Energy and AI 5 (2021) 100084

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.elsevier.com/locate/egyai

Increasing power plant efficiency with clustering methods and Variable Importance Index assessment

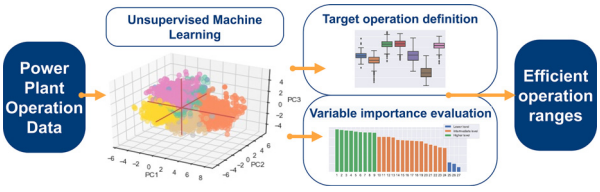
Jéssica Duarte[a,*], Lara Werncke Vieira[a], Augusto Delavald Marques[a], Paulo Smith Schneider[a], Guilherme Pumi[b], Taiane Schaedler Prass[b]

[a] Department of Mechanical Engineering, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
[b] Graduate Program in Statistics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

HIGHLIGHTS

- K-means and PCA applied to one-year real data of an existing power plant.
- Identification of the highest steam generator efficiency historical operating pattern.
- Evaluation of the process variables impact on the power plant.
- Definition of the Variable Importance Index to measure process variables impact.
- Definition of the target operating ranges to be set for each variable.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 29 January 2021
Received in revised form 21 April 2021
Accepted 21 April 2021
Available online 4 May 2021

Keywords:
Thermal power plant performance enhancement
Operating patterns identification
K-means clustering
Principal component analysis
Unsupervised machine learning

ABSTRACT

Power plant performance can decrease along with its life span, and move away from the design and commissioning targets. Maintenance issues, operational practices, market restrictions, and financial objectives may lead to that behavior, and the knowledge of appropriate actions could support the system to retake its original operational performance. This paper applies unsupervised machine learning techniques to identify operating patterns based on the power plant's historical data which leads to the identification of appropriate steam generator efficiency conditions. The selected operational variables are evaluated in respect to their impact on the system performance, quantified by the Variable Importance Index. That metric is proposed to identify the variables among a much wide set of monitored data whose variation impacts the overall power plant operation, and should be controlled with more attention. Principal Component Analysis (PCA) and k-means + + clustering techniques are used to identify suitable operational conditions from a one-year-long data set with 27 recorded variables from a steam generator of a 360MW thermal power plant. The adequate number of clusters is identified by the average Silhouette coefficient and the Variable Importance Index sorts nine variables as the most relevant ones, to finally group recommended settings to achieve the target conditions. Results show performance gains in respect to the average historical values of 73.5% and the lowest efficiency condition records of 68%, to the target steam generator efficiency of 76%.

1. Introduction

Fossil-fueled power plants are nowadays the main option for electricity generation, despite global initiatives on energy transition. Electricity demand is expected to grow by at least 25% by 2040 worldwide, and thermal power plants are to be considered on the global power mix. That predicted scenario justifies the efforts to attain efficient and responsible operation [1].

Plant performance decreases and loses its original design targets over years of operation, due to aging, maintenance issues, unexpected failures, and variability of operation patterns. Although the power plant operation includes a control system that handles plant stability, controllable losses are managed directly by operators. However, the power plant operators do not manage every aspect related to efficiency gains to guarantee the most advantageous approach. Their approach to the operation is subject to their individual experiences, due to a lack of directives to achieve high efficiencies. Their choice for conservative operation sometimes causes the system to decrease its performance. That opposi-

* Corresponding author.

https://doi.org/10.1016/j.egyai.2021.100084

Table 1
Number of variables adopted by the assessed references.

Number of Variables	References
Less than 15	[3,6,8–15,18–20,22]
More than 15	[4,5,7,16,17,21]

Table 2
Time interval adopted by the assessed references.

Time interval	References
No information	[4,11,16]
Less than 24h	[3,8,10,13–15,20]
Less than 4 months	[5–7,9,17,19,21,22]
A Year	[18]

tion establishes a need to merge these antagonistic strategies. Comprehensive knowledge of power plant operation evolving conditions allows preventing the system to deviate from its design performance, which concerns conversion efficiency, availability, safety, and emission levels.

Data-driven techniques can be an appropriate option to help operational decision-making, as they can identify plant patterns that promote the targeted performances while ensuring stable operation. Among these techniques, machine learning is an attractive option to handle such complex problems. Unsupervised machine learning techniques are widely applied for exploratory data analysis and pattern recognition purposes [2]. Among them, clustering analysis is a flexible tool that allows discovering unsuspicious hidden group patterns in the data.

A literature review was carried out to identify machine learning applications on thermal power plants operating conditions. Were selected papers that applied machine learning clustering techniques, presented a case study related to the operation of a thermal power plant and analyzed the plant operating conditions. As a result, 20 papers were selected and analyzed.

Many studies used pattern identification as an intermediate result to accomplish their research goal, while others would analyze deeper into the encountered conditions. References [3–8] developed fault detection and performance degradation methods, searching for high-efficiency or lower emissions. References [9,10] focused on multi model-method approaches. References [11–13] applied clustering techniques as an intermediary step to predict operation features values. References [14,15] approached Coordinated Control Systems (CCS). Reference [16] evaluated indicators to compare power plants competitivity. Reference [17] evaluated the evolution of a combustion process by trajectory analysis, to define the best path for achieving lower emissions. References [18–22] searched to optimize operating variable settings with different approaches, such as considering emission formation [21]. The wide scope of goals brought by these selected papers indicates that the identification of unlabeled operating conditions may promote diverse relevant insights on power plant operation.

Regarding the applied methodology, soft clustering methods were applied by Choi et al. [8], Xiaoying et al. [13], Hou et al. [14], Wu et al. [15], Chengbing et al. [18], Jia et al. [19], Liu et al. [20], such as Fuzzy C-means, when the authors considered the fuzzy membership degree results in their application. Meanwhile, [3–7,9,10,17,21,22] applied hard clustering methods, mainly the *k*-means technique, and each data point is associated with only one cluster.

The number of variables considered in the case studies may represent the extension of process aspects taken into account. The selected studies typically analyzed a limited selection of variables from the operation. As presented in Table 1, fourteen of the analyzed studies considered less than 15 variables, while 6 considered a larger range of variables. Most of them did not perform a variable selection analysis. In [3,5,7,18], the variable selection was based on the association of monitored input to output relations, such as the correlation coefficient.

