XLI CILAMCE
XLI Ibero-Latin American Congress on Computational Methods in Engineering
16-19 November, 2020 FOZ DO IGUAÇU PARANÁ - BRAZIL

DATA-DRIVEN IDENTIFICATION OF OPERATING PATTERNS IN A THERMAL POWER PLANT BY CLUSTERING METHODS

Jéssica Duarte[1], Lara Werncke Vieira[1], Augusto Delavald Marques[1], Paulo Smith Schneider[1], Guilherme Lacerda Batista de Oliveira[2]

[1]*Department of Mechanical Engineering, Federal University of Rio Grande do Sul*
R. Sarmento Leite, 425 - Centro Histórico, Porto Alegre, 90050-170, Rio Grande do Sul, Brazil
jessica.jd.duarte@gmail.com,lara.vieira@ufrgs.br,augustod.marques@gmail.com,pss@mecanica.ufrgs.br
[2]*Energy of Portugal - EDP*
Complexo Industrial e Portuário do Pecém (CIPP) – São Gonçalo do Amarante, Ceará, Brazil
guilherme.oliveira@edpenergiapecem.com.br

Abstract. Thermal power plant operation depends on the knowledge of a wide range of complex and cross dependent parameters. Information is usually captured through Distributed Control Systems (DCS) which allow to access up to date data but also long periods of recorded operation. Large and available data sets are decisive for plant operation, but they must be properly used and interpreted to achieve effectiveness. The purpose of the present paper is to present an identification of operational patterns from historical data from an actual thermal power plant based on unsupervised machine learning methods. The proposed methodology is applied to a long term data series from the 360 MW Brazilian coal-fired Pecem power plant, for 29 selected parameters, concerning its steam generator and associated mills. Dataset size and redundancy is treated by the Principal Component Analysis (PCA) approach, which defines a lower dimensional space, proper for clustering while preserving most of its variance. The K-means clustering method identifies operating point groups according to their degree of similarity. The appropriate cluster number is defined by means of the average silhouette coefficient, which measures the clusters consistency. Cluster parameter values and distribution are evaluated to verify result consistency. The assessment with the 29 parameters from the steam generator and mills system is presented, and the results show that the operation may be described globally by a 2 clusters analysis or, for refined observations, by a 10 clusters analysis. The different patterns encountered facilitate an understanding of the parameters arrangement and resulting performance, enabling the identification of low efficiency operation conditions and supporting practices to improve the plants operation.

Keywords: Power plant operation, Operation patterns, Operation parameters, K-means clustering, PCA.

1 Introduction

The future of energy generation is one of today's most important and complex challenges. Different scenarios on the future of energy were developped by [1], and an increase in energy demand is expected to grow by more than 25% until 2040. Thermal power plants are nowadays the main source of global energy generation, demanding research on an efficient operation.

The comprehension of its optimal operation could improve energy consumption and reduce harmful emissions. Data flow from the operating plants is accessed by a Distributed Control System (DCS), which records long periods of operation and generates an important amount of data. It must be properly interpreted to result in effective information.

Unsupervised machine learning methods are vastly applied for pattern identification in power plants processes. Kuriak *et al.* applied k-means clustering method to generate control signatures for real-time optimization of the combustion process [3], , being able to lead to operation control settings to optimize boiler efficiency. Hou *et al.* applies Fuzzy c-mean (FCM) for performance improvement of ultra-supercritical power plants by real-time running data [5]. The optimization of a desulfurization system operation [6] was also analysed by the FCM method, which was able determining the optimal parameter operation settings. Xiaoying *et al.* established by FCM a real-time predictive control of oxygen content in coal-fired power plants, combining it with a subspace

method, building a combustion process model that was verified to achieve an acceptable accuracy [7]. Wang *et al.* also modelled thermal power units by k-means clustering method with Spark-based FP-growth algorithm [8], to mine optimization targets values to improve its economical index.

This paper aims to recognize the different patterns that may occur on a thermal power plant operation, based on historical data from operation and steam generator parameters. The K-means clustering technique is applied to identify groups of datapoints within the data, with the support of Principal Component Analysis (PCA) to reduce the database dimensionality and of the silhouette coefficient to determine the appropriate number of clusters. The definitive cluster configuration is presented to enrich the comprehension of the operation patterns.

