

APPLYING PRINCIPAL COMPONENT ANALYSIS FOR AGGREGATION OF ENVIRONMENTAL VARIABLES AFFECTING THE OPERATIONS OF ELECTRICITY DISTRIBUTORS IN BRAZIL

Tatielle Menolli Longhini

Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais
R. Mato Grosso, 02 - Bairro - Vista Alegre, Ibirité – MG
tatielle.longhini@ifmg.edu.br

Marcelo Azevedo Costa

Universidade Federal de Minas Gerais
Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte – MG
macosta@ufmg.br

ABSTRACT

The National Electric Energy Agency (ANEEL) in Brazil regulates electricity tariffs using benchmarking methods like Data Envelopment Analysis (DEA) to assess the efficiency of operating costs among distributors. Recognizing that operational efficiency varies with geographic conditions, this study suggests incorporating environmental and operational factors into the DEA to avoid distortions. Principal Component Analysis (PCA) is introduced to integrate these environmental influences into the evaluation. By analyzing 21 environmental variables, PCA reduced them to four main components that significantly affect operating costs: Territoriality, Climatic Factor, Wildfires, and Accessibility/Infrastructure. These components enable adjustments to operating costs, ensuring a more level playing field by simulating similar environmental conditions across all companies. Applying this approach in a second-stage DEA model can enhance the accuracy and fairness of regulatory benchmarking and simplify result analysis by reducing the number of variables considered.

KEYWORDS. Principal Component Analysis (PCA). Operations of electricity distributor. Unsupervised analysis. Machine learning.

APLICAÇÃO DE ANÁLISE DE COMPONENTES PRINCIPAIS PARA AGREGAÇÃO DE VARIÁVEIS AMBIENTAIS QUE AFETAM AS OPERAÇÕES DE DISTRIBUIDORES DE ELETRICIDADE NO BRASIL

RESUMO

A Agência Nacional de Energia Elétrica (ANEEL) no Brasil regula as tarifas de eletricidade utilizando métodos de benchmarking, como a Análise Envoltória de Dados (DEA), para avaliar a eficiência dos custos operacionais entre as distribuidoras. Reconhecendo que a eficiência operacional varia com as condições geográficas, este estudo sugere incorporar fatores ambientais e operacionais na DEA para evitar distorções. A Análise de Componentes Principais (PCA) é introduzida para integrar essas influências ambientais na avaliação. Ao analisar 21 variáveis ambientais, a PCA reduziu-as a quatro componentes principais que afetam significativamente os custos operacionais: Territorialidade, Fator Climático, Incêndios Florestais e Acessibilidade/Infraestrutura. Esses componentes permitem ajustes nos custos operacionais, garantindo uma concorrência mais justa ao simular condições ambientais semelhantes para todas as empresas. Aplicar essa abordagem em um modelo DEA de segunda fase pode melhorar a precisão e a equidade do benchmarking regulatório e simplificar a análise dos resultados ao reduzir o número de variáveis consideradas.

PALAVRAS-CHAVE. Análise de Componentes Principais (PCA). Operações de distribuidora de energia elétrica. Análise não supervisionada. Aprendizado de máquina.

1. Introduction

Benchmarking methods are used in the electricity sector to estimate and compare individual operational costs with those of other companies in the industry [Llorca et al. 2016], [Haney and Pollitt 2013, 2009]. The Brazilian National Electric Energy Agency (ANEEL) regulates the Brazilian electricity sector by applying the DEA methodology since 2011, through which efficient cost frontiers are estimated [Costa et al. 2015].

It is common for input and output variables to be related to manageable factors that the operator can control to optimize efficiency. However, there is also a set of uncontrollable variables that can influence company performance, such as environmental or contextual variables. There is compelling evidence of a high correlation between DEA efficiency scores and environmental variables, leading to the common use of second-stage models to adjust for environmental heterogeneity [Silva et al. 2019a], [Gil et al. 2017], [Yu et al. 2009]. Instances of power interruption and voltage quality issues are primarily caused by environmental conditions. Factors such as adverse weather and geographical conditions have been identified as major influencers of electric service performance, requiring more empirical evidence and studies applied to benchmarking models [Growitsch et al. 2012], [Yu et al. 2009], [Yang and Pollitt 2009].

In the presence of a diversity of potential environmental factors, there are statistical procedures that can be used to estimate a reduced set or even a single environmental variable - the most common ones being principal component analysis and factor analysis, [Yu et al. 2009]. Principal Component Analysis (PCA) is a methodology used to condense the information contained in several original variables into a smaller set of variables. It is a multivariate technique that analyzes the interrelationships among a large number of variables, considering their correlations, to obtain components with minimal information loss. The linear combinations are called principal components and are uncorrelated with each other [Mingoti 2005].

The problem addressed in this work is associated with the fact that environmental variables are not considered in the current benchmarking model for calculating efficient operational costs, resulting in distortions due to the absence of adjustments for environmental particularities at the operating locations of distribution companies. This situation indicates the need for the use of additional techniques to assess environmental factors, especially in Brazil, where environmental heterogeneity is significant.

Therefore, the general objective of this work is to apply Principal Component Analysis (PCA) to the environmental variables of electric distribution companies in Brazil, in order to evaluate them more homogeneously, according to similarities in environmental conditions indicated by the obtained components. A set of environmental variables is currently provided by the regulator for studies.

