Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 28 June 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, and efficient means of analyzing those data are needed. The purpose of this research is to apply machine-learning methods to classification of point sources of light emission in nearby galaxies. An objects light emission over different wavelengths is the key data for classification as it indicates the composition of the object, along with its other physical attributes. Mean-shift, Affinity Propagation, and K-means, clustering methods were applied to observations of point sources in the M83 galaxy, to identify objects that emit similar combinations of light over multiple wavelength bands. The data was collected by the Wide Field Camera 3 on the Hubble Space Telescope. To identify which combination of bands was the best at separating different classes of objects, the strength of the clustering was tested using a silhouette score. This metric measures an objects distance from a cluster outside the one it was originally assigned to. The clustering results were also compared with the results of independent classification, to determine if each object was correctly identified. The results of this work will allow astronomers to plan observations that can be used to automatically classify objects in nearby galaxies, leading to a stronger understanding of how stars, and star clusters, form and evolve.

Key words: keywords here

1 INTRODUCTION

Galaxies are complex systems, comprised of numerous components with an 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 ().

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 nuclei with specific properties (e.g.), sometimes they are pure blank-field surveys. Broadband ($R = \Delta\lambda/\lambda < X$) filters are the most common imaging modality, although there have been a few attempts at narrow- or medium-band surveys as well (e.g. Wolf et al. 2003), Clustering in colour space

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

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

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

2 DATA

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 (XX true? XX) **And here we note some other interesting things about M83.**

The objective of the ERS observations as a whole was to probe star formation in galaxies. The observations of M83 were made in broad- and narrow-band filters in order to characterize both stellar and nebular properties. They cover a 3.6×3.6 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**. 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, H*ii* 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.

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

[to be re-organized] As a check on the catalog we used Sextractor to detect and photometer objects in the individual images. While the aperture photometry measurements matched well, the derived uncertainties were much smaller than those reported in the catalog. Indeed, the catalog uncertainties seem to be physically unreasonable, with median uncertainty values well above 1 magnitude in most 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 the individual filters.

Our analysis in this paper is primarily concerned with colours, rather than luminosities. Uncertainties in colours are computed as the quadrature sum of the relevant magnitudes. Observations in 10 bands allow the generation of 45 different colours, but not all of these colours are likely to be useful in characterizing components of the galaxy. As the F555W band has the most individual detections, we initially compute colours relative to this band. The last column of

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

Table 2.

Filter	$N_{\rm obj}$	$N_{\rm good}$	$m_{\rm good}$	$N_{\rm X-555}$
F225W	57585	1196	m	1149
F336W	61787	3572	m	3493
F373N	55908	236	m	229
F438W	64692	5528	m	5490
F487N	60956	429	m	427
F502N	61715	461	m	460
F555W	66539	13556	m	–
F657N	67819	3285	m	2406
F673N	61979	682	m	679
F814W	63759	11976	m	7199

Table 2 gives the number of objects for which a 'good' colour (uncertainty < 0.2 mag) is available.

3 METHODS

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

3.1 Princpal Component Analysis

Images of objects at various wavelengths share similar structures. These structures can be built using a subset of only the most significant of those images. Princpal Component Analysis (PCA) is a process of determining the most significant features of a high dimentional space in order to rebuild structures effectively Kuntzer et al. (2016). PCA projects a data set on its most significant basis, only keeping the dimensions that best explain the data. These dimensions are kept by imposing a maximum number of components on the algorithm, and it is forced to only select the most signifcant ones Kuntzer et al. (2016) PCA has been applied to many astronomical surveys... In this study, PCA is conducted to determine the most significant filters, and colours for classifying objects in M83. All ten filters were used in the inital analysis...

3.2 Mean Shift Clustering

Mean Shift is a non-parametric clustering technique that is based on probability density function estimates of each point in the data. Mean Shift is a very powerful algorithm, but has not been widely used in astronomy. At each point, the algorithm estimates the density around that point using a small sample of objects surrounding the point. The power of Mean Shift clustering is that the clusters are not confined to a particular shape. Because Mean Shift moves towards the local mode near the data on which it was initialized, it is useful for estimating the number of signifcant clusters in a dataset Comaniciu & Meer (2002) 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:

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

Where 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).

