Proceedings World Geothermal Congress 2020+1
Reykjavik, Iceland, April - October 2021

Data Mining Analytics for Geothermal Power Plants

Oscar Fernando Cideos Nunez

LaGEO, final 15 ave. Sur. Colonia Utila, Santa Tecla

ocideos@gmail.com

Keywords: data mining, big data, regression trees, random forest, neural networks, machine learning

ABSTRACT

Any industrial production plant in the world collects massive amounts of data daily. Geothermal power plants are no exception to this, assuming they follow the good practice of collecting data for later analysis. Geothermal developments tend to overlook the operational part of a field, given that it is the less risky part of geothermal development investment. While thermodynamic models are always reliable, once a power plant starts operation and then its aging process its particular thermodynamic model tends to be less accurate due to the small particularities of that power plant, in order to account for this big data models can be used to create a unique descriptor of a power plant.

1. INTRODUCTION

There is an ever-increasing trend in collecting data in our daily lives. With the recent advances in digital technology, this has become even easier than before; however, collecting data just for the sake of it is not helpful.

In general, the hypothesis would be for the data to answer a question, are consumers spending too much? Keep track of expenses. Is diet working? Weighing oneself once a week, one can continue proposing questions that can only be answered by keeping a good record of such data. Geothermal power plants are no stranger to this concept; the power output of the turbine is decreasing? Is the condenser pressure higher, or maybe the geothermal wells are producing less steam; in this case individuals can always check the thermodynamic model and see if it gives an answer. In general, this works very well, but later as the plant ages, the model is slightly off.

In a geothermal power plant, the data collected is generally only analyzed after an unprecedented event or anomalous operation due to the massive amounts of data collected, but if the process can be automated, the task becomes easier. In addition to the main equipment, there is data collection in the pipelines, wells, separating stations, etc. The idea of this research is to try to make use of the data collected over recent years for a geothermal power-plant in El Salvador and investigate if this data can be used to predict power production of the plant.

This research aims to have a different approach to the way data is usually processed in geothermal power plants. With an accurate predictor model of a power plant, the user does not need to have an engineering background in order to have a good descriptor of the power plant parameters, and a powerful decision-making tool for future projects.

2. POWER PLANTS DATA

For this research, the data used comes from Berlin geothermal power plant in El Salvador, operated by LaGEO S.A. de C.V.; this data has been previously used for another project. Due to data and information policies in the company, not all the data used to build the models can be published. Hence some of the data may be slightly adjusted, and a short clarification will be written anytime this is completed. Two different types of reports are used where data is stored, the operational report, and the field report. The operational report collects data from the sensors in the powerhouse daily at two-hour intervals. The field report stores the data from outside the field, wells data (temperature, pressure and mass flows) and operational events (scheduled maintenances, shutdowns, etc.).

There are four turbines currently operating in the Berlin geothermal field, 3 single flash units and 1 binary cycle power plant. This research focuses on Unit 1 and Unit 2, two single flash 28 MW identical turbines. All the models and analyses use only data from Unit 1 and Unit 2; the only exception is in the event analysis where there are events of Unit 1 or Unit 2 that are related to the operation of Unit 3.

The information comes from the control room of the power plant. Each parameter is logged into the report every two hours starting from 01:00 and finishing at 23:00. When a parameter is not logged into the report, the cause can be a fault in the sensor or a unit blackout.

Units 1 and 2 share the same logged information and sensor distribution; both units are also logged in the same worksheet of the report.

The data is stored with a short description of the logged parameter, the name of the sensor in the power plant which gets the data, the units in which the data is stored, the stored values, and a final column storing the average of the measurement. The average column uses the excel function for calculating averages, which includes only the cells with value. There may be a case where a cell may be left blank, which can be related to the causes stated above for parameters not logged.

3. PREPARING THE DATA

When preparing data for a predictive model, one usual approach is to observe the data, and know beforehand the type of data put into the models. The most usual approach is to make a graphic of the data; this helps in order to see how the data is distributed. There are some aspects to be considered when observing data:

- When there is a missing point in the data, unless it is a qualitative variable, all the related data points are usually discarded in order to obtain a more reliable data structure.

