Projektarbete i Mutltivariat analys

Cluster Analysis

Max Johansson, Michael Debebe, Oliver Rykatkin

Statistiska Instutionen

Samhällsvetenskapliga fakulteten

Uppsala universitet

2022-10-18

Contents

Abstract

In this paper we study if dry bean types can be clustered into meaningful subgroups based on measures of dimension and shape. The method we use is k-means clustering which seeks to split the observations in mind to k number of clusters by choosing the "true" value of the clusters in favor of the current numerical criterion. Our results indicate that clustering of dry beans may not be completely inappropriate. The difference between the clusters have been found among the dimensional variables, while the clusters have been quite alike considering shape variables.

1 Introduction

The common bean or Phaseolus vulgaris includes many well known bean types and is divided into green and dry beans based on the technique for growing them (FAO UN). For context, the produced quantity of dry beans (27,5 million) is higher than that of green beans (23,3 million) based on data from 2020 (FAOSTAT UN, see appendix 1). However, due to constraints of our data the focus of this paper is on the dry bean.

The dry bean is argued to be a sustainable source of nutrition, as well as being subject to serious university research (The U.S. Sustainability Alliance). Moreover, in combination with broader trends such as more plant oriented diets across several parts of the world in recent years (World Preservation Foundation), the relevancy of the dry bean has perhaps increased even further. In a study which collected data consisting of pictures and measurements of dry beans in order to classify the bean species, it is argued that the dry bean is the most important and produced legume in the world (Koklu and Ilkan, 2020).

All together, it appears that new insights with regards to dry beans could be of potential value. We aim to contribute to existing and novel research with knowledge, which also could be of interest to producers of dry beans. We do this by examining both dimensional and shape measures of dry bean types and thus concluding if there are any meaningful ways of grouping dry bean species. We argue that this has the potential to be useful descriptive information for studies as well as for producers of dry beans. The research question we are posing and are seeking to answer is:

`"Can any meaningful subgroups among dry bean types be identified?"`

2 Data

The data material is taken from UCLA's database and consists of 13611 observations where each row corresponds to an observed bean. This is the same data which was mentioned in the introduction chapter, in the paper by Koklu and Ilker (2020). The data material consists of 17 columns where the first 16 are numerical and the last is categorical. Within the framework of this work, only the variables that are not ratio measures of other variables and estimates of measures (the shape factor variables) will be used.

Dimensional Variables

1.) Area (A): The area of a bean zone and the number of pixels within its boundaries.

2.) Perimeter (P): Bean circumference is defined as the length of its border.

3.) Major axis length (L): The distance between the ends of the longest line that can be drawn from a bean.

4.) Minor axis length (l): The longest line that can be drawn from the bean while standing perpendicular to the main axis.

11.) Roundness (R): Calculated with the following formula: (4piA)/(P^2)

Shape Variables

6.) Eccentricity (Ec): Eccentricity of the ellipse having the same moments as the region.

9.) Extent (Ex): The ratio of the pixels in the bounding box to the bean area.

10.) Solidity (S): Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.

Categorical Variables

17.) Class: (Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira), which specie the dry bean belongs to.

Summary of data

Table 1: Summary of data

	Specie	Obs.
1	BARBUNYA	1322
2	BOMBAY	522
3	CALI	1630
4	DERMASON	3546
5	HOROZ	1928
6	SEKER	2027
7	SIRA	2636

Looking at table 1, one can conclude that the data consists of different amounts of each bean species. However, this should not be a problem for the cluster analysis as the sample size of each bean species is somewhat large.

3 Method

Distance Matrix

Distance Matrix is a matrix that shows the distance between pairs of objects. The distance matrix stores all the distances between all the observations using a $n \times n$. The diagonal entries in the matrix are always 0. This is because the distance between A and B is always the same as the distance between B and A. Hence the matrix is symmetrical. The distance measurement used for this report is Squared Euclidean distance.

Euclidean Square Distance

There are different ways to determine the within-cluster variation. There is the Euclidean distance, Mahalanobis distance and so forth. The most common choice when using the k-means method is the squared Euclidean distance. Mathematically the squared Euclidean distance is described as following:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Where $|C_k|$ is the number of observations in the th cluster. Therefore the within-cluster variation of the th cluster is the sum of all the pairwise squared Euclidean distance between the observations in the th cluster. The square distance is pretty much the same thing as the Standardized euclidean distance, the only difference is that the standardized don't take the square root.

