Precision Analysis of Semiconductor Channel Holes

FEBRUARY 8

Thermo Fisher Scientific
Authored by: Umesh Adiga

Precision Analysis of Channel Holes

1.1 Introduction

This dataset consist of several CSV files that contain number of geometric features of thousands of channel holes fabricated at nano-meter dimensions on a semconductor substrate. The data has tow parts

Original features i.e. the geometric features are measured from image data that are unprocessed, Enhanced features i.e. geometric features are collected from processed images.
Each part is subdivided further as features of a geometric fitted elliptical shape to the channel hole and features of the original shape.

Each image (of the same region) is acquired three times and the variation in geometric measurements of the same channel hole in three different images provide a measurement precision. The customer wants to know the best protocol to achive optimal precision. There are four possibilities.

Class 1: Original/unprocessed image features and measurements on actual shape of the channel holes
Class 3: Enhanced/processed image features and measurement on actual shape of the channel holes
Class 2: Original/unprocessed image features and measurements on fitted elliptical shape to the channel holes
Class 4: Enhanced/processed image features and measurements on fitted elliptical shape to the channel holes

we are not working on class 3 and class 4 here as that would be the repetition of the same python code applied on the data from different files.

In the data-wrangling step we focus on collecting data, organizing it, and making sure it's well defined. Some data cleaning and data exploration is also done here and described at corresponding steps.

1.2 Data Wrangling

This shows the number of entries in each of the triplets is not the same. We cannot calculate the standard deviation of the feature measurement for the same hole in the triplets (three measurements for each channel hole is made using three images) if we do not know which holes have missing measurements.

Column 1 in each file provides X_Y location of the channel hole. Ideally this X_Y location should be the same for all three channel holes that are considered triplicate data of the same channel hole. Unfortunately, due to process issues it is no so (imaging stage drifts over time). Fortunately, this drift is small and it is possible to find the triplicate measurements of the same channel hole by finding the distance between X_Y locations and constraining it to be within certain distance.

This process will also remove those channel holes with missing information (i.e. no triplicate measurements) from the further analysis process.

we do not have plausible column headings in the CSV files. So we must read the headings description file and provide column headings for further analysis.

1.2.1 Identify missing data

If the label is not present in all the files of the triplicate, then that row (channel hole) is considered missing crucial data for precision calculation and the corresponding label (and features) are not analyzed further.

We know bottom 25% and top 25% of the image is crappy (due to specimen issues) and hence the measurements from those regions have to be discarded. This means, all measures ments where Y location is less than 1024 or more than 3072 (image channel holes is of size 4096X4096) should be discarded.

Different features as they appear in the data-frames are as follows.

```
#   Column              Non-Null Count  Dtype
---  ------             --------------  -----
 0   file_name           760 non-null    object
 1   label               760 non-null    float64
 2   location (X_Y)      760 non-null    object
 3   area                760 non-null    float64
 4   area_fitted         760 non-null    float64
 5   perimeter           760 non-null    float64
 6   perimeter_Fitted    760 non-null    float64
 7   areaEnclosingCircle 760 non-null    float64
 8   areaFitEllipse      760 non-null    float64
 9   areaMinAreaRectangle 760 non-null   float64
 10  eccentricity        760 non-null    float64
 11  orientation         760 non-null    float64
 12  convexity           760 non-null    float64
 13  0degDiam            760 non-null    float64
 14  45degDiam           760 non-null    float64
 15  90degDiam           760 non-null    float64
 16  135degDiam          760 non-null    float64
```

Repeat the process for enhanced data

Now by displaying the histogram of each features in the data frame we can visualize, which feature is tightly clustered, which feature is missing and there may be files with no data that can be readily read!

