Article

Journal of the Brazilian Chemical Society

https://dx.doi.org/10.21577/0103-5053.20250006

J. Braz. Chem. Soc. **2025**, *36*, 5, e-20250006, 1-5

Sustainable Water Management for Steam Generation in Sugarcane Biorefineries: Applying PCA and MST Clustering in Sample Analysis

Érik Geraldo S. Souza[a] *and Fabiola M. V. Pereira*[*,a,b]

[a]*Grupo de Abordagens Analíticas Alternativas (GAAA), Instituto de Pesquisa em Bioenergia (IPBEN), Instituto de Química, Universidade Estadual Paulista (UNESP), 14800-060 Araraquara-SP, Brazil*

[b]*Instituto Nacional de Tecnologias Alternativas para Detecção, Avaliação Toxicológica e Remoção de Contaminantes Emergentes e Radioativos (INCT-DATREM), 14800-060 Araraquara-SP, Brazil*

Our study on sustainable water management in sugarcane biorefineries, which utilizes water as a primary resource for generating bioenergy through steam production, has employed a novel approach. High water quality is crucial for optimal efficiency, particularly in boiler operations. We have utilized unsupervised methods, such as principal component analysis (PCA) and minimum spanning tree (MST), alongside instrumental analysis data, to assess water quality in steam production. The PCA exploratory analysis identified three distinct clusters, with the relevant variables being conductivity and SiO_2 content, to differentiate the purity of a dataset of 120 samples. MST-based clustering corroborated the PCA findings, forming three clusters: sample 1 represented the purest water, while samples 3 and 6 were in different clusters, indicating less purity in boiler feedwater. These unsupervised methods are highly effective, providing accurate and reliable data analysis and significantly benefiting sugarcane biorefineries by eliminating subjective biases. The findings of this study promise to improve water management practices in sugarcane biorefineries, leading to more efficient and sustainable operations.

Keywords: purity water, boilers, bioenergy, chemometrics, sustainability

Introduction

Water quality throughout the process is crucial for sugarcane biorefineries since compromised water can influence the parameters of the sugarcane juice extracted. Furthermore, water quality guarantees the best efficiency, especially for boilers and other instruments used throughout the steam generation and use process. Sugarcane biorefineries can also produce energy using water as a primary source, generating condensed and steamed water.[1-5]

Analyzing data derived from reactions, processes, synthesis, and industrial plants across various sectors is extremely helpful in establishing standards, references, and concepts regarding uncertainties, analysis errors, and improving quality parameters.[6] For instance, one of these strategic non-supervised techniques is an exploratory analysis using principal component analysis (PCA), which can project the multidimensional data into a reduced number of variables known as principal components (PCs), as illustrated in equation 1:

$$\mathbf{X} = \mathbf{T}_a\mathbf{L}_a^T + \mathbf{E} \quad (1)$$

Matrix **X** contains the data of interest, decomposed into two matrices: score matrix **T** and orthonormal loading matrix **L**, with a matrix **E** representing errors for a specific number of principal components denoted as "a". The scores and loadings provide information about the samples and variables, respectively. Through exploratory analysis and data mining, it is possible to better understand the data, identify correlations between variables, and uncover underlying information. This approach allows for the identification of the main characteristics of the data, facilitating informed decision-making.[7,8]

Another technique is the minimum spanning tree (MST), which uses a graph where nodes represent stimuli and edges represent potential links, with weights typically employed to predict or reconstruct empirical dissimilarities data. The goal is to find the tree that spans the graph (ensuring a path

*e-mail: fabiola.verbi@unesp.br
Editor handled this article: Eduardo Carasek

exists between each pair of nodes without any cycles) such that the sum of the edge weights is minimized or maximized for dissimilarities or similarities, respectively. Solving the MST problem is formally equivalent to performing single-link clustering, and the relationship between clustering and spanning trees has proven highly valuable. However, it is assumed that interest in the MST first arose in engineering (e.g., in the layout of telephones, powerlines, and other networks). The main advantage over the different techniques is the application in Phyton,[9] which means there is no need for expensive software to run the code.[10-13]

This study uses mathematical and statistical techniques and instrumental analysis data[14] to evaluate the water quality in different steps of a power-generating plant dedicated to a biorefinery.

Experimental

Samples

Four water samples, four steam samples, and two condensed water samples were collected from April to June 2023, as shown in Figure 1. The number of analyses for each sample type was determined based on the requirements of the process. An average was calculated for each type, resulting in twelve replicates representing each sample, comprising 120 data samples.

