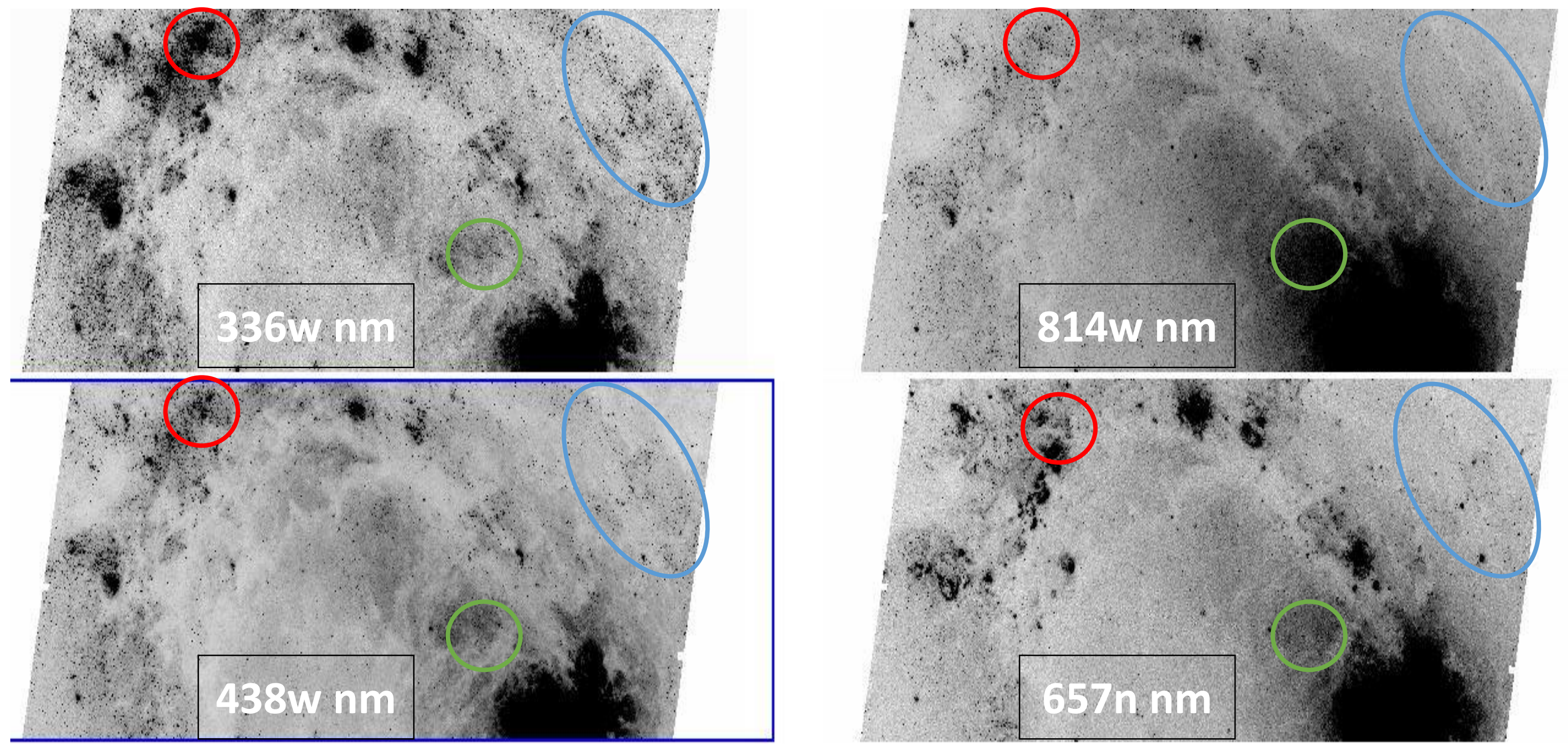
Clustering the M83 Galaxy

Alexander K. Kiar, Pauline Barmby
Department of Physics and Astronomy, Western University