Table 2 presents the time interval considered in the selected papers, divided into three categories. Four of the paper did not inform the interval, but most of those that presented it sampled data from an interval of up to 24 h. Only one paper analyzed a one-year period.

The previous review pointed out the ability of clustering techniques to identify process patterns, but none of these works proposed target operation set points while indicating the variables that should be monitored more effectively. An analysis over a one-year period of operation data considering a large range of variables could guide the definition of a high-efficiency operation directive. The need to establish a comprehensive method to keep system performance within a suitable range led to formulate the research question of the present work: are machine learning-based methods able to map operating conditions and related variables that guarantee efficient plant operation? The objective is to mitigate the operator's bias and enlighten the setting of operating parameters.

2. Applied techniques

The statistical tools chosen for the task were the Principal Component Analysis (PCA), the *k*-means clustering technique, and the Silhouette Coefficient. A variable importance index is proposed to define cause and effect relationships between operational variables and plant performance.

2.1. Principal component analysis

Principal Component Analysis (PCA) is a well-known and widely applied tool for data reduction from large sets of observed variables. PCA considers a small set of linear combinations of the original variables, the principal components, orthogonal to each other, that summarizes the data variability as best as possible, instead of working with the complete dataset. Technically, principal components are a sequence of projections of the data (linear combinations) which are mutually uncorrelated and ordered by variance. They are based on the singular value decomposition of a matrix. Suppose a dataset with N variables arranged in a matrix, say X. The goal is to obtain a set of $p < N$ principal components by using PCA. The first principal component, say $\boldsymbol{v}_1$, is the linear combination of the columns of X for which $X\boldsymbol{v}_1$ has the highest possible variance; the second principal component, $\boldsymbol{v}_2$, is the linear combination of the columns of X for which $X\boldsymbol{v}_2$ has the highest possible variance among all linear combination orthogonal to $\boldsymbol{v}_1$, and so on. After p steps, a set of p new orthogonal components is obtained, sorted from the highest to the lowest variance representation. More details can be found in [2,23].

2.2. *k*-means++ clustering

Clustering methods are unsupervised machine learning techniques aiming to divide unlabeled data into homogenous subgroups, according to their degree of similarity [23]. The *k*-means++ clustering method uses the Euclidian distance metric in such a way that each item is only assigned to one cluster, with the nearest cluster centroid. Its implementation requires the specification of the number of clusters beforehand, and finding a reasonable optimal clustering number is essential for the accuracy of the results. The *k*-means++ was proposed by [24] and has the advantage that tends to select centroids that are distant from one another, and this improvement makes the *k*-means algorithm much less likely to converge to a suboptimal solution [25]. A cluster of data points from operation variables is, herein, considered to represent an operation pattern.

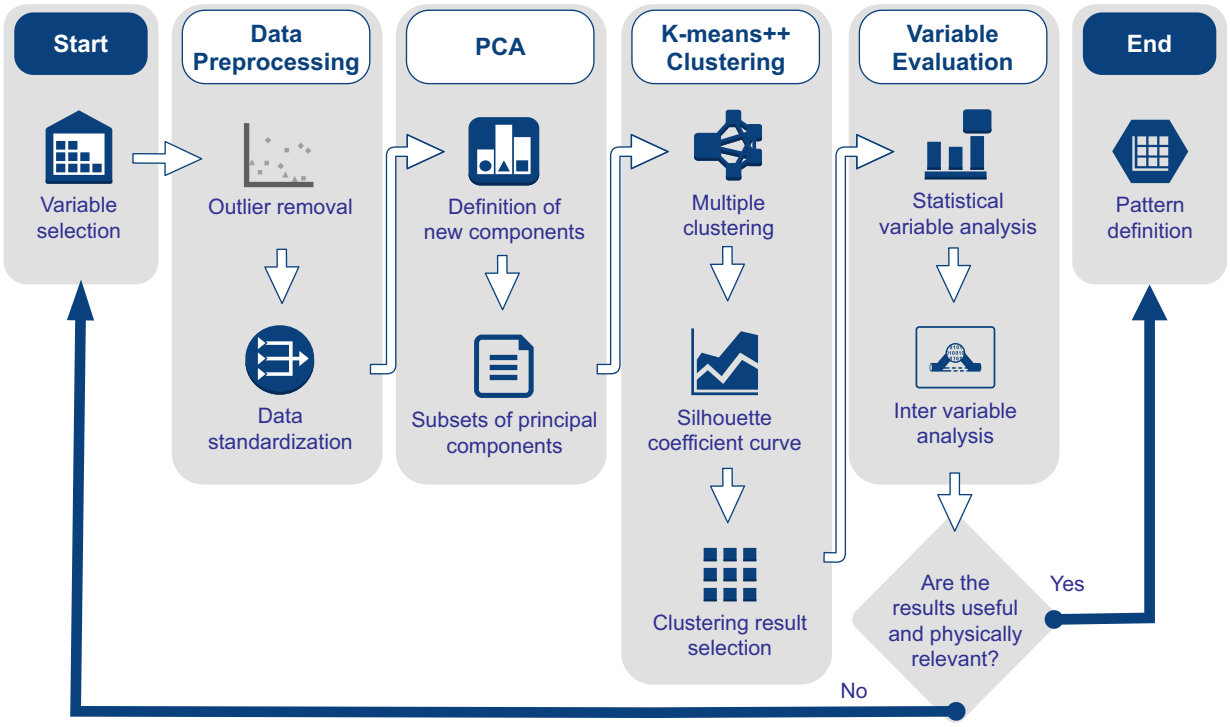


Fig. 1. Data set treatment and variable selection modeling.

