Galaxy cluster detection in the local universe, using machine learning methods.

Harry Psarakis[1]

[1]*Physics Department, University of Crete.*

In this thesis, i report the implementation of machine learning methods, i.e clustering algorithms such as DBSCAN and Hierarchical clustering, for the identification of galaxy clusters in 3D space in HECATE catalogue, which is a complete sample of nearby galaxies. The results were compared with known catalogs of galaxy clusters using standard methods.

I. INTRODUCTION

Clusters of galaxies, as the largest gravitationally bound structures in the universe, hold a unique position in astrophysical and cosmological studies. Their immense mass provides compelling evidence for the existence of dark matter, while their distribution offers crucial constraints on cosmological parameters. Serving as nodes in the cosmic web, clusters offer a window into galaxy evolution, influenced by unique intergalactic processes. Additionally, their role as natural astrophysical laboratories allows us to study phenomena like gravitational lensing, the hot intracluster medium, and galaxy interactions. In essence, the study of galaxy clusters offers insight into the universe's structure, its origin and the underlying processes driving its evolution.

The classification of clusters comes into two categories, groups or clusters, depending on the number of galaxies that are gravitationally bound. In both situations the galaxies are orbiting around the central point of the system's mass, the barycenter. Clusters contain from fifty galaxies (poor cluster) to thousands galaxies (rich cluster). They can further be classified into two main types based on their morphology and dynamical state: regular and irregular. Regular or symmetrical/spherical clusters are considered to be in a more relaxed state, meaning they've had more time to come to an equilibrium state, after any significant mergers or interactions. Proof of such statement is that regular clusters are completely dominated by early-type galaxies (elliptical). On the other hand, irregular clusters are still in the process of evolution and tend to have a higher fraction of spiral galaxies occupying them. Also, regular clusters appear to be compact, with the majority of elliptical galaxies harboring the center of the cluster, while irregular are significantly less dense in the center. Groups generally have less than 50 galaxy members to a lower limit of four members. Over cosmic time, galaxy groups will merge with other groups or fall into clusters and form bigger structures. The Virgo cluster is such an example, that can be classified as an open/irregular cluster that consists of smaller sub groups.

Today it is known that, although the galaxies determine the optical appearance of a cluster, the stellar mass contained in galaxies contributes only a small fraction to the total mass of a cluster, about $\sim$5%. In 1933 Zwicky measured the radial velocities of galaxies in Coma cluster from their Doppler shift spectra. From these observations he calculated the dispersion in their radial velocities. Then using the Virial theorem he estimated the mass of the cluster only to find out that the mass-to-light ratio was greater, by more than a factor of 10, than the typical values of its individual galaxies ($M/L \sim 1-20 M_\odot/L_\odot$). From this result he stated that Coma cluster contains significantly more mass than the sum of the masses of its galaxies.

Later X-ray observations revealed that X-ray emission from clusters is spatially extended and not just originated on individual galaxies. This indicates the presence of a hot gas, with temperatures ranging around $T \sim 10^8 K$ (thermal bremsstrahlung), occupying the space between galaxies (intracluster medium). For fully ionized hydrogen, the mass estimate of the gas is still only a few percent of the total mass of a cluster, almost $\sim$15%. The study of the X-ray spectra, such as the emission lines, can also reveal information about the abundance of heavy elements in galaxies, indicating their chemical evolution. That indicates that a significant amount of matter in clusters exists outside of galaxies and that aids in forming a broader perspective of the universe on large scales. For this hot intracluster gas to remain gravitationally bound within the cluster and achieve such high temperatures, there must be a strong gravitational force acting within. Since the total mass of galaxies is insufficient to provide that strong force necessary to confine this hot gas, dark matter must be present.

These significant results, i.e the mass of galaxy clusters exceed significantly that of the visible matter in stars and the mass of the intracluster medium, were the first indicators of the existence of dark matter. Dark matter as a significant portion, roughly 80%, exerts a massive gravitational pull, leading to phenomena like gravitational lensing. This lensing acts as a cosmic telescope, magnifying distant, faint galaxies, bringing them into observational reach and providing valuable insights into the early universe.

