A MORPHOMETRIC ANALYSIS OF CALIFA SURVEY GALAXIES THROUGH CLUSTERING AND PRINCIPAL COMPONENT ANALYSIS VISUALIZATION

Vanessa De Oliveira Gil\textsuperscript{1}, Fabricio Ferrari\textsuperscript{2}

\textsuperscript{1}Center For Natural And Exact Sciences, Federal University Of Santa Maria, Santa Maria, RS, Brazil.

\textsuperscript{2}Institute Of Mathematical Statistics And Physics, Federal University Of Rio Grande, Rio Grande, RS, Brazil.

DOI : https://www.doi.org/10.56726/IRJMETS47129

ABSTRACT

To comprehend the processes of formation and evolution of galaxies, it is essential to identify properties that allow their classification. The conventional approach involves subjective visual inspection of individual images, which is limited. Consequently, there is a necessity to seek an automatic method based on the quantification of its morphologies to find classes of galaxies.

We propose a classification method through the analysis of object groups within the morphometric parameter space, as measured by MORFOMETRYKA. This method proposes to identify clusters of strongly correlated observations that indicate the same physical processes of formation and evolution. The object of study is a subset of the CALIFA survey with galaxies at different morphological stages.

The clustering algorithm DBSCAN is employed for unsupervised classification. As this technique operates independently of predefined classes, the resulting morphological groups are based on the inherent characteristics of the survey. To enhance the precision of groupings, additional techniques such as principal components analysis and Delaunay decomposition are applied.

Keywords: Clustering, Unsupervised Learning, Morphology, Galaxies, Analysis.

I. INTRODUCTION

Some of the fundamental questions in Astrophysics involve understanding the intrinsic characteristics of galaxies as: structures, the process of formation and their evolution. Connecting physical processes with morphological structures is also a significant challenge. According to [1], understanding the morphology of galaxies is important to make reliable the study of the effects of the environment on galaxies, the segregation of types in clusters, and the fundamental factors that determine the type of galaxy during its formation, among other aspects. The relationships between the different types of galaxies are what direct the effectiveness of the study of the morphology to understand these problems [1].

Since [2], galaxies have been categorized by specialists based on their visual characteristics. The absence or presence of morphological structures placed the galaxies into specific groups. For instance, the difference between elliptical and spiral galaxies is quite evident. Spiral galaxies have spiral arms, whereas elliptical galaxies does not. However, it is not easy to definitively classify a spiral galaxy as Sa or Sb due to the morphological continuity between these subclasses. [3] revisited Hubble classification, proposing that the classification should depend on characteristics such as bars and rings in the structures. Other methods of classification were proposed by [4], [5] among others. They have in common the subjectivity present in the delimitation of the existing morphological types.

To answer these fundamental questions, firstly it is necessary to classify galaxies based on their intrinsic characteristics. To classify galaxies, the classification procedure must be objective and generic. It should be independent of the user and the tools available for the classification process.

There are two main ways to study galaxies properties: parametric and non-parametric. The parametric method consists of studying the light distribution of galaxies through predefined functions and fit profiles, such as the Sérsic’s profile. However, this approach introduces many errors through its steps due to the dependency of several free parameters. When non-parametric techniques are used we have a method more robust to quantify the morphology of galaxies because it is more direct. The required quantities are extracted directly from the image, eliminating dependence on free parameters.
As we need a robust method of galaxy separation, we chose non-parametric measures of galaxies' morphology. The most common morphometric classification system is CASGM ([6], [7], [8], [9]) which proposes to measure concentration, asymmetry, smoothness, Gini coefficient and moment of light. Other approaches are also made using spirality ($\sigma_\psi$) and entropy ($H$) [10].

For this purpose, we selected a method that separates objects into groups according to their similarities and dissimilarities. The clustering algorithm uses density difference as a way to create new clusters. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an unsupervised technique that discovers intrinsic patterns from the dataset [11]. Given that one of the main objectives is to automate the process of classifying galaxies into distinct classes, and we may not always know the shape of the dataset (spherical, linear, elongated), an approach based on density is suitable to address this issue, since it can be applied to any type of data.

Since the 1990s, various approaches involving machine learning to classify the types of galaxies have been under study. Among these techniques, the utilization of Artificial Neural Networks has been extensively explored by different authors, employing various types of data, as demonstrated by [12], [13], [14], and [15]. All these techniques show promise in replicating the classifications produced by specialists.