2.3. Silhouette coefficient

The Silhouette coefficient is a similarity measure of how close a given object is to the group it belongs, compared to other groups. It can be viewed as a consistency measure for clustering analysis. The Silhouette coefficient takes into account the cohesion of the group and the separation between the groups in a given cluster analysis. Suppose the data have been clustered into p groups, say $G_1, \dots, G_p$ with $n_1, \dots, n_p > 1$ elements in each group, respectively. Let $d : \mathbb{R}^2 \to [0, \infty)$ denote a metric (such as the Euclidean distance). For a point $x \in G_i$, Let

$$A(x) = \frac{1}{n_i - 1} \sum_{y \in G_i, y \neq x} d(x, y), \tag{1}$$

and

$$B(x) = \min_{k \neq i} \left\{ \frac{1}{n_k} \sum_{y \in G_k} d(x, y) \right\}. \tag{2}$$

One can observe that $A(x)$ is the mean distance between x and all other points in G_i, while $B(x)$ is the smallest dissimilarity between x to all other clusters. The Silhouette coefficient $s(x)$ of a data point x is defined as

$$s(x) = \frac{B(x) - A(x)}{\max\{A(x), B(x)\}}. \tag{3}$$

The Silhouette coefficient ranges from -1 to 1, and a positive value close to one means that x belongs to a very compact cluster far apart from any other cluster [26].

3. Model approach

The methodology employed in this work is summarized in the diagram presented in Fig. 1.

Data modeling **starts** by the variable selection, whose first set may be quite broad, due to the exploratory nature of unsupervised learning algorithms. An appropriate range of variables should cover all aspects of the studied process, considering the available information. If too many irrelevant variables are considered, they are to be estimated of low relevance in the next iterations of the methodology. **Data preprocessing** comprehended outlier filtering and data standardization. A given data point (vector) is identified as an outlier if any of its values is three standard deviations away from the mean. Standardization is required at this point. **PCA** is performed and those components for which the respective eigenvalue is greater than one are retained, as described in Section 2.1. With the definition of the lower bound, subsets of different numbers of PCA components are considered in the following steps. ***k*-means++ Clustering** is applied to obtain clusters, which are a group of points with similar operating conditions regarding the values of the variables. The number of groups k must be set beforehand, which is obtained throughout the following exploratory approach: data were clustered multiple times for different values of k, over the selected principal components subset. For each k, the average Silhouette coefficient is calculated and a Silhouette index curve is plotted. Results are compared concerning the highest Silhouette coefficient values. The definition of k should also consider its physical results. A smaller number of clusters favors the establishment of well-defined global patterns in the operation, while a high number of clusters will promote a more sensible identification of different patterns in the analysis.

The **variable evaluation** analyses all input variables. It starts by defining a variable importance index suitable for the task. The variable importance index helps to evaluate the effect or impact of each variable on a given cluster, as it is a percentage of one variable regarding all the considered ones. Supposed variables were chosen in the first step of the methodology and let $x_1, \dots, x_d$ be the observations of these variables, respectively, with $x_i = (x_{i,1}, \dots, x_{i,n})$, where n denotes the total number of observations. Let $G_1, \dots, G_k$ denote the k-means clusters with $n_1, \dots, n_k > 1$ elements, respectively. Let $\bar{x}_{i,s}$ be the mean of the elements in x_i that belongs to cluster G_s, that is, for $i \in \{1, \dots, n\}$ and $s \in \{1, \dots, k\}$, we define

$$\bar{x}_{i,s} = \frac{1}{n_s} \sum_{\{j : x_{i,j} \in G_s\}} x_{i,j}. \tag{4}$$

For two clusters G_r and G_s, $r, s \in \{1, \dots, k\}$, $r \neq s$, and a variable x_i, $i \in \{1, \dots, d\}$, let $\psi_{r,s}(i)$ denote the squared difference between the average of the components of x_i that belong to cluster G_r and G_s, that is,

