DIT406: Assignment 3: Clustering

Calvin Smith, Emir Zivcic

Chalmers University of Technology/University of Gothenburg

November 23, 2021

1. Show the distribution of phi and psi combinations using:

a. A scatter plot

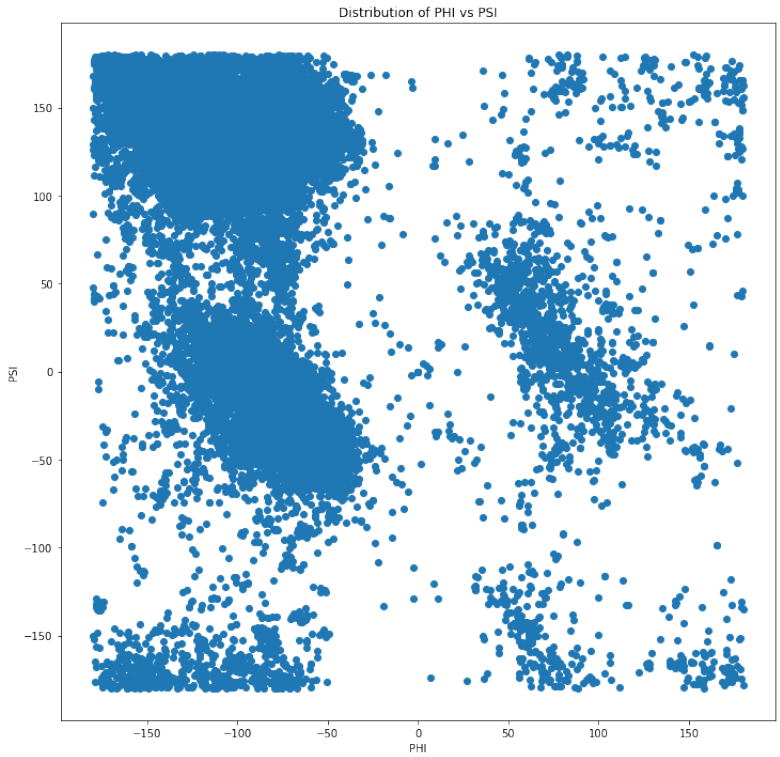

Figure 1: Distribution of PHI and PSI combinations.

b. A heatmap

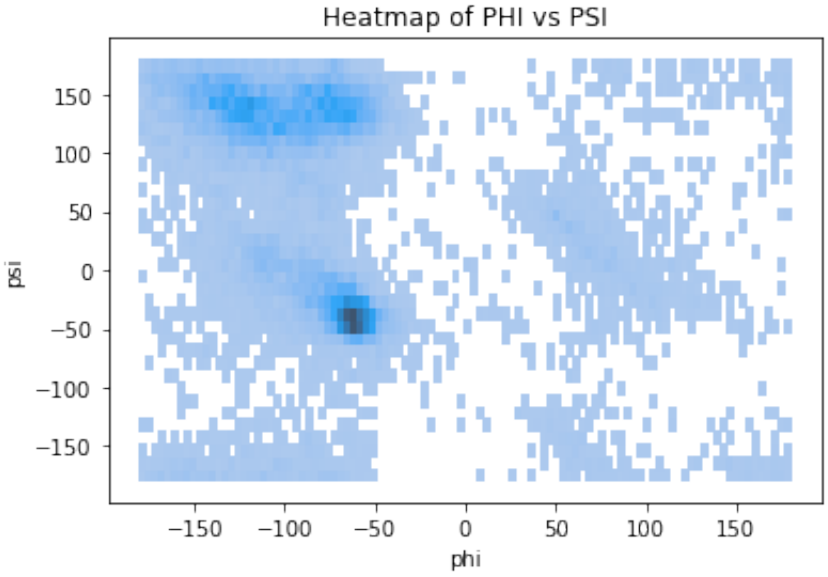

Figure 2: Distribution of PHI and PSI combinations.

2. Use the K-means clustering method to cluster the phi and psi angle combinations in the data file.

a. Experiment with different values of K. Suggest an appropriate value of K for this task and motivate this choice.

Figure 3 shows scatter plots for $k = 1$ to $k = 5$ using K-Means.