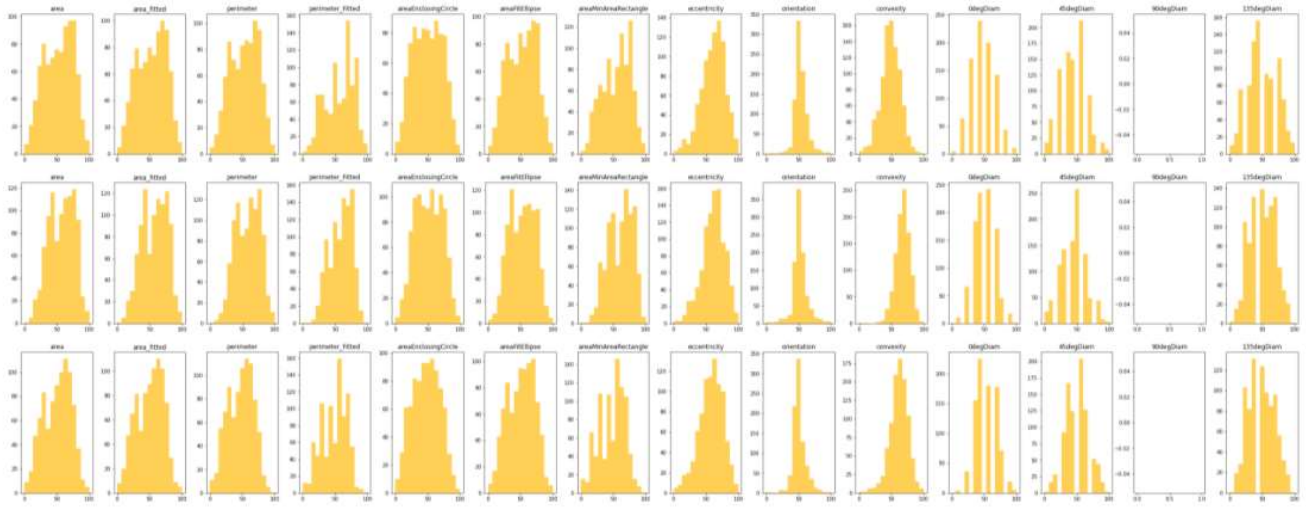

Figure1: Displaying the histogram of set of features from one of the feature files.

Display of scatter plot between features would provide information about inter feature correlation and would allow us use a representative of correlated features for further analysis

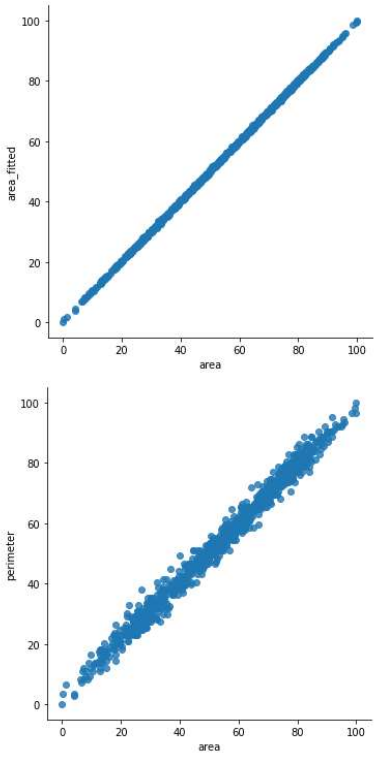

Figure 2: Scatter plot of a couple of features against each other shows how closely they are related.

We have read multiple CSV files with different feature measurement values. We have added appropriate column headings and indexed the rows. Using one of the columns (location X_Y) in the primary measurement file of the triplicates, we have recognised corresponding row indexes in the remaining two files of the triplicate. This also allowed us to deduce missing measurement in the triplicate measurement files. We removed all Nan and partially represented features. Every individual features

from all the dataframes are plotted as histogram to reveal failure to read/compute data from a couple of files and one file with inconstant data distribution. The scatter plots have shown the relationship between different features allowing us to select only a representative feature of the correlated data for further analysis.

The whole process was then repeated for features measured from enhanced image files.

1.3 Exploratory Data Analysis:

This dataset consist of several CSV files that contain number of geometric features of thousands of channel holes fabricated at nano-meter dimensions on a semconductor substrate. The data has two parts

Original features i.e. the geometric features are measured from image data that are unprocessed,
Enhanced features i.e. geometric features are collected from processed images.
Each image (of the same region) is acquired three times and the variation in geometric measurements of the same channel hole in three different images provide a measurement precision. The customer wants to know the best protocol to achive optimal precision. There are four possibilities.

Class 1: Original/unprocessed image features and measurements on actual shape of the channel holes
Class 2: Enhanced/processed image features and measurement on actual shape of the channel holes
Exploring information on the feature files extracted from fitted geometric shapes to the channel holes is a similar repittition of the exercise and hence we are not detailing it here.