The samples were collected in a sugarcane biorefinery located in Pitangueiras, São Paulo State, Brazil (–21.048431, –48.262912), where sugar, ethanol, yeast (*Saccharomyces cerevisiae*), and electrical energy are processed. Steam must be produced for this process, as all production stages require this resource. Therefore, water quality is essential for all systems to operate correctly. Thus, all samples are part of the steam generation process using high-pressure boilers at 67 kgf cm^{-2}.

Instrumental analysis

Several instrumental techniques were employed to monitor the water quality and composition of the samples. For pH and conductivity measurements, a mPA-210 pH meter (MS Tecnopon, Piracicaba, São Paulo State, Brazil) and a TEC-4MP digital conductivity meter (Tecnal, Piracicaba, São Paulo State, Brazil) were used, both at a controlled laboratory temperature of 25 ± 0.5 °C. For samples with pH values above 9.4, pH adjustment was performed beforehand using a neutralizing solution (Sinergia Científica, Campinas, São Paulo State, Brazil), added dropwise with continuous stirring until a color change from pink to colorless was achieved. Conductivity results were reported in μs cm^{-1}.

Two methods were used to determine the SiO_2 content based on sample composition, employing a DR 5000™ UV-Vis spectrophotometer (Hach, London, UK). The first method was used on most samples, while the second was explicitly applied to sample 9 (evaporative condenser).

Method 1

The sample was split into two 50 mL portions (sample and blank test). After adding specific reagents (molybdate 3, citric acid F, and amino acid F), the reaction progressed for set intervals before measurement. The spectrophotometer was set to the Silica ULR program, zeroed with the blank,

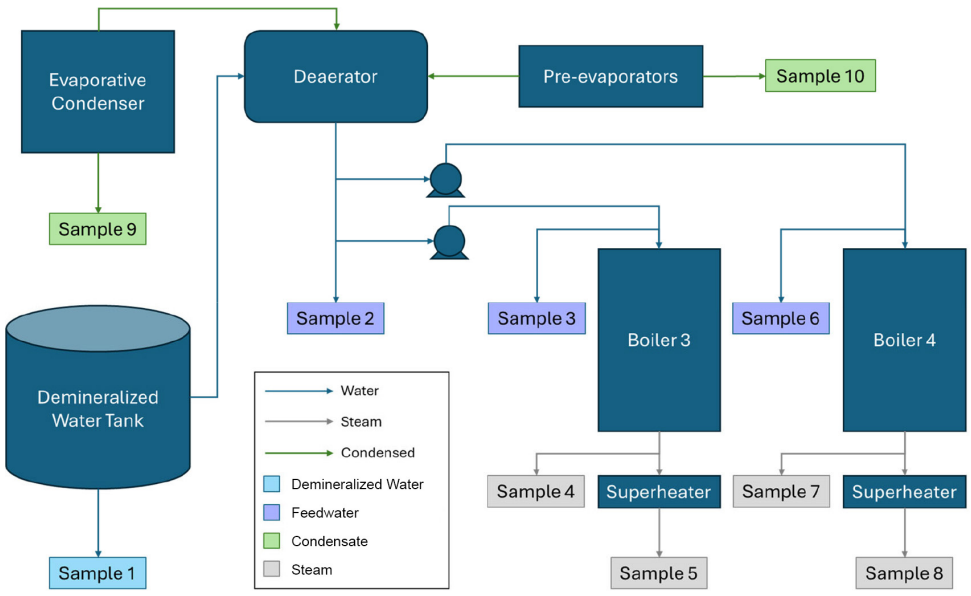

Figure 1. Locations of sampling points for water, steam, and condensed water in the steam generation process.

and the sample was read at 815 nm in a 10 mL cuvette, yielding results in mg L^{-1}.

Method 2

For sample 9, additional preparation was performed by adding specific reagents sequentially (2% hydrochloric acid, 10% oxalic acid, 10% ammonium molybdate, and 17% sodium sulfite), allowing the solution to rest for 10 min before analysis. The high silica program on the UV-Vis spectrophotometer was zeroed with a prepared blank test, and the sample was read at 460 nm with results recorded in mg L^{-1}.

This combined approach using UV-Vis spectrophotometry and other complementary instrumental techniques comprehensively analyzed water quality parameters, including pH, conductivity, and SiO_2 concentration.

