# Executive Summary

SO2 emission from recovery boiler is a regulated pollutant. This study investigates a process data set from a real pulp and paper mill in the United State to determine the root cause of emission using various statistical and machine learning algorithms including ordinary least square, principle component analysis, partial least square and regression tree.

Data was treated and transformed to maximize its potential for modeling. To account for time delay effect from one process to the next, lag time was determined using cross correlation and the data is lag adjusted to maximize correlation with SO2 emission.

Using frequency analysis it was found that secondary air flow, flue gas oxygen, dry solid density A&B, primary and secondary air temperature, black liquor pressure and indirect heater, and green liquor sulfidity share the same frequency as SO2 emission.

Regression methods were used to extract linear and nonlinear relationship between process variables and SO2 emission. These models are tested using a test set that the model have never seen before and predicted with accuracy ranged from 50% - 80%. The method with the highest accuracy is regression tree. From regression models, directional correlation is found, such as

higher primary air temperature reduce emission.

Combining result from spectral analysis, regression and process knowledge, it was found that secondary airflow, primary and secondary air temperature and black liquor indirect heater temperature are the causes and primary air temperature is the root cause.

# Table of Content

1. Background of recovery boiler and sulfur dioxide emission
2. Project Objective
3. Table of variables
4. Explanation of variables
5. Exploratory Analysis
   1. Univariate Analysis
      1. Time Trend
      2. Power Spectrum
      3. Autocorrelation
   2. Bivariate Analysis
      1. Correlation Analysis
      2. Spectral Correlation Analysis
      3. Cross-correlation and lag adjustment
6. Pre-processing
   1. Filling missing data
   2. Removal of Outliner
   3. Log and Box-Cox Transformation of Skewed Data
   4. Filtering noise using moving average filter
   5. Normalization
   6. Creation of training and test set
7. Regression
   1. OLS
   2. LASSO
   3. PCA
   4. Partial least square regression
   5. Decision Tree Regression
8. Conclusion
   1. Comparison of regression accuracy by models
   2. What is the root cause of emission?

Root Cause Analysis of SO2 Emission for Kraft Pulp Mill Recovery Boiler

# Introduction

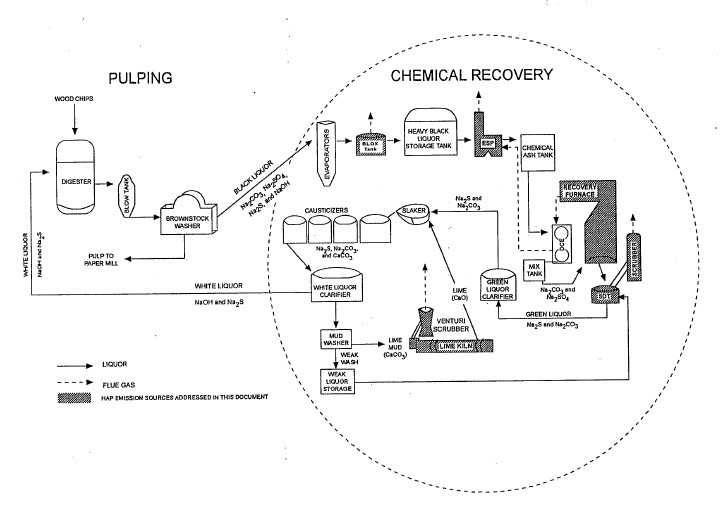


Figure 1: Process Flow Diagram of The Chemical Recovery Cycle[1]

In pulp and paper industry wood chips is cooked in digester at 12 pH to produce pulp. The pH was raised using caustic soda (NaOH) and sodium sulfide (Na2S). The wood fiber broken down is strained to make pulp, which is then bleached and process to make paper. The by product from the digester is called black liquor. Black liquor is rich in sodium, fine fiber and inorganic compound such as sulfur. To recycle sodium to make caustic soda and allows the mill to produce pulp continuously, and to use the energy stored in the black liquor, engineers designed the recovery boiler to burn black liquor to produce steam and sodium rich stream which is called green liquor. Figure 1 shows the process flow diagram of the chemical recovery cycle. To see an overview of the exterior of the Recovery boiler see Figure 1 in Appendix. Green Liquor is causticized at the causticizer and produce a NaOH rich and Na2S rich stream called white liquor, which is added to the digester to complete the cycle. Figure 2, shows a schematic diagram of a recovery boiler.

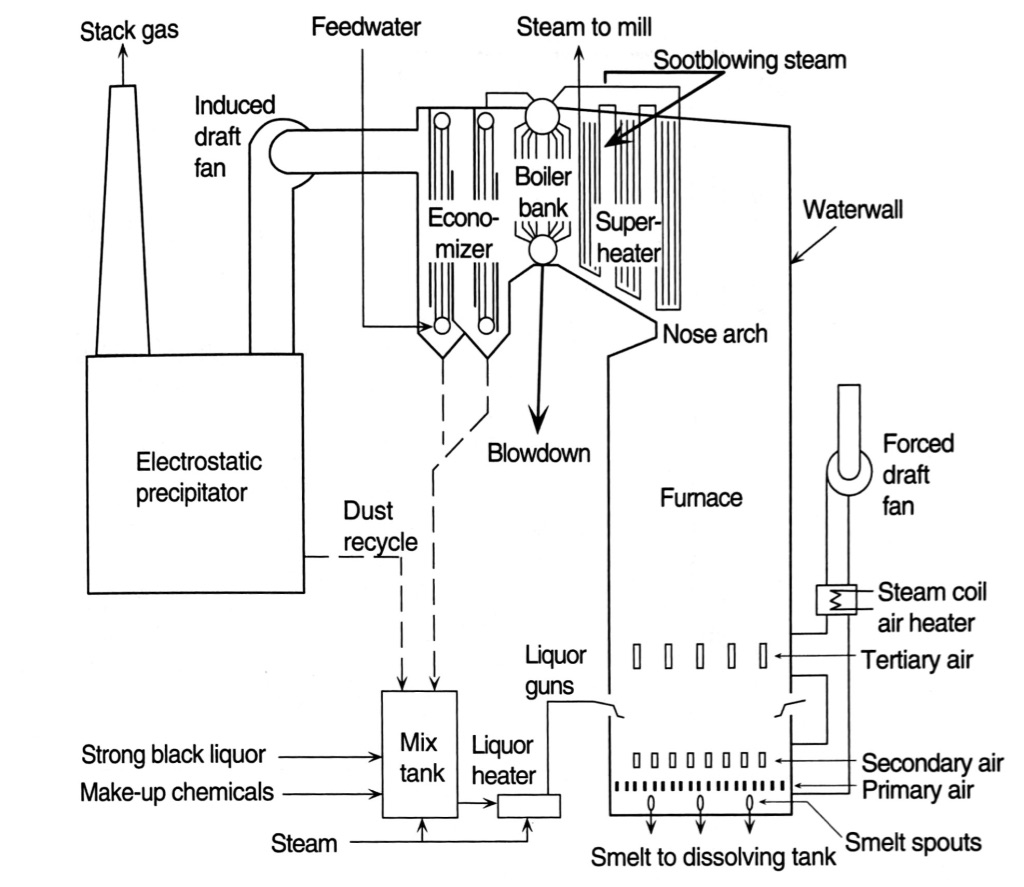


Figure 2: Schematic diagram of a recovery boiler

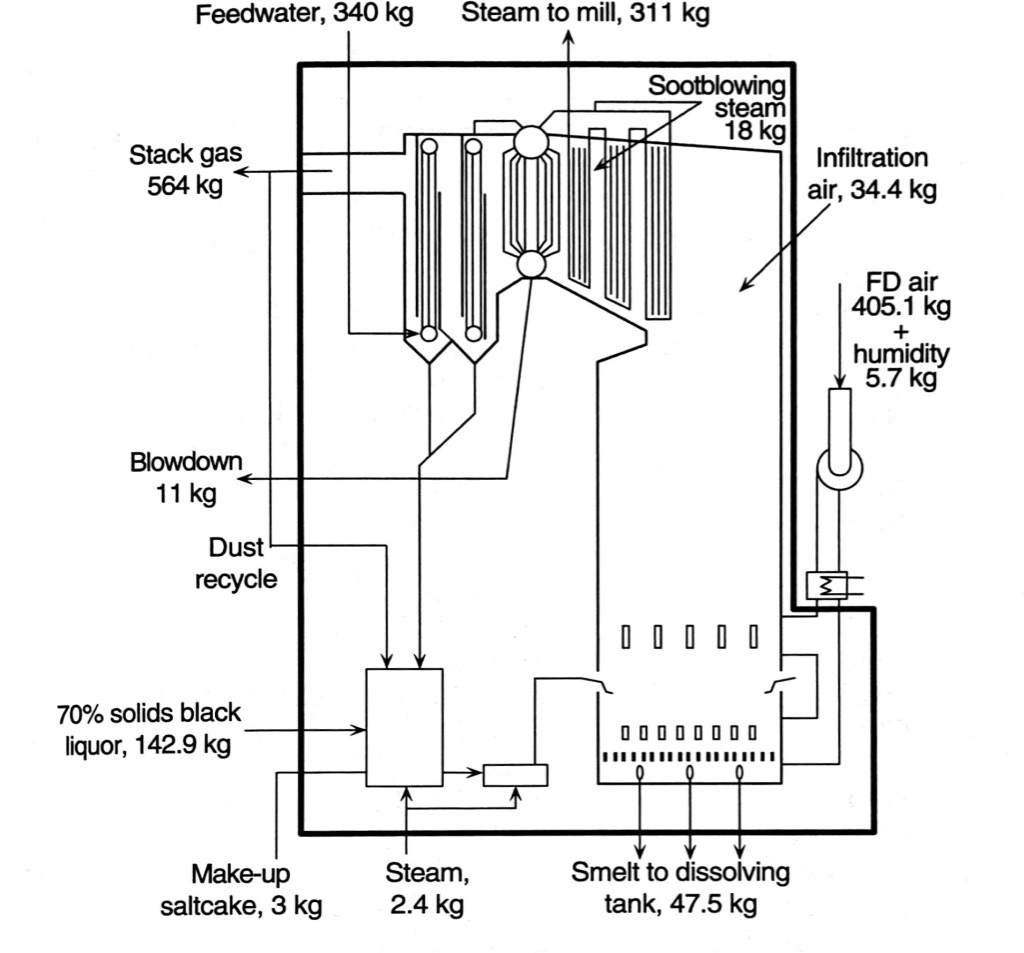


Figure 3: Material balance based on 100 kg of black liquor solids for a typical recovery boiler

The recovery boiler is a 13-floor tall boiler that produces thousand tons of black liquor per day. Black liquor is heated by indirect heater and sprayed into the hearth. High Black liquor pressure reduces the droplet size of the spray and improves combustion. The combustion raises internal temperature of the boiler to 800-1000 degree Celsius. Hot gases raise from char bed to pass through a bank of heat exchanger at the top of the recovery boiler to produce steam before scrubbing and leave to the atmosphere through stack. To provide oxygen for combustion, outside air is drawn in and injected at different height of the recovery boiler. By their position, the air is called primary air, secondary air and tertiary air. Since black liquor contains sulfur from wood residue in the as sodium sulfate (NaS) and sodium sulfide (Na2SO4), the combustion of black liquor produces sodium dioxide (SO2). Sodium dioxide is a toxic nerve agent at low concentration, explosives as high concentration and contributes to the formation of acid rain. Therefore, it is a regulated pollutant and the mill controls its emission using scrubber, electro-precipitators and optimization of the recovery boiler. This study aims to investigate process data from the mill to provide insight for optimization.

An important mechanism in reducing sodium dioxide is a reaction between gaseous sodium and sulfur dioxide at high temperature that produces sodium sulfate and carbon monoxide. The sodium sulfate would precipitate and reduces sodium emission. However, this phenomenon is affected by gas mixing at the tertiary air region. Therefore, the temperature and pressure of the three level of air supply affects SO2 production and precipitation. Figure 2 in Appendix shows the schematics of the internal of recovery boiler.

Black liquor is a mixture of sodium, sulfur, organics, fine fiber and water. The water content of black liquor affects the temperature of the char bed. Therefore, the black liquor is concentrated before spraying and combustion. A measure of sulfur in black liquor is called total reduced sulfur. It is an aggregated measure of sulfur in the form of hydrogen sulfide (H2S) , methyl mercaptan (CH3SH), Dimethyl sulfide (CH3SCH3), Dimethyl Disulfide (CH3S2CH3), Sulfur Dixoide.

There are 25 variables and 2328 rows of records in this dataset. The variables are listed in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| Tag number | Variables | Variable Description | Unit |
| 1 | Time stamp | Time stamp |  |
| 2 | PAF | Primary Air Flow | Kilopound per hour (KPPH) |
| 3 | SAF | Secondary Air Flow | Kilopound per hour (KPPH) |
| 4 | TAF | Tertiary Air Flow | Kilopound per hour (KPPH) |
| 5 | PWP | Primary Windbox Pressure | Inches of H2O |
| 6 | SWP | Secondary Windbox Pressure | Inches of H2O |
| 7 | TWP | Tertiarary Windbox Pressure | Inches of H2O |
| 8 | BR | Burn Rate | GPM |
| 9 | Sulfidity | White Liquor Sulfidity | % |
| 10 | BL\_Solid\_Test | Black Liquor Solid 50./50 Test | % |
| 11 | FGO | Flue Gas Oxygen | % |
| 12 | TRS | Total Reduced Sulfur | ppm |
| 13 | DSDA | Dry Solid Density Transmitter A | % |
| 14 | DSDB | Dry Solid Density Transmitter B | % |
| 15 | PAT | Primary Air Temperature | °F |
| 16 | SAT | Secondary Air Temperature | °F |
| 17 | Steam\_Flow | Steam Flow Rate | KLB/HR |
| 18 | BL\_Solid\_Flow | Black Liquor Solid Flow Rate | Kilopound per hour (KPPH) |
| 19 | Steam/DS | Steam to Dry Solid Ratio |  |
| 20 | SO2M | Sulfur Dioxide Concentration Minute Average | ppm |
| 21 | SO2HR | Sulfur Dioxide Concentration Hour Average | ppm |
| 22 | BLPressure | Black Liquor Pressure | PSIG |
| 23 | BL\_IHT | Black Liquor Indirect Heater Temperature | °F |
| 24 | SC\_RFS | Salt Cake Rotary Feeder Speed | Hz |
| 25 | LKGL\_SULFIDITY | Lime Kiln Green Liquor Sulfidity | % |
| 26 | NOXCORR | Nitrogen Oxides correlation | PPM |

## Variables Explanation

Recovery boiler operates at excessive oxygen supply to ensure perfect combustion. However, perfect combustion is not guarantee because there may be local deficiency of oxygen due to poor gas mixing. On the other hand, too much drafted air decrease boiler temperature and efficiency. Therefore, flue gas oxygen is monitored to determine the appropriate air flow rate.

Steam is produced from the recovery boiler to heat other processes. Steam flow rate is a measure of recovery boiler productivity. The higher the better. The heating value from biomass is relatively constant, subjected only to the wood species used in the digester. Therefore, a dip in steam to (black liquor) dry solid ratio signal that the boiler is not operating at optimum.

Smelt is the residue of combustion, made of mostly inorganic compound including sodium and sulfur. Smelt dissolved in weak wash is called green liquor. Green liquor sulfidity is a measure of the sulfur content in green liquor and by extension, sulfur content in smelt.

# Objective Statement

The goal of this study is to provide insight to the root cause of sulfur dioxide emission from recovery boiler. The result of the analytics project is to describe and provide recommendation to mill engineer to reduce sulfur dioxide emission through optimizing the boiler operating condition. Sulfur dioxide emission is contained in tag number 20 and tag 21.