In the data-wrangling step we have already completed collecting data, organizing it, and making sure it's well defined. Some data cleaning and data exploration was also done. In the exploratory data analysis step, we will explore the features, inter relationship between features by plotting the feature values and fitting standard models.

Objectives

1. Plot distribution of the features as histograms and note the files that doesn't conform to the expected distribution
2. Scatter plot features against one another to understand, highly correlated features
3. Decide which (can be more than one) is suitable feature(s) to be called as critical dimension

Let us plot mean of the feature values of the original data against that of enhanced data

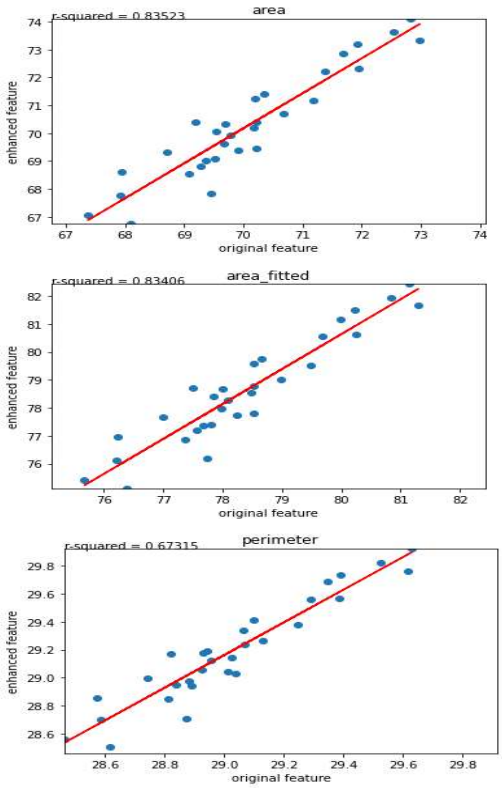

Figure 3: Mean values of original data Vs enhanced data for three features.

Below is the scatter plots of various features against each other.

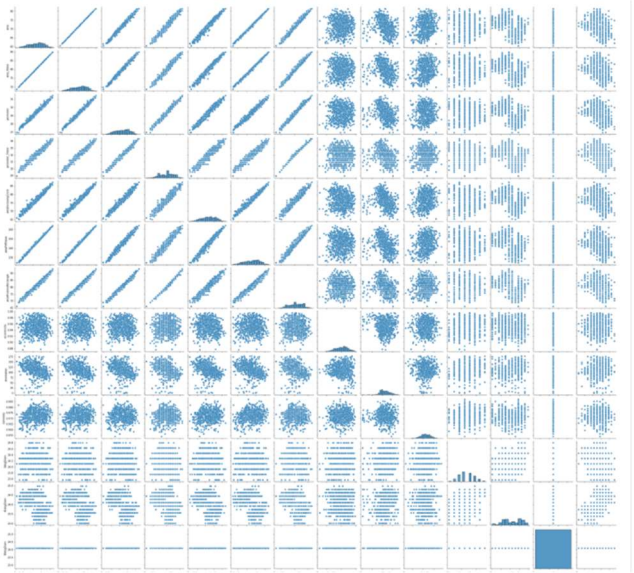

Figure 4: scatter plots of various features against each other.

What is concluded from the above histograms and the scatter plots is

1. Two sets of triplicate data are not read/wrangled properly as their histogram is not correctly displayed
2. One set of triplicates (histogram in RED) does not show unimodal property and this might be an indication of error in feature measurement or bad quality image acquisition. This set should not be used for further analysis.
3. This frequency distribution parameters (mean, sigma) shows that the enhanced data is more closer to normal (Gaussian) distribution. One can conclude from this that the "enhancement" has actually improved the feature clustering to be more tight.
4. Comparison of the mean of the features of original data and enhanced data shows that the mean value of the features has shifted somewhat.
5. 5.Comparison of the standard deviation (spread) of the features of original data and enhanced data shows that often the spread of the features is lower in the enhanced data compared to original data

Create an annotated heatmap of the correlations of the features with each other.

