# Assignment 5 - K-Means Clustering Analysis of the MAGIC Gamma Telescope Data

Tanushree Kumar

DATA 630 - Summer 2024

Professor Ami Gates

University of Maryland Global Campus

Due: July 30, 2024

## Introduction

The objective of this analysis is to perform a k-means clustering analysis on the MAGIC Gamma Telescope dataset to classify the different high-energy particles of gamma and hadron rays. Specifically, the aim is to answer the following question: Can gamma-ray events be effectively classified using k-means clustering to identify natural groupings in the data without prior labels? This analysis will help in understanding the characteristics of gamma events in comparison to hadron events, potentially enhancing the classification accuracy in high-energy physics research.

MAGIC (Major Atmospheric Gamma Imaging Cherenkov) is a system of two imaging atmospheric Cherenkov telescopes located at the Roque de los Muchachos Observatory on La Palma, Canary Islands (*MAGIC (Telescope)*, 2024). These telescopes detect gamma rays with photon energies indirectly by observing the Cherenkov radiation produced when gamma rays interact with the Earth's atmosphere. The interaction creates a cascade of secondary particles that travel through the atmosphere, producing a faint bluish light known as Cherenkov radiation. These gamma rays originate from some of the most extreme and violent events in the universe, such as supernova explosions, pulsars, black hole accretion, and active galactic nuclei, giving astronomers key insight into the lives of these celestial objects (*MAGIC | Max Planck Institut Für Physik*, 2021).

The MAGIC Gamma Telescope dataset contains observations of events recorded by the telescopes. Each event represents a potential gamma-ray interaction characterized by various features such as the length, width, and size of the event, concentration measures, asymmetry, and moments of the event. These features help differentiate between gamma rays (signal) and hadronic showers (background noise). The dataset's primary goal is to classify these events

accurately, distinguishing between genuine gamma-ray events, class 'g', and hadron events, class 'h'. This classification is crucial for further astronomical analysis and for improving the sensitivity and accuracy of gamma-ray observations.

K-means clustering is an unsupervised machine learning technique that partitions a dataset into K distinct clusters based on the features of the data. This method works in a way to group things such that the items in each group are similar but different from items in other groups. K-means clustering is chosen for this analysis due to its efficiency and simplicity in grouping similar data points. As an unsupervised learning algorithm, K-means does not require labeled data, making it suitable for exploring the intrinsic structures of the dataset. This method helps in identifying clusters of gamma and hadron events based on their feature similarities, providing insights into the natural grouping within the data. Furthermore, K-means clustering can assist in anomaly detection by highlighting data points that do not fit well into any cluster, potentially indicating unusual or interesting events.

## Analysis

The MAGIC Gamma Telescope dataset is sourced from the UCI Machine Learning Repository. The dataset contains 19,020 events with 11 features each, including physical and geometric characteristics of the detected events derived from the radiation images. Each instance represents a record of gamma-ray events observed by the MAGIC telescope. The key variables are:

- fLength: length of the gamma event; major axis of an ellipse, measured in millimeters
- fWidth: width of the gamma event; minor axis of an ellipse, measured in millimeters

- fSize: size of the gamma event; 10-log of the sum of the content of all pixels in the image, measured in the number of photons
- fConc: concentration of the event; ratio of the sum of the two highest pixels over fSize
- fConc1: first concentration measure; ratio of highest pixel over fSize
- fAsym: asymmetry of the event; distance from the highest pixel to center, projected onto the major axis, measured in millimeters
- fM3Long: long moment of the event; third root of the third moment along the major axis, measured in millimeters
- fM3Trans: the transverse moment of the event; third root of the third moment along the minor axis
- fAlpha: alpha angle of the event; angle of major axis with vector to the origin, measured in degrees
- fDist: distance measure of the event; distance from origin to center of the ellipse, measured in millimeters
- class: class of the event; gamma (signal), hadron (background) (Bock, 2007).

To understand the meaning of these features, it is important to note that the "image of a shower is an elongated cluster" (Bock, 2007), which is why there are variables denoting the length and width of the ellipse and the gamma event.

It should also be noted that the dataset contains an underestimated number of 'h' events due to technical reasons and real data has a majority of 'h' class events. The UC website also mentions in additional information that "classifying a background event (hadron) as signal (gamma) is worse than classifying a signal event as background" (Bock, 2007). Misclassification

of events can have different implications depending on whether the event is a signal or background. In gamma-ray astronomy, detecting true gamma-ray events (signals) from a large number of cosmic-ray events (background) is crucial. Accurate detection allows researchers to study cosmic sources of gamma rays, contributing to our understanding of the universe. Background events are generally far more numerous than signal events, thus, even a small false positive rate can lead to a large number of false signals, overshadowing true events. Classifying a background event as a signal is worse in this context because it can lead to significant resource wastage, incorrect scientific conclusions, and operational inefficiencies. Whereas, classifying a signal event as background, while it results in missed discoveries, maintains the overall integrity and reliability of the data, ensuring that the detections made are of high confidence.

Before building the model, exploratory data analysis (EDA) was performed to understand the distribution and relationships of the variables. Based on the EDA, data preprocessing was performed such as handling missing values, removing the target column, and scaling the variables. Using functions such as summary(), str(), and sum(is.na()) gave an overview of the dataset and showed that no missing values were present. The dataset description website also stated there were no missing values for any columns. All the variables were numerical, meaning little data preprocessing had to be performed.

The summary function provided a statistical summary for each variable in the dataset. The dataset contains 19,020 observations and 11 variables including the target variable, which is a factor with two levels of ‘g’ (gamma) and ‘h’ (hadron). The summary showed the data contains outliers, as seen in variables like ‘fAsym’, which ranges from -457.916 to 575.241. The presence of extreme values in other variables such as ‘fM3Long’ and ‘fM3Trans’ suggests a diverse set of events. Given the variability and range of the variables, outlier detection was necessary through

box plots. Figure 1 below shows a snippet of the box plots that display the outliers for the first three variables. The full figure can be seen below in Appendix A. Through these box plots, it can be observed that there are outliers present in the following features: ‘fLength’, ‘fWidth’, ‘fSize’, ‘fConc1’, ‘fAsym’, ‘fM3Long’, ‘fM3Trans’ and ‘fDist’.