# Cleaning Data

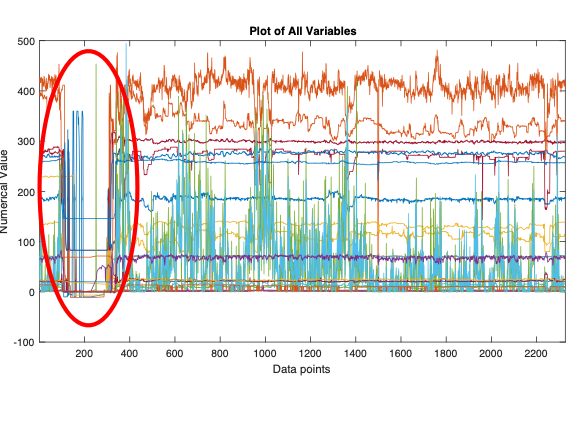
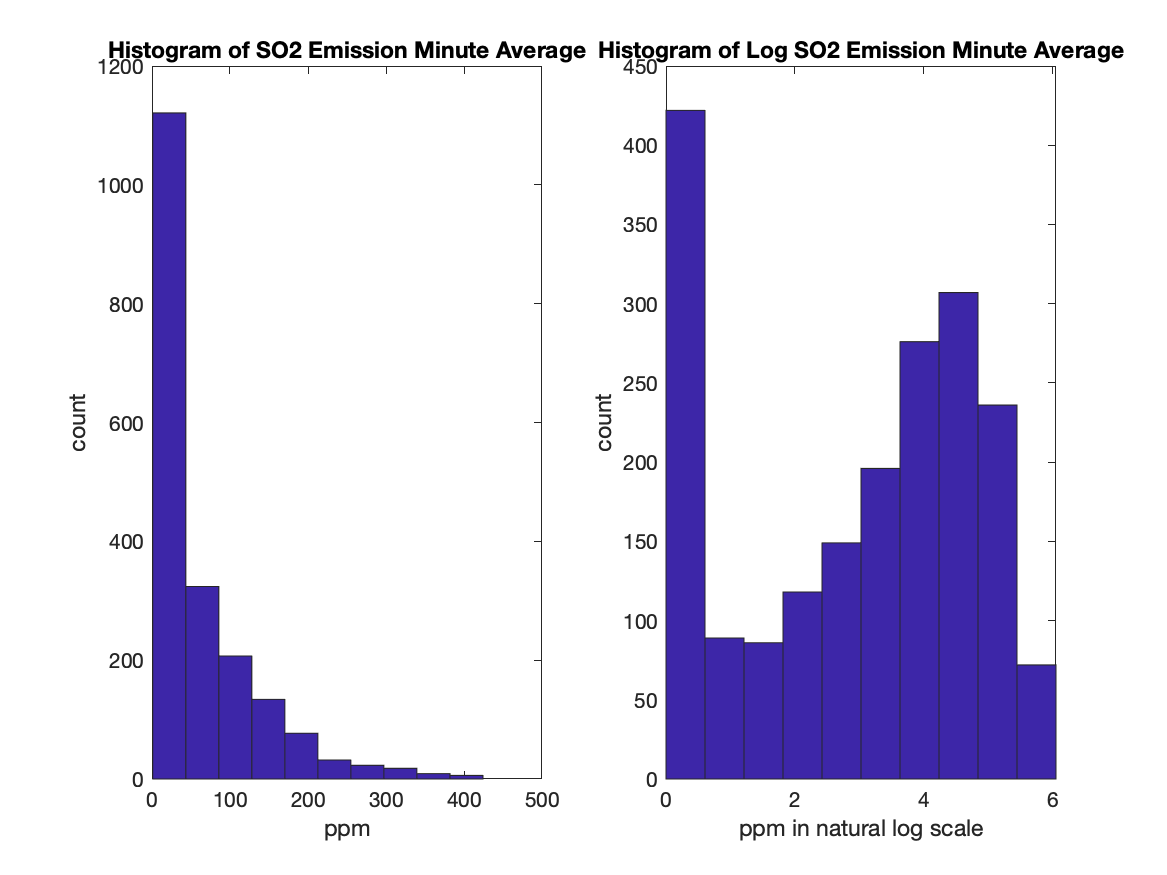
Figure 4 shows a period of no data from the recovery boiler, indicating a plant shutdown. This study focuses on steady state operation and the data before during and shortly after shutdown was removed. During startup of the recovery boiler and natural gas and other fuel is used to heat the boiler to the appropriate temperature. Therefore, the data during start up is not appropriate for this study. After removal, there are 1951 rows of record. Missing values were filled using moving average of 20 data points before and after the missing value

Figure 4: Plot of all variables showing plant shutdown

## Log transformation of SO2 emission minute and hour average

Since sulfur dioxide emission is left skewed, i.e, majority of the data concentrate on the low end, natural log transformed SO2 emission variable is created. After the transformation, the data is closer to normality. Normality is an important assumption in generalized regression. Therefore, Both minute and hour average is transformed.



## Time trend exploration

Figure 3 – 27 shows the time trend of all variables except time stamp. Primary Air Flow is steady. Secondary Air Flow drops in the first month of operation then remains steady. Tertiary Air Flow increases in the first month of operation then remains steady. Primary Air Pressure remains steady throughout the period. Secondary and Tertiary Windbox pressure shares similar trend. They decline and suddenly increase during the first two weeks of April, then decline steadily till the second week of May and then increase. Burn Rate remains steady between 260 and 280 GPM despite numerous sharp drops below 240. White liquor sulfidity is steady between 20% to 24% but the variance is greater in the first month than the second. Black Liquor solid test is steady at 70% except one data point at 0, probably an outlier. Flue gas oxygen is around 2% for the first month until first week of may, then steady at 3%. There is one peak at 16% on May 1st , potentially an disruption event or an outlier. Total Reduced Sulfur remains below 2 ppm except a few spike above 5 ppm and one spike above 20 ppm. Dry Solid Transmitter A and B shares the same trend. They largely center at 70%, which agrees with lab dry solid test. They shift upward in the first month to 71% then decline to 69% between May 1st and May 15th and increase again. Primary Air Temperature is stable and centers 298 F but the variance is higher before May 1st. Secondary Air Temperature increase from 265 F to 275 F from March 13th to April 10th, hold steady and then drops starting June 6th. Steam Flow Rate is steady around 410 KPH and shows one major dip to zero and 6 minor dip below 350 KPH. Black Liquor Solid Flow Rate is steady and center around 130 KPH and shows dips below 110 around the same time as steam flow rate does. Steam to dry solid ratio is steady at 3 except one point where steam flow rate is zero, indicating an outlier. Sulfur Dioxide minute and hourly average varies greatly 0 and 350 ppm. The variance is so great it is difficult to see trends. However, emission is visibly reduced after second week of May compared to before in terms of the peak emission and average emission. Black liquor pressure is steady and center between 20 and 25 PSIG. There is one peak at 35psig near June 5th. Black Liquor Indirect heater Temperature is steady and center around 257 F and shows some cyclical pattern in the range of 2 degree F. There is one dip to 225F at May 1st. Salt Cake Rotary Feeder Speed decreased from 14 Hz to 11 Hz between March 20th to April 24th and hold steady there since. The data is very noisy that it frequently drops to 2Hz over the data collection period, suggesting a sensor problem. Green Liquor Sulfidity drop from 26% to 23% from March 20th to April 3rd and increase to 24% and held steady there until a sharp drop to 22% on May 22th and increase to 24%. Nitrogen Oxides Correlation is steady at 69 ppm, but the data is pretty noisy. The measurement range from 55 to 75.

## Outlier Treatment

Outliers were discovered in a handful of variables during trend analysis. Outliers were filtered using DVAtool “outlier treatment” for all variables. Figure 28 in Appendix shows there most variables contains less than 5% outlier and none contains more than 10%.

## Compression and Quantification tests

Figure 29 in Appendix shows Secondary Air Flow may be over compressed and not suitable for modeling. Figure 30 in Appendix shows all variables past the quantification test. No data is lost due to quantification.

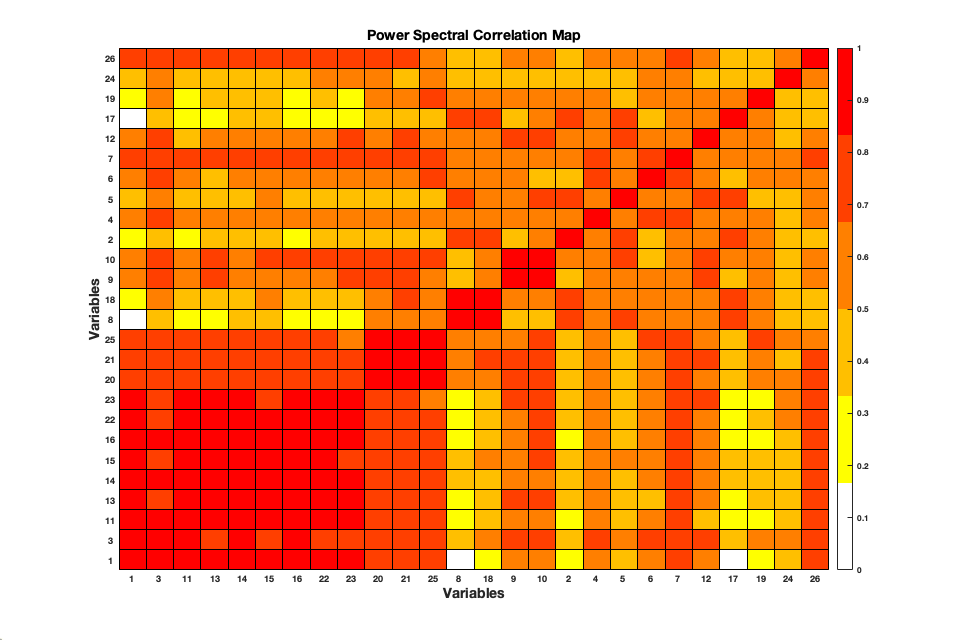


Figure 5: Power Spectral Corelation Map

# Correlation color map and Spectral Correlation Color Map

Figure 31 in Appendix shows a correlation color map that indicates the SO2 emission hour average and minute average are correlated, which is not surprising. However, there is no correlation except with time stamp and black liquor flow rate. The lack of correlation is due to time delay in the process. Mill operators recalls that changes in black liquor parameters takes minimum of 4 hours to manifest in other process data due to the large volume for heat and mass transfer to take place and slow combustion rate compared to gas boiler. Spectral analysis and lag adjustment are two useful methods to counteract this problem. Spectral analysis decomposes the data in to a series of sum of sinusoidal function with parameters such as amplitude for every frequency and phase. Spectral analysis unveiled periodic behavior in the data even if the correlation is out of phase or lagged.

The power spectral correlation map shows that sulfur emission is highly correlated with Seconary Air Flow (Tag 3), Flue Gas Oxygen (Tag11), Dry Solid Density Transmitter A & B (Tag 13,14), Primary Air Temperature (Tag 15), Secondary Air Temperature (Tag 16), Black Liquor Pressure (Tag 22), Black Liquor Indirect Heater Temperature (Tag 23), Lime Kiln Green Liquor Sulfidity (tag 25), and NOX Correlation (Tag 26).

## High Density Plot

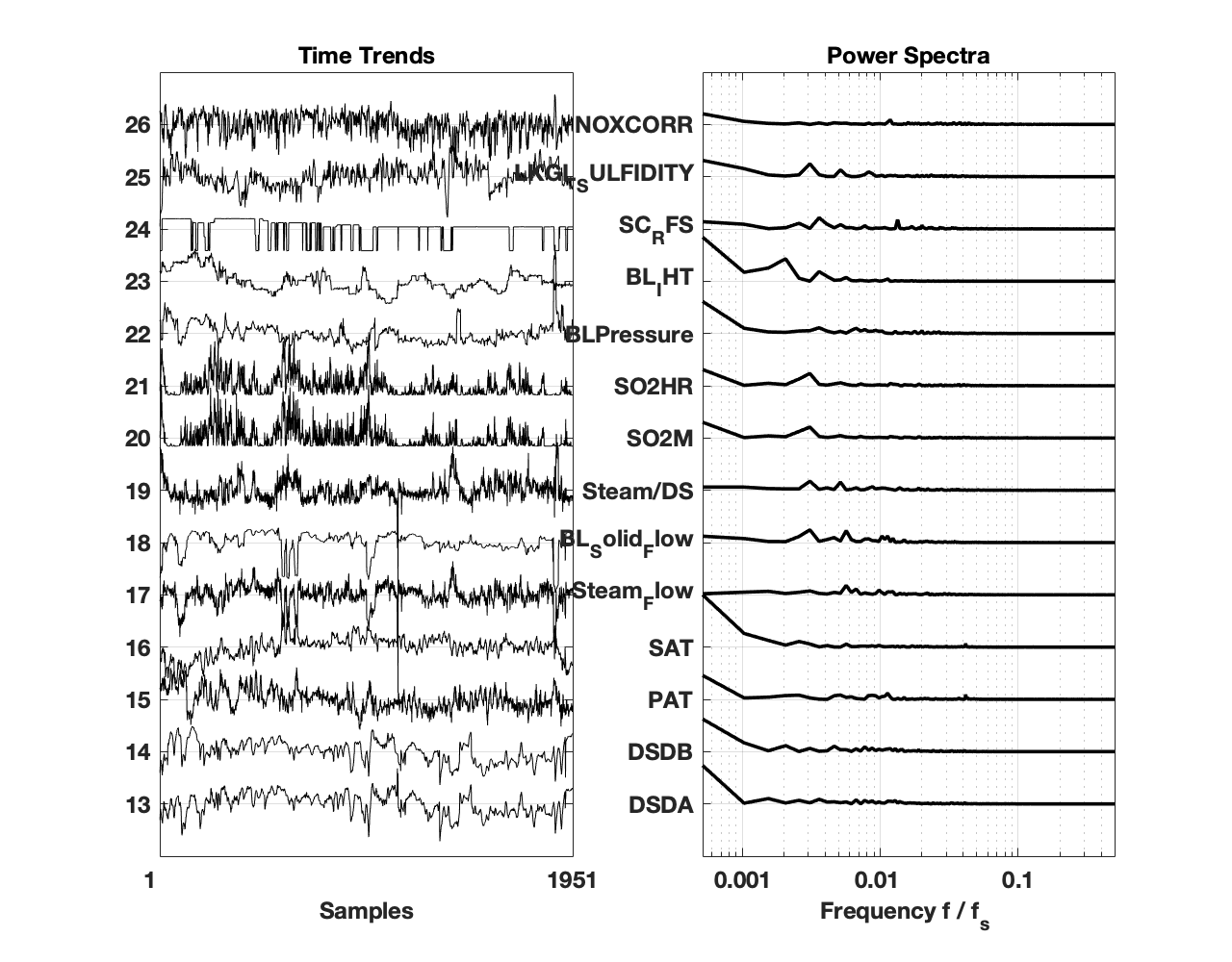
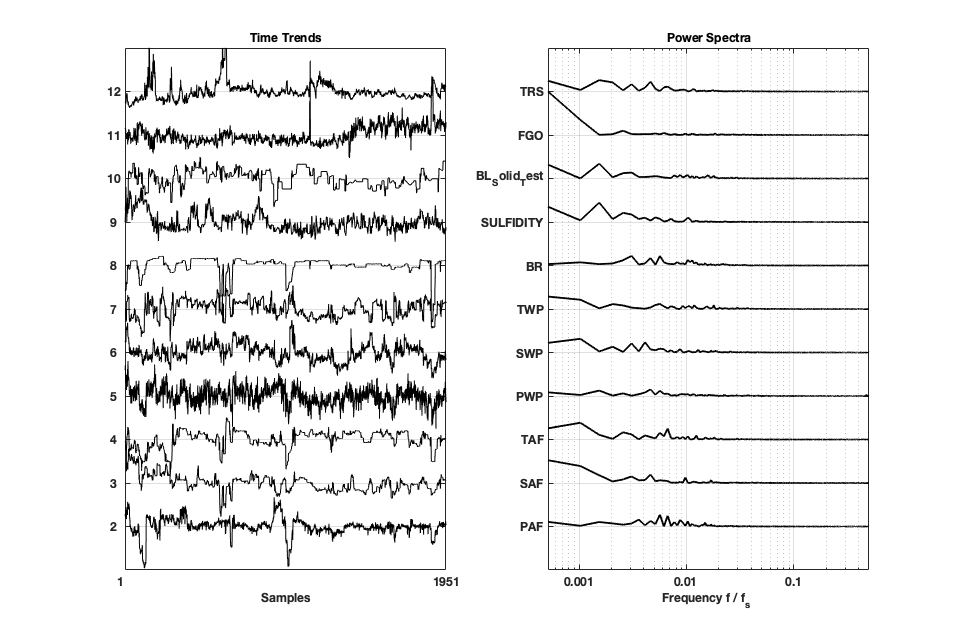


Figure 6: High Density Plot for all variables showing peak at similar frequency for Sulfidity, Black Liquor Solid Flow and Steam/Dry Solid Ratio and SO2 emission

Figure 6 shows a High Density Plot for all variables showing SO2 hourly and minute average Emission share the same frequency with green liquor sulfidity, black liquor solid flow rate, steam to solid ratio, burn rate and secondary windbox pressure. This means that these factors may be influencing each other in the boiler even though the correlation isn’t instantaneous.