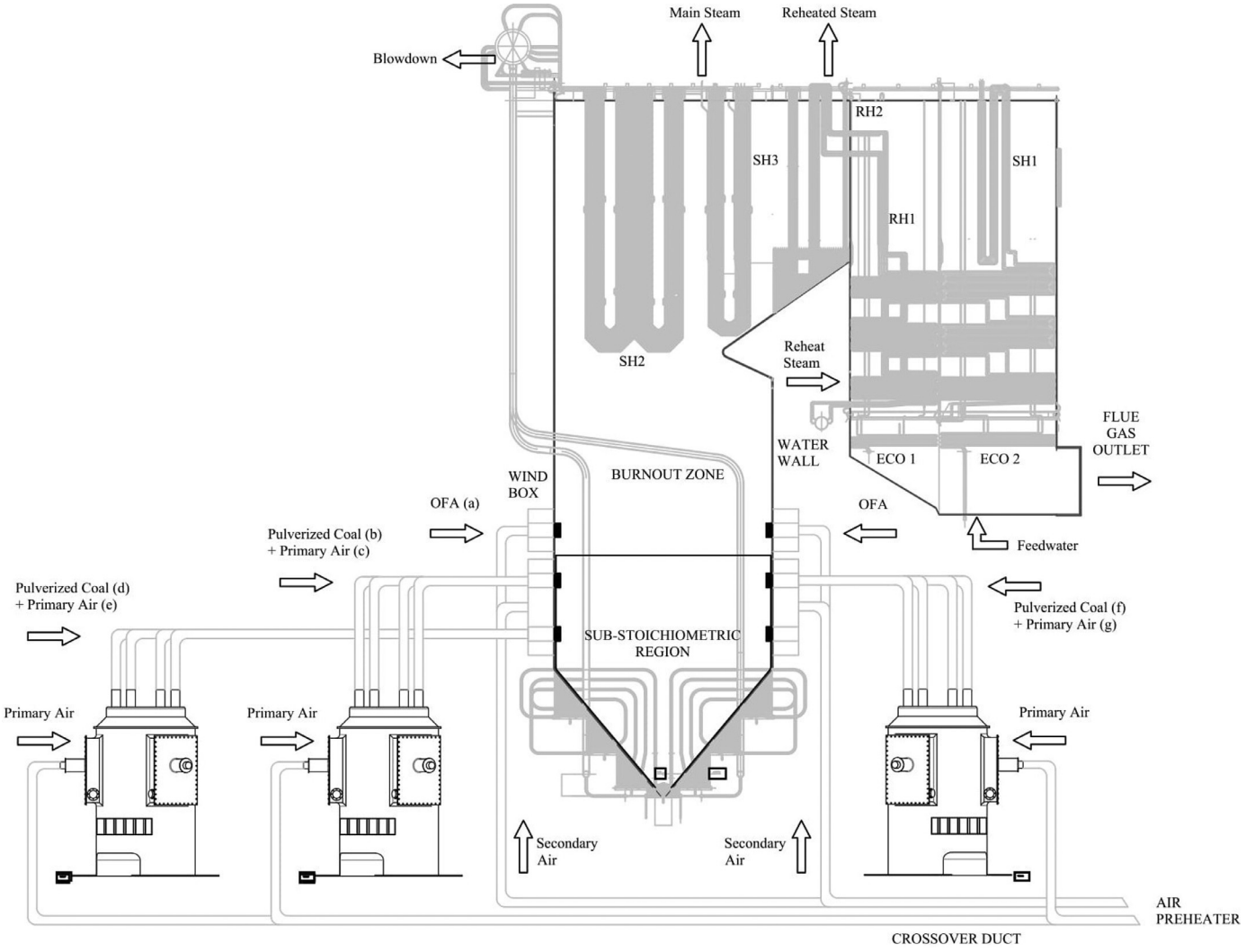


Fig. 2. PECEM power plant sub-critical steam generator and mills subset.

$$\psi_{r,s}(i) = \left(\bar{x}_{i,r} - \bar{x}_{i,s}\right)^2. \tag{5}$$

For $i \in \{1, \ldots, d\}$, the importance index of variable i is defined as

$$\mathcal{I}_i = \left[\frac{2}{(k-r)(k-1)} \sum_{r=1}^{k-1} \sum_{s=r+1}^{k} \psi_{r,s}(i)\right] \Bigg/ \left[\frac{2}{d(k-r)(k-1)} \sum_{j=1}^{d} \sum_{r=1}^{k-1} \sum_{s=r+1}^{k} \psi_{r,s}(j)\right], \tag{6}$$

$$= d\left[\sum_{r=1}^{k-1} \sum_{s=r+1}^{k} \psi_{r,s}(i)\right] \Bigg/ \left[\sum_{j=1}^{d} \sum_{r=1}^{k-1} \sum_{s=r+1}^{k} \psi_{r,s}(j)\right].$$

For a given variable x_i, the numerator in Eq. (6) represents the mean of the squared differences between the average of x_i among the groups, while the denominator represents the sum of the mean of the squared differences between the average of all x_j's among the groups. One can observe that if x_i produces homogeneous clusters, meaning that it does not significantly affect the groups, then the Eq. (6) numerator tends to be small compared to its denominator, and the variable importance index is low. On the other hand, if x_i promotes a very diverse set of clusters, then the ratio in Eq. (6) tends to be large and x_i is considered as an important variable. One can observe that $\mathcal{I}_i \in [0, 1]$ is a ratio and can be easily transformed into percentage just by multiplying by 100. Variables are then ranked by their importance index in decreasing order, allowing the identification of those variables affecting the operation the most.

The coherence of the clustering analysis with the considered set of variables was investigated, answering the methodology questioning step: Are the results useful and physically relevant? Boxplots can be useful in an inter-variable analysis, identifying both the individual and collective clusters behavior. If the clustering leads to physically absurd results or irrelevant/unlikely system configurations, or if a dominance of a subset of variables in the cluster formation was detected, the current analysis is discarded, another set of variables is considered in the variable selection step and the analysis is performed again.

The methodology **ends** by defining the target cluster accordingly with the analysis' goal. The cluster determines the operating ranges that must be set for each variable. The lower and the upper limits are defined according to the first and third quartile of the data associated with the target condition.

4. Case study

The clustering methodology was applied to the steam generator and its coal mills subset (Fig. 2), of one of the twin PECEM power plants, which is located at São Gonçalo do Amarante, State of Ceará, northwest of Brazil. Each sub-critical coal-fired power produces up to 360MW electric output independently.

The sub-critical coal-fired steam generator operates at 180 bara, 540 °C steam, comprising main steam generation, reheating, and economizers. Raw coal is pulverized by four independent mills, with three on

Table 3
Results for the 7th, 10th, 14th and 27th principal PCA components.

Cumulated Explained Variance	Principal Components	Eigenvalues
71.38%	7	1.00
80.85%	10	0.74
90.11%	14	0.55
100.00%	27	0.00

Fig. 3. Average Silhouette coefficient curve for different numbers of clusters and different numbers of PCA components.

continuous duty and one in stand by. The preheated air is split into two different streams, the primary and secondary air. The primary air feeds the mills to dry and transport pulverized coal into the steam generator burners, while the secondary sustains both the sub-stoichiometric combustion and the over-fired zone (OFA). Feedwater stream is preheated by the economizers to be then vaporized at the furnace water walls and finally superheated and delivered to the turbines, generating the power plant electrical output.

The methodology outlined in Section 3 was applied to that steam generator and mills subset. Twenty-seven variables were considered, which were all available variables from the referred processes of one generation unit for this analysis. The considered steam generator efficiency variable is based on DIN 1942. The 27 variables are presented in Appendix A. Data were taken hourly over a 14 month period, from 01 September 2018 (00:00:00) to 30 October 2019 (00:00:00), and concerned the 320 to 360 MW electric output range. That output range represented a 33% occurrence, which yielded a sample size of 3,357 initial observations for each of the 27 variables. The outlier removal procedure reduced the data set by 15%, resulting in a set of 2,846 observations. A sample size of 2,846 for 27 variables meets sample size condition for clustering problems [27].

PCA was applied after data standardization considering the full set of variables, generating 27 new orthogonal components. Some selected results are presented in Table 3. At least the 7 first principal components had to be retained since all of them displayed an eigenvalue greater or equal than 1. Sets of 10 and 14 components were also considered, while the total set of 27 variables was analyzed for comparison.

Data were clustered multiple times, considering k from 2 to 12 clusters, on the subsets of 7, 10, 14, and 27 principal components. The average Silhouette coefficient was calculated for each arrangement, as presented in Fig. 3.

