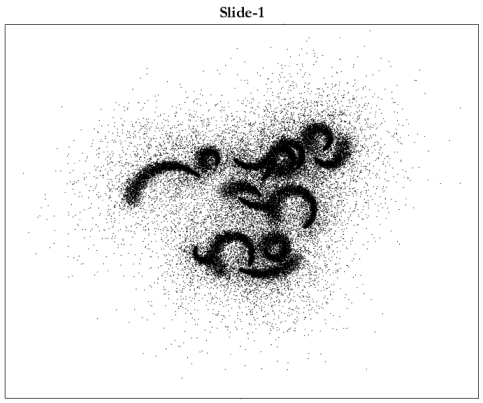
BME 6717 Dataset 5 - Clustering and Unsupervised Learning

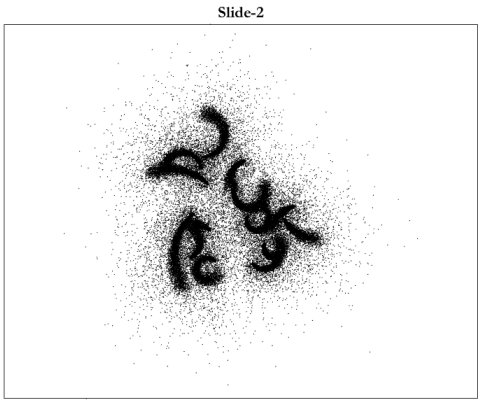
Gilgal Ansah

April 2022

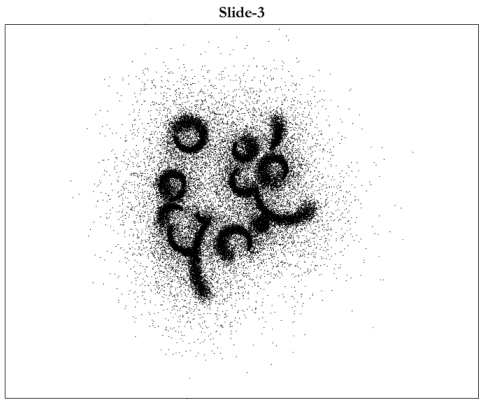
We begin by visualizing the different slides of nematodes. The images are shown below.



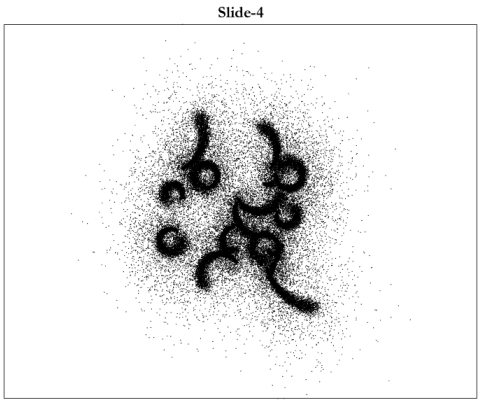
(a) Slide 1



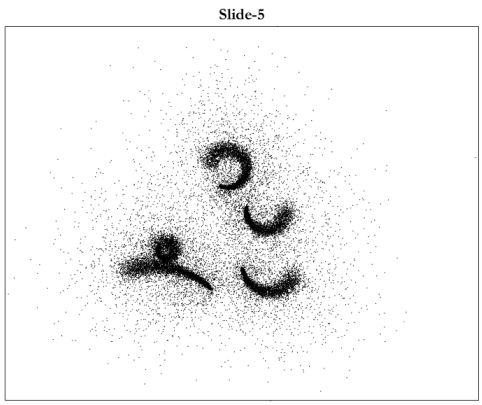
(b) Slide 2



(c) Slide 3



(d) Slide 4



(e) Slide 5

Figure 1: Five slide still-frame images of nematodes following pixel intensity thresholding

From the images, it is seen that the areas that have nematodes have a higher density of points. Unlike in datasets where each point represents an observation, a specific group density of points here represent an observation. As such, it will be prudent to choose a clustering algorithm that groups points based on region density instead of distance between points. By intuition (and from the above figure), a region with a high density of points can be thought of as a nematode (or part of a nematode).