Abstract

Space-based astronomical observatories generate vast quantities of data, and efficient means of analyzing those data are needed. The purpose of this research is to use machine-learning methods to classify point sources of light emission in nearby galaxies. An object's light emission over different wavelengths is the key data for classification as it indicates the composition of the object, along with its other physical attributes. Mean-shift and k-means clustering methods were applied to observations of point sources in the M83 galaxy, to identify objects that emit similar combinations of light over multiple wavelengths. The data was collected by the Wide Field Camera 3 on the Hubble Space Telescope. The strength of the clustering was tested using a silhouette score to identify which bands best separated different classes of objects. This metric measures an object's distance from a cluster outside the one it was originally assigned to. The clustering results were also compared with the results of independent classification, to determine if each object was correctly identified. The results of this work will allow astronomers to plan observations that can be used to automatically classify objects in nearby galaxies, leading to a stronger understanding of how stars and star clusters form, and evolve.

M83 & Photometry



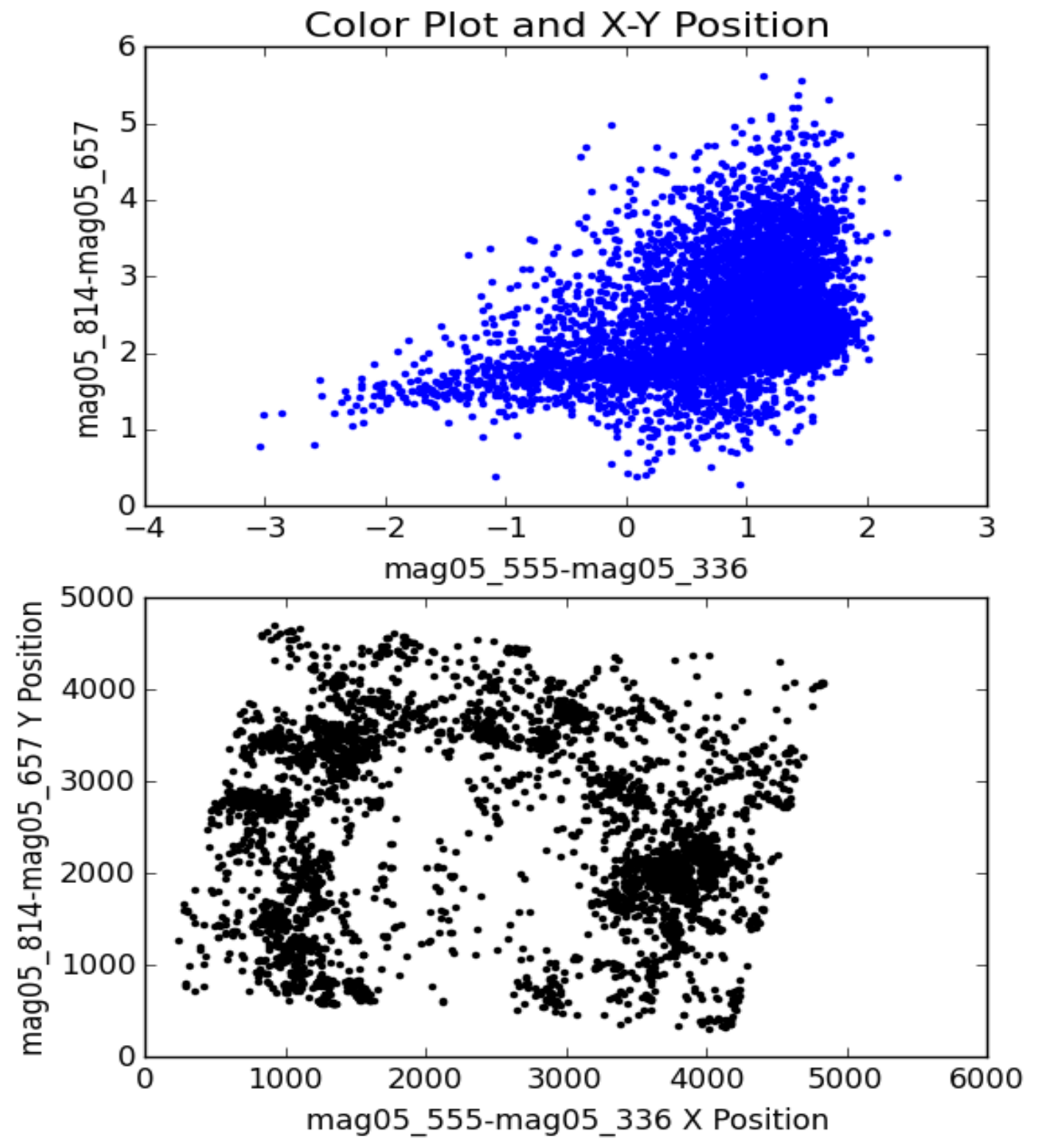
The M83 galaxy is located in the Centaurus A Group, 15 million light years away. M83 is a "barred spiral" galaxy, as it has a bar of stars through its center that push gas towards its center. This funnel of gas creates a very active star forming region near its center, creating many new stars.