The specific objectives of the work are: (i) to obtain the principal components associated with the environmental variables of electric distribution companies in Brazil; (ii) to interpret the obtained principal components. This study is organized as follows: besides the introduction, we have the literature review. Subsequently, the methodology of the work is presented, followed by the analysis and discussion of the results, where the obtained principal components are pointed out and interpreted. Finally, the final considerations of this study will be presented.

2. Literatura review

DEA is a comprehensive benchmarking model widely used by regulatory agencies worldwide. However, in Brazil, there are particularities that make the analysis of efficient operational costs complex, given the continental dimensions of the areas covered by concessions, environmental distinctions, and operational practices of transmission and distribution companies [Silva et al. 2019b]. It is noteworthy that some concession areas in Brazil are larger than some countries in Europe [Gil et al. 2017]. The environmental heterogeneity in the locations where Brazilian companies operate impacts the cost efficiency of energy transmission companies, which are of utmost importance for the modeling process [Silva et al. 2021]. In other words, there are

strong indications that the performance of these companies is not only affected by managerial factors, demonstrating the need for adaptation of the efficiency evaluation model.

To obtain reliable measures of inefficiency, it is necessary to control for the different environmental conditions under which each company operates, to assess whether inefficiency results from management failure or from such particularities [Llorca et al. 2016]. [Banker and Morey 1986] were pioneers in incorporating environmental variables into DEA models. [Ray 1988] further developed the second-stage analysis, where efficiency scores are initially estimated by DEA and subsequently correlated with environmental variables using linear regression models.

Yu et al. [2009] demonstrated that there are nine meteorological variables that tend to influence service interruptions, resulting in higher costs for the replacement of damaged equipment. Meanwhile, [Nillesen and Pollitt 2010] indicated that weather conditions have the most significant impact on transmission and distribution companies.

[Jamasp et al. 2010; 2012] identified that the inclusion of weather condition variables biases downwards the estimated coefficients of other relevant variables, mainly those related to the marginal cost of quality improvements. Concurrently, [Growitsch et al. 2012] pointed out that companies operating in favorable environmental conditions tend to have costs up to 30% lower than the average, while those operating in adverse conditions have costs 39% above the average.

[Gil et al. 2017] utilized geographic information in constructing a spatial statistics method to assess the performance of Brazilian electric distribution operators. Meanwhile, [Costa et al. 2019] proposed the development of geographical clusters, where the average efficient cost would be similar within the same cluster, suggesting segmenting the study area into smaller regions with similar environmental and socioeconomic factors, working with a smaller number of variables.

Periodically, public consultations are held to review and improve the methodology for calculating Regulatory Operational Costs and Tariff Regulation Procedures, with the latest being 062/2020. Historically, there have been attempts by the regulator to incorporate environmental variables into company assessments [Lopes et al 2021]. However, even though there is a predefined set of environmental variables, the methodology for environmental adjustment has not yet been presented by the regulator for the electric distribution sector. For future models aiming to incorporate environmental variables, the use of Principal Component Analysis (PCA) technique can aid in understanding environmental variables by condensing the information contained in several original variables into a smaller set of variables.

3. Methodology

In the present study, the Principal Component Analysis (PCA) technique is presented, used for data evaluation, as well as the research design. [Growitsch et al. 2012] used the technique of Factor Analysis (FA) to reduce environmental factors to a limited number of composite factors. The use of Principal Component Analysis (PCA) is also common for this type of situation. There is almost no difference between their statistical techniques [Costello and Osborn 2005]. For the present article, we will use PCA, a machine learning model, where an algorithm learns from experience to perform a task whose performance is measured.

It is an unsupervised technique, as the data is not labeled, and learning algorithms typically seek to group them by similarity [Johnson and Wichern 2007]. When the dataset is of large dimensionality, there is a tendency for points to be more sparse and the Euclidean distance loses its meaning, causing algorithms to lose performance [Timm 2002]. Therefore, it is common to use techniques to reduce the dimensions of the data through feature extraction or selection, allowing for an understanding of complex relationships between variables and providing insights [Skidmore and Thompson, 2010], [Hair et al. 2009].

PCA is an effective tool for reducing the dimensionality of data and identifying the main sources of variability in the system [Pearson 1901], [Hotelling 1933], [Barber 2020]. Where correlated attributes are transformed into uncorrelated sets through a simple orthogonalization procedure. Thus, obtaining the Principal Components (PCs), which result from the linear combination of the original attributes [Johnson and Wichern 2007], [Mingoti 2005]. With the

reduction in the number of variables, the aim is to interpret the constructed linear combinations. PCA uses the spectral decomposition of a covariance or correlation matrix to perform the relationship of Equation 1.

$$A = W \Lambda W^{-1} \text{ (Equation 1)}$$

The matrix A can be constructed from the covariances or correlations between the variables under study. And Λ is the diagonal matrix of eigenvalues of V , and W is the orthogonal $n \times n$ matrix of eigenvectors of V (which will be presented in Equation 2). [Mingoti 2005] defines X as a $T \times n$ matrix, where n variables are observed over T periods. The columns of X are denoted as $x_1, \dots, x_n$, representing data of variables correlated over time, with each variable having a mean of zero. Each x_j is a $T \times 1$ vector, where T represents the observations of each variable. The covariances of these data are summarized in the matrix V (Equation 2), which represents the correlations of the variables if the data are normalized (x_j with variance 1 and mean 0).

