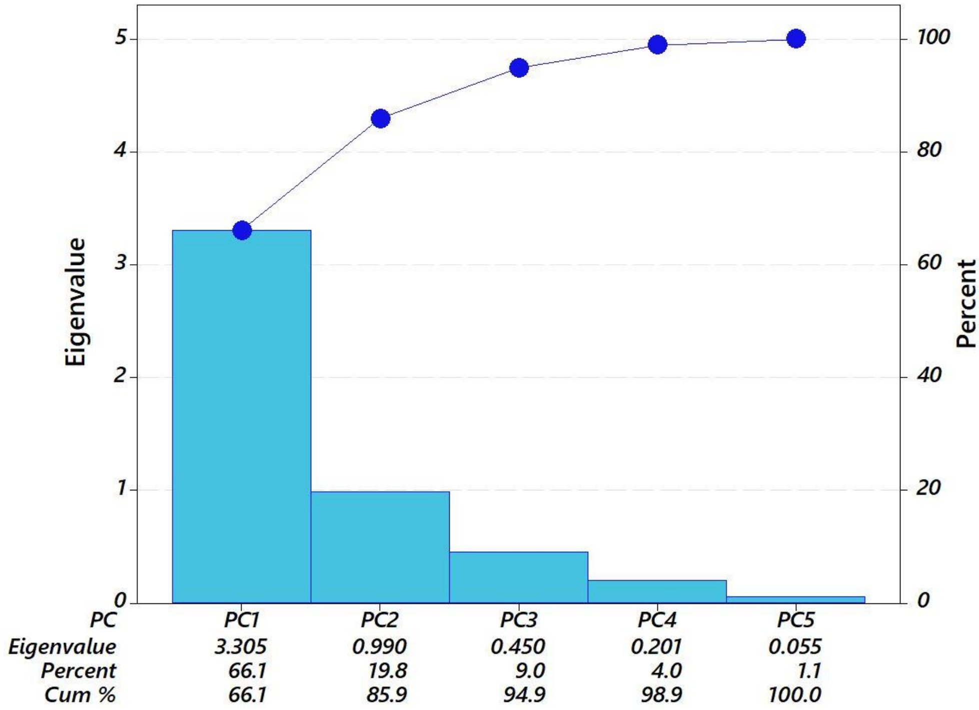


Figure 2: Scree plot

An alternative way to represent the behavior of both criteria (to choose the number of principal components) is through the Pareto chart (Figure 3). This graph illustrates the study's eigenvalue along with the explanation percentage of each component. In addition, the percentage of cumulative explanation is also presented, favoring decision making.

(Intentionally left blank)



PC	PC1	PC2	PC3	PC4	PC5
Eigenvalue	3.305	0.990	0.450	0.201	0.055
Percent	66.1	19.8	9.0	4.0	1.1
Cum %	66.1	85.9	94.9	98.9	100.0

Figure 3: Pareto chart with eigenvalues and percentage of explanation

Since PCA is the combination of a linear set for q random variables $Y_1, Y_2, \ldots, Y_q$, it can be inferred that the coordinate system represents a new set of coordinates before its original rotation, where the new axes hold the greatest data variability (Johnson, R.A., Wichern, 2007). Figure 4 shows the geometric interpretation of the axes.

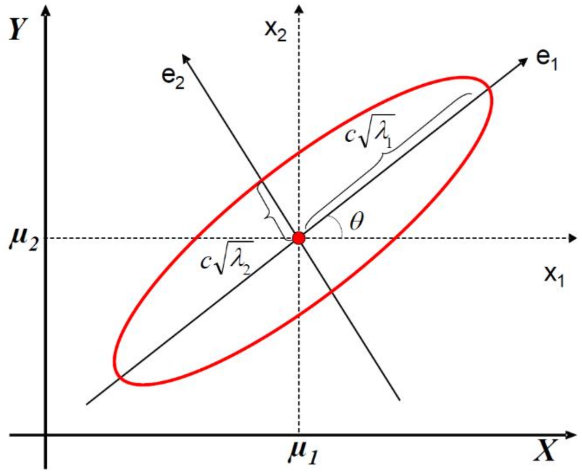


Figure 4. Constant density ellipsoid (Almeida et al., 2020).

PCA aims to find a combination of uncorrelated variables that adequately explains the original variables (Velasco et al., 2020). For this, we consider the random vector $\mathbf{X}^T = [X_1, X_2, \ldots, X_p]$ which has the covariance matrix $\mathbf{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p \geq 0$. Then, the linear combinations can be described as in Equation (2).

$$\begin{aligned}
Y_1 &= a_1^T X = a_{11}X_1 + a_{21}X_2 + \ldots + a_{p1}X_p \\
Y_2 &= a_2^T X = a_{12}X_1 + a_{22}X_2 + \ldots + a_{p2}X_p \\
&\cdots \\
Y_p &= a_p^T X = a_{1p}X_1 + a_{2p}X_2 + \ldots + a_{pp}X_p
\end{aligned} \tag{2}$$

Considering that Y_i for the i^{th} principal component, then, in Equation (3) and Equation (4):

$$Var(Y_i) = a_i^T \sum a_i = e_i^T \sum e_i; \qquad \forall i = 1, 2, \ldots, p \tag{3}$$

$$CoVar(Y_i, Y_k) = a_i^T \sum a_k = e_i^T \sum e_k; \qquad \forall i, k = 1, 2, \ldots, p \tag{4}$$

Thus, the principal components represent the uncorrelated linear combinations Y_1, Y_2, …, Y_p, in which the variances described in Equation (3) are the largest possible. That is, the i^{th} component can be defined from Equation (5), which was previously obtained through the formulation written in Equation (6) (Almeida, 2017).

$$PC_i = e_i^T Y = e_{1i}Y_1 + e_{2i}Y_2 + \ldots + e_{qi}Y_q \qquad i = 1, 2, \ldots, q \tag{5}$$

$$\begin{aligned}
&Max \qquad Var\left[e_i^T Y\right] \\
&subject\ to: \\
&\qquad e_i^T e_1 = 1 \\
&\qquad Cov\left[e_i^T Y, e_k^T Y\right] = 0 \\
&\qquad\quad k < i
\end{aligned} \tag{6}$$

Therefore, the original variables can be replaced by an uncorrelated linear set, i.e., the scores of the principal components. According to Johnson, R.A., Wichern, (2007), in possession of the standardized data matrix **Z** and the eigenvectors matrix **E** (from a multivariate set), the scores of the principal components can be obtained by Equation (7).

$$PC_{score} = \mathbf{Z}^T\mathbf{E} = \begin{bmatrix}
\left(\frac{y_{11}-\overline{y_1}}{\sqrt{s_{11}}}\right) & \left(\frac{y_{12}-\overline{y_2}}{\sqrt{s_{22}}}\right) & \cdots & \left(\frac{y_{1q}-\overline{y_q}}{\sqrt{s_{qq}}}\right) \\
\left(\frac{y_{21}-\overline{y_1}}{\sqrt{s_{11}}}\right) & \left(\frac{y_{22}-\overline{y_2}}{\sqrt{s_{22}}}\right) & \cdots & \left(\frac{y_{2q}-\overline{y_q}}{\sqrt{s_{qq}}}\right) \\
\vdots & \vdots & \ddots & \vdots \\
\left(\frac{y_{n1}-\overline{y_1}}{\sqrt{s_{11}}}\right) & \left(\frac{y_{n2}-\overline{y_2}}{\sqrt{s_{22}}}\right) & \cdots & \left(\frac{y_{nq}-\overline{y_q}}{\sqrt{s_{qq}}}\right)
\end{bmatrix}^T \times \begin{bmatrix}
e_{11} & e_{12} & \cdots & e_{1q} \\
e_{21} & e_{22} & \cdots & e_{2q} \\
\vdots & \vdots & \ddots & \vdots \\
e_{q1} & e_{q2} & \cdots & e_{qq}
\end{bmatrix} \tag{7}$$

Principal component scores are dimensionless, independent representations of an entire original dataset. These scores can be used in the most varied applications: from techniques such as artificial neural networks, cluster analysis to optimization techniques.

In the following topics, three real examples will be detailed using PCA on data from different multivariate processes in engineering. In such manner, applications of PCA will be addressed in data from a laser machining process (Belinato et al., 2019); in power quality sector (Almeida et al., 2021) and on data from NASA's Turbofan engine degradation analysis (Saxena & Goebel, 2008).

