# INM433 Visual Analytics Individual coursework submission:
# Visual Analytics with regard to predicting and optimising Iron Ore quality in Iron Ore flotation processes

By Niall Larkin

Niall larkin
Niall.Larkin@city.ac.uk

# 1. Motivation, data and research questions:

## 1.1. Motivation for study and domain specific research questions for investigation:

The motivation for this study is to assess the quality of Iron ore feed output from an Iron Mining Silica removal froth flotation unit operation(See Figure 1 below for process flow ) based on the quality of its feedstock and unit operation parameters. As a result of this there a number of analytical questions that come from this objective:

- What data from the manufacturing sensors are genuine feedback on the process?
- Is it possible to predict % Silica Concentrate every hour based on the process parameters and feed stream stock?
- What parameters can be altered to reduce the final %Silica concentrate.

This analysis is pertinent from a business standpoint as by optimising the processing with regards to understanding Silica concentration provides values in 2 ways:

1. Having the ability to reduced down the cost of finish product sampling reduces of the overall manufacturing process and decreases batch release time for each lot of ore.
2. Having a better characterisation of how different processing parameters influence the final silica output leads to the opportunity that silica removal can be further optimised improving downstream processing of iron ore by reducing the energy requirements during pig Iron production [1]

## 1.2. Data suitability:

The data set utilised for this analysis is directly from the manufacturing shop floor of Iron Ore froth flotation mining unit operation [2]. As a result it is suitable dataset to utilise for this analysis. One item to be conscious off the time series data output stream results for iron and silica are taken on an hourly basis and are meant to represent the subsequent hour prior of production data as a result.

## 1.3. Data Transformations

There were a number of data transformations that where required prior to performing any analysis on this dataset.

1. Removal of all commas in the numerical variables with decimal places in the dataset.
2. Change date timestamps as the time measurement for the process parameters where taken every 20 seconds and therefore need to be changed to reflect this for any time series analysis.
3. The data was aggregated up per hour and averaged on a mean value. This was done due to the fact that all final results where recorded on an hourly basis.

# 2. Tasks and approach

## Data Quality assessment What manufacturing instrumentation signals are genuine processing signals?

From a manufacturing standpoint typical issues there a myriad of issues encountered with regards to poor or noisy feedback from instrumentation due to manufacturing conditions [3]. As a result two activities need to be performed to address this:

1. Data quality assessments: For poor signal feedback for the process instrumentation or analytical errors are commonplace in a manufacturing setting. This is typically easy to identify by a lack of variance within the time series trend where it is physically impossible to have such phenomenon. In order to determine where this is present all time series trends will be plotted where static no variance in the trends will be identified and removed from the data set for analysis.
2. Process control systems for froth flotation control for primarily, ore pulp flowrate, column level and air flow rate [3] it is expected that the variability associated with these parameters will be relatively limited compared to other process parameters. However these instruments are exposed to significant physical disturbance and corruption of their feedback(Reference Table 1). As a result a Savtivzky Golay filter will be used to smooth high frequency noise from these variables to better visualise it initially. This approach has been used to smooth data from similar physical processes for visualisation [4].
3. ACF trends at multiple lags in conjunction with time series plots will be generated and reviewed by an analyst to determine where trend and seasonality is present in the dataset. Differencing will then be completed on any features where this is present to ensure stationarityy. Time series plots, histograms and ACF plots of the transformed datasets will be necessary in order to determine which differencing strategy has worked also. This visual approach is necessary due to the low statistical power that stationary statistical tests such as the Anderson-Fuller Dickerson tests provide [5]. Simple differencing strategies will be attempted first such as first-differencing, log plots, etc.

## Is it possible to predict % Silica Concentrate every hour based on the process parameters and feed stream stock and are there any parameters that influence silica concentration more than others.

Due to the process flow present of this unit operation(See Figure 1) the analysis of what elements of the process impact silica concentration can be broken down comparing the final silica concentrate to the different process parameters as outlined in figure 1.

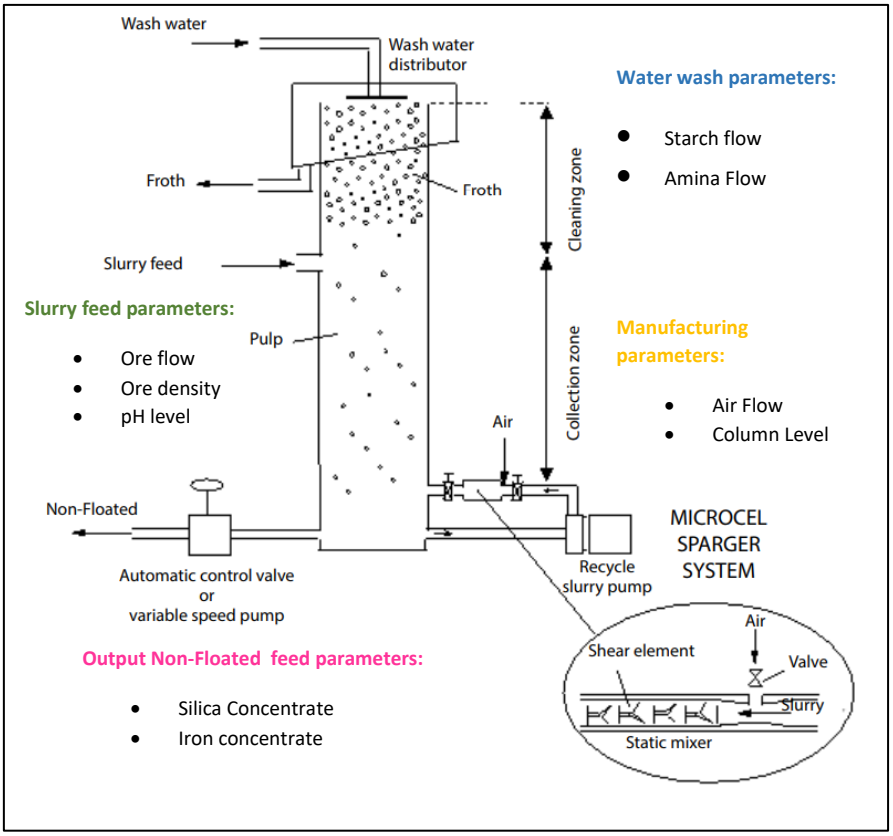


*Figure 1 Schematic of Microcel flotation column Iron Ore Froth Flotation [6]*

This can be assessed visually using direct /time-lagged pairplots vs silica concentrate and cluster analysis if clusters are visually found to be present. PCA analysis will also be performed to determine if there is any overall strong pattern within the data with a review of the eigenvectors to determine which parameter or cofactors could be the cause of significant variance.

Lagged pairplots will determine if historical processing data correlates and therefore predicts silica concentrate value. Considering the fact that the average residence time for iron ore particles in flotation columns is approximately 10 to 20 minutes [6] it can be expected that the maximum genuine lag will be 1 hour.

Multi-level regression will be performed against any clusters sets found to be of value by the analyst with the main parameters that can be found to predict silica concentration this will be done iteratively and interpreted by the analyst to confirm that clustering model has effectively captured clusters present in the initial plot. Regression and residual plots will be generated to confirm that the fit of the model is effective.

As can be seen above all visual tools being utilised are two dimensional plots with two planar variables and one retinal. Three dimensional plots will be utilised if strong co-factor behaviour is observed. These visual tools being used as per Bertin's image theory are the most effective as they satisfy this criteria and enable immediate perception of the data at a single glance [7].

# 3. Analytical steps:

## Data quality assessment:

Initial visualisation of the datasets features time series plots revealed a number of areas where data quality was called in question. Specifically this was found with regards to the quality control measurements taken from the finished product(refer to circled area highlighted in Figure 2 to Figure 3 below). The typical distributions of the finished product testing excluding these results can be seen in Figure 4. This is typical that repeated analytical tests of mineral samples take on a normal distributions in part due to the variability of the minerals and the analytical test itself. The fact that consistent outliers are present and relay a consistent value over two separate chemical compositions parameters in excess of 56.25 hours for both silica and iron assay testing at the same time is highly unlikely due to the variability in the analytical process. From this it can be observed these results are an outlier and the data associated with this was removed as a result.

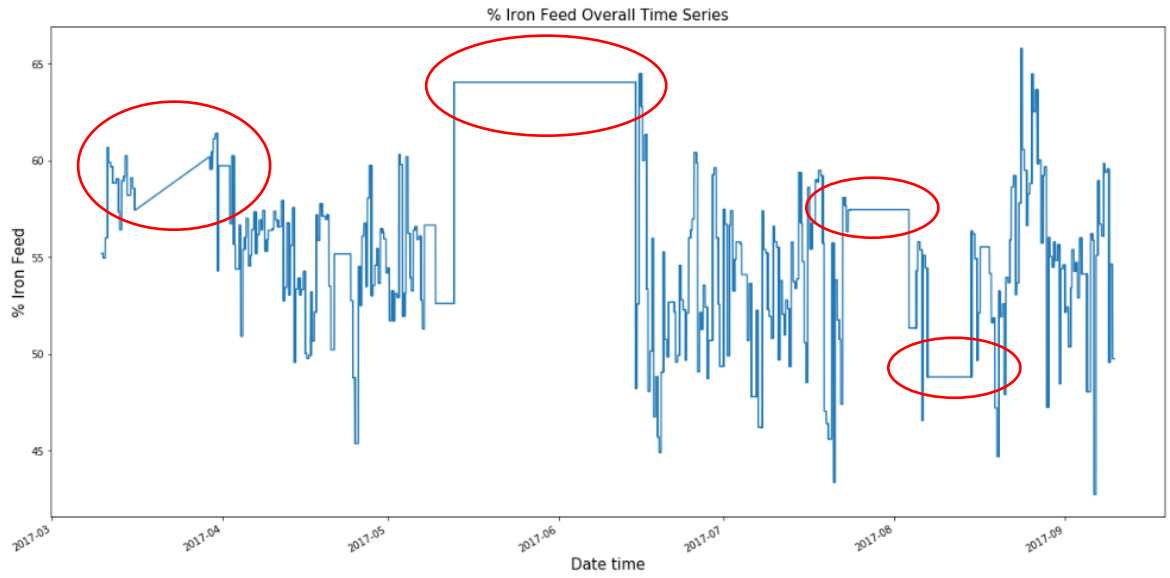


*Figure 2 Points of static feedback from quality control measurements of% w/w Iron feed into froth flotation tank*