$$V = T^{-1} X' X \text{ (Equation 2)}$$

It is desired that $T > n$, so that all n eigenvalues ($\lambda_1, \lambda_2, \dots, \lambda_n$) of V are positive and do not generate a singular matrix. Each PC is calculated as a linear combination of the columns of X , where the weights are selected so that: (i) the principal components are uncorrelated with each other; (ii) the first principal component explains the most variation in X , and so on. As Λ is the diagonal matrix of eigenvalues of V , and W is the orthogonal matrix of eigenvectors of V , the principal component matrix P is $T \times n$, considered as a product of the data matrix X with the eigenvector matrix of V (Equation 3) and the m -th principal component of V is defined as the m -th column of P (Equation 4):

$$P = X_{T \times n} W_{n \times n} \text{ (Equation 3)}$$

$$p_m = w_{1m}x_1 + w_{2m}x_2 + \dots + w_{nm}x_n \text{ (Equation 4)}$$

Where $w_m = (w_{1m} \ w_{2m} \ \dots \ w_{nm})'$ is the eigenvector corresponding to λ_m (the m -th largest eigenvalue of V), where p_1 belongs to the first and largest eigenvalue λ_1 , p_2 to the second largest eigenvalue λ_2 , and so on. Resulting in Equation 5, which demonstrates that: (i) the covariance matrix is diagonal (i.e., the principal components are uncorrelated); and (ii) the variance of the m -th principal component is λ_m , the m -th eigenvalue of V .

$$T^{-1} P' P = \Lambda \text{ (Equation 5)}$$

The sum of the eigenvalues of V represents the total variation in X , where the proportion of the total variation explained by the m -th component is given by Equation 6, and the original variables are represented by linear combinations of the principal components (Equation 7).

$$\frac{\lambda_m}{\lambda_1 + \lambda_2 + \dots + \lambda_n} \text{ (Equation 6)}$$

$$x_i = w_{i1}p_1 + w_{i2}p_2 + \dots + w_{in}p_n \text{ (Equation 7)}$$

The generated results are analyzed to obtain the number of principal components to be retained, where p is the number of original variables and k is the principal components (k should be smaller than p). In summary, the technique seeks a linear combination that has the highest possible variance. Next, it will seek a linear combination that is orthogonal to the first (independent) and captures the maximum remaining variance - the process continues until the number of available variables equals the number of combinations. Thus, the first combination (eigenvector) has the highest variability (eigenvalue) and the last one has the lowest. Evaluation criteria (Table 1) are used to follow these steps.

Table 1. Steps for Principal Component Analysis (PCA).

Step	Description
(1) Correlation and Covariance Matrix	It involves decomposing the covariance matrix or the correlation matrix. When using the covariance matrix for extraction, the components are influenced by variables with higher variance, which will be responsible for the main factors. However, when the values of different attributes are on different scales, it is preferable to use the correlation matrix instead of the covariance matrix. For this, Pearson's linear correlation is used to measure the degree of correlation between two variables (the coefficient, typically represented by ρ , takes values between -1 and 1).
(2) Overall Suitability	(2.1) Kaiser-Meyer-Olkin (KMO) Statistic The KMO statistic compares simple correlations with partial correlations. Its results range from 0 to 1, assessing the adequacy of the sample in terms of the degree of partial correlation between values, which should be small - a KMO value close to 0 indicates that factor analysis may not be suitable (weak correlation between variables), whereas the closer its value is to 1, the more suitable. (2.2) Anti-image It is a way to obtain evidence about the need to eliminate a particular variable from the model. The Measure of Sampling Adequacy (MSA) is obtained for each variable, similar to KMO. (2.3) Bartlett's Sphericity Tests Bartlett's sphericity test evaluates the hypothesis that the correlation matrix can be the identity matrix with determinant equal to 1 - if the correlation matrix is equal to the identity matrix, this means that factor analysis should not be used. The following hypotheses are considered: H_0: the correlation matrix is an identity matrix; H_1: the correlation matrix is not an identity matrix.
(3) Factor Creation	From the spectral decomposition of the covariance or correlation matrix, eigenvalues are obtained, which represent the total variance explained by each factor and are used to define the ideal number of factors. The first k factors contain the largest eigenvalues and thus the most variation in the data. After finding the eigenvalues and eigenvectors, factor scores must be calculated - these are used to create the factors.
(4) Factor Loadings	The goal is to make the empirical results more interpretable, understanding the relationships extracted in the analysis. To do this, it is necessary to analyze the factor loading matrix. The factor loading is a measure of the correlation of the variables with each of the factors; its squared value indicates what percentage of the variance of a variable is explained by a factor. The higher the factor loading, the greater the correlation with a certain factor.
(5) Communality	Refers to the amount of each variable explained by the factors - that is, it is the total shared variance of each variable across all the factors extracted from eigenvalues greater than 1. The higher the communality, the greater the explanatory power of the variable by PCA.
(6) Number of Factors	To determine the number of factors, the latent root criterion is used, which involves discarding all components with eigenvalues below 1 (an eigenvalue equal to 1 means that the factor is represented by a single variable). Thus, factors are selected to ensure a percentage of the total variance of the model, typically with cumulative variability above 70% or 80%. Emphasis is placed on the need for parsimony (higher percentage of variance with the smallest possible number of factors).
(7) Rotation	Rotation aims to increase the explanatory power of the factors by maximizing their independence, facilitating the interpretation of factors since, often, the analyzed variables are important in more than one factor. The varimax method is the most successful and therefore the most commonly used, consisting of determining a rotation angle in which pairs of factors are rotated equally.