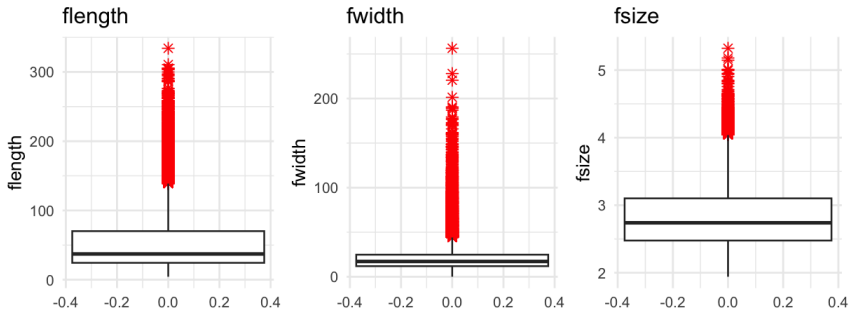


Figure 1. *Figure of the boxplots to detect outliers.*

The outliers in ‘fLength’, ‘fWidth’, and ‘fSize’ might indicate extreme cases where the detected particles’ image is different in size and/or in shape in comparison to the rest of the instances. The outliers in ‘fConc1’ might indicate instances where the image might be compact and dense or very dispersed. The outliers in ‘fAsym’ might indicate unusually asymmetric images. The outliers in ‘fM3Long’ and ‘fM3Trans’ also suggest extreme cases where the length and transformation of the image are different from the average value. The outliers in ‘fDist’ suggest the source of the image is either very close or much farther than the average distance.

To combat the wide ranges, a capping method was used to cap the extreme values within the specified percentile of 5th and 95th. Capping was preferred compared to other methods as this method is less aggressive than removal and can preserve data integrity while minimizing the impact of outliers. This ensured that the number of instances was still the same with the ‘class’ variable having nearly double the proportion of gamma events with 12,332 gamma instances (g) compared to the 6,688 hadron instances (h).

K-means clustering partitions the data into K clusters by minimizing the within-cluster variance. Each observation belongs to the cluster with the nearest mean. The algorithm works iteratively to initialize centroids randomly; assign each data point to the nearest cluster center/centroid; recalculate the cluster center based on the current cluster members; and repeat this process until convergence. Key input parameters include the number of clusters, K, and the initialization method for centroids. The objective function minimized by k-means is $J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$, where $\mu_i$ is the centroid of the cluster $C_i$ (Dabbura, 2018). To fit the k-means model, an appropriate number of clusters, K, is selected first. The Elbow method is used to determine the optimal K value by plotting the total within-cluster sum of squares (WSS) against the number of clusters.

To fit the model, the Elbow method was used to determine the optimal number of clusters. The Elbow method plot showed that the optimal 'k' values at k = 4, 8,14. The first model was built using 4 clusters, with exploration models being built with 8 and 11 clusters. A seed value was set which ensured the results would be reproducible when the model was run again. The components of the cluster were examined and a plot for 4 clusters was generated. To see how more clusters would perform, plots were generated for k = 8 and k= 14. Next, the sum of squared distances between clusters as the k value increases was plotted by using k between 2 and 15. Distances between each instance and the corresponding cluster were calculated to detect anomalies. The instances with the largest distances from the cluster center are the outliers.

## Results

For the clustering output with k = 8, the sizes of the clusters are as follows: Cluster 1 has 5932 instances, Cluster 2 has 7560 instances, Cluster 3 has 4004 instances, and Cluster 4 has

1524 instances. This distribution shows that Cluster 2 has the most instances, while Cluster 4 has the least.

The cluster centers represent the mean values of the variables for each cluster. Cluster 1 exhibits lower values for ‘fLength’, ‘fWidth’, and ‘fSize’, high values for ‘fAsym’ and ‘fConc1’, near-zero values for ‘fAsym’ and ‘fM3Ttrans’, moderate positive ‘fAlpha’, and negative ‘fAsym’. Cluster 2 shows slightly negative values for most features, a slightly positive value for ‘fAsym’, and near-zero values for ‘fM3Ttrans’. Cluster 3 has high positive values for ‘fLength’, ‘fWidth’, ‘fSize’, and ‘fM3Long’, negative values for ‘fConc’ and ‘fConc1’, slightly negative values for ‘fAsym’, and positive ‘fAsym’. Cluster 4 is characterized by very high positive values for flength, fwidth, and ‘fSize’, negative values for ‘fAsym’ and fconc1, very negative values for ‘fAsym’ and ‘fM3Long’, and positive ‘fAsym’.

The sum of squares analysis provides additional insights into the clustering. The Total Sum of Squares (totss) is 190190, representing the total variance in the dataset. The Within Cluster Sum of Squares (withinss) for the clusters are 19908.69, 29932.47, 31929.53, and 15542.51, representing the variance within each cluster. The Total Within Sum of Squares (tot.withinss) is 97313.19, which is the sum of within-cluster variances. The Between Cluster Sum of Squares (betweenss) is 92876.81, indicating the variance between the different clusters. The ratio of betweenss to totss is approximately 48.8%, indicating that about 48.8% of the variance is explained by the clustering.

The cluster-to-class evaluation shows the distribution of actual classes within each cluster. Cluster 2 contains a significant number of both ‘g’ (5643) and ‘h’ (1917) events, making it the largest and most mixed cluster. Cluster 1 is predominantly composed of ‘g’ events. Cluster 4, though the smallest, has a notable presence of ‘h’ events.

The dataset consists of features derived from a telescope's observations, classifying events into gamma rays and hadron events. The clustering analysis aimed to separate these events into distinct groups based on their features. Cluster 1 predominantly consists of gamma-ray events with high-concentration features. Cluster 2 is a mix of both gamma rays and hadron events, representing events with slightly negative and neutral values across most features. Cluster 3 is predominantly hadron events, characterized by high values for size-related features and low-concentration features. Cluster 4, though small, has a notable number of hadron events and is characterized by very high-size features and very low similarity and length-related features.

The objective was to use clustering to identify distinct groups of telescope events and understand their characteristics. By analyzing the cluster centers and sizes, it is evident that the clustering has effectively grouped events with similar characteristics, revealing underlying patterns in the data. The clustering results show distinct groups with meaningful differences in their feature values. Cluster 2 is the largest and most mixed, indicating it captures a diverse range of events. Clusters 1 and 3 highlight groups with high concentrations of gamma-ray and hadron events, respectively. The sum of squares analysis shows a reasonable variance explained by the clustering. These insights help in understanding the nature of the events captured by the telescope and can guide further analysis or classification tasks.

For the clustering output with k = 8, Cluster 1 contains 2000 instances, Cluster 2 contains 3381 instances, Cluster 3 has 2319 instances, Cluster 4 has 1384 instances, Cluster 5 has 2165 instances, Cluster 6 has 1774 instances, Cluster 7 has 1854 instances, and Cluster 8 has 4143 instances. This distribution indicates that Cluster 8 is the largest, while Cluster 4 is the smallest.