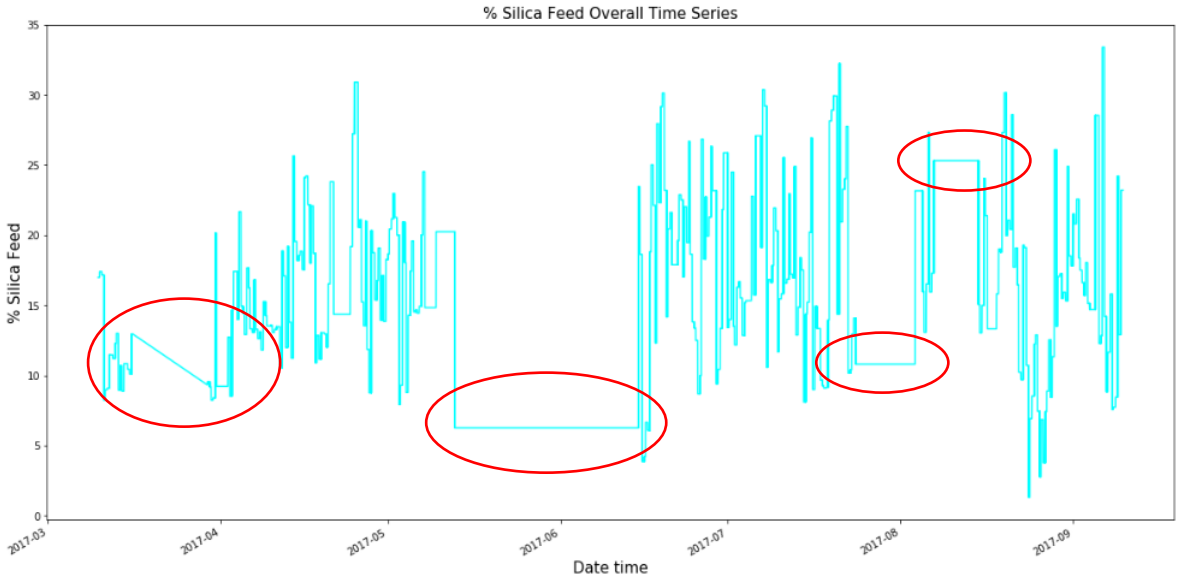


*Figure 3 Points of static feedback from quality control measurements of% w/w Iron feed into froth flotation tank*

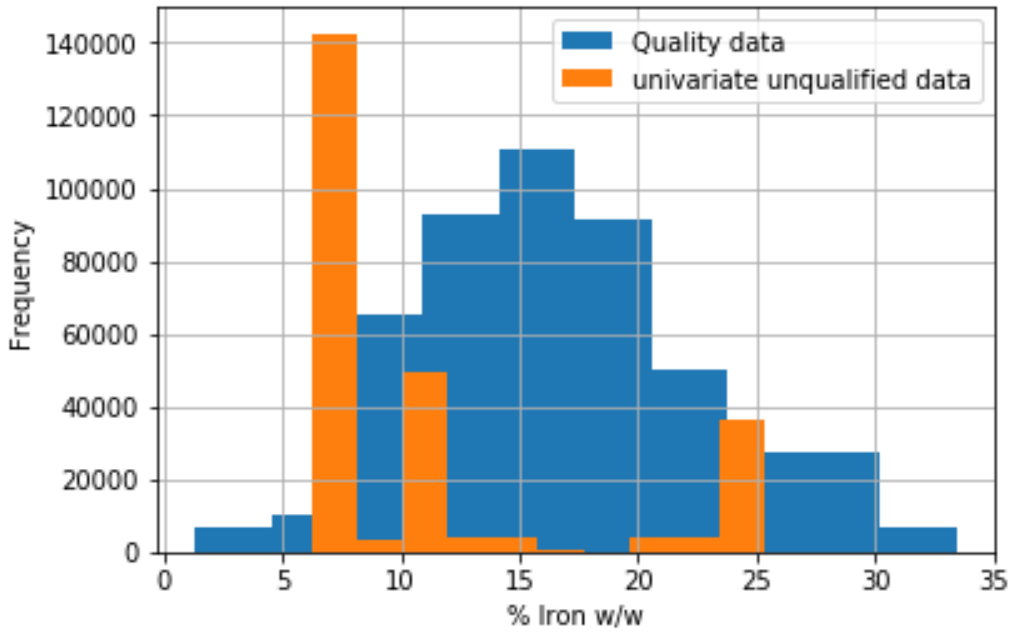


*Figure 4 Distribution of %Iron w/w assay values of feedstock quality distributed data and unqualified univariate data are plotted to show that the data is primarily outliers*

## Time series stationarity assessment:

From the initial Time series line plots (with Savitzky Golay filters to help visualise the trend from high frequency noise(See *Figure 5* to Figure 8 for examples and Table 1 explanation of causes of noise). ACF plots and histograms where generated (See Figure 9 to Figure 14 for examples) to determine what parameters needed to be made stationary prior to performing any further correlation analysis. From this the following parameters needed to be made stationary

1. Ore pH
2. Amina flowrate
3. Starch Flowrate

4. Ore pulp density

As a result these features where all de-trended using first differences approach. Following this all time series trends where replotted with subsections inspected by an analyst in conjunction with ACF and Histograms plots being generated again. From this, it was found from this that the features had been de-trended appropriately (See Figure 15 to Figure 20).

*Table 1 Cause of noise in process outliers*

| **Parameter** | **Cause of high frequency outlier** | **Outlier direction** |
|---|---|---|
| Amina Flowrate | Valve chatter, valve on flowrate line opening and closing rapidly in succession causing the flowrate to drop rapidly as a result. | Negative towards 0 and overshooting once valve reopens(See Figure 6 and Figure 8) |
| Starch Flowrate | | |
| Air Flow in all columns | | Negative towards 0 and overshooting once valve reopens(See figures *Figure 5*) |
| Tank level indicator | Cause is due to foaming at top of vessel from froth operation resulting in high low level signal interference | Negative and positive (Figure 7) |

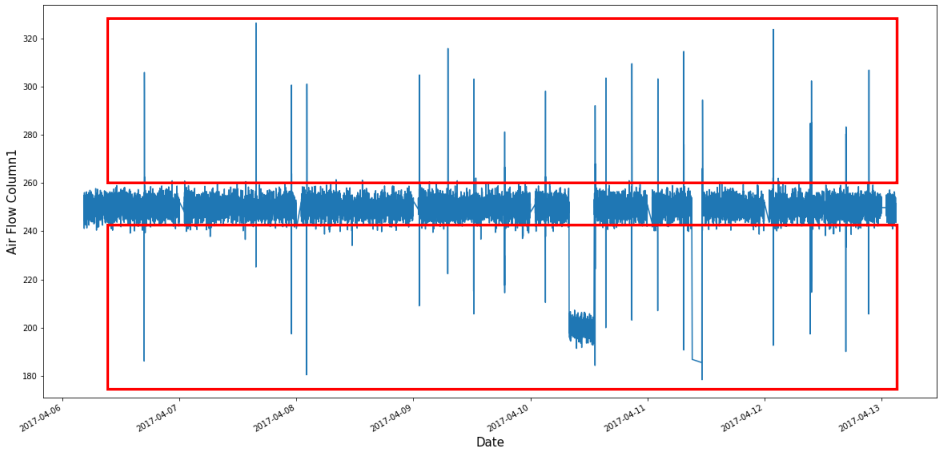


*Figure 5 Time series excerpt Air Flow Column 2 disturbance high frequency signals examples*