The lower the number of retained PCA components, the higher was the average Silhouette coefficients, indicating more consistent clustering results. This behavior was expected due to the so-called curse of dimensionality, which refers to the fact that data sparsity rapidly increases with dimensionality. The Silhouette coefficient presented a global maximum of $k = 2$ clusters for all considered subsets. For 7 and 10 principal components, a local maximum for $k = 7$ clusters was noticeable while for 14 and 27 principal components, a local maximum was observed for $k = 10$ clusters. That observation indicates that fewer principal components may be preferable in determining the most significant changes in operation, reducing from 10 to 7 identified clusters.

The analysis proceeded considering $k \in \{2, 7\}$ from the 10 first principal components. For presentation purposes, they are illustrated considering the 3 first PCA dimensions in Fig. 4.

It is possible to notice that $k = 7$ clusters are subdivisions of $k = 2$ clusters, which is important to identify finer variation on the operation. This is considerably more interesting given the goals of the present study, since the analysis of 7 operating patterns may be more informative on the variation of the variables than an analysis of only 2 global patterns. The variable importance index was obtained for each of the 27 input variables for $k = 7$ clusters. The results are presented in Appendix A and in Fig. 5.

The variable importance index values may be subdivided into three discernible plateaux, characterized by similar index values. That finding allowed to stratify variables into three groups, according to their variable importance index value. Nine out of 27 variables presented the highest indexes, while 15 presented intermediate values, followed by the remaining three. Variables with the highest variable importance indexes were the main steam flow (kg/s), average mill airflow (kg/s), average mill air temperature ($°C$), steam generator efficiency (%), total coal flow (kg/s), primary air collector pressure (mbarg), feedwater flow (kg/s), feed water pressure (barg), and reheated steam temperature ($°C$). Their variable importance indexes represented 43.17% of the sum of the variables' importance. These variables induce sparser clusters than the other ones, which means that they may have a greater impact on the power plant operation. The variables with a lower level of importance were the average CO furnace output (ppm), hot reheated steam temperature ($°C$), and main steam pressure (barg). This means that these variables are of little help in understanding the operation since they induce very homogeneous clusters. Thus it would not be of relevance to monitoring them to understand the operation conditions. Fig. 6(a)–(i) presents the boxplots of the 9 most important variables stratified by their cluster label.

It was found no dominance of any variables in the cluster formation since any of the boxplots presented strongly isolated clusters. Regarding the system configuration, one may observe the collective behavior of the nine variables according to the physical process. For instance, three of the clusters (3, 4, and 7) presented the highest levels of steam generator efficiency, from Fig. 6(d). The same clusters presented high values of main steam flow generation (Fig. 6(a)) and the lowest total coal flow (Fig. 6(e)), which is in line with expectations for high-efficiency conditions, higher steam generation for lower fuel consumption. The complementarity of variables within the clusters shows the good representativeness of the physical process by the model.

The three clusters that presented the highest levels of steam generator efficiency, clusters 3, 4, and 7, can indicate the best-operating conditions to aim for. However, the observations from Fig. 6 were not sufficient to choose an operation regime to follow, which led to assess the historical operating time of each cluster. The operating time in each cluster is obtained from the number of datapoints associated with them. The percentage of the time that the power plant spent under each of the 7 clusters is presented in Fig. 7.

It is possible to state that the system operated along with the four operating conditions (clusters 2, 4, 6, and 7) and the predominant one (cluster 3), with occasional changes to the less frequent clusters, 1 and 5. The operation in clusters 2, 4, 6, and 7 represents each 15% of the total time. Cluster 3 was the predominant cluster, representing over 25% of the operation.

The intervals were constructed based on historical data and correspond to the interquartile range for the variables in cluster 3. The low variability is a characteristic of the data in this cluster and is a reflection of the stability in the power plant operation.

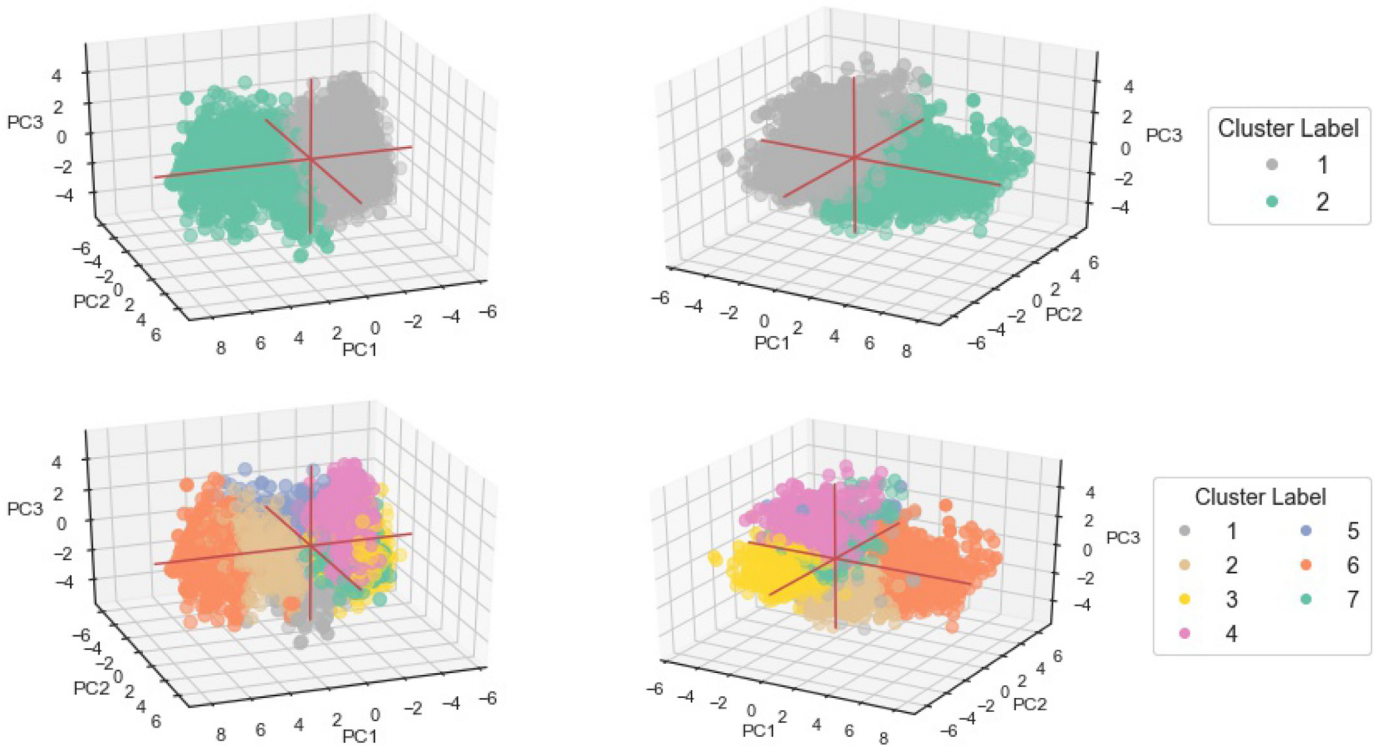


Fig. 4. Cluster results for $k = 2$ (top) and $k = 7$ (bottom), plotted considering the first 3 PCA dimensions under two different perspectives.