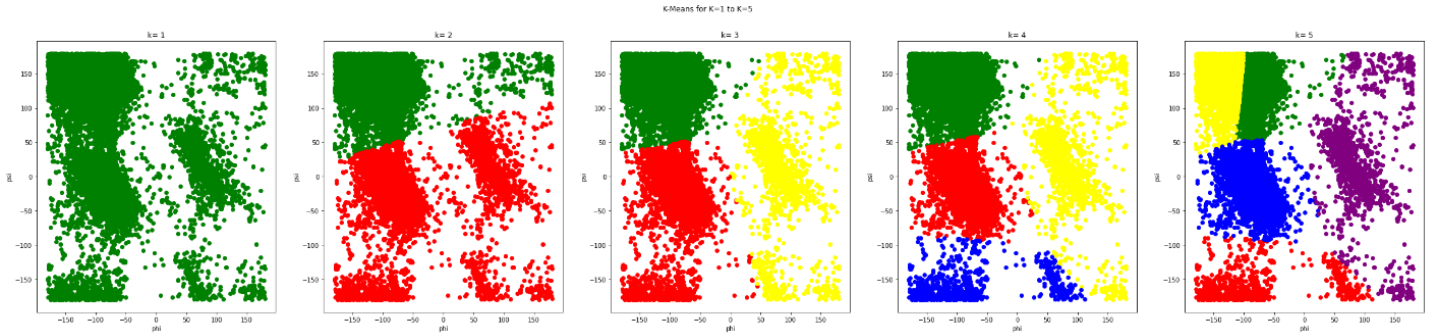

Figure 3: K-means clustering

It is hard to determine a proper clustering even by just looking, indicating that k-means is probably not the best method for clustering. The clusters behave fairly similar between runs even though there is randomization for the initial points. $K = 3$ looks to be the most sound, this can be verified with the elbow method.

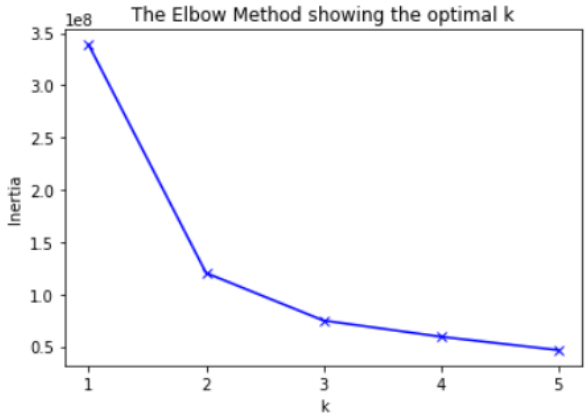

Figure 4: Elbow plot for K-means clustering

The elbow method show us the sum of the squared distances for each point from its centroid. We can see that the greatest drop off is when $K = 2$ or in our case $K = 3$, which also has a substantial drop off.

b. Validate the clusters that are found with the chosen value of K.

The clustering can be considered stable if removing random points does not affect the centroids significantly. In the graph below we can see the positions of the centroids between 20%, 40% and 80% does not change the centroid.

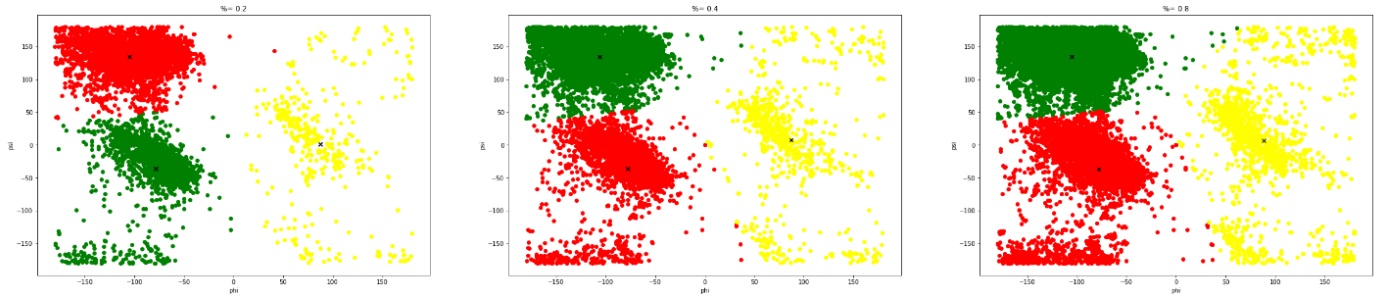

Figure 5: Subset validation for three clusters

Ideally the distances in a cluster would be 0 between all the points and infinite in the nearest cluster. A coefficient value of 1 is the best and -1 is the worst in a silhouette plot.