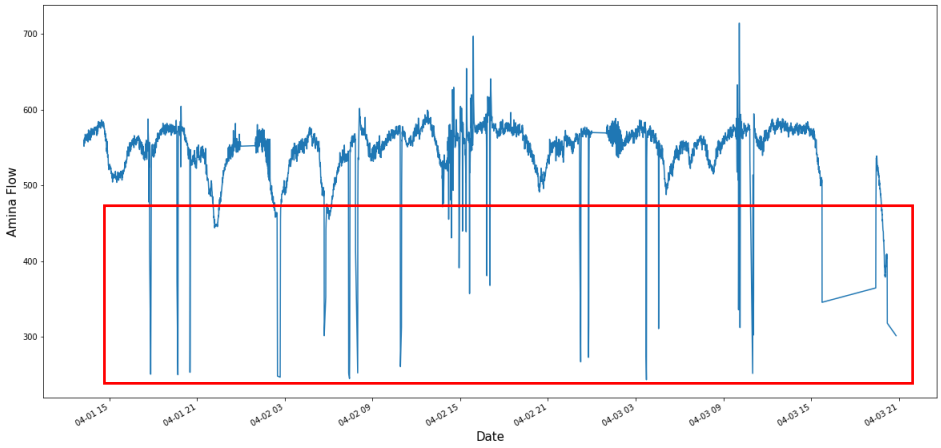


*Figure 6 Amina Flowrate high frequency disturbance signals*

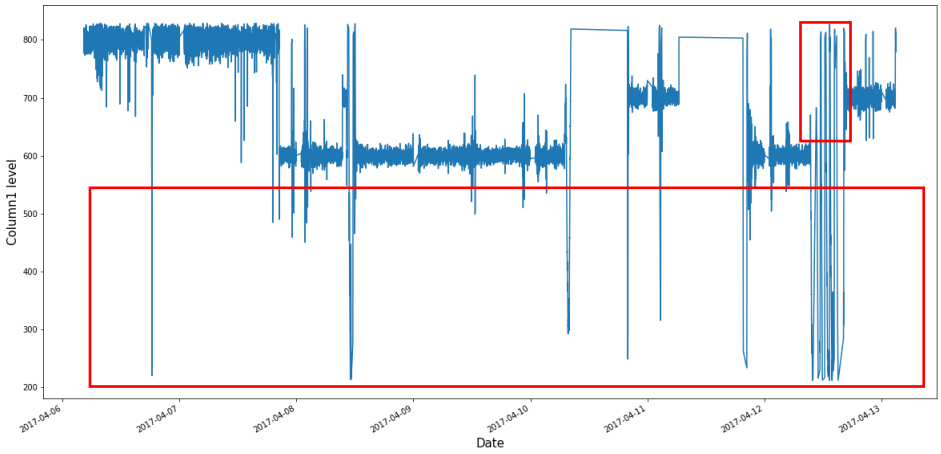


*Figure 7 Column1 liquid level high frequency disturbance signals present*

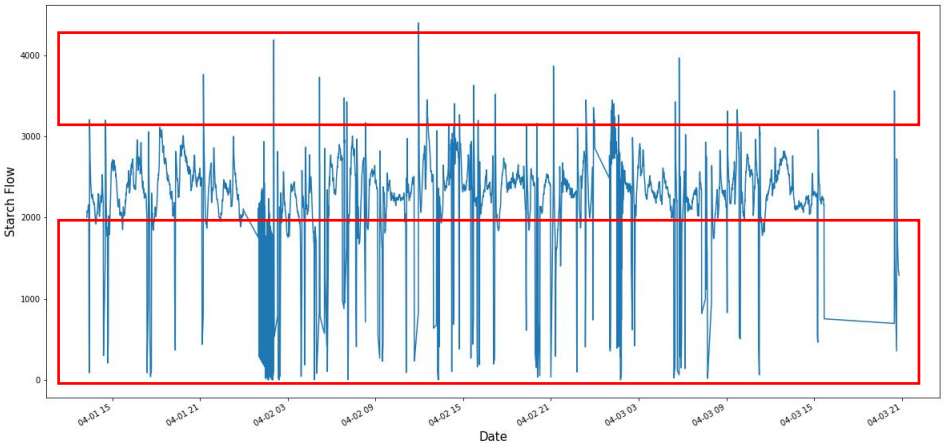


*Figure 8 Starch Flow with high frequency disturbance signals present*

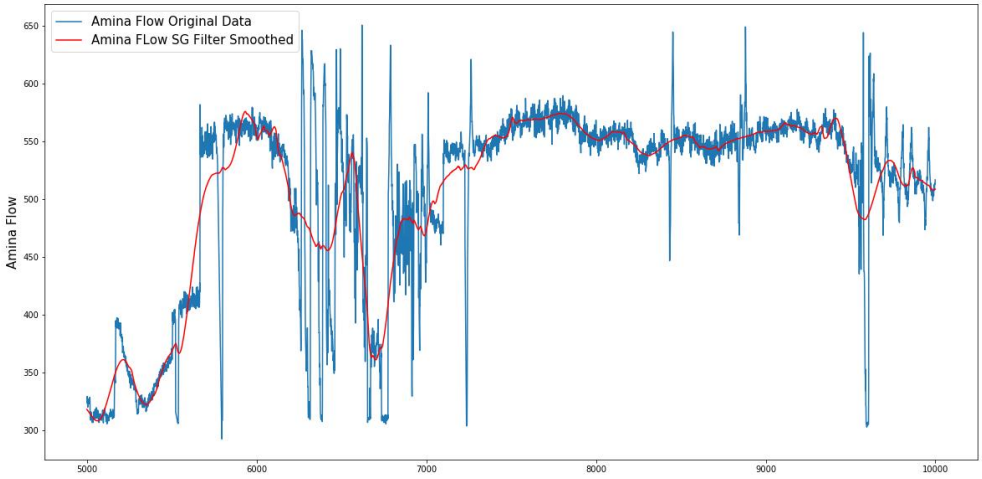


*Figure 9 Amina Flow Pre Detrending Time series plot excerpt*

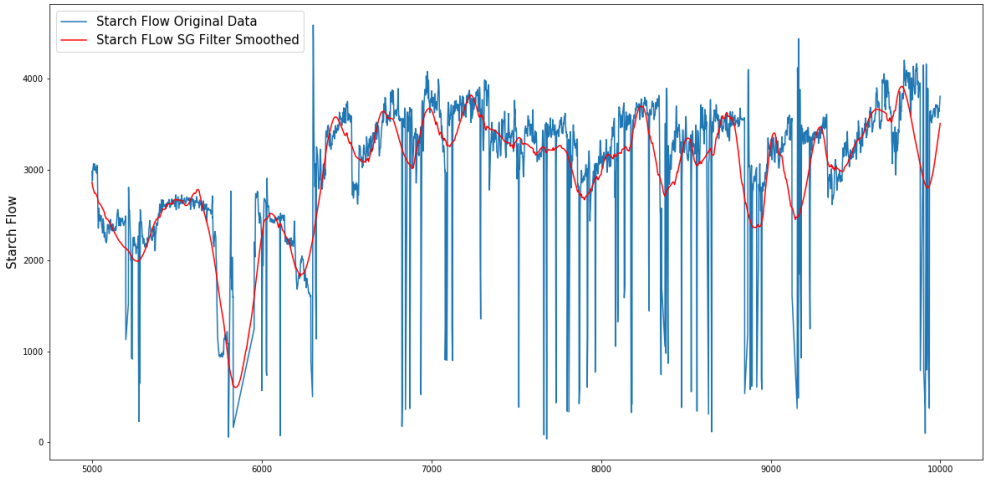


*Figure 10 Starch Flow Pre de trending time series plot excerpt*

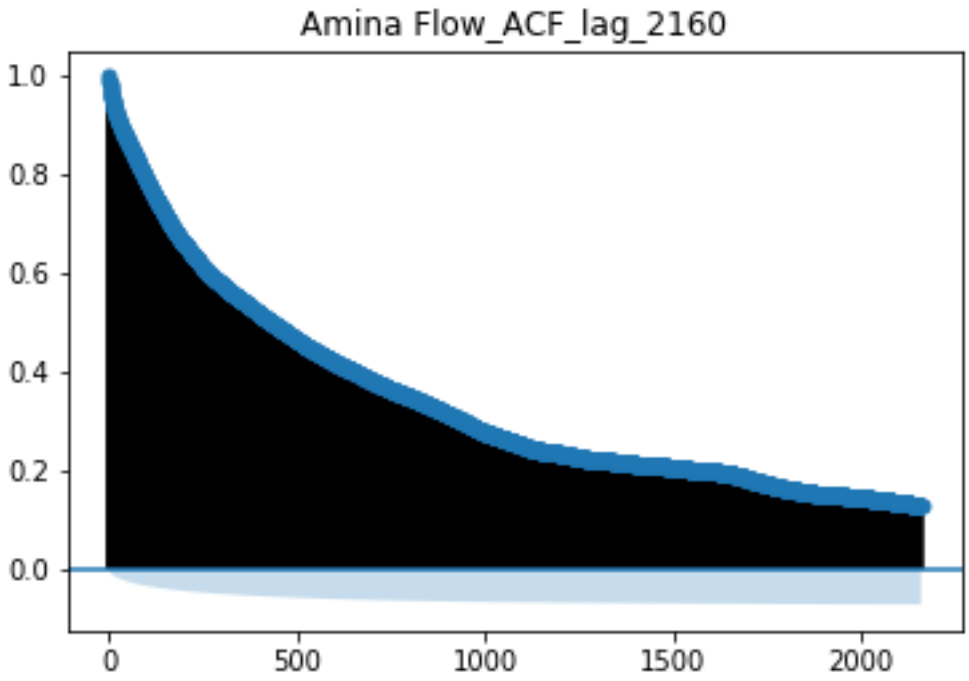


*Figure 11 Amina Flow ACF plot pre detrending*

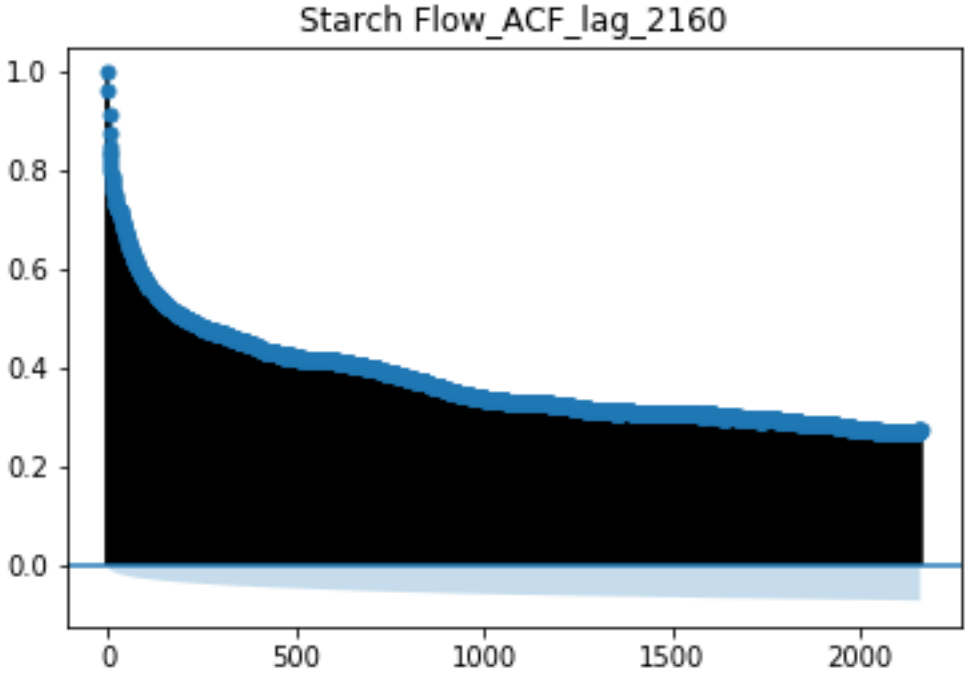


*Figure 12 Starch Flow Pre Detrending*

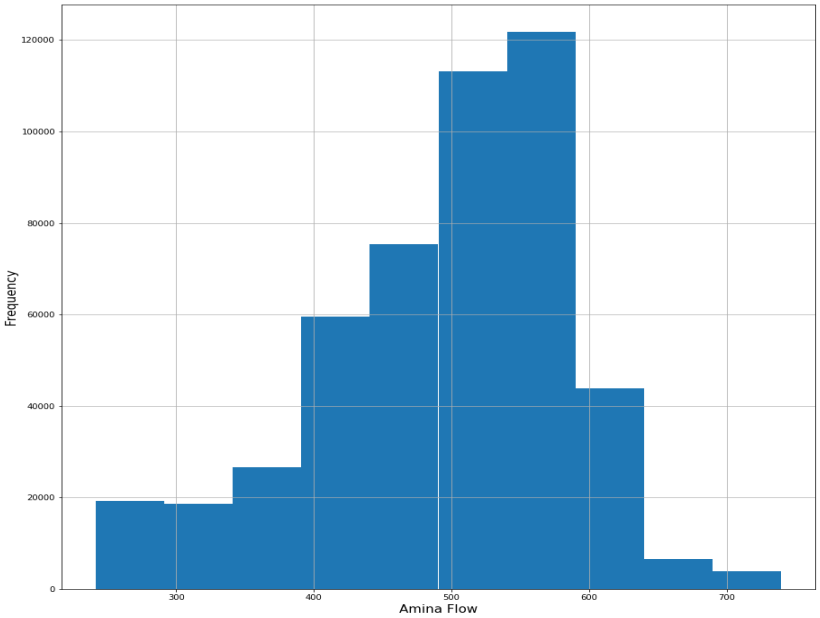


