Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 26 August 2016 (MN LATEX style file v2.2)

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. As technology advances, the size of surveys will grow, and efficient means of analyzing the data they produce are necessary. Machine learning methods present an effective way to handle large datasets, and are becoming a popular way to analyze large surveys in astronomy. The purpose of this research is to apply machine-learning methods to the classification of point sources in the nearby galaxy Messier 83 (M83). The Early Release Science Program survey took observations over ten filters in the UVIS channel, from the Wide Field Camera 3 on the Hubble Space Telescope, in the range of Optical to Near Infrared. Mean-shift, Affinity Propagation, and K-means, clustering methods were applied to observations of point sources in the M83 galaxy. Colour-colour combinations were created and clustered. Clustering was performed in two and three dimensions to determine the effectivness of clustering a typical survey. To identify which colour combination was most effective at separating different classes of objects, the strength of the clustering was evaluated and the results compared with independent classification, to determine if objects were correctly identified. The most successful combinations are discussed, and a process outlined for the application of these methods to future surveys. The results of this work will allow astronomers to plan observations that can be used to automatically classify objects in nearby galaxies, leading to more effective surveying, and effecient use of data.

Key words: keywords here

1 INTRODUCTION

Galaxies are complex systems, comprised of numerous components with an enourmous 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 (). **is this WFIRST and other observatories?**

What is the most efficient way to survey the subcomponents 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

† E-mail: pbarmby@uwo.ca
‡ E-mail: akiar@uwo.ca

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 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. Which papers should we use?**

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, and three-colour datasets. In conjunction with published catalogs of galaxy components, we identify the optimal process 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 $\sim$ 12 arcmin () is reasonably well-matched to the camera's field of view. **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 kpc^2 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 $0\farcs0396$ 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; Fouesneau 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).

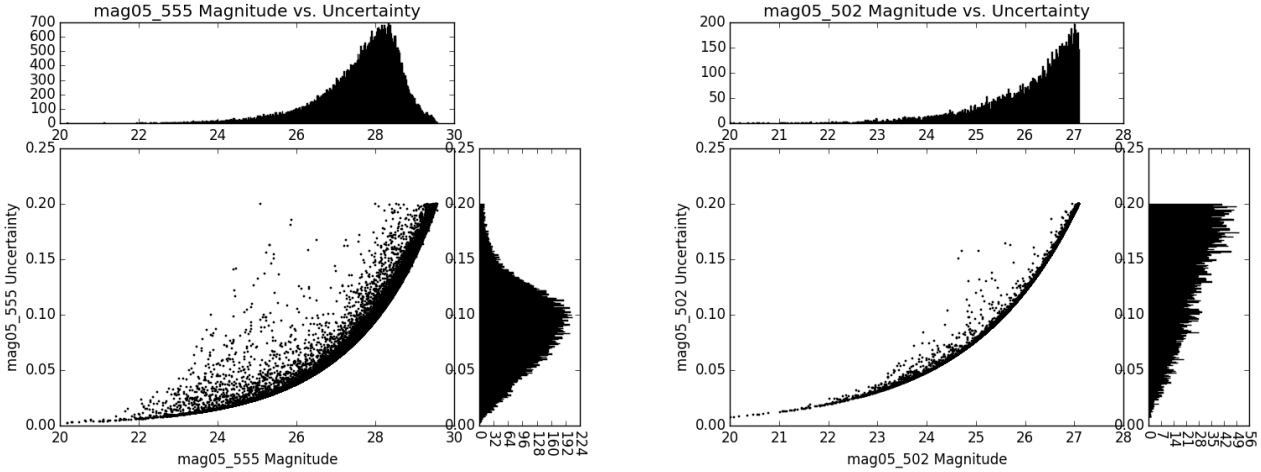
We analyze the catalog produced by Chandar et al. (2010) and made available via **REF**, hereafter referred to as the 'ERS catalog.' The objects in this catalog were detected on a 'white-light' image produced by a weighted combination of the $UBVI$ images. Photometry in 0.5- and 3-pixel radius apertures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

Table 1. Need to fix ion function List of filters from ERS survey with band names and exposure times.

Filter	Name	Exposure time
F225W	Wide UV	1800 s
F336W	U-band	1890 s
F438W	B-band	1180 s
F487N	Hβ	2700 s
F555W	V-band, South field	1203 s
F814W	I-band	1203 s

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 bandpasses, 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.



(a) Broad filter distribution.

(b) Narrow filter distribution.

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

Table 2. List of filter names with the number of objects detected, the number of objects with an uncertainty ¡ 0.2 mag, and the 0.5 px aperture magnitude for which the median uncertainty is 0.2.

Filter	$N_{\rm obj}$	$N_{\rm good}$	$m_{\rm good}$
F225W	57237	15011	25.159
F336W	62192	34129	26.574
F373N	55966	8878	24.752
F438W	66356	48858	28.048
F487N	63812	13335	25.77
F502N	64313	14654	26.424
F555W	67424	65652	30.059
F657N	67782	67634	26.855
F673N	65305	25295	26.284
F814W	67050	59600	27.8699

2.2 Colour Models

Discuss colour models. Figure is broad-band combination with model.

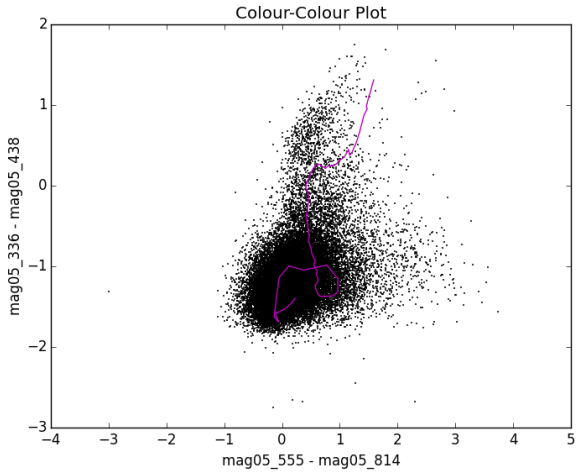


Figure 2. Colour-colour distribution of $V-I$ and $U-B$ with colour model plotted in pink.

3 METHODS

As the size of galactic surveys grows, the number of dimensions available for analysis increases. Clustering methods provide an efficient way of finding structure in high dimentional data by searching for structure in colour spaces that are difficult to visualize. The following techniques were used to cluster the data. All methods were implemented using the *sklearn.cluster* Python package (Pedregosa et al. 2011).

3.1 Mean Shift Clustering

Mean Shift is a non-parametric clustering technique that is based on probability density function estimates at each point in the data. Mean Shift is a very powerful algorithm, but has not been widely used in astronomy. 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 signifcant clusters in a dataset (Comaniciu & Meer 2002). At each point, the algorithm estimates the density around that point using a small sample of objects surrounding the point. The algorithm is based on two components: kernal 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, $\boldsymbol{H}$, which is assumed to be proportional to the matrix $\boldsymbol{H} = h^2\boldsymbol{I}$, with $h > 0$ Vatturi & Wong (2009). The desnity estimator for a multivariate density kernal 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 n are a set of independent d-dimensional data points, h is the magnitude of the bandwidth matrix, $k(x)$ is the profile of kernal $K(x)$, and $c_{k,d}$ is a constant making $K(x)$ integrate to oneVatturi & Wong (2009). 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. Conversly, 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).

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

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

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 kernal 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 interpretted as the center of a significant cluster in the data set, and is used to classify the objects that were shifted towards it.

3.2 Affinity Propagation Clustering

Affinity propagation (AP) is a relatively new clustering technique developed by Frey & Dueck (2007). Here, we will breifly describe the main compoments 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 first 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 eemplars Frey & Dueck (2007). 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)

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

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):

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

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 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 oscilations, 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).

3.3 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 dimentional 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:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} min\left(\|x_n - \mu_k\|^2\right) \quad (8)$$

Each point, x, is then assigned to the cluster center with the lowest distance in equation 8Tammour 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).

4 ANALYSIS

This section will outline colour construction, the clustering process, and the selection of the optimal clustering for each colour combination. Each colour combination was clustered using the same process, and in two and three dimensions.

4.1 Colour Selection

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. Since the average survey consists of four filters, different combinations of four filters were used to construct colours for clustering. Due to the large number of filters available in the ERS data, the combinations had to be narrowed down to a reasonable set. Two types of colour combinations were created: combinations of all broad band filters, and combinations with one narrow band and three broad bands. Additionally, two and three dimensions were considered for clustering, in order to maximize the use of the data in an average survey.