The breakdown of the 8 clusters is as follows: Cluster 1: Predominantly hadron events with 1265 hadrons and 735 gamma rays. Cluster 2: Predominantly gamma rays with 2911 gamma rays and 470 hadrons. Cluster 3: Predominantly gamma rays with 1648 gamma rays and 671 hadrons. Cluster 4: Predominantly hadron events with 1232 hadrons and 152 gamma rays. Cluster 5: Predominantly gamma rays with 1139 gamma rays and 1026 hadrons. Cluster 6: Predominantly gamma rays with 1220 gamma rays and 554 hadrons. Cluster 7: Predominantly gamma rays with 1229 gamma rays and 625 hadrons. Cluster 8: Predominantly gamma rays with 3298 gamma rays and 845 hadrons. The cluster plot shows how the clusters are distributed in the reduced two-dimensional space which can be seen in Figure 6 in Appendix A.

For k = 8, the clustering results indicate that gamma-ray events (g) and hadron events (h) are grouped into distinct clusters, with some clusters showing a mix of both types of events. This suggests that the clustering is capturing variations within each class of events, providing a finer granularity of the dataset's structure.

For the clustering output with k = 14, the sizes of the clusters vary, with some clusters being quite small and others larger. This increased number of clusters allows for a more detailed separation of the data. The breakdown of the 14 clusters is as follows: Cluster 1: Predominantly gamma rays with 1389 gamma rays and 210 hadrons. Cluster 2: Predominantly gamma rays with 603 gamma rays and 273 hadrons. Cluster 3: Mixed with 535 gamma rays and 702 hadrons. Cluster 4: Predominantly gamma rays with 1698 gamma rays and 319 hadrons. Cluster 5: Mixed with 464 gamma rays and 313 hadrons. Cluster 6: Predominantly hadron events with 561 hadrons and 25 gamma rays. Cluster 7: Predominantly hadron events with 1015 hadrons and 478 gamma rays. Cluster 8: Mixed with 32 gamma rays and 526 hadrons. Cluster 9: Predominantly gamma rays with 1646 gamma rays and 323 hadrons. Cluster 10: Mixed with 854 gamma rays

and 586 hadrons. Cluster 11: Mixed with 117 gamma rays and 354 hadrons. Cluster 12: Mixed with 605 gamma rays and 332 hadrons. Cluster 13: Predominantly gamma rays with 1340 gamma rays and 554 hadrons. Cluster 14: Predominantly gamma rays with 2546 gamma rays and 620 hadrons. The cluster plot for k = 14 shows a more complex distribution of the clusters in the reduced two-dimensional space which can be seen in Figure 8 in Appendix A.

For k = 14, the clustering results provide an even more detailed view of the dataset's structure. The increased number of clusters allows for capturing finer variations within each class of events. Some clusters are predominantly one type of event, while others are mixed, reflecting the inherent diversity in the dataset. The clustering results for k = 8 and k = 14 provide distinct groups with meaningful differences in their feature values. Cluster sizes and compositions vary, with some clusters predominantly containing gamma-ray events and others containing hadron events.

The sum of squared distances within clusters and the sum of squared distances between clusters was plotted which can be seen in Figure 9 in Appendix A. It can be seen that the sum of squared distances within the cluster drops as the k value is increasing and the sum of squared distances between clusters increases as k increases. There is an inflection point at k = 4 could be the ideal k value.

## Conclusion

The objective of this analysis was to classify high-energy particles detected by the MAGIC Gamma Telescope into gamma-ray and hadron events. The key findings show that by examining the features of these events, distinct groups can be identified. The clustering analysis revealed natural groupings within the data, successfully highlighting differences between

gamma-ray and hadron events. This helps in understanding the characteristics of these high-energy particles, providing valuable insights for further research in high-energy physics and astronomy.

While the analysis provided meaningful results, it faced several limitations. The dataset's inherent imbalance, with fewer hadron events, may have affected the accuracy of the classification. Additionally, the complexity of the features and their interactions may not have been fully captured, potentially leading to some misclassifications. The clusters formed may not be perfect representations of the true nature of the events due to these limitations in the data and the methods used.

To enhance the analysis, future work could focus on addressing the data imbalance by collecting more hadron event samples or using techniques to balance the data. Incorporating additional features or using more advanced methods could also improve the classification accuracy. Additionally, collaborating with experts in high-energy physics could provide deeper insights into the data, leading to more refined and accurate classifications.

This analysis demonstrates that gamma-ray and hadron events can be effectively grouped based on their characteristics, providing a clearer understanding of these high-energy particles. This classification is crucial for researchers studying cosmic phenomena, as it helps distinguish between genuine gamma-ray signals and background noise. By improving the accuracy and reliability of these classifications, astronomers can gain better insights into the universe's most extreme events, ultimately advancing the knowledge of cosmic sources and their behaviors.

## References


Bock, R. (2007, April 30). *MAGIC Gamma Telescope*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/159/magic+gamma+telescope

Dabbura, I. (2018, September 17). *K-means clustering: Algorithm, applications, evaluation methods, and drawbacks*. Medium; Towards Data Science. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

*MAGIC (telescope)*. (2024, May 16). Wikipedia. https://en.wikipedia.org/wiki/MAGIC_(telescope)

*MAGIC | Max Planck Institut für Physik*. (2021). Max Planck Institut Für Physik. https://www.mpp.mpg.de/en/research/magic-and-cta/magic

## Appendix A

All figures and visualizations mentioned in the report can be seen below. The R code is attached as a separate file.