The main difference between these techniques and the cluster analysis is that there is no need for a training set to divide the galaxies into groups. The division into these groups is based on characteristics derived from the morphometric data extracted from the images of galaxies in the corresponding sample to CALIFA. This implies that each data set will be divided based on the data used which produces a unique group layout independent of previous samples that might influence its allocation.

II. METHODOLOGY

Sample and Data

The analyzed sample consists of galaxies that are part of the CALIFA survey [26]. CALIFA's mother sample is composed of 939 galaxies which were selected from the Sloan Digital Sky Survey DR7 (SDSSDR7, [16]) photometric catalog [17]. SDSS is a survey that covers the northern sky and supply the photometry of these galaxies in five bands u, g, r, i, z [18]. The telescope used has a 2.5m mirror and it is located at Apache Point Observatory.

To collect the CALIFA data, 250 observation nights were dedicated to using the Potsdam Multi-Apertures Spectrophotometer (PMAS, [19]) in the PPak mode on the 3.5m telescope at the Calar Alto Observatory which is in the Centro Astronómico Hispano Alemán (CAHA) at Calar Alto, operated jointly by the Max-Planck-Institut für Astronomie and the Instituto de Astrofísica de Andalucía (CSIC). CALIFA aims to characterize galaxies of all morphological types across a broad spectrum of properties, including photometry, spectroscopy, dynamics, kinematics, and stellar population analysis.

The analyzed sample represents a subset of galaxies with an absolute magnitude limited by $-19 < M_r < -23.1$. Only the brightest galaxies were chosen within a redshift limitation of $0.005 \leq z \leq 0.03$ [20]. These constraints were applied to ensure that the galaxies must be observed under the same configuration, maintaining the consistency of the physical properties. Further details regarding the complete properties of the CALIFA are described in [20]. The morphometric parameters for these galaxies were measured using the r-band stamps from the Sloan Digital Sky Survey (SDSS) with the MORFOMETRYKA tool. Only results with quality flag equals zero were used to enhance the accuracy and reliability of this research. A subset of the mother sample was used and it has 267 galaxies considering all morphological types. The application of the quality flag restriction serves as a meticulous criterion, allowing us to exclusively incorporate the most robust and precise morphometric measurements in our analysis.

MORFOMETRYKA

MORFOMETRYKA is a robust computational tool designed to automate the quantification of a comprehensive array of morphometric and structural parameters in galaxies [10]. Its proficiency in systematically extracting such parameters contributes significantly to the efficacy and objectivity of our morphological analysis. The non-parametric indices are computed by this algorithm when a galaxy image is inputted, automatically extracting quantities from the image pertaining to concentration, asymmetry, and entropy. MORFOMETRYKA is capable of conducting measurements of large databases without human interference.
Cluster analysis is an unsupervised data mining technique with the objective of exploring and identifying groups within data. As stated by [21], "Cluster analysis can be used to classify astronomical objects and can often help astronomers find unusual objects within a flood of data."

Objects are divided based on their similarities, so they need high similarity scores within the same group and significant dissimilarities between objects of different groups. Typically, similarity measures are computed using distance metrics such as Mahalanobis, Euclidean, or Manhattan distances.

The objects are assigned to groups based on the parameters used for this study. Therefore, to partition the data into groups, it is necessary that the parameters employed have physical meaning. The more uncorrelated these parameters are, the more evident the clustering division becomes.

Visualizing multivariate data graphics can help us to perceive structures that are in evidence in the data, therefore it serves to suggest the existence of groups or the need to use auxiliaries tools to achieve the goal of generating coherent clusters.

Three primary types of clustering algorithms are: hierarchical, partition and based on density. In hierarchical algorithms, the data is partitioned successively, resulting in a hierarchical representation of the clusters [21]. The result of the hierarchical technique is a dendogram which illustrates how the groups are arranged.

Partitioning algorithms divide a dataset of \( n \) objects into \( k \) clusters. The number of clusters is required as an input parameter, having prior information about the data is recommended for making an appropriate choice.

An example of a partitioning algorithm is K-means [22]. It randomly initializes \( k \) centroids and the objects are arranged in groups based on their proximity to the centroid with the smallest distance. After that, the positions of the centroids are recalculated according to the objects within each cluster. These steps are repeated until the data converges. Another well-known algorithm is K-Medoids [23] which works with the Medoid concept, representing the most central object within a cluster. They are selected randomly and the other objects are assigned to the nearest Medoid cluster. After that the algorithm calculates the new Medoid position and reallocates the objects until the best result is found.