Limiting the number of band combinations reduced the number of dimensions required for clustering. However, in addition to the standard two dimensional colour combinations, clustering in three dimensions was investigated. Three dimensional colour combinations were created based on the combinations listed in Table **??** and Table 4. These tables list the two dimensional colour combinations, the number of objects dected in each colour and their mean uncertainty, and the number of objects detected in the colour combination. In the broad band combinations, three dimensional colour spaces were created by making colours with a common band, either B or V. These bands were selected in order to avoid creating the $U - I$ colour. **PB: Not sure if this next sentance explains why we chose to use a common band clearly. Just trying to say that the colours could be subtracted to transform back into the two dimensional space.** A common band was used in all colours in order to create a three dimensional space that of colours that could be transformed back into the original two dimensional space. In the narrow band combinations, the three dimensional colour spaces were created by making colours with the narrow band common between them. Similar to the broad band spaces, these combinations could be transformed into the original two dimensional space, and act as an extension of the two dimensional distribution. Clustering in three dimensions increased the complexity of the distribution, creating more information for the clustering algorithms to use. However, limiting the dimensionality of the problem to three allowed the analysis to stay within the number of bands in a common survey.

4.1.1 Broad Band Combinations

PB: Next sentance explains what we hope to find in these combos The first type of combination was comprised of the broad band filters: F336W (U), F438W(B), F555W(V), and F814W(I). These filters were used as they had the largest number of detections, and are common in most HST studies. They are indicative of the temperature of stellar objects, which relates to an objects mass and evolutionary state, describing the general state of the objects in the survey. **PB: Could you explain why we didn't use the UVW band or the U-I colour?** The F225W (UVW) filter was not included in this set as it is not a standard filter most surveys. The $U-I$ colour was omitted in the analysis because it was determined that this colour was not physically meaningful. This is because it is unlikely that an object would emit a detectable reading in both these bands due to their distance from one another in wavelength. Using the four filters, the broad band colour combinations created can be found in Table 3. These combinations were created in order to remove any obvious correlation between the colours that could occur by the inclusion of the same band in both colours.

4.1.2 Narrow Band Combinations

PB: Should we explain more about why we choose to do Broad - Narrow? And what objects we hope to find in these combos?

The second set of combinations included the narrow band filters: F373N (O_2), F487N ($H\beta$), F502N (O_3), F657N ($H\alpha$), F673N (S_2), and the broad band F225W (UVW). **Explain why narrow - broad** Colours were created with the narrow bands by pairing them with the broad band which covered their peak in wavelength space. This was done to separate objects that were emission line dominated, from objects that were continuum dominated. Making colours out of a combination of broad and narrow bands ensure that the objects clustered in these bands are physically meaningful, as it is likely that the object would emit in the broad band that contains the narrow band. **Not sure if that is a reason why we chose to construct them that way**. The second colour in each combination was created from two broad bands that did not overlap the first colour in wavelength space. Table 4 lists the narrow band colour combinations used for analysis. The number of objects in the narrow band combination, with the exception of the $H\alpha$ band, is significantly lower than the broad band combinations. These combinations were useful for analysis as their distributions were not as dense as the broad bands, and the clustering algorithms were able to detect interesting structure within them.

4.2 Clustering Process

Clustering was performed using all methods for each colour combination. The following process allowed the investigation

Table 3. Broad band colour combinations and the number of objects detected in each colour, and in each combination, with uncertainties less than 0.2.

Colour 1	Objects	Mean Uncertainty	Colour 2	Objects	Mean Uncertainty	Combined Objects
$U-B$	33523	0.1606 mag	$V-I$	57935	0.1334 mag	28931
$U-V$	33692	0.1429 mag	$B-I$	41413	0.1590 mag	28931
$B-V$	48660	0.1456 mag	–	–	–	–

Table 4. Narrow band colour combinations and the number of objects detected in each colour, and in each combination, with uncertainties less than 0.2.

$Narrow-Broad$	Objects	Mean Uncertainty	$Broad-Broad$	Objects	Mean Uncertainty	Combined Objects
$UVW-U$	14977	0.1539 mag	$B-V$	48660	0.1456 mag	14943
–	–	–	$B-I$	41413	0.1590 mag	14095
–	–	–	$V-I$	57935	0.1334 mag	14098
$U-O_2$	8675	0.1504 mag	$B-V$	48660	0.1456 mag	8657
–	–	–	$B-I$	41413	0.1590 mag	8558
–	–	–	$V-I$	57935	0.1334 mag	8559
$B-H\beta$	13269	0.1493 mag	$V-I$	57935	0.1334 mag	13147
O_3-V	14644	0.1418 mag	$U-B$	33523	0.1606 mag	13390
$H\alpha-I$	59465	0.1495 mag	$U-B$	33523	0.1606 mag	28920
–	–	–	$U-V$	33692	0.1429 mag	29060
–	–	–	$B-V$	48660	0.1456 mag	41317
S_2-I	25185	0.1535 mag	$U-B$	33523	0.1606 mag	14577
–	–	–	$U-V$	33692	0.1429 mag	14586
–	–	–	$B-V$	48660	0.1456 mag	18882

of the effect of all paramters on each clustering technique. This process identified the clustering that was most successful at identifying different segments of objects in the colour space.

4.2.1 Meanshift

Mean-Shift clustering was performed first by estimating the bandwidth paramter with the $estimate-bandwidth$ function in $scikit-learn$ (Pedregosa et al. 2011). This function estimates the bandwidth parameter based on the distances between points in the dataset, and determines if the distribution has high or low variance. Following the initial clustering, the bandwidth was varried and the clustering performed again to determine how sensitive a combination was to the parameter. The bandwidth values were changed on intervals of ± 0.1 or ± 0.05 from the estimated bandwidth value depending on a combination's sensitivity to the parameter. If a combination was very sensitive to bandwidth, then the number of clusters that meanshift would predict would vary greatly over a small range of bandwidth values. This type of combination usually resulted in poor segmentation, as the algorithm would not converge on a number of clusters. However, sensitivity could also be the result of the starting bandwidth estimate. If the original estimate was in an unstable bandwidth interval, then the hierarchy would reflect that, and the testing of multiple bandwidth values could result in convergence.

4.2.2 Affinity Propagation

Affinity Propagation clustering was performed after Meanshift. Affinity Propagation requires two parameters, the preferences, and the damping factor, outlined in Section 3.2. The initial clusterings were performed under two independent conditions. First, the preferences were set to the median value of the similarity between data points, and the damping factor was kept at the *scikit-learn* default value of 0.5. Second, the preferences were set to the minimum value of the similarities, and the damping factor kept at the default value of 0.5. These clusterings resulted in a segmentation with over 100 clusters in multiple colour combinations, which was clearly not meaningful.

Following the two initial conditions, the preferences were set to 10% of the number of objects in the data set, and the damping factor was set at 0.95. With these parameters, the clusterings varried significantly over differnet colours. The clusterings were repeated by varrying the damping factor and the preferences to try and reveal a trend in the parameters, but the algorithm was too sensitive for this size of dataset. Following the initial tests of Affinity Propagation, it was determined that this clustering method was not effective. Due to the number of computations required for the calculation of the messages passed between points on each iteration, the algorithm was very sensitive to the input parameters, and did not produce meaningful clusterings. The algorithm is effective for small and medium sized datasets, and was able to create some reasonable clusters when the uncertainty limit was set at $0.1mag$, which reduced the number

of objects significantly. After multiple clusterings, a systematic way of determine the correct number of clusters could not be determined, and the algorithm was not used further.

4.2.3 K-Means

K-Means clustering was performed last. The first clustering was performed using the number of clusters determined from the initial clusterings by Meanshift. Next, K-Means was performed with $K = \pm 4$ from the original clustering. This method of clustering was similar to the Meanshift approach, as it revealed how the dataset reacted to different values of K. K-Means was the most efficient algorithm of the three, as it produced clusterings quickly, and always produced clusters of reasonable size.

Each K-Means clustering was checked by plotting the sum-of-squares value for each value of K. As K increases, the inertia value decreases as the inertia value represent the total distance between the total distance between the points of each cluster. The value decreases with K as the total distance in each cluster decreases as more clusters are introduced. This distribution was used to check the clustering and ensure that the clusterings were successful.

Following the inertia test, the cluster centers were tested by running K-Means for 40 trials with the same value of K. This test was run to determine if the clusterings were stable as K-Means is initiallized randomly, and the clusters produced can depend on the starting position. It was clear that each initialization found different clusters first depending on the starting point. Despite the random initialization the final cluster centers did not change. A strong clustering was found when the variance in cluster centers was not large **quantify**. If the cluster centers vary across multiple initializations of K-Means, the clustering would not be reliable, and the combination would not be considered for analysis.

4.3 Characterizing the Clusterings

Determining the strongest clustering was the most difficult task of the analysis. Selecting the optimal clustering can often seem arbitrary, as no "right" answer is obvious. In order to characterize each clustering, a variety of metrics and statistics were calculated to evaluate each method. The relationships between a variety of clustering parameters were investigated to try and determine how they indicated the strongest clustering. Since the performance of the algorithms was directly related to the parameters used as input, those relationships were critical for characterizing the clustering. The objects in each cluster were then found in the white-light image of M83, to determine if there was a relationship between the objects assigned to the same cluster and their spatial position. **Not sure if we want to mention the colour models here** Finally, colour models were created and imposed on the cluster distribution to determine if the segmentation agreed with a model. **Need more on why the models were used**.

