

Deconstructing a galaxy: identifying components of M83 with photometric clustering^{*}

P. Barmby^{1†} and A. K. Kiar^{1‡}

¹*Department of Physics and Astronomy and Centre for Planetary Science and Exploration,
University of Western Ontario, London, ON, N6A 3K7, Canada*

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 apply machine-learning methods to classification of point sources of light emission in nearby galaxies. An objects 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, Affinity Propagation, 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 wavelength bands. The data was collected by the Wide Field Camera 3 on the Hubble Space Telescope. To identify which combination of bands was the best at separating different classes of objects, the strength of the clustering was tested using a silhouette score. This metric measures an objects 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.

Key words: keywords here

1 INTRODUCTION

Galaxies are complex systems, comprised of numerous components with an enormous range of size, mass, density, and composition. These components can be divided into baryonic (stars and their remnants, nebulae, star clusters, nucleus) and non-baryonic (dark matter); cataloging the components and describing the interactions between them is a key step in elucidating the natural history of galaxies. Only in nearby galaxies can individual sub-components be resolved. As observational technology has advanced, the definition of “nearby” has changed and will continue to do so, from Milky Way satellites and Local Group galaxies, to a few Megaparsecs (distance at which stars can be resolved with HST), to XX Mpc (distance at which stars can be resolved with JWST), to the entire observable universe with potential future facilities ().

What is the most efficient way to survey the sub-components of a nearby galaxy? Here we are discussing components detectable in imaging at ultraviolet through infrared

wavelengths, i.e. with effective temperatures in the range XX–XX K. Much cooler or hotter types of objects (molecular gas, accreting compact objects) are better-detected at other wavelengths. Particular stellar types, or star clusters, are often identified with broad-band colour-magnitude diagrams (e.g.). Narrow-band filters can also isolate special stellar types (e.g.) or objects prominent in emission lines such as planetary nebulae or supernova remnants (e.g.). Observations are typically designed with detection of particular classes in mind and sometimes re-used for additional purposes (e.g.). Spectroscopic follow-up is often required to confirm candidates. New observational facilities which provide spatially-resolved spectroscopy (, e.g.) may reduce the need for separate imaging and follow-up steps, but greatly increase the complexity of initial data analysis.

Multi-wavelength surveys are extremely common in studies of unresolved galaxies in the distant universe. While these are often designed to select galaxies or active galactic nuclei with specific properties (e.g.), sometimes they are pure blank-field surveys. Broadband ($R = \Delta\lambda/\lambda < X$) filters are the most common imaging modality, although there have been a few attempts at narrow- or medium-band surveys as well (e.g. Wolf et al. 2003), Clustering in colour space

† E-mail: pbarmby@uwo.ca

‡ E-mail: akiar@uwo.ca

can be used to select particular classes of objects from a survey, for example in selecting AGN via mid-infrared colours (e.g.), or high-redshift galaxies via Lyman-break dropouts (e.g.). **give some examples here of sophisticated analysis of colour spaces.**

The purpose of this work is to treat a nearby galaxy as if it were a blank field for surveys, and investigate the usefulness of different photometric colours for identifying sub-components. We make use of the Early Release Science (ERS) observations with the Wide-Field Camera 3 (WFC3) of the nearby spiral galaxy M83 () and in particular the catalog of point sources produced by . We form colours from the photometric measurements in the catalog and apply several clustering techniques to two-colour datasets. In conjunction with published catalogs of galaxy components, we identify the optimum parameters for clustering such a photometric dataset, and the best choices of filter.

2 DATA

2.1 Imaging dataset

The dataset used for this study is the Wide-Field Camera-3 Early Release Science (ERS) observations of the nearby spiral galaxy Messier 83 (M83). M83 is a grand-design spiral of type SAB, located at a distance of 4.66 Mpc (Tully et al. 2013) and the largest member of the M83 subgroup of the nearby Centaurus group of galaxies (Tully 2015). The galaxy’s apparent radius of ~ 12 arcmin () is reasonably well-matched to the camera’s field of view (XX true? XX) **And here we note some other interesting things about M83.**

The objective of the ERS observations as a whole was to probe star formation in galaxies. The observations of M83 were made in broad- and narrow-band filters in order to characterize both stellar and nebular properties. They cover a 3.6×3.6 kpc² region in the northern portion of the galaxy, including the nucleus, a portion of a spiral arm and an interarm region. The spatial resolution of the images is $0''.0396$ arcsec pixel⁻¹, corresponding to a linear scale of XX pc pixel⁻¹ at the 4.66 Mpc distance. A complete description of the observations and data processing is given by Chandar et al. (2010); our work here uses the observations in the UVIS channel, listed in Table 1. A number of previous studies have used the ERS M83 dataset for various purposes. These include studies of star clusters (Chandar et al. 2010; Wofford et al. 2011; ?; Bastian et al. 2011, 2012; Founesneau et al. 2012; Silva-Villa et al. 2013; Andrews et al. 2014; Chandar et al. 2014; Adamo et al. 2015; Ryon et al. 2015; Hollyhead et al. 2015; Sun et al. 2016), H II regions (Liu et al. 2013), supernova remnants and the interstellar medium (Dopita et al. 2010; Hong et al. 2011; Blair et al. 2014, 2015), resolved stars (Kim et al. 2012; Williams et al. 2015), and a super-Eddington off-nuclear black hole (Soria et al. 2014).

We analyze the catalog produced by Chandar et al. (2010) and made available via **REF**, hereafter referred to as the ‘ERS catalog.’ The objects in this catalog were detected on a ‘white-light’ image produced by a weighted combination of the *UBVI* images. Photometry in 0.5- and 3-pixel radius apertures at the positions of the detected

Table 1.

