Clustering: Determining Olive Oil Region Based on Acid Components

Jennifer Luu, Sara Rettus, Jelena Segan, Louis Tran

May 12th, 2021

1 Introduction

Olive oil is a liquid fat created by crushing olives and extracting the oil by some mechanical means. The majority of olive oil originates from the Mediterranean Basin, which is a collective term for the regions —Europe, Asia, and Africa — that surround the Mediterranean Sea. In this particular data set, we investigate 572 olive oil samples that come from Italy. [1]

This data set contains two categorical variables: *region* and *macro.area*. The column *region* includes the three major sections of Italy: Northern Italy, Southern Italy, and Sardinia. Sardinia is an island in the Mediterranean Sea that is a part of Italy. The column *macro.area* refers to the specific location of each region. Northern Italy consists of East Liguria, West Liguria, and Umbria. Southern Italy contains North Apulia, South Apulia, Calabria, and Sicily. Sardinia is its own major region, but it is divided into coastal and in-land.

The remaining 8 continuous columns record the percentage composition of fatty acids — palmitic, stearic, oleic, linoleic, palmitoleic, arachidic, and eicosenoic — found in the Italian olive oils. Each olive oil sample contains some percentage of up to 8 of the fatty acids. The measurements for each sample are percentages multiplied by 100 because some of the original percentages, such as eicosenoic, came out to be as low as 0.01%.

The goal of our project is to determine groups that capture the relationship between the acid compositions and regions and convey the differences through K-means clustering. We will measure the performance of this method with different clusters by creating confusion matrices to visualize the variation and providing accuracy scores.

[1] [Ric05]

2 Statistical Methods: Selecting the Number of Clusters

Climate, soil, drainage, and exposure are just a few of the important factors, which play a role in growing olives trees. If there is anything to be learned from the different amounts of fatty acids in olive oils, then something about their growing conditions must explain the differences in the component levels. [2] The clusters may provide some insight into the growing conditions of olive oil based on region.

While we were provided with the regions and macro areas from which these samples were taken, we decided against using those labels as our choice for cluster numbers. Regions and macro areas are social constructions in that borders and territories are drawn for political, economic, and historical reasons. Meaning that there is no reason to assume that separate regions will have separate climates or growing conditions.

It is not unreasonable to think that in some cases regions that belong to two different macro areas may have growing conditions more similar to each other than other regions in their macro area. For example, San Francisco and San Jose could be considered part of the same macro area, but the climate conditions in San Francisco are closer to those in the coastal cities of northern California than they are to the climate conditions of San Jose.

Despite not using our regions and macro areas to determine the number of clusters, we did use these labels as a way to provide meaning to the way in which the data was clustered, and to allow us to test our cluster for accuracy.

We assigned a *pseudo true* labels to our data after it was clustered based on max number of observations by region found in the cluster. [3] So, for example, assume we have three clusters and a total of 15 observations in Region One. If there was one observation from Region One in cluster 0, four observations in cluster 1, and ten in cluster 2, then we assign Region One's true label as cluster 2. Given fewer clusters than regional areas, some regions will be a part of the same clusters.

K-means clustering requires us to manually select the number of clusters. Since we were not using region, we needed another method with which to select our clusters. We used both the elbow and the silhouette methods for this process.

The elbow method helps visualize where there are diminishing returns on the amount of variance, which can be explained by additional clusters, while the silhouette method looks at how well each point is matched to its cluster, on a scale of -1 to 1. The silhouette score is the average of all these scores, where the higher the score the better the match.[4] [5]

[2] [ST02]

[3] We are calling these labels pseudo true because they are based on the assumption that the fatty acids in olive oil can be clustered according to groups of regions.

[4] Scikit-Learn's *inertia_* uses the sum of the squared distance between the observation and its cluster center to determine the point of diminishing returns for adding additional clusters.

[5] Silhouette score ranges from 1 to -1 where high values show that point is well matched to

Using these methods, we selected three different possible cluster arrangements. Because we are dealing with the same kind of measurement, i.e., percentage of fatty acid found in olive oil, and since our measurements can be classified as having a similar class, i.e., fatty acid, we first ran the elbow and silhouette method on unscaled data. The elbow method on the unscaled data suggested a three-cluster model, while the silhouette with the highest score suggested a two-cluster model.