K-means Method

This report is using the k-means clustering method for creating natural clusters of the observed data. Like other clustering methods the goal of the k-means method is to find similar observations, not to try to pair together variables There are two different properties that need to be fulfilled to be a proper K-means clustering. These are:

1. Each observation is in at least one cluster. That is $C_1 \cup C_2 \cup \ldots \cup C_k = \{1, \ldots, n\}$.
2. None of the clusters are overlapping. An observation cannot belong to more than one cluster. That is for all: $C_k \cap C_{k'} = \varnothing$ for all $k \neq k'$.

The k-means method is to minimize the within-cluster variation. Mathematically it is described as:

$$\text{Minimize}_{C_1,\ldots,C_K} = \min \sum_{k=1}^{K} W(k)$$

Where $W(k)$ is the within-cluster variation for each cluster. The measurement used for within-cluster variation differs. This report will use Euclidean square distance to define the within-cluster variation.

To determine how many different clusters to choose the report will use the following procedure.

1. Choose the number of initial clusters, k.
2. Reassign several observations so that the total variability within the cluster is smaller.
3. Repeat this procedure until the total variability increases instead of decreasing.

In the first step this report will have a hierarchy clustering approach to determine how many clusters to have initially. This means looking at a dendrogram of the data and finding big differences between the different branches created by the dendrogram.

Evaluation

To evaluate the k-means clusters this report will use the silhouette function introduced by Rousseeuw (1987). The method compares the closeness of an observation to other observations within its own cluster, as compared to observations in the closest other cluster. The method gives a standardized value between -1 and 1, the value is called Average silhouette widths (ACW). Where it compares the average distance of the observations within a cluster and the distance for each observation to the center of the cluster.

Kaufman & Rousseeuw (2005, p.88) has given the following interpretation of the ACW:

Table 2: Interpretation of AWC

ACW	Interpreation
0.71-1.00	A strong structure has been found
0.51-0.71	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial
=<0.25	No substantial structure has been found

4 Result

Correlation Matrix

Table 3: Correlation matrix over included variables

	Area	Perimeter	MajorALeng	MinorALeng	Eccent	Extent	Solidity	Roundness
Area	1							
Perimeter	0.97	1						
MajorALeng	0.93	0.98	1					
MinorALeng	0.95	0.91	0.83	1				
Eccent	0.27	0.39	0.54	0.02	1			
Extent	0.05	-0.02	-0.08	0.15	-0.32	1		
Solidity	-0.2	-0.3	-0.28	-0.16	-0.3	0.19	1	
Roundness	-0.36	-0.55	-0.6	-0.21	-0.72	0.34	0.61	1

Looking at the correlation matrix of the variables which are included in this paper, one can observe low, moderate and high correlations. High correlation between input variables is called collinearity and could be problematic for regression (James et al., 2013, p.99). However since it is not regression which is conducted, and since clustering is based on similar or dissimilar attributes variables we do not expected the existing collinearity to affect the clustering results.

