Data Analysis for Early Fault Detection

On-line monitoring approach for heat exchanger in solar-thermal plant

Javier M. Mora-Merchan, Enrique Personal, Antonio Parejo, Antonio García, Carlos León
Electronic Technology Department
Escuela Politécnica Superior, University of Seville,
Seville, Spain
e-mail: jmmora@us.es

***Abstract*—Nowadays, the reliability and robustness levels required for production systems are growing. These demands make philosophies, such as predictive maintenance essential in modern industry, among which the energy sector stands out. In this sense, this paper proposes an on-line monitoring system based on data mining models, which provides a useful tool to identify operation anomalies easily, being able to identify and prevent possible future failures. This approach has been applied successfully in a real case, where a performance analysis for the cooling systems of a solar-thermal power plant was implemented.**

Keywords-Predictive maintenance; fault detection; solar plant; data analysis.

I. Introduction

Nowadays, it is easy to see that the presence of continuous and more critical processes is growing in the industrial ambience. As their names suggest, these types of processes must run uninterruptedly and demand more in their reliability levels. Obviously, these demands require special treatment beyond a reactive or preventive maintenance. These needs are directly translated into an increment on the number of on-line monitoring or analysis systems, following what is known as a Predictive Maintenance (PdM) [1] philosophy. Specifically, PdM consists of a defect inspection strategy to prevent future problems using data analysis and identifying indicators for its detection.

The use of time series analysis to improve a process reliability is not a new approach [2][3]. However, the sensor number in processes, as well as their analytical capacity have grown systematically in the last years. An example of this evolution can be clearly seen in the energy industry, where machine learning and data mining analysis approaches are applied as useful tools to improve this service.

Specifically, these techniques have been applied by energy utilities at different levels, as can be seen in [4], where their authors propose methodologies based on data mining analysis, for maintaining the elements of smart grid distribution networks (cables, joints, manholes, and transformers), forecasting their failure probability. Specifically for cables isolation analysis, we can find [5], which proposes a partial discharge analysis, using the combination of wavelet packet transform analysis and a probabilistic neural network. Related to power transformers maintenance, [6] proposes a smart fault diagnostic approach combining five well-known methods based on dissolved gas studies, using several Artificial Neural Networks (ANNs) for their individual classification analysis and one more for the combination of their results.

Nevertheless, it is in the maintenance analysis of power generation systems where the application of intelligent approaches is more present in the last few years. As an example, [7] introduces an analysis framework for maintenance management of wind turbines, based on the characterization of correct operation settings, using ANNs. This framework directly obtains the information from the Supervisory Control and Data Acquisition (SCADA) system, combining it with an alarm and warning analysis. This approach allows the system to model the normal behavior of the gearbox bearing temperature. This estimation makes it possible to forecast possible damage in a gearbox earlier than traditional vibration-based approaches [8].

As can be observed above in the previously cited articles, most of the efforts are focused on the prediction of failures in the gearbox of wind generators that are the typical elements with greater cost and difficult substitution in them. However, there exist a multitude of elements in a generation plant, and without which its target could not be carried out either. This is what happens with the fluid condenser and cooling tower, essential in the refrigeration system of a solar-thermal power plant, and whose operation must also be monitored.

In this sense, this paper describes a Condition Monitoring System (CMS) to identify anomalies in the operation of both subsystems. For this, different approaches based on data mining have been evaluated to implement their operation models, determining the best option and validating their use for this purpose.

Specifically, the presented paper has been divided as follows; Section II describes the different necessary stages to make up a process model (data filtering, main variables identification, modeling technique election, etc.). After this, an on-line monitoring approach based on these models is described in Section III. Once the modeling and monitoring approaches have been described, Section IV performs a real application of them over some elements of a solar-thermal power plant. Finally, Section V lays out the conclusion of this work.