*Figure 13 Amina flow histogram pre detrending*

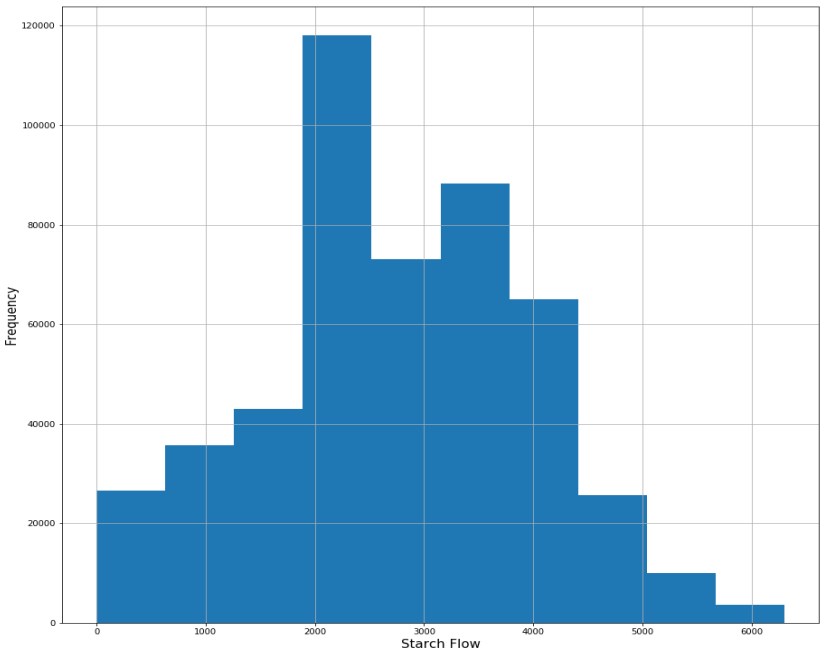


*Figure 14 Starch flow histogram pre detrending*

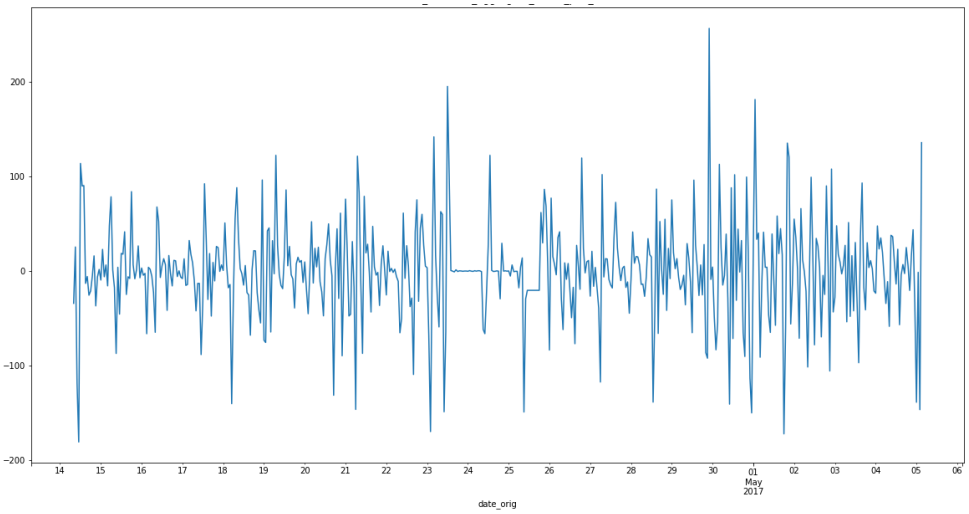


*Figure 15 Differenced at lag 1 Amina Flow rate inspection excerpt from total time series*

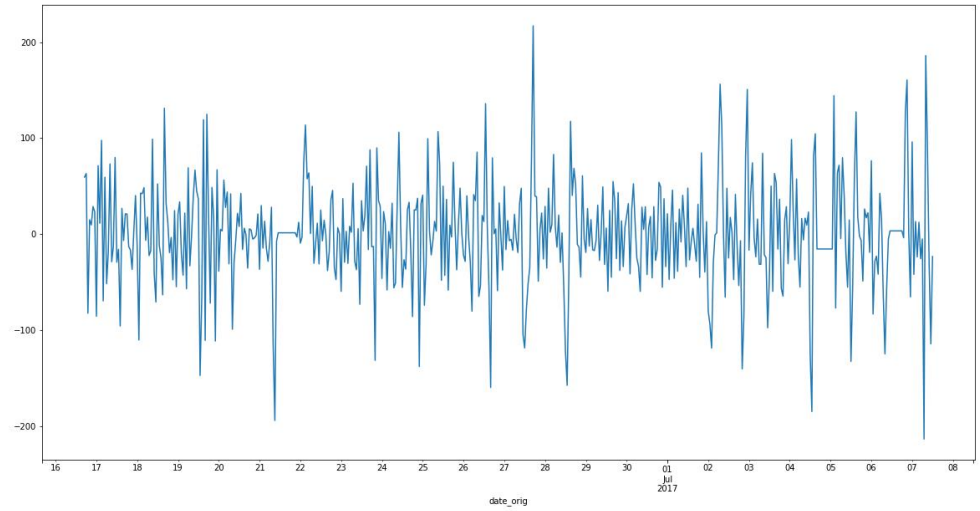


*Figure 16 Differenced at lag 1 Amina flowrate time series inspection excerpt from total time series*

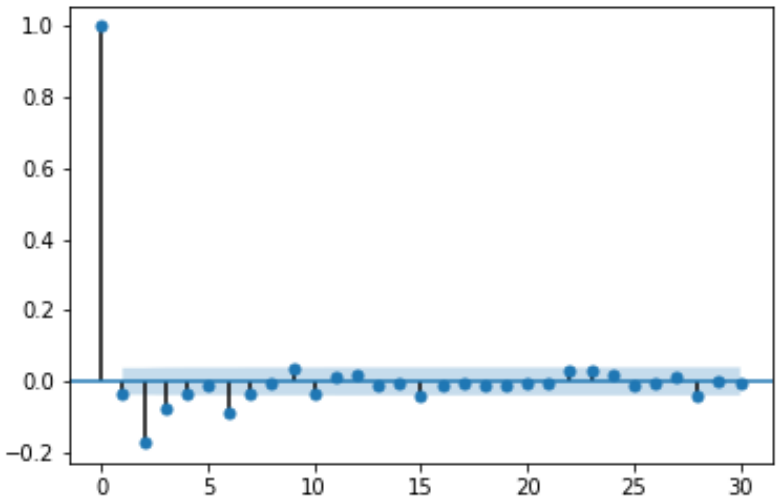


*Figure 17 Amina Flow Differenced at lag 1 ACF plot*

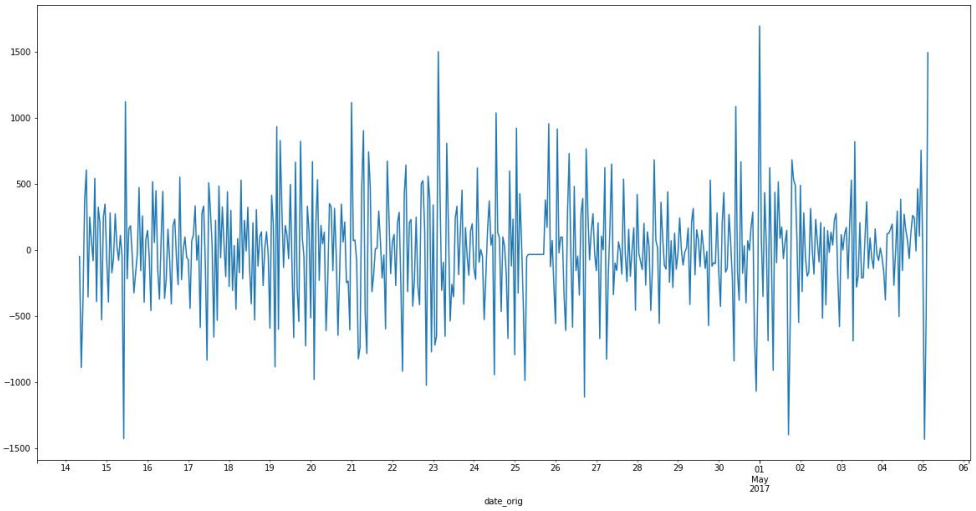


*Figure 18 Starch Flowrate Time series detrended at difference lag 1 inspection excerpt*

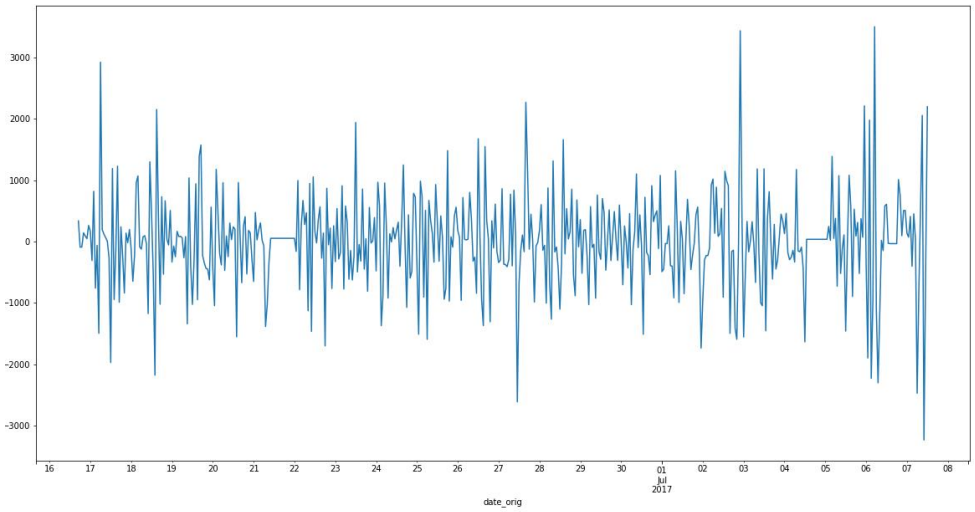


*Figure 19 Starch Flowrate Time series detrended at difference lag 1 inspection excerpt*