3 On the Use of PCA in Engineering Applications

3.1 PCA-based LBM process

The first example to be addressed will be based on the article by Belinato et al., (2019). In this study, the authors used the PCA technique with other statistical methods for experimentation and optimization in the machining process through laser beam machining (LBM). In general, this machining process has multiple objectives due to its quality characteristics. The authors performed a design of experiments (DOE) using the Response Surface Methodology (RSM) for the parameters of laser frequency (*f*), cut speed (*S*), laser power (*I*). The quality characteristics investigated were: material removal rate (MRR) and different roughness metrics (R_a, R_q, R_z, R_p and R_t). Table 1 and Table 2 present the parameter levels and the experimental matrix with the multiple quality characteristics analyzed, respectively.

The experiments were performed on a Deckel Maho Lasertec® machine model DML40SI (Figure 5a) and data collection used a Mahr® rugosimeter model M300 with an RD18 measuring device (Figure 5b). The workpieces used can be seen in Figure 6. All planning, experiment and data collection were carried out at Instituto Superior Técnico, University of Lisbon.

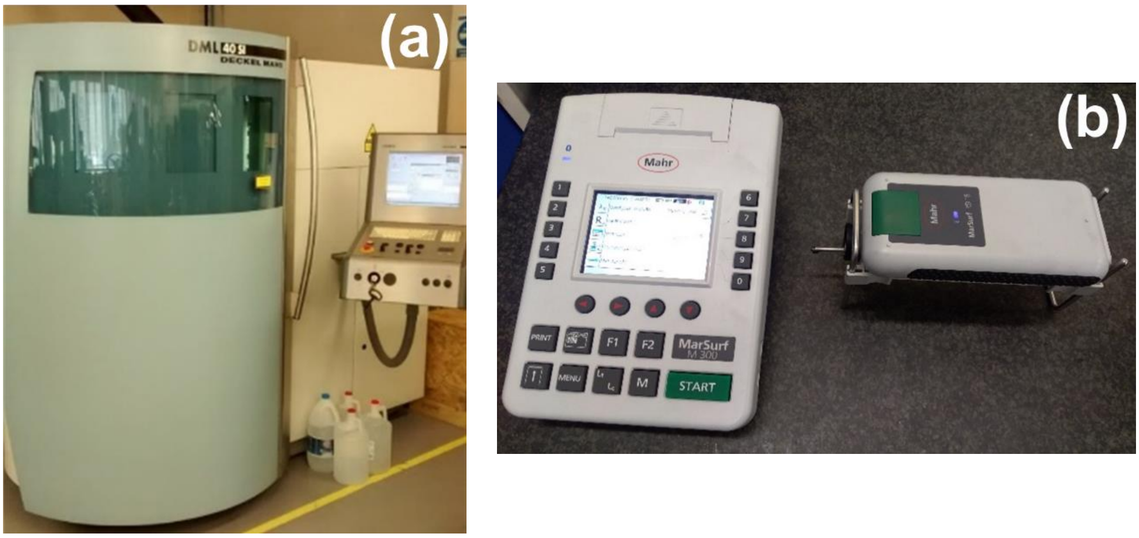


Figure 5: (a) DML40SI LBM machine and (b) Mahr® M300 rugosimeter (Belinato et al., 2019).

Table 1: Input parameters and levels (Belinato et al., 2019).

Input parameters	Level				
	-1.682	**-1**	**0**	**+1**	**+1.682**
f [kHz]	11.2	15	20.5	26	29.7
S [mm/min]	29.5	200	450	700	870.4
I [%]	26.3	40	60	80	93.6

Table 2: Experimental matrix (Belinato et al., 2019).

N	Setup			Responses					
	f	***S***	***I***	**R$_a$**	**R$_q$**	**R$_z$**	**R$_p$**	**R$_t$**	**MRR**
	[kHz]	[mm/min]	[%]	[µm]	[µm]	[µm]	[µm]	[µm]	[cm3/s]
1	15.0	200.0	40.0	4.54	5.85	31.30	14.46	39.58	5.95 x10^{-4}
2	26.0	200.0	40.0	2.12	2.65	12.77	6.70	14.00	4.20 x10^{-4}
3	15.0	700.0	40.0	7.27	9.03	42.06	22.59	47.90	7.81 x10^{-4}
4	26.0	700.0	40.0	3.68	4.53	17.47	8.82	18.86	4.21 x10^{-4}
5	15.0	200.0	80.0	12.38	15.07	66.30	33.68	78.49	1.84 x10^{-3}
6	26.0	200.0	80.0	5.28	6.91	35.33	20.30	45.70	6.57 x10^{-4}
7	15.0	700.0	80.0	11.82	14.66	63.70	31.40	79.66	2.05 x10^{-3}
8	26.0	700.0	80.0	6.08	7.25	36.26	17.15	52.88	2.33 x10^{-3}
9	11.2	450.0	60.0	11.93	14.93	61.52	33.55	80.09	1.77 x10^{-3}
10	29.7	450.0	60.0	3.14	4.71	14.96	8.04	17.93	1.21 x10^{-3}
11	20.5	29.55	60.0	11.91	15.91	69.30	33.67	95.14	2.24 x10^{-4}
12	20.5	870.45	60.0	3.07	4.14	26.24	12.64	30.49	1.24 x10^{-3}
13	20.5	450.0	26.36	3.64	4.36	17.57	8.68	22.42	3.07 x10^{-4}
14	20.5	450.0	93.64	10.73	12.92	71.64	32.47	89.21	2.70 x10^{-3}
15	20.5	450.0	60.0	6.22	7.62	36.61	17.18	47.00	1.54 x10^{-3}
16	20.5	450.0	60.0	5.90	7.38	36.71	17.81	42.50	1.57 x10^{-3}
17	20.5	450.0	60.0	6.17	7.51	36.35	16.91	38.37	1.56 x10^{-3}
18	20.5	450.0	60.0	6.23	7.71	36.14	18.10	41.29	1.56 x10^{-3}
19	20.5	450.0	60.0	5.88	7.55	36.52	17.95	44.76	1.58 x10^{-3}
20	20.5	450.0	60.0	6.17	7.60	35.60	17.33	38.25	1.60 x10^{-3}

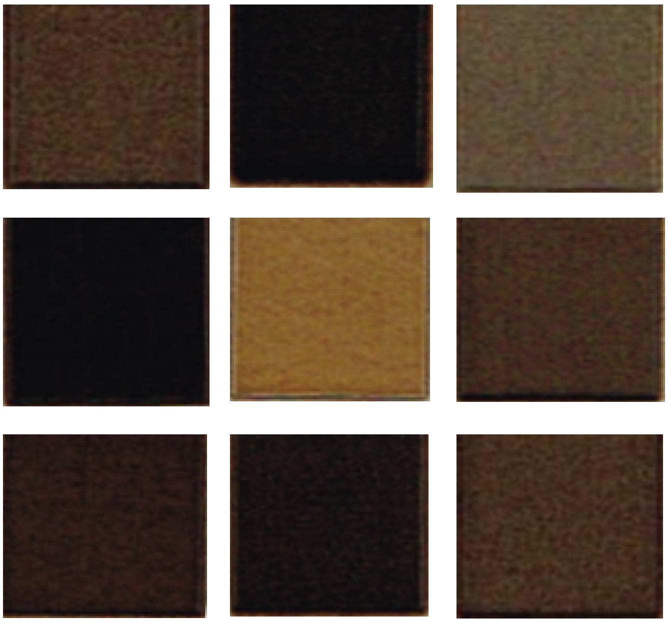


Figure 6: Machined surface workpieces (Belinato et al., 2019).

To emphasize principal components analysis, initially a correlation analysis is performed to verify the behavior of the quality characteristics. From Table 3, it appears that there is a significant correlation between the characteristics. This analysis can be verified using Pearson's correlation test and the relationship between the quality characteristics can also be verified through the correlation matrix illustrated in Figure 7. Then, information on eigenvalue and percentage of explanation is showed to define the number of components to be used in the analysis. Figure 8 presents the Pareto chart for the eigenvalues and cumulative percentage of explanation of the components. Through this analysis, it can be seen that the first component (PC_1) is the only one to present an eigenvalue greater than 1, in addition to explaining 86% of the original data. Thus, only 1 component is sufficient to represent all the quality characteristics analyzed.