4.3.1 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:

$$SilhouetteScore = \frac{b - a}{max(a, b)} \tag{9}$$

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. The score was used in two ways. First, the average score was calculated for a given clustering. This calculated the average score across all data in the sample. Next, the average score for each cluster was computed. The average cluster score evaluates the strength of a given cluster within a clustering. This metric allowed each cluster to be evaluated individually, and revealed which clusters were responsible for the average score of the clustering. Additionally, the average score allowed seemingly arbitrary clusters to be evaluated. If the segmentation did not seem meaningful, the average score would reveal if the cluster was isolated and compact. This revealed the significance of clusters that could have been viewed as noise or outliers. **This section needs more explaination**.

Struggling to explain why that is how the score works. Ideally, the silhouette score for the entire clustering should peak near the center of the distribution indicating the optimal clustering. High scores where the number of clusters is low often do not reflect the structure of the distribution, while high numbers of clusters are often imposed on the data as a result of the specified parameters. This is often not the case, as seen in Figure 3, which shows the distribution of the silhouette score against the number of clusters.

For the K-Means clusterings, the score does not peak in the center of the distribution. Instead of selecting the clustering with the highest score, the optimal clustering is found where the relation begins to elbow; between 4 and 5 clusters. This clustering is selected because any increase in K after this point does not affect the score, and does strengthen or weaken the clustering. This means that the algorithm has found the balance between the natural clusters in the distribution and artificially segmenting the data.

The distribution of score as a result of the Meanshift algorithm does not follow the same pattern. The silhouette score was not as successful at describing the strength of Meanshift clusterings as Meanshift often created one large cluster and several smaller ones, which is not considered strong by the score. In order to determine the optimal Meanshift clustering, other relationships were investigated.

The relations of the bandwidth parameter were investigated for each Meanshift clustering. Figure 4 shows the relation between the score and number of clusters for the three dimensional $U - O_2$ and $B - I$ combination. The blue dots represent the Meanshift clusterings.

It is clear that an optimal clustering cannot be determined from this relation, as the score decreases linearly with the number of clusters Meanshift predicts. The Mean-Shift scores do not follow a similar distribution as K-Means, as the accuracy of Meanshift is related to the bandwidth parameter, seen in Figure 5. The optimal Mean-Shift clustering was chosen by finding the bandwidth where the relation between the bandwidth and number of clusters reached an elbow, or where the relation between bandwidth and silhouette score elbowed. In both panels of Figure 5 that the

[H]

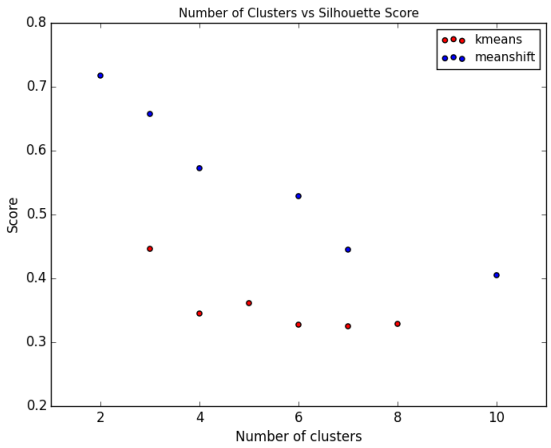


Figure 3. Distribution of the silhouette score as a result of the number of clusters imposed for the $UVW-U$ and $V-I$ colours. The *blue* points are the scores of Meanshift clustering, *red* points are scores of K-Means.

[H]

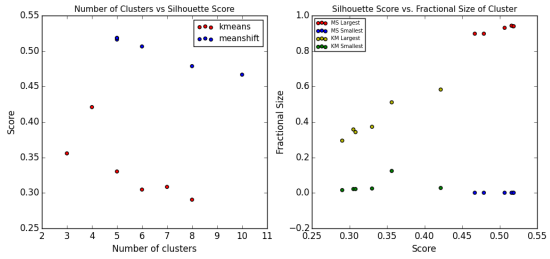


Figure 4. Distribution of the silhouette score as a result of the number of clusters imposed for the three dimensional $U-O_2$ and $B-I$ combination. The *blue* points are the scores of Meanshift clustering, *red* points are scores of K-Means.

bandwidth parameter predicts five clusters. Both distributions elbow at the same bandwidth interval, maximizing the score and predicting the same number of clusters for bandwidth valeus following the elbow. The bandwidth parameter was the primary indicator of the optimal Meanshift clustering. If a trend could not be found with the bandwidth parameter, then the silhouette score and number of clusters was investigated.

[H]

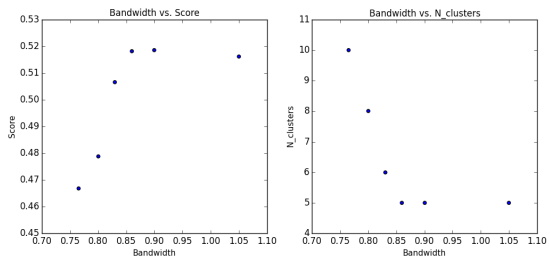


Figure 5. Distribution of the silhouette score as a function of bandwidth, and the distribution of the number of clusters as a function of bandwidth.

4.3.2 Cluster Statistics

Various statistics were calculated 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 helped 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 clusters whose averages varried significantly from the cluster centers were disgarded.

I'm not sure that this paragraph describes why we calculated the fractional size. The fractional size of each cluster was also calculated and described the distribution of objects between clusters. 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.

5 RESULTS

The section will outline the major results of the clustering. It is focused on the broad band clustering, and two combinations from the narrow band colours. The first narrow band combination is the most successful narrow band clustering, and the second is the least successful. A complete discussion of all the combinations can be found in Appendix 1 **Make appendix here**. **Do we need an appendix with all of that discussion?**

5.1 Broad - Broad Band Combinations

The broad bands were clustered using the colours found in Table 3. In both two and three dimensions, the $U-B$ and $V-I$ combination was clustered more effectively than the $U-V$ and $B-I$ combination. Both combinations had similar silhouette scores, however, the structure of the $U-V$ and $B-I$ combination was not identified by either clustering method. The $U-V$ and $B-I$ did not have clear branches of different objects, and the clustering methods segmented the distribution by colour. **not sure if we need more to describe why we chose the $U-B$ and $V-I$ combo**

5.1.1 2-Dimensions

Figure **??** shows the distribution of objects in the $U-B$ and $V-I$ combination. **I think the discussion of what we are looking for should go in section 4 with the colour combinations, but the distribution description should stay here** Three main features are apparent. The first feature is the concentration of obects at $-1U-B$ and $0V-I$. These objects are middle-aged objects, as their colours are neither blue nor red. **Not sure if thats right**.

[H]

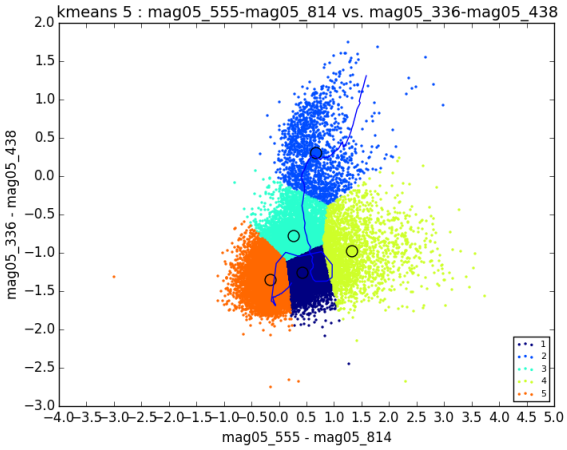


Figure 6. Colour-Colour distribution of the $U-B$ and $V-I$ colours, clustered using K-Means with $K=5$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

[H]

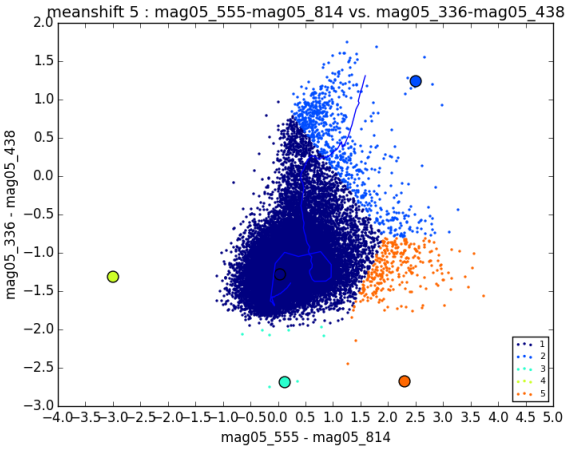


Figure 7. Colour-Colour distribution of the $U-B$ and $V-I$ colours, clustered using Meanshift with $h=0.6$ producing five clusters. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

The second feature is the group of objects that spread above 1 in the $V-I$ colour, while staying centered on -1 in the $U-B$ colour. **Can't remember which type of objects these are - younger objects since blue in $U-B$ but bright in I band since red in V - I** The last feature is the branch of red objects spanning from -0.5 to $+2$ in the $U-B$ colour. These objects are likely young objects that lie in the back of the galaxy or behind dust lanes as they are reddened in the $U-B$ colour, but are relatively neutral in the $V-I$ colour. **I think thats right**.