The case studied is the PECEM power plant, located at São Gonçalo do Amarante, in Ceará, which is responsible for 50% of the energy generation complex of the state. The power plant is composed by two independent sub-critical coal-fired power generating units, working in two different electrical regimes of 240MW and 360MW [EDP]. The methodology is applied to one generation unit at the 360MW electrical regime.

2 Applied Methods

Principal Component Analysis (PCA) is a statistical technique for data reduction with axis rotation [11]. The dataset is represented with fewer components, increasing its interpretability is a way to minimize information loss. The resulting components are sorted from the highest to the lowest variance representation, and a subset of the first number of components is selected to be the new dimensional space.

Clustering methods are unsupervised machine learning techniques aiming to divide unlabeled data into homogenous subgroups [12]. The K-means method uses the Euclidian distance metric, to assign each item 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 Silhouette Coefficient is an internal evaluation of the clustering consistency result and is considered in this paper to determine the more adequate number of clusters for a dataset. This cluster validation criteria considers the intracluster proximity and the intercluster separation. It ranges from -1 to 1, as a high positive value represents a compact cluster.

3 Methodology

The present work employed a methodology framework composed of 11 steps, that can be observed in Fig. 1.

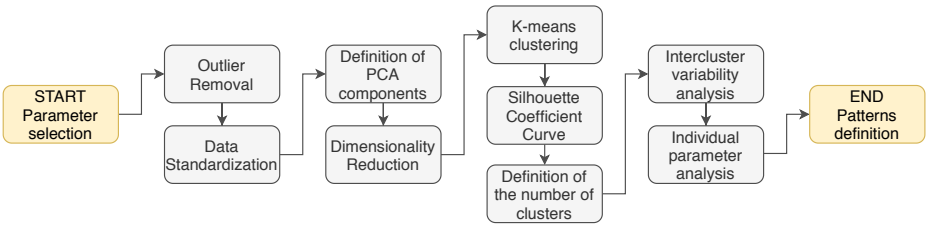

Figure 1. Proposed methodology framework.

The first step is the parameter selection. The analysis should contemplate process elements which may be relevant on the characterization of the operation to be studied, considering the available data.

The data is preprocessed following the parameter selection is the data preprocessin, by outlier removal and data standardization. For the data cleaning, the data point is identified as an outlier if any of its parameter values is three standard deviations away from the mean. Data standardization is then the conversion of the different data structures into a common data format.

The Principal Component Analysis defines new components for dimensionality reduction. The cumulative original variance for each component is obtained. The minimal number of PCA components for a defined cumulative variance is selected to be used in the analysis.

The K-means clustering starts with the application of the method to the dimensionally reduced data. The data is clustered multiple times for different numbers of clusters, in order to select the best fitted number in the next step. The average silhouette coefficient is calculated for each set of the multiple clustering from the previous phase. An average silhouette coefficient curve is plotted with the objective of evaluating the clustering consistency results, and the higher average silhouette coefficient values are identified. The corresponding number of clusters is adopted.

The results assessment starts with the the intercluster variability analysis step for the clustering results. For each parameter, boxplots for the different clusters are generated and compared. A visual inspection of these plots

may indicate the intercluster behavior of the selected parameters and, also, the coordination between different parameters. The following step is the individual parameter analysis, to inspect specific parameter behaviors.

Finally, the results are analysed for its usefulness and physical relevance in order to obtain process information. A comprehension of the process is required for such evaluation. It is then possible to define the power plant operation patterns.

4 Results and Discussion

The presented method is applied to identify operation patterns from 29 selected parameters of one generation unit at PECEM power plant. The selected parameters cover the mills and the steam generator processes. An extensive list of the parameters is presented in App. 1. The data is taken at one hour intervals over a 14 month period, from the beggining of September 2018 to the end of October 2019. The analysis was done to the plant's designed power generation, which is the 360MW operating load. The generation unit operates under this condition 33% of the time over the considered period, resulting in 3.357 samples.

The original dataset was reduced 15% by applying the outlier removal procedure. The dimensionality reduction by PCA is applied after the dataset standardization. The cumulative variance obtained by the selected number of components is presented in Tab. 1.

Table 1. PCA cumulative variance result for the case study.