Dendrogram

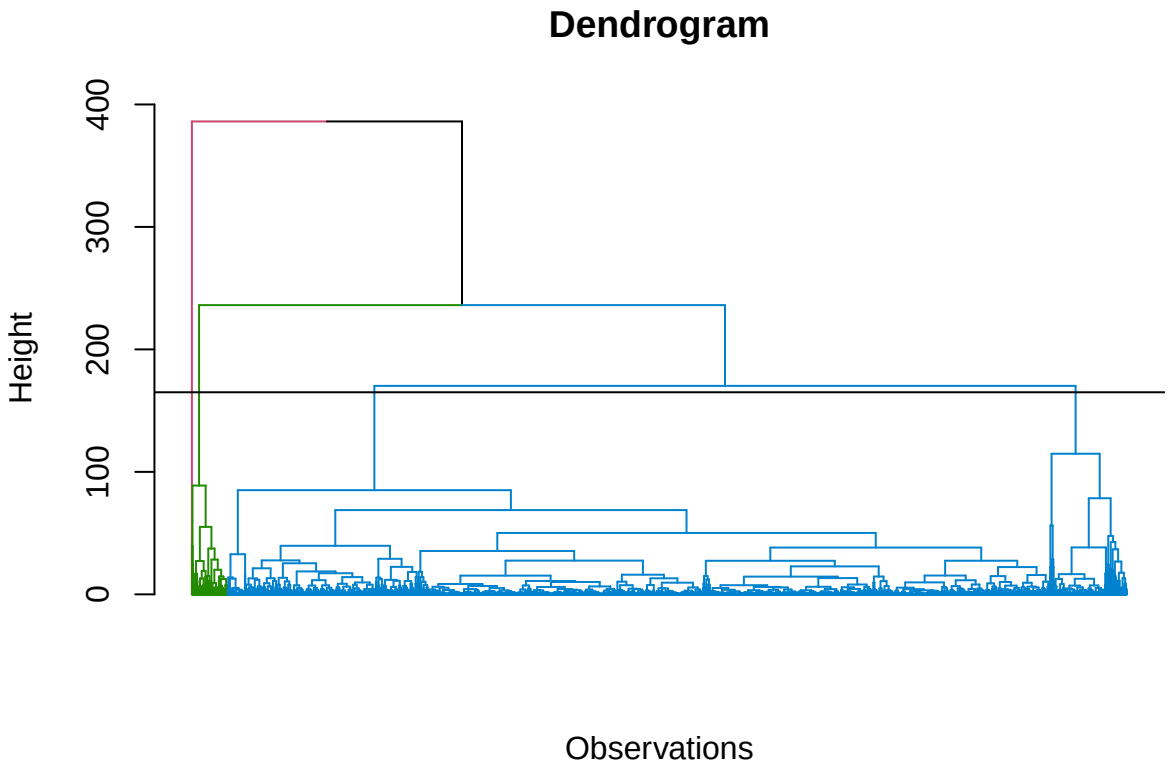


Figure 1: Dendrogram

From figure 1 above who shows the dendrogram, one can draw the conclusion that a initial k-value of 3 fits the data set.

Table 4: Cluster divison

Cluster	1	2	3
BARBUNYA	1282	39	1
BOMBAY	0	0	522
CALI	1611	17	2
DERMASON	49	3497	0
HOROZ	1889	39	0
SEKER	18	2009	0
SIRA	448	2188	0

From table 4 one can see that the clusters are not perfectly divided among the bean species, yet a majority of every bean species has been divided into only one cluster.

Describing statistics between clusters

Table 5: Describing statistics between clusters

Cluster	Area	Perimeter	MajorAleng	MinorAleng	Eccentricity	Extent	Solidity	roundness
1	63840	987.08	378.89	214.50	0.81810	0.73267	0.98449	0.81476
2	37614	716.52	261.82	182.39	0.70386	0.75954	0.98896	0.91369
3	173156	1584.15	592.59	373.94	0.77056	0.77648	0.98688	0.86420

According to the results presented in table 5, there seems to be some differences in the mean of variables between clusters which are more significant than others. The most apparent differences could be observed in the variables area, perimeter, major axis length and minor axis length which are all dimensional variables. Less apparent differences can be observed between the clusters when in regards to the shape variables of eccentricity, extent, solidity but also one dimensional variable which is roundness.