Filter	Name	Exposure time
F225W	Wide UV	1800 s
F336W	<i>U</i> -band	1890 s
F438W	<i>B</i> -band	1180 s
F487N	H β	2700 s
F555W	V-band, South field	1203 s
F814W	<i>I</i> -band	1203 s

sources was performed on the broad- and narrow-band images and tabulated in the Vega magnitude system. We apply the correction to the F657N magnitude zeropoint (from 20.72 to 22.35) noted in the header of the catalog. Chandar et al. (2010) discussed aperture corrections for this catalog, but since we are primarily concerned with colours and the aperture correction does not vary strongly with wavelength, we omit it. The catalog contains about 68000 objects which are expected to include individual stars, star clusters, stellar blends, supernova remnants, Hii regions, planetary nebulae, and background galaxies. Completeness and reliability of the catalog are not discussed by Chandar et al. (2010), but a visual inspection of the the detected sources on the white-light image suggests that a reasonable balance between completeness and reliability was achieved. Nine objects are flagged in the catalog as being problematic and we remove them from our analysis.

[to be re-organized] As a check on the catalog we used SExtractor to detect and photometer objects in the individual images. While the aperture photometry measurements matched well, the derived uncertainties were much smaller than those reported in the catalog. Indeed, the catalog uncertainties seem to be physically unreasonable, with median uncertainty values well above 1 magnitude in most band-passes, and the catalog notes do not recommend them for use except in a relative sense. Our comparison implied that recovering a more typical magnitude uncertainty distribution would be accomplished by dividing the 0.5-pixel magnitude uncertainties by 10 for the broad-band filters and 15 for the narrow-band filters. This allows us to use the catalog aperture magnitudes as an indicator of detected signal-to-noise: our analysis uses only objects with (scaled) 0.5-pixel magnitude uncertainties < 0.2 mag. For the remainder of the analysis we use magnitudes measured in the 0.5-pixel radius aperture, as these should be less affected by crowding and the variable galaxy background.

Table 2 and Figure 1 characterize the catalog in terms of measurements in individual filters. Not all objects are detected in all filters; Table 2 gives the number of objects for which photometry is reported in a given filter, the number for which scaled 0.5-pixel magnitude uncertainty is 0.2 mag or less, and the aperture magnitude at which the median magnitude uncertainty is 0.2 mag. Figure 1 shows the distributions of magnitudes and uncertainties in a broad and narrow filter.

Our analysis in this paper is primarily concerned with colours, rather than luminosities. Uncertainties in colours are computed as the quadrature sum of the relevant magnitudes. Observations in 10 bands allow the generation of 45 different colours, but not all of these colours are likely to be useful in characterizing components of the galaxy. As the

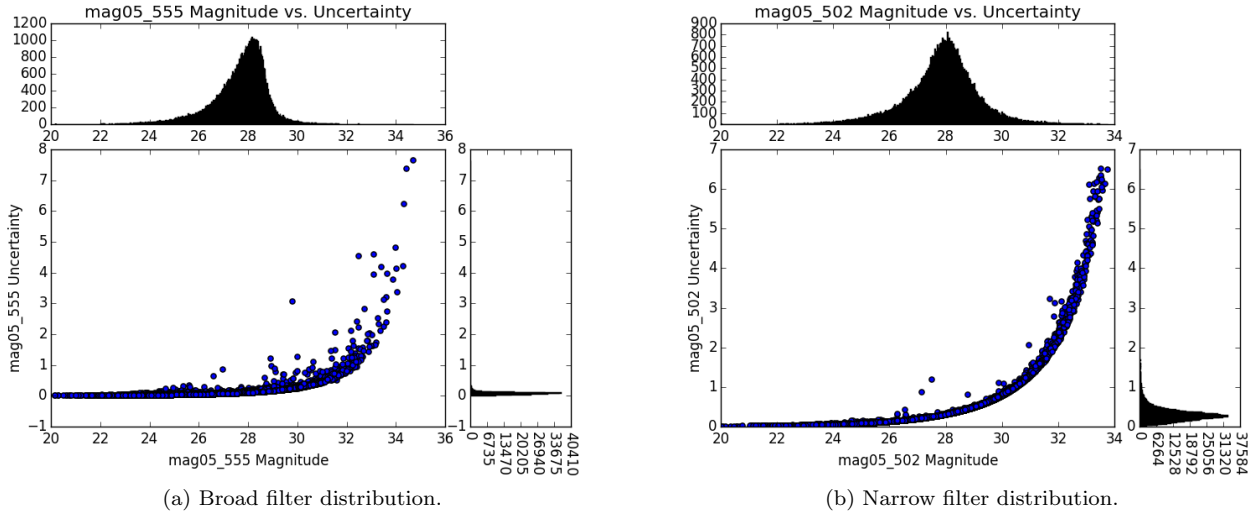


Figure 1. Distribution of magnitudes and uncertainties for objects in the Chandar et al. (2010) M83 ERS catalog.

Table 2.

Filter	N_{obj}	N_{good}	m_{good}	$N_{\text{X}-555}$
F225W	57237	15011	m	14976
F336W	62192	34129	m	33692
F373N	55966	8878	m	8853
F438W	66356	48858	m	48660
F487N	63812	13335	m	13322
F502N	64313	14654	m	14644
F555W	67424	65652	m	—
F657N	67782	67634	m	65502
F673N	65305	25295	m	65502
F814W	67050	59600	m	57935

F555W band has the most individual detections, we initially compute colours relative to this band. The last column of Table 2 gives the number of objects for which a ‘good’ colour (uncertainty < 0.2 mag) is available.