- When there are data points to far dispersed from the mean value of the graph (too high or too low) unless the user is sure the data is correct, it may mean a faulty measurement, and all the related data can be discarded.

- If the data is segmented, meaning that the data shows more than one clear mean (for instance a power plant producing 20 MW for 3 months and then 18 MW for 6 months) it is better to separate into two models until the user has a better understanding of the data structure.

- General knowledge of the process is usually desired when building a model, in order to better interpret the results initially given, once a reliable model has been built, it can be generalized.

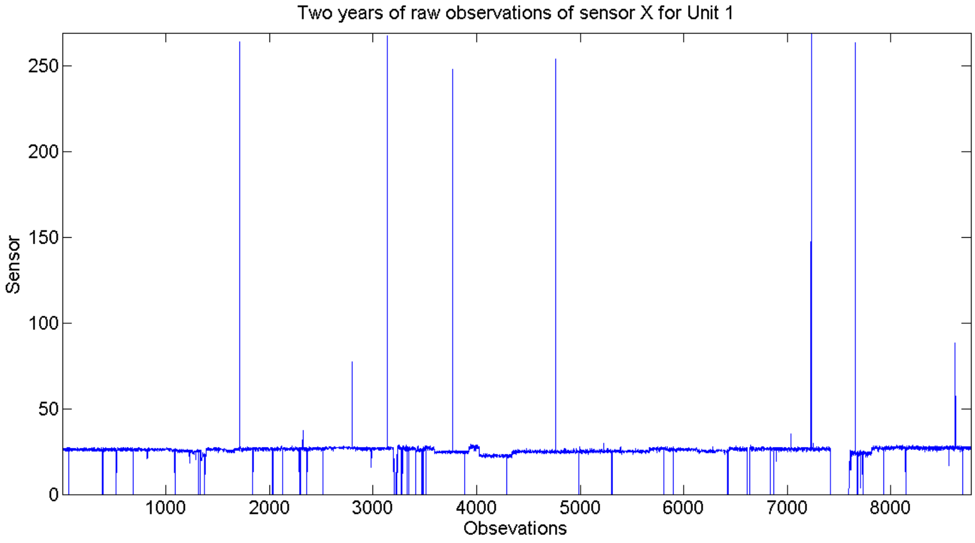

Figure 1: Graph of the sensor showing two years of raw data observations.

Two years of raw data is shown in Figure 1, as mentioned earlier in this paper, the data needs to be processed before applying any predictive or numerical models. In that figure, at least 10 peaks can be seen, but there are also several regions were the value drops to zero. Prior knowledge to this particular measurement is handy in this case, because this particular value cannot go above 30, so in this case one can easily discard all the values that are above this number, this can be done in any data processing software, or if the data is too much, a script in a programming language can be easily implemented.

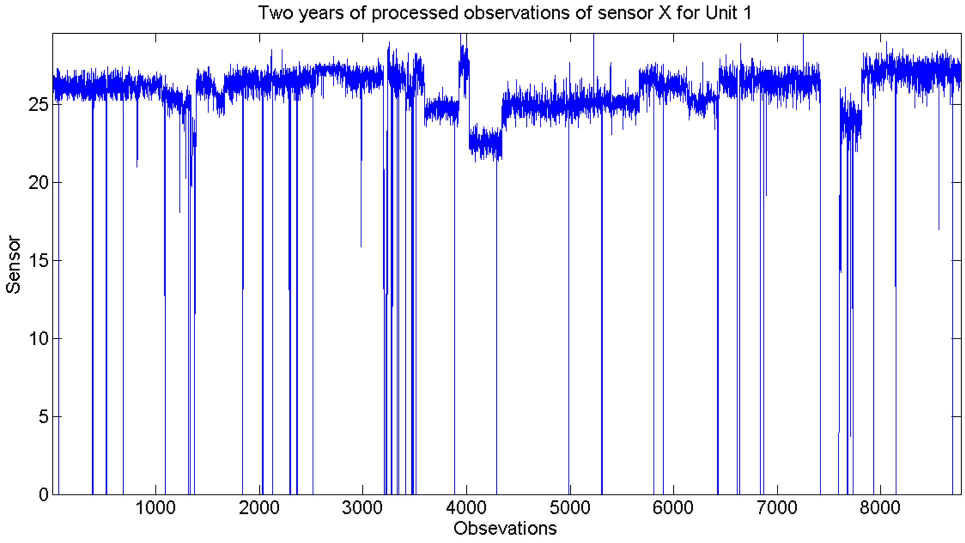

Figure 2: The same data as Figure 1 is shown in this figure, but with the peaks removed.