Knowing the number of components needed (only 1), in this case, it is enough to extract the component scores, which represent the original dataset in a dimensionless way. From this result, it is possible to verify that the multivariate PCA technique provides a data dimensionality reduction by 83.33%. This result favors a less complex analysis with less computational effort for the next steps (such as optimization, forecasting etc). For more details about the process and the approach performed by the authors, see the study by Belinato et al., (2019).

Table 3: Correlation analysis for the LBM process characteristics (Belinato et al., 2019).

	Ra				

Rq	0.996(1)
	0.000(2)
		Rq			

Rz	0.973(1)
	0.000(2)
		0.969(1)
		0.000(2)
			Rz		

Rp	0.979(1)
	0.000(2)
		0.979(1)
		0.000(2)
			0.989(1)
			0.000(2)
				Rp	

Rt	0.960(1)
	0.000(2)
		0.962(1)
		0.000(2)
			0.987(1)
			0.000(2)
				0.976(1)
				0.000(2)
					Rt

MRR	0.460(1)
	0.041(2)
		0.415(1)
		0.069(2)
			0.500(1)
			0.025(2)
				0.454(1)
				0.044(2)
					0.462(1)
					0.040(2)

(1) Pearson correlation
(2) P-Value

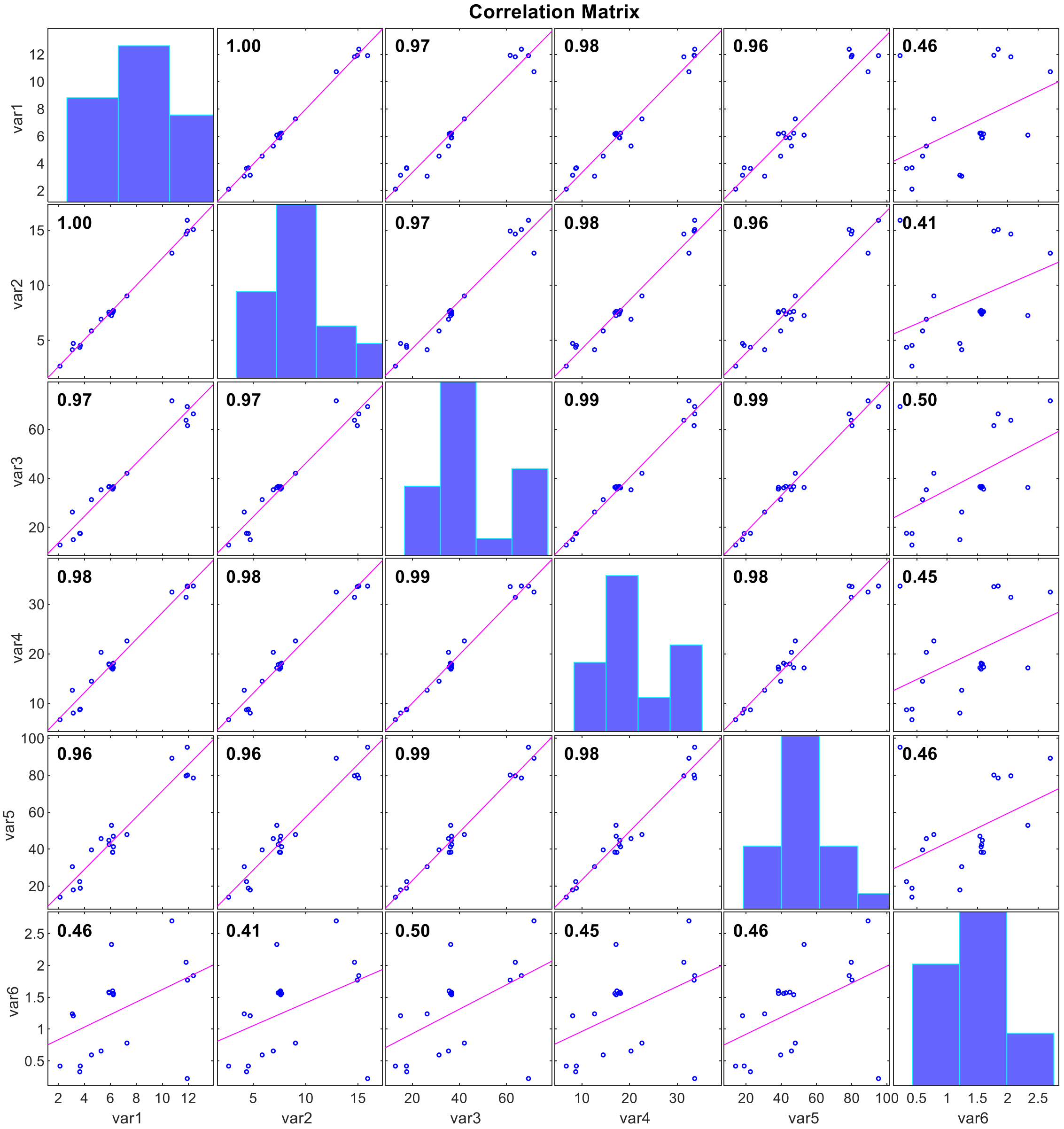


Figure 7: Correlation matrix from LBM process.

3.2 On the use of PCA in electric power substation quality indices

The second example to be addressed is derived from an analysis of power quality indices of electric power substations addressed by Almeida et al., (2021). In this study, the authors used data from 17 substations with 31 power quality characteristics. The substations are located in southeastern Brazil, corresponding to a total area of 41,241 km^2, about 90% of the state of Espírito Santo. Figure 9 illustrates, geographically, the location of the analyzed substations.

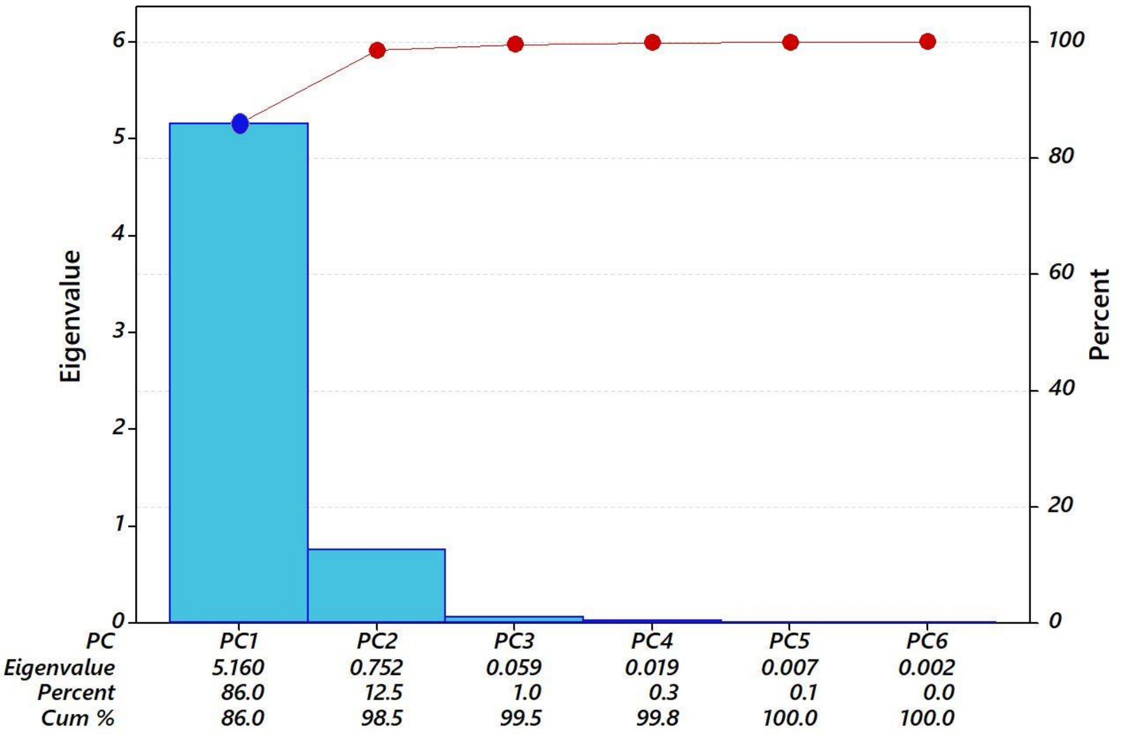


Figure 8: Pareto chart with eigenvalues and percentage of explanation (Belinato et al., 2019).