II. Process Model Generation

As can be seen in next sections, the proposed analysis will be applied on a solar-thermal application. However, this CMS approach can be applied in the characterization of more applications or environments for a performance or a PdM analysis. In this sense, this approach proposes a monitoring system that analyzes the correct behavior of a process comparing it to one or more (if different modes of operation are identified) pre-estimated operating models, all of them based on data mining. Obviously, a correct estimation of these models will be essential for a valid operation of this monitoring system. Due to this, the modeling task will be described in detail, dividing it into four stages:

A. Extraction of historical data

This stage consists of the extraction of historical data of the process, which should contain measurements and event logs (typically collected from the process SCADA). Obviously, the length and granularity of this historical data must be adequate and contain a representative sample of the behavior in the plant under study.

B. Identification of operating modes

In this stage, the data are split up according to the different operation modes, in which the plant was operating. From this division, the next stages of this section (Data filtering and Models implementation) will be performed with each of these sets independently.

C. Data filtering

Unfortunately, it is very common to find anomalies in historical data. Due to this, before starting the modeling process, it is necessary to carry out an integrity analysis over them. These analyzes usually require a preliminary visual inspection, later choosing the most appropriate statistical method. A typical approach to this end (when normal behavior follows a normal distribution), is the use of interquartile distance criterion, which allows the filter process to determine a limit to separate the outlier data from those are considered as correct.

D. Models implementation

Once the data has been separated for each operation mode and the anomalies have been eliminated for each one, the following step will be to make the process models up.

However, not all the information acquired from the SCADA (direct measurement, cross-effects between them and their non-linear effects) has the same effect over the parameter to be modeled, may not even be relevant. Therefore, to simplify the model, a sensitivity analysis based on Akaike Information Criterion (AIC) [9] has been carried out over the data. This process makes it possible to identify those variables without relevance, discarding them from the initial input set, simplifying the final input set of the model.

After selecting these relevant variables, the next step is to make up a model with a better fit. In this sense, up to five different modeling data mining techniques are proposed for this task, such as:

1) Linear model [10]

In this approach, it consists of estimating the coefficients (β_i) that represent the weight of each input (X_i), which try to fix the behavior of Y, following (1). This approach raises the drawback of not being able to model non-linear behavior. However, it is traditionally a good option for a large number of cases.

$$Y = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_n \cdot X_n \tag{1}$$

2) Linear model with quadratic and cubic terms [11]

This model is really an extension of the previous approach, incorporating also the quadratic ($\beta_{2,i}$) and cubic ($\beta_{3,i}$) terms of the model inputs (X_i). Equation (2) represents this type of model, making it possible to incorporate possible non-linear effects implicit in the process behavior.

$$\begin{aligned} Y = \beta_{1,1} \cdot X_1 + \beta_{2,1} \cdot X_1^2 + \beta_{3,1} \cdot X_1^3 + \ldots \\ \ldots + \beta_{1,n} \cdot X_n + \beta_{2,n} \cdot X_n^2 + \beta_{3,n} \cdot X_n^3 \end{aligned} \tag{2}$$

3) Linear model with second order combinations [11]

This model considers multiple interactions between variables up to the second degree, the relationship now being expressed by (3). This expression makes it possible to model the possible cross-interaction between the inputs (X_i).

$$\begin{aligned} Y = \beta_{1,1} \cdot X_1^2 + \ldots + \beta_{1,i} \cdot X_1 \cdot X_i + \ldots + \beta_{1,n} \cdot X_1 \cdot X_n + \ldots \\ \cdots \\ \ldots + \beta_{n,1} \cdot X_n \cdot X_1 + \ldots + \beta_{n,i} \cdot X_n \cdot X_i + \ldots + \beta_{n,n} \cdot X_n^2 \end{aligned} \tag{3}$$

4) Decision Tree [12]

A decision tree is a regression model represented as a binary tree. Each node contains a condition and the data traverses the tree according to the conditions that they fulfill. The leaves (nodes without descendants) include the regression formula to apply to each data that reaches it.

Therefore, a tree does not generate a linear model but a piecewise linear model.

5) Random Forest [13]

This last alternative consists of a classification and regression model based on a group of decision trees and a voting system.

Finally, all of these techniques are evaluated, only the best of them will be chosen for the monitoring systems.