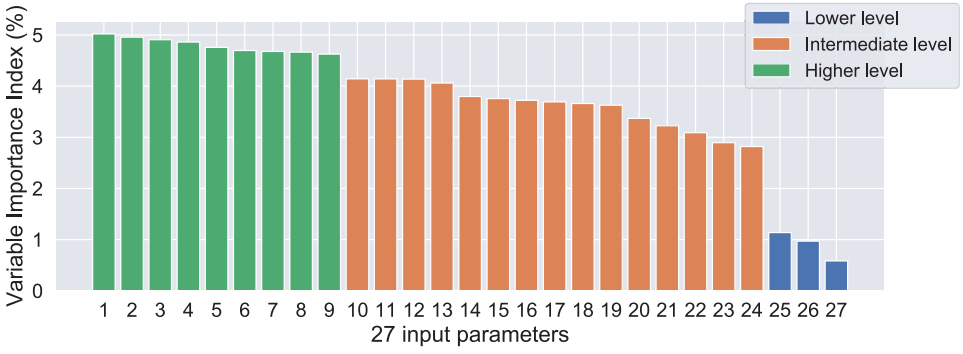


Fig. 5. Subset variable importance indexes.

Cluster 3 was defined as the target operating condition due to its highest steam generator efficiency and frequency. Based on the understanding that each of the encountered conditions has already been historically put in operation in the power plant, the target condition is assumed to be workable considering physical constraints. The target condition is expressed by operating ranges for each input variable. The upper and lower range limits were constructed based on historical data and correspond to the interquartile range for the variables in cluster 3. The results are presented for the 27 variables in Appendix A, and for the nine most important variables in Table 4. The low variability is a characteristic of the data in this cluster and is a reflection of the stability in the power plant operation.

The variables identified as the most important ones are those whose current variation influences the clusters partition. However, there are variables that are known to be highly relevant from the physical point of view and are controlled closely during the process. These well-controlled variables present stable values and therefore do not influence clustering results, with low importance index results. The variable ranking based on clustering aims to point out parameters that need more attention from operators.

Table 4
Lower and upper range limits for the nine most important variables to achieve the target operating condition of higher efficiency.

Variable	Unit	Target range	
		lower limit	upper limit
Main steam flow	t/h	1,164.17	1,175.11
Average mill air flow	kg/s	22.71	23.80
Average mill air temperature	°C	293.29	298.62
Steam generator efficiency	%	75	76
Total coal flow	t/h	132.08	134.44
Primary air collector pressure	mbarg	75.92	76.71
Feedwater flow	t/h	1,129.85	1,145.25
Feedwater pressure	barg	198.05	198.78
Steam to be reheated temperature	°C	325.86	327.73

The steam-generator efficiency increase for an operation in the presented target condition is determined from the historical operation values. From historical data, the average steam-generator efficiency for the 14-month period was of 73.5%, and the cluster of lowest efficiency,