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:

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

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:

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

$m_{h,K}$ 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:

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

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. Equation 1 introduces the bandwidth parameter h. Estimating the bandwidth correctly is critical to determining the correct number of clusters. If the bandwidth is too low, the density estimate will be undersmoothed, and Mean Shift will produce many small clusters Vatturi & Wong (2009). This is a result of the large density gradient resulting from a low bandwidth, causing many data points to be interpreted as local modes. 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).

3.3 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 only input required for AP are the *preferences* of each data point which describes the likelihood of a data point to be chosen as an exemplar Frey & Dueck (2007). The preferences are a measure of the similarity between a point i and a candidate exemplar k defined by:

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

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

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

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

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):

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

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

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

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

As the process iterates, for point i, the value of k that maximizes $a(i,k)+r(i,k)$ identifies i as an exemplar if $k = i$, or gives the exemplar of point i Frey & Dueck (2007). In order to ensure that the message passing does not cause numerical 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).

3.4 K-Means Clustering

K-Means clustering is one of the most widely used clustering methods and has been used to identify a wide range of interstellar and intergalactic objects. It is simple, robust, and easy to implement when analyzing high 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:

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

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

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

4 ANALYSIS

(i) PCA Analysis
(ii) Clustering process
(iii) Mean-Shift Analysis
(iv) Affinity Propagation Analysis

(v) K-Means Analysis

5 RESULTS

Well, what did you learn?

ACKNOWLEDGMENTS

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. We acknowledge the efforts of WFC3 Science Oversight Committee in conducting the Early Release Science program.

REFERENCES

Adamo A., Kruijssen J. M. D., Bastian N., Silva-Villa E., Ryon J., 2015, MNRAS, 452, 246
Almeida J. S., Prieto C. A., 2013, The Astrophysical Journal, 763, 50
Andrews J. E., Calzetti D., Chandar R., Elmegreen B. G., Kennicutt R. C., Kim H., Krumholz M. R., Lee J. C., McElwee S., O'Connell R. W., Whitmore B., 2014, ApJ, 793, 4
Bastian N., Adamo A., Gieles M., Lamers H. J. G. L. M., Larsen S. S., Silva-Villa E., Smith L. J., Kotulla R., Konstantopoulos I. S., Trancho G., Zackrisson E., 2011, MNRAS, 417, L6
Bastian N., Adamo A., Gieles M., Silva-Villa E., Lamers H. J. G. L. M., Larsen S. S., Smith L. J., Konstantopoulos I. S., Zackrisson E., 2012, MNRAS, 419, 2606
Blair W. P., Chandar R., Dopita M. A., Ghavamian P., Hammer D., Kuntz K. D., Long K. S., Soria R., Whitmore B. C., Winkler P. F., 2014, ApJ, 788, 55
Blair W. P., Winkler P. F., Long K. S., Whitmore B. C., et al., 2015, ApJ, 800, 118
Chandar R., Whitmore B. C., Calzetti D., O'Connell R., 2014, ApJ, 787, 17
Chandar R., Whitmore B. C., Kim H., et al., 2010, ApJ, 719, 966
Comaniciu D., Meer P., 2002, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 603
Dopita M. A., Blair W. P., Long K. S., Mutchler M., Whitmore B. C., Kuntz K. D., et al., 2010, ApJ, 710, 964
Fouesneau M., Lançon A., Chandar R., Whitmore B. C., 2012, ApJ, 750, 60
Frey B. J., Dueck D., 2007, Science, 315, 972
Hollyhead K., Bastian N., Adamo A., Silva-Villa E., Dale J., Ryon J. E., Gazak Z., 2015, MNRAS, 449, 1106
Hong S., Calzetti D., Dopita M. A., et al., 2011, ApJ, 731, 45
Kim H., Whitmore B. C., Chandar R., Saha A., et al., 2012, ApJ, 753, 26
Kuntzer T., Tewes M., Courbin F., 2016, ArXiv e-prints
Liu G., Calzetti D., Hong S., Whitmore B., et al., 2013, ApJ, 778, L41
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
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.