It is essential to remember that when cleaning the data from observations out of range, all of its associated data needs to be removed, this is done in order to prevent data ‘misalignment’ (points of data been correlated to observation not made at the same time). Even when the data has been cleaned from out of upper range observations, the lower range observations need to be carefully analyzed too, this means in this case, knowing which regions of the data contain observations that are out of range (plant shutdown or sensor out of service) and remove the information that no longer serves the purpose of analysis.

When starting a data mining analysis is an excellent practice to initially remove peak observations and observations with zero data, this can give a feeling of how the data is processed when analyzed, and when creating a more complex model, these data points can be analyzed too.

The sensors selected for the data analysis are shown below.

Table 1: Data used for the research.

Description	**Units**	**Variable name Unit 1**
Turbine output	MW	W1
Turbine steam mass flow at the inlet	Tons/hr	m1
Turbine steam inlet pressure	bar	P_in1
Turbine inlet temperature	°C	T_in1
Turbine chamber inlet pressure	bar	Pch_in1
Condenser pressure	bar	Pc1
Condenser temperature	°C	Tc1
Condenser water level	mm	CL1
Cooling water mass flow	Tons/hr	mc1
Cooling water inlet temperature	°C	Tcool_in1
Cooling water outlet temperature	°C	Tcool_out1
Left control valve position	%	CV_left1
Right control valve position	%	CV_right1

After processing the data, the next step done in this research, was to discover the relations between the collected data and the sensors, as with last time, prior knowledge of the data and thermodynamic background allows to have a better understanding of the data presented, in this case, a correlation plot was used to discover the relations.

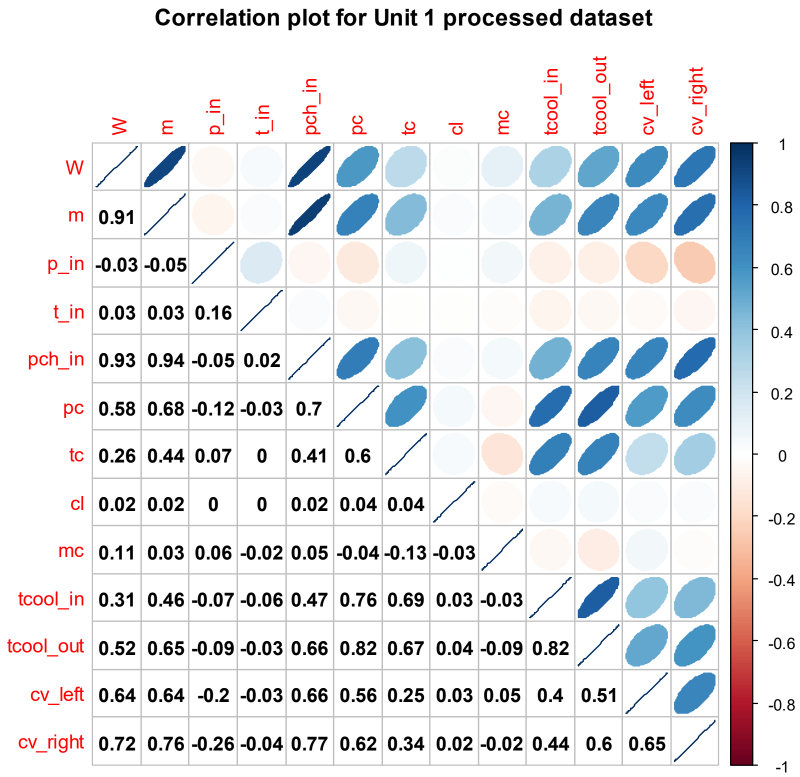

Figure 3: Correlation plot for unit 1 processed dataset.