2.2 Published catalogs

As one check on the results of our analysis, we use previously-published identifications of specific types of objects in M83. We compiled a ‘published catalog’ by combining the contents of the NASA Extragalactic Database (NED) and [what does it stand for?] (SIMBAD Wenger et al. 2000) and then adding the catalogs of Wolf-Rayet stars (Kim et al. 2012) and red supergiant candidates (Williams et al. 2015), which did not appear in either database. NED’s focus as an extragalactic database and SIMBAD’s focus on Galactic objects mean that their contents overlap but are not identical, and this is true of the area surrounding M83. A $3.3'$ radius region around the coordinates centered at (204.26761 deg, -29.839939 deg) contains 1553 NED objects and 1772 SIMBAD objects, of which 1220 are matched with each other at $1''$ tolerance. Although the two services use slightly different naming conventions, with human inspection the matches are generally recognizable as referring

to the same object. Interestingly, the databases do not always report the same object type even when the names are identical. The differences are reasonable in some cases (a supernova remnant can also be an X-ray source, for example), but not others (e.g. CXOU J133703.0-294945 is reported as a supernova remnant by SIMBAD and an *Hii* region by NED). A detailed study of the databases is beyond the scope of this work; for the purposes of this analysis, we kept the NED classification for objects which appeared in both databases. Objects which appeared in one database but not the other were primarily from recent work (e.g. Long et al. 2014), from older studies likely superseded by newer ones (e.g. Larsen 1999), or from studies in which only coordinates relative to the galaxy centre were given (de Vaucouleurs et al. 1983).

Our final combined catalog has 2425 objects of which 750**check** are in the region covered by the ERS catalog. The main classes are star clusters (350), X-ray sources (105), supernova remnants (86), *Hii* regions (81), and radio sources (36). Nearly every entry in the published catalog had an ERS catalog object within $1''$, and the mean distance between matched objects was $0''.26$. Given the nearly 100-fold difference in object density between the two catalogs, matching based on positions alone may result in spurious matches **REF**. *Some discussion of the exact matching procedure is warranted here, and a conclusion on what the best thing to do is.**

3 METHODS

As the size of galactic surveys grows, the number of dimensions available for analysis increases. In this survey, 45 different colour combinations are possible, creating a space of 45 possible dimensions. Clustering methods provide an efficient way of finding structure in high dimensional data by searching for structure in the feature spaces that cannot be visually inspected. A feature space is a set of n features that are associated with measurable quantities. The following techniques were used to cluster the data, and determine

the most significant features. All analysis was implemented using the *sklearn.cluster* Python package.

3.1 Principal Component Analysis

Images of objects at various wavelengths share similar structures. These structures can be built using a subset of only the most significant of those images. Principal Component Analysis (PCA) is a process of determining the most significant features of a high dimensional space in order to rebuild structures effectively Kuntzer et al. (2016). PCA projects a data set on its most significant basis, only keeping the dimensions that best explain the data. These dimensions are kept by imposing a maximum number of components on the algorithm, and it is forced to only select the most significant ones Kuntzer et al. (2016). Since the number of colour dimensions in this problem is not large, PCA was only considered when determining the most significant colours. The Explained Variance Ratio was used to determine each colour significance in each combination. PCA has been applied to many astronomical surveys... In this study, PCA is conducted to determine the most significant filters, and colours for classifying objects in M83. All ten filters were used in the initial analysis...

3.2 Mean Shift Clustering

Mean Shift is a non-parametric clustering technique that is based on probability density function estimates of each point in the data. Mean Shift is a very powerful algorithm, but has not been widely used in astronomy. At each point, the algorithm estimates the density around that point using a small sample of objects surrounding the point. The power of Mean Shift clustering is that the clusters are not confined to a particular shape. Because Mean Shift moves towards the local mode near the data on which it was initialized, it is useful for estimating the number of significant clusters in a dataset Comaniciu & Meer (2002). The algorithm is based on two components: kernel density estimation, and density gradient estimation. We will highlight the major components of the algorithm, for a full description of the, see Vatturi & Wong (2009).

The first element of Mean Shift is kernel density estimation. The major parameter of Mean Shift is bandwidth, \mathbf{H} , which is assumed to be proportional to the matrix $\mathbf{H} = h^2 \mathbf{I}$, with $h > 0$ Vatturi & Wong (2009). The density estimator for a multivariate density kernel is given by:

$$\hat{f}(x) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^n k\left(\left\|\frac{x-x_i}{h}\right\|^2\right) \quad (1)$$

Where h is the magnitude of the bandwidth matrix, $k(x)$ is the profile of kernel $K(x)$, and $c_{k,d}$ is a constant making $K(x)$ integrate to one Vatturi & Wong (2009).

The second element of Mean Shift is density gradient estimation. The density gradient is estimated from the gradient of equation 1 Vatturi & Wong (2009). The density gra-

dient is given by:

$$\nabla \hat{f}_{h,K}(x) = \frac{2c_{k,d}}{nh^d(d+2)} \left[\sum_{i=1}^n k' \left(\left\| \frac{x-x_i}{h} \right\|^2 \right) \right] \left[\frac{\sum_{i=1}^n x_i k' \left(\left\| \frac{x-x_i}{h} \right\|^2 \right)}{\sum_{i=1}^n k' \left(\left\| \frac{x-x_i}{h} \right\|^2 \right)} - x \right] \quad (2)$$

The second term of equation 2, is the Mean Shift; the difference between the weighted mean using k' , and x Vatturi & Wong (2009). Applying a normal kernel to the Mean Shift, the second term of equation 2 becomes:

$$m_{h,K}(x) = \frac{\sum_{i=1}^n x_i \exp\left(-\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^n \exp\left(-\left\|\frac{x-x_i}{h}\right\|^2\right)} - x \quad (3)$$

$m_{h,K}$ is the Mean Shift, and always points in the direction of largest ascent through the estimated density function Vatturi & Wong (2009).