Photometry is the measurement and study of the amount of electromagnetic radiation that telescopes receive from celestial objects. Each object emits a specific set of wavelengths of light, which may not be visible to the eye, but reveals clues about the physical properties of the object. The data set used contains 67,840 point sources over 10 wavelengths, each measured using two apertures.

Clustering Methods

Clustering

Photometry is performed in an n-dimensional colour space. The physical location of the objects is not the primary concern, as that location does not describe the composition of the object. In order to understand an object's composition, patterns in the colour space must be found using clustering algorithms.



Mean-shift

Mean-shift clustering finds local modes or peaks in a nonparametric density estimate. The method assigns each data point with the closest peak, and determines the number of clusters. When applied to the data set, the mean-shift method identified areas between two combinations of wavelengths that had a high density of objects.

K-Means

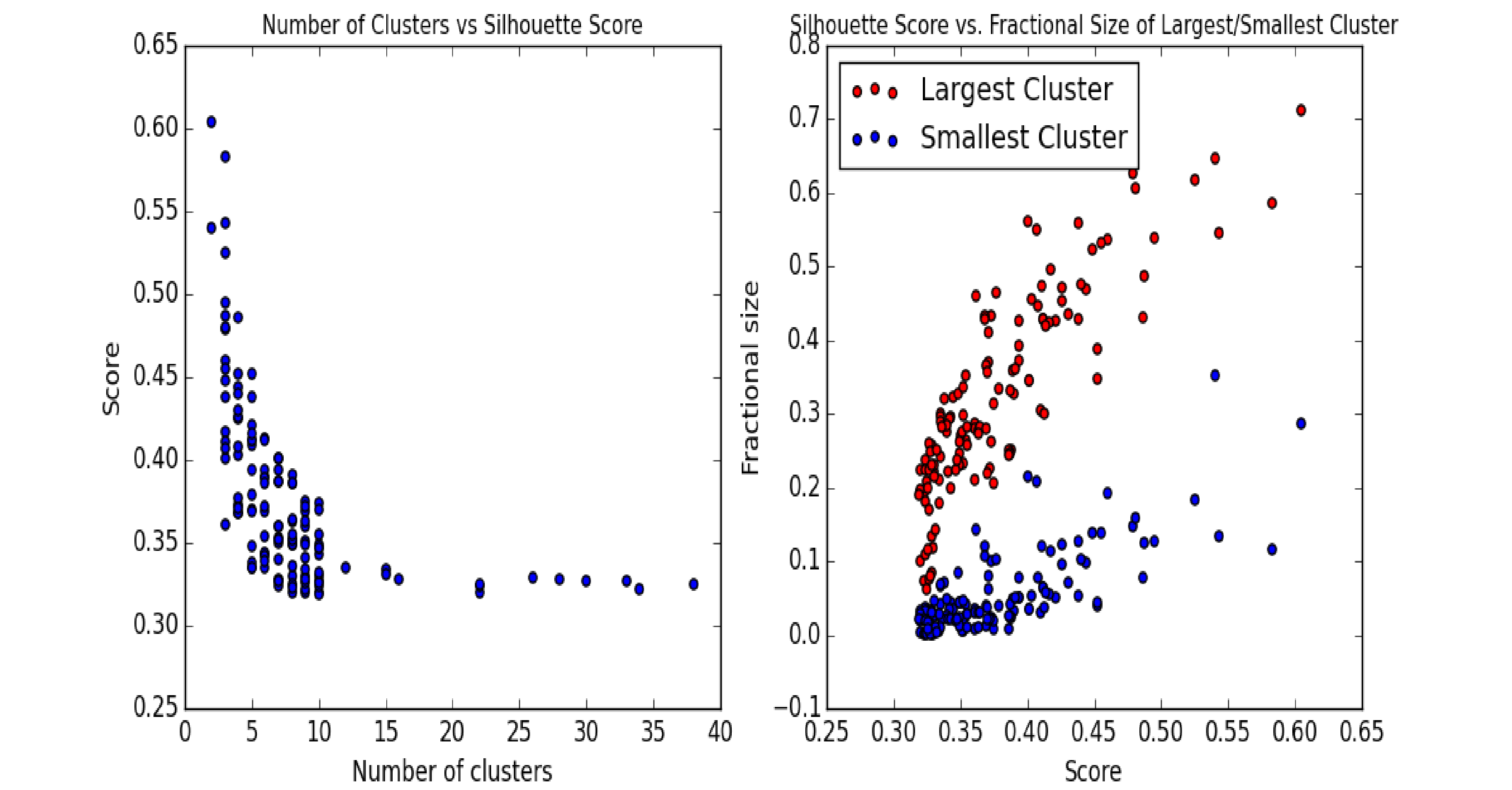
K-Means clustering minimizes the sum of squares objective function. The method picks a centroid of each *k* cluster, and assigns the closest points in the data set to each centroid. The mean-shift output was used as the number of clusters for the k-means algorithm, and the algorithm assigned the objects to each cluster.