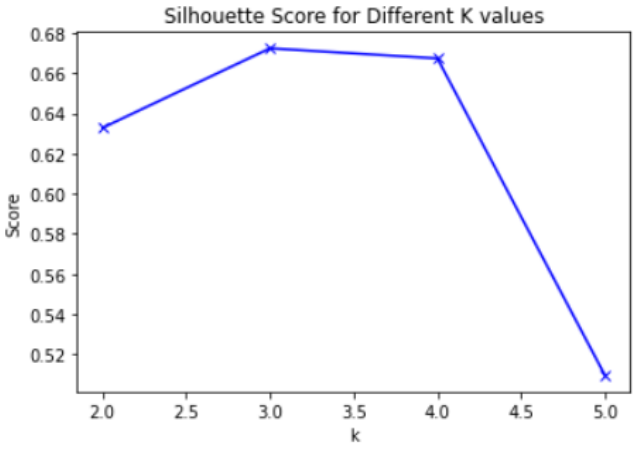

Figure 6: Finding optimal clusters

We can see, from the silhouette plot, that we receive the greatest score when we have 3 clusters.

c. Do the clusters found in part (a) seem reasonable?

By inspecting the result with three clusters visually, we can see that there is no clear reason to why the top left border between green and red was chosen to be there. When looking at a subset however then it is more clear why there are two clusters there. There clearly exists points that are outliers which should not be part of any cluster but K-Means will cluster every point, regardless if it is a reasonable distance or not.

d. Can you change the data to get better results (or the same results in a simpler way)? (Hint: since both phi and psi are periodic attributes, you can think of shifting/translating them by some value and then use the modulo operation.)

Yes! Just like centering Europe in the map makes more sense for Europeans, we should shift the data so that it is more centered, instead of cutting on the left side and continuing on the right. Plot below is achieved by transforming a -180 to 180 plot to a 0 to 360 plot instead, and shifting phi with 20 degrees and psi with 300 degrees.

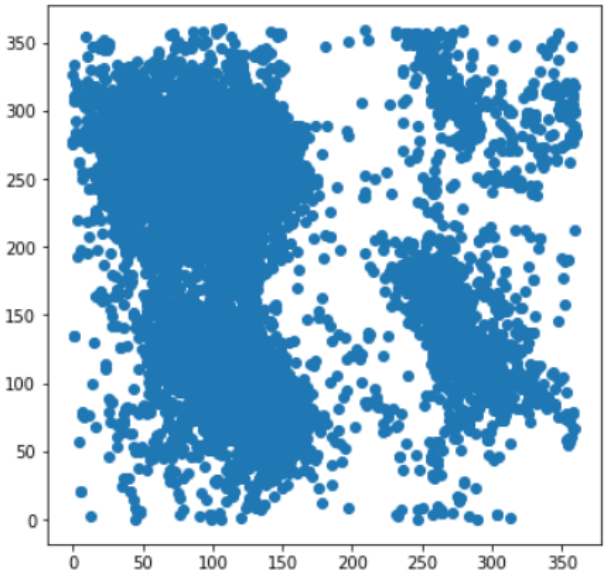

Figure 7: Psi and phi shifted

Shifting by trial and error ends in a scatter plot that shows entire clusters together instead of rolling over and continuing on the opposite side.

3. Use the DBSCAN method to cluster the phi and psi angle combinations in the data file.

A *StandardScaler* has been applied to two figures below.