## Is there periodic behavior in SO2 emission?

Autocorrelation is performed on log SO2 emission data over positive and negative 1000 lags, there is no periodic behavior in log SO2 minute emission during the period the data were collected.

## What is the lag time for SO2 emission and other variables?

Lag time is calculated using cross-correlation function for SO2 minute average and log SO2 minute average with all the other variables. Correlation is calculated between SO2 with shifted (lagged) copies of other variables for plus and minus 1000 hours. The final lag time is selected by maximizing correlation. Figure 7 shows the comparison of lag time found for SO2 Minute average and log SO2 minute average. Lag time calculated with SO2 Minute Average were -400 and +600 while lag time calculated are at most 34 hours. Table 1 shows the actual values. The lag time calculated using is log SO2 minute average is more reliable and more realistic. First, since emission is happens after every other processes, it is reasonable that the lag is negative for all variables, i.e. emission correlated better with previous measurement than the current measurement. Second, the lag time makes physical sense. For example, Flow of air lags 5 hours while steam production lags 1 to 2 hours. This is because changes in air supply affects combustion and heat convection to heat exchanger bank. It takes more time because it is further upstream in the process. It may be possible that it takes 34 hours for salt cake feed rate to affect emission because it is added to a mixing tank between black liquor concentrator and the recovery boiler. However, not everything makes physical sense. For example, Primary, Secondary and Tertiary Windbox pressure lag time is 0, -26 and -10 while they are located close in the process. Air temperatures lag time is zero, indicating that it correlates the best with current temperature, which is unlikely.

Cross-correlation function reports

Where m is the index for variable, t is the original time, tau is lag.

Higher value of R means similarity is higher and higher value is better. However, cross-correlation is not comparable with correlation. Correlation coefficient range from -1 to 1 and 0 means no correlation and 1 means always positively correlated. Cross-correlation ranges from -0 to positive infinity because it is a series of sum. Cross-correlation coefficient normalizes the sequence so that the autocorrelations at zero lag equal 1.

Figure 8 shows the comparison of maximum correlation calculated using SO2 Minute Average and log SO2 Minute Average. The correlation using log SO2 Minute Average is around 0.7-0.85 while using SO2 Minute Average the correlation is consistently low excepts correlating with itself and SO2 Hourly Average. This shows that regression should be performed on log SO2 emission to get the most out of the data.

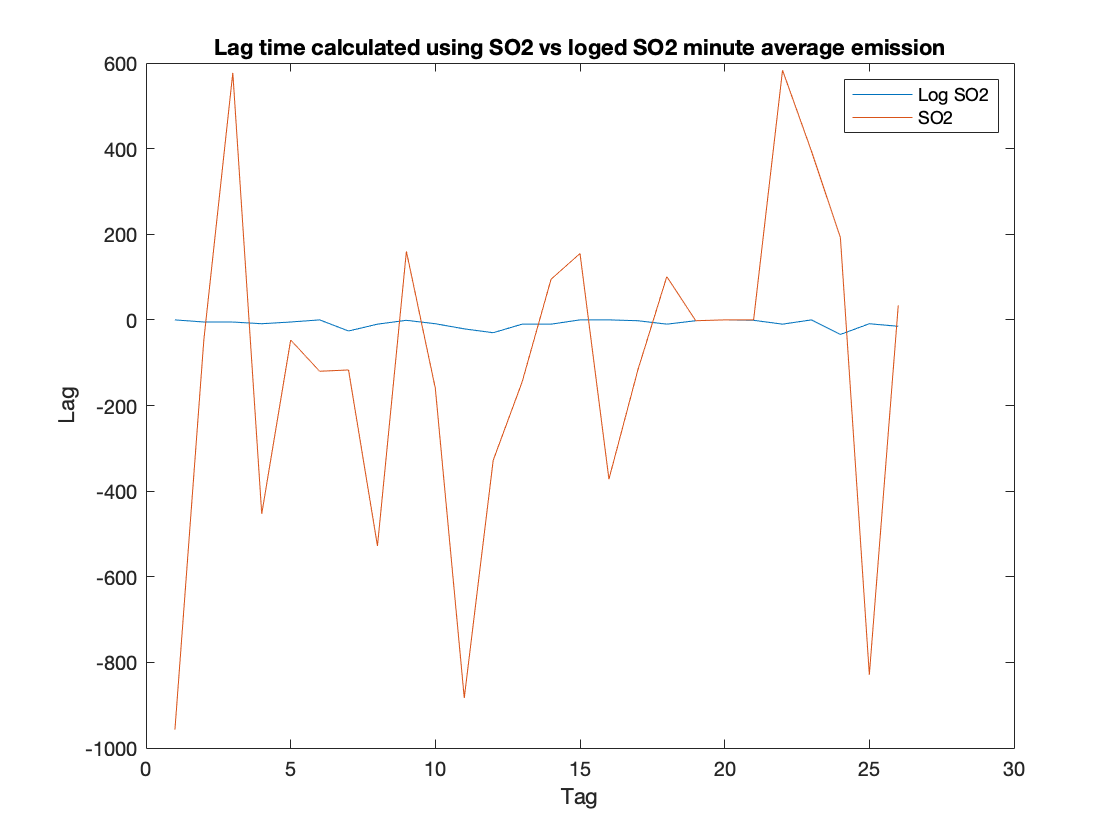


Figure 7: Lag time calculated using SO2 vs log SO2 minute average emission

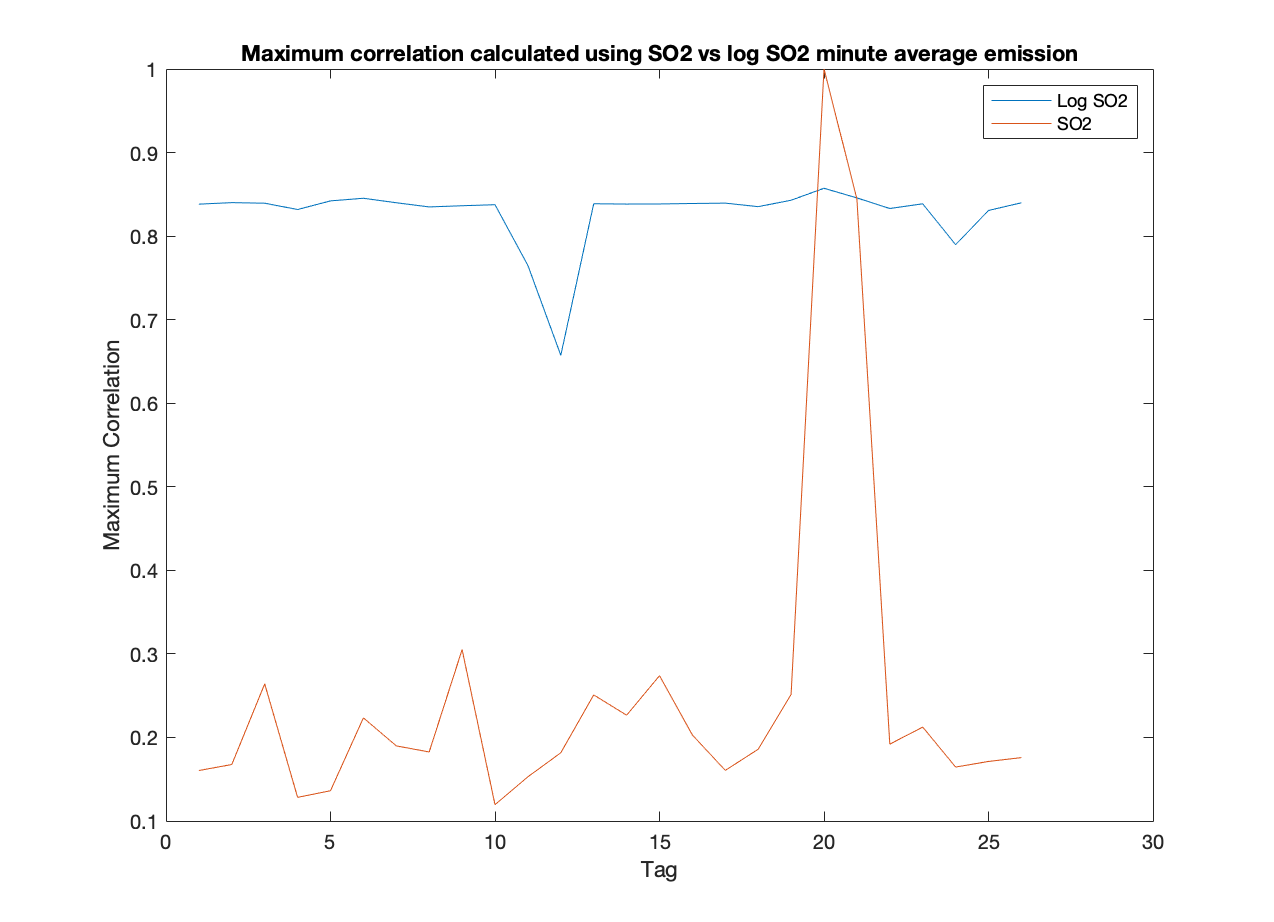


Figure 8: Maximum Cross-correlation calculated using SO2 vs log SO2 minute average emission

Table 1: Comparison of Lag time and Maximum Correlation calculated using SO2 Minute Average vs log SO2 Minute Average

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Lag time with Log | Lag time without log | Maximum Cross-correlation with Log | Maximum Cross-correlation without Log |
| Time stamp | 0 | -957 | 0.84 | 0.16 |
| PAF | -5 | -46 | 0.84 | 0.17 |
| SAF | -5 | 577 | 0.84 | 0.26 |
| TAF | -9 | -453 | 0.83 | 0.13 |
| PWP | -5 | -47 | 0.84 | 0.14 |
| SWP | 0 | -120 | 0.85 | 0.22 |
| TWP | -26 | -117 | 0.84 | 0.19 |
| BR | -10 | -528 | 0.83 | 0.18 |
| Sulfidity | -1 | 160 | 0.84 | 0.31 |
| BL\_Solid\_Test | -9 | -160 | 0.84 | 0.12 |
| FGO | -21 | -883 | 0.76 | 0.15 |
| TRS | -30 | -328 | 0.66 | 0.18 |
| DSDA | -10 | -145 | 0.84 | 0.25 |
| DSDB | -10 | 95 | 0.84 | 0.23 |
| PAT | 0 | 155 | 0.84 | 0.27 |
| SAT | 0 | -372 | 0.84 | 0.20 |
| Steam\_Flow | -2 | -117 | 0.84 | 0.16 |
| BL\_Solid\_Flow | -10 | 101 | 0.84 | 0.19 |
| Steam/DS | -2 | -2 | 0.84 | 0.25 |
| SO2M | 0 | 0 | 0.86 | 1.00 |
| SO2HR | -1 | 0 | 0.85 | 0.84 |
| BLPressure | -10 | 583 | 0.83 | 0.19 |
| BL\_IHT | 0 | 394 | 0.84 | 0.21 |
| SC\_RFS | -34 | 192 | 0.79 | 0.16 |
| LKGL\_SULFIDITY | -9 | -829 | 0.83 | 0.17 |
| NOXCORR | -15 | 34 | 0.84 | 0.18 |

Filtering outlier

Outliers are filtered using DVAtool and saved as “SO2data\_outlier\_filtered.mat” for further analysis.

## Box-Cox Transformation of Skewed data

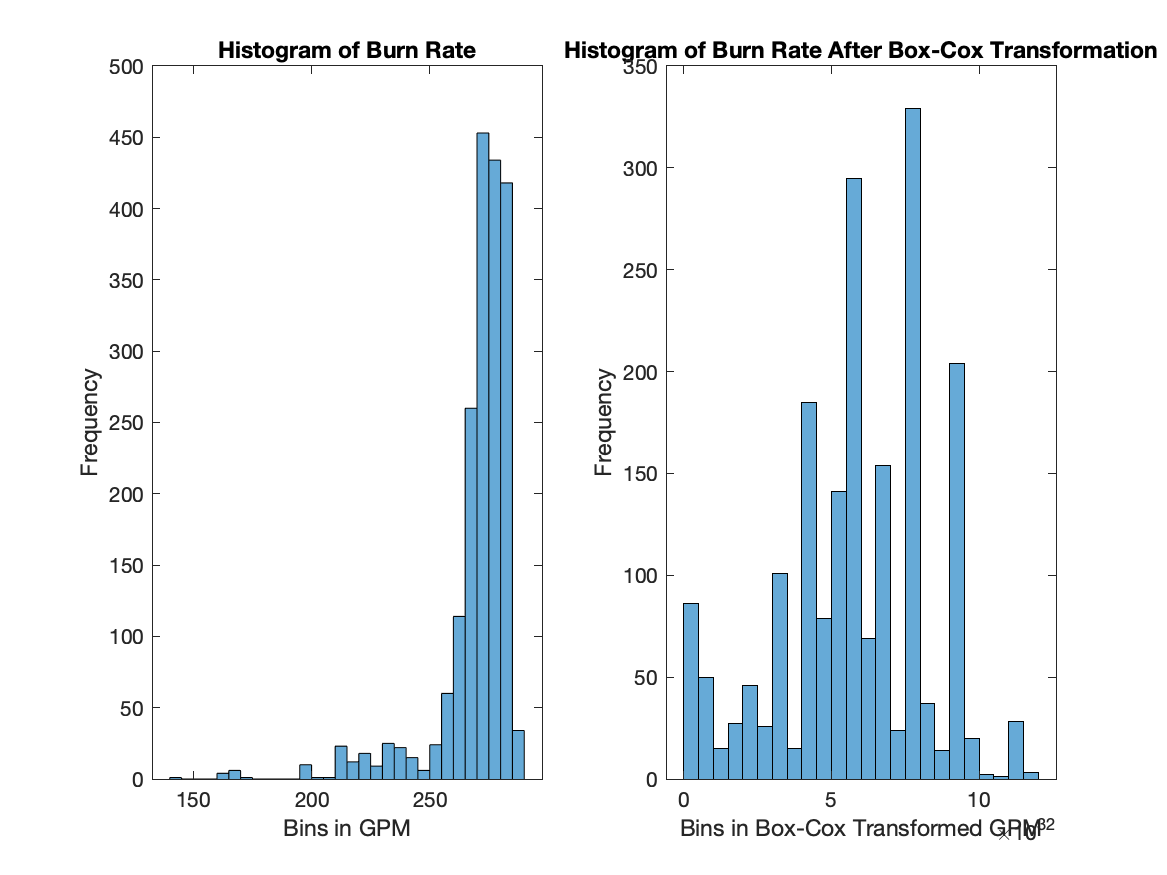


Figure 9: Comparison of Histogram of Burn Rate Before and After Box-Cox Transformation

Burn Rate was found skewed during exploratory analysis using histogram. Since linear regression models assume normality from its independent variables, transforming Burn Rate to near normal distribution is preferred. Box-Cox Transformation is a family of power transformation that use the formula

The algorithm search for parameter lambda by maximizing the log likelihood function. The transformation is reversible when lambda is found. Figure 9 compares the distribution of Burn Rate before and after transformation. The transformed Burn Rate is visibly less skewed.