III. Monitoring System

The modeling process is a complex task that typically involves a lot of data and requires a high computation cost. However, this process is only done once (as off-line task), or with low periodicity to obtain models that reflect some possible changes in the process.

Conversely, on the on-line monitoring system, the pre-estimated model is faced with the direct measurements of the parameter to be evaluated, using this error normalized by its standard deviation as a performance indicator of the correct operation (see Figure 1).

This tool allows the user to identify anomalous trends easily, using standard deviation analysis, identified by the following color code:

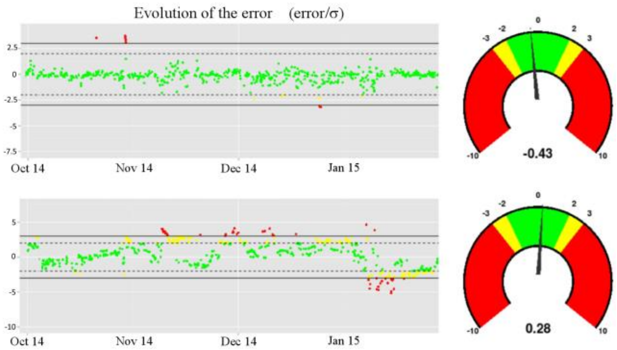

Figure 1. Monitoring interface of the proposed solution.

- *Correct operation* (green points). Cases whose predictions do not differ by more than 2σ.
- *Warning* (yellow points). Cases whose predictions have an error between 2σ and 3σ.
- *Offlimit* (red points). Cases whose predictions exceed 3σ.
- *Out* (black points). Cases whose distances exceed 10σ.

This analysis is able to anticipate a possible failure. For example *warning points*, although still in the normal data range, require greater observation to estimate if the system has a tendency to leave such margins. *Offlimit points* may represent a clear anomaly case (and a possible failure cause). And *Out points*, which are clearly out of the model and are operation modes completely out of training. These points should be studied separately to find possible failures.

IV. Study Case

Once the proposed approach to monitoring and PdM has been described, this section shows its application over a solar power plant based on heliostats.

As a brief description, this type of power plant uses mobile mirrors (or heliostats) that are oriented reflecting and concentrating sunlight toward a specific spot (typically located on a tower). This concentrated radiation generates heat energy that will be converted into motion through a turbine and various fluid circuits (with molten salts and fluids like water). Later, this rotating energy will be converted to electricity through a power generator (as can be seen in the Figure 2).

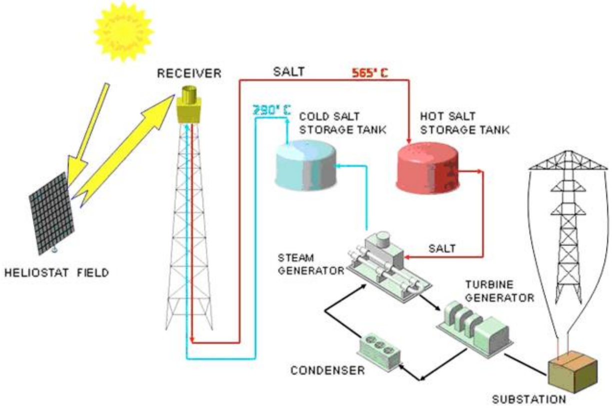

Figure 2. Solar flow basic shema [14].

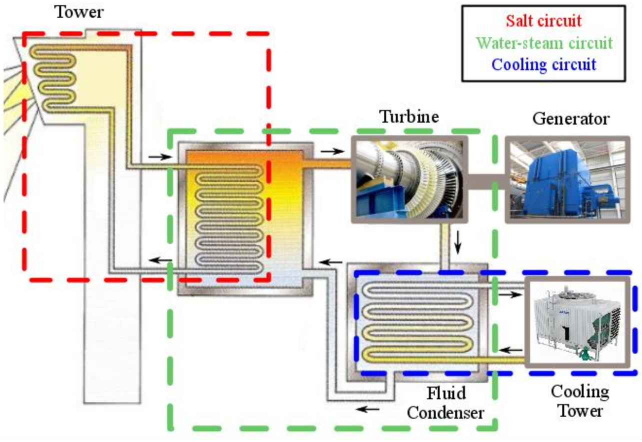

Figure 3. Basic schema of the heat exchange processes.