The density-based algorithms have some advantages over partitioning algorithms. When DBSCAN is used, there is no need to input the number of clusters because the algorithm finds the correct number based on values of the radius of search and the minimum quantity of points. Another advantage concerns the data format while K-means works better with spherical data, DBSCAN can be applied efficiently to data of arbitrary shape and size. When the data points have a dense distribution in comparison with the neighborhood it is called a cluster. Noise areas are less dense than the clusters. With DBSCAN, it is possible to identify outliers, and these points are not assigned to the clusters, which is different from the other algorithms.

Despite DBSCAN not requiring knowledge of the exact number of clusters, it is important to choose an appropriate value for the neighborhood radius, where the algorithm will search for a minimum number of points inside the circles to form clusters. Determining these quantities is not so complicated, but it is important to choose a value that is neither too large nor too small. A radius that is too large would result in all objects being grouped into a single cluster, while a radius that is too small might lead to the algorithm generating numerous small clusters probably meaningless.

To address this challenge, the use of Delaunay triangulation (DT) [24] has been proposed. DT subdivides a set of points into triangles and it has a special property that the circumsphere of the triangles cannot have any points of the triangulation so they have to be empty. This triangulation is significant as it maximizes the smallest angle among all existing triangulations. Additionally, if there are more than three points within the same circumsphere, the triangulation becomes unique [25].

In many cases when astrophysical problems are being solved there are a large number of variables to consider. Principal Component Analysis (PCA) is a great statistical technique which helps to choose the most relevant variables. PCA operates by linearly transforming a dataset to generate uncorrelated variables. These variables, termed principal components, are designed to encapsulate the most significant information about the data. It is a requirement that these principal components maintain orthogonality among themselves. To satisfy the objective to find clusters in CALIFA data sample the algorithm DBSCAN was used and PCA to visualize clusters.
III. RESULTS AND DISCUSSION

Analyzing the density of the morphometric indexes present in the CALIFA sample, two primary peaks of density are discernible between the values corresponding to $C_1$, $H$, and $q_{F/2D}$, as illustrated in Figure 1. It is noteworthy that a morphometric continuity is observed in these parameters; however, the presence of two distinct density peaks is quite evident. The sample of CALIFA dataset was subjected to cluster analysis using the DBSCAN algorithm. As DBSCAN identifies regions with varying densities, the parameters mentioned hold promise for the comprehensive analysis of galaxies within the sample. In Figure 1, the representation of parameters $C_1$ and $H$ exhibits two regions with high densities. The relationship between entropy and axis ratio appears less distinct. In the configuration corresponding to $C_1$ and $q_{F/2D}$, a region with a prominent density peak is evident, while a smaller region is situated at the opposite extreme of the graphic. This information suggests the potential to identify distinct groups of galaxies within this parameter space.

Figure 2, clearly demonstrates the presence of two well-separated and distinct groups within the analyzed dataset. The black dots in the figure symbolize the central points of the identified groups. The black circle in the diagram represents the size of $\varepsilon$, the parameter utilized by the algorithm to detect these galaxy groups. The relationship between $C_1$ and $H$ suggests that the blue group represents late-type galaxies. Typically, galaxies with low light concentration are situated in the late-type branch of Hubble’s Diagram. The red group, characterized by high central light, represents early-type galaxies. The green dots represent objects that the algorithm cannot categorize into any of the existing groups. Despite this, the green objects were not considered as a unique group; therefore, they can be regarded as outliers by DBSCAN. This is likely due to their position in the border regions between the two main groups, causing the algorithm to encounter difficulties in accurately grouping them. This factor may suggest that the galaxies in this region have characteristics corresponding to both the early and late-type branches of galaxies morphology.

Figure 1: Density plots of $C_1$, $H$, and $q_{F/2D}$ reveal two distinct density peaks within the CALIFA dataset sample.

$C_1$ and $H$ show a clear delimitation between the two groups near the (0,0) position. This result is in agreement with the information presumed by the density graphic in Figure 1, as it faithfully reproduces the groups in relation to the density peaks. When studying the relationship between $C_1$ and $q_{F/2D}$, the group of late type galaxies is more compact, whereas the red cloud is more dispersed. In this space, there is a difficulty in grouping objects with low concentration and a higher axis ratio. In the analysis of the relationship between $H$ and $q_{F/2D}$, there is a slight density difference between the two groups, and their centers are positioned in locations distant from each other.