Nonetheless, because the fatty acid oleic makes up over 73% of the total fatty acid for all our samples combined, with the next highest acid concentration attributed to palmitic at approximately a paltry 12%, we also decided to scale the data so that the concentration of oleic in our samples would be lessened in its effect on our model. After scaling the data, both the elbow method and silhouette method suggested five clusters.

3 Statistical Methods: Test, Train and Random State

Our data was split into training and testing data, where 30% of our data was reserved for testing. All our test and clustering configurations were completed on the same test and train set.

Because K-means randomly selects which observations with which it begins its distance measurements, different observations can be put into different groups on different runs. To make sure that our clusters were relatively stable, we ran our clusters on three different random states where we concluded that a cluster was stable if, after determining the *pseudo true* label, the same regions were clustered together.

After running random state 10, 11, and 12 to perform K-means clusters, we were able to determine that each of our cluster groups were stable. Because random state 10 was the first state we specified for K-means, we maintained this state for the analysis used in our cluster models.

4 Statistical Methods: Analysis

We used several methods in which to analyze our models. We performed an accuracy test on each of our three cluster models using our *pseudo true* labels along with creating a heat map, which shows the accuracy level for each cluster.

We determined the two best models were those of the two clusters and five clusters. However, we could not determine which one was best. The model

its cluster and low values show that it is not well matched.
a(i) = The average distance of that point with all other points in the same clusters.
b(i) = The average distance of that point with all the points in the closest cluster to its cluster.
s(i) = silhouette coefficient or i'th point using below mentioned formula.
$s(i) = \frac{b(i)-a(i)}{max(b(i),a(i))}$
Silhouette score = $1/n \sum_{i=1}^{n} s(i)$

with only two clusters had a higher accuracy score, and does not risk overfitting the data. However, the five cluster model appeared to make more sense of the regional maps. We concluded that we were unable, without additional information, to determine one model's supremacy over the other.

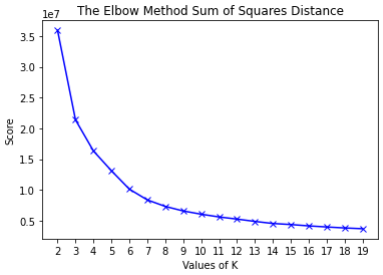
We completed our analysis by assuming that the five cluster model was an accurate depiction of the clusters, which could be generalized to these regions in Italy, and visualized the data using the Principle Component Analysis (PCA).

We chose Principle Component Analysis (PCA) as our dimension reduction method to represent our data in 2-dimension instead of 8-dimension. The PCA method transforms the data that we have into a new set of variables by using eigenvectors and eigenvalues from the correlation matrix between the original variables. Each variable in this new set of features contains a certain percentage of information from all 8 original variables, and the 2 variables that contain the highest information will be chosen to represent our data in 2-dimension.

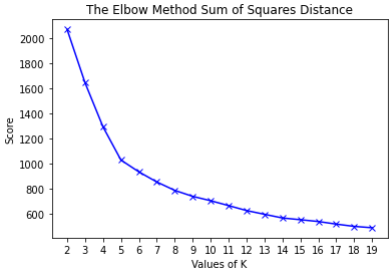
In nontechnical terms, the PCA method constructs 2 new variables to visualize the data in low dimension but still keep most of the information. This PCA plot combined with a regional map of the area, would allow us to describe which linear combinations of fatty acids were attributable to each region.

5 Elbow Method Results

In using the elbow method, we are looking for a sharp point, or *elbow*, in the graph, which signifies a slowing rate of additional clusters explaining the variation in the data. The sharp point for the unscaled data appears at k = 3, while the sharp point for the scaled appears at k = 5.



Unscaled



Scaled

6 Silhouette Method Scores

In running the elbow method first, we were given a visual of where our differences in variances per cluster become very small, which allowed us to shorten the range for the computed scores. In computing scores for between 2 and 7 clusters, using unscaled data, the highest silhouette score was $\approx .5455$ for k = 2. The highest

score using scaled data, on a range of k = 2 to k = 11, the highest score was ≈ .3827 for k = 5.