Cumulative Variance	Number of Components
100.00%	29
90.41%	15
82.01%	11
62.35%	5

The number of components is here defined to be 11, since it represents 82,01% of the variance while substantially reducing the original 29 components dimensionality.

The data is clustered multiple times, from 2 to 13 clusters, which is done for each number of components considered at the PCA reduction. To better understand the dimensionality results and to verify that the definition of 11 components is adequate, this is done for each considered number of components PCA reduction in Tab. 1, from 2 to 29 components, each one presenting a different curve. The average silhouette coefficient is determined for each clustering result, and Fig. 2 presents its results.

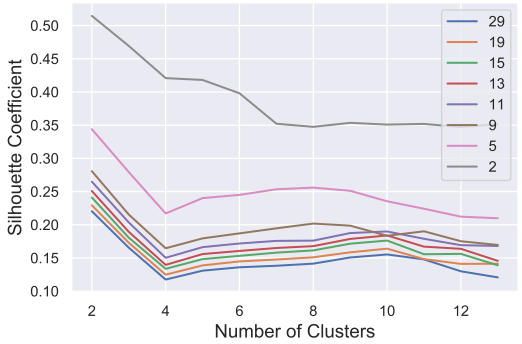

Figure 2. The average silhouette coefficient from different number of clusters for each number of PCA components for analysis B.

It can be observed that the lower the number of components, the higher is the average silhouette coefficient, indicating more consistent clustering results. The data sparcity increases rapidly with the volume of high-dimensional spaces. On the other hand, as presented in Tab. 1, a low number of components wouldn't represent adequately the original dataset variance and the shape of the curve and its results are affected. The defined number of 11 PCA components sufficiently conserves the higher-dimentional average silhouette coefficient curve shape. Moreover, the higher average silhouette coefficient of 0.26 is observed occurring for 2 clusters and a local maximum average silhouette coefficient of 0.19 is observed for 10 clusters, whose analysis may distinguish subtle

operation patterns than for 2 clusters. Figure 3 and 4 present the clustering results with different colors for the 2 and the 10 clustering analysis, respectively.

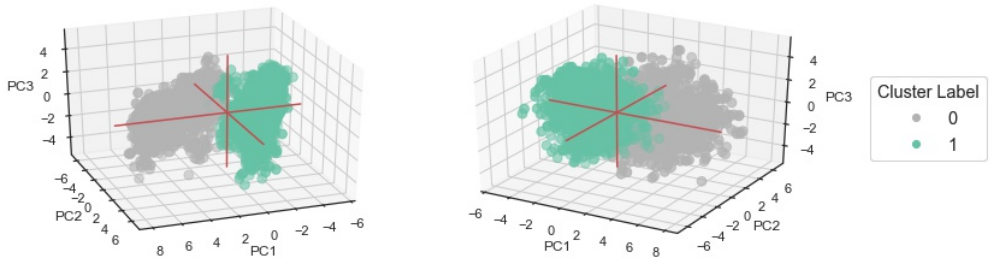

Figure 3. Cloud of datapoints representing the 2 clusters on the first 3 PCA components.

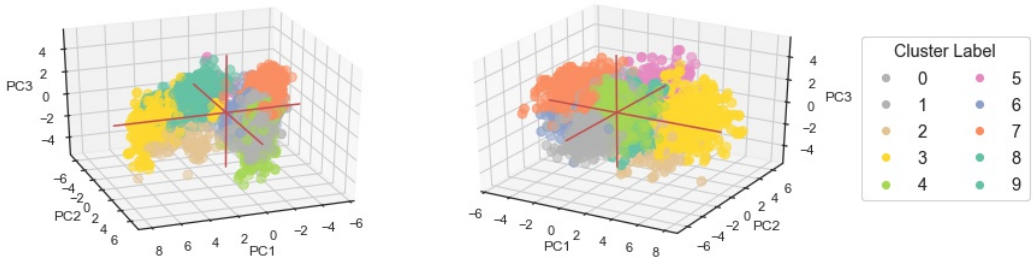

Figure 4. Cloud of datapoints representing the 10 clusters on the first 3 PCA components.

Its important to remark that, comparing these figures, the 10 different clusters from the Fig. 4 are subdivisions from the two clusters from Fig. 3. Table 2 presents these subdivision relations.

Table 2. Subdivision from the 2 cluster analysis to the 10 cluster analysis.