**Figure 1**

*Figure of the boxplots to detect outliers.*

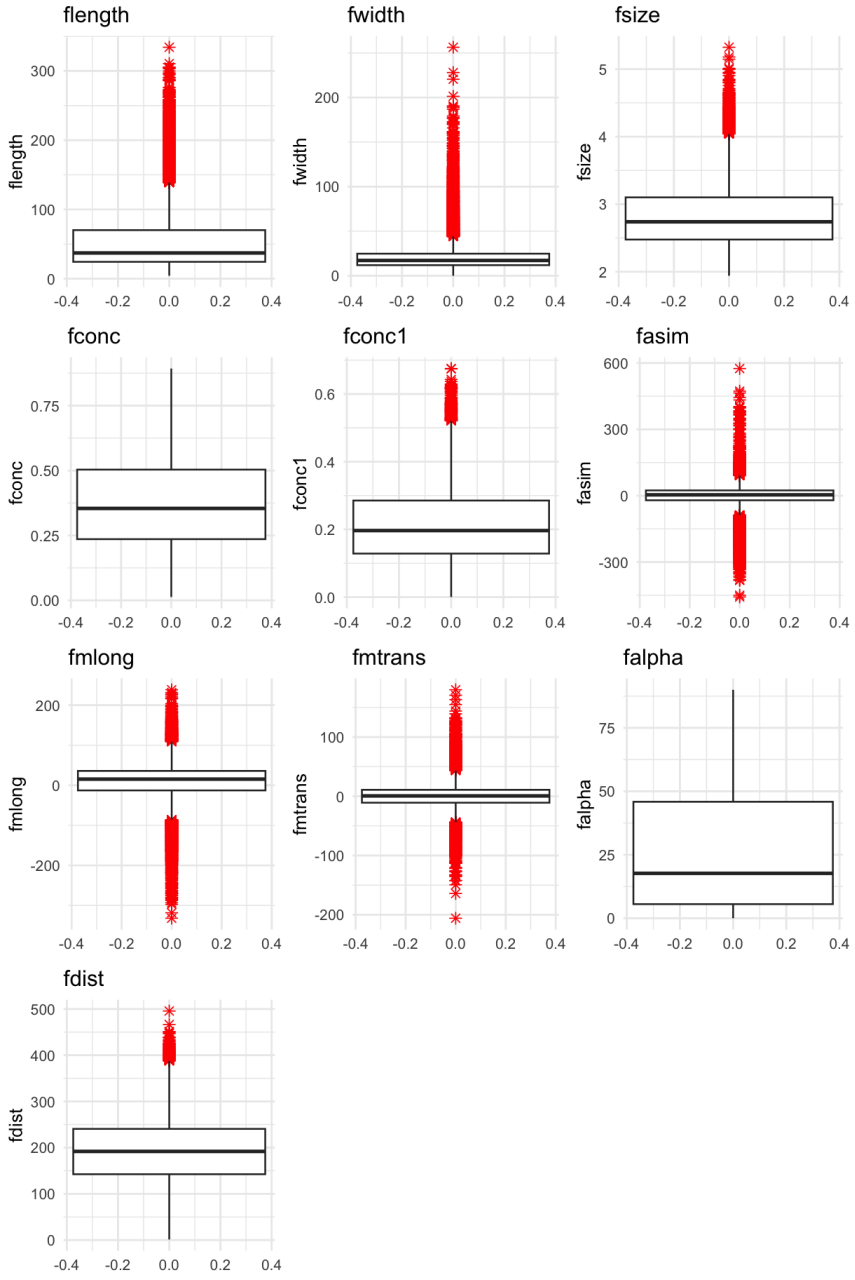

**Figure 2**

*Figure of the Elbow Method to determine the optimal number of clusters.*

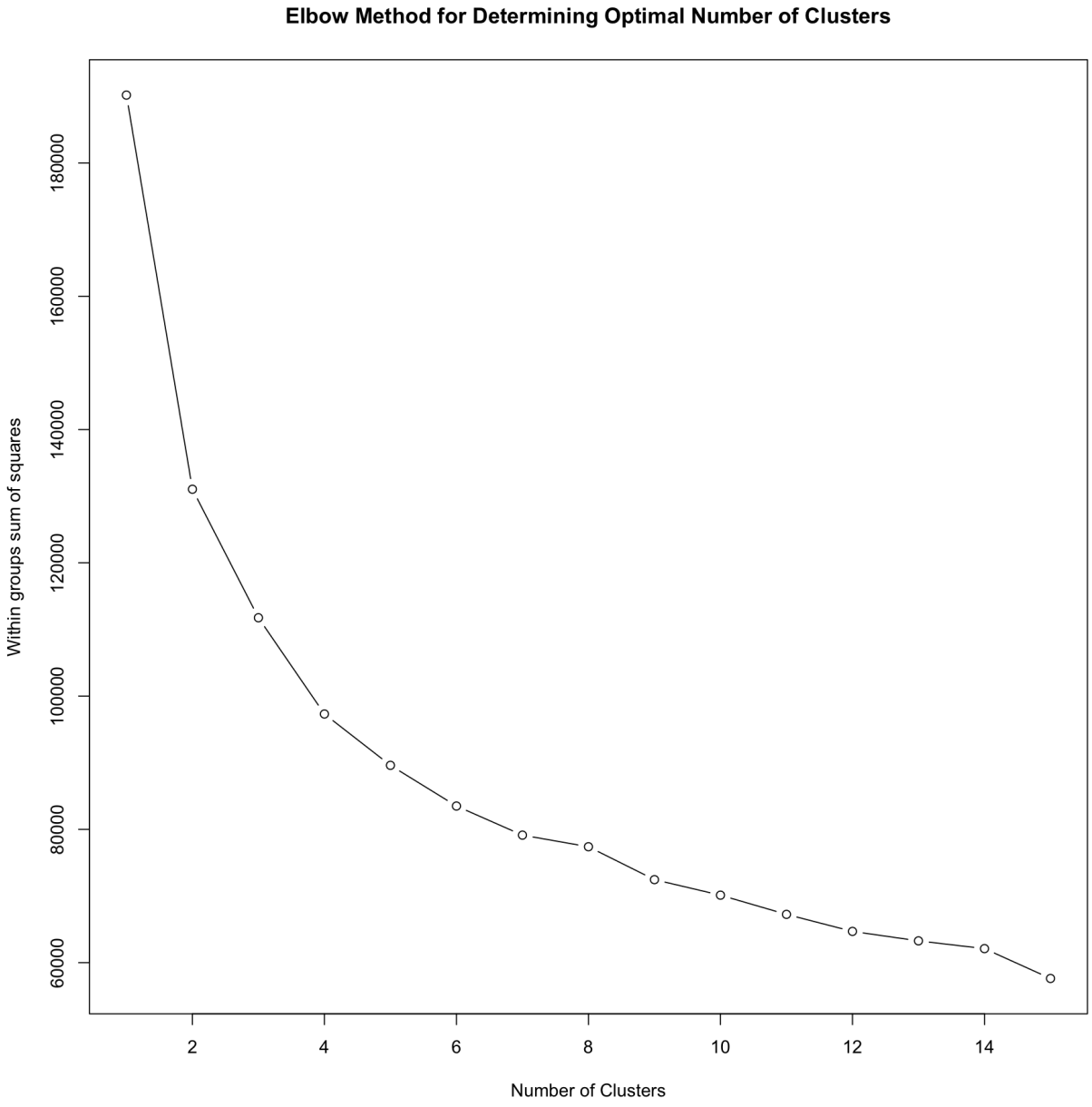

**Figure 3**

*Figure of the cluster components for the value k = 4.*

```
> # Cluster assignment for each instance
> tele_cap_kc$cluster
   [1] 2 1 4 1 3 2 2 2 3 2 3 1 2 1 2 2 2 2 2 2 3 3 3 1 3 3 2 2 1 2 2 3 2 2 3 3 2 2 2 3 3 1 2 1 3 1 2 1 3 2 2 2 2 2 2 3 1 3 3 1 3 2 2 1 3 2 1 2 2 2 1 3 2 1 2
  [76] 2 1 3 3 2 2 3 1 1 1 3 1 2 2 2 3 3 3 2 1 2 1 2 1 2 4 2 2 2 2 3 2 2 3 1 2 1 2 2 1 1 2 3 2 1 1 1 3 3 2 2 2 3 3 2 2 3 1 1 2 2 2 2 1 1 1 1 1 4 1 2 1 1 1 2
 [151] 2 3 1 2 2 3 3 1 3 1 1 1 3 2 2 2 3 1 3 3 1 1 2 2 1 1 2 3 1 1 3 1 2 2 2 2 1 2 2 3 3 3 1 2 3 1 2 2 3 3 2 2 2 3 1 3 2 2 2 2 1 1 2 2 2 2 3 2 2 1 3 2 3 1
 [226] 3 2 3 1 1 3 2 2 3 2 1 3 2 2 2 3 4 1 1 2 3 2 2 3 3 3 2 1 3 1 2 1 3 2 1 1 2 2 4 2 1 3 1 2 1 1 3 2 3 2 2 2 3 1 2 2 2 2 2 1 1 2 2 3 2 1 3 2 2 2 3 2 1 2 2
 [301] 2 3 2 2 2 2 2 1 2 2 1 3 1 1 2 1 2 1 1 2 3 1 2 3 2 3 1 1 3 2 1 1 2 3 3 2 1 2 2 2 3 3 2 2 3 2 1 2 4 2 2 3 2 1 3 2 1 2 1 2 2 2 3 1 2 2 2 1 2 1 2 3 3 1 2
 [376] 2 1 1 2 1 2 1 2 1 2 2 1 2 2 3 1 3 2 2 2 2 1 3 3 3 3 2 1 2 2 2 2 2 2 2 2 3 2 3 2 2 2 3 2 1 2 2 2 2 2 1 2 2 3 3 2 3 2 2 2 1 2 2 2 3 2 3 1 2 2 2 2 2 3
 [451] 1 1 1 2 3 1 3 3 2 1 1 1 2 1 3 3 1 1 4 2 1 2 2 2 3 2 1 3 3 3 1 3 1 1 2 2 3 1 2 2 1 1 2 2 2 1 2 2 3 2 1 1 3 3 2 2 2 1 1 2 1 2 2 3 2 2 1 1 3 2 1 3 1 2 1
 [526] 3 2 1 1 2 1 3 1 1 2 3 2 3 2 2 3 2 1 2 2 2 2 1 1 3 1 1 2 1 2 3 2 2 2 2 1 3 3 3 1 2 1 1 2 2 2 2 2 2 2 4 3 3 3 2 2 1 1 1 3 2 2 3 2 1 2 2 1 1 3 1 1 1 1 2
 [601] 1 2 3 1 2 1 3 2 2 3 2 2 3 2 1 3 1 2 2 2 1 2 2 2 3 2 1 2 2 2 1 2 3 2 2 3 2 3 2 2 1 2 2 2 2 1 1 2 1 2 2 1 3 2 2 1 1 1 2 3 1 2 3 2 3 2 2 2 1 3 2 1 2 1 1
 [676] 2 2 2 2 1 1 2 2 1 1 3 3 1 2 3 2 3 2 2 2 3 3 2 2 1 2 3 1 1 3 1 1 2 2 3 4 3 2 2 1 4 2 2 1 3 2 2 1 1 1 2 2 1 2 3 2 2 1 3 1 2 3 1 2 2 2 3 1 1 1 2 2 2 2 2