Thus, neither the power generation levels nor critical operations were analyzed for this example. Actually, the proposed analysis consists of monitoring two elements of the cooling system; the fluid condenser and the cooling tower. Specifically, the fluid condenser is the first stage of the cooling system, and it is responsible for carrying out the heat exchange between the water-steam and the cooling circuits (see Figure 3). The next element, the cooling tower is the part of the cooling process where the water of this circuit is cooled down in other heat exchangers, to be finally poured into a water tank.

Regardless of the temperature ranges in which both process operate, both have the same target (reduce the temperature of the fluid that is flowing through it). In this sense, a good indicator to evaluate this target fulfillment could be the difference between the input and output (thermal jump) of each one. Due to this, both processes have been analyzed, following in both the same approach (obviously varying the input data set for each).

The available information and its reliance for each process is summarized in Table I, which has provided up to 522,664 observations (one year of data approximately). Each of them was divided into two operation modes (day and night modes), as can be seen in the example shown in Figure

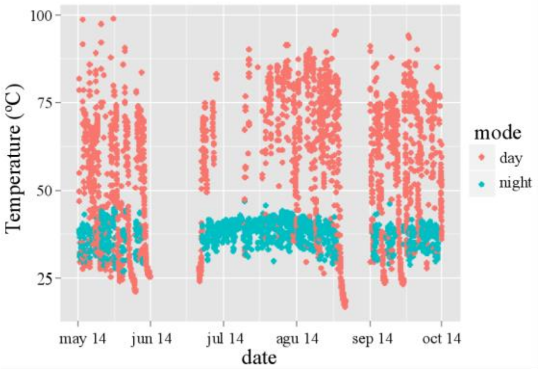

Figure 4. Condenser input temperature (in day and night modes).

TABLE I. COMPLETE SET OF INPUT PARAMETERS

Available measurements	Relev.*
Temperature at steam input in the fluid condenser	
Pressure at steam input in the fluid condenser	(1)
Level at steam input in the fluid condenser	
Temperature at cooling water input in the fluid condenser	(2)
Pressure at cooling water input in the fluid condenser	(1)
Temperature at cooling water output in the fluid condenser	(2)
Pressure at cooling water output in the fluid condenser	(1),(2)
Flow of the input cooling water in the fluid condenser	
Motor current of the cooling tower	(2)
Temperature of water tank	
Level of water tank	
pH of water tank	
Ambient temperature	
Relative humidity	
Atmospheric pressure	
Active power generated	(1)

* Note: (1) relevant for the fluid condenser model, (2) relevant for the cooling tower model.

4. Thus, following the previously described AIC selection method, it is possible to identify four relevant variables for each model (see details in Table I). From this information, and applying the procedures described in previous sections, it is possible to infer up to four system models (one for each subsystem and each mode), necessary for the proposed monitoring.

As can be seen in Table II, Table IV, Table VI and Table VIII, the best evaluated techniques in the four cases is the random forest, using an implementation with five trees.

In this sense, a cross-validation technique was proposed to validate each model. This test allows each method to show how well they function and how good they are.

TABLE II. COMPARISON BETWEEN MODELING METHODS (FLUID CONDENSER, DAY MODE)

Method	σ (ºC)
Linear model	1.488
Model with quadratic and cubic terms	1.309
Second order cross model	1.313
Decision tree	1.364
Random forest (five trees)	**0.598**

TABLE III. OBTAINED RESULTS WITH SELECTED MODEL (FLUID CONDENSER, DAY MODE)

Category	Percentage of cross-validation subset (%)	Percentage of complete filtered set (%)	Percentage of complete set (%)
Correct	95.36	95.57	89.37
Warning	2.54	2.51	6.23
Offlimit	2.01	1.83	4.30
Out	0.09	0.09	0.10