Mean Shift clustering involves the application of equation 3 to shift the points of a data set towards the direction of the Mean Shift vector Vatturi & Wong (2009). The points are shifted by:

$$x^{i+1} = x^i + m_{h,K}(x^i) \quad (4)$$

Shifting the data points by equation 4 ensures that when the points converge, the center is the area of highest local density, or density “mode”. The density mode can be interpreted as the center of a significant cluster in the data set, and is used to classify the objects that were shifted towards it. Equation 1 introduces the bandwidth parameter h . Estimating the bandwidth correctly is critical to determining the correct number of clusters. If the bandwidth is too low, the density estimate will be undersmoothed, and Mean Shift will produce many small clusters Vatturi & Wong (2009). This is a result of the large density gradient resulting from a low bandwidth, causing many data points to be interpreted as local modes. Conversely, if the bandwidth is too large, a small number of large clusters will be detected, resulting in groupings of data that may blur the underlying structure Vatturi & Wong (2009).

3.3 Affinity Propagation Clustering

Affinity propagation (AP) is a relatively new clustering technique developed by Frey & Dueck (2007). Here, we will briefly describe the main components of AP, for a full description of the technique, see Frey & Dueck (2007). AP takes the similarities between the data points as input for clustering, and uses a series of “messages” between data points to determine the number of clusters and their centers. The centers of AP clustering are actual data points, called exemplars, which make it useful for clustering as it does not create average centers for each cluster. The only input required for AP are the *preferences* of each data point which describes the likelihood of a data point to be chosen as an

exemplar Frey & Dueck (2007). The preferences are a measure of the similarity between a point i and a candidate exemplar k defined by:

$$s(i, k) = -\|x_i - x_k\|^2 \quad (5)$$

Similarity values influence the number of clusters AP identifies, as the larger similarity values are likely chosen as exemplars Frey & Dueck (2007). If all data points are equally suitable to be used as exemplars, the preference value could be common among all data points. This technique was adopted as no prior knowledge of the data set was used to determine where clusters were centered. Preference values could be estimated using the median value of the similarities, the minimum value, or randomized to see the effects over various clusterings Frey & Dueck (2007).

Once the preference value is determined, two messages are computed between all the data points. The first message is the "responsibility" $r(i, k)$, which is sent from point i to candidate exemplar k : Frey & Dueck (2007)

$$r(i, k) \leftarrow s(i, k) - \max_{k' \text{ s.t. } k' \neq k} \{a(i, k') + s(i, k')\} \quad (6)$$

Responsibility measures the evidence of how suitable point k is to be an exemplar of point i Frey & Dueck (2007), after considering other potential exemplars for point i . The "availability", $a(i, k')$ in equation 6, is sent from candidate exemplar k to point i to compute the evidence for how appropriate it would be for point i to choose candidate k as an exemplar, considering evidence from other points that believe candidate k should be their exemplar Frey & Dueck (2007):

$$a(i, k) \leftarrow \min\left\{0, r(k, k) + \sum_{i' \text{ s.t. } i' \notin \{i, k\}} \max\{0, r(i', k)\}\right\} \quad (7)$$

The availabilities of all points are initialized to zero, and the first iteration of responsibilities are set to the input preferences Frey & Dueck (2007). Each iteration updates equation 6 and equation 7 to determine the optimal exemplars for the data. As the availabilities (Equation 7) are updated, a threshold is calculated to ensure that the availability for a given candidate does not become positive:

$$a(k, k) \leftarrow \sum_{i' \text{ s.t. } i' \neq k} \max\{0, r(i', k)\} \quad (8)$$

Equation 8 is called the "self-availability", and reflects the evidence that k is an exemplar, based on the responsibilities sent to k from other points Frey & Dueck (2007). The self-availability limits the influence of strong candidates, to ensure that false exemplars are not chosen.

As the process iterates, for point i , the value of k that maximizes $a(i, k) + r(i, k)$ identifies i as an exemplar if $k = i$, or gives the exemplar of point i Frey & Dueck (2007). In order to ensure that the message passing does not cause numerical oscillations, the messages are damped as they are updated. The previous message value is multiplied by a damping-factor λ , and $1 - \lambda$ multiplied by the update value is added. The damping-factor has a value between zero and one, with a default value of 0.5 Frey & Dueck (2007).

3.4 K-Means Clustering

K-Means clustering is one of the most widely used clustering methods and has been used to identify a wide range of interstellar and intergalactic objects. It is simple, robust, and easy to implement when analyzing high dimensional spaces, making it a powerful way to analyze galactic surveys. Generally, k-means begins by selecting k data points at random and deems these points cluster centers. Each object in the data set is then assigned to a cluster center by computing the least-squares distance to each center. K-Means aims to minimize the sum of squares within each cluster given by:

$$J = \sum_{n=1}^N \sum_{k=1}^K \min(\|x_n - \mu_k\|^2) \quad (9)$$

Each point, x , is then assigned to the cluster center with the lowest distance in equation 9 Tammour et al. (2016). Once all data points have been assigned, the centers are re-calculated by taking the average of all the points in each cluster. This process continues until the centers do not change after two consecutive iterations Almeida & Prieto (2013).

- (i) description of clustering and classification
- (ii) description of PCA
- (iii) description of Mean-Shift
- (iv) description of Affinity Propagation
- (v) description of K-Means
- (vi) experiments with how to apply the techniques
- (vii) final parameters used

4 ANALYSIS

In this section we will outline the process used for each of the clustering methods. An aim of this work was to help astronomers determine which filters were best at identifying different types of objects in a survey. Since the average survey is limited to four filters, different combinations of four filters were used to construct colours for clustering. Colours were constructed using the difference between two wavelengths. For a given combination of four filters, all possible colour combinations were used. The feature space was constructed in two and three dimensions, to add a layer of analysis that isn't typically found in colour-colour space.

4.1 Clustering Process

Clustering was performed using all three methods for each colour combination. The methods were used independently and in unison in order to determine the optimal clustering. The following process allowed the investigation of the effect of all parameters on each clustering technique, leading to an optimal clustering.

(i) **Mean-Shift:** Mean-Shift clustering was performed first by estimating the bandwidth parameter with the *estimate_bandwidth* function in *scikit-learn*. Following the initial clustering, a "bandwidth hierarchy" was created by performing the clustering again with bandwidth values between \pm intervals of 0.1 from the estimated bandwidth.