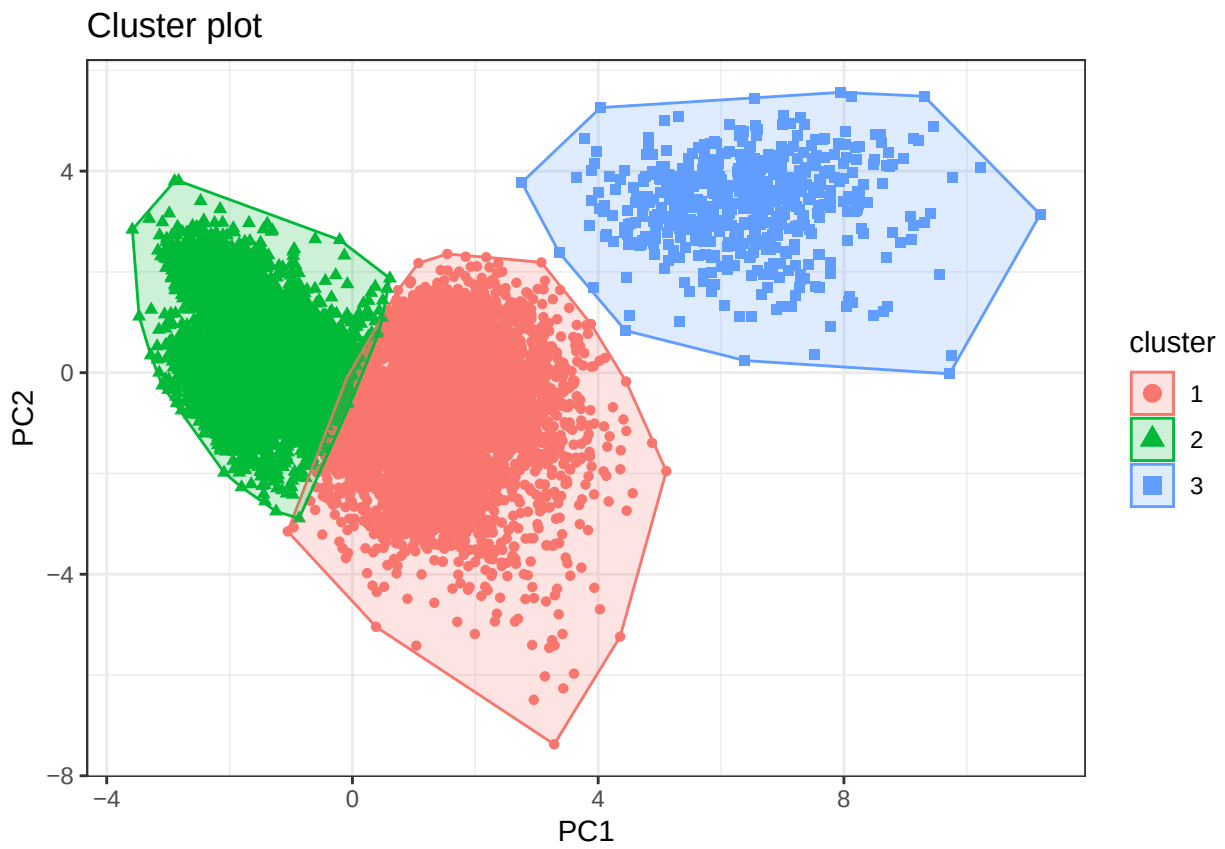


Figure 2: Clusterplot showing division among clusters and their loadings

An ocular analysis of the cluster plot in figure 2 shows that there is only really one actual separation between the clusters. While cluster 1 and 2 are not really separated from each other by any space, cluster 3 is separated from the other clusters. Moreover, a few observations from cluster 1 can be identified as overlapping into cluster 2.