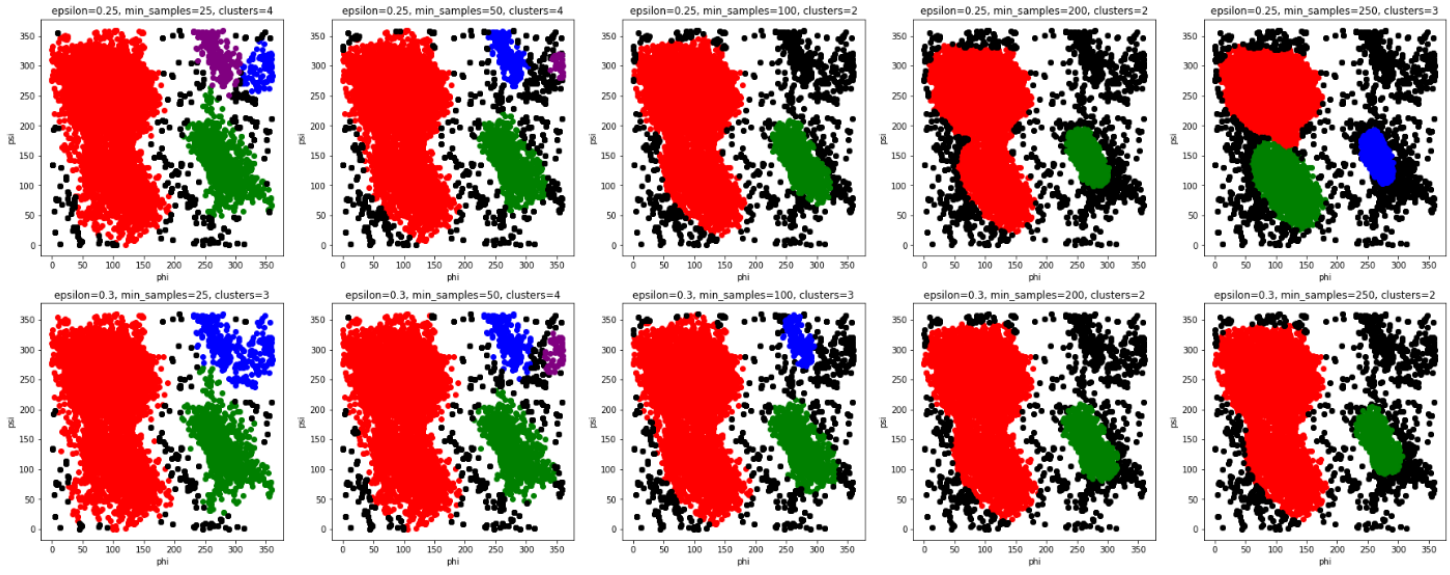

Figure 8: Combinations of epsilon = 0.25 to 0.3 and min samples = 25 to 250

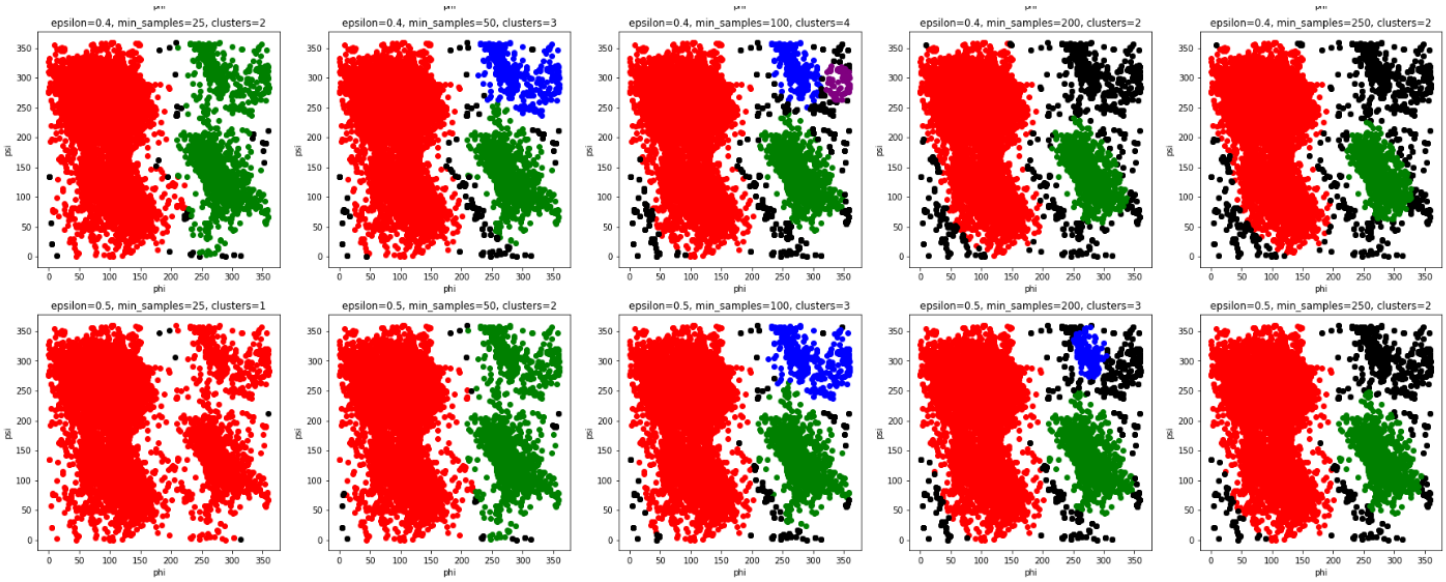

Figure 9: Combinations of epsilon = 0.4 to 0.5 and min samples = 25 to 250

a. Motivate: i. the choice of the minimum number of samples in the neighbourhood for a point to be considered as a core point, and ii. the choice of the maximum distance between two samples belonging to the same neighbourhood ("eps" or "epsilon").