The strongest K-Means clustering was selected at $K=5$ (Figure 6) as it is the point where the silhouette score began to plateau. K-Means was able to identify the branch of red $U-B$ objects at low values of K. As K increased, the algorithm was also able to identify a group of objects that are red in $V-I$, and blue in $U-B$.

The boundaries of each cluster approximately represent magnitude intervals in each colour, identifing groups of objects that are relatively blue and red, seperating them from objects with extremes of either colour. The clusters also trace sections of the modelled colours. The centers of clusters 2 and 5 line up well with the models, indicating that these clusters could be representative of different ages of objects in this distribution. **Not sure if that was the right implication or if there is something else to their alignment**.

The clusters from the two dimensional clustering show distinct locations in M83. The branch of red $U-B$ objects found in Cluster 2 of Figure 6 are objects that lie loosly around the spiral arms. These objects are generally not found in the concentrated regions of the arms, and lie to the left and right of these areas. On the whitelight image, these objects appear to be isolated, dim point sources that could be lying behind clouds of dust, in the back of the galaxy, or on their own outside the arm. **PB: What could these objects actually be? Young star clusters, clouds, background sources/galaxies?**

The branch of red $V-I$ objects found in Cluster 4 of Figure 6 are objects that lie primarely in the dense regions of the spiral arms, with a few objects lying in the nucleus, and in the region south of the nucleus. These objects also appear to be quite dim point sources, or no detection in the whitelight image. This could mean that the objects are some form of cloud, nebula, or background galaxy. These objects are interesting as their $V-I$ colour stretches to values over 3, indicating very red emission. **PB: is this all true...**

Despite the seemingly arbitrary segmentation of clusters 1 and 3, the objects seem to inhabit different regions of M83. These two clusters generally trace each others location in the galaxy. Cluster 1 is generally confined to the denser regions of the arms, while cluster 3 fills in the areas between the objects in cluster 1. Cluster 1 objects appear to be bright point sources on the whitelight image, while cluster 3 objects are dim or non-existant. **PB: what could these objets be?**.

Despite their difference in colour, cluster 5 also traces clusters 1 and 3 around the galaxy. Cluster 4 objects stay mainly concentrated in the spiral arms, filling in the dense area in the interarm regions. Cluster 4 objects are not as apparent in the interarm region as clusters 1 and 3, but still appear in denser areas. **PB: what could these objects be?**

The strongest Meanshift clustering was selected as the point where the relation between bandwidth and the number of clusters plateaued, and can be seen in Figure 7. Its bandwidth was $h=0.6$, which produced five clusters.

This clustering failed to identify the branch of red objects in Figure **??**. The cluster that contained the some of the objects from the red branch also contained objects which belonged to the red branch of the $V-I$ colour, which are objects with different properties. Clusters 3 and 5 contain only three total objects. These are not meaningful clusters, as it is unlikely that these objects are significant detections. The clusters and their centers do not align with the model colours, indicating that Meanshift was unable to identify different segments of the stellar lifecycle.

Figure 8 shows the comparison between the clustering methods.

[H]

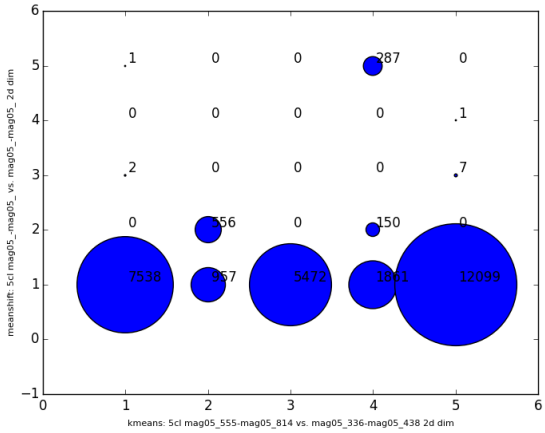


Figure 8. Comparison of object cluster assignment between Meanshift and K-Means clusterings.

[H]

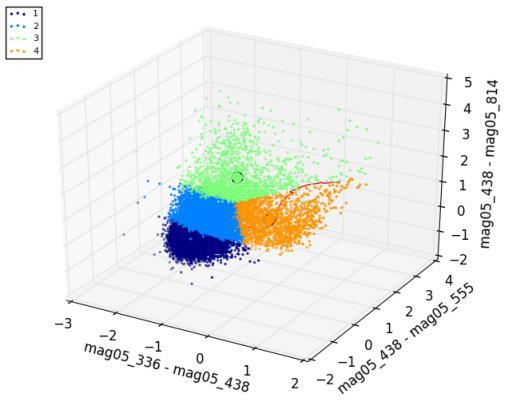


Figure 10. Colour-Colour-Colour distribution of the $U-B$, $B-V$, and $B-I$ colours, clustered using K-Means with $K=4$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

The axes are the respective clusterings, and the size of the bubble is related to the number of objects that belongs to both clusterings. It is clear that there is little agreement between the cluster assignment between the two methods. Cluster 2, the red branch of objects, is the only cluster where a significant portion of objects were assigned to the same cluster. **Implication** K-Means distributes objects far more evenly between clusters, and creates more meaningful segmentation as seen in Figure 6. Desipte K-Means' success in identifying different features of the distribution, Meanshift's silhouette score was 0.4037, and K-Means was 0.3577. This difference is a reflection of the separation of the individual clusters within each clustering. The score is biased towards clusterings that create one large cluster and several smaller ones as it measures the likelihood of an object being placed in cluster other than the one it was assigned to. It is clear that the Meanshift clusters are well separated, as clusters 2 and 5 could be viewed as outliers. **Trying to explain that the score is biased towards one large cluster with several smaller clusters**. However, these clusters to not contain all of the objects from the features they identify and the clustering is not more accurate than the K-Means segmentation. Due to this comparison, the K-Means segmentation was selected as the optimal clustering for this combination, and was used for the completion of the results.

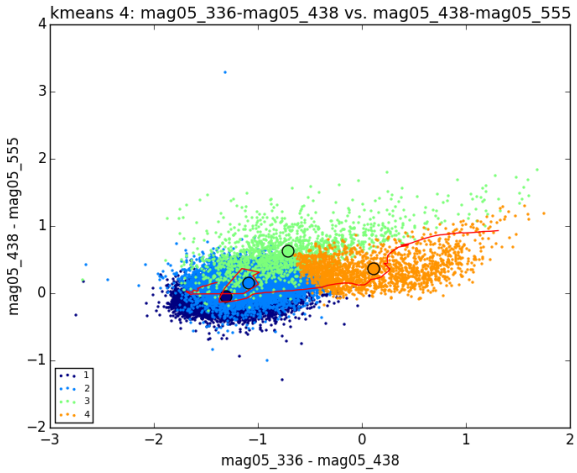
5.1.2 3-Dimensions

Three dimensional clustering was performed with colours created from the $U-B$ and $V-I$ two dimensional combination. The colours used were: $U-B$, $B-V$, and $B-I$. The distribution has three clear features. **Not sure if we need to describe these objects in the same detail as 2d** The first feature is a group of objects that are neutral in the $B-V$ and $B-I$ colours, but are quite blue (-1 and beyond) in the $U-B$ colour. The second feature is a branch of objets that is neutral in the $U-B$ and $B-I$ colours, but quite red ($0-+4$) in the $B-V$ colour. The last feature is a group of objects that are quite blue (-1 and beyond) in the $U-B$ colour, and quite red in the $B-V$, and $B-I$ colours ($0-+5$).

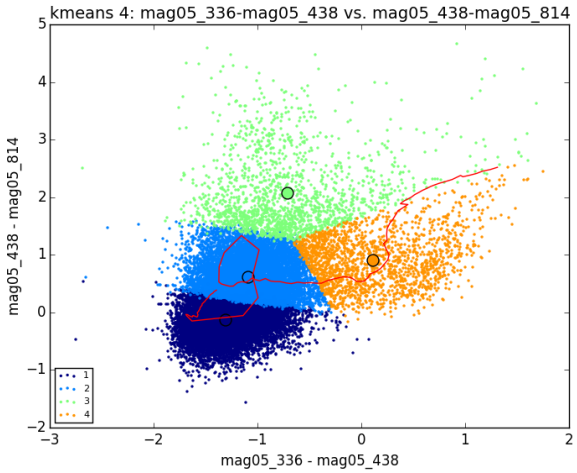
K-Means was able to identify objects that were in specific regions of the colour space. The optimal K-Means clustering was found at $K=4$ (Figure 10), as the score peaked, and the red branch of objects was identified. Figure 9 displays the projection of the three dimensional clustering into each of its two dimensional components. It is clear that the three dimensional clustering was driven by the $B-I$ vs. $U-B$ distribution, as the clusters were almost completely seperated in that space. The clusters overlapped extensively in the $B-V$ vs. $U-B$ space. **PB: Do you know why this would be? Would it be from the range of colour in the B-I dimension?** This could be a result of the colour distribution in the $B-I$ space, as the range of object colour in this dimension is much larger than in the $B-V$ colour. This relationship was found through all colour combinations.