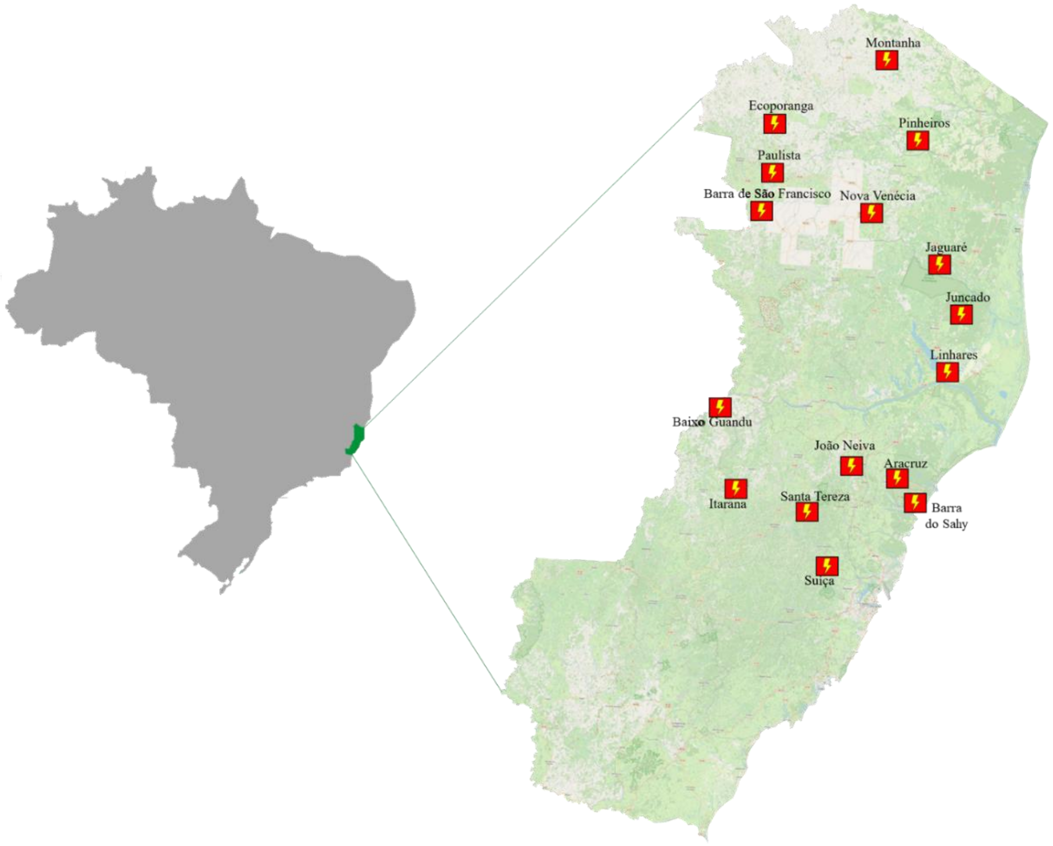


Figure 9: Location of the investigated substations in the State of Espírito Santo (Almeida et al., 2022).

Power quality measurements were collected over a year, in order to cover different seasonalities that influence the performance of the electricity distribution network (such as rain, winds, among other phenomena). In addition, 30 power quality monitors from Schweitzer Engineering Laboratories, model SEL 734, were used to acquire this data. The behavior of the dataset towards the substations is illustrated in Figure 10. The set were applied in a method to find the best clustering technique for this set and, consequently, classify the substations based on the power quality. Due to the correlated structure of the data (originally available in Miranda et al., (2016), the use of exploratory techniques such as PCA is recommended.

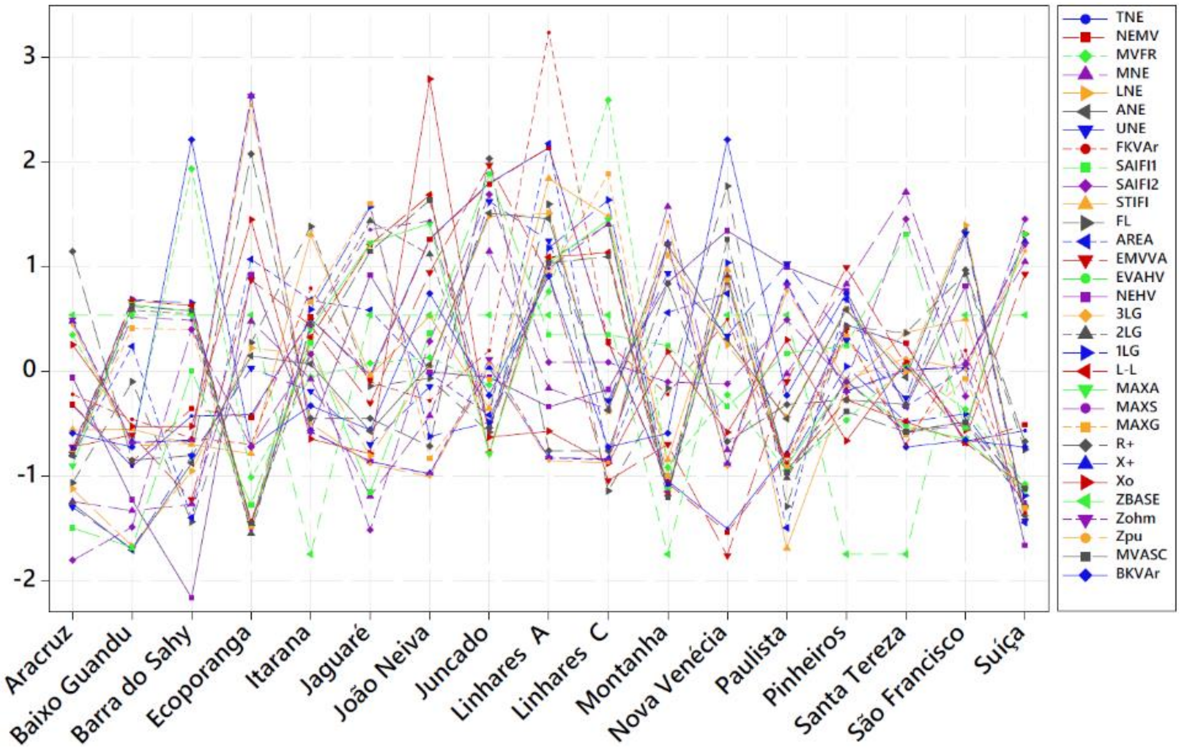


Figure 10: Relationship between power quality indices and substations (Almeida, 2021).

After test and confirm that data structure presents a significant correlation, the eigenvalues and the percentage of explanation of each component are calculated. This procedure is intended to determine the number of principal components needed for the study. As an alternative view of the first example (applied to the process by LBM), the behavior of the eigenvalues can also be visualized through the Scree plot (Figure 11). In this graph it is possible to verify that the first six components present eigenvalues greater than 1 (λ ≥ 1). Figure 12 illustrates the behavior of eigenvalues 1 and 2 of their respective principal components.

(Intentionally left blank)

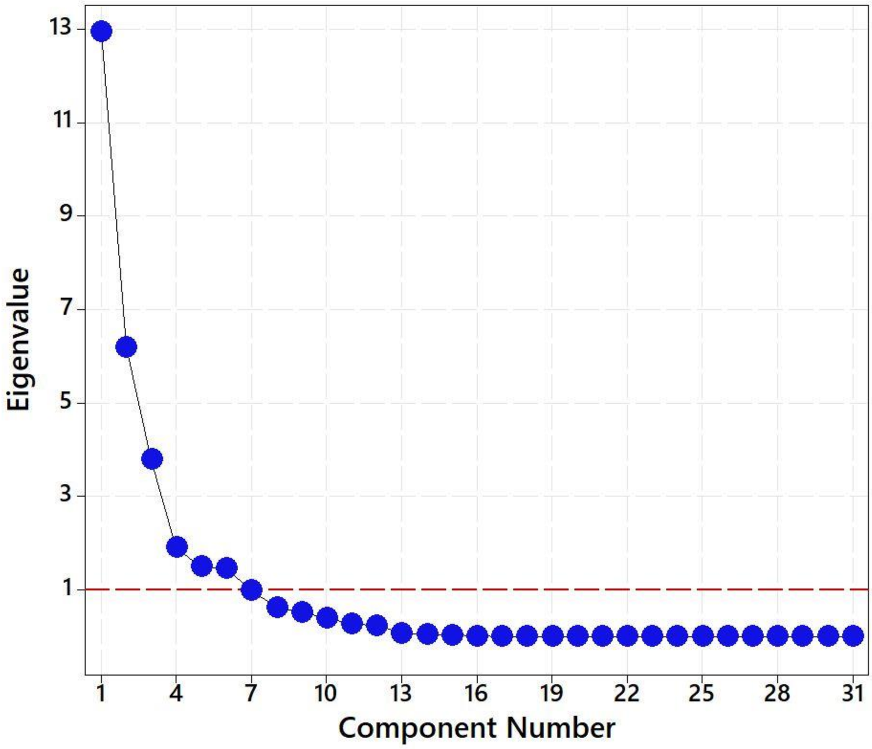


Figure 11: Scree plot of eigenvalues from power quality indices (adapted from Almeida, 2021).