Note. [Ferreira 2008], [Johnson and Wichern 2007], [Timm 2002].

The research is quantitative in nature, where the 21 environmental variables of electric power distribution companies, obtained from Technical Note 176 of 2022 (NT176/2022), are average values for the years 2018 to 2021 and were analyzed using the Principal Component Analysis (PCA) technique, with R® software (Table 2).

Table 2. Environmental variables.

Abbreviation	Variable Name
PrecipMedAnnualMin	Minimum annual average precipitation
PrecipMedAnnualMax	Maximum annual average precipitation
PrecipMedAnnualMed	Annual average precipitation
FocQueim	Average annual number of heat spots (average number of spots)
DensFocQueim	Average annual density of heat spots (spots per km ²)
PercEstraPavQA	% paved roads over the total - number of accesses in the service area
PercEstraPavIN	% paved roads in relation to the total
PavMed_UR	Average urban unit paving
PavMed_RU	Average rural unit paving
TempMedAn	Annual average temperature
AreaAtuacao	Service area of the electric utility company
UCs	Consumer units
DensUCs	Density of consumer units (uc/km ²)
DensRede	Network density (network/km ²)
DeclMedPerc	Average slope (percentage)
DescAtmMed	Average incidence of lightning strikes (average per km ²)
DescAtmMax	Maximum lightning strikes
UnidCons_TI	Conservation units and indigenous lands (percentage)
AreaTotalVeg	Total area of vegetation
AltMedVeg	Average height of vegetation
VegAlta	Incidence of tall vegetation (percentage)

Note. own authorship (2024), obtained from Technical Note 176 of 2022 (NT176/2022).

In total, there are data from 52 distribution companies located in 26 states of the country - except Amapá. There are variables with very similar definitions among them. A characteristic that can be considered for PCA if there are few independent sources of variation and most of the system's variation can be represented by the most important components.

4. Presentation and discussion of results

There is significant variability in the values of environmental variables across different regions of the country, which can impact the efficiency of companies, especially in their operating costs. The maps for the environmental variables (Table 2) are presented in Figure 1.

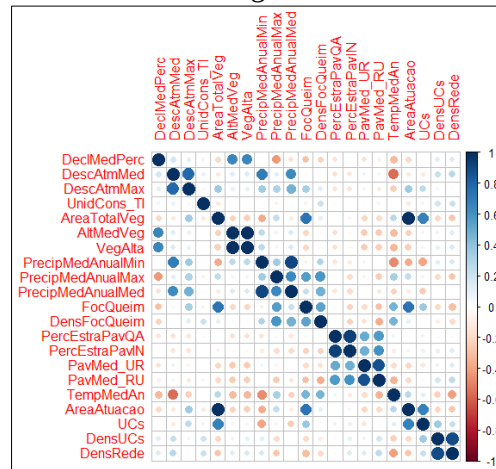
Figure 1. Environmental variables in the regions of Brazilian energy distributors.



Note. own authorship (2024).

Figure 2 displays the correlation matrix chart among pairs of environmental variables. Eight significant interactions (above 0.7) are observed, as indicated by [Hair et al. 2009], which are: (i) DescAtmMax x DescAtmMed, (ii) VegAlta x AltMedVeg, (iii) FocQueim x AreaTotalVeg, (iv) PrecipMedAnualMed x PrecipMedAnualMin, (v) PerEstraPavIN x PerEstraPavQA, (vi) PavMed_UR x PavMed_RU, (vii) AreaAtuacao x FocQueim, (viii) DensRede x DensUCs.

Figure 2. Correlations among the environmental variables.



Note. own authorship (2024).

For the development of the study, an analysis of the Kaiser-Meyer-Olkin (KMO) statistic was performed to assess the adequacy of the sample regarding the degree of partial correlation between the values. The values assumed by the variables were observed from the Measure of Sampling Adequacy (MSA) statistic as a way to obtain indications about the need to eliminate a certain variable from the model – according to Fávero and Belfiore (2017), variables should be eliminated if the MSA values are below 0.5.

Consequently, seven of the environmental variables were excluded, as their MSA values were low: AltMedVeg (0.4585023), VegAlta (0.4622058), PercEstraPavQA (0.4341284), PercEstraPavIN (0.4383455), DensRede (0.3815192), UnidCons_TI (0.4242100), DensUCs (0.3812257). As a result, the KMO, which previously had a value of 0.573021, increased to 0.6727562, having a classification very close to adequate. It is important to emphasize that AltMedVeg and VegAlta, although excluded to obtain the principal components, are environmental variables that directly influence the operational efficiencies of electricity distribution companies. It is likely that there is multicollinearity between these variables and those that remained in the DEA model.