 [751] 3 1 2 2 2 2 2 1 3 2 2 1 1 2 2 3 2 3 2 1 1 1 3 3 2 3 2 2 3 3 2 1 3 1 2 3 2 3 2 2 2 2 1 2 2 1 4 2 2 3 1 1 1 2 4 3 2 2 2 3 3 4 2 2 4 2 3 1 3 3 2 1 2 2 2
 [826] 2 2 2 2 1 3 1 2 3 1 2 1 3 2 3 3 2 2 1 1 3 2 2 3 2 1 2 1 2 2 3 1 2 3 2 2 2 3 2 3 2 2 2 3 1 2 1 2 2 2 2 2 2 2 1 2 2 1 2 3 2 1 3 3 1 3 1 2 1 1 2 2 1 1 2
 [901] 2 2 1 1 3 3 2 3 2 3 3 2 2 2 1 2 1 2 1 2 1 2 3 2 2 3 2 2 1 3 2 3 1 3 1 2 2 1 1 2 3 2 2 1 3 4 2 3 1 2 2 1 1 2 1 2 1 2 2 2 2 2 3 2 3 1 2 1 2 1 1 3 2 3 1 1
 [976] 3 2 2 2 2 1 3 2 1 1 2 2 3 2 1 3 4 3 1 3 1 1 2 2 2
 [ reached getOption("max.print") -- omitted 18020 entries ]
> # Cluster centers matrix.
> tele_cap_kc$centers
     flength     fwidth       fsize      fconc     fconc1       fasim      fmlong      fmtrans     falpha      fdist
1 -0.8040482 -0.7471032 -0.97881261  1.2020467  1.1928760  0.12042800 -0.23323626 -0.004821731  0.5121257 -0.4517093
2 -0.3025471 -0.2396107 -0.09170569 -0.2329033 -0.2509142  0.24282417 -0.07836283 -0.018872134 -0.2389510 -0.2591913
3  1.0434113  0.8452669  1.16471928 -0.9895374 -0.9610678 -0.09070459  1.25603330  0.019104642 -0.4588238  0.8079729
4  1.8891414  1.8758692  1.20477392 -0.9236772 -0.8734341 -1.43500555 -2.00339689  0.062192160  0.3974217  0.9211957
> # Sum of squared distances within cluster+sum of squared distances between other clusters
> tele_cap_kc$totss
[1] 190190
> # Sum of squared distances within cluster by cluster
> tele_cap_kc$withinss
[1] 19908.69 29932.47 31929.53 15542.51
> # Total sum of squared distances within cluster.
> tele_cap_kc$tot.withinss
[1] 97313.19
> # Sum of squared distances between instances and other cluster centers
> tele_cap_kc$betweenss
[1] 92876.81
> # Number of instances in a cluster
> tele_cap_kc$size
[1] 5932 7560 4004 1524
> # Number of iterations the method took to run
> tele_cap_kc$iter
[1] 4
> # Reruns 0 if the method ran without problems
> tele_cap_kc$ifault
[1] 0
> # Cluster to class evaluation
> table(tele$class, tele_cap_kc$cluster)

      1    2    3    4
  g 3671 5643 2815  203
  h 2261 1917 1189 1321
```