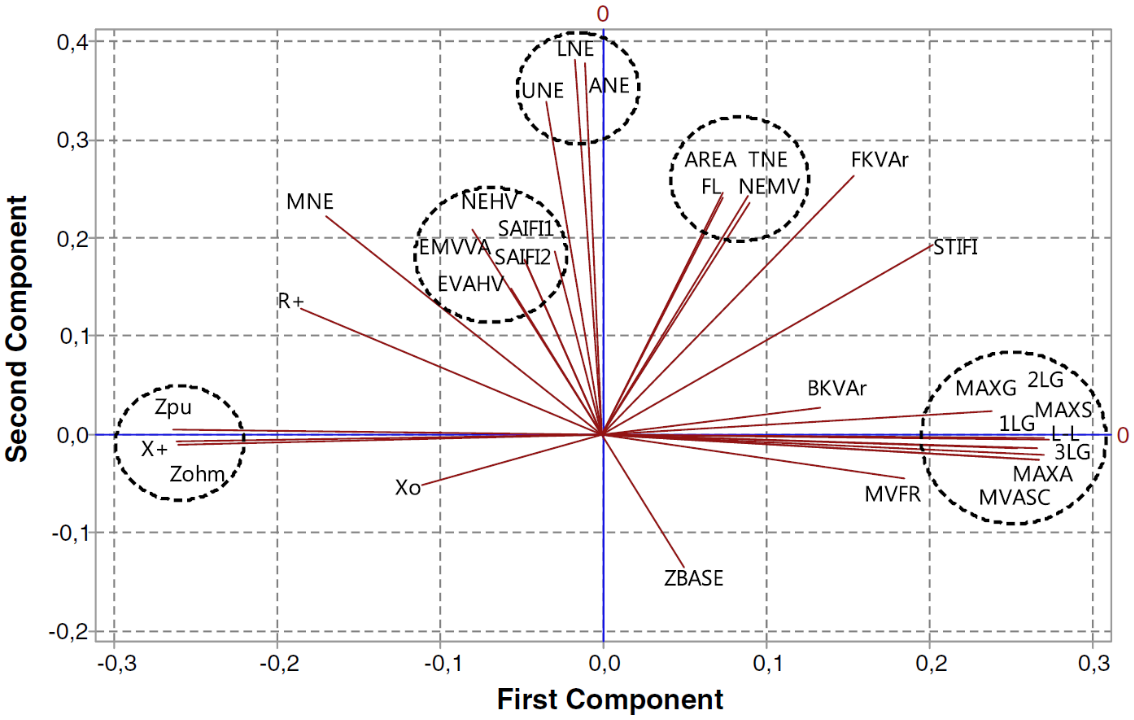
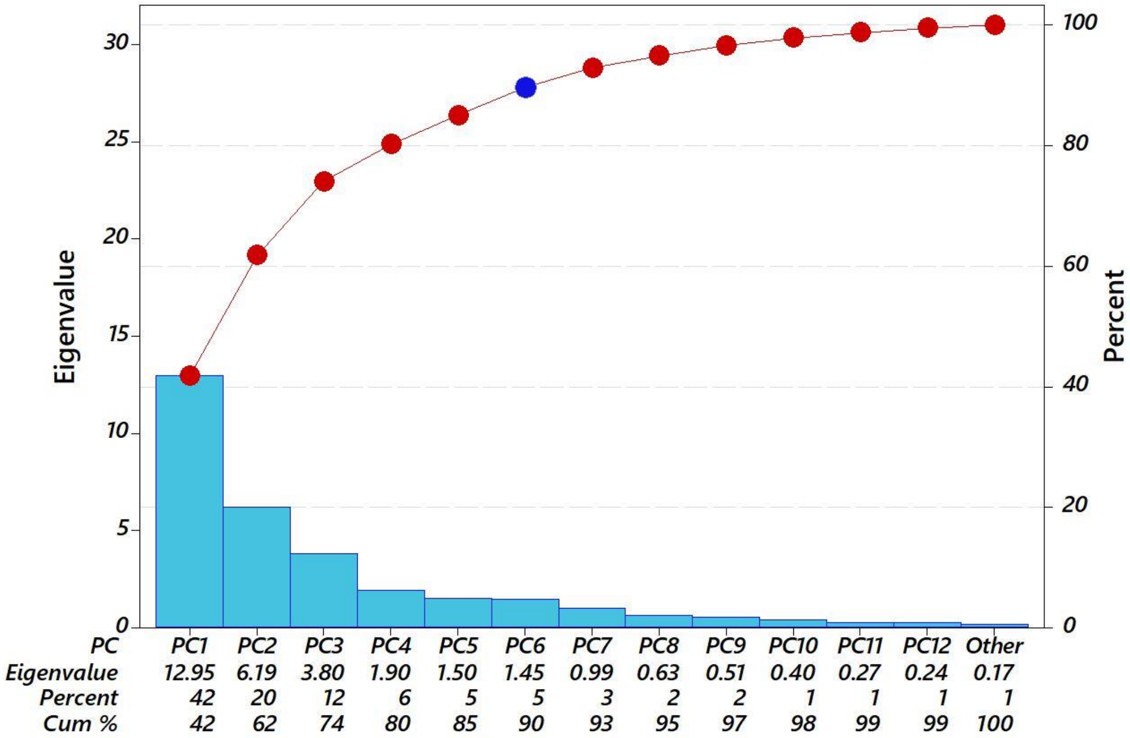


Figure 12: Plot of eigenvectors 1 and 2 (Miranda et al., 2016).

Complementarily, the Pareto chart (Figure 13) also provides information on the explanation percentage of each principal component. In this graph it is possible to verify that the fourth component presents a cumulative percentage of 80%. In this case, it is advisable to use the number of components necessary to respect the eigenvalue criterion. By using 6 components instead of 4, there data explanation contribution increases from 9.53%, in addition to properly respecting the Kaiser criterion discussed above.



PC	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	Other
Eigenvalue	12.95	6.19	3.80	1.90	1.50	1.45	0.99	0.63	0.51	0.40	0.27	0.24	0.17
Percent	42	20	12	6	5	5	3	2	2	1	1	1	1
Cum %	42	62	74	80	85	90	93	95	97	98	99	99	100

Figure 13: PCA Pareto chart for substation data.

There is a discussion in literature about the number of components to be used in exploratory analysis studies, both for principal components and factor analysis (another widely used exploratory technique). Some of these discussions can be verified in studies such as Visinescu & Evangelopoulos (2014) and Almeida (2021).

Considering the six principal components, the extraction of scores can be performed, thus generating 6 vectors of dimensionless and independent scores. These values adequately represent and explain the 31 original variables of the study. From this application, there is an 80.5% reduction in the data dimensionality. This result favors further analyzes that involve computational effort, as in the application of cluster algorithms used in the aforementioned power quality studies. The following studies bring more details and information about the object of study: (Almeida et al., 2022); (Almeida et al., 2021) and (Miranda et al., 2016).

3.3 On the use of PCA in Turbofan engine degradation data

The analysis and investigation of aeronautical engine data is widely investigated by industries and researchers (Chatterjee & Litt, 2003; Deng et al., 2020; Goebel et al., 2007; Kurosaki et al., 2004; Listou Ellefsen et al., 2019; Saxena et al., 2008; Xu et al., 2020), comprising the collection of numerous information generated by several different sensors in specific positions. The diagnosis of this magnitude generates a large amount of data that usually present a multivariate characteristic. Thus, the PCA strategy can also be explored in this context. Data referring to turbofan engine degradation predictions will be analyzed. This set can be found in the public archives of the National Aeronautics and Space Administration (NASA), referring to the study by Saxena et al., (2008) and available at Saxena and Goebel, (2008). The data were generated using C-MAPSS – Commercial Modular AeroPropulsion System Simulation software, based on the behavior of turbofan engines, shown in the diagram of Figure 14.