Subsequently, with the other 14 environmental variables retained, the Bartlett's test of sphericity was conducted. As the obtained p-value was 1.29e-294 and is below the test statistic of 0.05, we say that the matrix is suitable for analysis by principal components. It is noted that there are four principal components with eigenvalues greater than 1 that capture the minimum percentage of 70% data variability, as suggested by [Mingoti 2005]. Therefore, the adjustment of the model with four principal components is capable of explaining 83.34% of the variability, as evidenced by the cumulative var (Table 3).

Table 3. Percentage of variability explained by the Principal Componentes (PC).

Percentage of variability	PC1	PC2	PC3	PC4
SS loadings	3.41	3.29	2.90	2.07
Proportion Var	0.2434079	0.2348413	0.2073121	0.1479318
Cumulative Var	0.2434079	0.4782491	0.6855613	0.8334930
Proportion Explained	0.2920335	0.2817555	0.2487269	0.1774841
Cumulative Proportion	0.2920335	0.5737890	0.8225159	1.00

Note. own authorship (2024).

To facilitate the interpretation of the results and highlight the relationship between variables within the principal components, a varimax rotation was performed - without changes to the analyses already conducted. Subsequently, factor scores were obtained to create linear combinations of variables and generate the functions for each of the four principal components under analysis (Table 4).

Table 4. Factor scores of the four Principal Components (PC).

	PC1	PC2	PC3	PC4
DeclMedPerc	0.005900905	0.0974005506	-0.285274423	-0.241884805
DescAtmMed	0.048255573	0.3027738715	-0.088685096	0.006088787
DescAtmMax	0.179984119	0.2701488960	-0.077316274	-0.018057151
AreaTotalVeg	0.287323288	0.0199773033	-0.009747924	0.027937824
PrecipMedAnnualMin	-0.120004033	0.2214271373	0.090677657	-0.007125287
PrecipMedAnnualMax	-0.007564402	0.0449781036	0.288174947	0.035953571
PrecipMedAnnualMed	-0.069280965	0.2078499303	0.174269639	0.003130964
FocQueim	0.156870766	0.0095730744	0.169053598	0.001892639
DensFocQueim	-0.075885082	-0.0109525908	0.271964354	-0.091428541
PavMed_UR	0.016793588	-0.0004473682	0.088785395	0.477015145
PavMed_RU	0.048863559	0.0087531676	0.001363602	0.464975082
TempMedAn	-0.028755648	-0.2179094770	0.221996310	-0.061782780
AreaAtuacao	0.287324170	0.0199763109	-0.009743982	0.027931448
UCs	0.266927014	0.0195650111	-0.098659907	0.033988665

Note. own authorship (2024).

Complementing the findings, the factor loading matrix is analyzed. The higher the factor loading, the greater the correlation with a specific component (Table 5). The most significant variables of each PC are highlighted in bold.

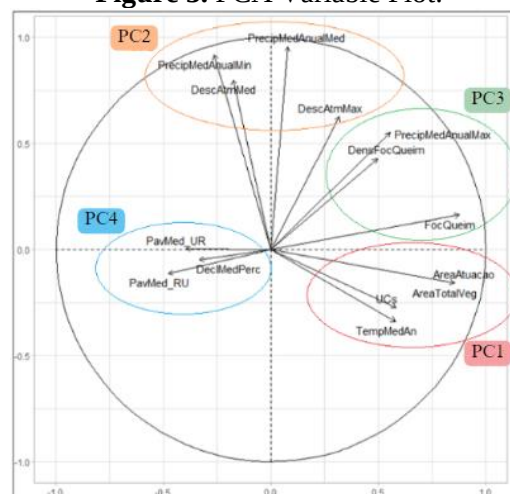
Table 5. Factor Loadings.

Factor Loadings	PC1	PC2	PC3	PC4
DeclMedPerc		0.195	-0.655	-0.35
DescAtmMed		0.94	-0.1	
DescAtmMax	0.46	0.785		
AreaTotalVeg	0.95		0.171	
PrecipMedAnnualMin	-0.428	0.813	0.273	
PrecipMedAnnualMax	0.145	0.271	0.83	
PrecipMedAnnualMed	-0.195	0.783	0.541	
FocQueim	0.651		0.606	-0.167
DensFocQueim		0.105	0.779	-0.295
PavMed_UR	-0.125			0.932
PavMed_RU			-0.205	0.937
TempMedAn	0.179	-0.614	0.566	-0.235
AreaAtuacao	0.95		0.171	
UCs	0.814		-0.105	

Note. own authorship (2024).

A PCA variable plot was also generated (Figure 3) to facilitate the interpretation of factor loadings and the composition of each component (Table 5). As can be seen from Table 5, TempMedAn has a low loading in CP1 compared to other values; however, the variable is assigned to this component based on the interpretations obtained from Figure 2. The original variables can be represented by linear combinations of the principal components (Table 6), as indicated in Table 4. Based on the most significant factor loadings, as determined by Table 5 and Figure 3, it was possible to interpret the generated principal components.

Figure 3. PCA Variable Plot.



Note. own authorship (2024).