**Figure 4**

*Figure of the cluster plot for value k = 4.*

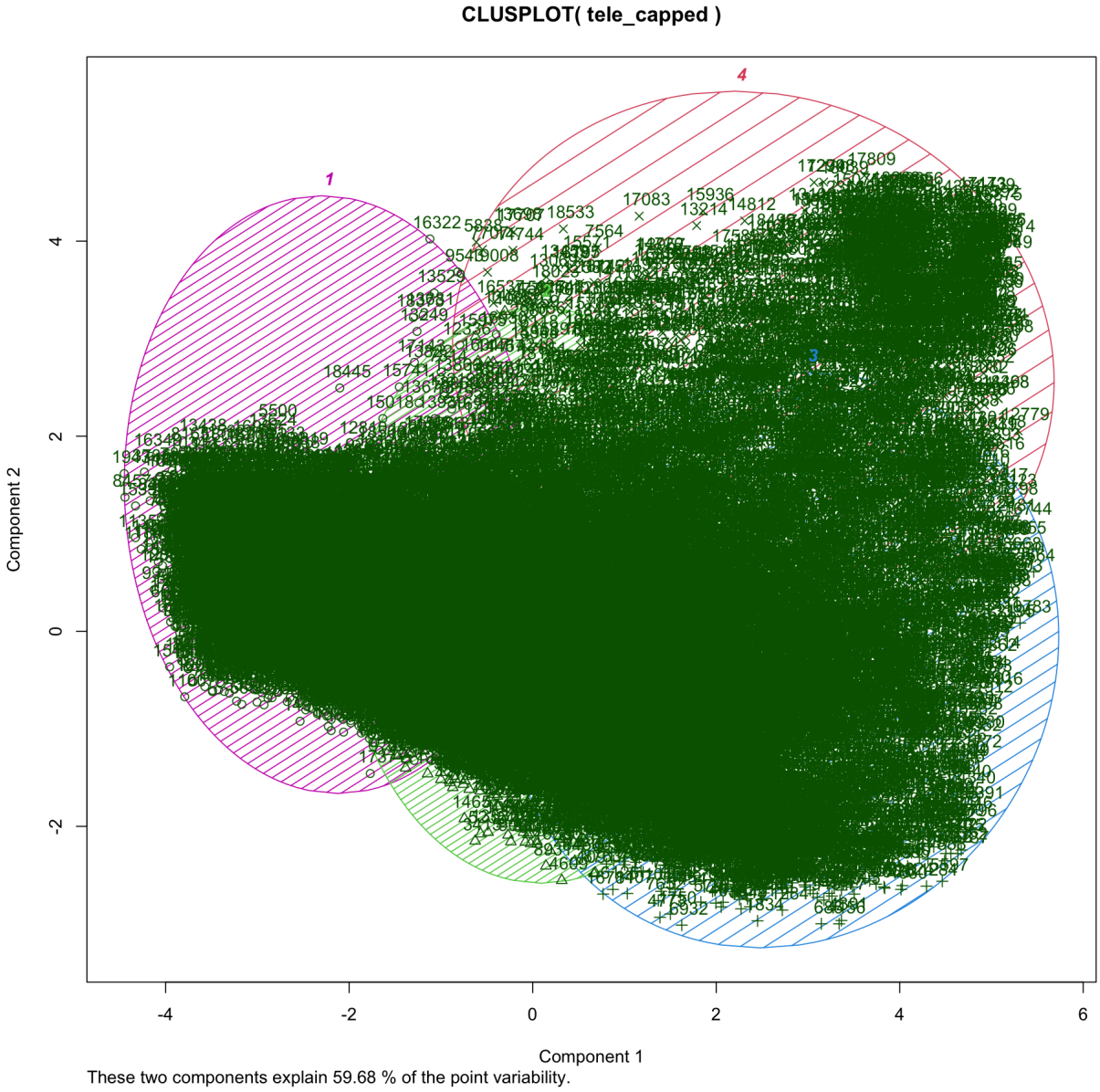

**Figure 5**

*Figure of the cluster table for value k = 8.*

```
> # Trying k = 8
> tele_cap_kc <- kmeans(tele_capped, 8)
> table(tele$class, tele_cap_kc$cluster, dnn=c('Class in the dataset', 'Cluster mumber'))
                    Cluster mumber
Class in the dataset    1    2    3    4    5    6    7    8
                   g  735 2911 1648  152 1139 1220 1229 3298
                   h 1265  470  671 1232 1026  554  625  845
```