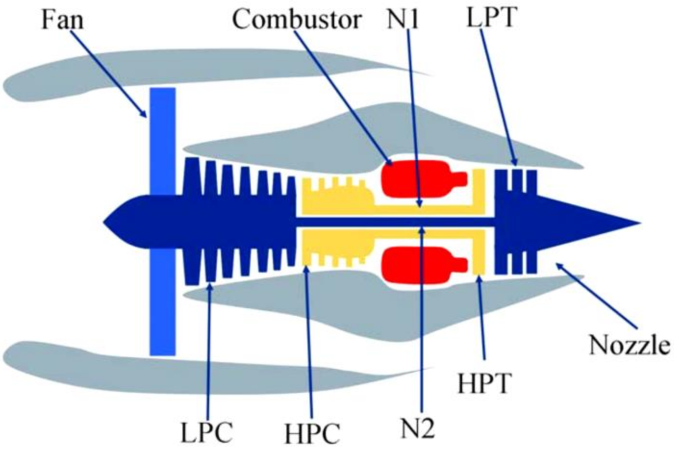


Figure 14: Diagram of engine simulated (Saxena et al., 2008).

For this example, a set will be selected that describes the simulation of 218 motors in constant execution until the moment of failure (between 127 and 356 cycles). The set presents information for analysis, using 21 C-MAPSS outputs with 33,991 collections for each characteristic (totaling 713,811 data). Table 4 describes the characteristics of the C-MAPSS outputs. Figure 15 illustrates the behavior of the sensors as a function of the engine's remaining useful life (RUL). The complete dataset and more details about the collection are available in the articles and database mentioned above.

Through correlation analysis it is possible to assume that there is a significant variance-covariance structure for the whole set, considering a confidence interval of 95%. Given the large amount of data, the correlation analysis will not be presented here, but it can be easily replicated with the help of any statistical software.

From the correlation level, the data are able to be used in the PCA strategy. Thus, the eigenvalues and contribution percentage of each component are initially verified to estimate the ideal amount for this data set. Through this initial analysis, it is possible to verify that only two principal components present $\lambda \geq 1$ (16,908 and 3,578, respectively). The behavior of the eigenvalues can be verified through Figure 16. In addition, the second component presents an accumulated explanation percentage of 97.55% of the data. Such results infer that only two components are sufficient to adequately represent all 21 quality characteristics collected through the sensors. The

Pareto chart illustrates the behavior of both results mentioned above, as shown in Figure 17.

Table 4: C-MAPSS outputs (Adapted from (Saxena et al., 2008).

Symbol	Description	Units
T2	Total temperature at fan inlet	°R
T24	Total temperature at LPC outlet	°R
T30	Total temperature at HPC outlet	°R
T50	Total temperature at LPT outlet	°R
P2	Pressure at fan inlet	psia
P15	Total pressure in bypass-duct	psia
P30	Total pressure at HPC outlet	psia
Nf	Physical fan speed	rpm
Nc	Physical core speed	rpm
epr	Engine pressure ratio (P50/P2)	--
Ps30	Static pressure at HPC outlet	psia
phi	Ratio of fuel flow to Ps30	pps/psi
NRf	Corrected fan speed	rpm
NRc	Corrected core speed	rpm
BPR	Bypass Ratio	--
farB	Burner fuel-air ratio	--
htBleed	Bleed Enthalpy	--
Nf_dmd	Demanded fan speed	rpm
PCNfR_dmd	Demanded corrected fan speed	rpm
W31	HPT coolant bleed	lbm/s
W32	LPT coolant bleed lbm/s	lbm/s

(Intentionally left blank)

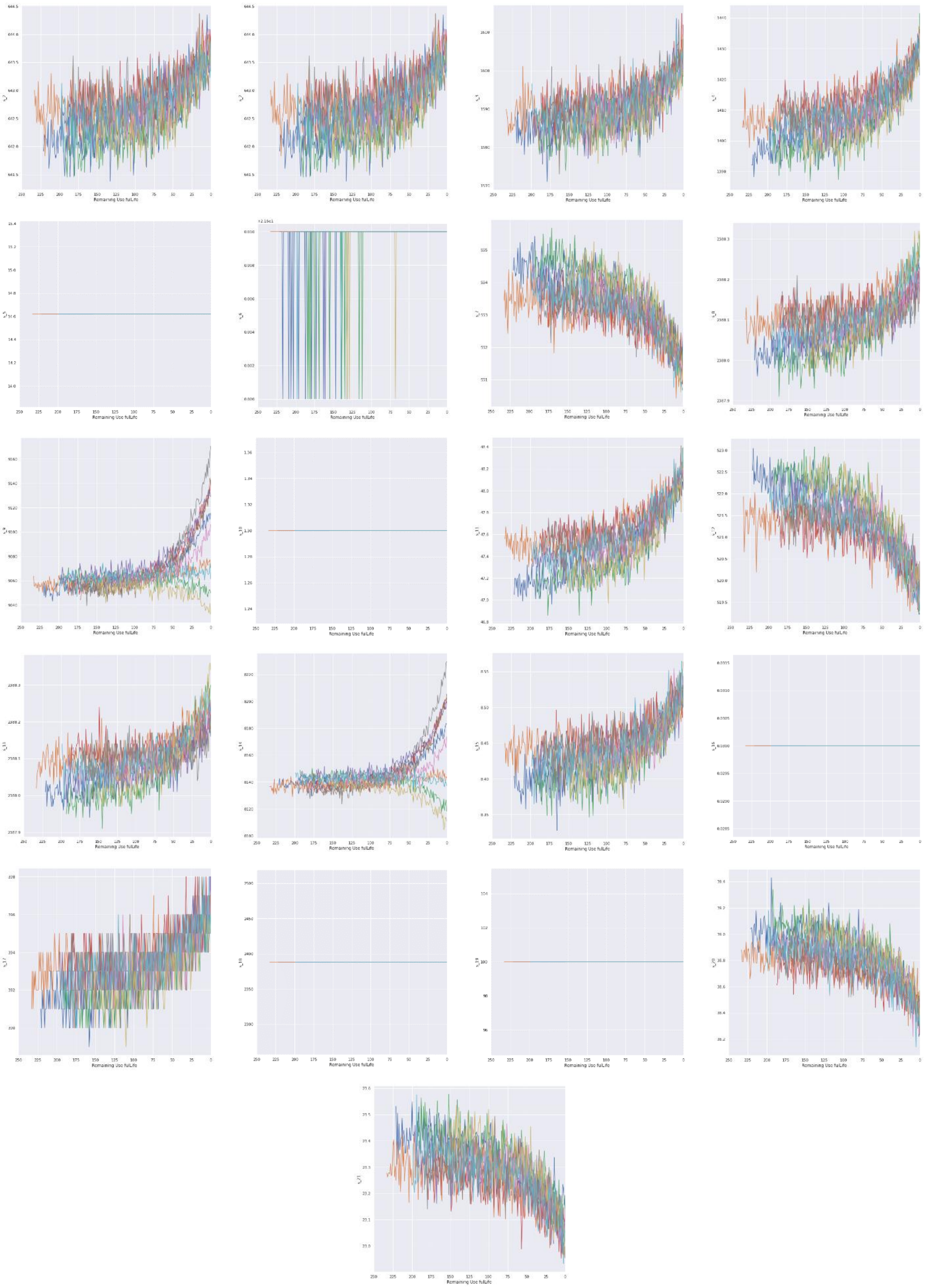


Figure 15: Sensors signal (#1 to #21) in function of the engine's RUL.