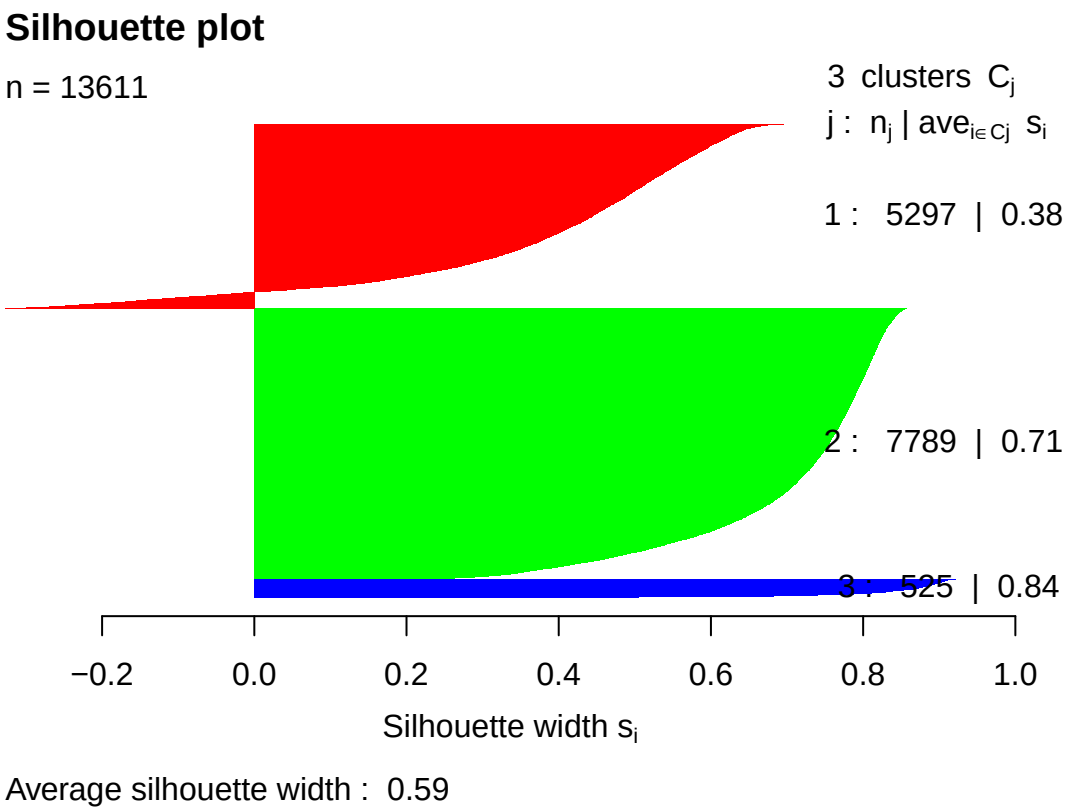


Figure 3: Silhouette diagram for 3 clusters

In figure 3 above, the average silhouette width is plotted, from which one can see that both cluster 2 and 3 get relatively high scores meanwhile cluster 1 get a low score, this indicates that observations within cluster 1 are poorly matched to their cluster which can be seen in the figure. However the overall average silhouette width accumulates to 0.59.

5 Conclusion

Taken as a whole, the results presented above indicate that clustering is reasonable. According to the average silhouette widths, the average silhouette width for the model indicating that three clusters should be used returns a ACW value of 0.59 which indicates a reasonable structure.

The cluster plot can be interpreted as some support for existing subgroups. From table 4 we know that the majority of observations of the species Barbunya, Cali and Horoz are divided into cluster 1. In a similar fashion the majority of Dermason, Seker and Sira observations are divided into cluster 2. The main difference between cluster 1 and 2 is that they differ in perimeter and area dimensions, according to table 2. From table 4 we learn that cluster 3 is made up almost entirely by the Bombay bean which does not appear in any other cluster. This cluster or more specifically the Bombay bean is characterized by its relatively large dimensions.

6 Bibliography

Books

James, G., Witten, D., Hastie, T., Tibshirani, R. (Eds.), 2013. An introduction to statistical learning: with applications in R, Springer texts in statistics. Springer, New York

Kaufman, L. & Rousseeuw, P.J. (2005). Finding groups in data: An introduction to cluster analysis. John Wiley & Sons, Inc. Hoboken, NJ.

Digital

Archive.ics.uci.edu. (n.d.). UCI Machine Learning Repository: Dry Bean Dataset Data Set. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset.

Food and Agriculture Organization of the United Nations. Land and Water - Crop information - Bean. [WWW Document], n.d. URL https://www.fao.org/land-water/databases-and-software/crop-information/bean/en/ (accessed 10.11.22)

Food and Agriculture Organization of the United Nations. FAOSTAT - Crops and livestock products. [WWW Document], n.d. URL https://www.fao.org/faostat/en/#data/QCL (accessed 10.11.22)

Koklu, M. and Ilker A.O. (2020). Multiclass Classification of Dry Beans using Computer Vision and Machine Learning Techniques. Computers and Electronics in Agriculture, Volume 174. https://doi.org/10.1016/j.compag.2020.105507

The U.S. Sustainability Alliance. U.S. Dry Beans Fact Sheet. [WWW Document], n.d. URL https://thesustainabilityalliance.us/u-s-dry-beans-fact-sheet/ (accessed 10.11.22).