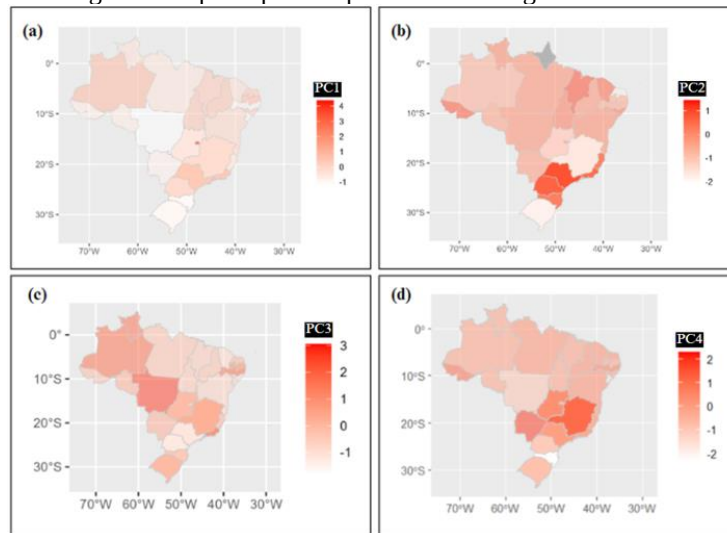
Table 6. Linear combinations of principal components (PC), original variables with the highest correlation with each PC, and interpretation of each PC.

PC	Linear Combination	Most significant variables	Interpretation of the PC
PC1	$0.005900905 \times \text{DecImedPerc} + 0.048255573 \times \text{DescAtmMed} + 0.179984119 \times \text{DescAtmMax} + 0.287323288 \times \text{AreaTotalVeg} - 0.120004033 \times \text{PrecipMedAnualMin} - 0.007564402 \times \text{PrecipMedAnualMax} - 0.069280965 \times \text{PrecipMedAnualMed} + 0.156870766 \times \text{FocQueim} - 0.075885082 \times \text{DensFocQueim} + 0.016793588 \times \text{PavMed_UR} + 0.048863559 \times \text{PavMed_RU} - 0.028755648 \times \text{TempMedAn} + 0.287324170 \times \text{AreaAtuacao} + 0.266927014 \times \text{UCs}$	AreaAtuacao AreaTotalVeg UCs TempMedAn	Territoriality
PC2	$0.0974005506 \times \text{DecImedPerc} + 0.3027738715 \times \text{DescAtmMed} + 0.2701488960 \times \text{DescAtmMax} + 0.0199773033 \times \text{AreaTotalVeg} + 0.2214271373 \times \text{PrecipMedAnualMin} + 0.0449781036 \times \text{PrecipMedAnualMax} + 0.2078499303 \times \text{PrecipMedAnualMed} + 0.0095730744 \times \text{FocQueim} - 0.0109525908 \times \text{DensFocQueim} - 0.0004473682 \times \text{PavMed_UR} + 0.0087531676 \times \text{PavMed_RU} - 0.2179094770 \times \text{TempMedAn} + 0.0199763109 \times \text{AreaAtuacao} + 0.0195650111 \times \text{UCs}$	DescAtmMed PrecipMedAnualMin PrecipMedAnualMed DescAtmMax	Climatic factor
PC3	$-0.285274423 \times \text{DecImedPerc} - 0.088685096 \times \text{DescAtmMed} - 0.077316274 \times \text{DescAtmMax} - 0.009747924 \times \text{AreaTotalVeg} + 0.090677657 \times \text{PrecipMedAnualMin} + 0.288174947 \times \text{PrecipMedAnualMax} + 0.174269639 \times \text{PrecipMedAnualMed} + 0.169053598 \times \text{FocQueim} + 0.271964354 \times \text{DensFocQueim} + 0.088785395 \times \text{PavMed_UR} + 0.001363602 \times \text{PavMed_RU} + 0.221996310 \times \text{TempMedAn} - 0.009743982 \times \text{AreaAtuacao} - 0.098659907 \times \text{UCs}$	FocQueim DensFocQueim PrecipMedAnualMax	Wildfires
PC4	$-0.241884805 \times \text{DecImedPerc} + 0.006088787 \times \text{DescAtmMed} - 0.018057151 \times \text{DescAtmMax} + 0.027937824 \times \text{AreaTotalVeg} - 0.007125287 \times \text{PrecipMedAnualMin} + 0.035953571 \times \text{PrecipMedAnualMax} + 0.003130964 \times \text{PrecipMedAnualMed} - 0.001892639 \times \text{FocQueim} - 0.091428541 \times \text{DensFocQueim} + 0.477015145 \times \text{PavMed_UR} + 0.464975082 \times \text{PavMed_RU} - 0.061782780 \times \text{TempMedAn} + 0.027931448 \times \text{AreaAtuacao} + 0.033988665 \times \text{UCs}$	PavMed_RU PavMed_UR DecImedPerc	Accessibility/ Infrastructure

Note. own authorship (2024).

For each of the generated principal components, the respective incidences in the regions of Brazilian energy distributors were obtained (Figure 4). Such visualizations will support the interpretations of each generated principal component and the impacts on each locality.

Figure 4. Incidence of the generated principal components in the regions of Brazilian energy distributors.



Note. own authorship (2024).

Caption. PC1 - Territoriality, PC2 - Climate Factor, PC3 – wildfires, PC4 - Accessibility/Infrastructure.