**Figure 6**

*Figure of the cluster plot for value k = 8.*

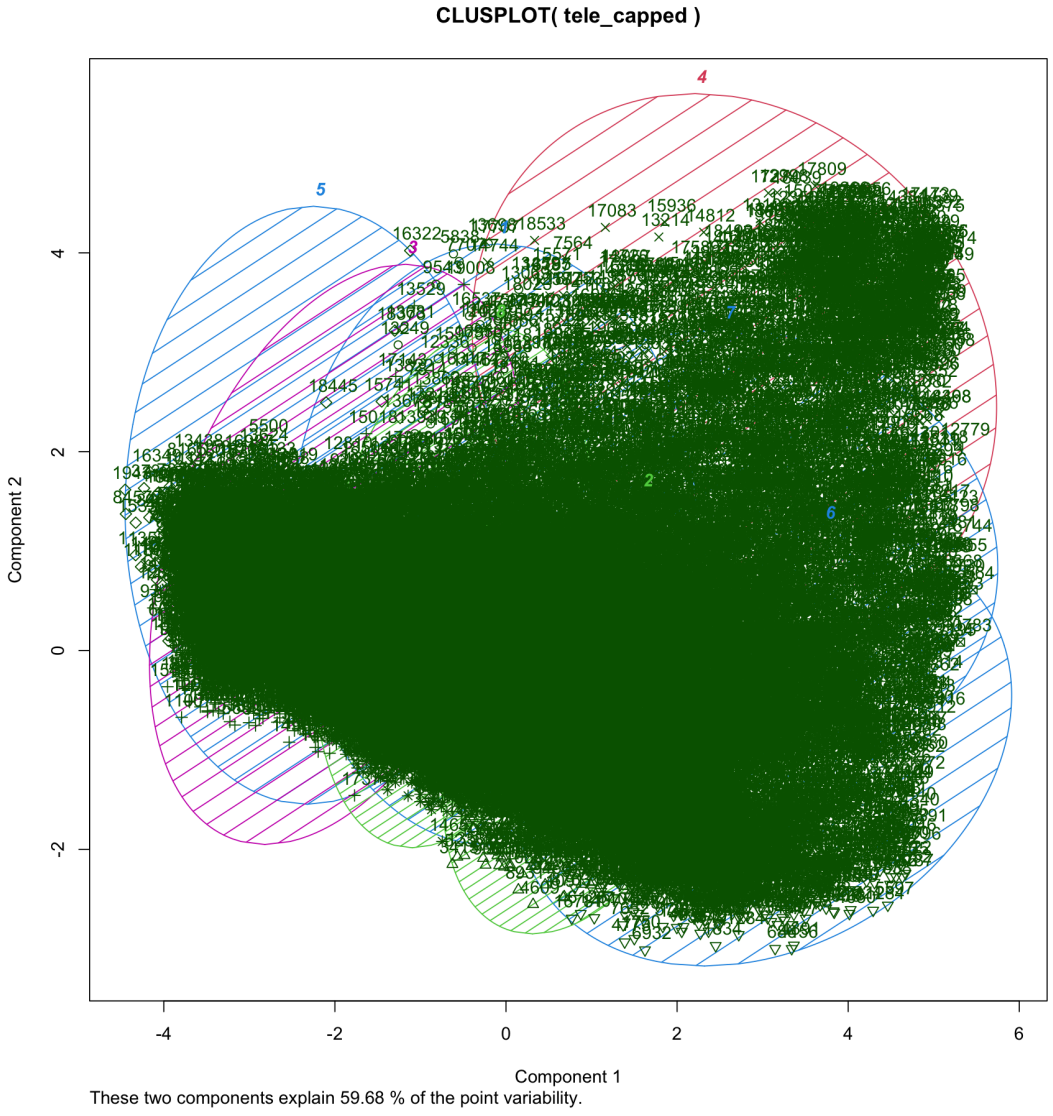

**Figure 7**

*Figure of the cluster table for value k = 14.*

```
> # Trying k = 14
> tele_cap_kc <- kmeans(tele_capped, 14)
> table(tele$class, tele_cap_kc$cluster, dnn=c('Class in the dataset', 'Cluster mumber'))
                   Cluster mumber
Class in the dataset    1    2    3    4    5    6    7    8    9   10   11   12   13   14
                   g 1389  603  535 1698  464   25  478   32 1646  854  117  605 1340 2546
                   h  210  273  702  319  313  561 1015  526  323  586  354  332  554  620
```

**Figure 8**

*Figure of the cluster plot for value k = 14.*

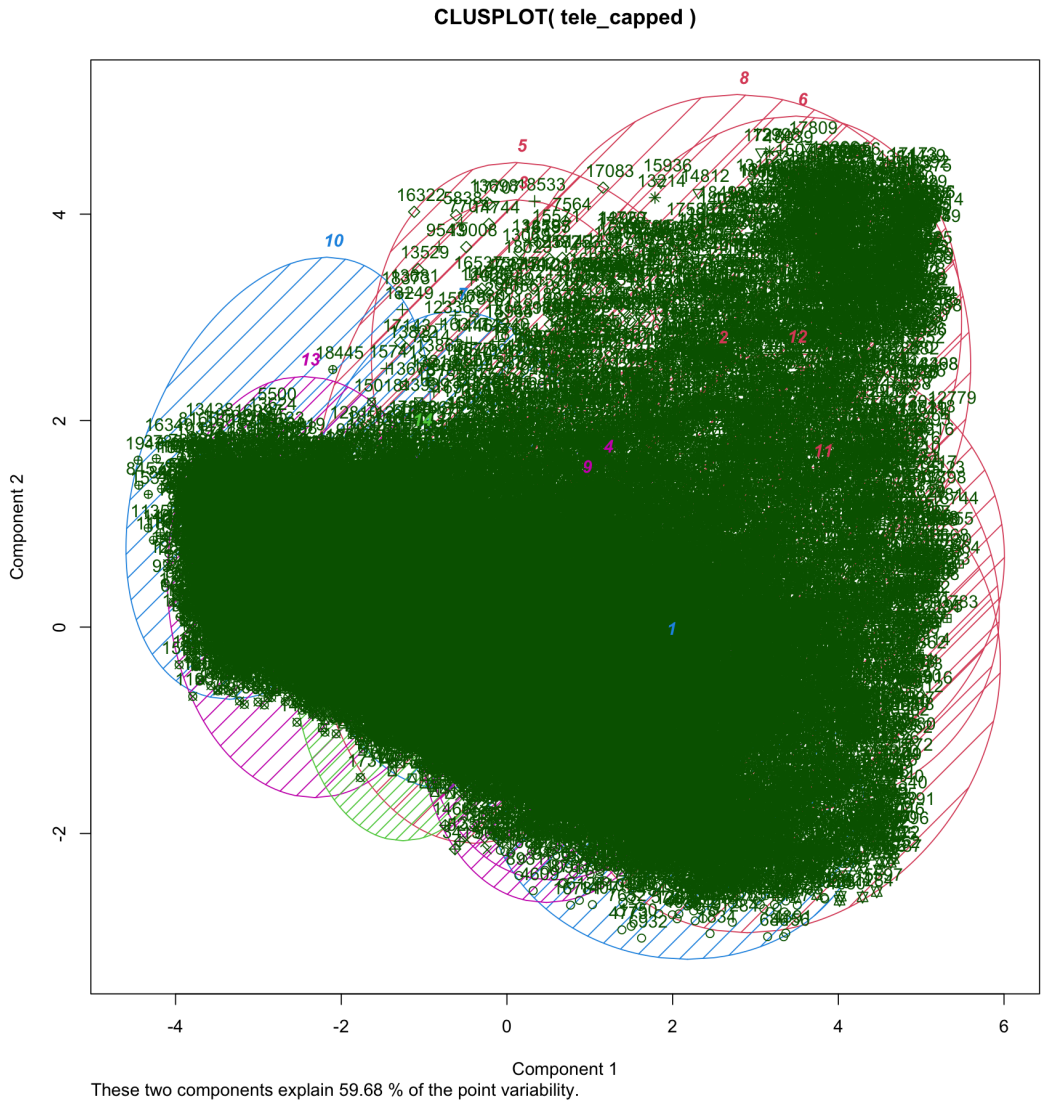

**Figure 9**

*Figure of the plot the sum of squared distances within clusters and the sum of squared distances between clusters.*