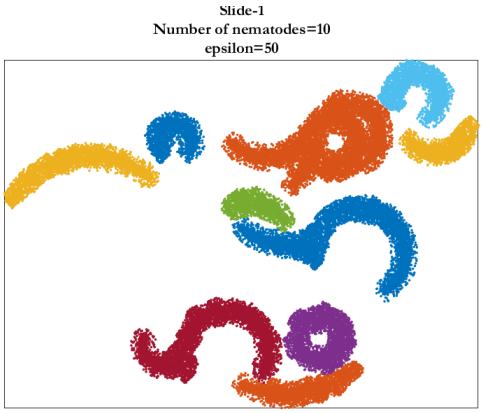
The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was used as it was the best suited for such an application. Firstly because, as implied in the name, it identifies clusters based on the density of points in cluster regions. Secondly, it takes noise into consideration and clusters noise regions also (even if they are not close to each other). Thirdly, it does not necessarily cluster based on distance between points. Lastly, it identifies arbitrary cluster shapes. Since the orientation of nematodes is arbitrary and their shapes are not regular, the arbitrary clustering ability of DBSCAN makes is able to capture these shapes and orientations.

However, although DBSCAN captures high density regions, it had to be accompanied with constant visualization of clusters to arrive at 'optimal' cluster numbers. This is also due to the arbitrary number of nematodes per slide, their orientations, and noise. The nature of each slide affected the parameters that the algorithm needed to make correct classifications.
The distance measure used for all the computations after a few iterations was the *cityblock* measure. This was chosen somewhat randomly after iterating through a couple of distance measures. The minimum number of points in each neighborhood was set at 100 and parameter tuning was done on the epsilon value.

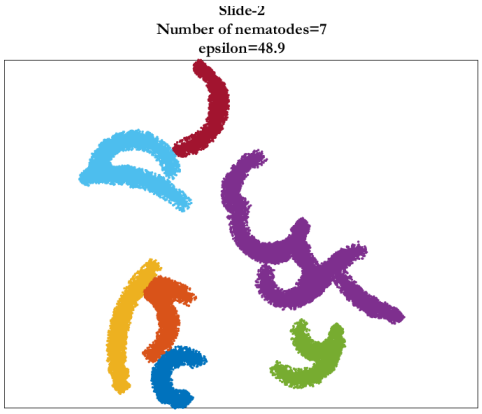
An optimal classification was arrived at by combining visualization with the silhouette measures. It was identified that DBSCAN does a really good job at identifying noise and hence most of the noise was captured in the data and silhouette measure. The noisy data had negative silhouette values and was removed from the clustering.
Visualization and silhouette measures were observed only for the identified clusters.
Figure 2 shows the result of the clustering.
From the boxplots shown in figure 3, it is observed that the silhouette values had medians above the 0.4 for each of the slides. This gives an approximate 70% (0.4 on a scale of -1 to 1) confidence of the classification.
The number of outliers (usually a small percentage) was also useful in asserting the 'accuracy' of the classification.
It is worth noting that these classifications are not very accurate for instance with slide 5 where there is a really high median silhouette value suggesting 4 nematodes, but by inspection (Fig 1) there are 5.

Insights

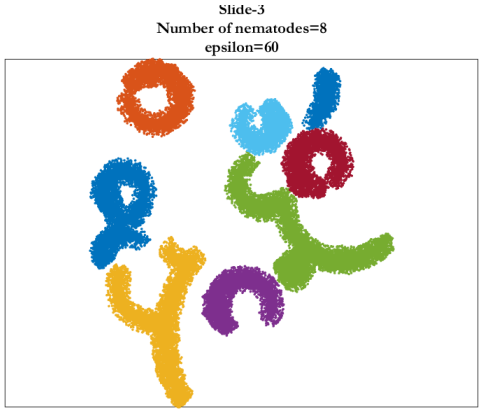
Density based clustering proved very useful in identifying regions that may contain nematodes. However, when nematodes are very close to each other, it fails to accurately segment them. Decreasing the neighbourhood of close points tends to break intact clusters instead of further clustering close, but separate observations (nematodes).
It is worth noting also that, DBSCAN does a better job at removing noise than it does clustering. However it is better suited at clustering regions on density than other algorithms, such as kmeans.
K nearest neighbours seemed to be a good algorithm, but required a query to classify and hence was not used in this analysis. The distance measure used though could be better suited to work is DBSCAN.
Hyper-parameter tunning will be required to estimate accurate parameters for the DBSCAN algorithm that can be used to automate clustering for different instances of a task (eg. slides containing different nematodes).