Data processing

The recorded data were evaluated using the Matlab® 2023b (MathWorks, Natick, MA, USA)[15] routines developed in our research group and Pirouette 5.0 (Infometrix, Bothell, WA, USA) software.[16,17] The variables pH, conductivity, and SiO_2 content were pre-processed by autoscaling, i.e., mean equals 0 and standard deviation equals 1, for PCA calculations. Python codes were prepared to calculate the MST data evaluation.[18]

Results and Discussion

Exploratory analysis

PCA projects the multidimensional spectral information into compact matrices, termed scores (**T**) and loadings (**L**), arranged in descending order of explained variance, as shown in equation 1.[7,8]

The PCA exploratory analysis revealed three distinct clusters with 100% explained variance for three principal components (PCs), 62% for PC1, and 28 and 10% for PC2 and PC3, respectively.

Figure 2a presents the score plots, revealing three main clusters. The first cluster, shown as blue circles, represents sample 1, corresponding to the demineralized water tank. In pink circles, the second cluster includes samples 3 and 6, representing the feedwater of boilers 3 and 4, respectively.

The remaining samples form the third cluster, represented by gray circles. This group includes:

(*i*) Sample 2, from the feedwater deaerator,
(*ii*) Sample 4, representing saturated steam from boiler 3,
(*iii*) Sample 7, representing saturated steam from boiler 4,
(*iv*) Samples 5 and 8, representing superheated steam from boilers 3 and 4, respectively,
(*v*) Sample 9, corresponding to the evaporative condenser and
(*vi*) Sample 10 represents the condensate from the pre-evaporators.

Figure 2a shows the relationships among these clusters based on their water or steam sources. In Figure 2b, the loading plots revealed that the PC1 was responsible for its differentiation, with 62% explained variance visualized in the scores plot (Figure 2a). The essential variables were conductivity and SiO_2 content to differentiate the most (sample 1) to least (samples 2 and 6) pure samples. The pH parameter was associated with the differentiation of the cluster of other samples (gray circles) and 3 and 6 (pink circles) concerning the cluster of samples 1 (blue circles), with 28% of explained variance along PC2, shown in Figure 2b.

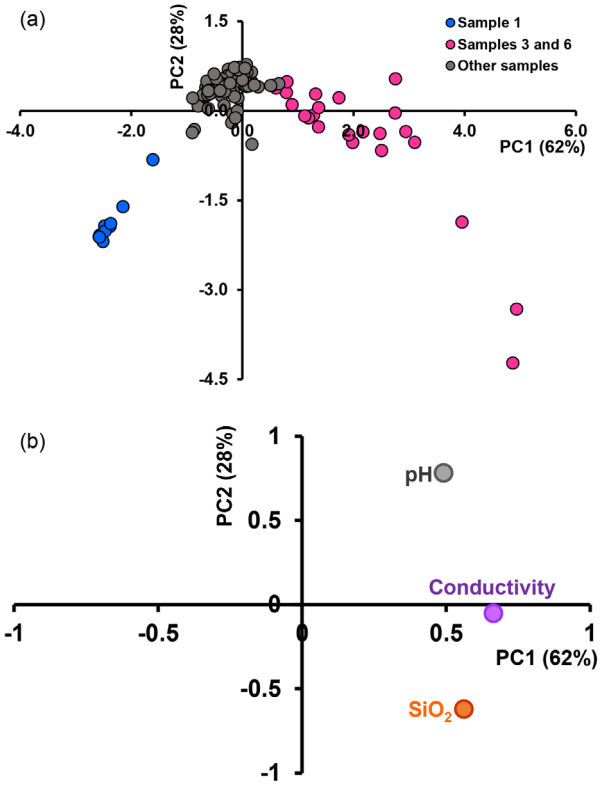

Figure 2. Score (a) and loading (b) plots calculated using principal component analysis (PCA) for data matrix samples (120 × 3).

Samples 3 and 6 showed vast differences in conductivity, 16-fold higher than sample 1, and SiO_2 content, 201-fold compared to sample 1. Analogize to other samples (2, 4, 5, 7-10), the conductivity is 8-fold, and SiO_2 content is 23-fold over sample 1. It is emphasized that the sample 1 cluster (blue circles) had the lowest values for conductivity, from 1.1 to 3.5, SiO_2 content, 0.002 to 0.008, and pH, between

5.5 and 7.5. The cluster for samples 3 and 6 (pink circles) ranged from 12.6 to 54.9 (conductivity), 0.1-1.5 (SiO_2 content), and 9.0-10.4 (pH), respectively. The cluster for other samples (gray circles) had values from 4.7 to 26.2, 0.002 to 0.2, and 8.2 to 9.8 for the same parameters, respectively.