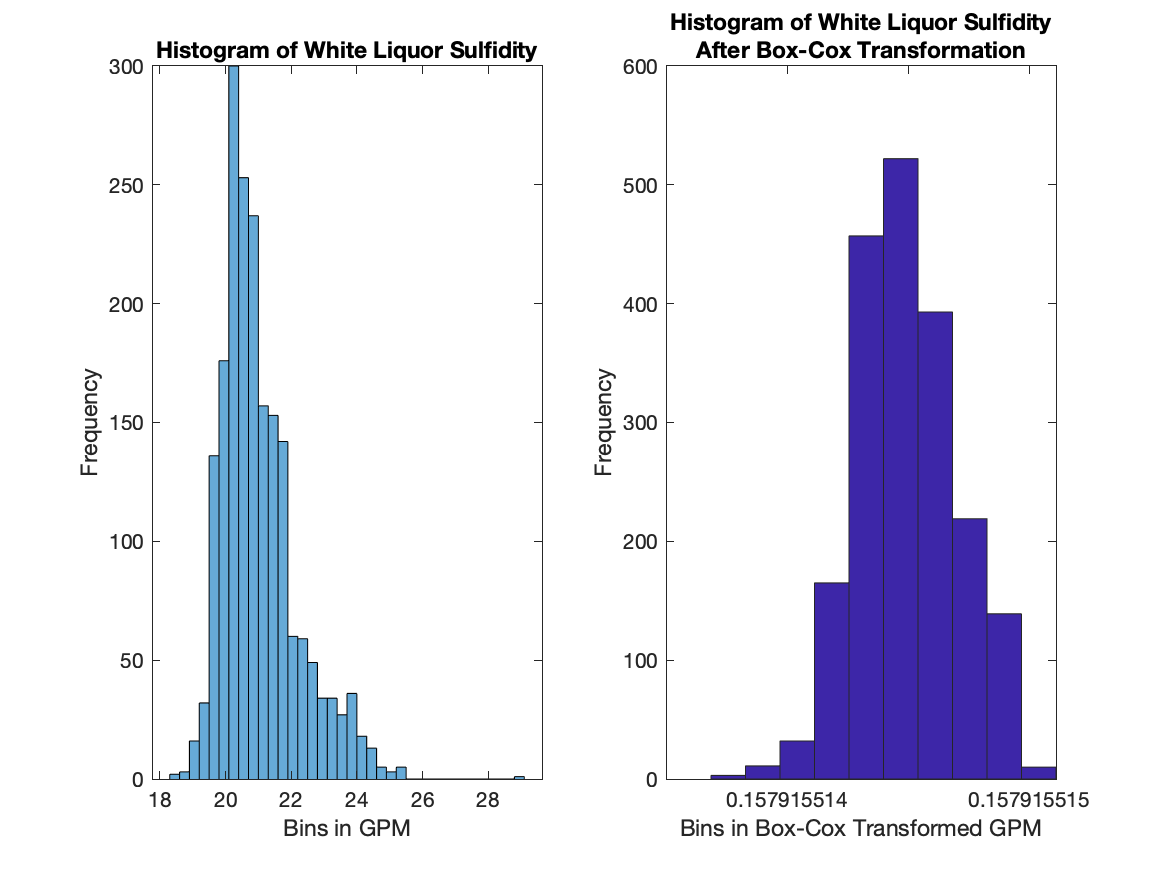


Figure 10: Comparison of Histograms of White Liquor Sulfidity Before and After Box-Cox Transformation

Similarly, White Liquor Sulfidity is left skewed, as shown in Figure 10. It was transformed to a less skewed distribution using Box-Cox transformation.

## Filtering noise using moving average filter

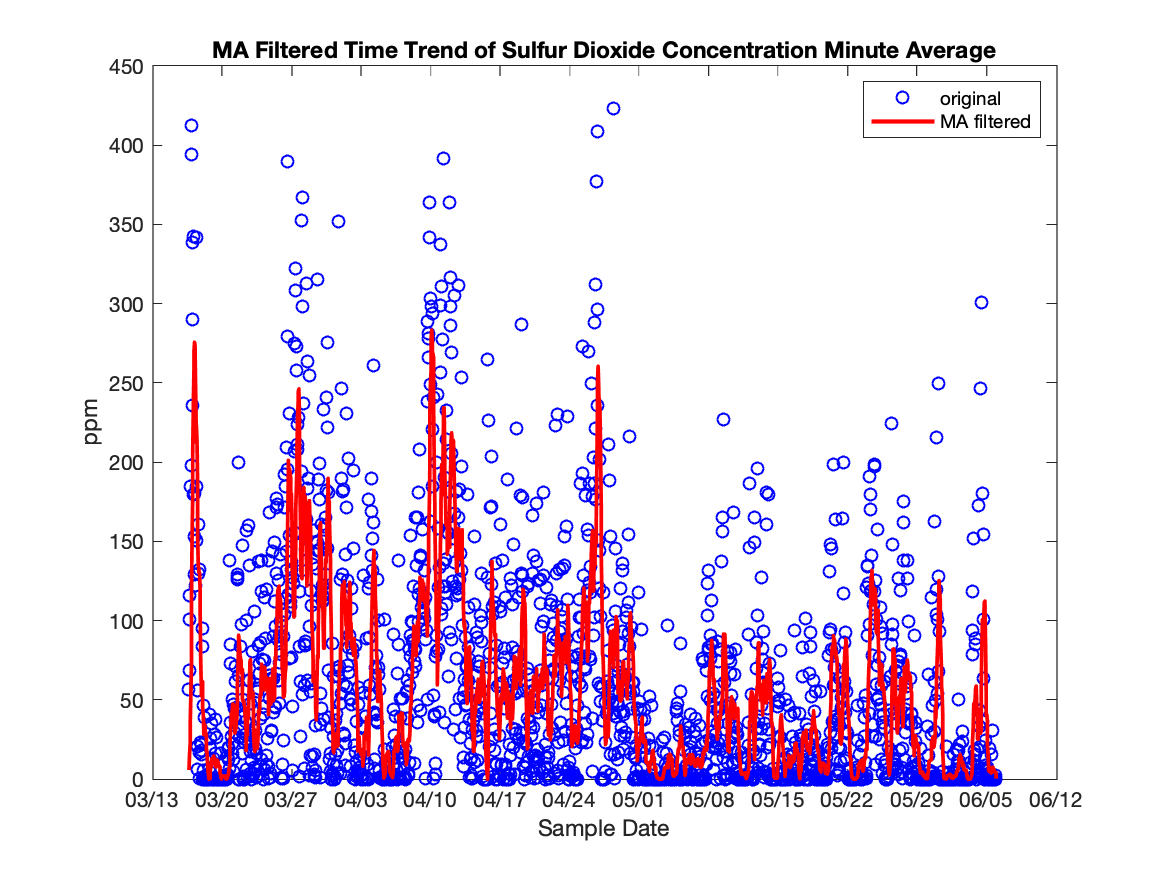


Figure 11: Moving Average Filtered Time Trend of Sulfur Dioxide Concentration Minute Average with window size = 10

To remove noise in the data set a moving average filter is employed for all variables with window size = 10. Since the data is hourly, the filter calculates the average value of the past 10 hours. Figure 11 shows the moving average values along with the original data. The peak values is reduced from 400s ppm to300s ppm. With reduced noise, the linear model will be more capable of capturing the true correlation. Will the loss in data quality improve the accuracy of the model prediction? We will find out during testing phase.

## Lag adjustment of data

From cross-correlation analysis, the lead-lag time where cross-correlation was maximized was founded. Using this result, each variable is shifted by the lead-lag time to improve correlation during linear regression analysis.

## Creating test set and training set.

The pre-processed data was randomly divided without replacement into training set and test set. The training set contains 70% of the data and the test set contains 30%. The total number of records in training and test set are 1366 and 585.

### Simple linear Regression

Backslash or divide was used to perform least square regression model training on training set treated with moving averaging filter size of 20 without an intercept.

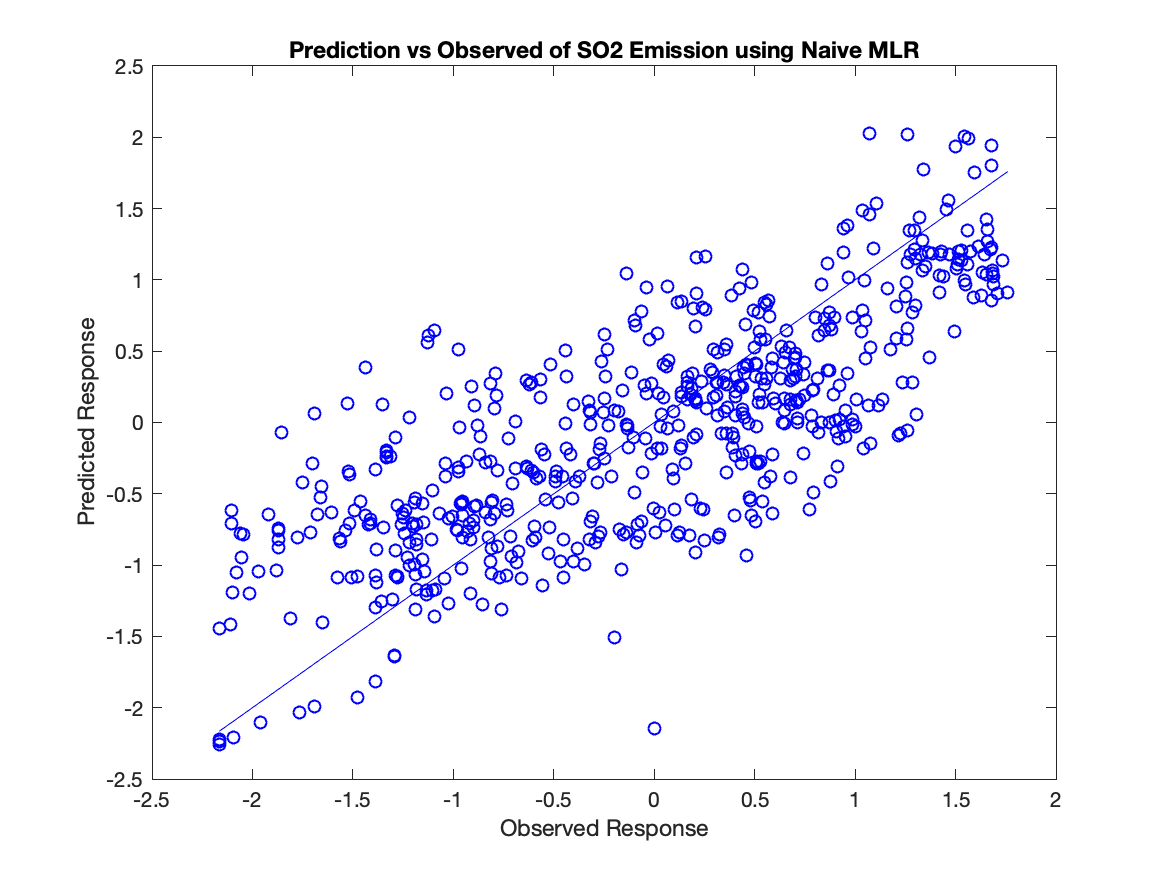
The model is then used to predict normalized SO2 emission minute average based on all 23 variables. Figure 12 shows the comparison between predicted response and the observed response. The model was able to predict the trend in the emission but with a wide margin of variance. SSE is 237.8710 and RMSE was 0.6512. R2 is 0.5866.

Figure 12: Prediction vs Observed Response for SO2 Emission using Naive MLR

Table 2 shows the model parameters for each variable. The parameters with high absolute beta are White liquor Sulfidity, Primary Air Temperature and Secondary Air Temperature. The model shows that Emission increase with Secondary Air Tempearture, Black Liquor Indirect Heater Temperature and decrease with White Liquor Sulfidity, Primary Air Tempearture, and with black liquor solid 50/50 test. The beta values for Primary Windbox Pressure, Nitrogen Oxide, and Dry Solid density Transmitter A is low, which suggestthat these factors does not contribute to SO2 emission. The result is counter intuitive to process knowledge because high white liquor sulfidity and solid density means more sulfur is available for combustion in the boiler and thus more SO2 can be potentially produce.

Table 2: Table of model parameters for Ordinary Least Square

|  |  |  |
| --- | --- | --- |
| index in Time Table | Variable Description | Beta |
| 1 | Primary Air Flow | 0.13 |
| 2 | Secondary Air Flow | 0.48 |
| 3 | Tertiary Air Flow | -0.09 |
| 4 | Primary Windbox Pressure | 0.01 |
| 5 | Secondary Windbox Pressure | 0.16 |
| 6 | Tertiarary Windbox Pressure | 0.23 |
| 7 | Burn Rate | -0.05 |
| 8 | White Liquor Sulfidity | -1.15 |
| 9 | Black Liquor Solid 50./50 Test | -0.56 |
| 10 | Flue Gas Oxygen | -0.14 |
| 11 | Total Reduced Sulfur | 0.11 |
| 12 | Dry Solid Density Transmitter A | 0.04 |
| 13 | Dry Solid Density Transmitter B | 0.32 |
| 14 | Primary Air Temperature | -1.88 |
| 15 | Secondary Air Temperature | 1.91 |
| 16 | Steam Flow Rate | 0.14 |
| 17 | Black Liquor Solid Flow Rate | -0.19 |
| 18 | Steam to Dry Solid Ratio | 0.28 |
| 21 | Black Liquor Pressure | -0.07 |
| 22 | Black Liquor Indirect Heater Temperature | 1.15 |
| 23 | Salt Cake Rotary Feeder Speed | 0.16 |
| 24 | Lime Kiln Green Liquor Sulfidity | -0.40 |
| 25 | Nitrogen Oxides correlation | 0.02 |

LASSO

LASSO is a method of regression that minimize the number of coefficient and variable used when possible. It is a useful tool for finding out what variables are the key predictors for SO2 emission.

LASSO training was performed on training set using 10-fold cross validation and the best outcome was chosen. Figure 13 shows the trace plot of coefficient fit by LASSO. Lambda, on the x axis, is the regularization parameter. The higher the lambda the fewer variables are included in the model. The Y axis is the magnitude of coefficient for each variables. LASSO algorithm computes all possible combination of models with different number of variables included. As we go from left to right to the plot, the regulariation parameter decrease and more coefficients becomes non zero, i.e. more variables are included in the model. The algorithm then MSE for these combination, find the minimum MSE and suggests the largest lambda that is within 1 standard error from the minimum MSE. Figure 14 shows how MSE as lambda decreases and the model includes more variables. The error bar of the MSE was calculated using cross-validation . The green lines in both Figure 13 and Figure 14 indicates the minimum MSE and the purple line indicates the suggested lambda. The suggested lambda value is 0.0052. which correspond to 22 degree of freedom. Minimum RMSE is 0.0112.