Figure 2: Graphics indicating the results obtained through DBSCAN for the CALIFA sample using parameters...
Red dots represent early-type galaxies, blue dots correspond to late-type galaxies, and green dots do not fit into either pre-established group. The black dots indicate the center of each group, and the black circle represents the size of $e$, the radius of search that the algorithm uses to group the data.

How can we illustrate more easily how groups of galaxies fit into this space of parameters? Figure 3 presents a histogram that facilitates the identification of similarities and distinctions between the groups. As discussed earlier, a morphometric continuity between the groups is evident, aligning with expectations based on the analysis of density charts and the groups obtained through DBSCAN. In all regions of the histogram, it is evident that the groups share some characteristics, explaining the overlap observed in both the identified groups and the outliers.

![Figure 3](image-url)  
**Figure 3:** Histogram representing each group of galaxies obtained.

When principal component analysis is applied in this parameter space, it becomes less intuitive to find distinct density peaks in the graphs of Figure 4. Between components 1 and 2, there is a slight difference in density between two regions that stand out in the graphic. Nevertheless, when the third major component is used in combination with components 1 and 2, there is a continuous density range. The inclusion of this third component could suggest that there is only one group in this set. In Figure 5, the first two principal components generate a distribution similar to that found with the parameters $C_1$ and $q_{Hir2D}$. However, for $PC_2$ and $PC_3$, the groups overlap and thus do not provide relevant information for identifying sample groups.

![Figure 4](image-url)  
**Figure 4:** Density graphics of principal components $PC_{1,2}$ and $PC_3$ which shows peaks of density according to the established relations.

When we talk about principal components, it is difficult to discern how the parameters that gave rise to them influence the distribution of clusters. In Figure 6, it is possible to understand which parameter contributes more or less for each group. The concentration index and entropy are great allies to find early and late types of galaxies in data. They are fundamental to separate galaxies in clusters. Also, observing $PC_{1,2}$ and $PC_2$ it is interesting that the green group concentrates in $q_{Hir2D}$ axis.
Figure 5: Graphics that indicate results obtained through DBSCAN for the three principal components. Red dots indicate early type galaxies, blue dots refer to late type galaxies, and green dots do not fit into either pre-established group. The black dots indicate the center of each group and the black circle refers to the size of the \( \varepsilon \), radius of search, that the algorithm uses to group the data.

Figure 6: Graphics of clusters distribution and directions of parameters variance.

Figure 7: Histogram representing each group of galaxies obtained after PCA.

We found that the application of the parameters concentration, entropy, and axis ratio produces meaningful results to identify different types of galaxies in two clusters. This result is obtained with or without the use of PCA technique. However, as we can see in Figure 7, the groups are more clearly separated when PCA is applied. In the Figure 2, Figure 5 and Figure 6 that represent the clusters, it is possible to notice that the early-type galaxies have a higher concentration and a lower entropy. For late-type galaxies, the concentration is lower and the entropy higher.

IV. CONCLUSION

The utilization of principal components analysis (PCA) further contributed to the exploration of parameter space. While the PCA results may appear less intuitive in terms of distinct density peaks, the histograms and clustering graphics provide a comprehensive view of how galaxies are distributed across the principal components. This approach enhances our understanding of the contribution of each parameter to the overall classification process.

The comparison between the parameter space without PCA and that with PCA emphasizes the efficacy of the latter in refining the separation between clusters. The histogram analysis clearly illustrates the improved
separation of groups when PCA is applied, reinforcing the utility of this technique in enhancing the classification performance.

Automated analysis removes the potential for human bias and subjectivity, ensuring consistent and reliable classification. Besides, the automated method can analyze large datasets rapidly, saving valuable time and resources.

The results demonstrate the proposed method’s efficiency and objectivity in categorizing galaxies, revealing distinct early and late-type clusters. Our automated classification method offers a valuable tool for astronomers, opening doors to deeper insights about galaxy morphology. The results obtained provide a foundation for further studies in galactic morphology, by applying the method to larger and more diverse datasets, including galaxies from different environments and epochs, could refine the classification system and provide a more comprehensive understanding of galactic evolution.

ACKNOWLEDGEMENTS

The first author acknowledge CAPES who financed the research and SDSS to make data available.

V. REFERENCES