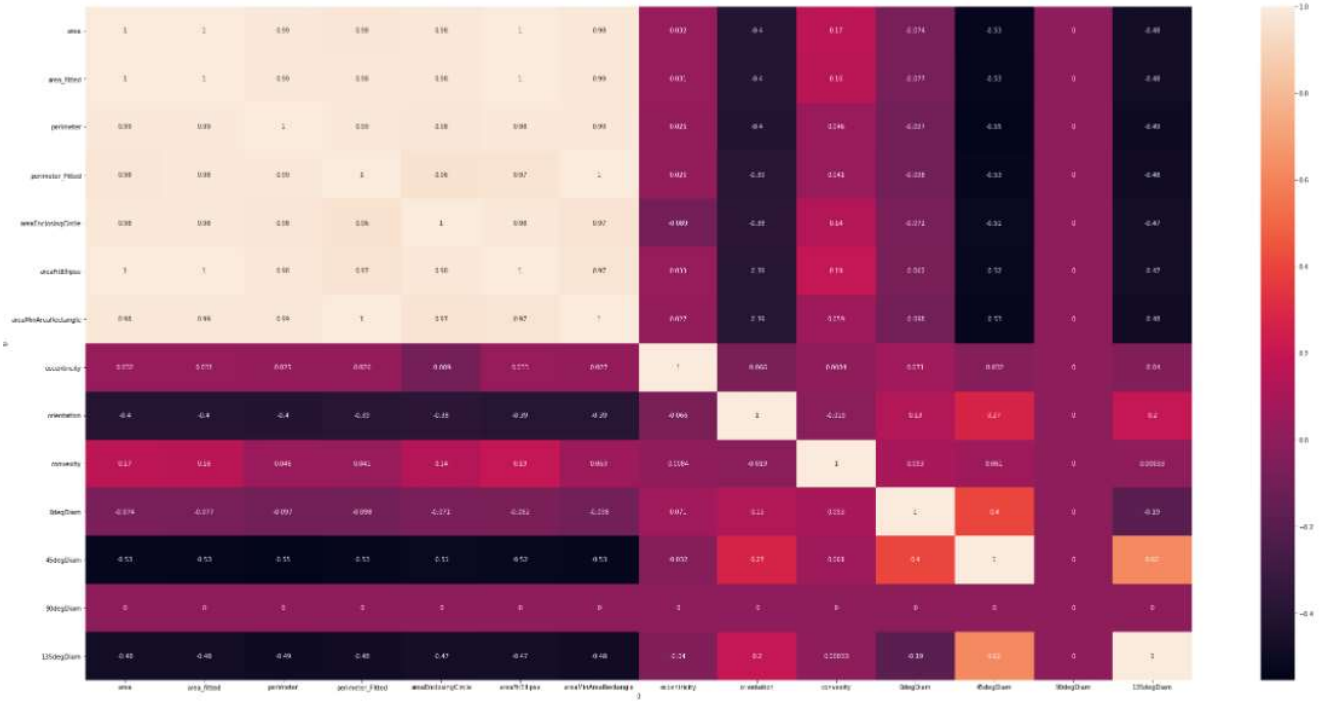

Figure 5: Heatmap of the correlation of the features.

Conclusions from the above pairplot, heatmap and the scatter plots of features Vs features from the same class of data

1. Size defining features such as 'area', 'area_fitted', 'perimeter', 'perimeter_Fitted', 'areaEnclosingCircle', 'areaFitEllipse', 'areaMinAreaRectangle'show high correlation i.e. area and the area related features (perimeter, etc.,) show high correlation
2. There is a small but significant trend in relation between 0deg, 45deg and 135deg diameter that implicitely indicate the channel holes are more elliptical than circular

3. Shape defining features such as 'eccentricity', 'orientation', 'convexity', do not correlate with area based features or diameter
4. Feature "90degDiam" has the same value for all channel holes. This shows that there is a bug in the feature engineering code (outside the scope of this work)

1.4 Processing and Modelling:

This dataset consist of several CSV files that contain number of geometric features of thousands of channel holes fabricated at nano-meter dimensions on a semconductor substrate. The data has two parts

Original features i.e. the geometric features are measured from image data that are unprocessed,
Enhanced features i.e. geometric features are collected from processed images.
We use various clustering techniques to explore possible inter-relationship between features, formation of clusters, localization of feature clusters with respect to X and Y location, etc. This will implicitly tell the stake holder whether there is any inherent fault with the imaging instrument or imaging protocol that produces feature variations and its dependency on location.