2 clusters	**Subdivision in 10 clusters**
0	2, 3, 5, 8 and 9
1	0, 1, 4, 6 and 7

Accordingly, the results assessment is performed, examining the clustering results. For each of the 29 parameters, the 10 clusters boxplots are visually inspected. The observations for a selection of the parameters is presented at Fig. 5, for an analysis of the cluster's different operation performances. The boxplots colors represent the correspondancy to the 2 clusters as presented in Tab. 2, as the clusters presented in grey boxplots are subdivisions of cluster 0 and the clusters presented in green boxplots, the subdivisions of cluster 1.

The clusters subdivisions of cluster 0, presented in grey boxplots, operate with a lower average steam generator efficiency than the clusters subdivisions of cluster 1, presented in green boxplots.

From the 10 clusters, cluster 3 presents the lower steam generator efficiency (P11), which is coherent due to its high total coal flow consumption (P12) and low power generation (P13). Clusters 2 and 9 are also characterized by lower efficiency, while clusters 0, 4 and 7 present the highest steam-generator efficiency values (P11). The main steam flow (P14) boxplots patterns are similar to the steam generator efficiency (P11), which is coherent since the main steam flow transports energy from the steam generator to the turbines.

The time the power plant has operated at each pattern (cluster) over the considered period is obtained. Table 3 presents the percentage of the time operated at each of the 10 operation patterns from the encountered clusters.

Table 3. Percentage of the time operated at each of the 10 clusters.

Cluster	0	1	2	3	4	5	6	7	8	9
Occurance	8.41%	12.77%	7.88%	10.87%	7.84%	3.80%	9.78%	19.59%	2.67%	16.39%

It is possible to observe that there are patterns at which the power plant operates for short periods of time, by up to 10% of the considered condition, and others that operate for relatively longer periods, such as cluster 7.

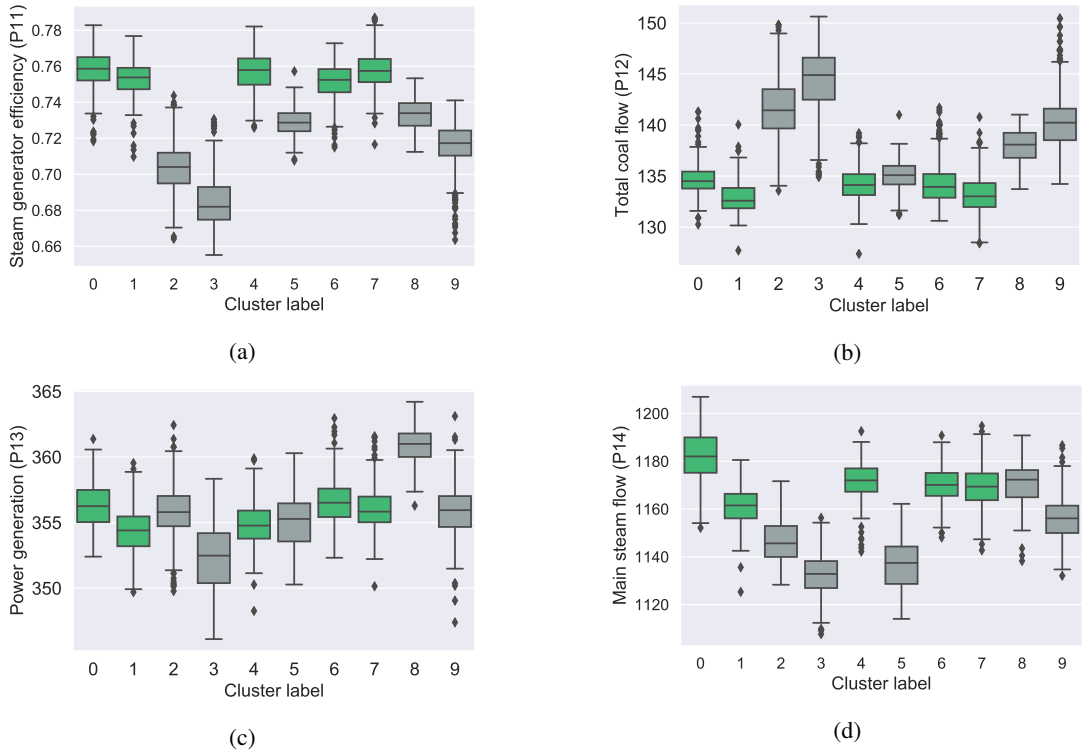

Figure 5. Boxplots representing each of the 10 clusters for the selected parameters: (a) steam generator efficiency (P11), (b) total coal flow (P12), (c) power generation (P13) and (d) main steam flow (P14).