This correlation shows how the values of any two variables correlate with each other, the upper off-diagonal of the plot show a “graph” where a blue ellipse show positive correlation and a red ellipse shows negative correlation (that means that if one value increases the other decreases and vice versa), the lower off-diagonal show the Pearson correlation coefficient, which is a numerical representation of the value shown in the elliptical graph. This information gives us a good idea of how good the data used for the evaluation correlates with each other, allowing the creation of an accurate model.

4. DATA MINING MODELS

What is commonly known as data mining is simply finding data patterns in datasets that are practically too big to be discovered by a “simple” observation, in this sense, even if the data used for this research doesn’t reach more than 200 mb of data after processing, a simple look at the data couldn’t give us the insight of a data mining predictive model.

After preparing the dataset as explained in the section before, a good practice is to split the data into three different datasets, usually a training set (where most of the data is used) a validation set and a testing set, the sets were divided as follows 70% for training, 15% for validation and 15% for testing.

Three types of models were created in order to model the data. The first one is a thermodynamic model, which will allow us to have a base model for comparison, the other types of models were regression type models (regression trees and random forest) and linear models (minimum squared and Akaike Information Criterion). The thermodynamic model was created in EES, and the rest of the models were created using custom scripts in R programming language.

Four data mining models were used for this research; the models are:

Regression Trees

This type of model, also known as Decision Trees, is divided into two subcategories, classification trees (for categorical variables) and regression trees (for numerical variables). This model selects the variables linked to the target that has the most significant influence and based on selection criteria give each of the variables a threshold value in order to deliver an estimated target value. In general, where the variability of the target value is too high, this type of model tends to be less accurate.

Random Forest

This type of model, as its name implies is a forest of regression trees. Essentially random forests are a collection of decision trees that individually predict the value of the target function, and then the whole forest votes for the final predicted value.

The next two models are created by using a combination of all the variables involved in the process, and the testing all the models created over the validation dataset, after that, two validation methods are used to select two different models.

Minimum Validation Error.

In this method the sum of the absolute value of the difference between the observed values and the calculated values is stored, and the model with the smallest validation accuracy is selected.

Akaike Information Criterion

This method is used to compare models with different complexities, used to select a model that makes a good “explanation” of the data while penalizing the model’s complexity in order to prevent poor generalization.