The three dimensional clustering illustrates how the algorithm is able to identify two distinct branches of objects beyond the dense center of the distribution. Cluster 4 is a branch of objects that is quite red in both the $U-B$ and $B-V$ colours, but is neutral in the $B-I$ colour. Cluster 3 is a branch of objects quite blue in the $U-B$ colour, but red in the other two. **what is the implication of this?** The algorithm then segments the dense section of the distribution between the bluer and redder objects in the $B-V$ and $B-I$ colours. These two clusters are not ideal, as the dense portion of the distribution could be argued as the same cluster. However, at $K=3$, the branches were not identified, so the segmentation of the dense region is necessary. The clusters projected into the base colours can be seen in Figure **??**, where the branches of objects are clearly identified.

The cluster centers of clusters 1, 2, and 4, match the model colours predicted. The red branch of objects traces the older stellar population, while clusters 1 and 2 trace the young and intermediate ages of the population. Cluster 3 is positioned near the loop in the stellar model, however there is a disagreement in the $U-B$ colour. **Any other implications of the model matching?**



(a) $U-B$ vs. $B-V$ Distribution.



(b) $U-B$ vs. $B-I$ Distribution.

Figure 9. $U-B$ vs. $B-V$ and $B-I$ projections from three dimensional clustering.

[H]

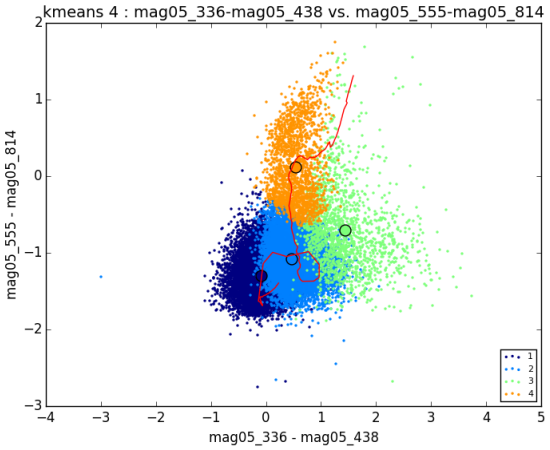


Figure 11. Colour-Colour distribution of the $U-B$, and $V-I$ colours, projected from the three dimensional clustering using K-Means with $K=4$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

[H]

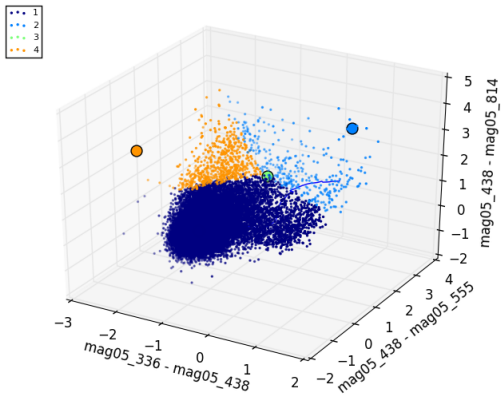


Figure 12. Colour-Colour-Colour distribution of the $U-B$, $B-V$, and $B-I$ colours, using Meanshift with $h=0.75$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

In the K-Means clustering (Figure **??**), clusters 4 and 3 follow similar patterns found in two dimensions. These clusters are the red $U-B$ branch of objects and the red shelf of $V-I$ objects that span the whole range of $U-B$ colour. Cluster 4 objects are generally located in the spiral arms, but their location is not concentrated, and they are fanned across the entire width of the arms. This is in agreement with the cluster that identified the red branch in two dimensions, however, three dimensions seems to include more objects that could be considered redder than the rest of the distribution. Cluster 3 objects trace cluster 4, but are in the dense regions of the spiral arms, and do not fan out. **PB: What could these be?** Clusters 1 and 2 in three dimensions do not provide the same detail as two dimensions as they lump all the objects in the center of the distribution into two groups. There are no patterns in object location in these two clusters.

Meanshift was also able to produce interesting clusters in three dimensions. The optimal clustering was chosen at $h=0.75$ which produced 4 clusters. This clustering was the peak score, but was not the number of clusters for most intervals of bandwidth. Most bandwidth values predicted 6 clusters, however, the score and cluster seperation at those intervals was poor. Figure 12 shows the three dimensional clustering at $h=0.75$.

Meanshift identified two groups of objects that lie above the dense area of the distribution in Figure 12. Figure **??** shows the projection into the original space, where the cluster location is easier to identify. The clustering identified two groups of ojbects which are quite red in the $V-I$ colour, but have different $U-B$ colours. Cluster 4 is blue in the

[H]

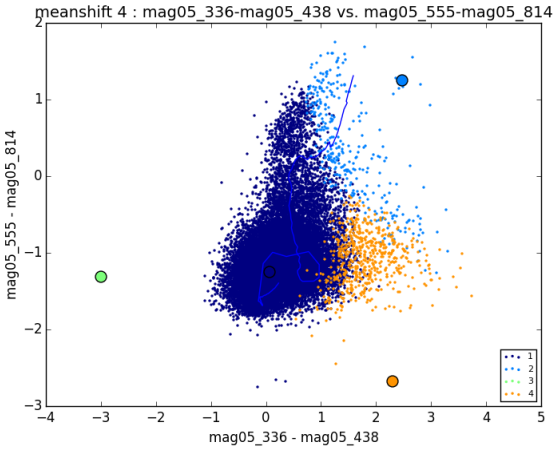


Figure 13. Colour-Colour distribution of the $U-B$, and $V-I$ colours, projected from the three dimensional clustering using Meanshift with $h = 0.75$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

$U-B$ colour while Cluster 2 is redder. This identification highlights Meanshift's ability to find outliers in the distribution, as it does not pick out the large branch of objects that are red in $U-B$. However, these two clusters combined only hold 3% of the objects used for clustering. These clusters would have been considered insignificant, however, they are also found at $h = 0.55$. This means that the clusters are identified at various bandwidth intervals, which means they are likely significant clusters. **Add cluster statistics table**

The outlying clusters identified by Meanshift in three dimensions are located in interesting areas of M83. There is almost no overlap between the locations of the objects in these two groups. Cluster 2 objects are fanned out through the spiral arms, and in the core of the galaxy. These objects appear to be dim point sources on the whitelight image, and could be background objects. Cluster 4 objects are also found in the spiral arms, however, they are almost only concentrated in the dense areas. These objeccts are found in the same regions of the arms as cluster 2, but they are not close to one another. This could indicate that these two classes of objects are different physical objects in M83. **Not sure if this is the right result of their locations** Lastly, cluster 3, a single object, does not appear to be a meaningful object. It does not seem to be detected in the whitelight image, as it is in an area without a clear, independent point source. Since this object has very blue colours, it is not clear what it could be, and may be noise. **Not sure if this is true!**

Comparing the two clusterings resulted in a similar comparison to the two dimensional clustering. No significant distribution between clusters was apparent, and all the K-Means clusters were placed in Cluster 1 of the Meanshift clustering. However, since the three dimensional Meanshift segmentation was more meaningful than the two dimensional clustering, it was kept for further analysis.

[H]

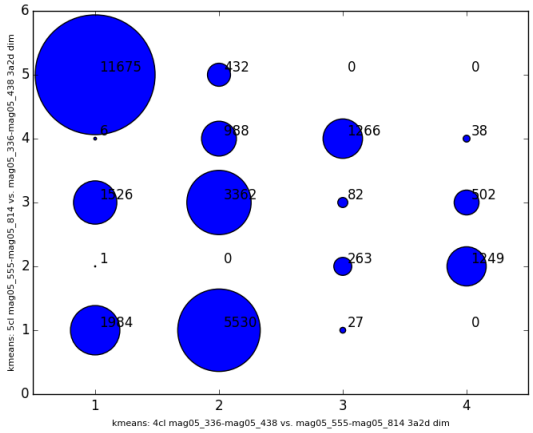


Figure 14. Comparison of object cluster assignment between K-Means in two dimensions and K-Means in three dimensions.

5.1.3 Two and Three Dimension Comparison

Figure 14 shows the comparison between the two and three dimensional clusterings.

There is significant overlap in each of the five clusters from the two dimensional clustering. This shows agreement between the two clusterings for which objects belong in the same cluster. It is difficult to compare clusterings with a different number of clusters, but it is clear that Cluster 2 in Figure 10 is the cluster that is used to add the additional cluster in two dimensions. With this agreement, it is reasonable to assume that both clusterings are strong, and since the three dimensional clustering had the higher score, it was selected as the optimal clustering.

5.2 U - OII vs. B - V: Successful Clustering

The U - OII combination was clustered with the B-V, B-I, and V-I colours using Meanshift followed by KMeans. **More about what we are looking for in this combination** The $U-OII$ vs. $B-V$ combination was selected for discussion as its K-Means score was the highest in two and three dimensions. Meanshift did not perform well in two dimensions, but in three dimensions it was able to identify structure in the distribution like K-Means.