(ii) **AP:** Affinity Propagation clustering was then performed by setting the preferences to 10% of the number of objects in the data set, and setting the damping factor to

0.95. Following that clustering, an "Affinity hierarchy" was created by varying the damping factor, and preference value to determine the effect of each parameter.

(iii) **K-Means:** K-Means clustering was performed last. The first two clusterings were performed using the number of clusters determined from the initial clusterings by Mean-Shift and AP. Next, K-Means was performed with $K = \pm 4$ from the original clustering.

4.2 Selecting the Clustering

In order to determine the optimal clustering, a series of metrics were used. For each clustering, the silhouette score, cluster centers, root mean square, average and standard deviation of each colour, and the distances between points in each cluster were calculated.

4.2.1 Silhouette Score

The silhouette score is a metric used to describe the compactness of a cluster in a given clustering and is calculated as an average of all samples in a clustering. The silhouette score is given by:

$$\text{SilhouetteScore} = \frac{b - a}{\max(a, b)} \quad (10)$$

where a is the mean intra-cluster distance, and b is the distance between a point and the nearest cluster that point is not a member of. In addition to the average score of the clustering, the average score for each cluster within the clustering was computed. The cluster score determines what is driving the average score, and uncovers which clusters are most compact.

4.2.2 Cluster Statistics

Various statistics were calculated for each cluster within a clustering to help describe the similarity between the objects in a given cluster. The standard deviation and average colour was calculated for each colour and each cluster within a clustering. These metrics describe the distribution of the objects in the colour-colour space within a cluster. Clusters that had large standard deviations were viewed as too dissimilar to be a meaningful cluster, and the clustering parameters were changed or the clustering was removed. Clusters whose averages and medians were not close were also discredited.

The fractional sizes of each cluster were also calculated. This metric describes the distribution of objects between clusters, and provided a complete way to evaluate a given clustering. If a clustering segmented the objects into a large cluster followed by several smaller ones, the clustering was investigated further, as this segmentation could mean one of two things. This type of clustering could be a result of the identification of interesting objects, in which case the clustering algorithm was able to identify the objects and place them in the same cluster. However, this type of clustering could also be a result of the underlying distribution of the data, as the clustering techniques are largely drawn to areas of high density. If this is the case, the clustering only created the smaller clusters as a result of the parameters imposed on the clustering.

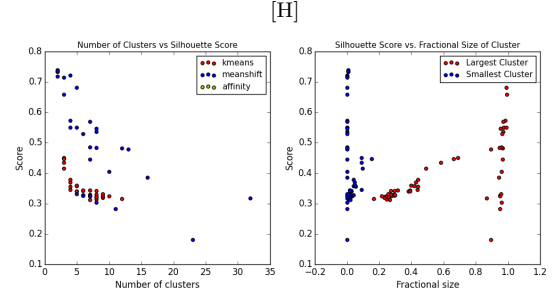


Figure 2. Distribution of the silhouette score as a result of the number of clusters imposed. The *blue* points are the scores of Mean-Shift clustering, *red* points are scores of K-Means, and *yellow* points are scores of AP.

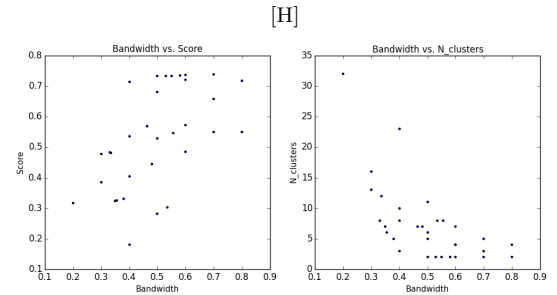


Figure 3. Distribution of the silhouette score as a function of bandwidth, and the distribution of the number of clusters as a function of bandwidth.

4.2.3 Parameter Relationships

In addition to metrics, the relationships between various metrics were investigated to determine the optimal clustering. Figure 2 shows the relationship between the silhouette score and the number of clusters imposed on the data for each type of clustering. The optimal number of clusters is found where the relation flattens. For K-Means, this point is between 5-10 clusters.

The Mean-Shift scores do not follow a similar distribution as K-Means, as the accuracy of Mean-Shift is more directly related to the bandwidth parameter, seen in Figure 3. The optimal Mean-Shift clustering was chosen by finding the bandwidth where the relation between the bandwidth and number of clusters reached an elbow, which was usually between 3 - 5 clusters.

5 RESULTS

Tobereorganized

The results of the analysis are grouped into the type of band used to make each colour.

5.1 Broad & Broad Band Combinations

The four broad bands were used to create different colours. The only colour that was omitted from the analysis was the $U - I$ colour. This was omitted because the range between those wavelengths is large enough that it is unlikely

the ratio between the flux of those bands has any physical meaning.

5.1.1 $U - B$ vs. $V - I$

5.2 Broad & Narrow Band Combinations

Each broad and narrow band combination was tested against every broad-broad band combination that did not boarder the narrow band.

5.2.1 $UVW - U$

The $UVW - U$ combination was tested clustered with the $B-I$, $V-I$, and $B-V$ colours.

5.2.1.1 Mean-Shift When clustered using Mean-Shift, a similar pattern of clustering was seen in all combinations. Due to the structure of the Mean-Shift algorithm, it is drawn towards areas of high density in the distribution. Since the distribution of the $UVW-U$ combinations were generally concentrated around zero in the colour-colour space, the Mean-Shift algorithm would pick out one large cluster with many smaller ones. Each smaller cluster had to be investigated to determine if the algorithm had found a meaningful cluster, or if it had just clustered noise. In order to determine this, a cluster hierarchy was created with different bandwidth values to see how long each cluster lasted in the bandwidth space. If a small cluster was created at a low bandwidth level and stayed alive until the number of clusters became 2-3, then it is reasonable to assume that the cluster was meaningful. Figure 4 shows the result of one trial of Mean-Shift clustering with $h = 0.6$ which created 4 clusters. Figure 5 shows the result of one trial of Mean-Shift clustering with $h = 0.4$, creating 10 clusters. The structure of cluster 2 in both Figure 4 and Figure 5 can be seen distinctly. This cluster would have been viewed as noise if the hierarchy was not created, but after testing multiple bandwidth values, it can be seen that these objects are significant. Varying the bandwidth did not have a large affect on the number of clusters produced. Bandwidth values from 0.5 to 1.0 only created a difference in three clusters generated.