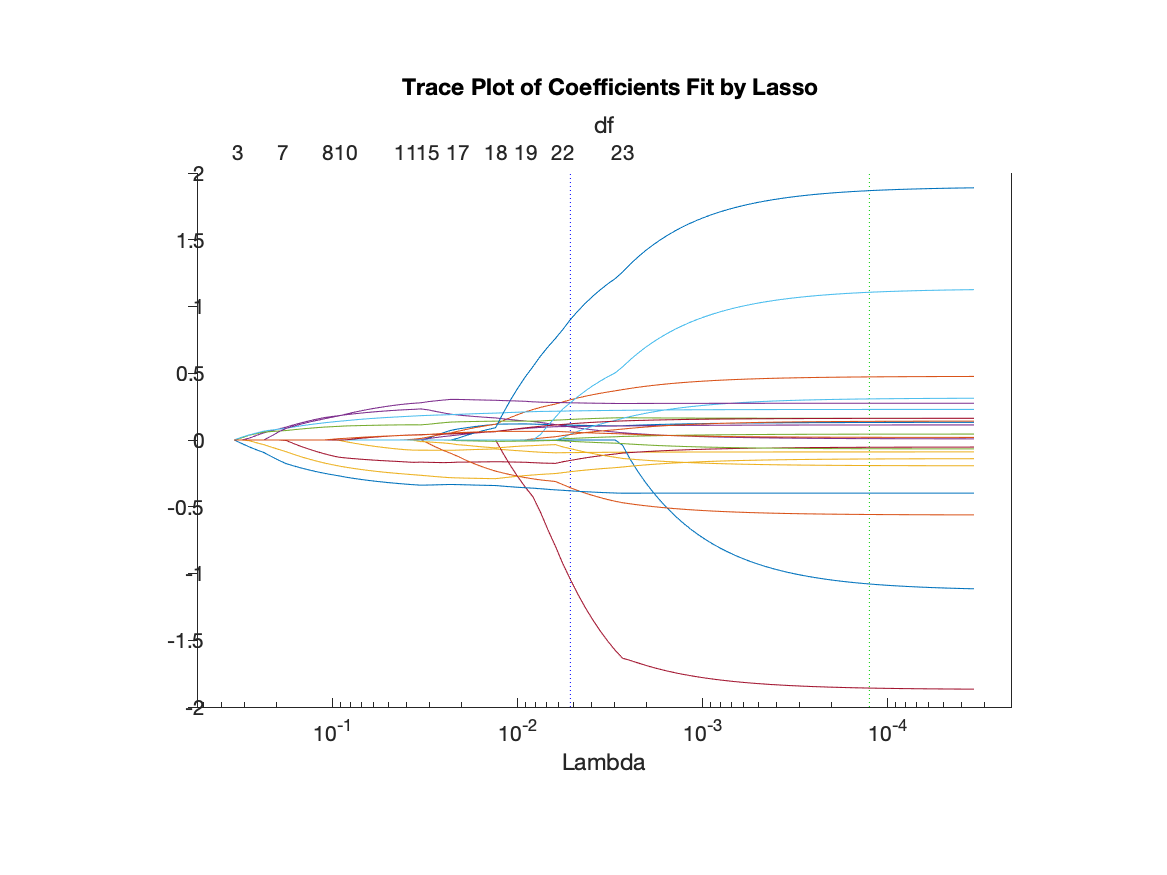


Figure 13: Trace Plot of coefficients Fit by Lasso

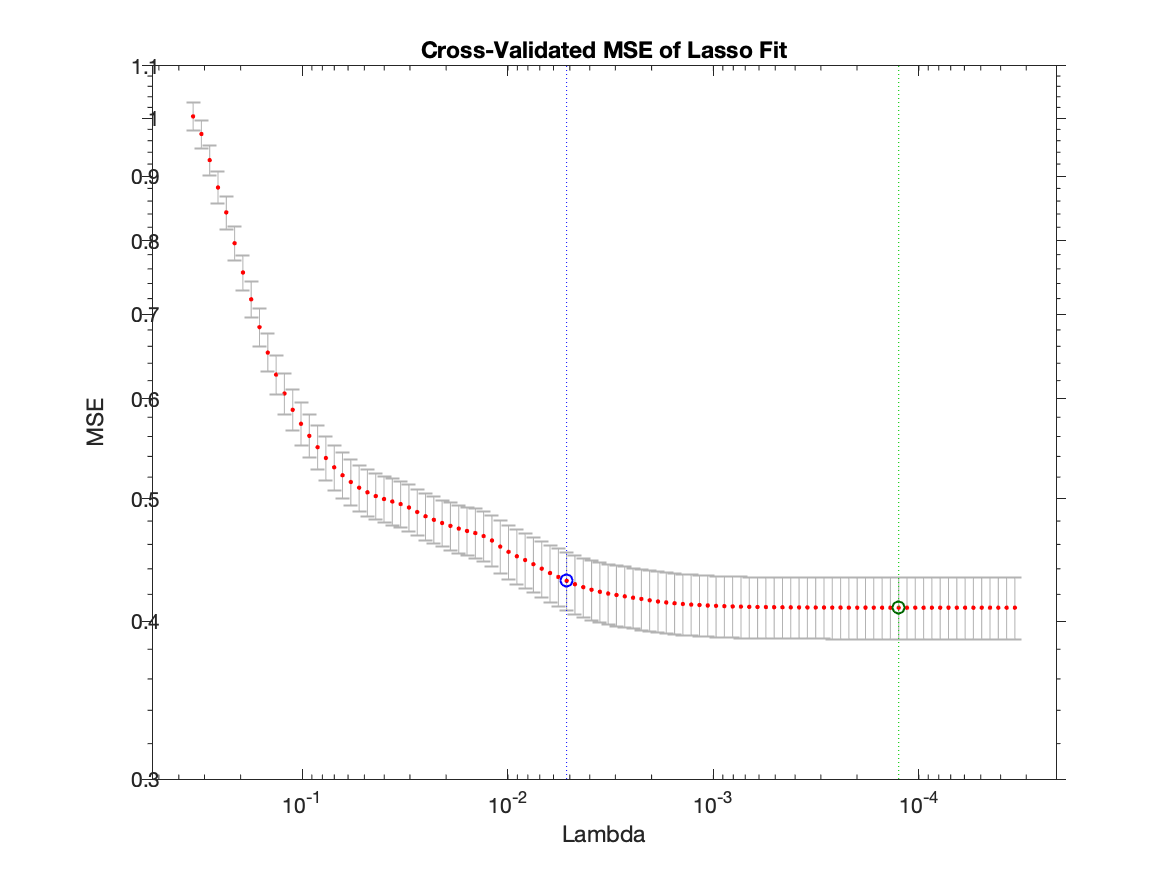


Figure 14: Cross-Validated MSE of LASSO Fit

The suggested LASSO model with 22 variables is used to test its prediction accuracy using the test set. Figure 15 shows the comparison of predicted normalized SO2 Emission minute average and the observed normalized SO2 Emission minute average. The model was able to predict the emission trend at the high end but at the low end it overestimates the emission. Unfortunately, LASSO was only able to drop the number of parameters in the model by one as it see most variables helpful in reducing prediction error.

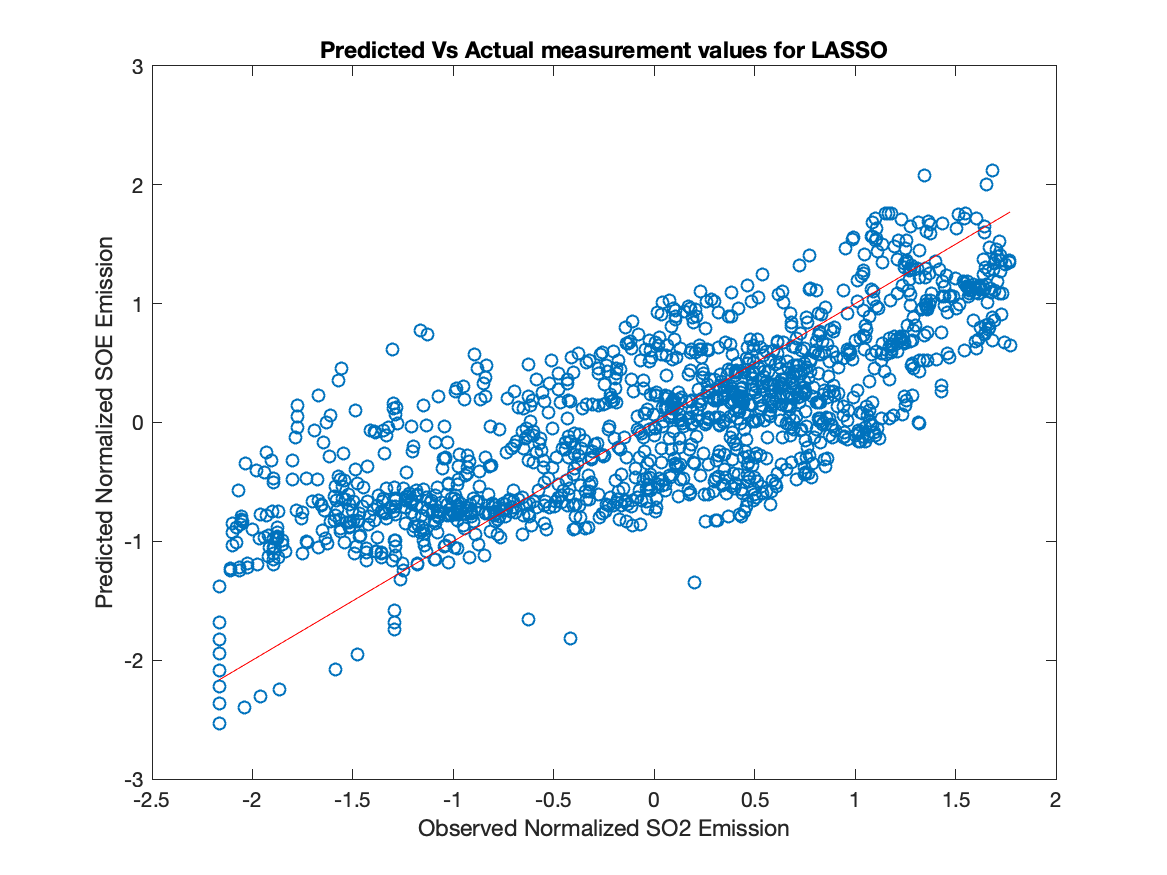


Figure 15: Predicted vs Actual Measurement values for LASSO

PCA

Principle Component Analysis is a method of dimension reduction by creating latent variables from linear combination of features in X. The method determines weather the data is along a line, a plane or hyperplane and therefore possible to describe the data using fewer variables. A dimensionally reduced dataset does not lose information but speed up calculation in the model training, especially for computationally heavy iterative model such as artificial neural network.

PCA was performed using singular value decomposition method in matlab.

Matrix u, s and v are calculated such that S is the square root of the eigenvalues of the X’X matrix.

Using diag(s) we can determine if the X matrix is full rank. Figure 16 shows the diagonal values of the matrix s. Since there is no zero in the diagonal, the matrix is full rank. There is no column that are linear combination of other column. Therefore, we cannot express a column in terms of other column. However, in the presence of noise in the data, it is difficult to determine how small the s diagonal should be to consider it essentially zero.

However, it is still possible to create latent variables to describe the data.

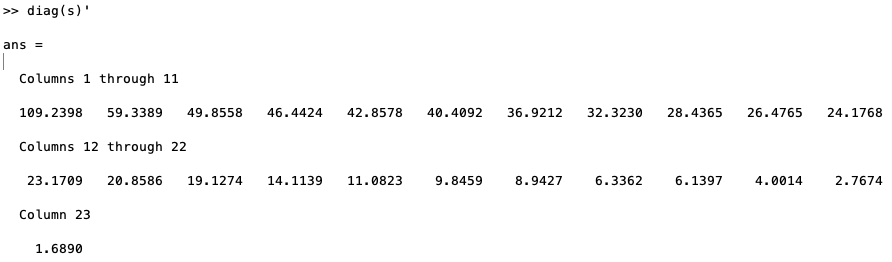


Figure 16: diagonal values of the s matrix

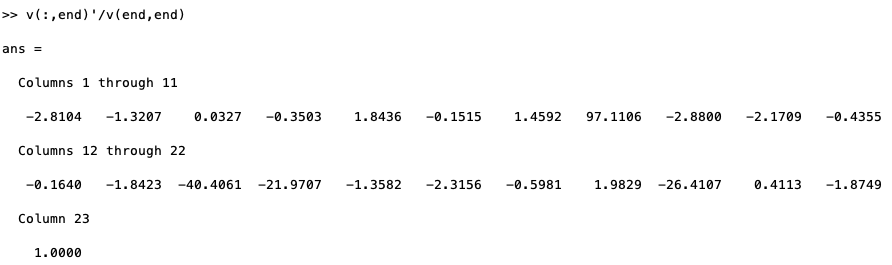
The covariance matrix of the new latent variable is

Df is 1366 for the training set.

Score matrix is U\*S. If the columns of score matrix is zero or near zero, the matrix is not full rank.

Using virtual = V(:,end)/v(end,end) The coefficient for each column in X can be calculated. The virtual variable is therefore

The value of beta is



The mean of the virtual variable is approximately zero. In this case, the mean of the virtual variable is -0.0427.

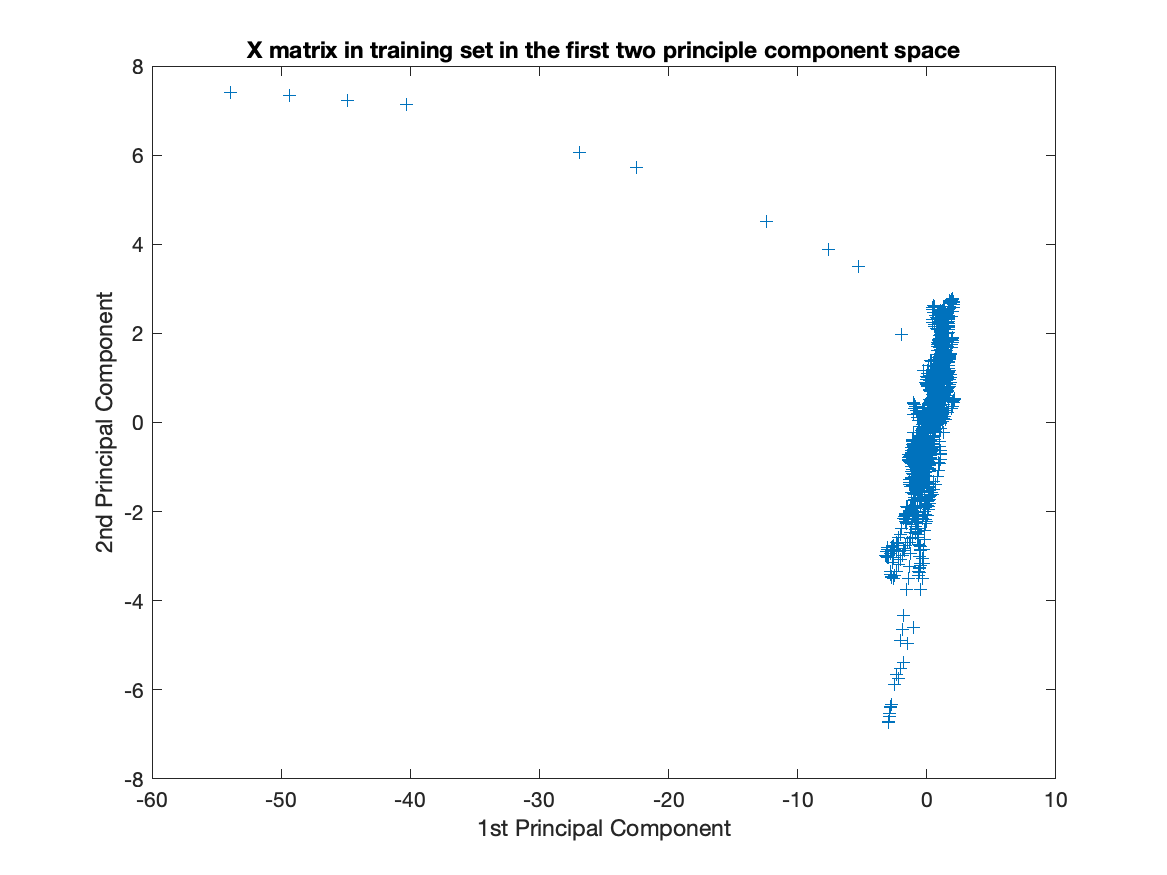


Figure 17: X matrix in training set in the first two principle component space

How does the X matrix in the principle component space looks like? Figure 17 shows the data plotted on the first two principle component.

Next the function PCA from matlab is used to find the principle component.

[coeff,score,latent,tsquared,explained,mu] = pca(X\_train);

Coeff is a p by p matrix. Each column contains the coefficient for one principle component and the columns are sorted in decreasing order of component variance. Score is the representation of X in the principle component space. Latent is the principle component variance and the eigenvalue of covariance matrix of X.