5.2.1 2-Dimensions

K-Means performed stronger than Meanshift in two dimensions, and was selected for most of the results. **Need some introduction to the two dimensional clustering**

The K-Means algorithm produced more reliable results, as it produced clusters of relatively similar sizes. The silhouette score elbowed at $K = 5$, and was selected as the optimal clustering, see Figure 15. At $K = 5$ K-Means segmented the data based on integer colour values, and identified the section of data that was significantly red in the U-OII colour. The centers of clusters 3, 4, and 5 align well with the model colours. These clusters identify different ages of the the model stellar population. However, clusters 1 and

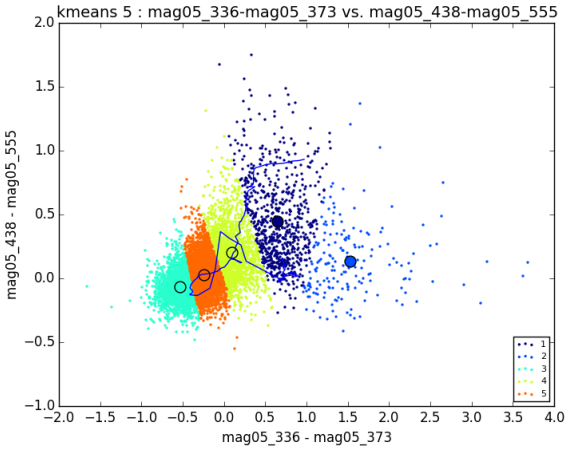


Figure 15. Colour-Colour distribution of the $U - O_2$ and B-V colours, clustered using K-Means with $K = 5$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

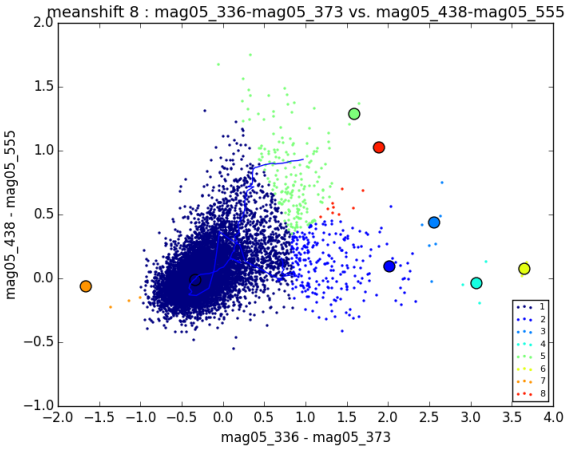


Figure 16. Colour-Colour distribution of the $U - O_2$ and B-V colours, clustered using Meanshift with $h = 0.33$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

2 do not align well with the model, despite identifying seperate parts of the distribution.

Need help identifying these objects The K-Means clustering was the only segmentation considered on M83, as Meanshift's segmentation was not meaningful. Cluster 2, the red branch, seemed to be a combination of dim point sources and objects in the back of the galaxy or behind clouds. A significant portion of cluster 1 objects were found in the core of the galaxy. The remainder of the objects were distributed throughout the areas surrounding dense regions in the spiral arms, and consisted of bright point sources, and point sources behind clouds or in the back of the galaxy. **Not sure what the significance of that is since they are red in both colours** Cluster 3, the objects blue in both colours, are consentrated in the dense regions of the spiral arms. The few objects that are located in the core are concentrated in the very center, and there are no objects located in the dust lanes between the core and the spiral arm. Cluster 5 objects are located in the dense regions of the spiral arms, but are not concentrated in the centers of these areas like the objects in cluster 3. Similar to cluster 5, cluster 3 objects are located sparsly along the dense regions of the spiral arms, and seem to trace the position of the objects in cluster 3.

This combination was more sensitive to bandwidth selection than others. With bandwidth $h = 0.2$, 32 clusters were produced, while $h = 0.4$ produced 3. Due to this sensitivity, the bandwidth hierarchy was created on much narrower increases in h, to produce more meaningful clusters. After producing the narrow hierarchy, the meanshift algorithm predictad a range of clusters from 3 to 13. In each clustering, the algorithm did not seem to segment the data significantly, as 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 16. This clustering segmented the data into three main groups. Cluster 1 was the densist region of the distribution, and clusters 2 and 5 were 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. Additionally, the outer areas of the arms were segmented into their own clusters. These clusters are not meaningful as these objects would have similar properties to each arm. This segmentation was the strongest two dimensional Meanshift candidate. However, since it was still poor, the two dimensional Meanshift clustering was not considered for the rest of the analysis.

Not sure if we should talk about locations in M83 at all

5.2.2 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 U-OII, OII-B, and OII-V. The performance in three dimensions was stronger for all clustering parameters. The clustering algorithms were able to identify a large branch of objects that was red in the U-OII colour, and very blue in the other two colours, see Figure 17. 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 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. The score peaked at $K = 3$ (Figure **??**), which was much higher than any other value of K. This was caused by the clear segmentation of the branch of blue objects in three dimensions. When projected 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 18. Additionally, all three cluster centers align with segments of the model colours. Cluster 1 is a branch of young stars, and each cluster pro-

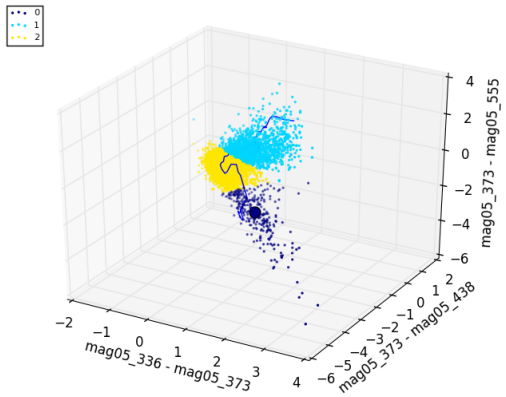


Figure 17. Colour-Colour-Colour distribution of the $U - O_2$, $O_2 - B$, and $O_2 - V$ colours, clustered using K-Means with $K = 3$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

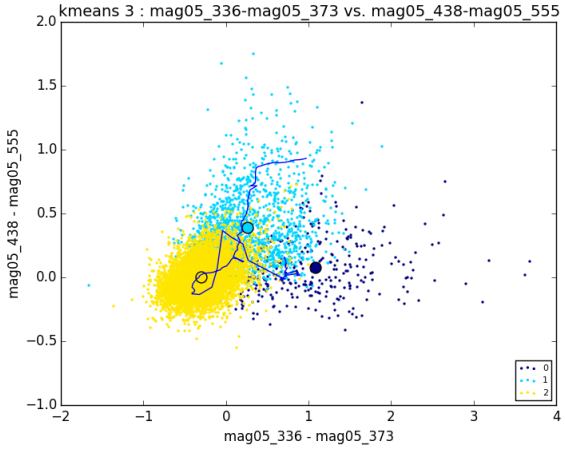


Figure 18. Colour-Colour distribution of the $U - O_2$ and B-V colours, projected from the 3D clustering using K-Means with $K = 3$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

gresses in the age of the stellar population. This pattern is even more pronounced for higher values of K, but the clustering accuracy is reduced significantly as K increases.

Clusters 1 and 2 were almost entirely objects in the back of the galaxy or behind clouds. The objects in cluster 3 were located in the densist areas of the spiral arms. These locations agree with the two dimensional clustering, however, clusters 1 and 2 are much more pronounced in three dimensions. The three dimensional clustering was stronger as the objects in each cluster were not similar to the objects in the other two clusters, which was not the case in two dimensions.

The optimal meanshift clustering was not as apparent in three dimensions. The bandwidth did not converge at a number of clusters, or the score. However, a weak plateau was found when 6 clusters were produced, at $h = 0.62$. This bandwidth value was the point before a steep drop in the relation between bandwidth and score, and was selected for the optimal clustering. Figure **??** shows the three dimensional clustering. The three dimensional clustering was driven by the $O_2 - B$, and $O_2 - V$ colours equally, contrary to the pattern found in the broad band clustering. This means that despite the different range in colours, both colours had attributes to add to the distribution, and the objects in this narrow band have elements in both broad bands. **Not sure if that is the implication**.