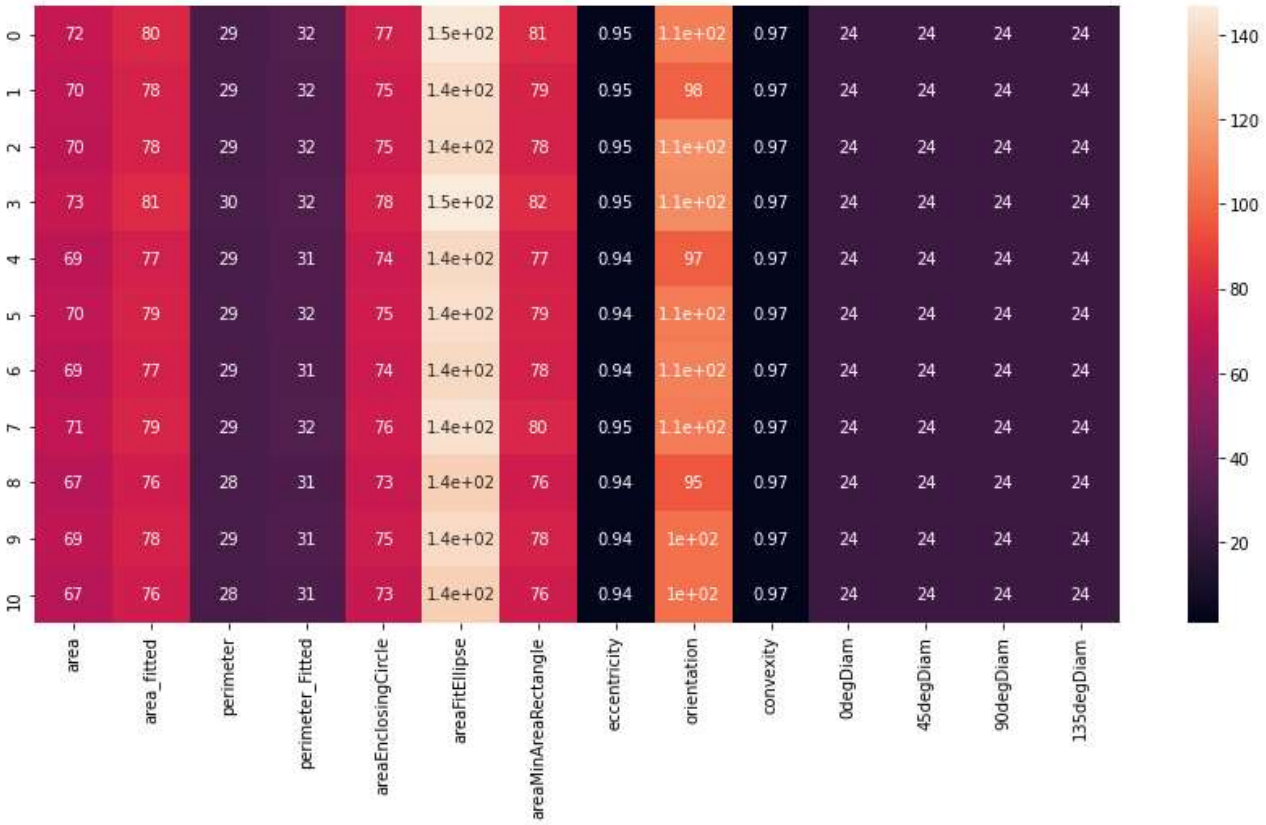

Figure 6: Average of the features from ten feature files across the triples.