7 Analysis: Accuracy Scores, Regional Maps, and Heat Maps

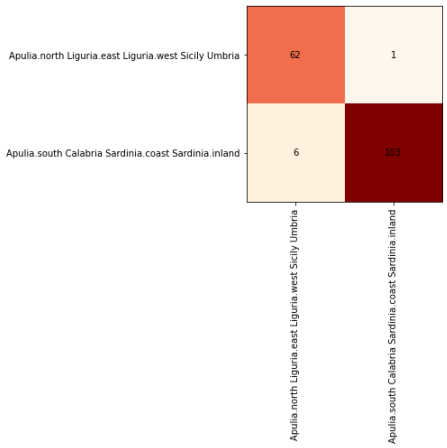
7.1 Two-cluster model

When the data are put into only two clusters, all the regions, except for Calabria are clearly in one or the other cluster, with 0 to 1 test observations being put into a different cluster. While more Calabria observations are in cluster one over cluster zero, cluster 1 only accounts for 68% of the observations.

Cluster 0: North Apulia, East Liguria, West Liguria, Sicily, and Umbria.
Cluster 1: South Apulia , Calabria, Sardinia (Coast), Sardinia (Inland)
Accuracy Score: 0.9593023255813954

Cluster 0: ≈ 99% accurate	Cluster 1: ≈ 91% accurate

Heat Map



The regional areas, which are clustered together is shown on the regional map below. If not for Sicily falling into cluster 0, the map would have indicated that areas to the South and with large coastlines formed one cluster and the areas with less coast line and farther to the North were part of another cluster. Under such a situation, it would not be unreasonable to think that perhaps coastal southern climates effected the fatty acid components of olive oil.

However, Sicily is one of the southern most regions, and its borders enjoy a full coastline. While we cannot conclude that this is the best model, it does provide us with an avenue of further investigation, to see if the conditions under

which Sicily produces its olives, i.e., perhaps they are grown inland, can seen in other regions in its cluster.

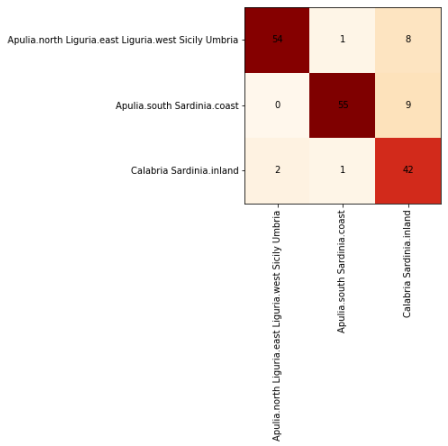
7.2 Three-cluster model

The three-cluster model maintained the first cluster, but split the second cluster into two different clusters, where South Apulia and and the Sardinia coast were part of one cluster, and Calabria and inland Sardinia broke off into a third cluster. The three-cluster model had the worst accuracy score, of all of our models, which was the ultimate reason for not selecting this model.

Cluster 0: North Apulia, East Liguria, West Liguria, Sicily, and Umbria
Cluster 1: South Apulia and Sardinia (Coast)
Cluster 2: Sardinia (Inland) and Calabria
Accuracy Score: 0.877906976744186

Cluster 0: $\approx 96\%$ accurate	Cluster 1: $\approx 96\%$ accurate	Cluster 2: $\approx 71\%$ accurate

Heat Map



In addition, to the lower accuracy scores, it is far more difficult to make meaning from the regional areas, which clustered. So, for example, inland Sardinia is grouped with Calabria, which may indicate that Calabria's olives are grown more inland. Yet, Sicily is not placed in either group, again being placed with the northern regions.

7.3 Five-cluster model

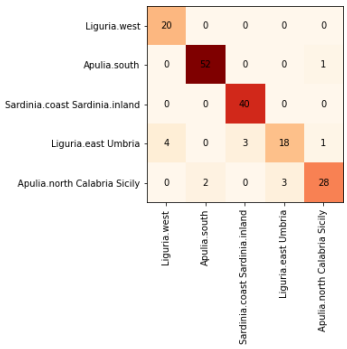
In the five cluster model, West Liguria and South Apulia each have their own cluster, and the entire macro region of Sardinia is put into its own cluster. The cluster, which previously contained East Liguria, North Apulia, Umbria, and Sicily was split into two, with East Liguria and Umbria being clustered together and North Apulia and Sicily part of another cluster. The mystery is that Calabria, which had been clustered with South Apulia or Sardinia previously is now clustered with North Apulia and Sicily. While the accuracy score for the five cluster model was less than the two cluster model, it was relatively high. In addition, there was some sense to be made of the regional maps.