Min-points should be chosen, as a rule of thumb when domain knowledge is lacking, as two times the dimension of the data. In this case it would be 4 but a range from 25 to 250 was chosen instead due to the high density of the data. Finding the optimal epsilon was found by looping through different values and analyzed visually to see which produced the best result. Min-points neighbors is unspecified due to it not changing the outcome. The same graph was produced from different n nearest neighbors.

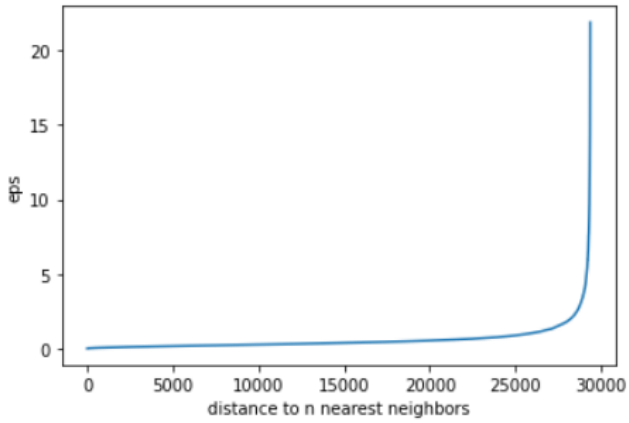

Figure 10: Finding optimal epsilon

Epsilon could have also been found by analyzing a plot of epsilon and how it relates to the n nearest neighbors for each data point. By sorting the distances we can see that there will be about a few hundred points that will not be a part of any cluster since the distance they have to their n closest neighbors is too large. A good value for epsilon ends up being about 1.8 for 28 000 points or even up to 3.8 for 29 000 data points.

b. Highlight the clusters found using DBSCAN and any outliers in a scatter plot. How many outliers are found? Plot a bar chart to show which amino acid residue types are most frequently outliers.