So the study of clusters of galaxies is crucial for several reasons, and that led scientists to the creation of galaxy cluster catalogs. The first attempts to create major galaxy cluster catalogs were thrown by George Abell and Fritz Zwicky. The catalogue of rich galaxy clusters

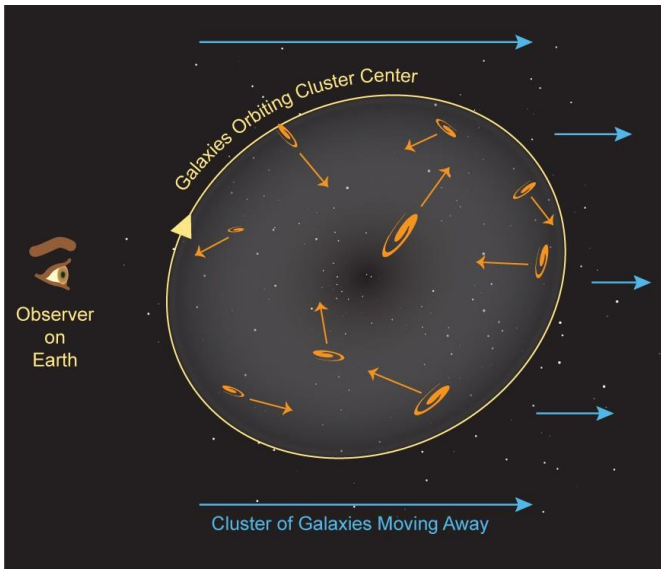

Figure 1. Illustration of a galaxy cluster's sources.

by Abell (1958) has been widely used as a primary source list for different type of astronomical studies (G. Abell, H. Corwin et al. 1989). Its an all-sky catalogue of 4073 rich clusters of galaxies, each having at least 30 members, at the local universe ($z < 0.2$). Abell was looking for regions in the sky that showed an overdensity of galaxies by using the Palomar Observatory Sky Survey's photographic plates, which covered the majority of the sky. Abell's criteria for the identification of clusters were based on an increased concentration of galaxies within a circle of angular radius, called the Abell radius

$$\vartheta_A = \frac{1'.7}{z}$$

where z is the estimated redshift and within a magnitude interval $m_3 \leq m \leq m_3+2$, where m_3 is the apparent magnitude of the third brightest galaxy in the cluster. The Catalogue of Galaxies and Clusters of Galaxies (Zwicky Catalogue) is a list of more than 29000 galaxies and 9000 galaxy clusters identified from Palomar Observatory Sky Survey in 1960. Zwicky's criteria for identifying clusters relied on visual inspection of photographic plates, similar to Abell, but without the strict richness criteria that Abell employed. The Zwicky catalogue is notable for its breadth, including both rich and poor clusters. Since publication, the catalogue has been updated with corrections, using new observations and data corrections.

Despite the great importance of such galaxy cluster catalogs, they come with limitations. Ideal catalogs should meet two primary standards. They should be complete, meaning that they include all the galaxy members of a cluster, and they should be reliable by ensuring that only objects meeting the criteria are listed, without any inaccuracies. The aforementioned catalogs and the ones used in this thesis, are neither complete nor pure. First of all measurements acquired from photographic plates are subjective and such manual cluster identification cannot be applied to modern large cluster catalogs. Secondly, galaxy counts on images are strongly affected by projection effects. Since the observations are two dimensional representations of the three dimensional universe, projected overdensities on the celestial sphere can be classified as clusters. For example, a random alignment of galaxies along the line of sight might be mistakenly identified as part of a cluster, even if those galaxies are not gravitationally bound to each other and are at different redshift. Modern catalogs have limitations as well, despite being developed by advanced observational techniques and tools, due to projection effects, redshift uncertainties etc.