5.2.1.2 K-Means When clustered using K-Means, two types of results were seen. Figure 6, shows the result of K-Means clustering for $K = 5$ against the $B-V$ colour.

The algorithm split the data into groups based on its $UVW - U$ colour. The pattern continued for all values of K , and the silhouette scores of the combination elbowed at $K = 5$. The second type of result for K-Means can be seen in Figure 7, against the $B-I$ colour. This clustering segmented the data into circular groups within the distribution. Similar to the $B-V$ combination, the silhouette score elbowed at $K = 5$.

5.2.1.3 Astronomy Implications After the clustering was performed, the location of the cluster members were imposed on the whitelight image of M83. The algorithms were able to identify sources that are located in unique sections of the galaxy. K-Means segmented the data into objects that are located along the spiral arms exclusively, clustered in the

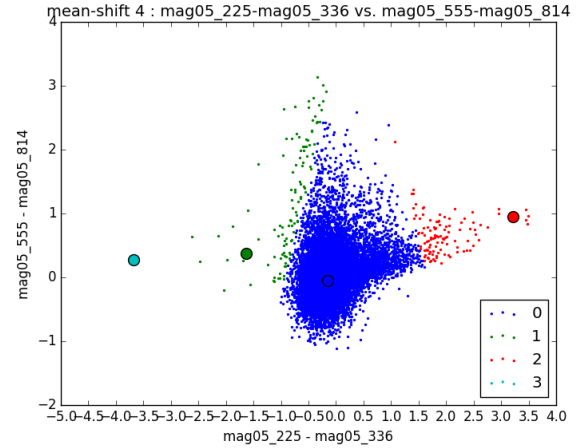


Figure 4. Colour-Colour distribution of the $UVW-U$ and $V-I$ colours, clustered using Mean-Shift with $h = 0.6$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

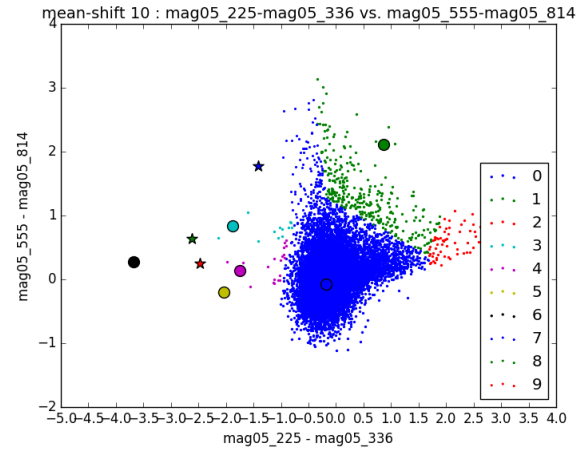


Figure 5. Colour-Colour distribution of the $UVW-U$ and $B-I$ colours, clustered using K-Means with $h = 0.4$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

denser regions of the arms and nucleus, and in the intra-arm region. This pattern of object location was found through all combinations of broad - broad colours. The Mean-Shift segmentation was able to identify objects that were located in the spiral arms. The objects that were not part of the main cluster appear to be point sources that fall in the back of the galaxy, or are some form of cloud or dust.

5.2.2 $U - OII$

The $U - OII$ combination was clustered with the $B-V$, $B-I$, and $V-I$ colours using Meanshift followed by KMeans.

5.2.2.1 2-Dimensions This colour seemed to be much more sensitive to bandwidth selection than other combinations. With the $B-V$ colour, $h = 0.2$ produced 32 clusters, while $h = 0.4$ produced 3. With the $V-I$ colour, $h = 0.35$

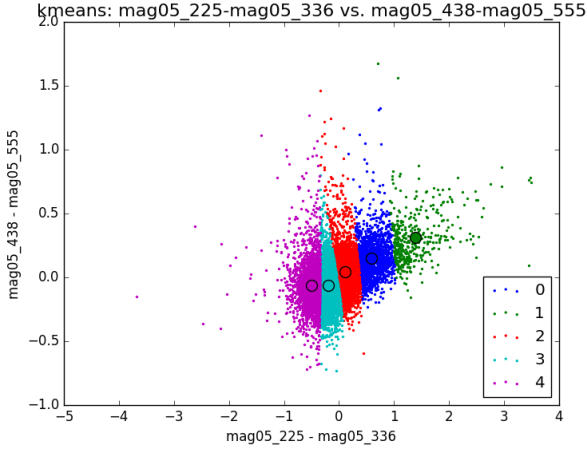


Figure 6. Colour-Colour distribution of the UVW-U and B-V colours, clustered using K-Means with $K = 5$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