Minimum spanning tree-based clustering

A k-nearest neighbors (k-NN) graph connecting each sample to its k-nearest neighbors in the input space was created, as shown in Figure S1 (Supplementary Information (SI) section). With the help of Figure S1, the resulting k-NN graph for the samples, considering k = 5, showed the clustering tendency for the replicates of sample 1 (from 1 to 12, except for 3), which differ because they are positioned in distinct parts of the k-NN graph.

Minimum spanning tree-based clustering pertains to graph theory, which involves mathematical structures representing binary relationships among elements of a finite set. Typically, a graph comprises a set of vertices connected pairwise by edges. When an edge connects two vertices, they are considered neighbors, as shown in Figure S2 (SI section). Given G = (V, E, w), where w: $E \rightarrow R^+$ is the edge weighting function, obtain the spanning tree T that minimizes the following criterion as shown in equation 2. Note that finding the MST of a graph is an optimization problem.[10-12]

$$w(T) = \sum_{e \in T} w(e) \quad (2)$$

Kruskal's algorithm[13] constructs the minimum spanning tree (MST) of an arbitrary graph in a computationally efficient manner. The algorithm begins with a tree consisting of n vertices and no edges. Each iteration adds the lowest cost edge to the tree, ensuring no cycles are formed, as trees are acyclic structures. The following section presents Kruskal's algorithm for obtaining an MST.

The divisive MST-based algorithm was applied to the dataset to find clusters in the samples. The most well-known class of MST-based clustering algorithms are the divisive algorithms that employ the single linkage strategy. In this approach, the distance between two clusters is determined by the closest pair of elements from each cluster. The main idea of this algorithm is to remove edges from the MST to minimize the sum of the intra-cluster spreads of the partitions, a criterion similar to that used in the K-means algorithm. Figure S3 (SI section) shows the results for two clusters when the first largest edge was removed.

Figure 3 shows that the second largest edge was removed, resulting in three clusters, a pattern consistent with the clustering observed in the PCA data analysis. In that analysis, sample 1 was identified as the purest water, while samples 3 and 6 formed separated clusters, representing the least pure feedwater from boilers 3 and 4. Figure 3 highlighted the clusters for samples 1, 3, and 6, and Figure S4 (SI section) showed the original graph.

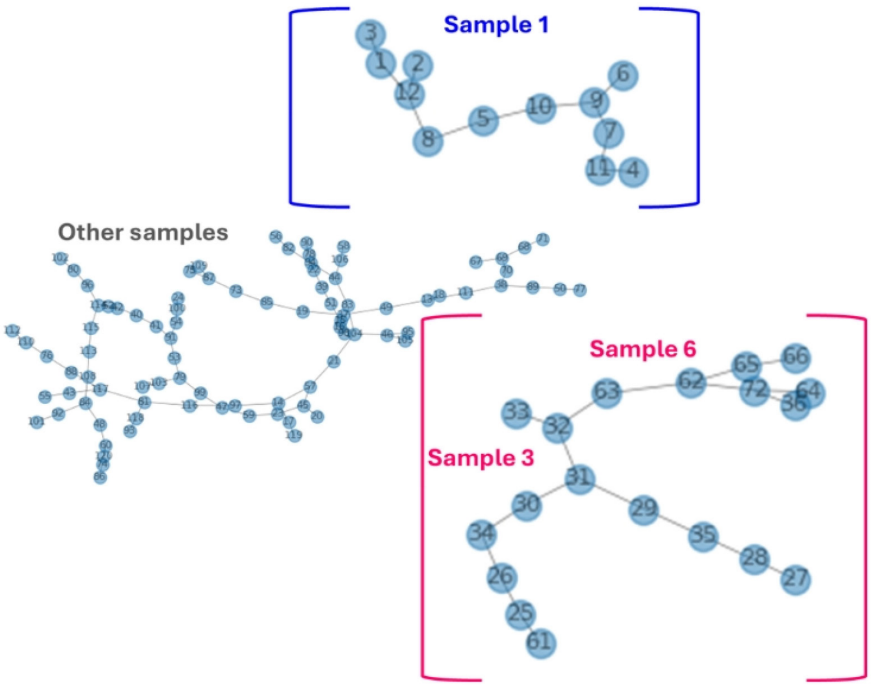

Figure 3. The divisive minimum spanning tree (MST)-based clustering found in the three clusters. Note that the first smallest cluster comprises the replicates of samples 1 (1-12), the second for samples 3 (25-36), and 6 (62-66 and 72).