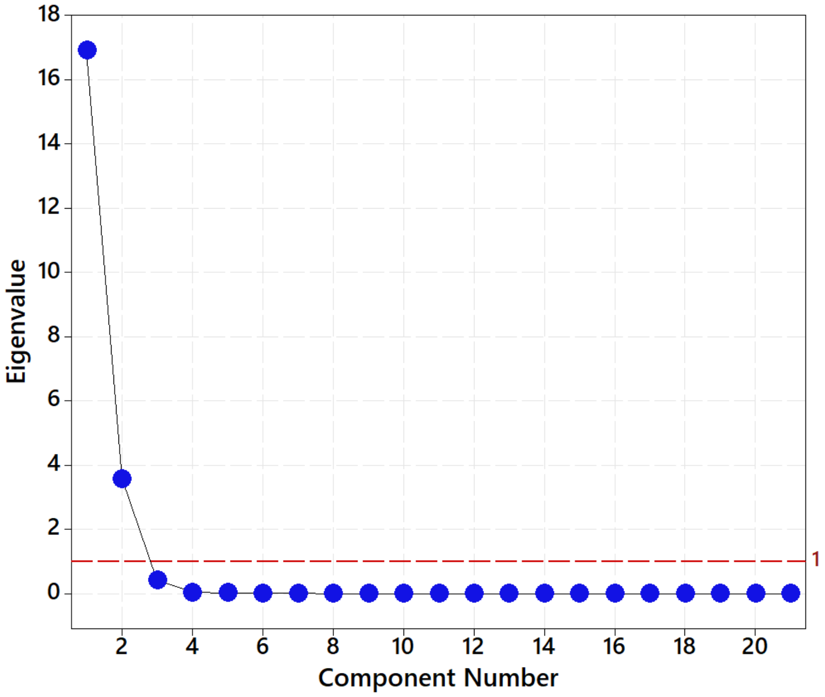
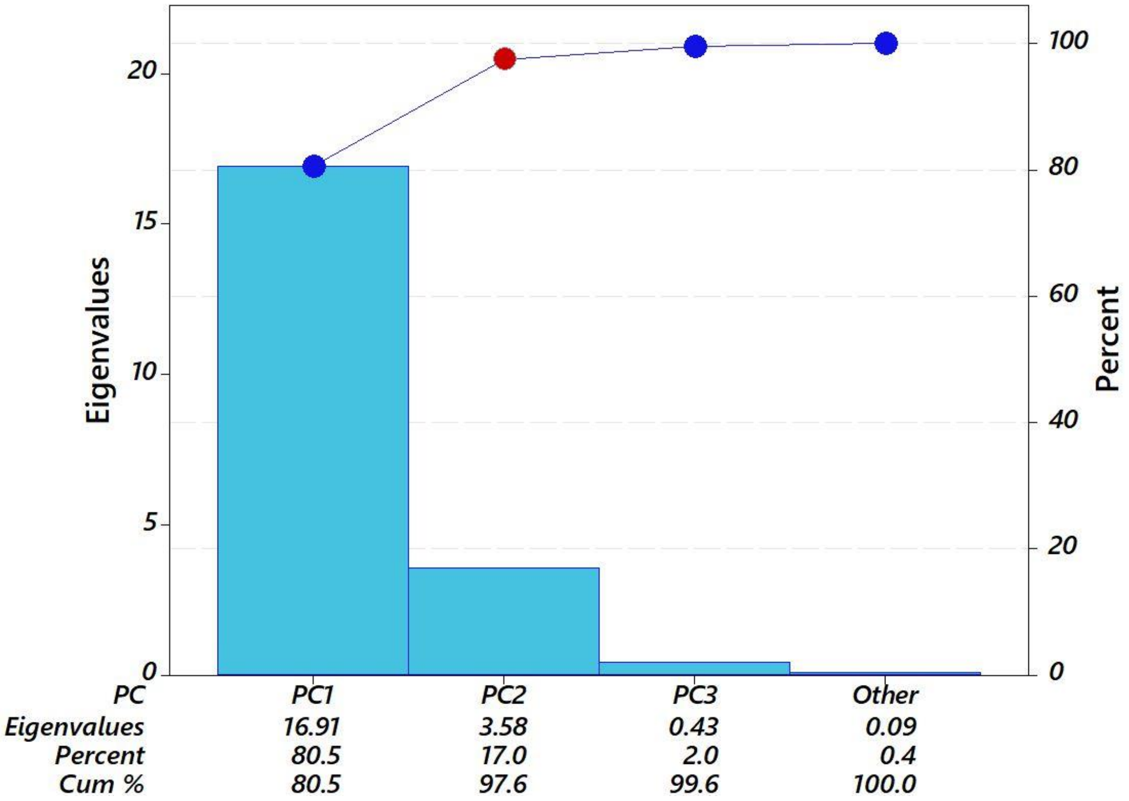


Figure 16: Scree plot of eigenvalues from C-MAPSS outputs.



PC	PC1	PC2	PC3	Other
Eigenvalues	16.91	3.58	0.43	0.09
Percent	80.5	17.0	2.0	0.4
Cum %	80.5	97.6	99.6	100.0

Figure 17: PCA Pareto chart for C-MAPSS outputs.

In a complementary way, it is possible to verify which variables have the greatest effect on these components. This relationship can be seen through a loading plot, as described in Figure 18. In this graph, the variables closer to 0 have a weak influence and, antagonistically, those close to 1 or -1 have a strong influence on the component.

From the ideal number of components (in this case, only 2), it is possible to extract the scores that will represent all the original variables through independent and dimensionless vectors. From the application of the analysis, it is possible to reach a reduction of the data dimensionality of 95.25%, i.e., the original set that had 713,811 data points can be properly represented by 67,822 data points. This significant reduction favors the computational and analytical performance of the data, reducing the time and helping for more accurate evaluations, since the PCA considers the correlated structure of the data. The following studies bring more details and information about the object of study: Saxena et al., (2008); Saxena and Goebel, (2008).

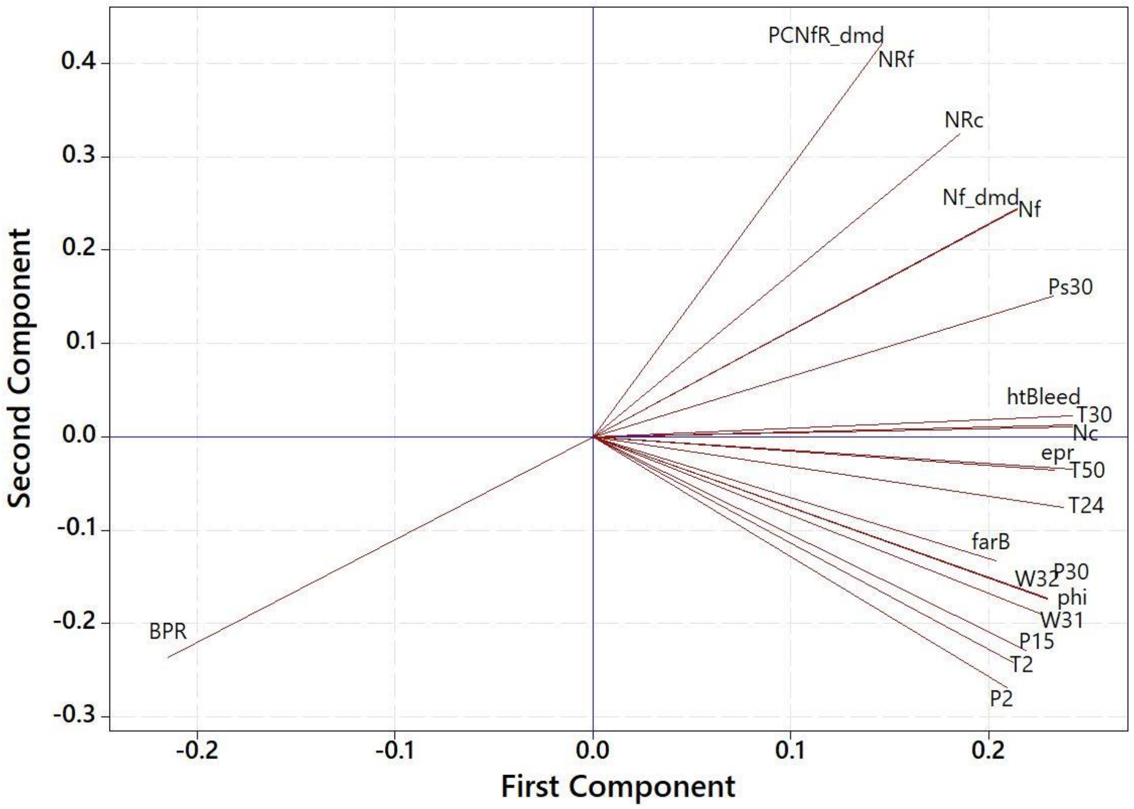


Figure 18: Loading plot from the C-MAPSS outputs

4 Conclusion

In this chapter we present a general and brief overview of the principal component analysis technique. This technique is characterized as a powerful tool to reduce the data dimensionality and create uncorrelated response vectors, being widely used in several segments. In addition to an explanation of this strategy, this chapter described 3 examples that cover different areas of engineering, showing the possibilities of using PCA. Finally, it is hoped that this chapter will help readers understand PCA, as well as how to apply this strategy to other multivariate data sets.