Table 3: Summary of the PCA Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| PC# | Eigenvalue | %Variance Explained | Cumulative % Variance |
| 1 | 8.7 | 39.8 | 39.8 |
| 2 | 2.6 | 11.7 | 51.5 |
| 3 | 1.8 | 8.3 | 59.8 |
| 4 | 1.6 | 7.2 | 67.0 |
| 5 | 1.3 | 6.1 | 73.1 |
| 6 | 1.2 | 5.4 | 78.5 |
| 7 | 1.0 | 4.5 | 83.1 |
| 8 | 0.8 | 3.5 | 86.6 |
| 9 | 0.6 | 2.7 | 89.2 |
| 10 | 0.5 | 2.3 | 91.6 |
| 11 | 0.4 | 1.9 | 93.5 |
| 12 | 0.4 | 1.8 | 95.3 |
| 13 | 0.3 | 1.5 | 96.8 |
| 14 | 0.3 | 1.2 | 98.0 |
| 15 | 0.15 | 0.7 | 98.7 |
| 16 | 0.09 | 0.4 | 99.1 |
| 17 | 0.07 | 0.3 | 99.4 |
| 18 | 0.06 | 0.3 | 99.7 |
| 19 | 0.03 | 0.1 | 99.8 |
| 20 | 0.03 | 0.1 | 99.9 |
| 21 | 0.01 | 0.1 | 100.0 |
| 22 | 0.01 | 0.0 | 100.0 |
| 23 | 0.00 | 0.0 | 100.0 |

Table 3 shows the summary of Principle Component Analysis. The first 5 principle component captured three quarter of the variance. The 95% of the variance in matrix X\_train can be represented in a 12 principle component space and 99% can be represented using 16 principle components.

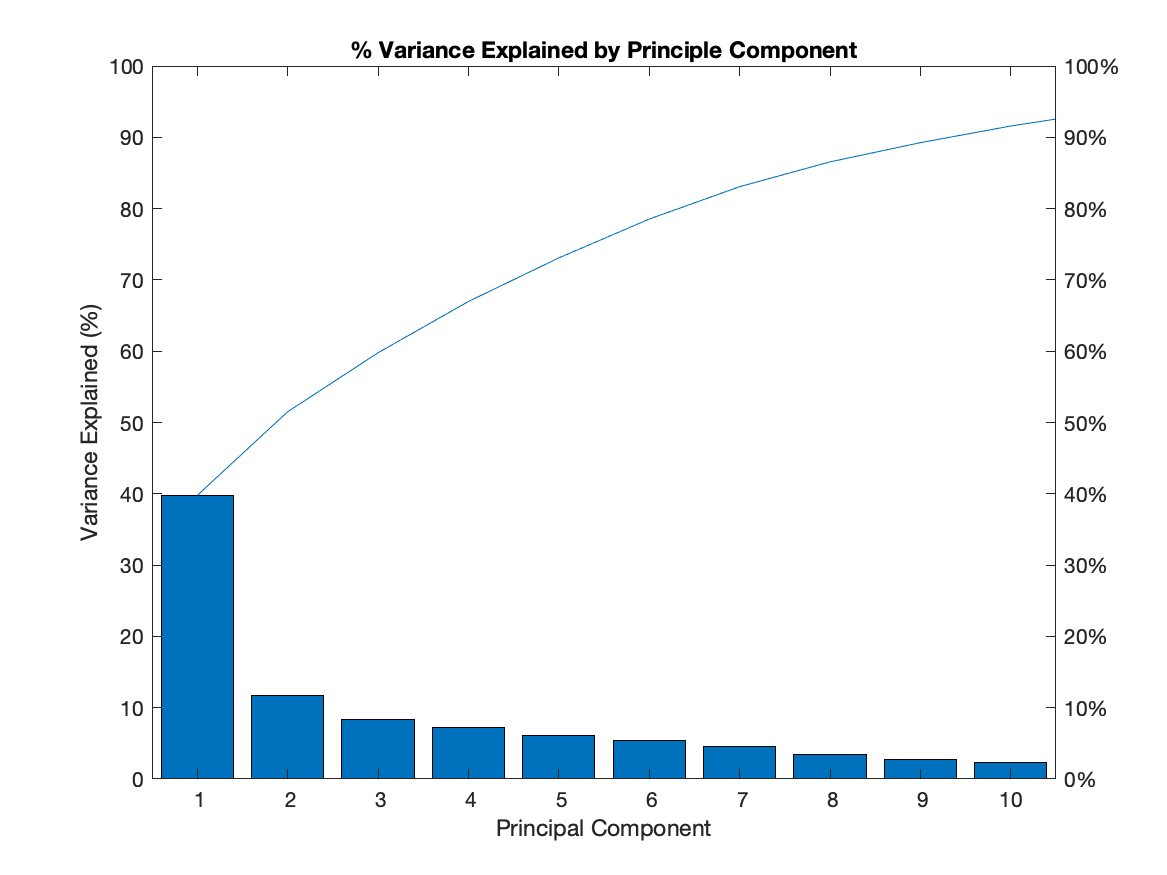


Figure 18: Percent of Variance Explained by Principle Component

Figure 18 shows how much variance is explained by each principle component and the blue line shows the cumulative variance explained. Noted that the first two PCs captured half of the variance and all the other components capture less than 10 % of variance on their own.

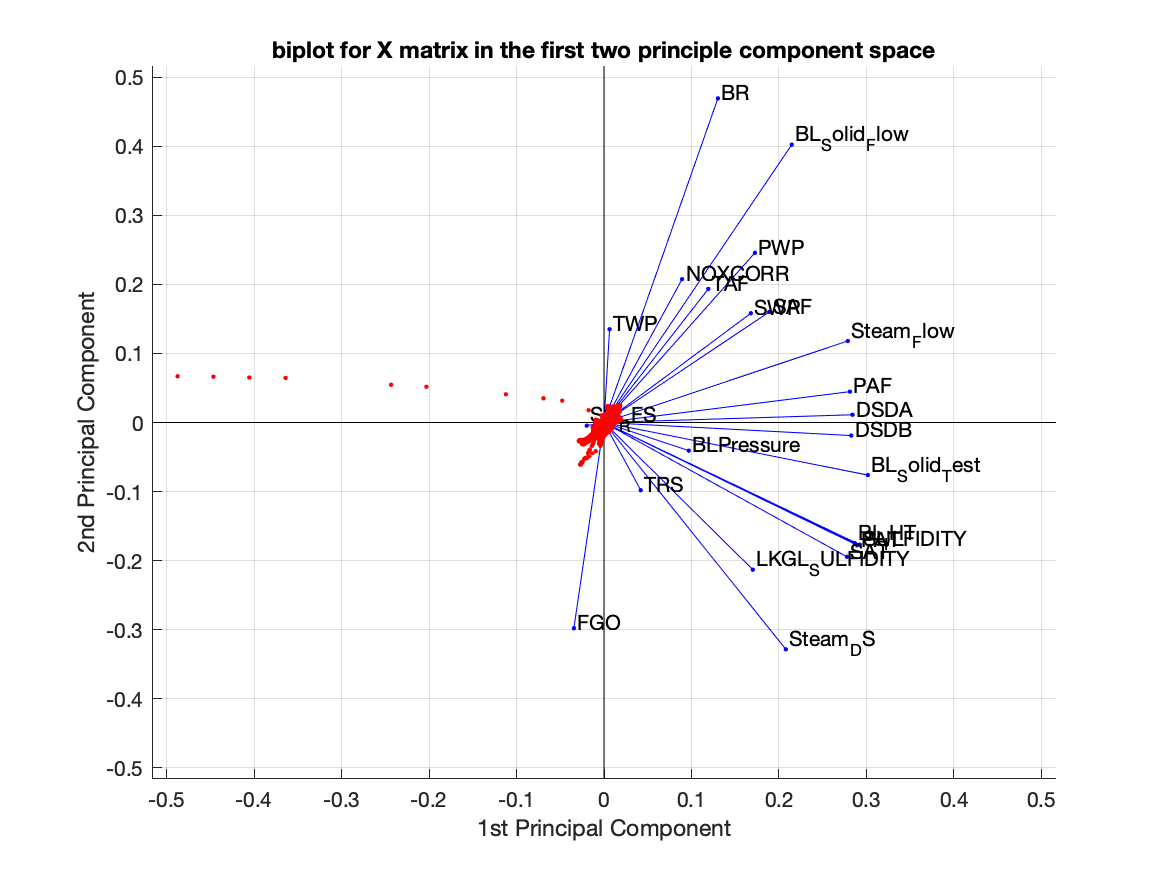


Figure 19: Biplot for X Matrix in the first two principle component space

In addition to the X\_train in the PC space, Figure 19 included the contribution of each variable to the principle component.

All 23 variables are represented in this bi-plot by a vector, and the direction and length of the vector indicate how each variable contributes to the two principal components in the plot. Note that only salt cake rotary feed speed and flue gas oxygen contributed negatively to the PCs and flue gas oxygen is significant in scale. All the other variables contributed positively to the first principle component. In terms of magnitude, burn rate, black solid flow and black liquor heater temperature and white liquor sulfidity are significant contributors to the first two PCs.

The result of principle component analysis is used to perform principle component regression. All 23 principle components were used in regression. Figure 20 shows The prediction vs observed value for normalized so2 emission using PCR. The model was able to predict the linear trend in the data set as the points scattered along the 45 degree red line. Numerical analysis shows that PCR‘s goodness of fit is 0.5866 and RMSE is 0.6512, which are the same as using ordinary least square. Using fewer components

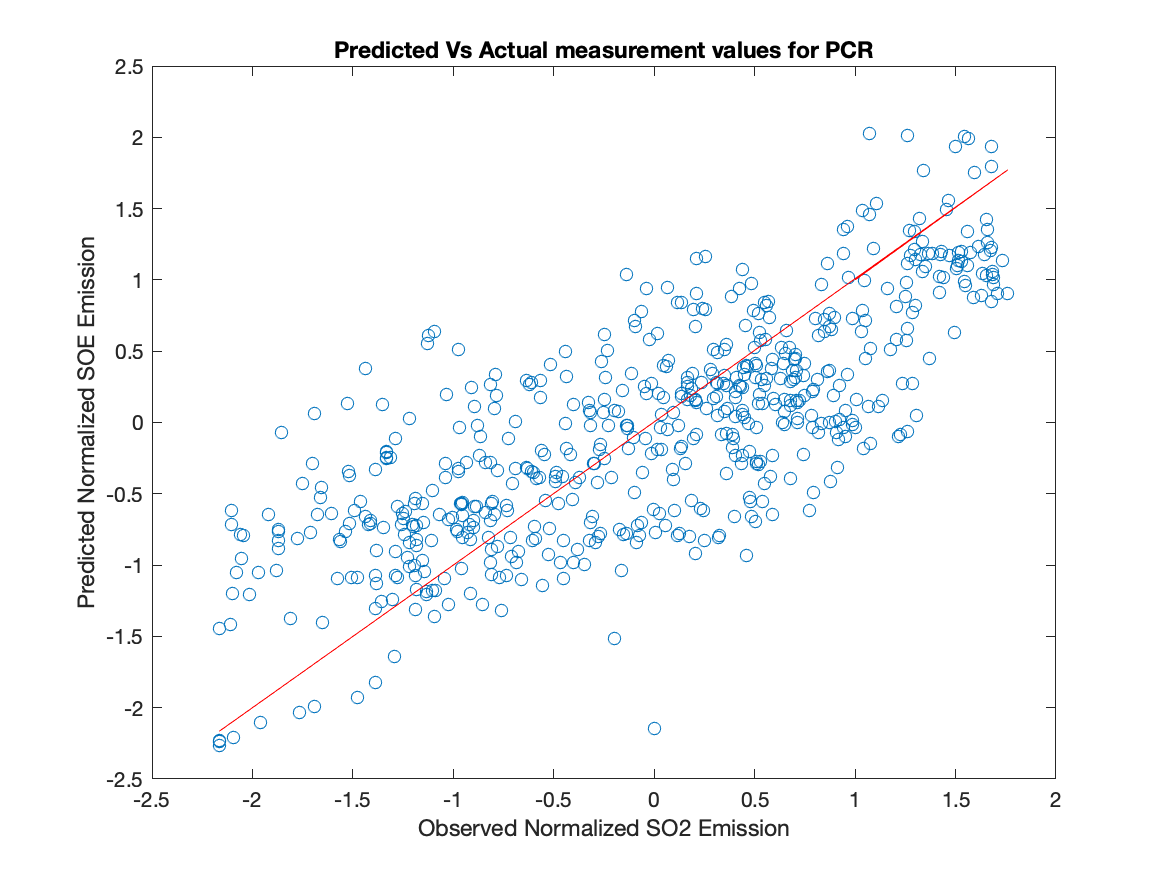


Figure 20: Predicted vs Observed SO2 Emission for PCR

Partial Least Square

PLS is a method of feature compression that also consider the Y component. It converts a MIMO problem into a series of SISO and piece the result of SISO regression together to arrive at the result for the MIMO problem.

PLS was performed on the training set using 10 fold cross validation.Figure 20 shows the percent variance explained in Y vs Number of PLS components. The first two component was able to explain around 50% of variance and with 10 PLS components it was able to explain 57% of variance in Y.

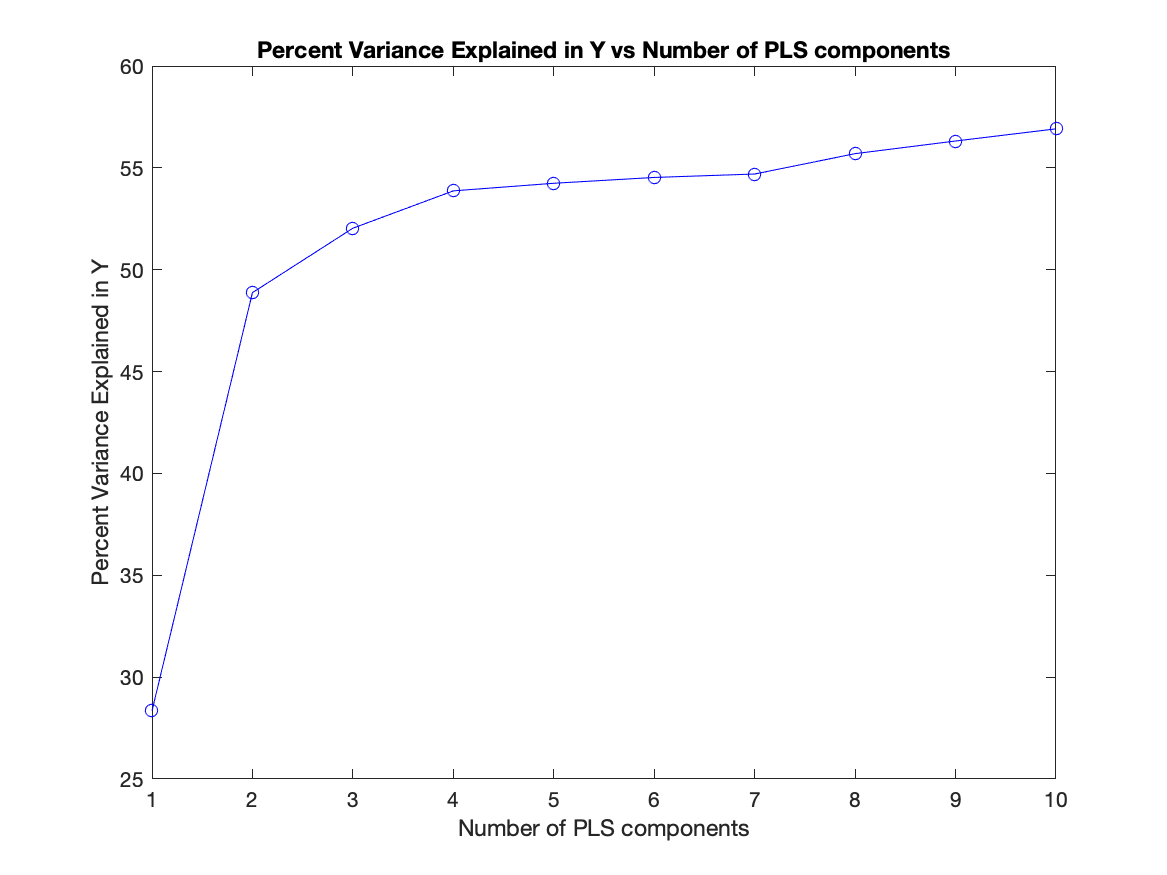


Figure 21: Percent Variance Explained in Y vs Number of PLS components

PLS was used to predict SO2 emission minute average from X\_test. Figure 21 shows the residuals from PLS prediction. The distribution of residual on the positive and negative side appears even and there is no linear trend in the residuals. Therefore, the residuals are randomly distributed and the model captured all linear pattern in the data. Figure 22 shows the comparison of predicted normalized SO2 emission vs the measured value. The red line indicates the 45 degree line. If the model predicts everything correctly, all dots will lie on the line. If the data point is above the line, the model overestimated the value. If the data point is below the line, the model underestimated the value. The model was better at prediction than only using the average values and the error on each side is even.