Building on the limitations of traditional cluster identification methods, automated methods, such as machine learning techniques, can offer enhanced reliability in identifying genuine clusters, minimizing misclassifications caused by projection effects or random galactic alignments. This is especially important in the context of current, extensive galaxy catalogs, based on observations of million galaxies. By systematically integrating multi dimensional data from diverse astronomical observations, machine learning techniques can provide robust approaches to cluster identification, thereby improving the accuracy and efficiency of astronomical classification tasks. Moreover, the predictive capabilities of these techniques are instrumental in unveiling potential undiscovered galaxy clusters offering hints and directions for further exploratory initiatives. Finally the merging of machine learning methods and astronomical observations can yield substantial insights, and bridge the existing gaps in our understanding, allowing for more informed interpretations of the universe.

II. DATA

The goal of this thesis was to identify clusters in HECATE (The Heraklion Extragalactic CATaloguE). This is an all-sky value-added galaxy catalogue of 204,733 individual galaxies within a radius of ~200 Mpc. HECATE is based on the HyperLEDA catalogue and is enriched with additional information from other extragalactic and photometric catalogues. The catalog aims to support contemporary and upcoming multi-wavelength investigations of the nearby universe. It offers lots of information such as positions, distances, sizes, photometric measurements etc. (K. Kovlakas, A. Zezas, J. Andrews et al. 2021), but not cluster associations for its members. The study involved a clustering analysis of these galaxies based on their spatial arrangement, with the intention of assigning cluster identifiers to each galaxy candidate.

In order to develop and test the methods presented in this thesis, local universe galactic catalogs of known clusters were also used, like Northern/Southern Abell catalogue (G. Abell, H. Corwin et al. 1989), Compact

Groups of Galaxies (A. McConnachie, D. Patton et al. 2009), Hickson Compact Groups of Galaxies (P. Hickson 1982), Extended Virgo Cluster Catalog (K. Suk, R. Soo-Chang et al. 2014), which contained the cluster associations for their galaxy members. Because the Abell catalogue by itself does not contain information about its cluster members, a complete search was performed on NED (NASA/IPAC Extragalactic Database) with 'ABELL' as a cross-ID. A total of 131,313 galaxies were collected through that search, while 463 obtained from Hickson catalog, 1589 from EVCC and 323,221 from compact groups (A. McConnachie).

A. Selection of galaxies

In order to select which HECATE galaxies were suitable for the purposes of this study, a series of data selection techniques were applied on the catalog.

Firstly, through TOPCAT (Tool for OPerations on Catalogues And Tables), which is an interactive graphical program that can examine, analyse, combine, edit and write out tables, all HECATE galaxies were cross matched with the galaxies obtained by the cluster/group catalogs mentioned before, to check which HECATE galaxies overlapped with a known cluster/group. If they were identified a new feature was added to HECATE's galaxies with the label of the cluster/group they belonged.

Regarding the cross match criteria, the algorithm used by TOPCAT, compares elliptical regions on the sky. The search radius of each HECATE galaxy was the D_{25} ellipse, which represents the region of the galaxy where the surface brightness falls to 25 magnitudes per square arcsecond, a standard level used to define the size of galaxies in photometric studies. So, the position of each HECATE galaxy was characterized by: RA (degrees), DEC (degrees), D_{25} semi-major axis (arcmin), D_{25} semi-minor axis (arcmin) , and the position angle (degrees). When information about the axis of D_{25} ellipse was missing, galaxies assumed to have major and minor radii around 0.5 *arcmin*.

For the galaxies obtained by the other catalogs, a cross match was performed between them and their clusters, for the first to obtain their cluster association. During this cross match, the galaxies assumed to be point sources while the angular size of the clusters/groups were provided in the corresponding catalogs.

Then a cross match between all HECATE galaxies and the galaxies obtained by the other catalogs was performed. The algorithm compared the ellipses of HECATE galaxies and the ellipses of 0.5 arcmin, both as semi-major and semi-minor axis, for the other galaxies. A match between galaxies occurred if there was an overlap between these ellipses.

Secondly, the final labeled galaxy set, containing all the cross matched galaxies, was further examined for the fulfillment of the following criteria.

- The velocity dispersion of the galaxies in each cluster was computed and the ones exceeding the $mean \pm 3std$ limit were dropped from the dataset.
- Galaxies that had difference, between their velocity and their cluster's provided in the relevant catalog, above $4000 km/s$ were characterized as outliers and they were discarded.