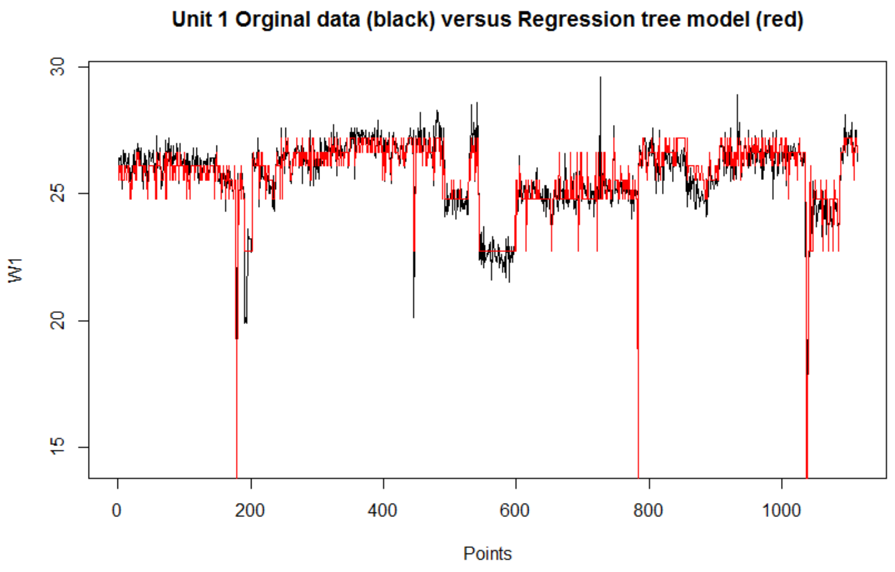

Figure 4: Regression tree model of Unit 1 power output

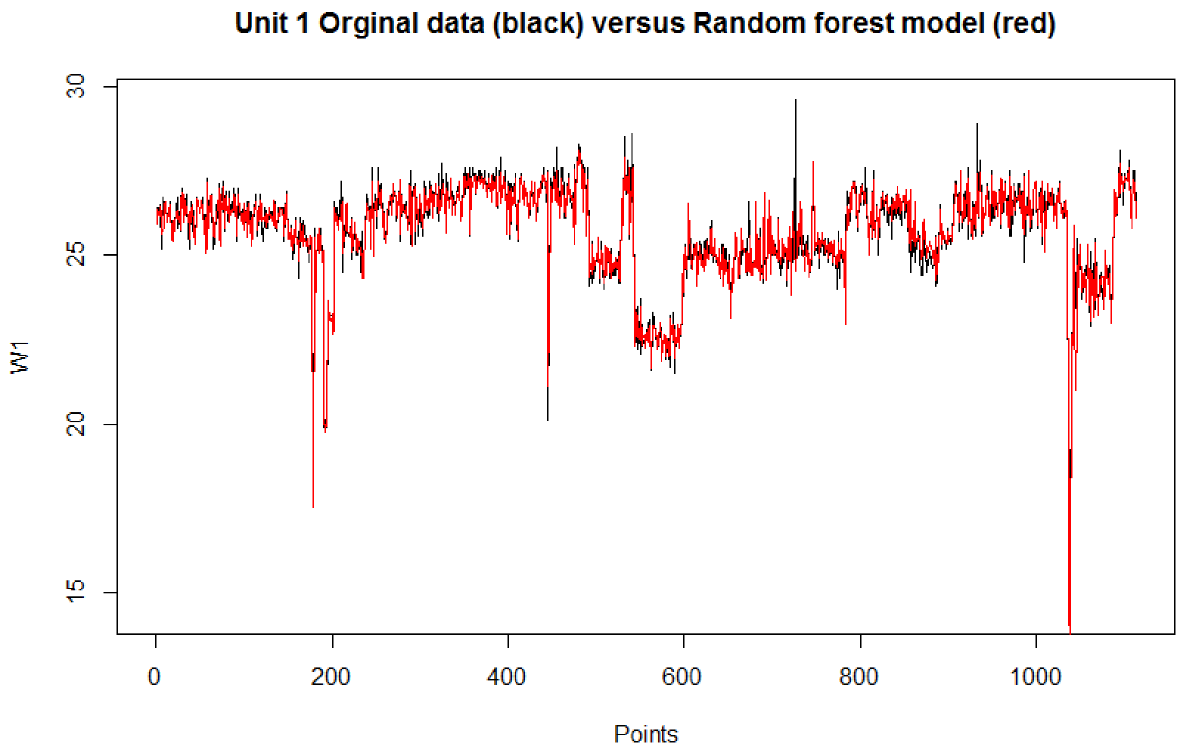

Figure 5: Random forest model of Unit 1 power output.