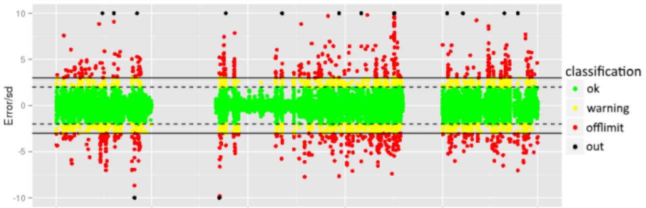

Figure 5. Distribution for complete set of filtered historical data with selected model (fluid condenser, day mode).

TABLE IV. COMPARISON BETWEEN MODELING METHODS (FLUID CONDENSER, NIGHT MODE)

Method	σ (ºC)
Linear model	1.497
Model with quadratic and cubic terms	1.403
Second order cross model	1.454
Decision tree	1.360
Random forest (five trees)	**0.635**

TABLE V. OBTAINED RESULTS WITH SELECTED MODEL AND UNFILTERED DATA (FLUID CONDENSER, NIGHT MODE)

Category	Correct	Warning	Offlimit	Out
Percentage (%)	95.50	2.48	1.94	0.08

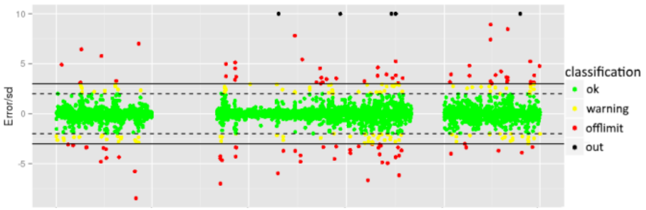

Figure 6. Distribution for complete set of unfiltered historical data with selected model (fluid condenser, night mode).

On the one hand, once the more adequate technique has been chosen, a fluid condenser model raises a standard deviation of 0.61589ºC in day mode and 0.62475ºC in night mode. Additionally, on the other hand, the cooling tower model raises a standard deviation of 0.38035ºC in day mode and 0.62475ºC in night mode. Therefore, it is possible to conclude that these proposed models are a valid estimation for the studied subsystems. This conclusion is also validated

TABLE VI. COMPARISON BETWEEN MODELING METHODS (COOLING TOWER, DAY MODE)

Method	σ (ºC)
Linear model	0.940
Model with quadratic and cubic terms	0.906
Second order cross model	0.717
Decision tree	1.269
Random forest (five trees)	**0.380**

TABLE VII. OBTAINED RESULTS WITH SELECTED MODEL (COOLING TOWER, DAY MODE)

Category	Percentage of cross-validation subset (%)	Percentage of complete filtered set (%)	Percentage of complete set (%)
Correct	94.70	94.96	84.87
Warning	3.21	3.06	5.88
Offlimit	2.06	1.96	8.61
Out	0.03	0.02	0.64

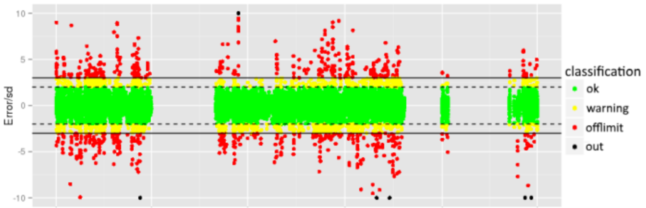

Figure 7. Distribution for complete set of filtered historical data with selected model (Cooling tower, day mode).