A *StandardScaler* has been applied in figure 11

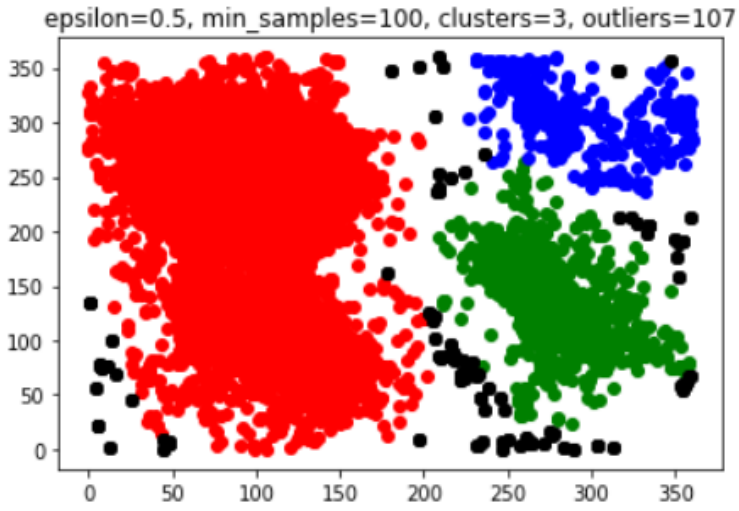

Figure 11: Optimal DBSCAN

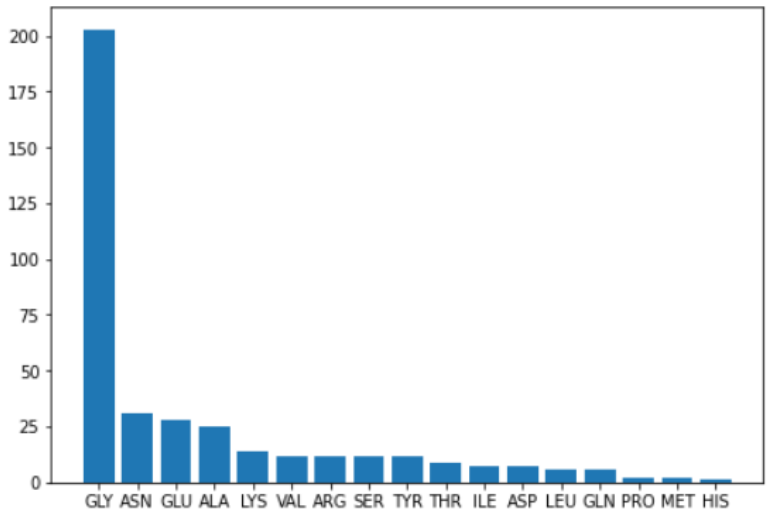

Figure 12: Optimal DBSCAN residue

c. Compare the clusters found by DBSCAN with those found using K-means.

K-Means is easier to implement if a person has no domain knowledge of the data-set, this does however mean that data points need to be filtered if they look like outliers since K-Means will cluster every point, regardless of distance. A dendrogram could be utilized to disallow points to be part of a cluster if they exceed a certain distance.

Even without domain knowledge we could determine an epsilon and min-points that results in a nicer clustering compared to K-Means.

d. Discuss whether the clusters found using DBSCAN are robust to small changes in the minimum number of samples in the neighbourhood for a point to be considered as a core point, and/or the choice of the maximum distance between two samples belonging to the same neighbourhood ("eps" or "epsilon").

A small change in in min-points will not have a great effect due to the high density of data points. Epsilon however will have a great difference due to the same reason, the data is dense.

4. The data file can be stratified by amino acid residue type.

a. Use DBSCAN to cluster the data that have residue type PRO. Investigate how the clusters found for amino acid residues of type PRO differ from the general clusters (i.e., the clusters that you get from DBSCAN with mixed residue types in question 3).

As previously mentioned, determining the optimal parameters for DBSCAN can be quite arbitrary and requires domain knowledge. The residue type PRO contains 1596 observations and the data is two-dimensional. Taking this into account while also observing from figure 13 that the data seems to be centered around two distinct clusters with some points bridging the gap between (almost looks like a cluster of its own) the clusters and some points scattered around. With this in mind, we have chosen $minPts = 30$. To determine the parameter *epsilon* we have taken the average distance of each point from its 100 nearest neighbors and plotted the distances in ascending order. The result of this can be seen from figure 14, and a suitable value for epsilon is chosen to be $epsilon = 15$.

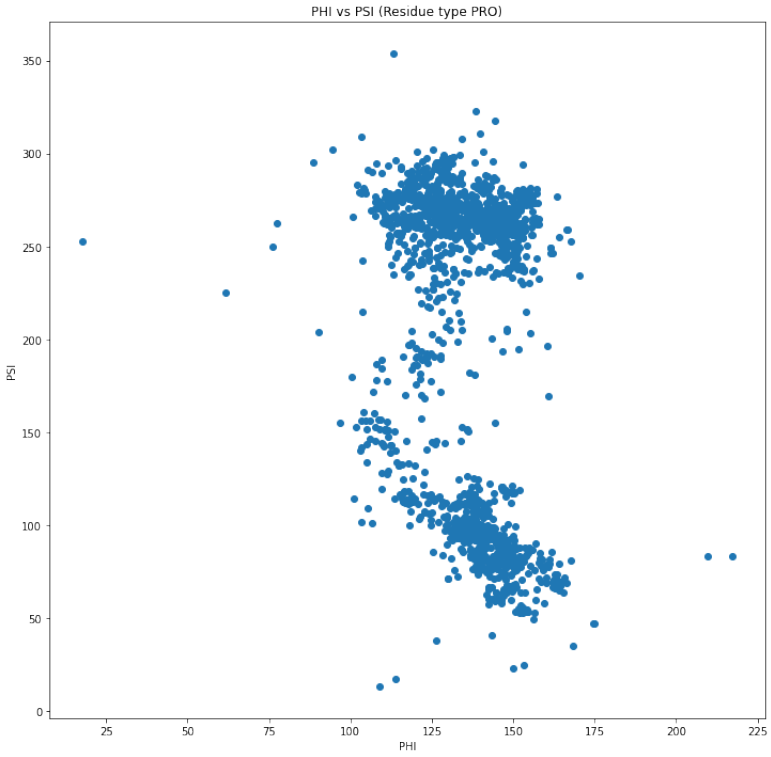

Figure 13: Distribution of PHI vs PSI (residue type PRO)