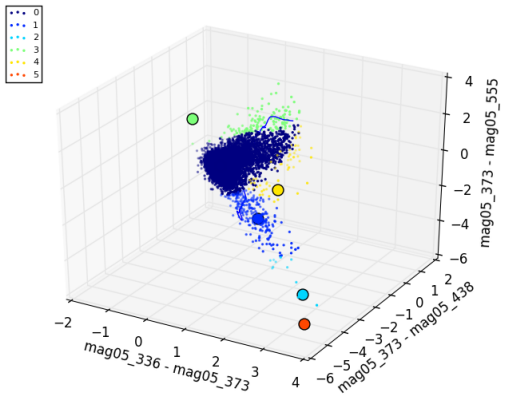


Figure 19. Colour-Colour-Colour distribution of the $U - O_2$, $O_2 - B$, and $O_2 - V$ colours, clustered using Meanshift with $h = 0.62$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

Meanshift identified groups of objects that did not lie in the dense center of the distribution. Several clusters can be seen that are well defined in three dimensions, but are not apparent in the two dimensional projection (Figure **??**). The clusters that seem obvious in three dimensions overlap significantly in two dimensions. However, this clustering is stronger than the two dimensional clustering. **Need more implications**

Not sure if we should talk about locations in M83 at all

Figure 21 shows the comparison between the two and three dimensional K-Means clusterings. There is clear agreement between cluster 1 in three dimensions and cluster 2 in two dimensions, as these are the extremely red objects in the $U - O_3$ colour. Clusters 2 and 3 in three dimensions and 4 in two dimensions do not agree. This can be seen in the apparent overlap between clusters 2 and 3 in Figure 18. This illustrates the additional information gained from clustering in three dimensions, as clusters 2 and 3 are clearly seperated in Figure 17.

5.3 $O_3 - V$ vs. U - B: Unsuccessful Clustering

The $O_3 - V$ colour was clustered with the U - B colour in two dimensions and the U-O_3, and B-O_3 colours in three dimensions. The clustering methods were not successful at identifying structure in this combination in both two and three dimensions. This was a result was a combination of the colour distribution, and the ability of the clustering methods to identify structure in the distribution. In two dimensions,

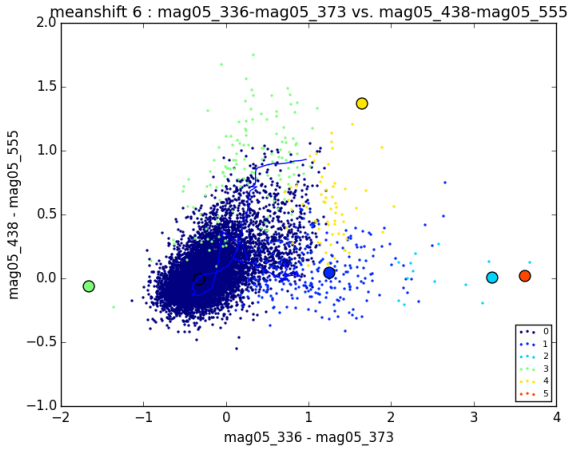


Figure 20. Colour-Colour distribution of the $U-O_2$, and $B-V$ colours, projection from the three dimensional clustering using Meanshift with $h = 0.62$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

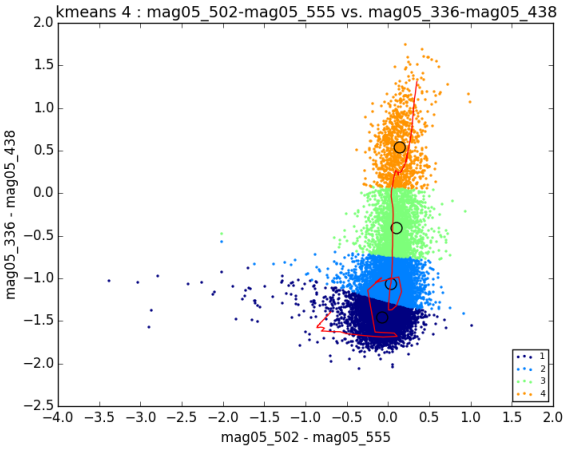


Figure 22. Colour-Colour distribution of the O_3-V and U - B colours, clustered using K-Means with $K = 4$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

[H]

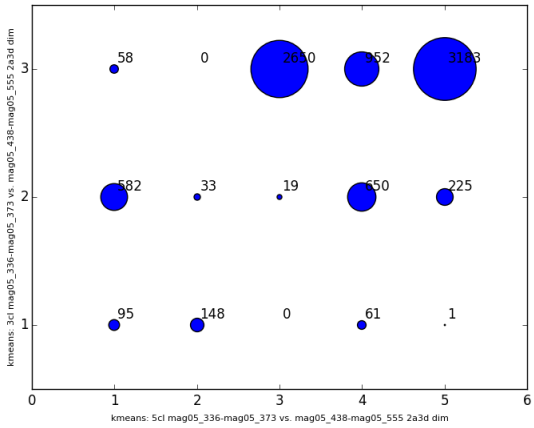


Figure 21. Comparison of object cluster assignment between K-Means in two dimensions and three dimensions.

the majority of the objects had $O_3 - V$ colours between -0.5 and $+0.5$. This distribution did not provide enough variance for the clustering algorithms, and caused the poor segmentation.

5.3.1 2-Dimensions

The $O_3 - V$ and $U - B$ combination did not produce a distribution with as much information as other colours. Two main features were visible in the distribution. The first feature was the majority of objects centered at 0 in the $O_3 - V$ colour. These objects only varried in the $U - B$ colour. Even with that variance, the range in $U - B$ colour was only between -1.5 and $+1.5$ colour magnitudes. The second feature was a small number of objects that were extremely blue in the $O_3 - V$ colour. These objects had colours ranging from -1 to -3.5 magnitudes. However, the number of objects in this branch was significantly less than that of the main branch.

The K-Means two dimensional clustering segmented the data into sections of U-B colour. As K increased, K-Means was able to identify a branch of objects that are bluer in both colours. The segmentation in the colour-colour space translated into the U-B vs B CMD. Each clustering segmented the CMD by U-B colour, and the cluster of bluer objects appears to be a group of objects bright in the U band with a brightest B magnitude of approximately 25. The K-Means score begins to plateau at K=4, however K=5 has a slightly higher score than the rest of the plateau as this is the first clustering to identify the branch of bluer objects. Despite the plateau, the clustering scores are not as high as other combinations, and the segmentation appears to be arbitrary. K-Means optimal clustering was $K = 4$ (Figure 22), which did not identify the clear branch of blue objects. Despite the poor segmentation, the cluster centers aligned well with the colour models, tracing them through the distribution. The clusters selected different ages of the stellar population, but the segmentation was still poor, and a result of the distribution.

The K-Means clusters did not show distinct patterns through M83. Since the segmentation does not pick out specific structure in the distribution, it was unlikely that the spatial relationships were as significant as other colour combination. This was confirmed as there did not seem to be a difference in location between the objects in cluster 1 and 4 in Figure 22. The objects Meanshift clusters 2, 3, 4, 5, 7, and 8 identify objects that are spread out through the spiral arms, without any objects in the core. Despite their locations, these clusters combined consisted of less than 1% of the data in the combination, and the other clusters did not create meaningful patterns.

The Meanshift clustering did not seem to create meaningful segmentation. The Meanshift parameters did not display the same patterns as other combinations. The Meanshift score did not plateau at any number of clusters, and the large center cluster contained almost all of the objects in

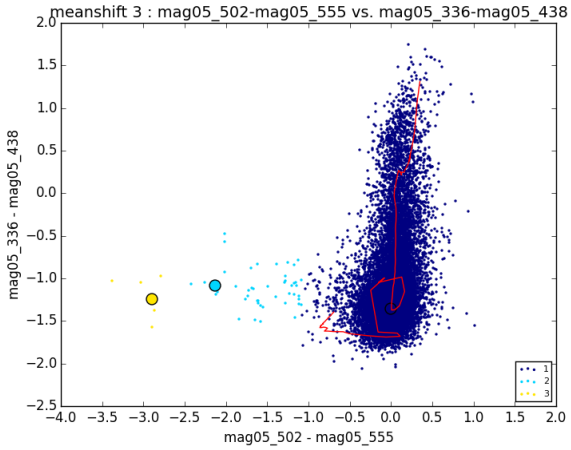


Figure 23. Colour-Colour distribution of the $O_3 - V$ and U - B colours, clustered using Meanshift with $h = 0.4$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

[H]

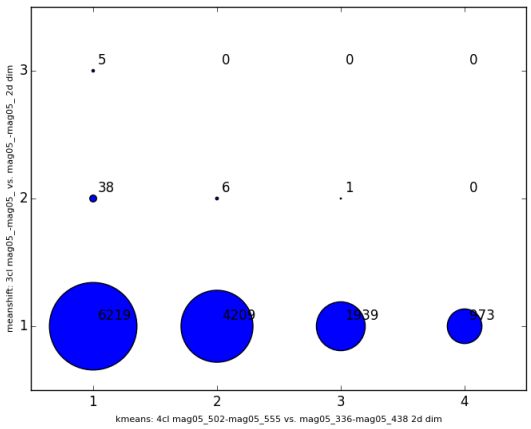


Figure 24. Comparison of object cluster assignment between K-Means at K=4 and Meanshift at h=0.4 in two dimensions.

each segmentation. As the bandwidth was increased, Meanshift increased the segmentation amoung outliers, increasing the number of clusters which only contained one object The bandwidth relationships were not found in this combination, making the optimal clustering selection difficult. The difference in score between $h = 0.39$ and $h = 0.4$ was the largest of any interval, and $h = 0.4$ was selected as the optimal clustering with three clusters (Figure 23). Meanshift was able to identify a small group of objects that were blue in B - V colour (cluster 2), but this cluster represented less than 1% of the objects in the clustering.