Regarding the timeline of the operation conditions, the order at which the operation alternates between the different clusters is analysed. The clustering in two clusters has essentially divided the operation in two periods, finding the difference that operation has gone through along the analysed year. Cluster 0 is observed predominant from November 2018 until May 2019, and cluster 1 is predominant from May 2019 until October 2019.

Meanwhile, the operation alternates between the different patterns encountered at the 10 clusters. The operation alternates between clusters 2, 3, 5 and 9 from September 2018 until May 2019, and, from May until October 2019, the operation alternates between clusters 0, 1, 4, 6 and 7. These informations along with the time the power plant has operated at each cluster from Tab. 3, its noteworthy that the operation may be characterized by one predominant condition for a determined period of time, such as operating predominantly at cluster 9 until May 2019 or at cluster 7 from May 2019 until October 2019.

With the understanding that the operation conditions from cluster 7 pattern is one of the predominant pattern and that it presents a high steam-generator efficiency average, the operation ranges for each parameter is presented. Table 4 presents the lower and upper range limit for all considered 29 parameters for cluster 7, considering the first and third quartiles, whose parameters details are presented at Appendix 1.

Table 4. Lower and upper limit for each parameter operation ranges for cluster 7 pattern.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**
Lower	102.20	66.62	0.80	1.96	15.94	75.89	2.00	344.50	13.82	32.95
Upper	105.39	69.54	0.80	2.37	17.30	76.42	3.00	349.59	14.44	33.56
	P11	**P12**	**P13**	**P14**	**P15**	**P16**	**P17**	**P18**	**P19**	**P20**
Lower	0.75	131.97	355.03	1163.71	167.33	536.57	36.03	325.59	1129.30	198.12
Upper	0.76	134.31	356.98	1174.90	167.82	538.48	36.30	327.73	1144.77	198.93
	P21	**P22**	**P23**	**P24**	**P25**	**P26**	**P27**	**P28**	**P29**	
Lower	270.41	31.93	537.30	357.98	76.32	293.28	22.64	341.85	333.00	
Upper	271.97	32.19	541.85	358.31	78.64	298.55	23.86	348.99	337.30	

Conditions with higher steam generator efficiency are desirable for a profitable operation. The operational ranges described in Tab. 4 provide preferable arrangements to assure the higher steam generator efficienciesThese conditions were already attained for 19.59% of the analysed period. Further analysis should evaluate the reasons why this condition is not kept and support practices to improve the power plant operation.

5 Conclusions

This paper applies unsupervised machine learning methods to recognize different operation patterns at a 360 MW thermal power plant based on one year historical data, from September 2018 to October 2019. The methodology based at the principal component analysis and K-means clustering methods was implemented to 29 selected operation parameters. It was identified a partition in 2 and a subdivision in 10 clusters. The 2 clusters have identified two period-based patterns dividing the analysed year operation. The first cluster represents the operation until May 2019, while the second cluster represents the operation from May until October 2019 and presents higher average efficiencies, up to 8% higher than clusters from the first period. The analysis of its subdivisions in 10 different clusters identified different patterns that alternate along the operation, exposing the behavior of the analysed parameters and enabling the identification of conditions for maintaining a high steam generator efficiency operation. The defined operation ranges for each parameter for such condition is obtained. The understanding of these configurations and its resulting performance on the plant may support practices to improve its operation and to have an efficient decision making in the plant.

Acknowledgements. Authors acknowledge Energy of Portugal EDP for the financial and technical support to this project; Duarte acknowledges the financial support from CNPq 154147/2020-6 for her undergraduate scholarship; Vieira acknowledges the INCT-GD and the financial support from CAPES 23038.000776/2017-54 for her PhD grant; Marques acknowledges the financial support from CNPq 132422/2020-4 for his MSc grant; Smith Schneider acknowledges CNPq for his research grant (PQ 305357/2013-1).