The criteria above were primarily established to ensure that galaxies obtained by NED database were indeed real, and not possible outliers stored as "ABELL" galaxies. It is also worth noting that galaxies which appeared to be members to both a compact group and a cluster (duplicates) were assigned with their cluster label instead. Following the procedure above the final labeled galaxy set contained 1584 galaxies from EVCC catalog, 353 galaxies belonging to Hickson's compact groups, 2142 to the McConnachie et al. compact groups and 7438 belonging to Abell clusters, resulting in total 11517 HECATE galaxies with their cluster label assigned.

III. METHODS

That labeled galaxy set, i.e HECATE galaxies found to already reside well known groups and clusters, was used for spatial clustering in 3D in a semi-supervised manner. By semi-supervised, the domain knowledge, notably the cluster label for each galaxy was leveraged. The advantage of having cluster labels for a small subset of HECATE led to the methodologies detailed in this section.

Each galaxy, on the labeled set, was characterized by four features. RA (Right Ascension), DEC (Declination), and Redshift described the spatial information of each galaxy in spherical coordinates. Because redshift by itself was not contained as an information in HECATE, all velocities were converted through the Doppler relationship

$$z = \frac{u_{rs}}{c}$$

where u_{rs} is the recessional velocity of a galaxy and c is the speed of light. The fourth feature provided was the cluster label, indicating the specific cluster or group, to which each galaxy belong. To ascertain the spatial relationships between galaxies, pairwise distances were computed for every galaxy pair using a custom distance function adapted for spherical coordinates,

$$D = \sqrt{r^2 + r'^2 - 2rr'(\sin\vartheta\sin\vartheta'\cos(\varphi - \varphi') + \cos\vartheta\cos\vartheta')} \tag{3.1}$$

where (r, ϑ, φ) and $(r', \vartheta', \varphi')$ are the spherical coordinates of a galaxy pair. The distance r and r' of galaxies was calculated through Hubble's Law for small redshift

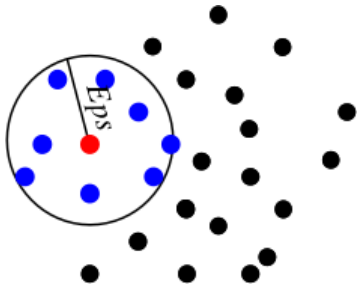

Figure 2. Illustration of *eps* and *min_samples* (blue data points).

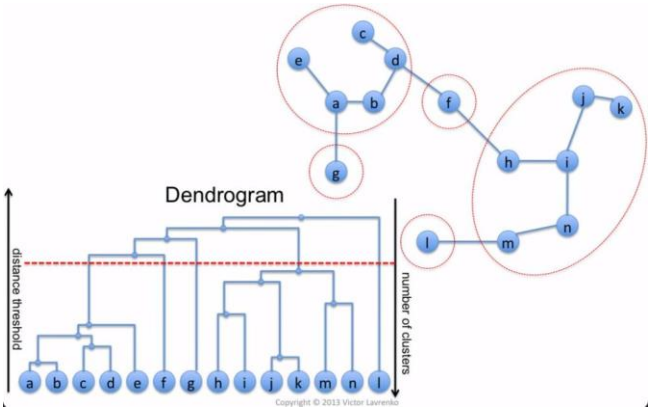

Figure 3. Construction of *dendrogram* and application of *distance threshold*.

values

$$r = \frac{u_{rs}}{H_0}$$

where H_0 is the Hubble constant, which was set to $67 km/s/Mpc$.

Given that galaxies in HECATE catalog exist in the local universe, and thus have small redshift values, the expansion of the universe can be approximated as linear. This linear approximation consequently ensures that distances derived from redshift measurements will satisfy the triangular inequality. Keeping this in mind, the proposed distance metric for calculating galactic distances, between individual galaxies, is justified.

Clustering techniques in machine learning traditionally refers to unsupervised learning methods used to group data points into distinct clusters based on some measure of similarity without prior labels. However in reality, data can sometimes include some labels, or even details like constraints indicating which items should or shouldn't be linked together. This added information can offer valuable insights for both clustering and identifying outliers. Building on this concept, semi-supervised clustering is a technique that leverages these labeled items while also grouping the unlabeled ones.