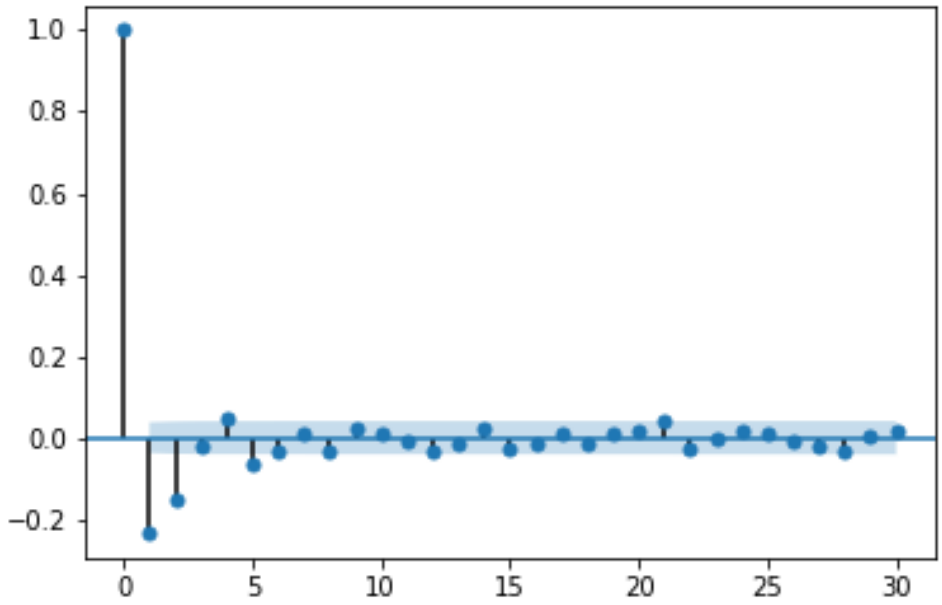


*Figure 20 ACF plot Starch Flowrate time series differenced at lag 1*

## Correlation and cluster analysis of pair plots

Pair plots where generated across all parameters towards the final feed excerpts along with histograms being generated of the final output streams to determine if the feed parameters could be classified based on the presence of no-normal multimodal distributions.

From the histogram analysis it was found that for the silica concentrate at the end of the process feed there was a large right hand tail present in the distribution (See Figure 21). As a result an additional feature was generated where the entire dataset was divided between the finalsilica concentration levels above below levels of 2.5% to see if there was any particular trend with respect to the right hand tail. From this analysis no real significant variance was observed. PCA analysis with colour overlay between the two groups was generated also (See Figure 22) to determine whether or not any specific cluster could be observed with respect to the rest of the general final silica concentrate. No significant variance could be observed however Air Column level and Air Flow where found to be the most significant cause of variance by comparison based on eigenvectors(See

Figure 23) which resulted in these parameters being investigated further in the next stage of this analysis.

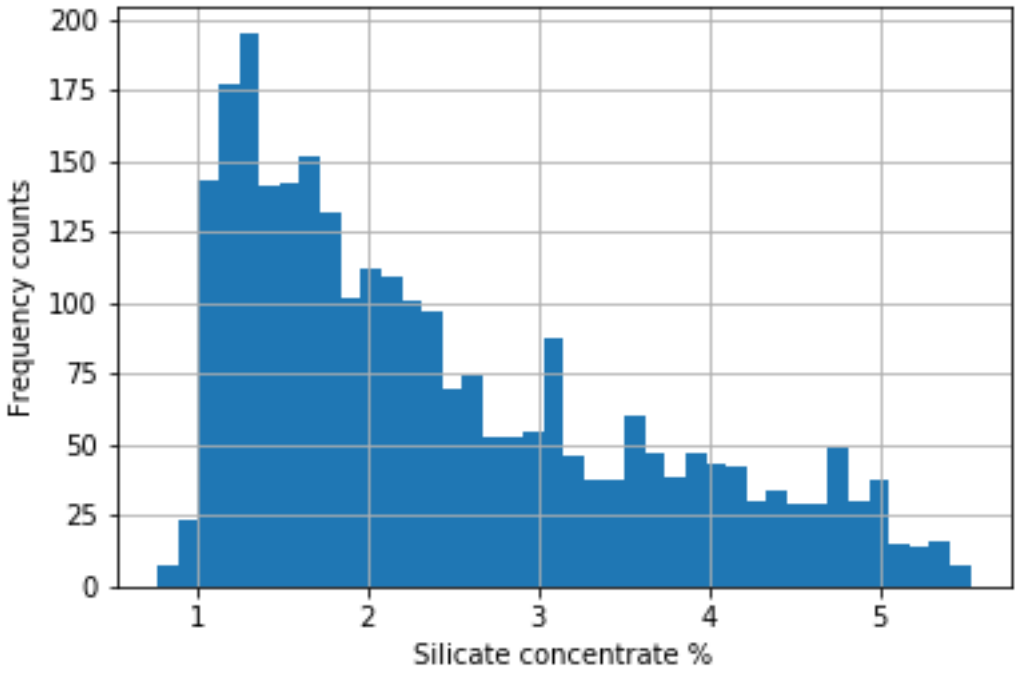


*Figure 21 Histogram of final output feed silica concentration*

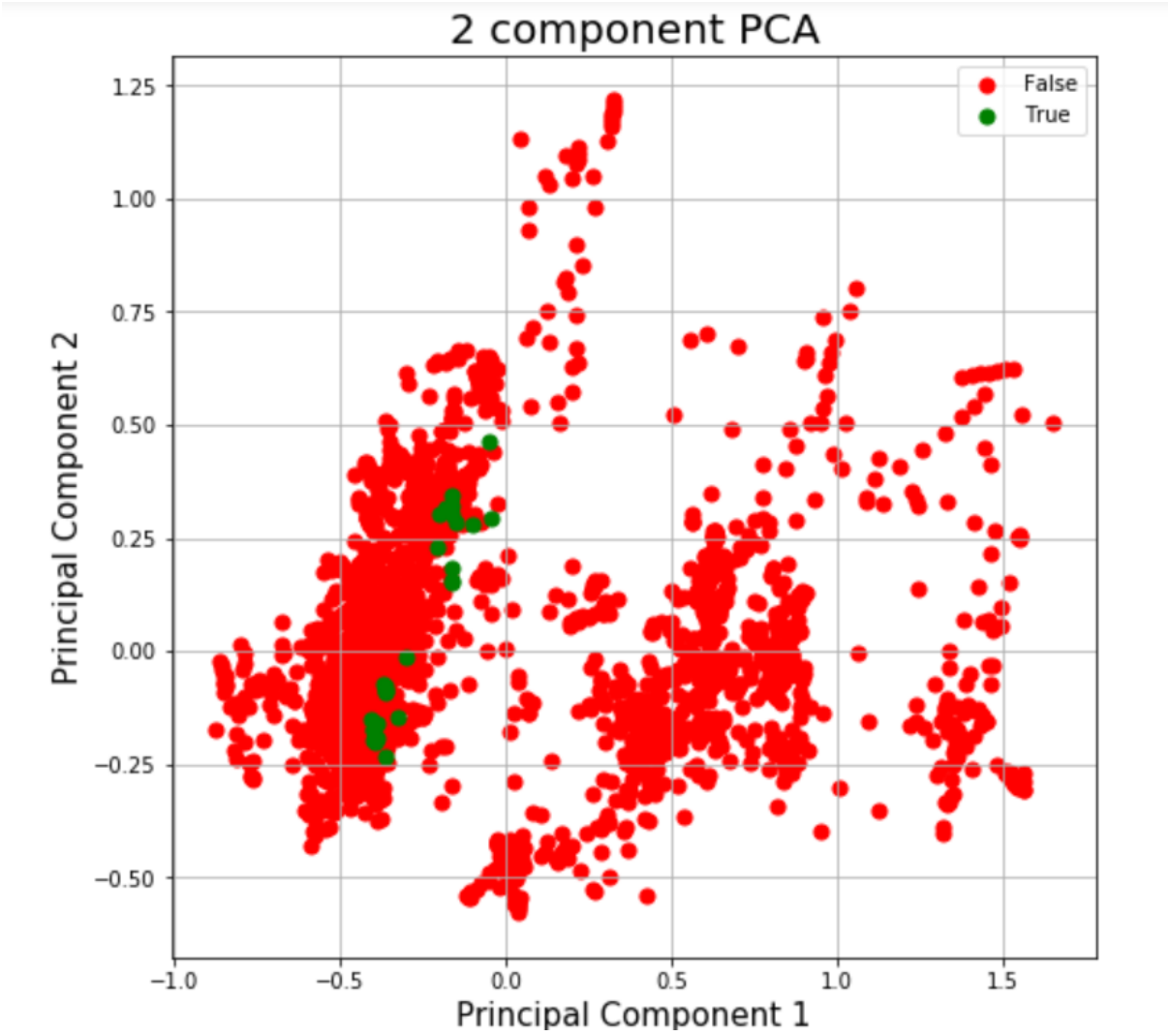


*Figure 22 Plot of discretized data between low silica concentrate False values <2.5 and high silicate concentrate True values*

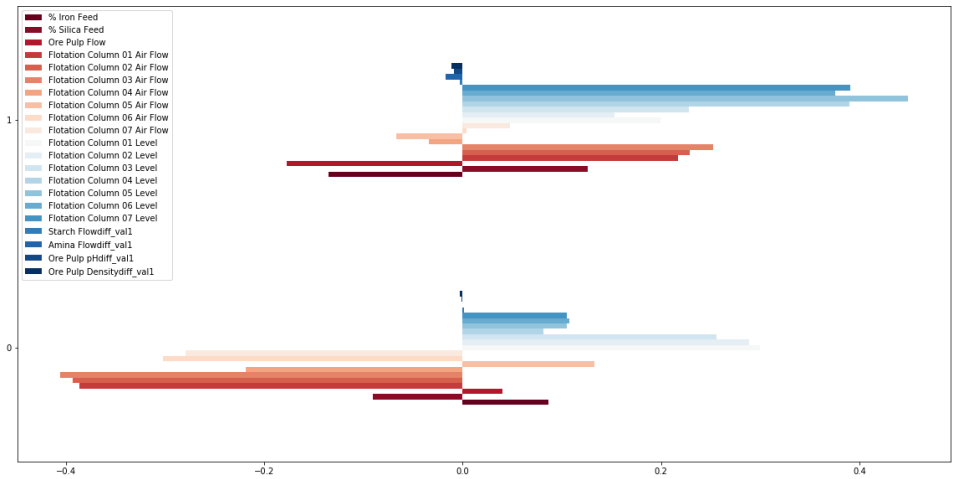


*Figure 23 PCA Eigenvector values for each component in analysis*

The direct and lagged Pairplots between the silica concentration versus the other differenced time series where generated iteratively. From this it was determined that there was no significant correlations found any of the parameters in the dataset with white noise correlations being found for most parameters(See Figure 24 to Figure 25).