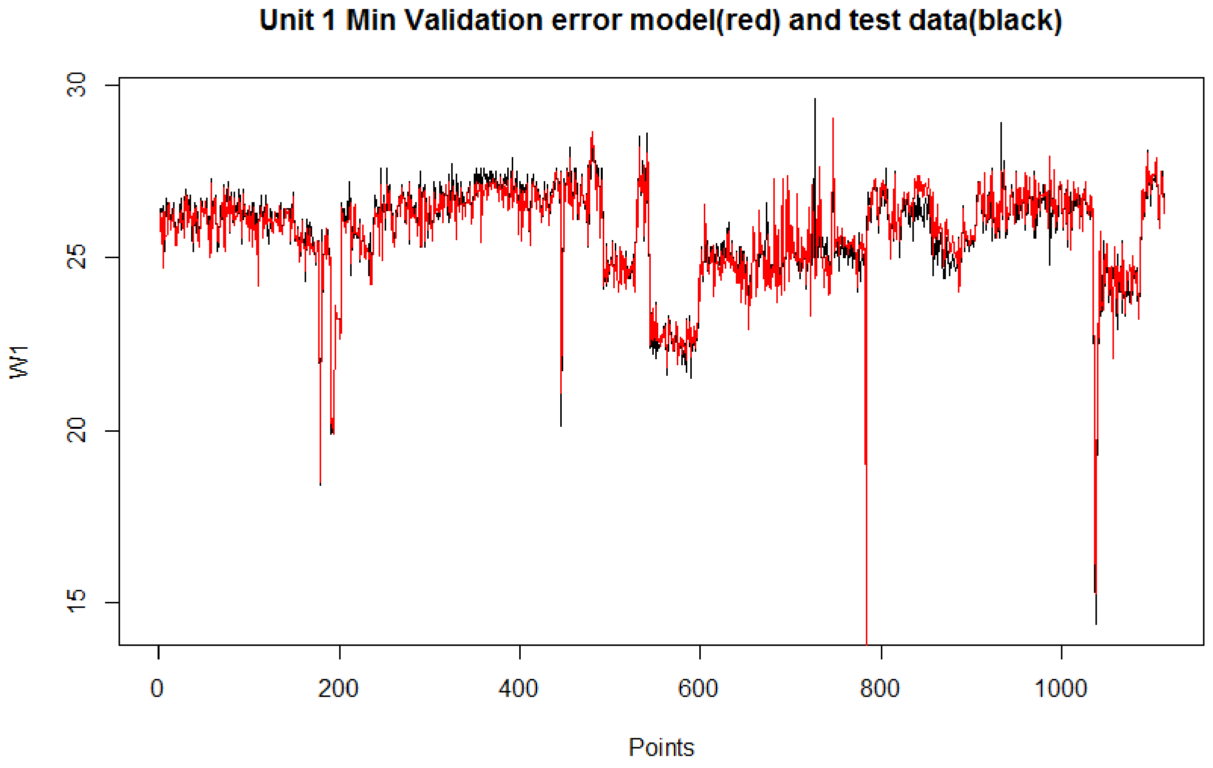

Figure 6: Min validation error model for the Unit 1 power output.