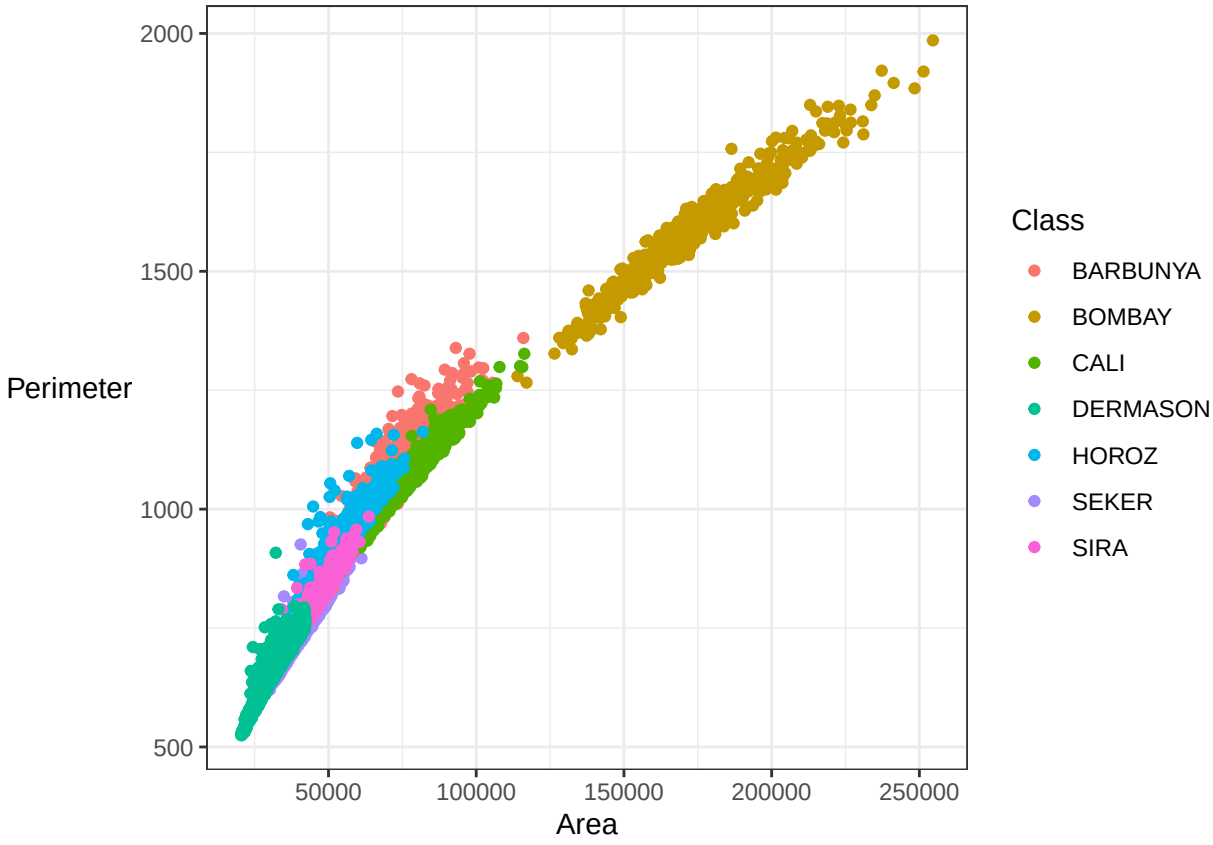
World Preservation Foundation. Vegetarian Vegans Growth. [WWW Document], n.d. URL https://worldpreservationfoundation.org/business/vegetarian-vegans-growth/ (accessed 10.11.22).