Figure 24 shows the comparison between the optimal K-Means and Meanshift clusterings. There is no agreement between the two methods on which clusters the objects belong to. This is because Meanshift was only able to identify the extremely blue $O_3 - V$ objects as an independent cluster. This is a result of the distribution of the $O_3 - V$ objects. Since the density of the objects around 0 in the $O_3 - V$ colour is so high, Meanshift is not able to segment the objects in a meaningful way.

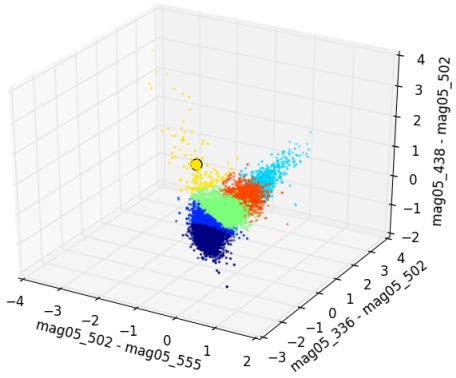


Figure 25. Colour-Colour-Colour distribution of the U-O_3, B-O_3, and O_3-V colours, clustered using K-Means with $K = 6$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

5.3.2 3-Dimensions

The three dimensional distribution displayed more structure than two dimensions. Two clear features were visible. The first was a branch of objects that are red in the U-O_3 colour, and neutral in the rest. The second feature was a branch of objects that were blue in the O_3-V colour, red in the B-O_3 colour, and neutral in the U-O_3 colour.

At all values of K, K-Means is able to identify the first branch of objects. However, it is not until K=6 that the algorithm was able to identify the second branch as its own cluster, see Figure **??**. By this point, the algorithm has segmented the dense area of the distribution by its U-O_3 colour. When projected into two dimensions, there is significant overlap between the clusters that were segmented by colour, and the first branch of objects does not seem to be its own cluster in two dimensions. The score at K=6 causes a slight peak, which shows the effect of picking out both branches of objects. However, the score is still significantly lower than the clusterings that do not identify these branches.

The Meanshift score plateaued clearly at 7 clusters. There is a large drop in score between 5 and 7 clusters, and both clusterings were able to identify both branches. Meanshift was able to identify sub-clusters within the blue branch of objects, that are objects with extremely blue colours, see Figure **??**. The Meanshift clustering was more effective than K-Means, as it did not segment the dense area after it had identified each branch. The clustering with 5 clusters was chosen as the optimal clustering as the clustering with 7 clusters divided the red branch in two, causing significant overlap in the two and three dimensional spaces. When projected into two dimensions, the blue branch of objects is not completely identified. A large portion of objects are assigned to the large cluster that lie at the base of the blue branch. This segmentation does not identify the branches of objects as well as other combinations. Additionally, the number of

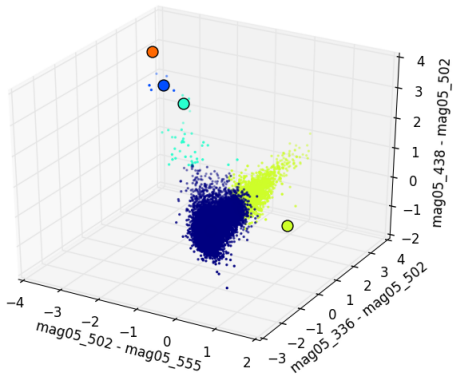


Figure 26. Colour-Colour-Colour distribution of the U-O_3, B-O_3, and O_3-V colours, clustered using Meanshift with $h = 0.5992$. The colour of each point corresponds to the cluster the point was assigned to. Cluster numbers can be seen in the legend.

clusters where the bandwidth and score relation plateaus is between seven and nine clusters, does not align with the optimal clustering.

Needs more discussion After investigating each clustering on the whitelight image, most segmentations did not identify sets of objects that were located in specific areas of the galaxy. Cluster 4 of the strongest Meanshift clustering identified objects that were located in the less dense regions of the spiral arms of M83. This cluster isolated the branch of red objects in the colour distribution. Additionally, this cluster was clearly defined in the CMD and split the objects at colour 0. The largest cluster in the blue branch of objects picked isolated objects in the spiral arm, with only one object located in the nucleus. All of these objects appear to be background galaxy objects or objects behind clouds, as few of the objects appeared in the whitelight image. The other two clusters that segmented the blue branch were also objects that appear to be background or covered by clouds, indicating that these objects are quite bright in the O_3 band, and not in the V band.

6 DISCUSSION

This should be where we present a process for future surveys.

After analyzing the results of all filter combinations it was determined that clustering in colour spaces presents an effective way of analyzing photometric surveys.

The most effective clustering method was K-Means. It consistently produced reliable clusters in two and three dimensions, and could be used to segment blanket surveys regardless of the underlying distribution. K-Means was able to segment data in to equal groups, but it often missed the detail of the distribution. Determining the optimal clustering from the K-Means method was more methodical than others. The silhouette score was a clear indicator of when the optimal clustering had occured, as the peak score or elbow corresponded to the clustering the had identified the most structure in the distribution. The center and inertia tests also aided the clustering evaluation, as they provided a fast way of identifying reliable colour combinations.

Meanshift was able to determine which objects were outliers from the central distribtion. The Meanshift method performed well in three dimensions, but was less reliable in two dimensions. One of Meanshift's benefits was that the cluster shapes were not confined to circles and spheres as K-Means were. It was able to create clusters of uneven sizes, which proved beneficial when the distribution was extremely dense, with branches of objects spanning a large range of colour. The Meanshift algorithm was able to create more isolated clusters than K-Means, and proved effective in most colour combinations. Meanshift struggled when the distribution was relatively flat throughout the range of colours. In this case, the algoithm would pick out small groups of objects that did not seperate themselves from the main distribution. Overall, Meanshift was not as reliable as K-Means, but given the proper distribution, it was able to identify meaningful clusters. Meanshift should be used when the density of the distribution is uneven, and when observers are searching for significant groups of outliers in colour space. **Should we talk about what types of objects these could be?** Determining the optimal clustering was more difficult with Meanshift, as many distributions were sensitive to the bandwidth parameter. The relation between the bandwidth and the silhouette score proved most effective at identifying the optimal clustering, as the elbow in that distribution identified the number of clusters that Meanshift converged on.

Affinity propagation performed well with smaller datasets, but was not able to handle a 0.2 uncertainty limit. Affinity propagation was very sensitive to its parameters, and time should be taken to determine the optimal **preference** values and damping factor. The message passing of Affinity Propagation makes it a very strong clustering method as it relies directly on the data in order to determine the number of clusters. When the size of the dataset was appropriate, Affinity Propagation was able to identify meaningful clusters that were similar to the segmentations of K-Means. The method should be used for datasets with a small number of samples. **What type of surveys would these be?**

In most colour combinations (**Not** $H\alpha$), the three dimensional colour combinations produced more effective clusterings. This is a result of the additional information revealed from the higher dimensional space. The three dimensional space highlighted the structures revealed from two dimensions. This was clear from the narrow band combinations. In most narrow band combinations, it was clear that two branches of objects seperated themselves from the dense center of the distribution, but neither clustering method was able to identify the whole branch. In three dimensions, these branches were clearly seperated from the rest of the distribution, and the clustering methods were able to identify them. **Do we need to give examples in the discussion section?**

The broad band colour combinations that was most effective at identifying different colour classes of objects was $U - B$ and $V - I$. In two and three dimensions this distribution was more effective than $U - V$ and $B - I$. **not sure why this is the case - not really a difference in the distributions. the successful combination had a**

smaller colour range. The unsuccessful had less distinguissed outliers. This combination had a clear branch of objects that were redder in the $U - B$ colour, and the clustering methods were able to identify these objects.

The narrow band colour combinations that were most effective at identifying different colour classes of objects were $U - O_2$, $H\beta - B$, $H\alpha - I$ $(U - B)$. These combinations all contained clear branches of objects that the methods identified in two and three dimensions. Objects in these clusters were located in distinct locations in M83, and were generally different types of point sources. The $H\alpha - I$ and $U - B$ combination was the only combination that performed better in two dimensions. The three dimensional distribution did not reveal any more information than two dimensions, and the clear branch of objects in two dimensions was lost.

Overall, the clustering algorithms were able to identify specific classes of objects in two and three dimensional colour spaces. These classes either categorized objects by integer colour intervals or identified outliers from the distribution. In both cases, the objects in each cluster shared similar colour. A systematic way of identifying the optimal clustering was found, by comparing the parameters of each clustering method and comparing the segmentations each method presented. **Not sure what else to include here**

7 CONCLUSION

Summarize paper

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.

APPENDIX A: PUBLISHED CATALOGUES

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 $3\farcm3$ 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 H*ii* 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), H*ii* 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 $0\farcs26$. 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.**

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
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
Pedregosa F., Varoquaux G., Gramfort A., et al., 2011, Journal of Machine Learning Research, 12, 2825
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., Jasniewicz 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 TEX/ LATEX file prepared by the author.