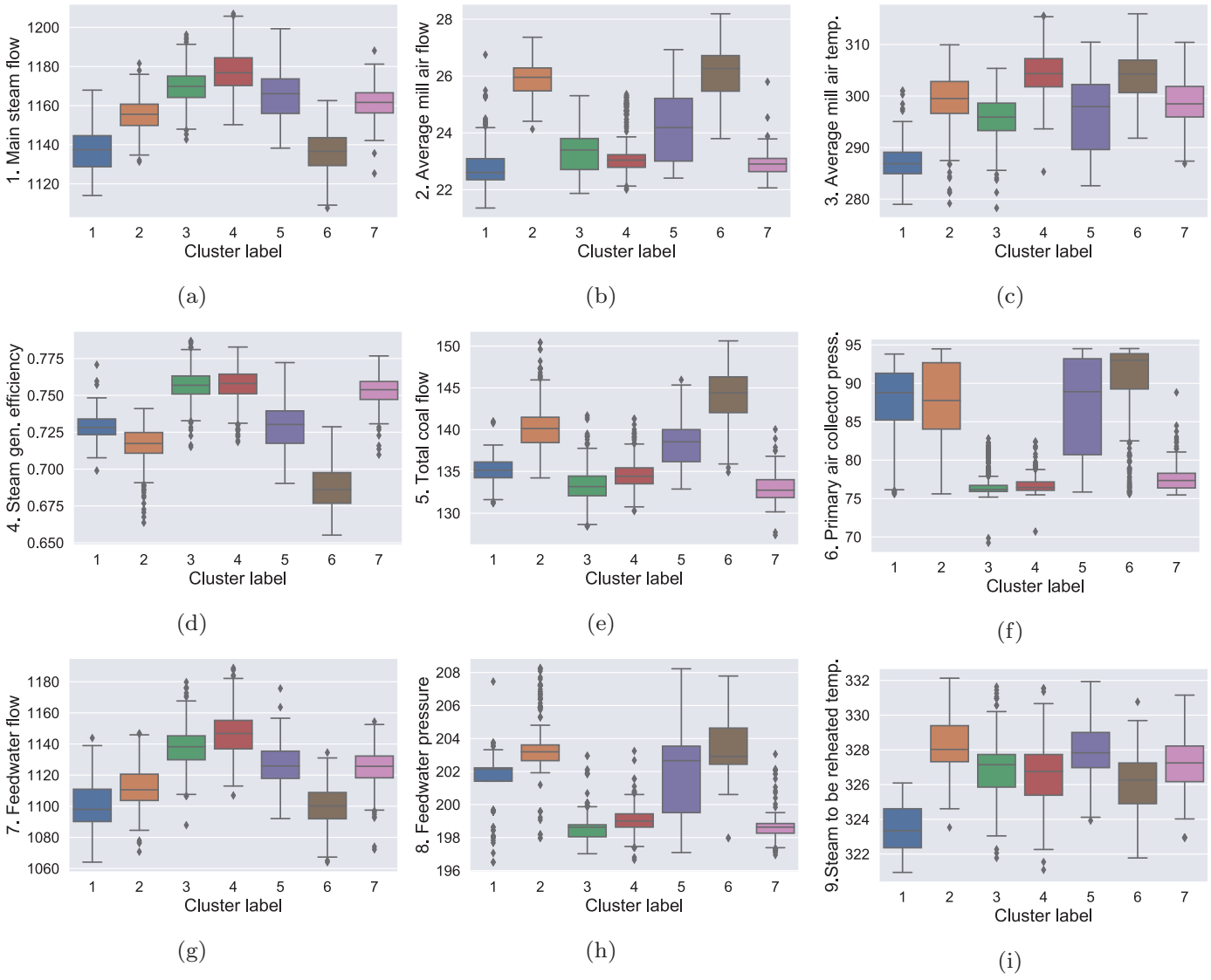


Fig. 6. Boxplots representing each of the 7 clusters for the respective variable.

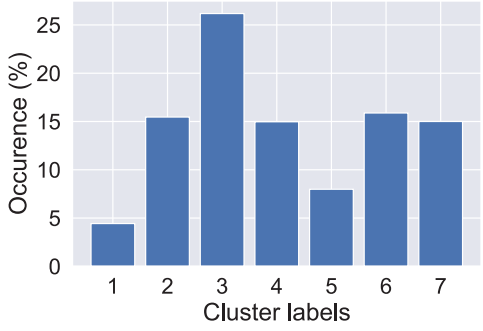


Fig. 7. Percentage of time that the power plant spent under each of the 7 clusters.

cluster 6, resulted in 68% steam-generator efficiency. From Fig. 6(d), an operation set within the presented target ranges would achieve a steam-generator efficiency of approximately 76%. For instance, if the operator maintain the target condition for 80% of the time, efficiency and resources would be optimized. For the analyzed 14 month period, coal consumption would have been reduced up to 8,157 tonnes in the 360MW operating load.

Considering the formulated research question in Section 1: Are machine learning-based methods able to map operating conditions and related variables that guarantee efficient plant operation? This machine learning-based methodology defined operating conditions, allowing the identification of a target operating condition supporting higher steam generator efficiencies. It provides the power plant operator the information of which values should be set and which variables should be monitored closely.

5. Conclusions

This paper applied unsupervised machine learning methods to recognize different operating conditions at a 360MW thermal power plant based on one year of historical data, identifying the configuration that yields the most efficient operation to support operating practices. The methodology based on principal component analysis (PCA) and k-means++ clustering method was implemented to 27 selected operation variables. A partition of the data in 7 clusters was selected for analysis since it recognizes fine variations in operation without overestimating nonessential data. The input variables were evaluated by their degree of impact on the operation, measured by a proposed variable importance index. This evaluation allows for an identification of the variables that need more attention for a high-efficiency operation, since their variabil-

ity impacts the overall process. Nine of the 27 variables were identified as the most relevant to the case study. These variables enabled the identification of the condition that maintains a high steam generator efficiency operation. An operation within the proposed arrangements would achieve typical steam generator efficiencies of 76%, which show performance gains in respect to the average historical values of 73.5% and the lowest efficiency condition records of 68%. This increase in steam-generator efficiency represents a savings of 8,157 tonnes in the 360MW operating load during a year. The employment of the present methodology guarantees high efficiencies within conservative variable ranges, promoting a periodical revision of the operating setpoints. The associated range of operation for each variable is presented, along with the indication of the most relevant variables to monitor.

Acknowledgments

Authors acknowledge Energy of Portugal EDP for the financial and technical support to this project; J. Duarte acknowledges the financial support from CNPq 154147/2020-6 for her undergraduate scholarship; L.W.Vieira acknowledges the INCT-GD and the financial support from CAPES 23038.000776/2017-54 for her Ph.D. grant; A.D.Marques acknowledges the financial support from CNPq 132422/2020-4 for his MSc grant; P.S. Schneider acknowledges CNPq for his research grant (PQ 301619/2019-0). T.S. Prass acknowledges the support of FAPERGS (ARD 01/2017, Processo 17/2551-0000826-0).

Appendix A

The complete list of selected variables from the PECÉM power plant is presented along with its measure unit. Agreeing to the Memorandum of Understanding (MOU) between the researchers and EDP PECEM, the complete one-year-long data set should not be provided.

The third column presents the resulting variable importance index. The two last columns indicate the target condition operating ranges, defined by the results from cluster 3, represented by the lower and upper limits to be maintained for each variable.

Code	Variable	Unit	Importance	Target range	
			Index (%)	lower limit	upper limit
1	Main steam flow	t/h	5.02	1,164.17	1,175.11
2	Average mill air flow	kg/s	4.96	22.71	23.80
3	Average mill air temperature	°C	4.91	293.29	298.62
4	Steam generator efficiency*	%	4.86	0.75	0.76
5	Total coal flow	t/h	4.76	132.08	134.44
6	Primary air collector pressure	mbarg	4.70	75.92	76.71
7	Feedwater flow	t/h	4.68	1,129.85	1,145.25
8	Feedwater pressure	barg	4.66	198.05	198.78
9	Steam to be reheated temperature	°C	4.63	325.86	327.73
10	Hot reheated steam pressure	barg	4.14	31.97	32.21
11	Stoichiometric ratio	%	4.14	0.80	0.80
12	Average O2 excess	%	4.14	1.99	2.45
13	Steam to be reheated pressure	bara	4.06	36.06	36.32
14	Mill A dynamic classifier speed	rpm	3.80	100.46	105.42
15	Feedwater temperature	°C	3.76	270.51	271.97
16	Total of primary and secondary air flow	kg/s	3.72	341.83	350.01
17	Average heated air temperature	°C	3.69	333.73	338.01
18	Power generation	MW	3.66	355.04	357.10
19	Average coal temperature	°C	3.63	76.47	78.78
20	Secondary air flow	kg/s	3.37	66.34	69.43
21	Average furnace combustion gas temperature	°C	3.23	345.02	350.13
22	Secondary air collector pressure	mbarg	3.09	15.96	17.24
23	Drum water temperature	°C	2.90	356.94	358.31
24	Main steam temperature	°C	2.82	536.87	539.36
25	Average CO furnace output	ppm	1.14	2.00	3.00
26	Hot reheated steam temperature	°C	0.97	537.60	542.14
27	Main steam pressure	barg	0.58	167.29	167.77

* Steam generator efficiency based on DIN 1942.