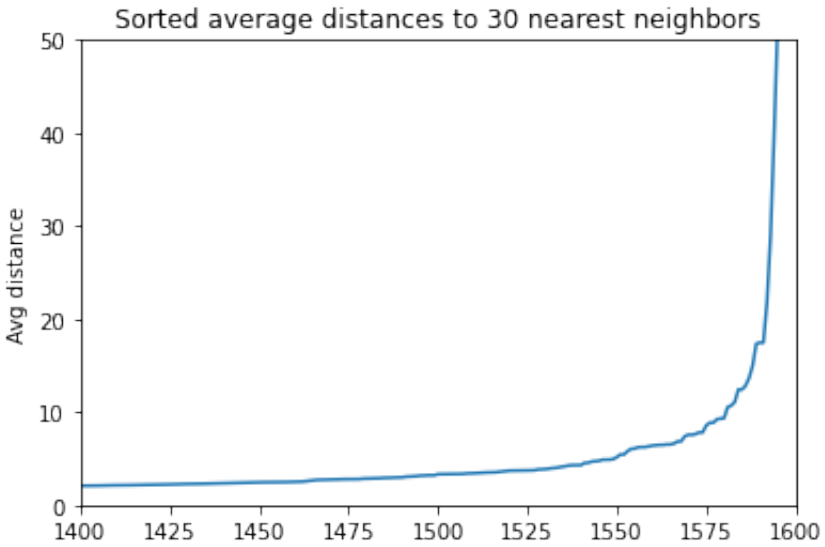

Figure 14: Determining epsilon through an 'elbow' plot.

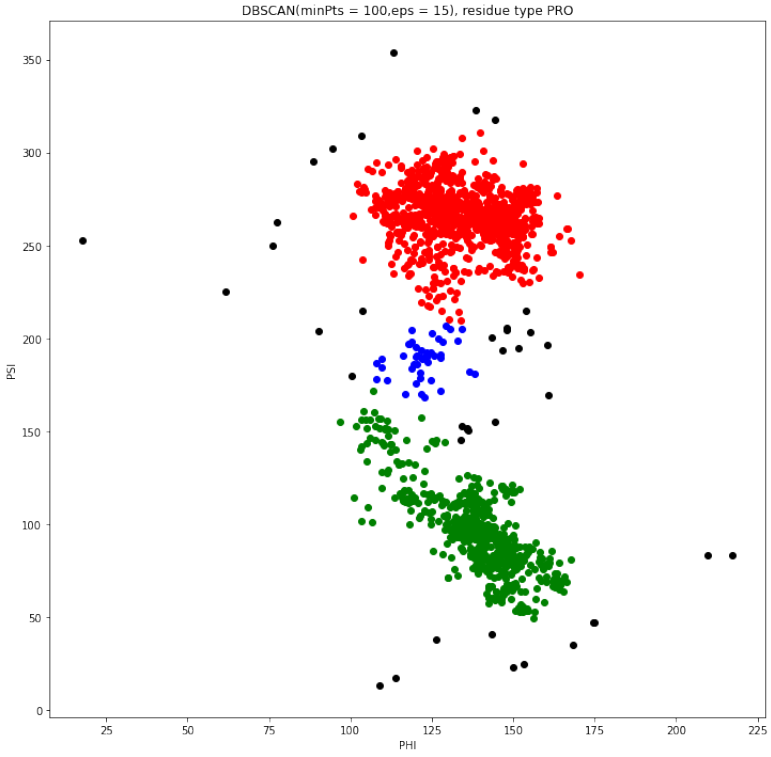

Figure 15: Clustering results from DBSCAN with residue type PRO

Figure 15 shows the result of running DBSCAN($minPts = 30$, $epsilon = 15$). The algorithm produces three clusters and outliers (plotted in black) spread around our clusters. Out of 1596 data point, 1557 point are assigned to a cluster and 37 points are considered outliers (2.3%). This seems to resonate quite well compared to the result from running DBSCAN on the full dataset, where few points of residue type PRO are clustered as outliers.

b. Now use DBSCAN to cluster the data that have residue type GLY. Investigate how the clusters found for amino acid residues of type GLY differ from the general clusters.