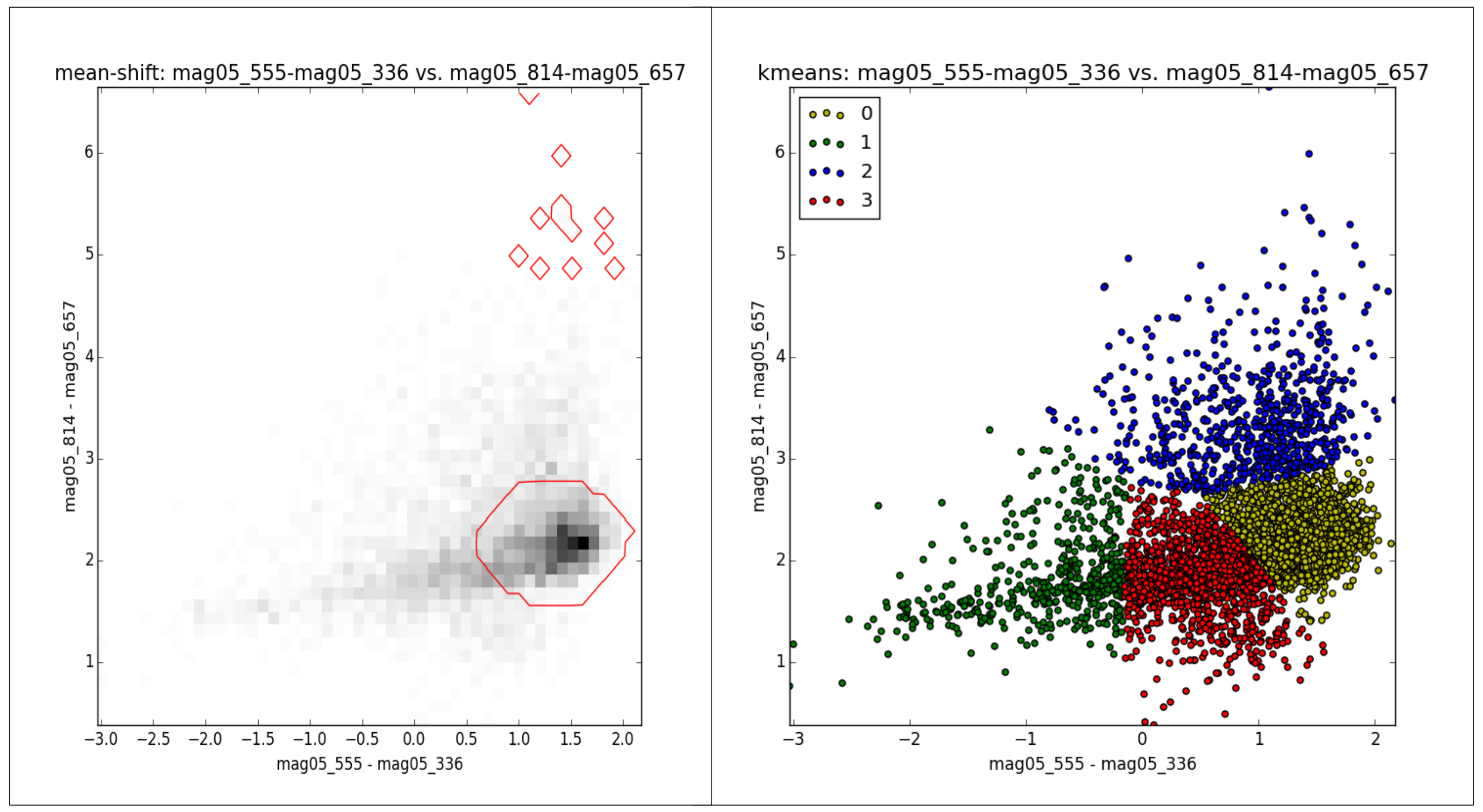
Results

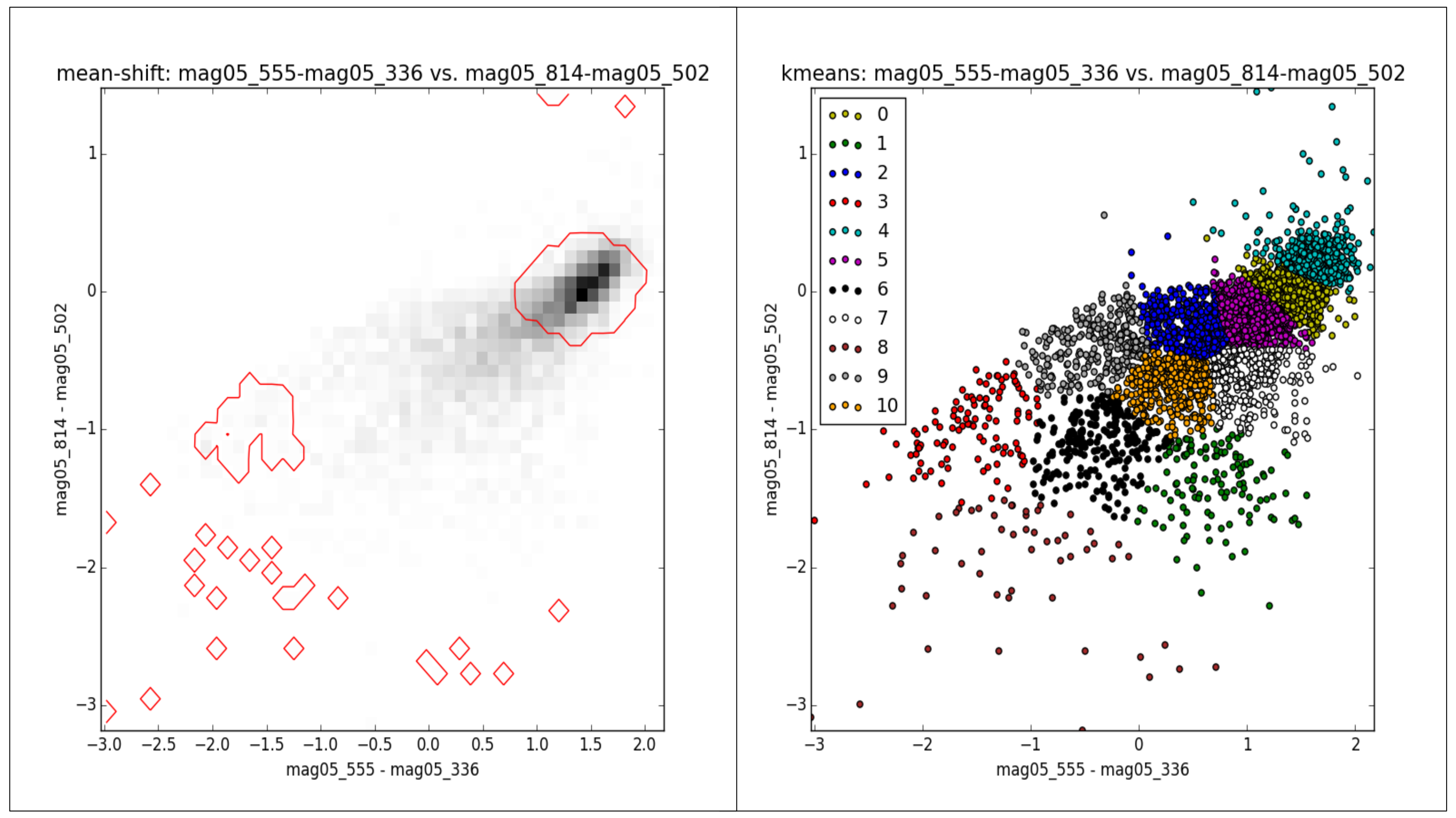
Right: The first panel shows the optimal number of clusters generally lies between 5 and 10. The second panel shows that as the accuracy increases, the lowest proportion of objects in each cluster increases as well. The clustering tends to become less accurate as the number of clusters imposed on the data increases.

Bottom Left: Mean-shift (1) and K-means (2) clustering performed on a combination of wavelengths. These panels show relationship between the methods. The number of clusters predicted by mean-shift accurately describes the data as the K-means plot shows.

Bottom Right: Mean-shift (1) and K-means (2) clustering performed on a different combination of wavelengths. These panels show the loss of accuracy as the number of clusters increases. The number of clusters predicted by mean-shift divides the data excessively where the data is densest. However, k-means is able to effectively identify clusters as the data disperses.


Number of Clusters vs Silhouette Score
Score
Number of clusters
Silhouette Score vs. Fractional Size of Largest/Smallest Cluster
Largest Cluster
Smallest Cluster
Fractional size
Score


mean-shift: mag05_555-mag05_336 vs. mag05_814-mag05_657
mag05_814 - mag05_657
mag05_555 - mag05_336
kmeans: mag05_555-mag05_336 vs. mag05_814-mag05_657
mag05_814 - mag05_657
mag05_555 - mag05_336


mean-shift: mag05_555-mag05_336 vs. mag05_814-mag05_502
mag05_814 - mag05_502
mag05_555 - mag05_336
kmeans: mag05_555-mag05_336 vs. mag05_814-mag05_502
mag05_814 - mag05_502
mag05_555 - mag05_336