The principal components (PCs) were interpreted based on Figure 4 along with the incidences of each environmental variable, as indicated in Figure A1:

(1) **PC1 - Territoriality:** Its most significant variables in composition were the company's operating area (AreaAtuacao), vegetation area (AreaTotalVeg), consumer units (UCs), and average annual temperature (TempMedAn). It is noteworthy that TempMedAn presents the lowest factorial load. Generally, distributors operating in larger areas tend to have a greater total vegetation area and more consumer units as well – regarding the latter, it is important to observe the issue of operating zone (rural or urban). It is worth mentioning that the Amazon region is characterized by a large vegetation area, and São Paulo, Rio de Janeiro, and the Federal District have a high number of consumer units.

(2) **PC2 - Climate Factor:** There is a direct influence of the average atmospheric discharges (DescAtmMed), minimum average annual precipitation (PrecipMedAnualMin), average annual precipitation (PrecipMedAnualMed), and maximum atmospheric discharges (DescAtmMax).

(3) PC3 - Wildfires: The most significant variables were fire hotspots (FocQueim), fire hotspot density (DensFocQueim), and maximum average annual precipitation (PrecipMedAnualMax). It is observed that the largest fire hotspots are in regions of high vegetation and larger operating areas, such as the North and the Midwest (the cerrado has seen an increase in incidents over the years). Additionally, there is a similarity between the map of the principal component and the South Atlantic Convergence Zone (SACZ), characterized as one of the main meteorological systems influencing the rain regime in the Central-West and Southeast regions of Brazil, occurring more predominantly from late spring to summer. It is an area with a cloud corridor that crosses from the south of the Amazon Region to the Atlantic Ocean, passing through the central region of the country. This system is formed due to the behavior of winds at high and low levels of the atmosphere, which can create a cold front, a low-pressure center, or an elongated area of low pressure on the surface (INMET, 2024).

(4) PC4 - Accessibility/Infrastructure: There is a direct influence of variables such as average urban pavement (PavMed_UR), average percentage of pavement (PavMedPerc), and average percentage of slope (DeclMedPerc). It is noted that Minas Gerais, one of the regions with the highest prominence in the analysis, is marked by high slope due to being a mountainous region and by a large rural area (although they have less rural pavement and more urban pavement).

Reducing the analysis of 21 environmental variables into 4 principal components simplifies the interpretation and synthesis of information about environmental factors. Brazil is a country of continental proportions, covering an area of 8.5 million km², making it the 5th largest country in the world, composed of 27 states, many of which are larger than some European countries. The country spans various climatic zones impacting a wide ecological, geographical diversity, and propensity for rainfall. Moreover, each locality has different characteristics regarding the electrical grid and accessibility infrastructure.

The consideration of environmental variables is still recent, and the study by [Gil et al. 2017] demonstrates how significant their inclusion in the efficiency model is. Environmental factors are generally treated by electric sector companies as operating costs - companies located in adverse environments tend to adopt better management practices [Nillesen and Pollitt 2010].

It is noted that correcting efficiency scores based on environmental variables and the latent spatial structure reproduces significant differences between the original and corrected efficiency scores, especially among companies that received lower ratings in the original model. There is evidence of a cluster formation of companies with low scores in the Northern region, characterized by hostile environmental conditions - due to high rainfall, high humidity, and the presence of the tropical forest with tall vegetation -, and are mostly public companies [Gil et al. 2017].

Final considerations

The urgency of integrating environmental variables into the operational cost analysis methodologies of electric distribution companies in Brazil has become increasingly evident. Given the climatic effects and the diversity of environmental dynamics in the concession area of electric distribution companies, it becomes necessary to adjust the regulatory operational costs by the environmental components. This study aimed to apply Principal Component Analysis (PCA) to the environmental variables of electric distribution companies in Brazil. With this, it sought to evaluate them more homogeneously, according to environmental condition similarities indicated by the obtained components.

Out of the 21 environmental variables analyzed, the PCA synthesized into 4 main components, making the adjusted model capable of explaining 83.84% of the variances of the variables. The main components obtained were: (1) Territoriality; (2) Climatic Factor; (3) Wildfires; (4) Accessibility/Infrastructure.

As a suggestion for future studies, it points to the individual analysis of each of the environmental variables, according to the location of the electric distribution companies in Brazil, as well as of the main components obtained. Studies proposing the development of a second-stage

DEA, starting from the introduction of environmental variables, have become even more common, as exposed by [Silva et al. 2021].

Thus, it is concluded that the study achieved its objectives, managing to: (i) obtain the main components of the environmental variables of the electric distribution companies in Brazil; (ii) interpret the main components obtained from the environmental variables. Principal Component Analysis is configured as an efficient strategy to synthesize information without loss of data representativeness.