One time lagged parameters was found to negatively correlate to the final silica concentrate(See Figure 26):

- Iron concentrate at a lag of 1

Clusters where found with respect to two of the other parameters :

- Column Air Flow from columns 1 to 3(See Figure 27)
- Column liquid level from columns 1 to 3(See Figure 28)

Due to the fact that there appeared to be significant variance in cluster size on the y axis for the column level scatter plot (See Figure 28) at 700mm and that these parameters where the source of the largest variance in the dataset from PCA analysis, cluster analysis was then performed. This was done so as to better help establish their position and variance with respect to silica concentration level.

Agglomerative clustering was performed for this analysis as from the initial time series plots it was known there was a specific hierarchal structure to the data as for the process specific column liquid levels where held for long periods of time at specific liquid levels (Reference Figure 29 to Figure 31). This was done iteratively with 6 clusters but increased to 7 and 8 to ensure no sub-clusters where prevalent within the original visualisation there where not obvious on the initial visualisation(See Figure 32 to Figure 34). The stopping condition for this approach was the silhouette score based on Euclidean distance and also a qualitative assessment of the clusters by the analyst as a sanity check. From this analysis a total of 8 clusters was selected with a silhouette score of approximately 0.85(see Figure 34). Cluster 3 was disregarded as noise however. Cluster 4 was found to be of particular interest as it had lower overall variance and lower overall mean silica concentration which is the gangue(mining waste product term) product.

As a result multi level regression was performed on cluster 4 versus the remainder of the dataset with respect to the lagged iron concentration values to determine how they correlate against silica concentration. The results from this analysis can be seen below in Figure 35 to Figure 36. This was done iteratively also in order to determine the intercept of the line that would arrive at the residuals with the stopping conditions being the lowest residuals square values that could be achieved in conjunction with visual review of the trend line by the analyst. With this the final element of analysis was completed for this process.