References

[1] IEA. World energy outlook 2018 executive summary. International Energy Agency; 2018.
[2] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer New York; 2013. doi:101007/978-1-4614-7138-7.
[3] h Kim K, seok Lee H, h Kim J, Ho Park J. Detection of boiler tube leakage fault in a thermal power plant using machine learning based data mining technique. In: 2019 IEEE International conference on industrial technology (ICIT); 2019. p. 1006–10. doi:10.1109/ICIT.2019.8755058.
[4] Peng B, Xia H, Ma X, Zhu S, Wang Z, Zhang J. A mixed intelligent condition monitoring method for nuclear power plant. Anna Nucl Energy 2020;140:107307. doi:10.1016/j.anucene.2020.107307.
[5] Tingting Y, Yang B, Weichun G, Huanhuan L, Guiping Z, You L. Early warning method for power station auxiliary failure considering large-scale operating conditions. In: 2019 IEEE 2nd International conference on power and energy applications (ICPEA); 2019. p. 11–16. doi:10.1109/ICPEA.2019.8818537.
[6] Yu J, Jang J, Yoo J, Park JH, Kim S. A clustering-based fault detection method for steam boiler tube in thermal power plant. J Electr Eng Technol 2016;11(4):848–59.
[7] Xu J, Bi D, Ma S, Bai J. A data-based approach for benchmark interval determination with varying operating conditions in the coal-fired power unit. Energy 2020;211:118555.
[8] Choi SW, Yoo CK, Lee I-B. Overall statistical monitoring of static and dynamic patterns. Ind Eng Chem Res 2003;42(1):108–17.
[9] Yanqiao C, Liheng L, Hainan C. Research and designing of boiler-turbine model linearization balance operating point based on energy-saving and non-linear measurement analysis. In: IECON 2012-38th Annual conference on ieee industrial electronics society. IEEE; 2012. p. 1145–9.
[10] Li X, Liu Q, Wang K, Wang F, Cui G, Li Y. Intelligent partition of operating condition-based multi-model control in flue gas desulfurization. IEEE Access 2020;8:149301–15.
[11] Barbasova T, Filimonova A, Zakharov A. Energy-saving oriented approach based on model predictive control system. In: International Russian automation conference. Springer; 2019. p. 243–52.
[12] Jiang R, Wang Y, Yan X. Density clustering analysis of fuzzy neural network initialization for grinding capability prediction of power plant ball mill. Multimed Tools Appl 2016;76(17):18137–51. doi:10.1007/s11042-016-4089-4.
[13] Xiaoying H, Jingcheng W, Langwen Z, Bohui W. Data-driven modelling and fuzzy multiple-model predictive control of oxygen content in coal-fired power plant. Trans Inst MeasControl 2017;39(11):1631–42.
[14] Hou G, Gong L, Huang C, Zhang J. Novel fuzzy modeling and energy-saving predictive control of coordinated control system in 1000 MW ultra-supercritical unit. ISA Trans 2019;86:48–61.
[15] Wu X, Shen J, Li Y, Lee KY. Data-driven modeling and predictive control for boiler–turbine unit using fuzzy clustering and subspace methods. ISA Trans 2014;53(3):699–708.
[16] Schang L, Wang S. Application of the principal component analysis cluster analysis in comprehensive evaluation thermal power. units 2015:0–4.
[17] Liukkonen M, Hiltunen T, Hälikkä E, Hiltunen Y. Modeling of the fluidized bed combustion process and NOx emissions using self-organizing maps: an application to the diagnosis of process states. Environ Modell Softw 2011;26(5):605–14. doi:10.1016/j.envsoft.2010.12.002.
[18] Chengbing H, Gang L, Dakang S, Shunka C. An optimization study on energy consumption benchmarks for boilers and their auxiliary systems.. Int J Simul-Syst Sci Technol 2016;17(47).

[19] Jia J, Lu Y, Chu J, Su H. The application of fuzzy pattern fusion based on competitive agglomeration in coal-fired boiler operation optimization. In: International conference in swarm intelligence. Springer; 2015. p. 76–83.
[20] Liu S, Sun L, Zhu S, Li J, Chen X, Zhong W. Operation strategy optimization of desulfurization system based on data mining. Appl Math Modell 2020;81:144–58.
[21] Liukkonen M, Heikkinen M, Hiltunen T, Hälikkä E, Kuivalainen R, Hiltunen Y. Artificial neural networks for analysis of process states in fluidized bed combustion. Energy 2011;36(1):339–47. doi:10.1016/j.energy.2010.10.033.
[22] Wang H, Jia L. Big data knowledge mining based operation parameters optimization of thermal power. In: 2019 IEEE 8th Data driven control and learning systems conference (DDCLS). IEEE; 2019. p. 338–43.
[23] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.
[24] Arthur D, Vassilvitskii S. *k*-means + +: the advantages of careful seeding. Tech. Rep.. Stanford; 2006.
[25] Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. O'Reilly Media; 2019.
[26] Igual L, Segui S. Introduction to data science: a python approach to concepts, techniques and applications. Barcelona: Springer; 2017.
[27] Dolnicar S, Grün B, Leisch F, Schmidt K. Required sample sizes for data-driven market segmentation analyses in tourism. J Travel Res 2013;53(3):296–306. doi:10.1177/0047287513496475.