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] IEA, 2018. *World Energy Outlook 2018 Executive Summary*. International Energy Agency.
[2] Barbasova, T. A., Filimonova, A. A., & Zakharov, A. V., 2020. Energy-saving oriented approach based on model predictive control system. In *Lecture Notes in Electrical Engineering*, pp. 243–252. Springer International Publishing.
[3] Kusiak, A. & Song, Z., 2006. Combustion efficiency optimization and virtual testing: A data-mining approach. *IEEE Transactions on Industrial Informatics*, vol. 2, n. 3, pp. 176–184.
[4] Song, Z. & Kusiak, A., 2007. Constraint-based control of boiler efficiency: A data-mining approach. *IEEE Transactions on Industrial Informatics*, vol. 3, n. 1, pp. 73–83.
[5] Hou, G., Gong, L., Huang, C., & Zhang, J., 2019. Novel fuzzy modeling and energy-saving predictive control of coordinated control system in 1000 MW ultra-supercritical unit. *ISA Transactions*, vol. 86, pp. 48–61.
[6] Liu, S., Sun, L., Zhu, S., Li, J., Chen, X., & Zhong, W., 2020. Operation strategy optimization of desulfurization system based on data mining. *Applied Mathematical Modelling*, vol. 81, pp. 144–158.
[7] Xiaoying, H., Jingcheng, W., Langwen, Z., & Bohui, W., 2016. Data-driven modelling and fuzzy multiple-model predictive control of oxygen content in coal-fired power plant. *Transactions of the Institute of Measurement and Control*, vol. 39, n. 11, pp. 1631–1642.
[8] Wang, H. & Jia, L., 2019. Big data knowledge mining based operation parameters optimization of thermal power. In *2019 IEEE 8th Data Driven Control and Learning Systems Conference (DDCLS)*. IEEE.
[9] Igual, L. & Seguí, S., 2017. *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applicaitons*. Springer, Barcelona.
[EDP] EDP. Ute porto do pecém. `https://brasil.edp.com/pt-br/ute-pecem`. Accessed: 2020-08-10.
[11] Aggarwal, C. C., 2015. *Data Mining*. Springer International Publishing.
[12] Kubat, M., 2017. *An Introduction to Machine Learning*. Springer International Publishing.

1 APPENDIX

The extensive operation parameter list presents the 29 selected parameter names and measure unit. The last two columns indicate the lower and upper limits for the operating range for cluster 7.

Table 5. Extensive operation parameter list

Name	Operation Parameter	Unit	Lower limit - cluster 7	Upper limit - cluster 7
P1	Mill A dynamic classifier speed	rpm	102.20	105.39
P2	Secondary air flow	kg/s	66.62	69.54
P3	Stoichiometric ratio	-	0.80	0.80
P4	Average O2 excess	%	1.96	2.37
P5	Secondary air collector pressure	mbar	15.94	17.30
P6	Primary air collector pressure	mbar	75.89	76.42
P7	Average CO furnace output	ppm	2.00	3.00
P8	Average furnace combustion gas temperature	°C	344.50	349.59
P9	Intern consumption A	MW	13.82	14.44
P10	Intern consumption B	MW	32.95	33.56
P11	Steam generator efficiency	-	0.75	0.76
P12	Total coal flow	t/h	131.97	134.31
P13	Power generation	MW	355.03	356.98
P14	Main steam flow	t/h	1163.71	1174.90
P15	Main steam pressure	barg	167.33	167.82
P16	Main steam temperature	°C	536.57	538.48
P17	Steam to be reheated pressure	bar_a	36.03	36.30
P18	Steam to be reheated temperature	°C	325.59	327.73
P19	Feedwater flow	t/h	1129.30	1144.77
P20	Feedwater pressure	barg	198.12	198.93
P21	Feedwater temperature	°C	270.41	271.97
P22	Hot reheated steam pressure	barg	31.93	32.19
P23	Hot reheated steam temperature	°C	537.30	541.85
P24	Drum water temperature	°C	357.98	358.31
P25	Average coal temperature	°C	76.32	78.64
P26	Average mill air temperature	°C	293.28	298.55
P27	Average mill air flow	kg/s	22.64	23.86
P28	Total of primary and secondary air flow	kg/s	341.85	348.99
P29	Average heated air temperature	°C	333.00	337.30

$*$ short for Heat Exchanger