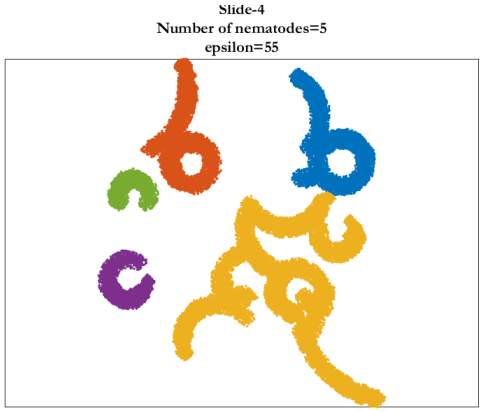
(a) Slide 1



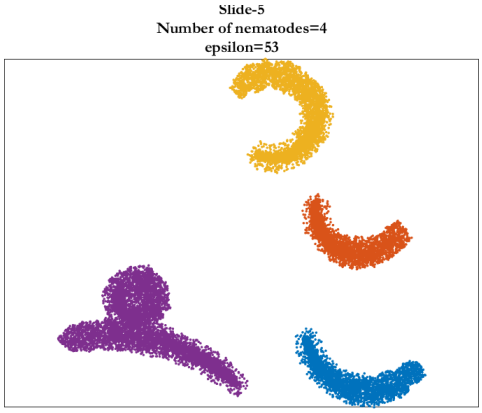
(b) Slide 2



(c) Slide 3

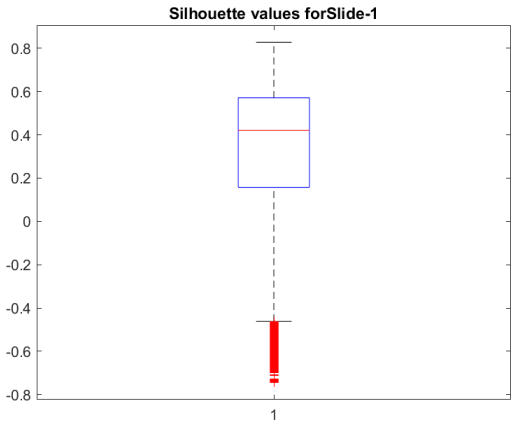


(d) Slide 4

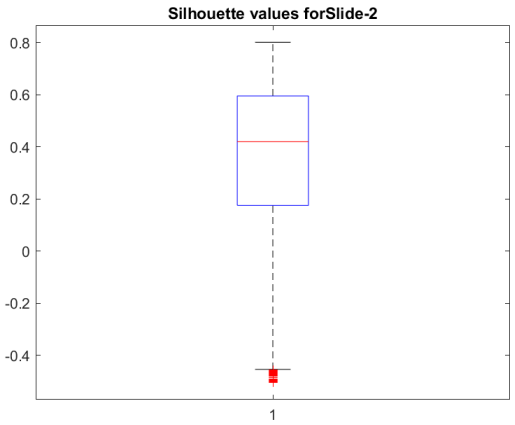


(e) Slide 5

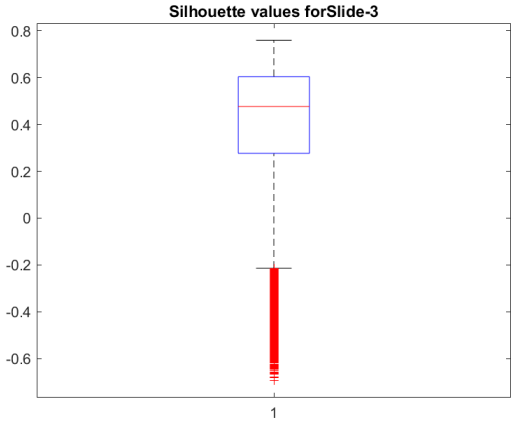
Figure 2: Clustered slides of nematodes



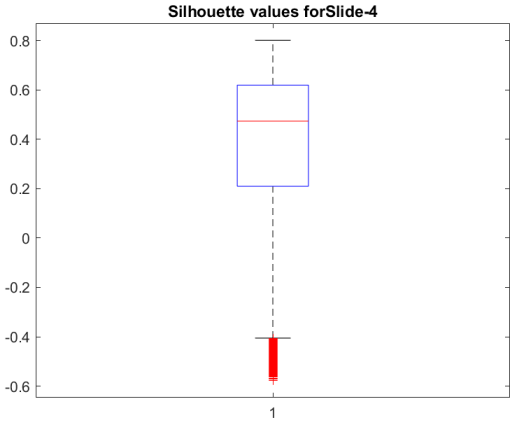
(a) Slide 1



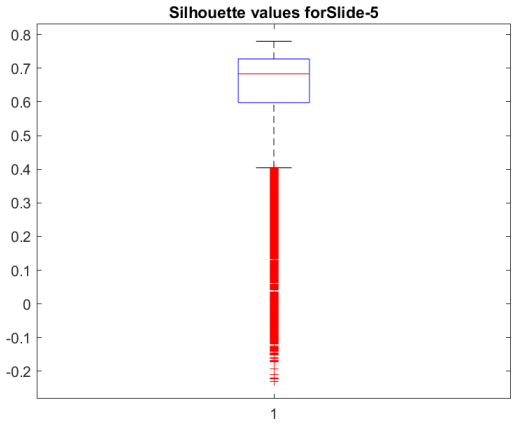
(b) Slide 2



(c) Slide 3



(d) Slide 4



(e) Slide 5

Figure 3: Box-plots showing the silhouette measures of the clusters

MATLAB

```
%% BME 6717: Clustering and Unsupervised Learning
clear
close all

load('NematodeImagesThresholded.mat');
%% Data Visualization

n=5; %slide number

slide = SlideGrab{n};

slideName=strcat('Slide-',num2str(n));

figure;
plot(slide(:,1),slide(:,2),'.k','Markersize',2)
title(slideName)
set(gca,'box','on','YTick',[],'Xtick',[],'FontName','Garamond'); axis equal

%saveas(gcf,strcat(slideName,'.png'))

%% k-distance graph from DBSCAN documentation
minpts=100;

kD = pdist2(slide,slide,'euc','Smallest',minpts);
figure;
plot(sort(kD(end,:)));
title('k-distance graph')
xlabel('Points sorted with 50th nearest distances')
ylabel('100th nearest distances')
grid
%% Clustering
epsilon = 53; %

idx = dbscan(slide,epsilon,minpts,Distance="cityblock");

clusterNum=length(unique(idx));


figure;
gscatter(slide(idx≠-1,1),slide(idx≠-1,2),idx(idx≠-1))
% gscatter(slide(:,1),slide(:,2),idx)
title({slideName,strcat('Number of nematodes=',num2str(clusterNum-1)),...
    strcat('epsilon=',num2str(epsilon))})
set(gca,'box','on','YTick',[],'Xtick',[],'FontName','Garamond','FontSize',10); axis equal
set(legend,'Visible','off');
saveas(gcf,strcat('clustered',slideName,'.png'))
%% Silhouette Measure
figure;
s=silhouette(slide,idx,'cityblock');
boxplot(s(idx≠-1));
title(strcat('Silhouette values for ',slideName))
saveas(gcf,strcat('box',slideName,'.png'))
```