Before the clustering procedures, it was necessary for the total number of ground truth clusters to be calculated. The basic assumption was that the minimum number of galaxies in order for a cluster/group to be identified was four. That resulted on 190 different clusters existing on the labeled galaxy set.

Three different clustering methods were then applied to the labeled data to obtain spatial clustering results, all provided by the scikit-learn python library. These methods were optimized, to check how well they predicted the ground truth cluster labels while also achieving a high performance.

1. DBSCAN clustering

   DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that groups together data points that are in close proximity based on a distance measure (eps) and a specified density, i.e minimum number of points (min_samples) (Fig. 2). Unlike many other clustering methods, DBSCAN does not require the number of clusters to be specified a priori. This allows the algorithm to identify clusters of varying shapes, making it especially useful for data with non-globular and irregular structures. The most important of its parameters, is metric, which defines the method to calculate the distances between different points in the dataset. The most commonly used metric, suitable for datasets where features are continuous, is *Euclidean distance* given by

$$Distance = \sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2}$$

   for two points in a 2-D space represented as (x_1,y_1) and (x_2, y_2). Moreover, another notable feature of DBSCAN is its ability to distinguish noisy or outlier data points, classifying them as points that do not belong to any cluster due to a lack of sufficient density. Its capability to isolate and distinguish outliers adds a layer of robustness, allowing for more accurate clustering results and making it more reliable in real-world applications where noise and anomalies are present.

2. Agglomerative clustering

   On the other hand, Hierarchical clustering is a methodology that creates a tree of clusters known as a dendrogram (Fig.3), illuminating the hierarchical structure of data points based on their similarity. Among the different approaches is hierarchical clustering, Agglomerative clustering is the most common, which adopts a "bottom-up" strategy. Starting with each data point as a single cluster, it recursively merges the closest pair of clusters into a single cluster, continuing this process until only one large cluster remains or a stopping criterion is met (distance_threshold) (red dotted line,

Fig.3). This enables the exploration of data groupings at various levels of precision, affording a better understanding of the inherent structures within the dataset. The metric which specifies the distance metric used between individual data points is called affinity, where also in this case the most common is *Euclidean*. Another requirement is the selection of an appropriate linkage criterion which determines the proximity between two clusters. Linkage criteria, such as single-linkage, complete-linkage, and average-linkage, dictate whether the distance between the closest points, farthest points, or the average distance between all points in the clusters will be used as the merger criterion respectively. This choice significantly influences the shape and depth of the clusters formed. In several cases the total number of clusters for the algorithm to predict, can also be defined, by configuring the n cluster attribute.

3. Running them sequentially.

At first, DBSCAN and Agglomerative clustering were run separately. Then the labeled dataset was split into train and test samples, 80% and 20% respectively. The same galaxies belonging to train set for one procedure, was the same for the other. In that way a fair and accurate comparison is ensured among the algorithms evaluation performance and clustering results. When DBSCAN and Agglomerative clustering were combined on a sequential approach, they had to be trained on different training sets in a way to avoid overfitting. So the initial dataset was first divided into 80%-20% train and test samples, and then the train set was further split into 40%-40%. Each half was used to train each clustering algorithm. On all cases the best models were evaluated on the 20% test sample.

Upon dividing the initial dataset into an 80%-20% split, a notable issue emerged. Consider a scenario involving a group of four galaxies. The splitting process could potentially allocate one galaxy to the test sample, leaving the remaining three for training, or the other way around. Under such circumstances, given the present assumption that a cluster needs 4 galaxies to be identified, DBSCAN is likely to classify these isolated galaxies as outliers, leading to the loss of that particular group. So to avoid missing any small clusters of galaxies, the best model acquired by both clustering algorithms was re-fit on the complete set of galaxies. In that way all galaxies were assigned their predicted match. However, this approach inherently risked overfitting, because the proper approach would require the model to be applied on a separate test set. For the purposes this thesis, the primary focus was to examine the behavior of the labeled data and to create a model that best describes them. In this specific context, the potential overfitting is not of great concern.