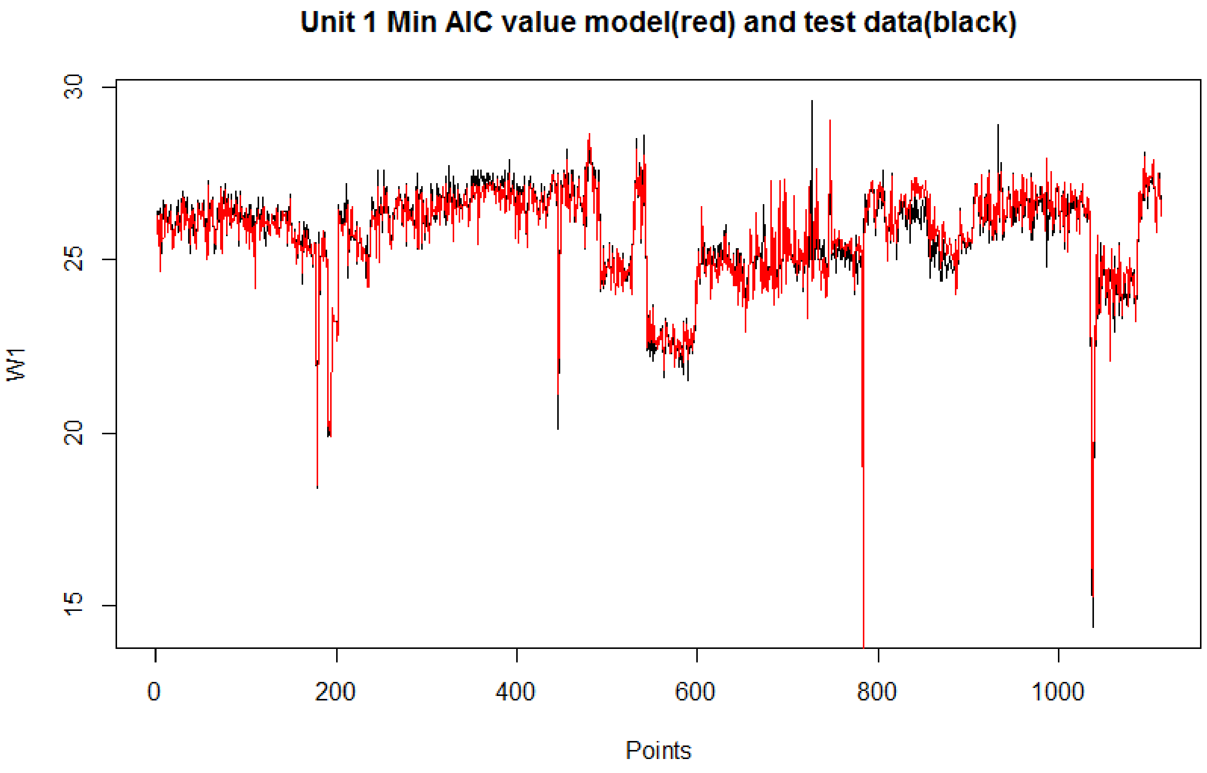

Figure 7: Minimum AIC value model for the Unit 1 power output.

As can be seen from all the model graphs, it is very difficult to discern what model is the best performer of all the models, for comparison, below are shown the thermodynamic model and the random forest model side by side.

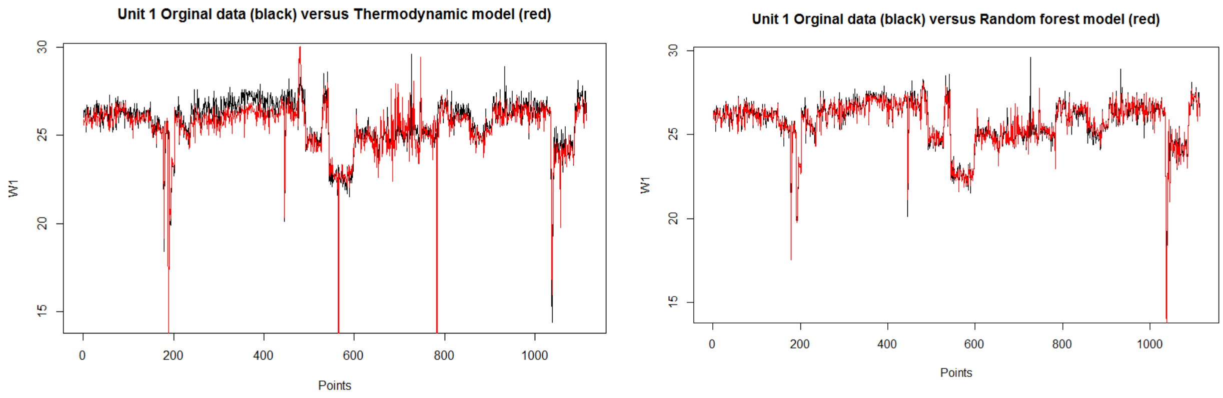

Figure 8: Comparison side by side of the thermodynamic model and the random forest model.

It is clearly shown in Figure 8, that the random forest model is a better predictor of the power plant output than the thermodynamic model, but between the data mining models, a better selection criteria needs to be taken into account than just comparing two graphs side by side.

A very important step to take when splitting the data is to randomize the selection, meaning that for instance if the data comes from 12 months the selection shouldn’t be from Jan-Oct for the training set, Nov for validation, and so on, instead there should be a randomized selection of the data points and let the software do the splits based on the number of observed points.

5. SELECTING THE BEST MODEL

For the selection of the model, four methods were used in order to have an idea of the performance of the models and be able to compare them qualitatively. These four tools are: Mean Absolute Error, Mean Square Error of Prediction, Coefficient of Model Determination and Modelling efficiency.

Table 2 Score values from the evaluation methods for each of the models.