Numerical analysis shows that the model goodness of fit (R square) is 0.55 with RMSE 0.66 for the normalized data. Increasing the number of components from 10 to 23 in PLS results in goodness of fit to around 0.58 and RMSE at 0.64. That is to say that the first 10 components already captured most of the trend and further increase in number of components only provide limited improvement in accuracy. Figure 24 captures the variable weights from the first 10 predictors. Primary Air Temperature(top red), Black Liquor Solid Test(top orange), Secondary Air Temperature(bottom blue) is visibly carrying more weight than others.

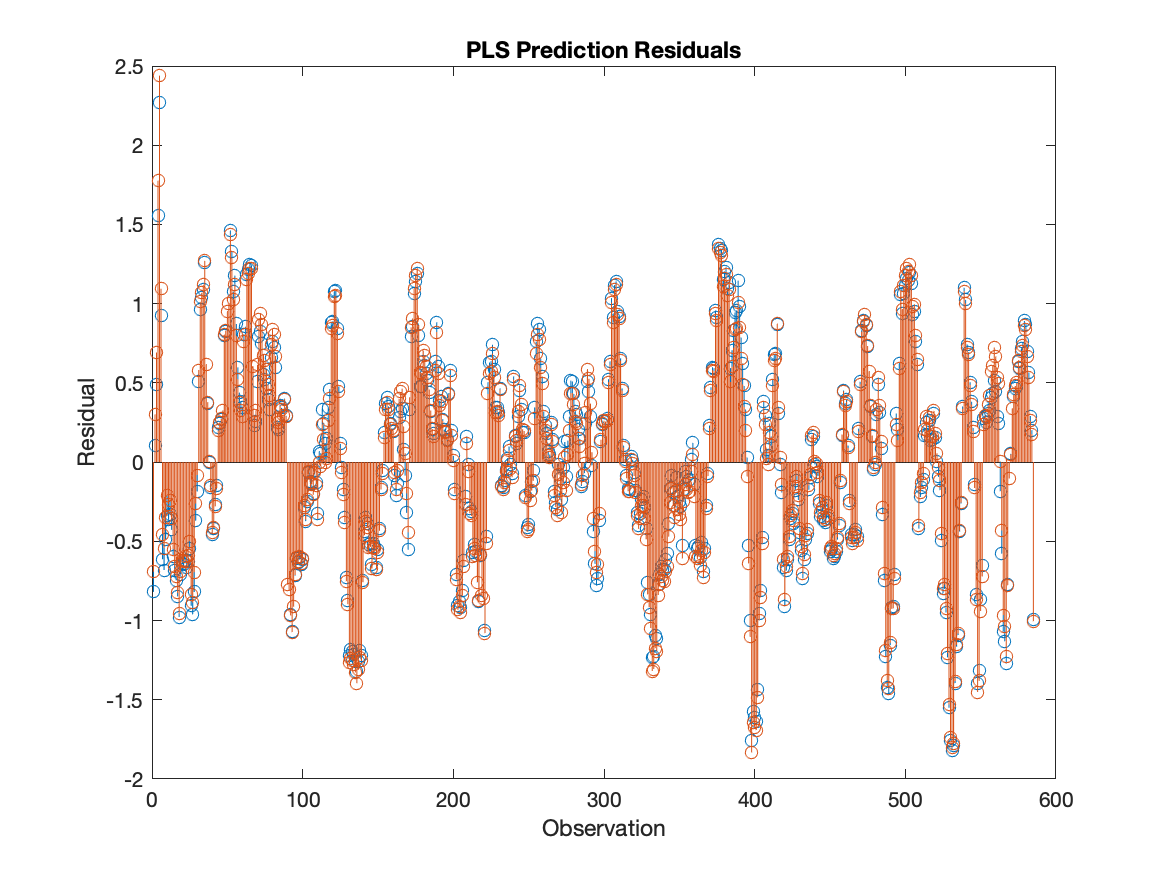


Figure 22: PLS Prediction Residuals

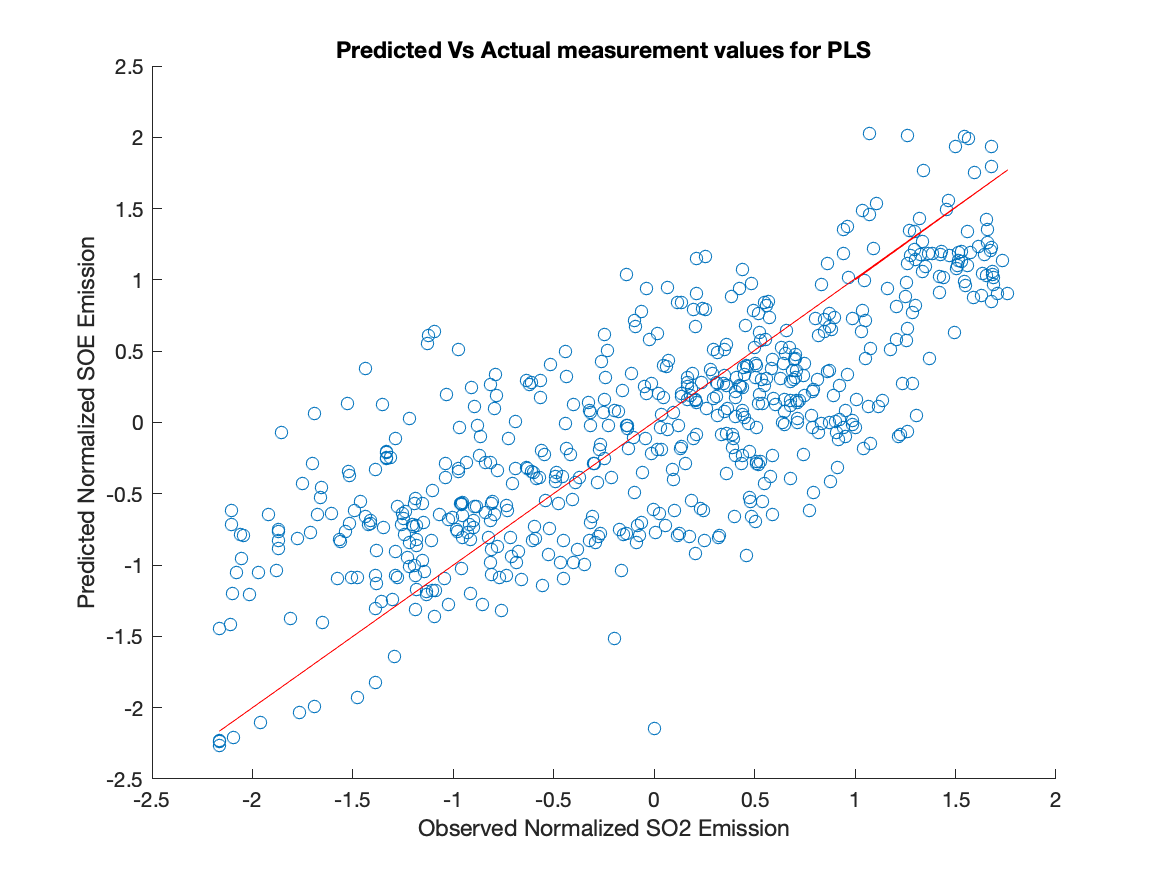


Figure 23: Predicted Normalized SO2 Emission vs Observed for PLS

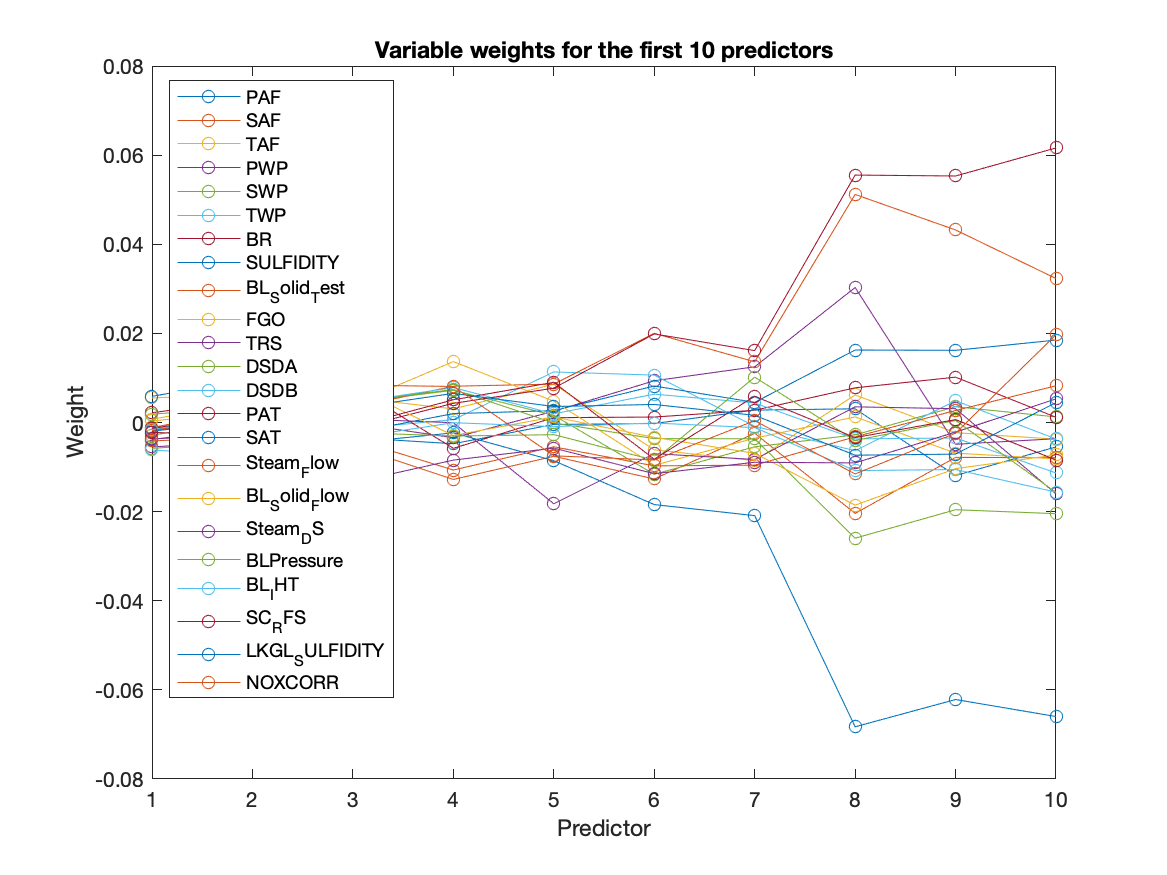


Figure 24: Variable weights for the first 10 predictors for PLS

Regression Tree

Regression Tree is used to train on the SO2 emission training data with 10-fold cross validation and predicts the emission in the test set. Without limits to the number of splits, the model achieved RMSE of 0.3289 and goodness of fit of 0.8921 in prediction. The performance of regression tree is better than all the other methods. However, there are many branches and won’t fit in a plot. How many branches are in the tree? The text form of the tree contains 467 split and nodes. Figure 24 shows the prediction vs observed normalized SO2 emission. The red line is the 45 degree line. The fact that data points are closely aligned on the line is an indicator of better accuracy in prediction.

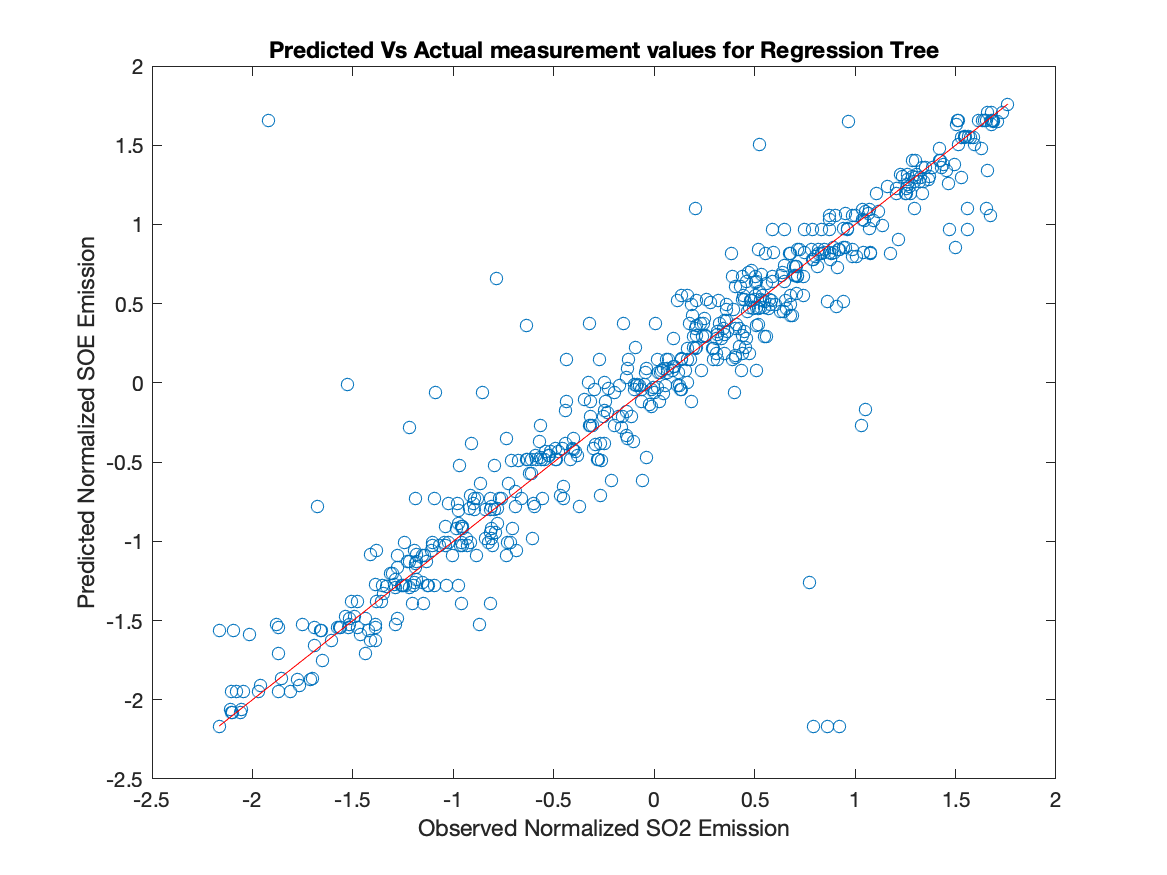


Figure 25: Prediction vs Observed SO2 Emission Using Decision Tree

Does the minimum number of the leave in node matters? Optimization of parameter for regression tree was performed using 'OptimizeHyperparameters','auto' option. The method calculate 30 cases with different minimum number of leave sizes. Figure 24 shows the objective function vs different minimum leaves size. Figure 25 shows the minimum objective function for different function iterations. The analysis shows that one or two minimum number of leave is the best parameter for this data. I will use the default setting where minimum number of leave is one.

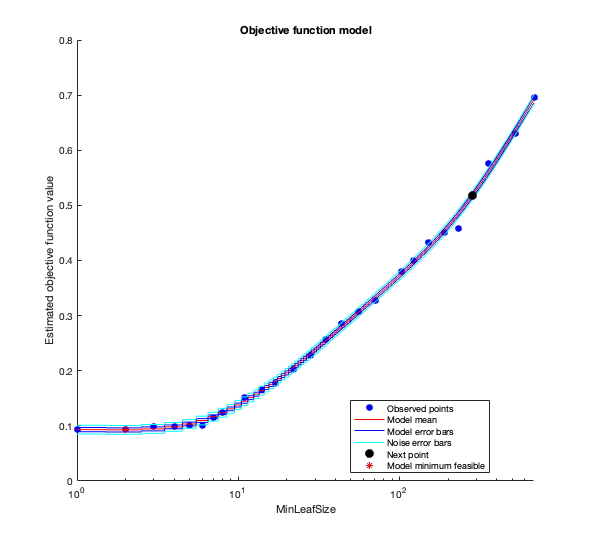


Figure 26: Regression Tree Estimated objective function values vs minimum leaf size

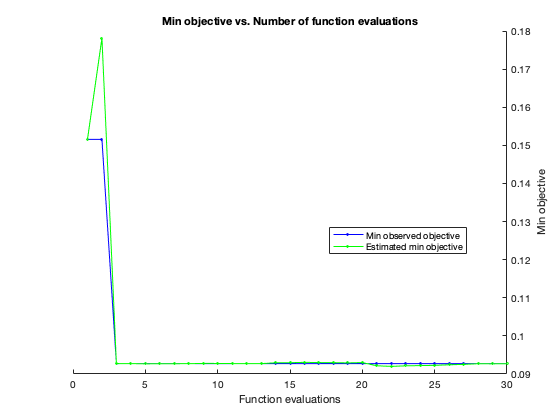


Figure 27: Minimum objective vs number of function evaluations

What is the cross-validation error for the 10 fold cross validation? The cross-validation error is 0.0830.