Acknowledgements

The authors would like to acknowledge the financial support from the Brazilian agency CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) under process number 150117/2021–3, CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), and FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais - APQ-00385-18).

References

Almeida, F. A. de. (2021). Aprimoramento do poder discriminatório de funções elipsoidais modificadas por cargas fatoriais rotacionadas na formação otimizada de agrupamentos. Universidade Federal de Itajubá (UNIFEI).

Almeida, F. A. De, Mello, L. G. D., Romão, E. L., Gomes, G. F., Gomes, J. H. D. F., Paiva, A. P. De, Filho, J. M., & Balestrassi, P. P. (2021). A PCA-Based Consistency and Sensitivity Approach for Assessing Linkage Methods in Voltage Sag Studies. *IEEE Access*, *9*, 84871–84885. https://doi.org/10.1109/ACCESS.2021.3088436

Almeida, F. A. de, Romão, E. L., Gomes, G. F., Gomes, J. H. de F., Paiva, A. P. de, Miranda Filho, J., & Balestrassi, P. P. (2022). Combining machine learning techniques with Kappa–Kendall indexes for robust hard-cluster assessment in substation pattern recognition. *Electric Power Systems Research*, *206*, 107778. https://doi.org/https://doi.org/10.1016/j.epsr.2022.107778

Almeida, F. A. (2017). *Análise Multivariada do Sistema de Medição de um Processo de Solda a Ponto por Resistência Elétrica utilizando Componentes Principais Ponderados.* Universidade Federal de Itajubá.

Almeida, F. A., Leite, R. R., Gomes, G. F., Gomes, J. H. de F., & de Paiva, A. P. (2020). Multivariate data quality assessment based on rotated factor scores and confidence ellipsoids. *Decision Support Systems*, *129*, 113173. https://doi.org/10.1016/j.dss.2019.113173

Belinato, G., de Almeida, F. A., de Paiva, A. P., de Freitas Gomes, J. H., Balestrassi, P. P., & Rosa, P. A. R. C. (2019). A multivariate normal boundary intersection PCA-based approach to reduce dimensionality in optimization problems for LBM process. *Engineering with Computers*, *35*(4). https://doi.org/10.1007/s00366-018-0678-3

Bounoua, W., & Bakdi, A. (2021). Fault detection and diagnosis of nonlinear dynamical processes through correlation dimension and fractal analysis based dynamic kernel PCA. *Chemical Engineering Science*, *229*, 116099. https://doi.org/https://doi.org/10.1016/j.ces.2020.116099

Chatterjee, S., & Litt, J. (2003). Online Model Parameter Estimation of Jet Engine Degradation for Autonomous Propulsion Control. In *AIAA Guidance, Navigation, and Control Conference and Exhibit*. American Institute of Aeronautics and Astronautics. https://doi.org/doi:10.2514/6.2003-5425

Deng, Y., Bucchianico, A. Di, & Pechenizkiy, M. (2020). Controlling the accuracy and uncertainty trade-off in RUL prediction with a surrogate Wiener propagation model. *Reliability Engineering & System Safety*, *196*, 106727. https://doi.org/https://doi.org/10.1016/j.ress.2019.106727

Ferreira, D. F. (Federal U. of L. (2018). *Estatística Multivariada* (3th ed.). UFLA.

Gaudêncio, J. H. D., Almeida, F. A., Turrioni, J. B., Quinino, R. C., Balestrassi, P. P., & Paiva, A. P. (2019). A multiobjective optimization model for machining quality in the AISI 12L14 steel turning process using fuzzy multivariate mean square error. *Precision Engineering*, *56*(December 2018), 303–320. https://doi.org/10.1016/j.precisioneng.2019.01.001

Goebel, K., Qiu, H., Eklund, N., & Yan, W. (2007). Modeling Propagation of Gas Path Damage. *2007 IEEE Aerospace Conference*, 1–8. https://doi.org/10.1109/AERO.2007.352835

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(7), 498–520. https://doi.org/10.1037/h0070888

Johnson, R.A., Wichern, D. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). Prentice-Hall.

Jolliffe, I. T. (2010). *Principal Component Analysis* (2nd ed.). Springer Series in Statistics.

Kurosaki, M., Morioka, T., Ebina, K., Maruyama, M., Yasuda, T., & Endoh, M. (2004). Fault Detection and Identification in an IM270 Gas Turbine Using Measurements for Engine Control . *Journal of Engineering for Gas Turbines and Power*, *126*(4), 726–732. https://doi.org/10.1115/1.1787515

Listou Ellefsen, A., Bjørlykhaug, E., Æsøy, V., Ushakov, S., & Zhang, H. (2019). Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture. *Reliability Engineering & System Safety*, *183*, 240–251. https://doi.org/https://doi.org/10.1016/j.ress.2018.11.027

Mahmoudi, M. R., Heydari, M. H., Qasem, S. N., Mosavi, A., & Band, S. S. (2021). Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries. *Alexandria Engineering Journal*, *60*(1), 457–464. https://doi.org/https://doi.org/10.1016/j.aej.2020.09.013

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1995). *Multivariate Analysis* (5th ed.). Academic Press Limited.

Miranda, J., Maria, J., Filho, D. C., Paulo, A., Vitor, P., Souza, G. De, & Tomasin, S. (2016). A PCA-based approach for substation clustering for voltage sag studies in the Brazilian new energy context. *Electric Power Systems Research*, *136*, 31–42. https://doi.org/10.1016/j.epsr.2016.02.012

Nhu, V.-H., Samui, P., Kumar, D., Singh, A., Hoang, N.-D., & Tien Bui, D. (2020). Advanced soft computing techniques for predicting soil compression coefficient in engineering project: a comparative study. *Engineering with Computers*, *36*(4), 1405–1416. https://doi.org/10.1007/s00366-019-00772-7

Pearson, K. (1901). *On Lines and Planes of Closest Fit to Systems of Points in Space*. University College. https://books.google.com.br/books?id=uGt_YgEACAAJ

Saxena, A., & Goebel, K. (2008). *Turbofan Engine Degradation Simulation Data Set*. NASA Ames Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA. http://ti.arc.nasa.gov/project/prognostic-data-repository

Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. *2008 International Conference on Prognostics and Health Management*, 1–9. https://doi.org/10.1109/PHM.2008.4711414

Song, J., & Li, B. (2021). Nonlinear and additive principal component analysis for functional data. *Journal of Multivariate Analysis*, *181*, 104675. https://doi.org/https://doi.org/10.1016/j.jmva.2020.104675

Velasco, J. A., Amaris, H., & Alonso, M. (2020). Deep Learning loss model for large-scale low voltage smart grids. *International Journal of Electrical Power & Energy Systems*, *121*, 106054. https://doi.org/https://doi.org/10.1016/j.ijepes.2020.106054

Visinescu, L. L., & Evangelopoulos, N. (2014). Orthogonal rotations in latent semantic analysis: An empirical study. *Decision Support Systems*, *62*, 131–143. https://doi.org/10.1016/J.DSS.2014.03.010

Wang, F.-K., & Chien, T.-W. (2010). Process-oriented basis representation for a multivariate gauge study. *Computers & Industrial Engineering*, *58*(1), 143–150. https://doi.org/https://doi.org/10.1016/j.cie.2009.10.001

Xu, H., Fard, N., & Fang, Y. (2020). Time series chain graph for modeling reliability covariates in degradation process. *Reliability Engineering & System Safety*, *204*, 107207. https://doi.org/https://doi.org/10.1016/j.ress.2020.107207

Yu, Y., Peng, M., Wang, H., Ma, Z., & Li, W. (2020). Improved PCA model for multiple fault detection, isolation and reconstruction of sensors in nuclear power plant. *Annals of Nuclear Energy*, *148*, 107662. https://doi.org/https://doi.org/10.1016/j.anucene.2020.107662