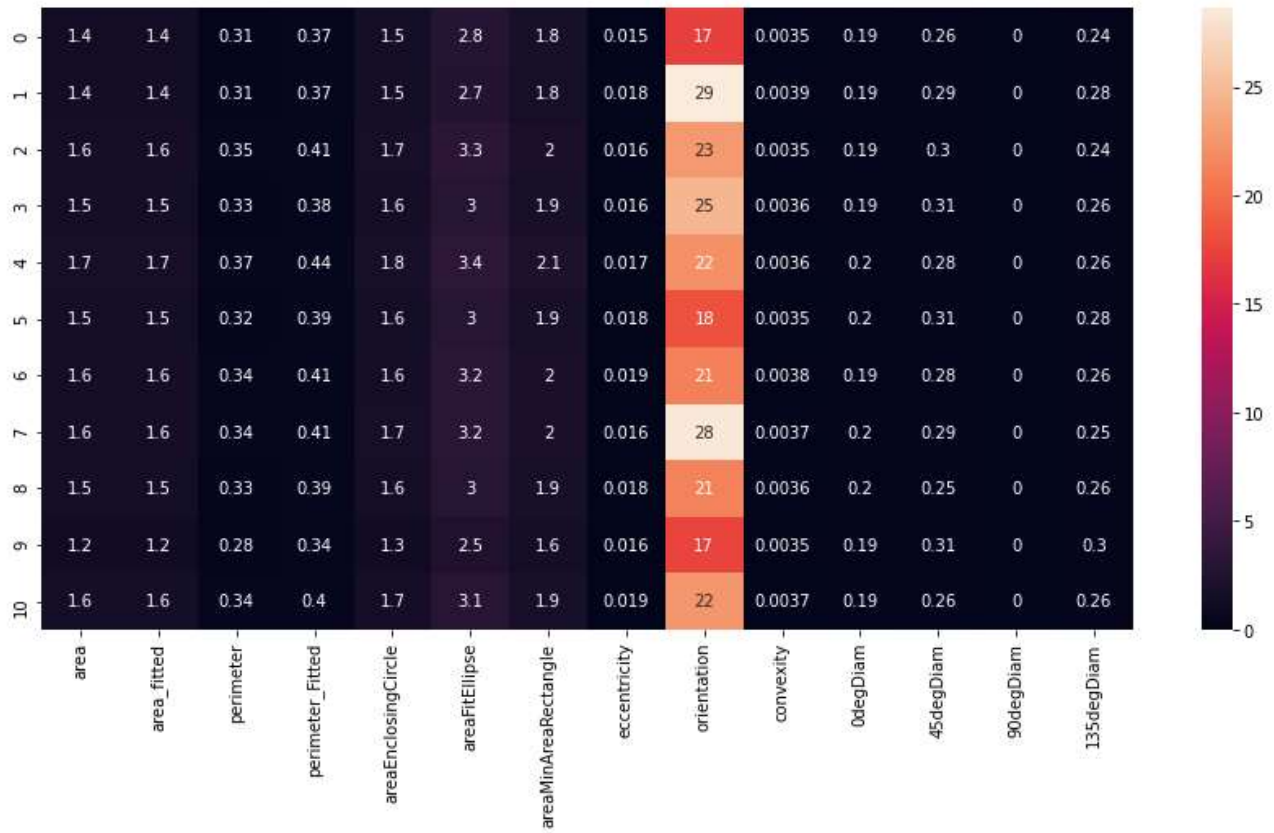

Figure 7: Standard deviation of the various features across the triplets for all the ten feature files.

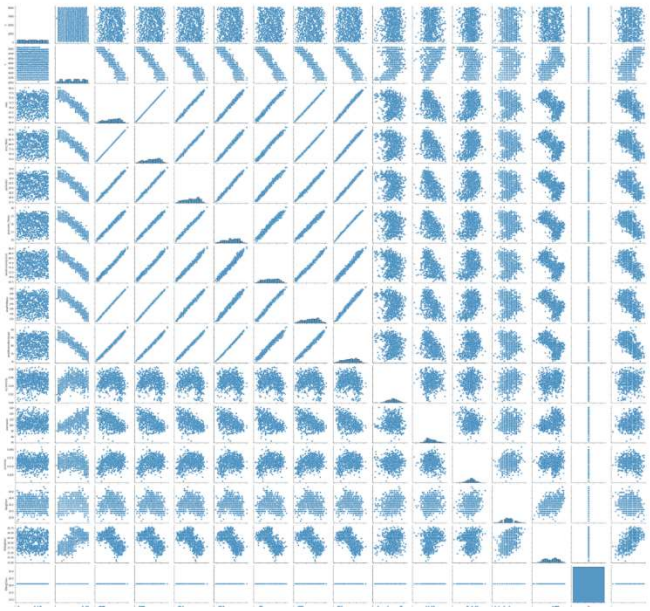

Figure 8: Scatter plot of Y location of the channel holes Vs the average feature values

what we can conclude from the above:

1. Location X has no explicit relation with the variation in the feature values (for all the features)

2. Location Y on the other hand shows a specific pattern with respect to features related to the area/size family
3. It appears that features such as area, area_fitted, perimeter, perimeter_fitted, enclosingCircle area, bestfitEllipse area, and enclosing rectangle area, decrease in their values as the Y increases. Increase in Y means traversing from top to bottom of the image (raster scan). This is an interesting observation indicating some fault in imaging or orientation of the specimen with respect to imaging direction.
4. Interestingly, the diameters (which are proportional to the area/size) appears to show a diffrent relation. But that can be explained with varying orientation of the channel holes.

Let us checkout the clustering and how it might be related to location.

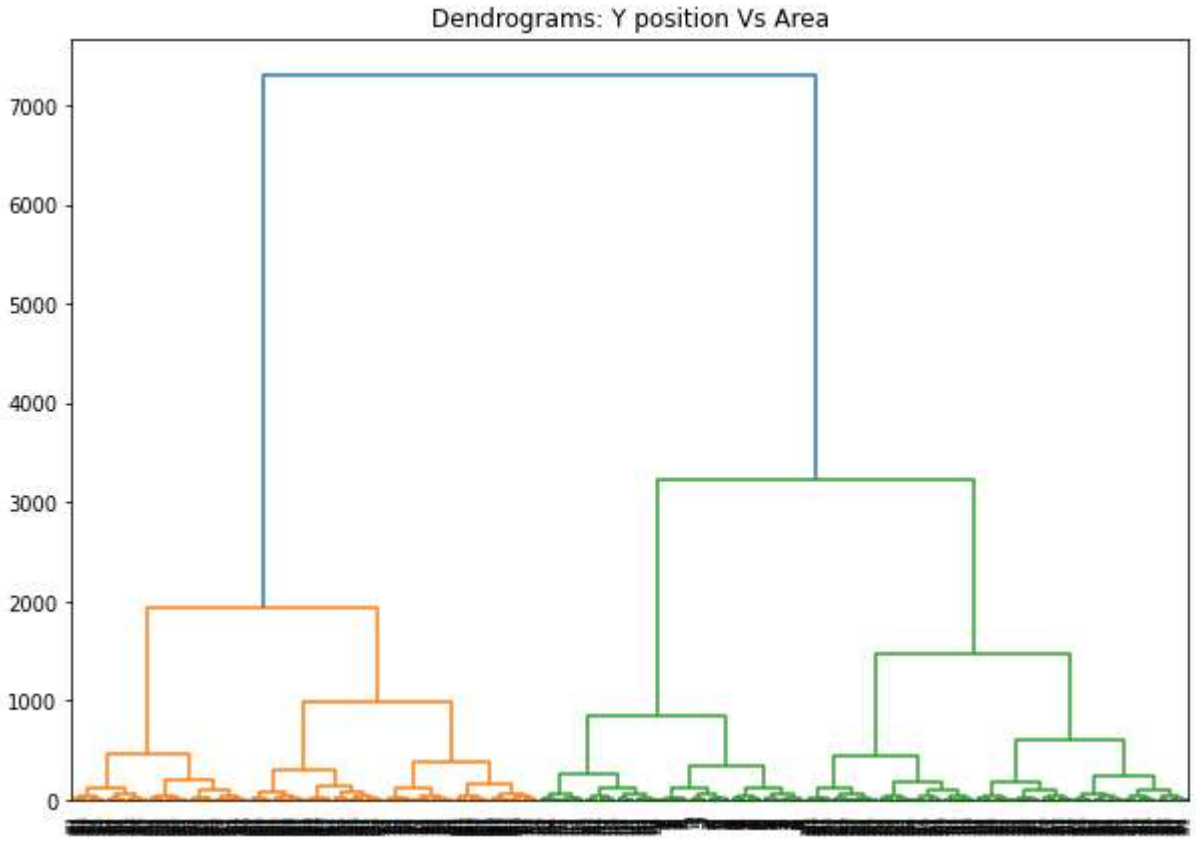

Figure 9: Dendograms showing possible clusters at various inter cluster distances. Here the Y position of the channel hole vs the area of the channel hole are clustered.