What are the contributing parameters in predicting SO2 emission? The Predictor Selection option estimate predictor importance values by summing changes in the risk due to splits on every predictor and devides the sum by the number of branch node. Figure 28 shows the predictor importance estimates and that Primary Air Flow, black liquor indirect heater temperature, and Secondary Windbox Pressure are the top three important predictors in the model. It was followed by Primary Air Temperature and Tertiary windbox pressure. The predictor importance is calculated by summing difference in MSE before and after the split for every predictors.

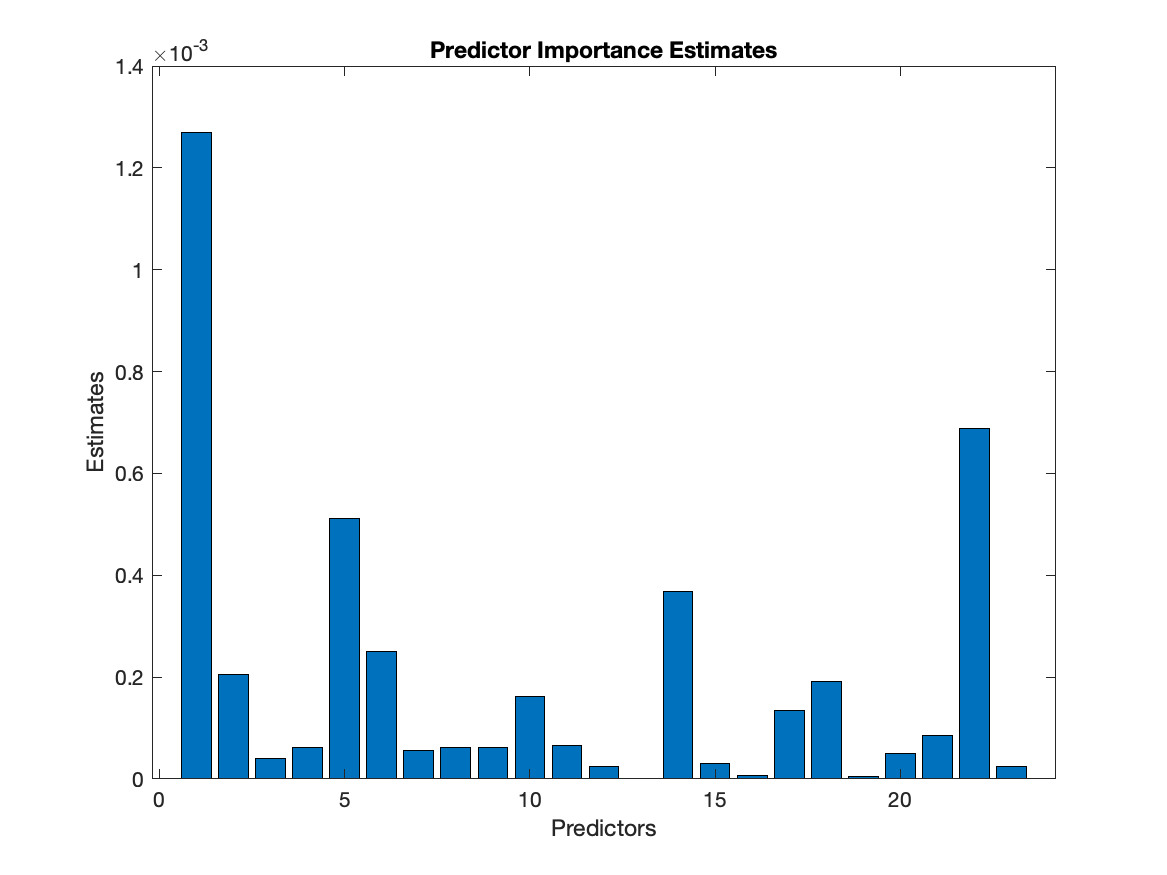


Figure 28: Predictor Importance Estimates

Conclusion

Ordinary least square, elastic net regularisation (LASSO), principle component regression, partial least square and regression trees were used to model the SO2 emission minute average data from a pulp and paper mill. The data was preprocessed by removing section that are related to shutdown, fill missing using moving averages, find lag and adjust for lag to maximize correlation, log and Box-Cox transformation for skewed data, filtering noise in in the signal and normalization. Training set and tests set are created by random sampling without replacement and training set contains 70% while test set contains 30%. The model is then used to predict SO2 values given all the other variables. Table 4 shows the comparison of regression methods. Ordinary least square uses all 23 variables was able to achieve R-square values of 0.58 and RMSE at 0.65. LASSO is method that can select fewer features to achieve the same prediction. However, in this case LASSO was only able to reduce number of variable by one. One possible explanation of LASSO failure to reduce the number of variables is the level of noise in the data. It is difficult to tell how much variance is due to noise and how much variance is caused by factors that wasn’t measured or available in the data set. I feel that a moving average with window size 20 hours is already very generous and result produced with window size larger than a day becomes difficult to interpreted in a physical sense. Nevertheless, I expect LASSO performance will improve if the filtering is more aggressive and LASSO is less likely to regress noise.

Principle component analysis performance converge with OLS at high number of component. In fact, the method produced the identical r-square values and RMSE as OLS despite regress in a different space. This is expected because the data didn’t change and only the axis is reassigned. At lower number of principle components, PCR performs reasonably well. With 10 PCs, it achieved Rsquared value of 0.47 and RMSE of 0.73. Compared this to LASSO in Figure 14, for 10 variables which is equivalent to lambda equal to 10, gives a MSE of 0.55, or RMSE of 0.74. PCR is expected to perform better than LASSO or OLS because its components includes variables that captures most variation in the independent variables.

Partial Least Square is a better method at feature reduction and prediction. This is because it consider both dependent and independent variables when creating its component. Using PLS with 10 components, I was able to predict with goodness of fit at 0.55 and RMSE at 0.66, which is comparable to the performance of OLS using all variables

Finally, regression tree has the best performance of all method at making accurate predictions because it is not bound to continuous linear models. It can group predictors into discrete branches and calculate averages for each leaf. During testing, it gives a goodness of fit of 0.86 and RMSE at 0.36. As the number of leaf increase, prediction improves to a point then decrease due to overfitting. If the number of split is limited at 30, R-square decreased to 0.74 and RMSE at 0.5, still better than OLS. Figure 27 shows the visualization of the a tree with 30 splits.

In production, these predictions need to be convert to their original units using the mean and standard deviation of the training set. Then, the predication will track the 20 hour moving average emission 50-80% of the time, according to their R square value.

## Comparison of regression accuracy by models

Table 4: Comparison of Regression Methods and Accuracy

|  |  |  |  |
| --- | --- | --- | --- |
| Method | RMSE | R square | number of variables/ components/ split for trees |
| OLS | 0.6512 | 0.5866 | 23 |
| LASSO | 0.6695 | 0.5623 | 22 |
| PCR | 0.6512 | 0.5866 | 23 |
| PCR | 0.7369 | 0.4706 | 10 |
| PLS | 0.6699 | 0.5524 | 10 |
| Regression Trees | 0.3669 | 0.8657 | max |
| Regression Trees | 0.5056 | 0.7451 | 30 |

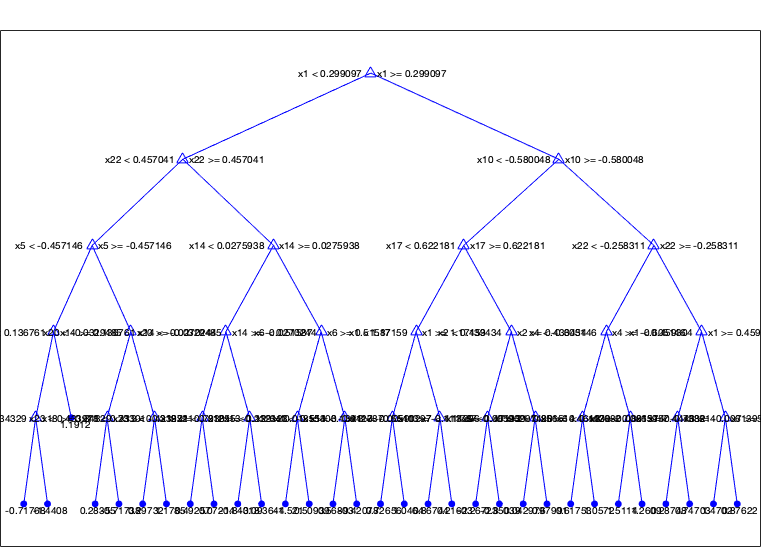


Figure 29: Regression Tree with 30 splits for Normalized SO2 Emission

# Root Cause of SO2 Emission

To find the root cause of SO2 Emission in recovery boiler, variables importance indicators are summarized in Table 5. Beta values is used for OLS, PCR and PLS. The relative magnitude of beta for each variable indicates their relative importance. The sign of the beta indicates the direction of the correlation. Predictor importance is used for regression tree. Predictor importance is always a positive value and no directional information can be obtained. Finally, the weight of first PCA component is shown as a reference. Because it captures 40% of the variance in X, it shows what varies the most in X. The top 6 parameters with highest absolute value for each method is highlighted in red. Therefore, a variable that are red for most method is highly likely to be the dominant cause of the emission. Out of all methods, The variables deemed importance by regression deserves extra attention because the method predicted almost 90% of variance in testing.

Primary Air Temperature is an important predictor for SO2 emission in all methods. Its beta value is negative meaning higher Primary Air Temperature reduces SO2 emission. Black liquor indirect heater temperature is an important parameter for regression tree, OLS and PCR. It has a positive sign in beta values meaning higher temperature lead to higher emission. It is also one of the high variance parameter in the system. Primary Air Flow is the most important predictor for regression tree and its importance is missed by other linear method.

From Frequency Analysis it was found that secondary air flow, flue gas oxygen, dry solid density A&B, primary and secondary air temperature, black liquor pressure and indirect heater, and green liquor sulfidity share the same frequency as SO2 emission.

Elimination by process knowledge

Note that Green Liquor and SO2 emission are both the output from the recovery boiler. It makes sense that SO2 emission correlate with sulfur content in green liquor, but it is not the cause. Steam is the third product stream from recovery boiler. It is unlikely steam flow rate and steam to dry solid ratio be a real cause of emission. Flue gas oxygen leaves the stack along with SO2 and unlikely to be the root cause.

Table 5: Contributors to SO2 Emission by method. The top 6 with highest absolute value for each method is highlighted in red

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | OLS Beta | PCA first component | PCR Beta | PLS Beta with 10 components | Regression Tree |
| Intercept |  |  | -0.0033 | -0.002 |  |
| Primary Air Flow | 0.13 | 0.2811 | 0.1338 | 0.1736 | 1.25E-03 |
| Secondary Air Flow | 0.48 | 0.1889 | 0.4791 | 0.3014 | 8.33E-05 |
| Tertiary Air Flow | -0.09 | 0.1192 | -0.0891 | -0.106 | 1.24E-05 |
| Primary Windbox Pressure | 0.01 | 0.1725 | 0.0063 | 0.0692 | 4.35E-05 |
| Secondary Windbox Pressure | 0.16 | 0.1678 | 0.1629 | 0.1517 | 4.62E-04 |
| Tertiarary Windbox Pressure | 0.23 | 0.0064 | 0.2291 | 0.2278 | 2.92E-04 |
| Burn Rate | -0.05 | 0.1302 | -0.0517 | -0.2344 | 1.10E-05 |
| White Liquor Sulfidity | -1.15 | 0.2923 | -1.1528 | -0.2019 | 3.55E-05 |
| Black Liquor Solid 50./50 Test | -0.56 | 0.3015 | -0.5635 | -0.5889 | 1.15E-04 |
| Flue Gas Oxygen | -0.14 | -0.0344 | -0.1381 | -0.2517 | 6.98E-05 |
| Total Reduced Sulfur | 0.11 | 0.042 | 0.1131 | 0.1305 | 5.07E-05 |
| Dry Solid Density Transmitter A | 0.04 | 0.284 | 0.0446 | 0.2063 | 1.25E-05 |
| Dry Solid Density Transmitter B | 0.32 | 0.2826 | 0.3174 | -0.0241 | 1.12E-05 |
| Primary Air Temperature | -1.88 | 0.2892 | -1.8789 | -0.6623 | 4.04E-04 |
| Secondary Air Temperature | 1.91 | 0.2777 | 1.9104 | 0.7832 | 6.65E-05 |
| Steam Flow Rate | 0.14 | 0.2786 | 0.1429 | 0.103 | 1.87E-05 |
| Black Liquor Solid Flow Rate | -0.19 | 0.2146 | -0.1942 | -0.0042 | 8.37E-05 |
| Steam to Dry Solid Ratio | 0.28 | 0.2078 | 0.2757 | 0.3191 | 1.62E-04 |
| Black Liquor Pressure | -0.07 | 0.0969 | -0.0679 | 0.0106 | 4.80E-05 |
| Black Liquor Indirect Heater Temperature | 1.15 | 0.2867 | 1.1518 | 0.1438 | 1.75E-04 |
| Salt Cake Rotary Feeder Speed | 0.16 | -0.0197 | 0.163 | 0.1107 | 7.59E-07 |
| Lime Kiln Green Liquor Sulfidity | -0.4 | 0.1701 | -0.3978 | -0.3204 | 7.19E-04 |
| Nitrogen Oxides correlation | 0.02 | 0.0893 | 0.0184 | 0.0697 | 1.39E-04 |

Combining result from spectral analysis, regression and process knowledge, it was found that secondary airflow, primary and secondary air temperature and black liquor indirect heater temperature are the likely cause and primary air temperature and black liquor indirect heater temperature the likely root cause. Table 6 summarized the root cause screening process.

Table 6: Summary of Root Cause Screening

|  |  |  |  |
| --- | --- | --- | --- |
|  | Spectral Analysis | Regression | Process Knowledge |
| Primary Air Flow |  | Yes |  |
| Secondary Air Flow | Yes | Yes | Yes |
| Tertiary Air Flow |  |  |  |
| Primary Windbox Pressure |  |  |  |
| Secondary Windbox Pressure |  |  |  |
| Tertiarary Windbox Pressure |  |  |  |
| Burn Rate |  |  |  |
| White Liquor Sulfidity | Yes | Yes | Yes |
| Black Liquor Solid 50./50 Test |  | Yes |  |
| Flue Gas Oxygen | Yes |  |  |
| Total Reduced Sulfur |  |  |  |
| Dry Solid Density Transmitter A | Yes |  |  |
| Dry Solid Density Transmitter B | Yes |  |  |
| Primary Air Temperature | Yes | Yes | Yes |
| Secondary Air Temperature | Yes | Yes | Yes |
| Steam Flow Rate |  |  |  |
| Black Liquor Solid Flow Rate |  |  |  |
| Steam to Dry Solid Ratio |  |  |  |
| Black Liquor Pressure | Yes |  |  |
| Black Liquor Indirect Heater Temperature | Yes | Yes | Yes |
| Salt Cake Rotary Feeder Speed |  |  |  |
| Lime Kiln Green Liquor Sulfidity | Yes | Yes |  |
| Nitrogen Oxides correlation |  |  |  |

# Reference

[1] : Thompson Equipment Company blog. blog.Teco-inc.com

[2] : Kraft Recovery Boilers. Terry N. Adams,TAPPI Press. 1997