Hyperparameter tuning process occurs with the use of RandomizedSearchCV python library. Instead of trying out every possible combination of hyperparameters, as in GridSearch, it samples a fixed number of sets from specified distributions, param distribution. The main advantage of this approach is that it can be faster and more efficient than an exhaustive search, especially when the hyperparameter space is large. Another important usage is that it incorporates cross-validation, a vital technique in ensuring the robustness of the model's evaluation. Cross-validation involves partitioning the training data into a specified number of folds. Instead of training the model on the entire train set and then evaluating on a separate test set, the train set is divided e.g in 5 folds (cv=5) (Fig. 4), and the model is trained on all but one of these folds, and validated on the remaining fold. This process is repeated until each fold has served as a validation set. By doing so, it reduces the risk of overfitting and provides a more comprehensive assessment of the model's performance on unseen data. Last but not least, it requires a scoring parameter which will evaluate the performance of the hyperparameter combinations.

Table I. Notation Table

Notation	Description
eps	Minimum density threshold
min_samples	Minimum number of points
metric	Method to calculate distances
distance_threshold	Linkage distance to stop merging clusters
n_cluster	Number of clusters found
affinity	Metric to compute distances
linkage	Proximity between observations
param_distribution	List of parameters
cv	Number of cross-validation splitting
scoring	Evaluation of performance

Another way for someone to get insights about the influence of different values of hyperparameters on the training and validation set is by plotting the validation curves. A validation curve is a tool used in machine learning to provide helpful information about the models complexity, expressed through hyperparameters, and the models performance. By analyzing it, it becomes clear how well the model behaves on unseen data (validation score). If both train and validation scores are low, then the model is *underfitting*, meaning that its to simple or has been regularized too much. When the training score is much higher than the validation, the model fits very well the training data but fails to generalize to new input data. In this case the model is *overfitting*. The ideal

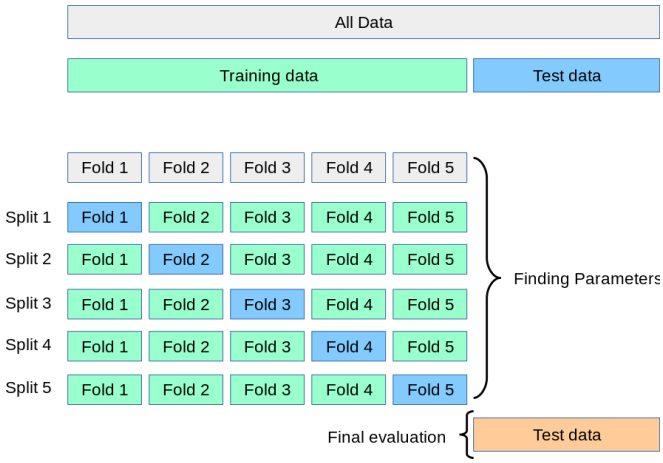

Figure 4. Illustration of cross-validation procedure.

case is when the model is *just right*, where it fits the training set very well while also generalizing new input data. In such cases the scores between the two curves are comparable. In summary, validation curves can be used, only, to illustrate the impact of hyperparameters, it should not be used to tune the model.

IV. Evaluation

Given the luxury provided by the ground truth labeled galaxies, a suitable assessment metric would be the creation of a custom scoring function which would evaluate the agreement between ground truth and predicted galaxy cluster labels.

Here is an overview of the function's implementation

- For a true cluster in the labeled set, the function finds the galaxies residing in it.
- It extracts the predicted cluster labels of these galaxies.
- It scans all the predicted clusters.
- Computes the true positives, false positives and false negatives by assuming two classes, one for the specific predicted cluster being investigated and all the others as a second class.
- Computes the F_1score for each predicted cluster and stores the predicted cluster with the highest value.
- When the scan is completed, the function continue to the next true cluster following the same procedure.
- Finally, for all true clusters in the labeled set, it calculates the mean between all maximum $F_1scores$.

where F_1score is a measure of a test's accuracy and was calculated by

$$F_1score = \frac{2 * TP}{2 * TP + FP + FN}$$

As true positives were defined the predicted galaxies that were correctly identified by the algorithms to be part of a specific true cluster. As false positives denoted the predicted galaxies that the algorithm incorrectly identified as part of a specific true cluster. And as false negatives were the predicted galaxies that the algorithm failed to identify as part of a specific true cluster.