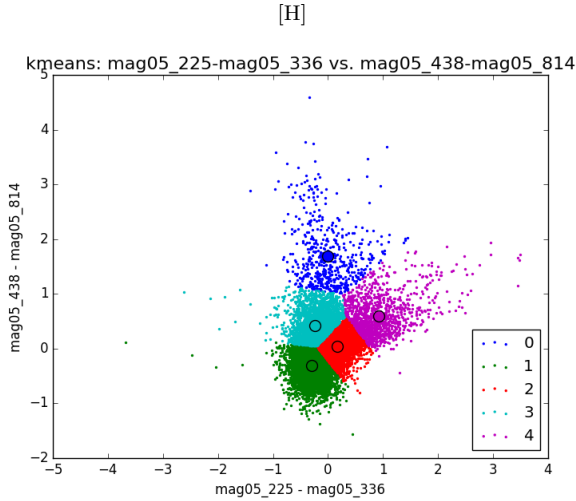


Figure 7. Colour-Colour distribution of the UVW-U and B-I colours, clustered using K-Means with $K = 5$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

produced 17 clusters, while $h = 0.6$ produced 3. Due to this sensitivity, the bandwidth hierarchy was created on much narrower increases in h , which produced more meaningful clusters. After producing the narrow hierarchy, the mean-shift algorithm predicted a range of clusters from 3 to 13. In each clustering, the algorithm did not seem to segment the data significantly, as similar to UVW - U, it produced one large cluster with several smaller ones. The number of clusters predicted reduced linearly with the bandwidth selected, however, the silhouette score saw a sharp drop at $h = 0.33$, which produced 8 clusters, see Figure 8. This clustering segmented the data into three main groups, which were two "arms" in the distribution that spread to the redder areas of both colours. Despite picking out these two groups, the two arms contained only approximately 5% of the data and required further investigation.

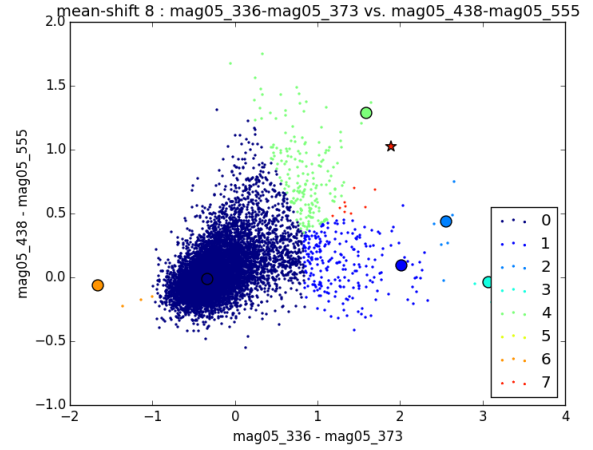


Figure 8. Colour-Colour distribution of the $U - O_2$ and B-V colours, clustered using Meanshift with $h = 0.33$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

The K-Means algorithm produced more reliable results, as it produced clusters of relatively similar sizes. As K increased, the sum of squares value for each clustering decreased, a trend that is expected. The silhouette score was a maximum at $K = 3$, and elbowed at $K = 5$. Both clusterings were investigated to determine which was optimal. At $K = 3$, the distribution was segmented according to its U - OII colour. At $K = 5$, the segmentation was similar, however, the section of data that was significantly red in the U-OII colour was given its own cluster.

With each combination of broad bands, the same patterns existed. This combination in 2-Dimensions did not seem to uncover any more detail or interesting objects than the UVW-U combinations.

5.2.2.2 3-Dimensions Following the initial clustering, the colours were each broken down into a combination of the OII band and each other band. The colours used in three dimensions were a combination of U-OII, OII-B, OII-V, and OII-I.

The clustering performance in three dimensions was generally better for almost all clustering parameters. With all combinations, the clustering algorithms were able to identify a large branch of objects that was fairly red in the U-OII colour, and very blue in the other two, see Figure 9. This branch was identified at all values of K , and most values of h . The added complexity of three dimensions removed the restrictions of only using two dimensions, and allowed the algorithms to cluster the distributions more accurately.

The optimal meanshift clustering was not as apparent in three dimensions. In the OII - B vs. OII - V combination, the score and number of clusters did not plateau, and the Meanshift clustering was not considered for the optimal clustering. However, in the OII - B vs. OII - I and OII - V vs. OII - I combinations, the number of clusters elbowed at an h value that maximized the silhouette score. The number of clusters elbowed at 5, over a range of h values for both combinations. In each case, the algorithm was able to pick out groups of outliers more clearly than in two dimensions,

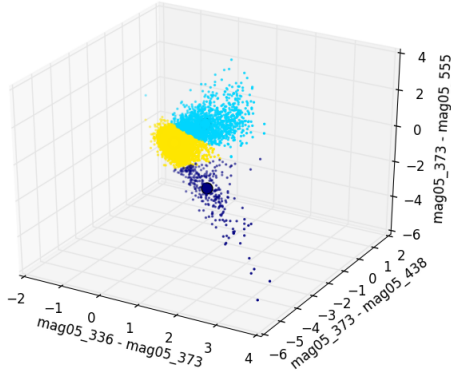


Figure 9. Colour-Colour distribution of the $U - O_2$, $O_2 - B$, and $O_2 - V$ colours, clustered using K-Means with $K = 3$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

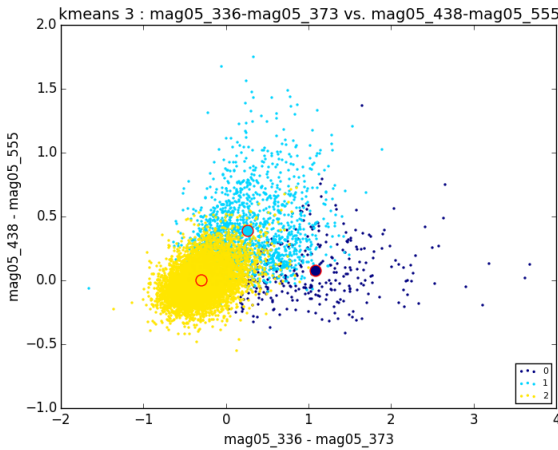


Figure 10. Colour-Colour distribution of the $U - O_2$ and B-V colours, projected from the 3D clustering using K-Means with $K = 3$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

and the elbow point was taken as the optimal meanshift clustering.

The K-Means algorithm was superior to meanshift for picking out evenly sized groups in all combinations, however, it was not able to pick out some of the detail lying in the groups of outlier data. In the combinations that meanshift was successful in, the score peaked at $K = 4$, and this was chosen for the optimal clustering. In the last combination, the score peaked at $K = 3$. In addition, each K value was able to pick out the clear branch of objects. When projected back into two dimensions, the successful segmentation of K-Means can be seen, as it identifies both branches of red objects, and the dense area around zero, see Figure 10.