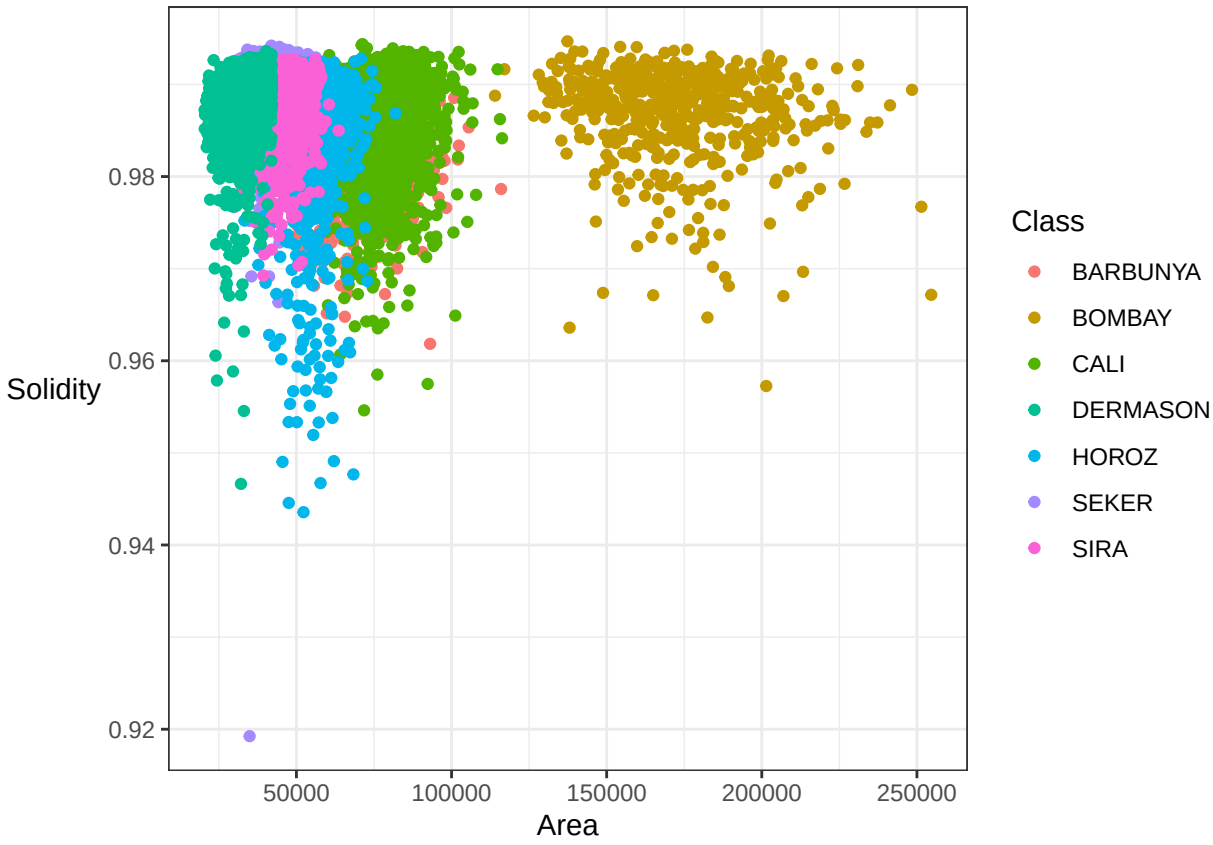
Appendix

Production data (2020) of beans dry and other beans green. Food and Agriculture Organization of the United Nations - FAOSTAT Input: Regions: World + (Total) Elements: Production Quantity Items: “Beans, dry”, “Other beans, green”. Years: 2020

Domain Code	Domain	Area Code (M49)	Area	Element Code	Element	Item Code (CPC)	Item	Year Code	Year	Unit	Value	Flag	Flag Description
QCL	Crops and livestock products	001	World	5510	Production	01701	Beans, dry	2020	2020	tonnes	27545942	E	Estimated value
QCL	Crops and livestock products	001	World	5510	Production	01241.90	Other beans, green	2020	2020	tonnes	23276716	E	Estimated value

7 Attachments






```
kk <-aggregate(df,list(df$Class),mean)
kk <-kk[,1:9]
knitr::kable(kk)
```

Group.1	Area	Perimeter	MajorAxisLength	MinorAxisLength	Eccentricity	Extent	Solidity	roundness
BARBUNYA	53980.04	1046.11	370.04	240.31	0.75466	0.74927	0.98280	0.80020
BOMBAY	173485	1585.62	593.15	374.35	0.77052	0.77656	0.98690	0.86442
CALI	75538	1057.63	409.50	236.37	0.81480	0.75895	0.98502	0.84593
DERMASON	32019	665.21	246.56	165.66	0.73663	0.75295	0.98823	0.90811
HOROZ	53649	919.86	372.57	184.17	0.86744	0.70639	0.98548	0.79442
SEKER	39881	727.67	251.29	201.91	0.58478	0.77167	0.99035	0.94451
SIRA	44729	796.42	299.38	190.80	0.76728	0.74944	0.98797	0.88465

Table 7: Describing statitiscs among species

	Specie	Area	Perimeter	MajorALength	MinorALength	Eccent	Extent	Solidity
1	BARBUNYA	69804.13	1046.11	370.04	240.31	0.75	0.75	0.98
2	BOMBAY	173485.06	1585.62	593.15	374.35	0.77	0.78	0.99
3	CALI	75538.21	1057.63	409.50	236.37	0.81	0.76	0.99
4	DERMASON	32118.71	665.21	246.56	165.66	0.74	0.75	0.99
5	HOROZ	53648.51	919.86	372.57	184.17	0.87	0.71	0.99
6	SEKER	39881.30	727.67	251.29	201.91	0.58	0.77	0.99
7	SIRA	44729.13	796.42	299.38	190.80	0.77	0.75	0.99