References

- Banker, R. D.; & Morey, R. C. (1986). Efficiency analysis for exogenously fixed inputs and outputs. *Operations research*, 34 (4). p. 513-521.
- Barber, D. (2020). Bayesian Reasoning and Machine Learning. UCL.
- Costa, M. A., Mineti, L. B., Mayrink, V. D., & Lopes, A. L. M. (2019). Bayesian detection of clusters in efficiency score maps: An application to brazilian energy regulation. *Applied Mathematical Modelling*, 68, p. 66–81.
- Costa, M.; Lopes, A. L.; & Matos, G. D. P. (2015). Statistical evaluation of Data Envelopment Analysis versus COLS Cobb–Douglas benchmarking models for the 2011 Brazilian tariff revision. *Socio-Economic Planning Sciences*, 49, p. 47-60. doi: <https://doi.org/10.1016/j.seps.2014.11.001>.
- Costello, A. B.; Osborne, J. W. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment*, 10(7):1-9.
- Gil, G. D. R.; Costa, M. A.; Lopes, A. L. M.; & Mayrink, V. D. (2017). Spatial statistical methods applied to the 2015 brazilian energy distribution benchmarking model: Accounting for unobserved determinants of inefficiencies. *Energy Economics*, 64. p. 373–383.
- Growitsch, C.; Jamsb, T.; & Wetzel, H. (2012). Efficiency effects of observed and unobserved heterogeneity: Evidence from Norwegian electricity distribution networks. *Energy Economics*, 34 (2). p.542-548.
- Haney, A. B.; & Pollitt, M. G. (2013). International benchmarking of electricity transmission by regulators: A contrast between theory and practice?. *Energy Policy*, 62. p. 267–281.
- Hair Jr., J.F.; Black, W.C.; Babin, B.J.; Anderson, R.E.; & Tatham, R.L. (2009). Análise multivariada de dados. 6.ed. Porto Alegre, Bookman.
- Hotelling, H. (1933). Analysis of a Complex of Statistical Variables Into Principal Components. *Journal of Educational Psychology*, 24 (6). p. 417-441.
- Jasmab, T.; & Pollitt, M. (2003). International benchmarking and regulation: an application to European electricity distribution utilities. *Energy policy*, 31(15). p. 1609-1622.
- Jamsb, T.; Orea, L.; & Pollitt, M. (2010). Weather factors and performance of network utilities: A methodology and application to electricity distribution. *EPRG Working Paper 1020*, Cambridge Working Paper in Economics 1042, University of Cambridge.
- Jamsb, T.; Orea, L.; & Pollitt, M. (2012). Estimating the marginal cost of quality improvements: The case of the UK electricity distribution companies. *Energy Economics*, 34 (5). p.1498-1506.
- Johnson, R. A.; & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Prentice-Hall Inc., 6 ed.
- Llorca, M.; Orea, L.; & Pollitt, M. G. (2016). Efficiency and environmental factors in the US electricity transmission industry. *Energy Economics*, 55. p. 234–246. doi 10.1016/j.eneco.2016.02.004
- Lopes, A. L.; Costa, M. A.; & Silva, A. V. (2021). Contribuição à Consulta Pública 062/2020. Lopes-Ahn & consultores associados. Web page. https://www.aneel.gov.br/consultas-publicas?p_auth=YFYW0rfJ&p_p_id=participacao_publica_WAR_participacao_publica_portlet&p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&p_p_col_id=column-2&p_p_col_pos=1&p_p_col_count=2&_participacao_publica_WAR_participacao_publica_portlet_id=ParticipacaoPublica=3477&_participacao_publica

ca_WAR_participacaoPublicaportlet_javax.portlet.action=visualizar
ParticipacaoPublica. Acessado: 2024-05-31.

Mingoti, S. A. (2005). Análise de dados através de métodos de estatística multivariada: Uma abordagem aplicada. Belo Horizonte: Editora UFMG. 297 p.

Nillesen, P.; & Pollitt, M. (2010). Using regulatory benchmarking techniques to set company performance targets: The case of US electricity. *Competition and Regulation in Network Industries*, 11(1): 50-85.

Pearson, K. (1901). Onlines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2 (11): 559-572.

Ray, S. C. (1988). Data envelopment analysis, nondiscretionary inputs and efficiency: an alternative interpretation. *Socio-Economic Planning Sciences*, 22(4): 167-176.

Silva, A.V.; Costa, M.A.; Lopes, A. L.M.; & Carmo, G. M. (2019a). A close look at second stage data envelopment analysis using compound error models and the Tobit model. *Socio-Economic Planning Sciences*, 65: 111-126. doi: <https://doi.org/10.1016/j.seps.2018.04.001>

Silva, A. V.; Costa, M. A.; Ahn, H.; & Lopes, A. L. M. (2019b). Performance benchmarking models for electricity transmission regulation: Caveats concerning the Brazilian case. *Utilities Policy*, 60 (100960). doi: <https://doi.org/10.1016/j.jup.2019.100960>.

Silva, A. V.; Costa, M. A.; & Lopes, A. L. M. (2021). Accounting multiple environmental variables in DEA energy transmission benchmarking modelling: The 2019 Brazilian case. *Socio-Economic Planning Sciences*, 101162. doi: <https://doi.org/10.1016/j.seps.2021.101162>.

Timm, N.H. (2002). Applied Multivariate Analysis. Springer: Verlang.

Yang, H.; Pollitt, M. (2009). Incorporating both undesirable outputs and uncontrollable variables into DEA: The performance of Chinese coal-fired power plants. *European Journal of Operational Research*, 197(3): 1095-1105.

Yu, W.; Jamasb, T.; & Pollitt, M. (2009). Does weather explain cost and quality performance? An analysis of UK electricity distribution companies. *Energy Policy*, 37(11): 4177-4188.