5.2.2.3 Astronomy Implications Both 2 and 3 dimensional clusterings were investigated in the whitelight image. The clusterings in three dimensions were able to segment ob-

jects more diffinitively than two dimensions. This was most noticable in the red branches of the distribution. In two dimensions, the clusters that segmented the red branches seemed to be a combination of dim point sources and objects in the back of the galaxy or behind clouds. In three dimensions, these clusters were almost entirely objects in the back of the galaxy or behind clouds instead of the combination. Additionally, the clusters in three dimensions were better able to detect the boundary between the dense center of the distribution and the outlying branches. When projected onto the galaxy, the objects dense center of the distribution were located in the densist areas of the spiral arms. The three dimensional clusterings were able to identify these objects and keep them as a separate cluster.

ACKNOWLEDGMENTS

The authors acknowledge financial support from the Natural Science and Engineering Research Council (NSERC) of Canada. This research has made use of the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. This research has made use of the SIMBAD database, operated at CDS, Strasbourg, France. We acknowledge the efforts of WFC3 Science Oversight Committee in conducting the Early Release Science program.

REFERENCES

- Adamo A., Kruijssen J. M. D., Bastian N., Silva-Villa E., Ryon J., 2015, MNRAS, 452, 246
- Almeida J. S., Prieto C. A., 2013, The Astrophysical Journal, 763, 50
- Andrews J. E., Calzetti D., Chandar R., Elmegreen B. G., Kennicutt R. C., Kim H., Krumholz M. R., Lee J. C., McElwee S., O’Connell R. W., Whitmore B., 2014, ApJ, 793, 4
- Bastian N., Adamo A., Gieles M., Lamers H. J. G. L. M., Larsen S. S., Silva-Villa E., Smith L. J., Kotulla R., Konstantopoulos I. S., Trancho G., Zackrisson E., 2011, MNRAS, 417, L6
- Bastian N., Adamo A., Gieles M., Silva-Villa E., Lamers H. J. G. L. M., Larsen S. S., Smith L. J., Konstantopoulos I. S., Zackrisson E., 2012, MNRAS, 419, 2606
- Blair W. P., Chandar R., Dopita M. A., Ghavamian P., Hammer D., Kuntz K. D., Long K. S., Soria R., Whitmore B. C., Winkler P. F., 2014, ApJ, 788, 55
- Blair W. P., Winkler P. F., Long K. S., Whitmore B. C., et al., 2015, ApJ, 800, 118
- Chandar R., Whitmore B. C., Calzetti D., O’Connell R., 2014, ApJ, 787, 17
- Chandar R., Whitmore B. C., Kim H., et al., 2010, ApJ, 719, 966
- Comaniciu D., Meer P., 2002, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 603
- de Vaucouleurs G., Pence W. D., Davoust E., 1983, ApJS, 53, 17
- Dopita M. A., Blair W. P., Long K. S., Mutchler M., Whitmore B. C., Kuntz K. D., et al., 2010, ApJ, 710, 964

- Fouesneau M., Lançon A., Chandar R., Whitmore B. C., 2012, *ApJ*, 750, 60
- Frey B. J., Dueck D., 2007, *Science*, 315, 972
- Hollyhead K., Bastian N., Adamo A., Silva-Villa E., Dale J., Ryon J. E., Gazak Z., 2015, *MNRAS*, 449, 1106
- Hong S., Calzetti D., Dopita M. A., et al., 2011, *ApJ*, 731, 45
- Kim H., Whitmore B. C., Chandar R., Saha A., et al., 2012, *ApJ*, 753, 26
- Kuntzer T., Tewes M., Courbin F., 2016, *ArXiv e-prints*
- Larsen S. S., 1999, *A&AS*, 139, 393
- Liu G., Calzetti D., Hong S., Whitmore B., et al., 2013, *ApJ*, 778, L41
- Long K. S., Kuntz K. D., Blair W. P., Godfrey L., Plucinsky P. P., Soria R., Stockdale C., Winkler P. F., 2014, *ApJS*, 212, 21
- Ryon J. E., Bastian N., Adamo A., Konstantopoulos I. S., Gallagher J. S., Larsen S., Hollyhead K., Silva-Villa E., Smith L. J., 2015, *MNRAS*, 452, 525
- Silva-Villa E., Adamo A., Bastian N., 2013, *MNRAS*, 436, L69
- Soria R., Long K. S., Blair W. P., Godfrey L., Kuntz K. D., Lenc E., Stockdale C., Winkler P. F., 2014, *Science*, 343, 1330
- Sun W., de Grijs R., Fan Z., Cameron E., 2016, *ApJ*, 816, 9
- Tammour A., Gallagher S. C., Daley M., Richards G. T., 2016, *Monthly Notices of the Royal Astronomical Society*, 459, 1659
- Tully R. B., 2015, *AJ*, 149, 171
- Tully R. B., Courtois H. M., Dolphin A. E., et al., 2013, *AJ*, 146, 86
- Vatturi P., Wong W.-K., 2009, pp 847–856
- Wenger M., Ochsenbein F., Egret D., Dubois P., Bonnarel F., Borde S., Genova F., Jasiewicz G., Laloë S., Lesteven S., Monier R., 2000, *A&AS*, 143, 9
- Williams S. J., Bonanos A. Z., Whitmore B. C., Prieto J. L., Blair W. P., 2015, *A&A*, 578, A100
- Wofford A., Leitherer C., Chandar R., 2011, *ApJ*, 727, 100
- Wolf C., Meisenheimer K., Rix H.-W., Borch A., Dye S., Kleinheinrich M., 2003, *A&A*, 401, 73

This paper has been typeset from a \LaTeX file prepared by the author.