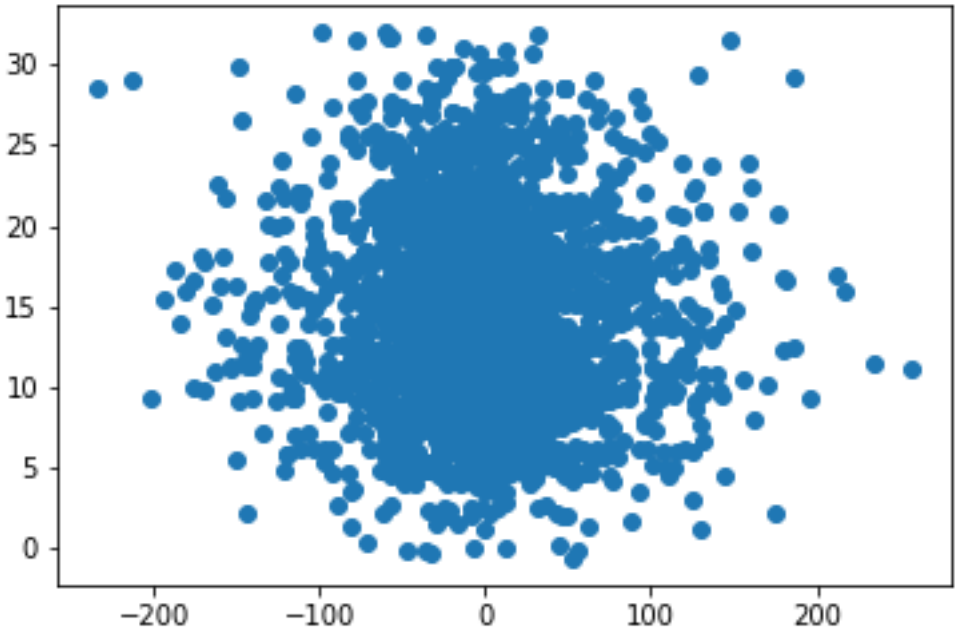


*Figure 24 % Silica removed versus differenced Amina Flowrate plot at lag 1*

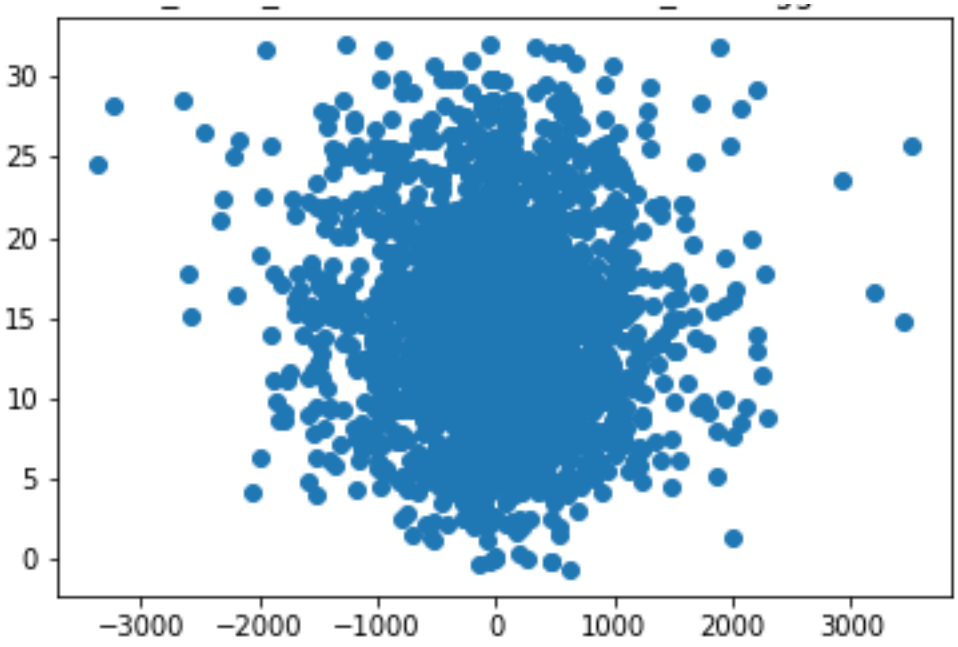


*Figure 25 % Silica removed versus differenced Starch Flowrate plot at lag 1*

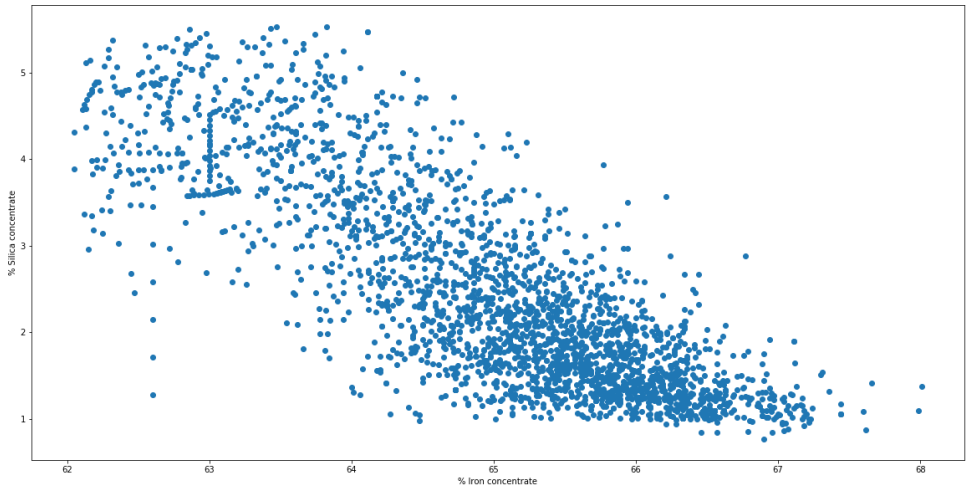


*Figure 26 Iron concentrate lag 1 vs Silica concentrate values*

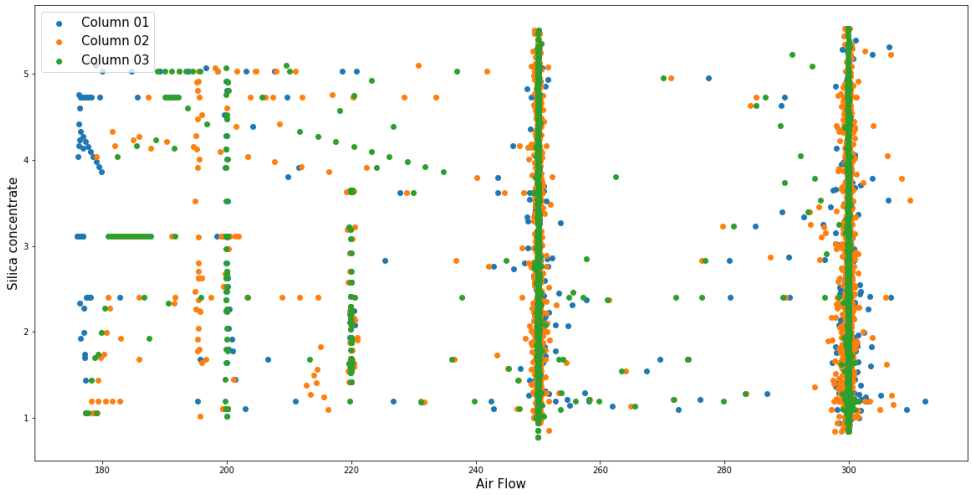


*Figure 27 Column 1to 3 Air Flow versus Silica concentrate*

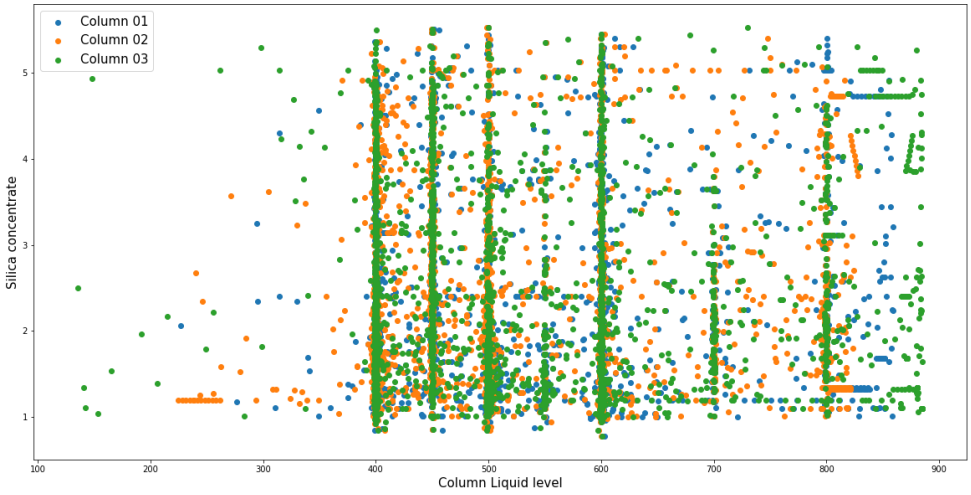


*Figure 28 Columns 1 to 3 Column level vs silica concentrate*

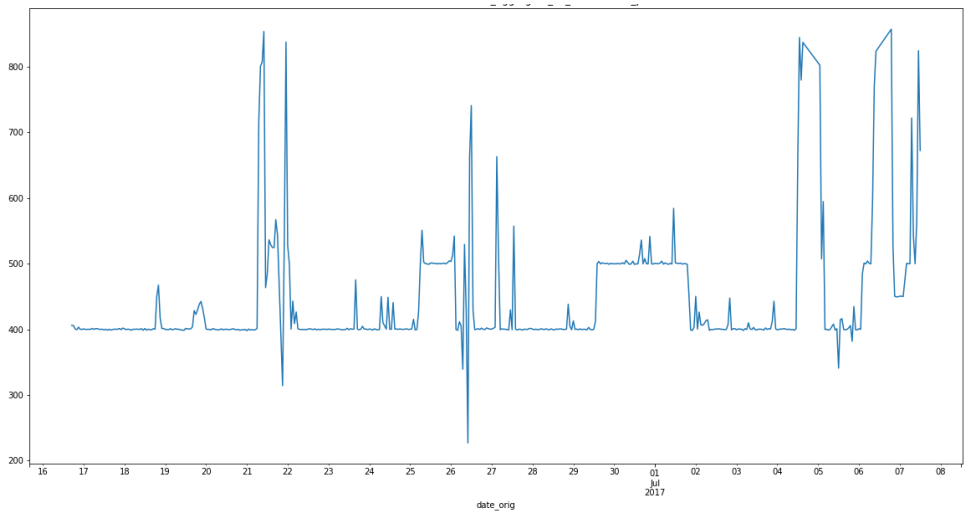


*Figure 29 Column 1 Column level time series plot excerpt*

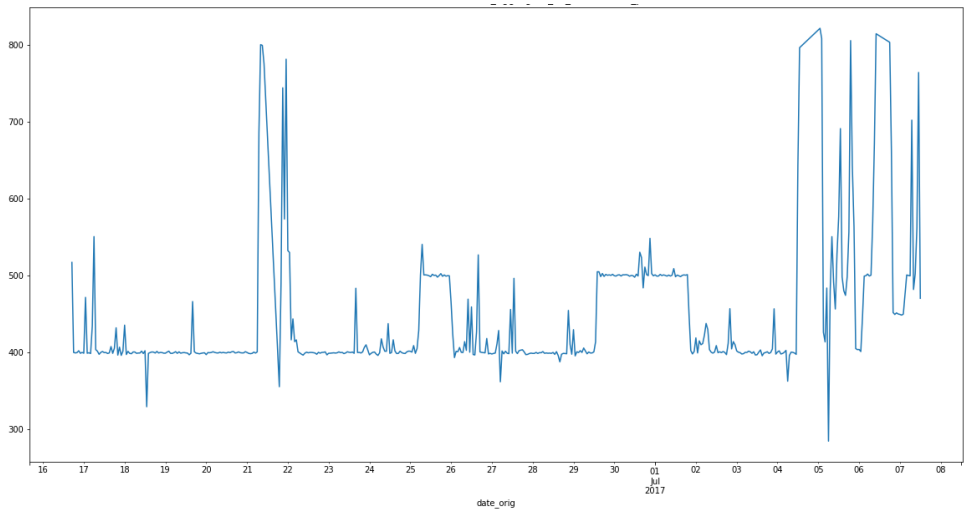


*Figure 30 Column 2 Column level time series plot excerpt*

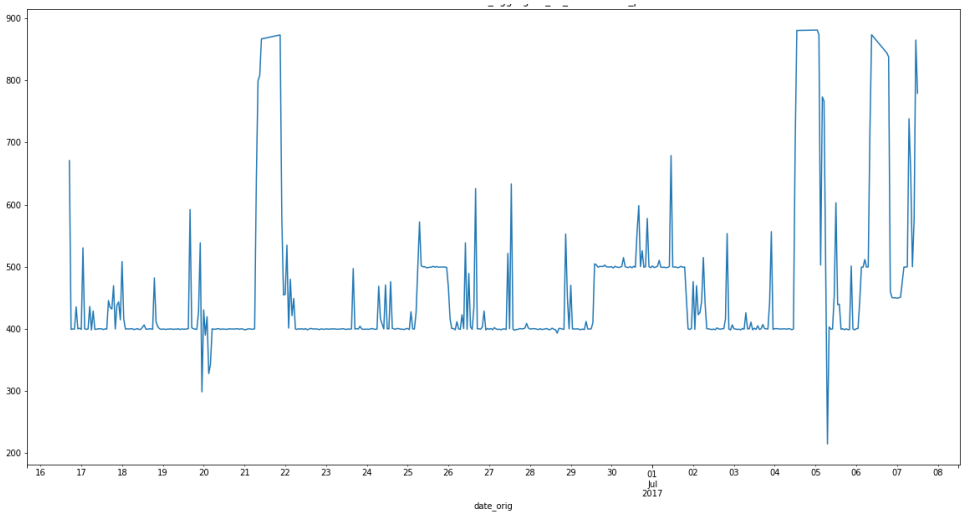


*Figure 31 Column 3 Column level time series plot excerpt*

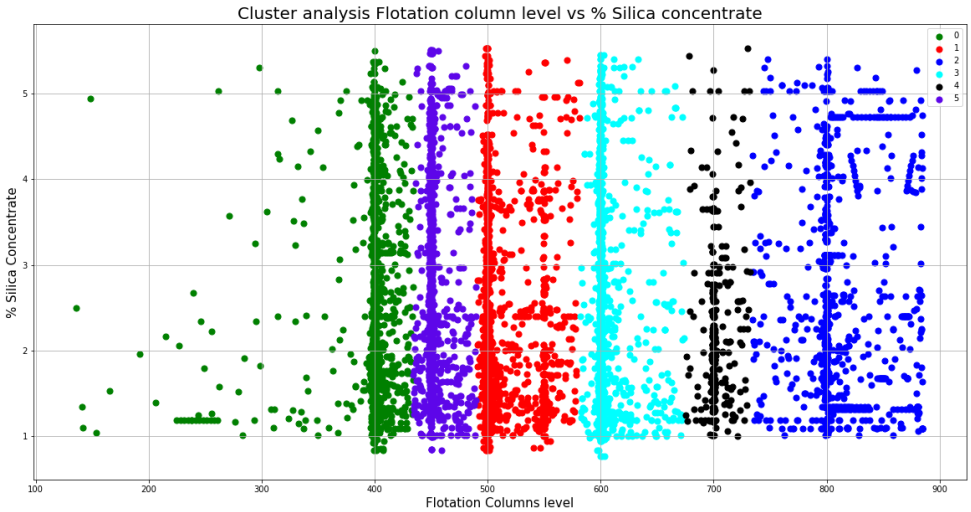


*Figure 32 Silica concentration vs Columns 1-3 column level Cluster analysis with 6 clusters*

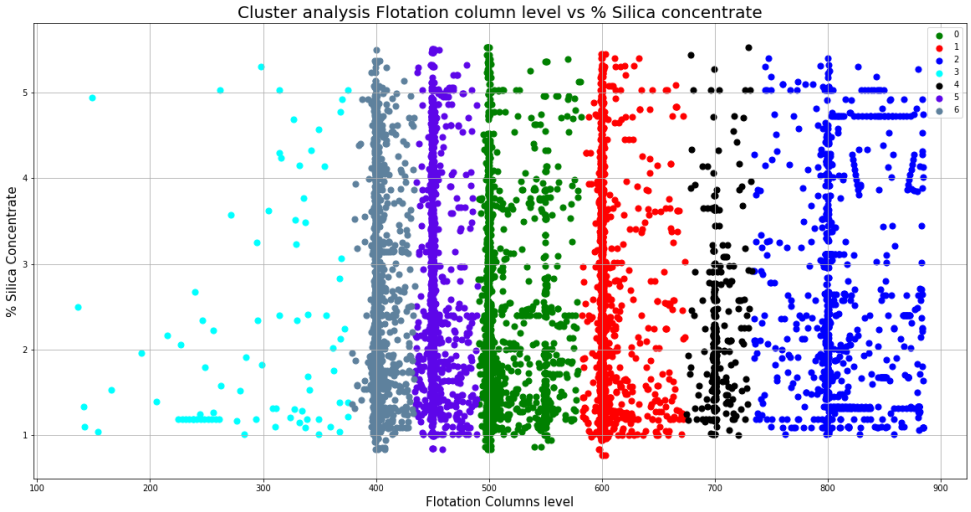


*Figure 33 Silica concentration vs Columns 1-3 column level Cluster analysis with 7 clusters*