TABLE VIII. COMPARISON BETWEEN MODELING METHODS (COOLING TOWER, NIGHT MODE)

Method	σ (ºC)
Linear model	0.965
Model with quadratic and cubic terms	0.952
Second order cross model	1.281
Decision tree	1.376
Random forest (five trees)	**0.422**

TABLE IX. OBTAINED RESULTS WITH SELECTED MODEL AND UNFILTERED DATA (COOLING TOWER, NIGHT MODE)

Category	Correct	Warning	Offlimit	Out
Percentage (%)	94.92	3.09	1.96	0.03

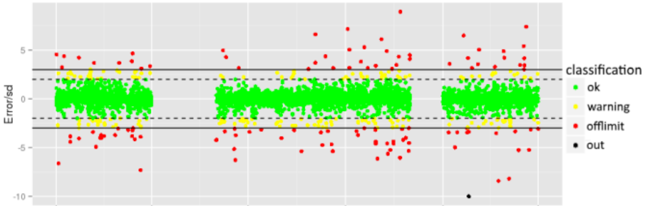

Figure 8. Distribution for complete set of unfiltered historical data with selected model (Cooling tower, night mode).

by Table III, Table V, Table VII and Table IX, and also from Figure 5 to Figure 8 that show the distribution of the error between real an estimated historical data.

V. CONCLUSIONS

As was commented in this paper, the accuracy requirements of current systems have grown enormously in the last few years. This evolution makes the monitoring and on-line analysis (such as PdM) essential in the production systems, among which the energy industry stands out.

In this sense, a PdM approach based on data mining techniques is proposed in this paper. This approach brings up a comparison between the real performance of a plant and an estimation (based on a model) of it, identifying a possible deviation from it as an anomaly, which could lead to future failure.

This approach has been evaluated over the cooling subsystems of real solar-thermal power plant. In this way, the analysis of this application made possible to validate its usefulness, comparing different modeling techniques and identifying the more appropriate of them for this application.

ACKNOWLEDGMENT

This research has been supported by the Ministry of Economy and Competitiveness of Spain through the SIIAM project (Reference Number: TEC2013-40767-R).

REFERENCES

[1] H. M. Hashemian and W. C. Bean, “State-of-the-Art Predictive Maintenance Techniques*,” *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 10, pp. 3480–3492, Oct. 2011.

[2] H. Lu, W. J. Kolarik, and S. S. Lu, “Real-time performance reliability prediction,” *IEEE Transactions on Reliability*, vol. 50, no. 4, pp. 353–357, Dec. 2001.

[3] D. B. Durocher and G. R. Feldmeier, “Predictive versus preventive maintenance,” *IEEE Industry Applications Magazine*, vol. 10, no. 5, pp. 12–21, 2004.

[4] C. Rudin *et al.*, “Machine Learning for the New York City Power Grid,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 328–345, Feb. 2012.

[5] D. Evagorou *et al.*, “Feature extraction of partial discharge signals using the wavelet packet transform and classification with a probabilistic neural network,” *IET Science, Measurement Technology*, vol. 4, no. 3, pp. 177–192, May 2010.

[6] S. S. M. Ghoneim, I. B. M. Taha, and N. I. Elkalashy, “Integrated ANN-based proactive fault diagnostic scheme for power transformers using dissolved gas analysis,” *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 23, no. 3, pp. 1838–1845, Jun. 2016.

[7] P. Bangalore and L. B. Tjernberg, “An Artificial Neural Network Approach for Early Fault Detection of Gearbox Bearings,” *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 980–987, Mar. 2015.

[8] A. Zaher, S. D. J. McArthur, D. G. Infield, and Y. Patel, “Online wind turbine fault detection through automated SCADA data analysis,” *Wind Energy*, vol. 12, no. 6, pp. 574–593, 2009.

[9] J. M. Chambers, *Statistical Models in S; Chapter 6: Generalized linear models*, 1st ed. Chapman and Hall/CRC, 1991.

[10] J. M. Chambers, *Statistical Models in S; Chapter 4: Linear Models*, 1st ed. Chapman and Hall/CRC, 1991.

[11] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2015.

[12] T. Therneau, B. Atkinson, and B. Ripley, *Recursive Partitioning and Regression Trees*. 2017.

[13] A. Liaw and M. Wiener, “Classification and Regression by randomForest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[14] J. I. B. J.I. Ortega and F. M. Tellez, “Central Receiver System Solar Power Plant Using Molten Salt as Heat Transfer Fluid,” *Journal of Solar Energy Engineering*, vol. 130, no. 2, pp. 1-6, 2008.