Conclusions

The outcomes of the PCA and MST analyses highlighted the differentiation among the water samples, mainly their purity based on an unsupervised pattern. Sample 1 from the demineralized water tank was the purest, while samples 3 and 6 were the least pure from the feedwater of boilers 3 and 4, respectively. This differentiation was evident from the conductivity, SiO_2 content, and pH data, and it was verified by analyzing all data simultaneously. Using two unsupervised algorithms-principal component analysis (PCA) and minimum spanning tree (MST)-offers a significant advantage for data analysis in biorefineries. These methods objectively reveal patterns and relationships within the data without introducing subjective bias, highlighting impactful insights for optimizing biorefinery processes.

Supplementary Information

Supplementary data are available free of charge at http://jbcs.sbq.org.br as PDF file.

Acknowledgments

This study was supported by the São Paulo Research Foundation (FAPESP) under grant No. 2014/50945-4; National Council for Scientific and Technological Development (CNPq) grants Nos. 302085/2022-0, and 465571/2014-0; and the Coordination for the Improvement of Higher Education Personnel (CAPES)-Finance Code 001 and grant No. 88887136426/2017/00.

Author Contributions

The authors contributed equally to conceptualization, data curation, and formal analysis.

References

1. Demadis, K. D.; Mavredaki, E.; Stathoulopoulou, A.; Neofotistou, E.; Mantzaridis, C.; *Desalination* **2007**, *213*, 38. [Crossref]
2. Andrade, D. F.; Guedes, W. N.; Pereira, F. M. V.; *Microchem. J.* **2018**, *137*, 443. [Crossref]
3. Guedes, W. N.; Pereira, F. M. V.; *Microchem. J.* **2018**, *143*, 331. [Crossref]
4. Guedes, W. N.; Pereira, F. M. V.; *Comput. Electron. Agric.* **2019**, *156*, 307. [Crossref]
5. Guedes, W. N.; Santos, L. J.; Filletti, E. R.; Pereira, F. M. V.; *Food Anal. Methods* **2020**, *13*, 140. [Crossref]
6. Olivieri, A. C.; Faber, N. M.; Ferré, J.; Boqué, R.; Kalivas, J. H.; Mark, H.; *Pure Appl. Chem.* **2006**, *78*, 633. [Crossref]
7. Wold, S.; Esbensen, K.; Geladi, P.; *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37. [Crossref]
8. Bro, R.; Smilde, A. K.; *Anal. Methods* **2014**, *6*, 2812. [Crossref]
9. *Python*, 7.0.8; Jupyter Team, USA, 2023. [Link] accessed in January 2025
10. Arabie, P.; Hubert, L. J.; *Annu. Rev. Psychol.* **1992**, *43*, 169. [Crossref]
11. Gower, J. C.; Ross, G. J. S.; *Appl. Stat.* **1969**, *18*, 54. [Crossref]
12. Hubert, L. J.; *Br. J. Math. Stat. Psychol.* **1974**, *27*, 14. [Crossref]
13. Kruskal Jr., J. B.; *Proc. Am. Math. Soc.* **1956**, *7*, 48. [Crossref]
14. Souza, E. G. S.; Pereira, F. M. V.; *Braz. J. Anal. Chem.* [Crossref]
15. *Matlab*, 2023b; MathWorks, Natick, MA, USA, 2023.
16. *Pirouette*, 5.0; Infometrix, Bothell, WA, USA 2023.
17. Sperança, M. A.; Nascimento, P. A. M.; Olivieri, A. C.; Pereira, F. M. V.; *Biofuels, Bioprod. Biorefin.* **2022**, *16*, 758. [Crossref]
18. Castello, H. F.; Silva, F. L. R.; Ferreira, D. S.; Levada, A. L. M.; Pereira-Filho, E. R.; Pereira, F. M. V.; *J. Chemom.* **2024**, *38*, e3575. [Crossref]

Submitted: October 1, 2024
Published online: January 10, 2025