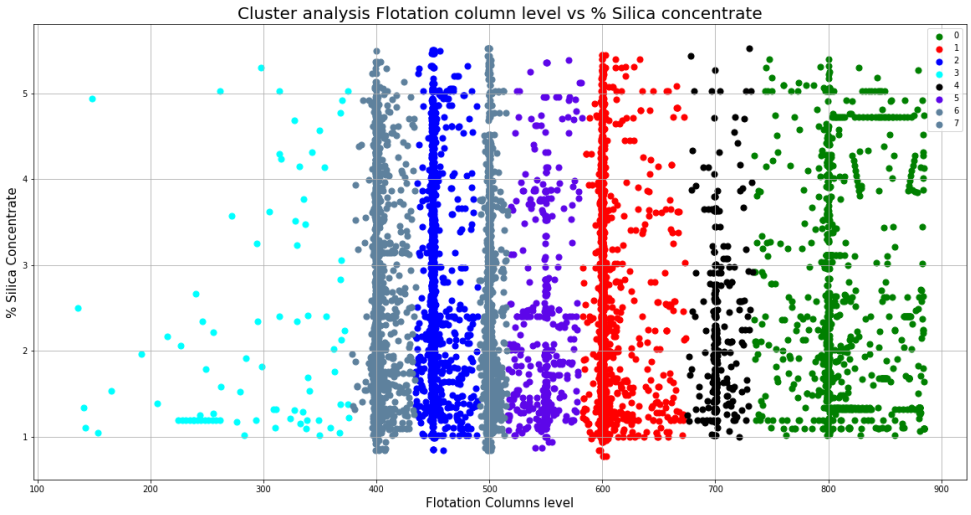


*Figure 34 Silica concentration vs Columns 1-3 column level Cluster analysis with 8 clusters*

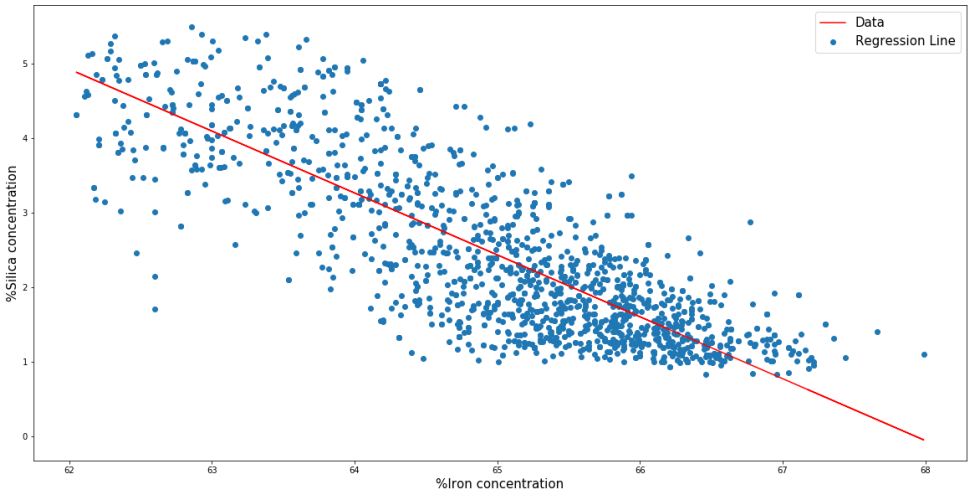


*Figure 35 All  Non cluster 4 Lagged Iron concentration vs Silica concentration*

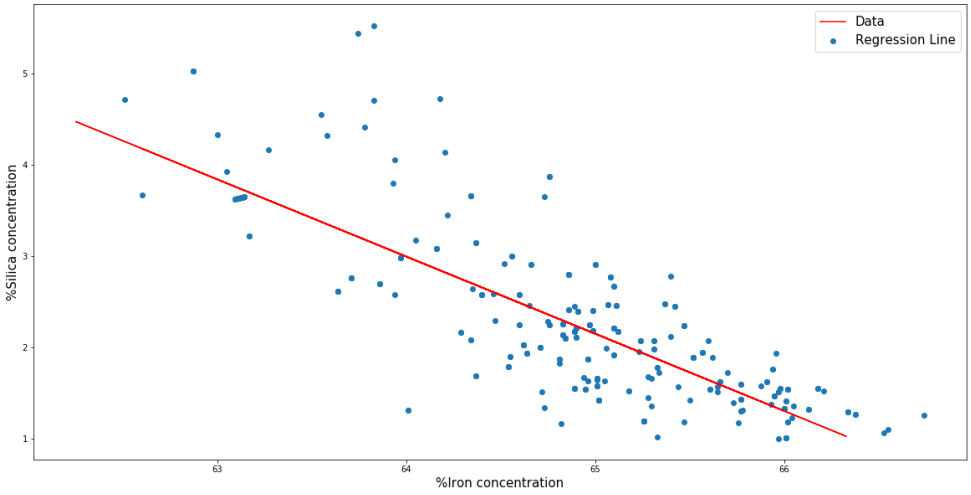


*Figure 36 Cluster 4 Lagged Iron concentration vs Silica concentration on test data*

# Finding:

## Is it possible to predict % Silica Concentrate every hour based on the process parameters and feed stream stock?

Based on the analysis performed above using the lagged Iron concentrate it is possible to determine the an approximate concentration of the silica present in the final product with an $R^2$ value of 0.6. Due to the fact that an industrial setting it is required to have an accuracy closer to an $R^2$ approaching 0.95 [8] due to ore grade constraints this model could not be used to predict the silica concentration in the final product without finished product testing. This could however be used as an investigative tool to assess spurious analytical results to confirm if they are representative of typical values observed from the process.

## What parameters can be altered to reduce the %Silica concentrate?

Overall it was found that Columns 1 to 3 liquid level when ran within a specific range of 700mm the distribution of silica concentration was overall lower compared to other levels utilised. One item to note however is the overall density of cluster 4 which was at found at this approximate level of 700mm was lower(373 data points approximately 15 days of production) compared to other clusters found in the system (Average 1420 data points). As a result the reduction in variation could be due to the lack of data points as opposed to the process itself. Therefore additional testing at this setpoint would be suggested prior to utilising this setpoint as a permanent process parameter.

## Critical Reflection:

Overall the implications of the findings with regards to what are optimum process parameters based on the manufacturing data is typical of what can be found from performing the DMA of the DMAIC(Define Measure Analyze Improve Control) analysis used in the lean six sigma system [9]. How these findings from this analysis can then be implemented and controlled for in the long term would require further analysis outside the scope of this project which pertains to other requirements in the manufacturing process such as process safety and cost –benefit analysis.

What is unique to this type of analysis however to other typical methods used in the literature for process industries is the utilisation of cluster and PCA analysis to assist in the visualisation of unique data structures within manufacturing process. Typically these methods are not utilised for the DMAIC problem solving optimisation approach in industry and would be of benefit to utilise these tools going forward in industry [10] [11].

Overall the visual analytics tools utilised to solve the research questions that where created at the start of this process where satisfactory. However there could have been a number of roadblocks with regards to the dataset and methods utilised that could have made the analysis more challenging. Specifically from the dataset perspective the interpretation of low quality data was far from optimum but it is typical of manufacturing systems that this is an issue [12]. Utilisation of silhouette plots and dendograms visualisations would have assisted with the visualisation of the agglomerative clustering further then just reviewing the final clustering plot with silhouette scores especially, if smaller subclusters where found. These methods can be incorporated into this process flow however as part of future work in this field.

With regards to the applications of the tools above to other fields outside of the process industry (bulk chemical, pharmaceutical and Oil and Gas) industries this approach has merit to being applied to other continuous processes in terms of relaying financials or economic time series variables to one another or other time dependent processes.

# References


[1] L. L and H. RJ, “Effects of alumina on sintering performance of hematite iron ores.,” *ISIJ international,* vol. 47, no. 3, pp. 349-358, 2007.

[2] E. M. Oliveira, “Quality Prediction in a Mining Process,” IHM Stefanini , 23 Nov 2017. [Online]. Available: https://www.kaggle.com/edumagalhaes/quality-prediction-in-a-mining-process. [Accessed 25 Nov 2018].

[3] J. C. Shean BJ, “"A review of froth flotation control.",” *International Journal of Mineral Processing ,* vol. 100, no. 3-4, pp. 57-71., 2011.

[4] G. J. J. Peter M Ireland, “Collision of a rising bubble–particle aggregate with a gas–liquid interface,” *International Journal of Mineral Processing,* vol. 130, no. 2, pp. 1-7, 2014.

[5] S. M.-M. W. Pandit, Time series and system analysis with applications., New York : Wiley, 2000.

[6] M. Inc, “Section 4 – Separations, Metso,” in *Basics in Minerals Processing (5th Edition)*, Helsinki, Metso, 2006, p. 117.

[7] N. G. A. Andrienko, Exploratory analysis of spatial and temporal data: a systematic approach, Berlin: Springer Science & Business Media, 2006.

[8] E. A. Z. R. A. C. L. M. C. A. Ricardo Saldanha, “"Prior assay as an approach to flow titrations. Spectrophotometric determination of iron in alloys and ores.",” *Analytica Chimica Acta ,* vol. 416, no. 2, pp. 231-237, 2000.

[9] S. R. D, “"Six-Sigma: the evolution of 100 years of business.",” *Int. J. Six Sigma and Competitive Advantage ,* vol. 1, no. 1, pp. 4-20, 2004.

[10] B. I. F. W., Implementing six sigma: smarter solutions using statistical methods, New York: John Wiley & Son, 2003.

[11] J. L. De Mast Jeroen, “"An analysis of the Six Sigma DMAIC method from the perspective of problem solving.",” *International Journal of Production Economics ,* vol. 139, no. 2, pp. 604-614., 2012.

[12] K. Zaman, “www.kaggle.com,” Fifth Tribe , [Online]. Available: https://www.kaggle.com/fifthtribe/how-isis-uses-twitter/kernels. [Accessed 28 11 2018].

[13] K. D. I. W. Walid Magdy, “#FailedRevolutions: Using Twitter to Study the Antecedents of ISIS Support,” Qatar Computing Research Institute, Qatar, 2015.