Compared to residue type PRO, the GLY data has a lot more spread and is not as centered at PHI. From figure 16 it looks like we have four distinct areas where most of our data is located. Similarly to the previous section, we have chosen the parameters based on the dimensions of the data (2), the number of observations (2176), a distance plot as well as looking at

the spread of the data. Just as with residue type PRO, it seems reasonable to choose $minPts = 30$ and $epsilon = 15$, and obviously for comparable purposes.

Figure 16: Distribution of PHI vs PSI (residue type GLY)

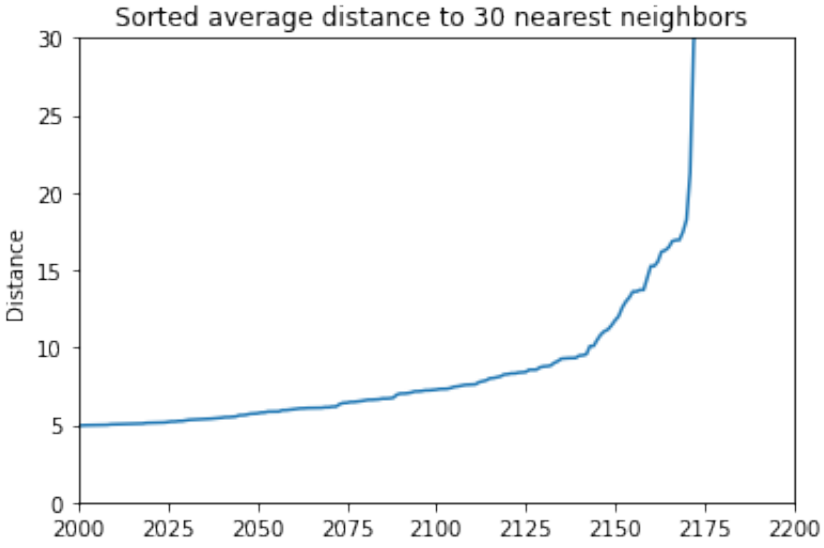

Figure 17: Determining epsilon through an 'elbow' plot.

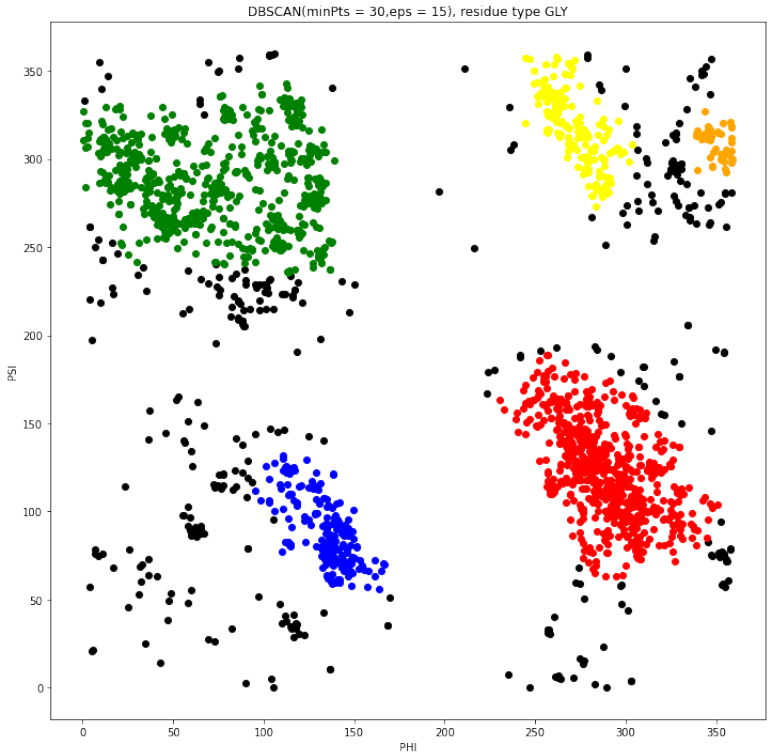

Figure 18: Clustering results from DBSCAN with resiude type GLY.

The clustering result is shown in figure 18 and produced 5 clusters. Out of 2176 observations, 1811 are assigned to a cluster and 365 are considered as outliers. Once again, this seems to resonate well with the results from DBSCAN on the full dataset, where GLY was the most frequent residue type among outliers.