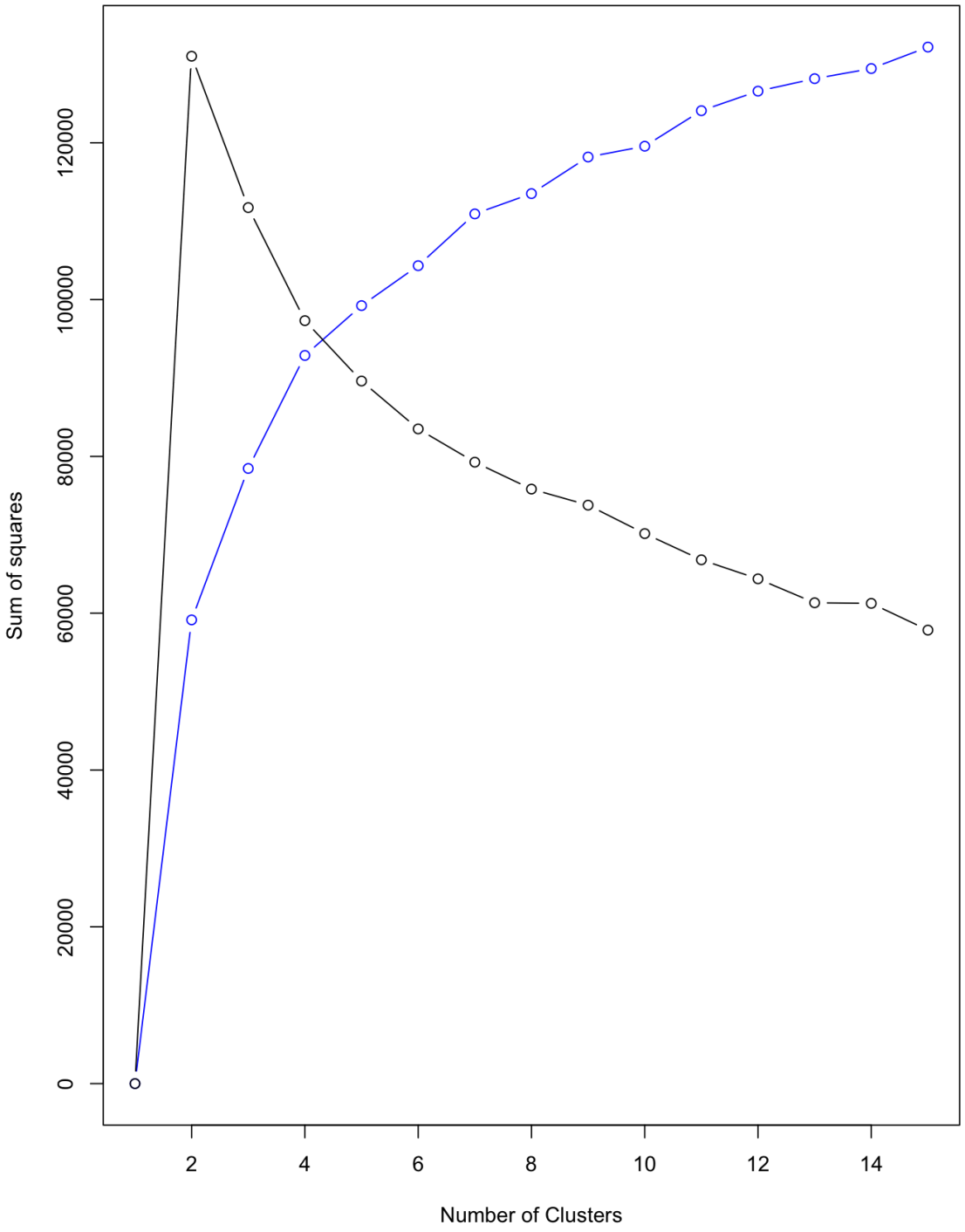

**Figure 10**

*Figure of the anomaly detection.*

```
> # Anomaly detection
> centers <- tele_cap_kc$centers[tele_cap_kc$cluster, ]
> head(centers, 15)
      flength      fwidth      fsize      fconc      fconc1       fasim      fmlong      fmtrans     falpha      fdist
3  -0.2218018 -0.18545829 -0.0769452 -0.2493505 -0.27801489  0.06314244 -0.30208799 -0.123864673  1.3284172 -1.2362611
13 -0.8757026 -0.85338170 -1.1352155  1.5934555  1.61245624  0.14725308 -0.21497465  0.004546763 -0.2781504 -0.3343640
8   2.0732419  2.31265117  1.4744064 -1.1278855 -1.07952699 -1.43346179 -2.08962389 -1.805737530  0.5438379  0.9285876
13 -0.8757026 -0.85338170 -1.1352155  1.5934555  1.61245624  0.14725308 -0.21497465  0.004546763 -0.2781504 -0.3343640
12  1.1657074  0.91997937  1.0711276 -0.9036670 -0.84976515 -1.49992355  1.20147906  1.337312557 -0.4760116  1.1380085
4  -0.2474000 -0.09131979  0.1549955 -0.4947576 -0.50776472  0.40647488  0.01717815  0.810831726 -0.5637259 -0.1702033
4  -0.2474000 -0.09131979  0.1549955 -0.4947576 -0.50776472  0.40647488  0.01717815  0.810831726 -0.5637259 -0.1702033
14 -0.6487201 -0.56007148 -0.6602143  0.4380704  0.40357143  0.17966348 -0.18386154 -0.011353087 -0.4622011 -0.4140968
11  1.7867753  2.40750135  1.5093185 -1.3088550 -1.28128399  1.13228629  1.63011441 -0.175896749  0.6508514  0.3523649
9  -0.2201064 -0.09082717  0.1846520 -0.5027038 -0.51714621  0.42373222  0.11570809 -0.836553564 -0.5492056 -0.1796340
12  1.1657074  0.91997937  1.0711276 -0.9036670 -0.84976515 -1.49992355  1.20147906  1.337312557 -0.4760116  1.1380085
14 -0.6487201 -0.56007148 -0.6602143  0.4380704  0.40357143  0.17966348 -0.18386154 -0.011353087 -0.4622011 -0.4140968
9  -0.2201064 -0.09082717  0.1846520 -0.5027038 -0.51714621  0.42373222  0.11570809 -0.836553564 -0.5492056 -0.1796340
3  -0.2218018 -0.18545829 -0.0769452 -0.2493505 -0.27801489  0.06314244 -0.30208799 -0.123864673  1.3284172 -1.2362611
5   0.7826424  0.07419231  0.1532586 -0.1253880 -0.05625432 -1.23556654 -1.21758839 -0.120042443 -0.5355533  1.0986359
```

**Figure 11**

*Figure of the outliers.*

```
> outliers <- order(distances, decreasing=T)[1:5]
> outliers
[1] 18293 13459 17942 12582 18931
> tele[outliers,]
       flength  fwidth  fsize  fconc fconc1     fasim    fmlong  fmtrans  falpha    fdist class
18293 184.1744 70.3681 3.2656 0.0991 0.0563 -378.9457  180.7741 -42.7278 89.5442  39.5465     h
13459 144.6510 40.3478 2.9380 0.6182 0.4988  -74.6539 -155.4890 -42.0003 38.0104 280.0510     h
17942 141.0382  8.6551 2.4930 0.5991 0.3821 -184.3575  115.2470 -10.1714 10.7012 160.0322     h
12582 119.6410 39.7995 2.8701 0.6069 0.4525  -70.4334 -114.2610 -43.2190 36.5720 381.6780     h
18931  80.5290 55.3783 3.6765 0.0763 0.0332  116.7082  -56.4538 -42.6596 41.8947  40.8251     h
```

**Figure 11**

*Figure of the scatterplot of the outliers.*