Cluster 0: West Liguria
Cluster 1: North Apulia, Calabria, and Sicily
Cluster 2: South Apulia
Cluster 3: Sardina (coastal and inland)
Cluster 4: East Liguria and Umbria
Accuracy Score: 0.9186046511627907

Cluster Accuracy

Cluster 0: $\approx 83\%$	Cluster 1: $\approx 96\%$	Cluster 2: $\approx 93\%$	Cluster 3: $\approx 86\%$	Cluster 4: $\approx 93\%$

Heat Map

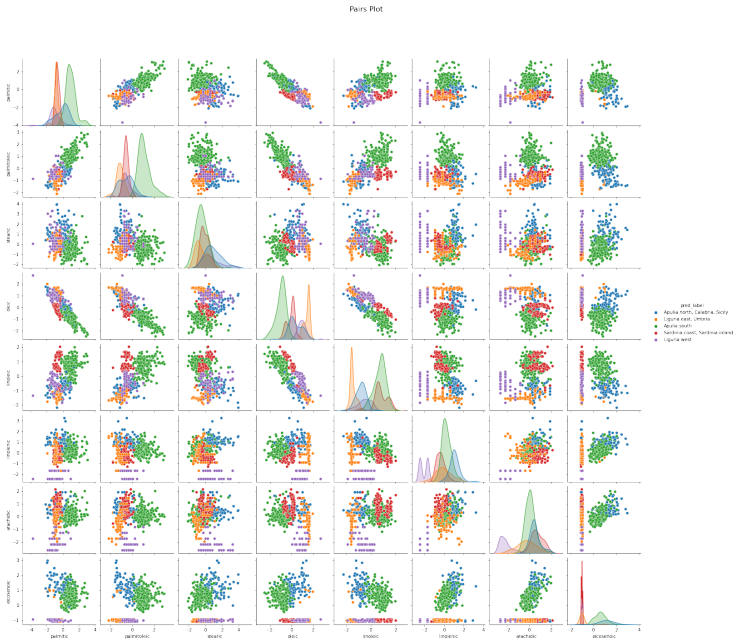


This clusters understood from the point of the regional map are clustered within their macro areas, except for West Liguria and South Apulia forming their own clusters. These clusters might be an indication that these regions within the macro area have very different growing conditions.

8 PCA Analysis on Five-Cluster Model

To further investigate the influence of fatty acid components on the separation of the clusters on scaled data, we visualize the pairwise relationships between 8 fatty acid components with the predicted label from the K-means model. The pairs plot below displays the combination of fatty acid components, the distribution of the components along the diagonals, and the resulting separation among the five clusters in the off-diagonals.

For example, examine the subplot between linoleic and eicosenoic acid components. The green cluster [South Apulia] is higher in eicosenoic measurements, whereas the blue cluster [North Apulia, South Apulia, Calabria, and Sicily] has a lower concentration of linoleic acid and a slightly higher concentration of eicosenoic acid; this causes a separation between the two clusters. In the same subplot, we cannot separate the clusters [Sardinia coast, Sardinia inland], [East Liguria, Umbria], and [West Liguria] if we only look at their eicosenoic concentration but we can see the differences in linoleic concentration.



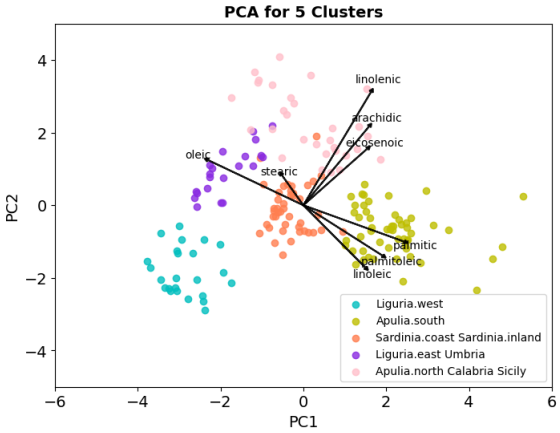
The pairs plot demonstrates the limitations of visualizing 2 fatty acid components at a time. It is important to consider all 8 fatty acid components to better capture the separation among the 5 clusters.