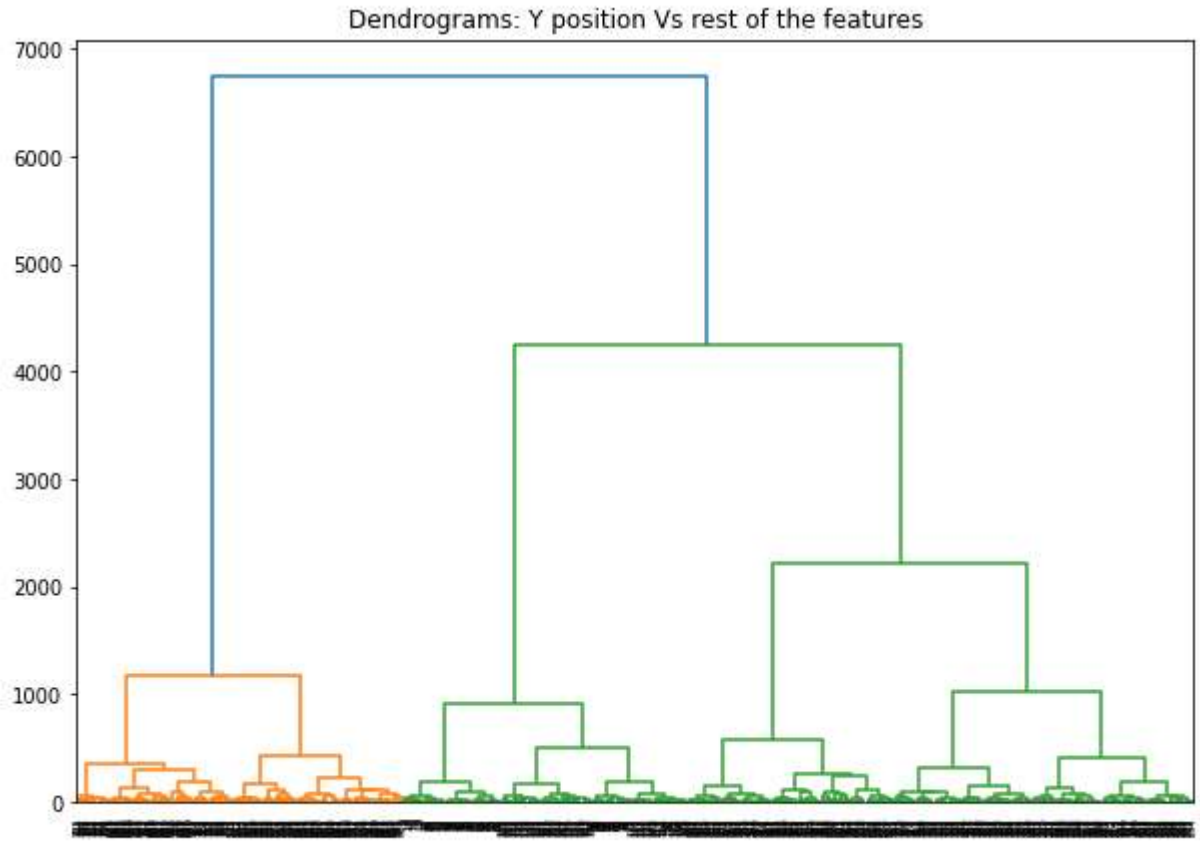

Figure 10: Dendograms showing possible clusters at various inter cluster distances. Here the Y position of the channel hole vs all other features together of the channel hole are clustered.

The cluster analysis of the data confirms our original belief that there is largely two clusters of shapes and if we force for high clustering, we may find the third cluster. Number of clusters calculation also show this as 2.5 clusters i.e. there is a significant and clean clustering that can be obtained at K=2 and also at K=3. The Dendograms also support this hypothesis.

So, our scatter plot analysis that the area related features of the channel hole decrease as the Y position increases hold true. This confirms some kind of slant or tilt in the specimen with respect to the incident electron beam.

Actual proposal did not really required this clustering analysis done but only to calculate the precision of the channel hole measurement which we have done in the previous step.