Model	**MAE (Mean Absolute Error)**	**MSEP (Mean Square Error of Prediction)**	**CD (Coefficient of Model Determination)**	**MEF (Modelling efficiency)**
Minimum validation accuracy	0.01623	0.29304	0.01884	0.8559
Minimum AIC value	0.01621	0.29262	0.01985	0.8561
Regression tree	0.02036	0.46145	0.03319	0.77308
Random forest	**0.01026**	**0.11708**	**0.00244**	**0.94243**
Thermodynamic model	0.03784	1.59344	0.34116	0.21643

The mean absolute error compares the distribution of the differences between the predicted and the observed values against zero. The lower the MAE, the more accurate the model.

The mean square error of prediction (MSEP) calculates the squared sum of the difference between the observed values between the model predicted values divided by the total number of observations. The model with the smallest value shows that it is a better predictor than the others.

The coefficient of model determination (CD) is the ratio of the total variance of observed data to the squared of the difference between model-predicted and the mean of the observed data. The smallest the value, the better the predictor.

The modeling efficiency (MEF) is interpreted as the proportion of variation explained by the predicted values. If the model prediction were perfect, the value of MEF would be equal to one, and if the MEF is lower than zero the fitted values predict the data worse than using just the mean.

As shown above, the model with the best accuracy is the random forest in all cases. This type of model can be implemented.

7. FUTURE WORK

These types of models can be straightforward to implement if there is good information available. The model will work in any power plant, the better the quality of the data, the better the predictor will work. The main reason for creating a model like this has to be, what type of data is being measured? Is this data accurate enough for a good model? Is the data all collected at the same type by the same type of device?

As shown previously, the data mined model gave a very good description of the target function, which means that the model works and can be expanded to include other variables in the same power plant, and include external variables that can be correlated to any plant parameters.

Future work related to this research will include:

- Usage of other types of data mining and machine learning models.
- Usage of a bigger dataset, including atmospheric variables and geothermal field data (temperature, pressure and flow rates in production and reinjection wells).
- Fine-tuning of prediction models in order to detect inefficiencies in the processes.
- Correlation with financial data.

8. CONCLUSIONS

After developing the models, the main conclusions of this research are:

- Knowledge of the process is always desirable for the person creating the models and this particular case (geothermal development) and thermodynamic background helps to select the best variables that intervene in the process and can affect the target variable.
- A big part of creating a good model is the careful preparation of the dataset that will be used for creating the model. Also, a good familiarization of the machine learning and data mining models available at the time is necessary in order to make the model creation process a bit easier.
- The bigger cost in developing this type of model comes from very early on in plant development, given that the data has to be collected over a very good period of time and the sensors and data measurement points have to be representative of the process as a whole.

- Once the data is verified to be of good quality and a good model is created. The next step is to look for more opportunities to increase efficiency in the process.
- Even if there are thermodynamic models available, and these types of models may seem redundant. Even the best thermodynamic model cannot take into account the particularities of an already ongoing process, which in turn can be obtained using data mining.

REFERENCES

Bradley, J., and Amde, M., 2015: Random Forests and Boosting in Mllib, website: https://databricks.com /blog/2015/01/21/random-forests-and-boosting-in-mllib.html

Breiman L., 2001: Random forest – Random features, Technical Report 567, Sept., 1999, Statistics Department University of California Berkeley, 29 pp.

Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J., 1984: Classification and Regression Trees. Chapman and Hall, London, New York, Washington, D.C. 359 pp.

Meisner, M. P., Wachs, M., Sambasivan, R. R., Zheng, A. X., and Ganger, G. R., 2009: Modeling the Relative Fitness of Storage, Carnegie Mellon University, 12 pp. website: http://pdl.cmu.edu/PDL-FTP/SelfStar/sigmetrics07.pdf

Tedeschi, L. O., 2004: Assessment of the Adequacy of Mathematical Models. Workshop on Mathematical Model Analysis and Evaluation, Sassari, Italy, 28 pp.

Therneau, T. M., Atkinson, E. J. and Mayo Foundation, 2015: An Introduction to Recursive Partitioning Using the RPART Routines. Rstudio documentation, 62 pp.