Therefore, we use the Principal Component Analysis (PCA) method to reduce the dimension of the data while preserving as much of the data's variation as possible. In the PCA for 5 clusters plot below, there are 8 vectors representing the influence of 8 fatty acid components on the data in two dimensions. The direction of the vectors convey the correlation between each acid component and the 2 calculated principle components. The length of the vectors tells us how strong the correlation is.

Meaning, the vector for each acid component points in the direction of a region for increasing values. For example, in the PCA plot for 5 clusters, the vector associated with oleic acid component points toward low values of the first principle component (PC1), so we know that the lower the value of PC1, the higher the oleic acid concentration is. Using the same method, we can see that the higher the value of PC1, the higher the palmitic, palmitoleic, and linoleic acid concentrations are. If we look at the second component (PC2) in the plot, the higher its values get, the higher the linolenic, arachidic, eicosenoic and stearic acid concentrations generally are.

From the position of the clusters in this plot, we can see that the South Apulia region approximately has a higher concentration of palmitic, palmitoleic, and linoleic acid than the other regions. We can also see that all 3 clusters [North Apulia, Calabria, Sicily], [East Liguria, Umbria], and [West Liguria] has a higher concentration of oleic acid than the other regions. One of the biggest differences between them is the linolenic acid concentration. The cluster [North Apulia,

Calabria, Sicily] has the highest concentration, and the cluster [East Liguria, Umbria] does not contain as much of the concentration. We can also see that the cluster [West Liguria] has the loWest concentration of linolenic acid.



9 Conclusion: Possible Directions with PCA

From the visualization, the oleic acid component appears to be one of the most prominent fatty acids in this olive oil sample. It influences the separation between the clusters the most in our analysis. Therefore, we wanted to know more about this oleic acid.

García-Inza et al.[6] studied the effect of temperature on olive oil accumulation. He found that the oleic acid concentration decreased linearly when the temperature got higher from 16 to 32 Celsius. Nissim et al.[7] stated in their research that olive oil accumulation procedure is known to begin in the second half of the summer in Israel.

An interesting fact we found during research is that the summer of Liguria and Umbria, the regions that have higher concentrations of oleic acid, have a cooler average temperature than the other regions in the data set. While we do not have any direct proof that the temperature of the regions in Italy has an effect on the fatty acid components of the olive oil, we suspect that there is a connection between the temperature, or maybe other environmental elements, in each region and the chemical components in the examined olive oil.

Therefore, if we have more data on the olive oil observations such as temperature or the time in the year the oil accumulation procedure takes place, or other possible environmental impact like soil and water, we can further investigate the relationship between the regions and the fatty acid component in olive oil.

[6] [Gar+14]
[7] [Nis+20]

References

[ST02] H. Sadeghi and A.R. Talaii. "IMPACT OF ENVIRONMENTAL CONDITIONS ON FATTY ACIDS COMBINATION OF OLIVE OIL IN AN IRANIAN OLIVE, CV. ZARD". In: *Acta Hortic* (2002), pp. 579–581. DOI: 10.17660/ActaHortic.2002.586.121.

[Ric05] Vito Ricci. "Fitting distributions with r". In: *Contributed Documentation available on CRAN* (2005).

[Gar+14] G. Garcia-Inza et al. "Responses to temperature of fruit dry weight, oil concentration, and oil fatty acid composition in olive (Olea europaea L. var. 'Arauco')". In: *European Journal of Agronomy* 54 (Mar. 2014), pp. 107–115. DOI: 10.1016/j.eja.2013.12.005.

[Nis+20] Yael Nissim et al. "High temperature environment reduces olive oil yield and quality". In: *PLOS ONE* 15 (Apr. 2020), pp. 1–24. DOI: 10.1371/journal.pone.0231956. URL: https://doi.org/10.1371/journal.pone.0231956.