The custom scoring function was designed to assess the efficacy of clustering predictions on labeled data. The objective here was to find the optimal set of model parameters or "hyperparameters" that maximize the agreement between the predicted and true clusters. For each true cluster, the function identifies which predicted cluster has the highest overlap of data points, quantified by the highest F_1score. After calculating the highest scores for each true cluster, the function computes their average, returning this average score as a comprehensive metric for evaluating the performance of a clustering model. The optimal hyperparameter set is then determined as the one that maximizes this mean score across all true clusters, ensuring the most accurate representation of the underlying data structure is achieved. This entire process constitutes an optimization problem aimed at maximizing the model's predictive accuracy in assigning galaxies to their true clusters.

V. RESULTS

In this section, the clustering procedures that resulted in 3D spatial galaxy clustering for all three clustering techniques are presented. Two different metrics were tried to be optimized during each clustering procedure. The first was particularly concentrated on the optimization of the clustering algorithms performance (F_1score) , by finding their best hyperparameter combinations. The second was that, the model with the best hyperparameter set should reproduce the same number of clusters, as know by ground truth. The results on how the different clustering techniques classify the predicted galaxies, is presented on figures 11-14 (Fig. 11-14), on a selection of five well known clusters, such as Virgo, Coma, Hydra, Abell 2197 and Abell 2199.

A. DBSCAN

The primary goal was initially to conduct hyperparameter tuning on DBSCAN and explore its hyperparameter space in order to optimize the algorithm's performance according to the custom scoring function. The minimum number of galaxies was determined to be 4

(min_samples=4), a decision grounded in the assumption that a minimum of four galaxies is required for the identification of a group. The distance metric that has been used between galaxies is the one mentioned before on methods section (3.1). The exploration was particularly focused on finding the most effective value for the eps parameter.

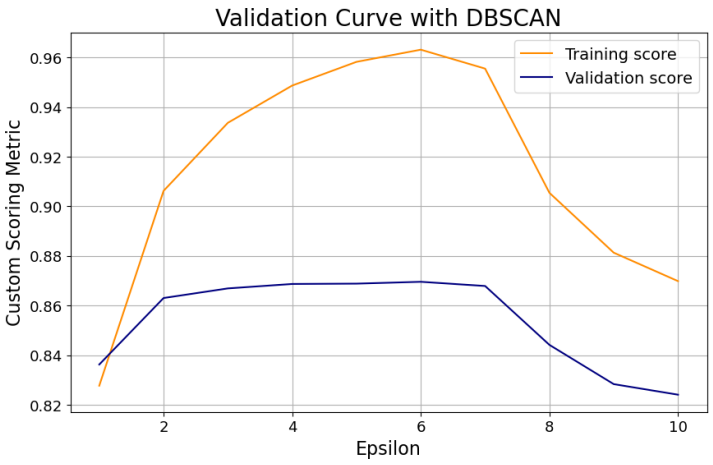

Figure 5. Validation curve of DBSCAN.

As illustrated on the validation curve (Fig. 7), the model peaks on epsilon values ranging from 2 to 6. The fact that the range between training and validation curves from epsilon values of 2 to 10 is comparable, should be noticed as well. A cross-validation process was then implemented on the training set (80% of whole labeled set). For a number of folds equal to five, $cv = 5$, the best value of epsilon obtained was 6, a value that peaked the validation score's curve. Despite of that, a value of 2 seemed to be a more appropriate choice, since the model there starts to converge.

After the best hyperparameter combination was obtained, the model was fit again and predicted the labels of the whole labeled data set (100%). This resulted on a total number of predicted cluster labels equal to 190, as expected from the ground truth. The reason why the best model was fit and predicted on the whole labeled data is the one described on the previous section as well. For a 80%-20% split, clusters with 4 members will probably not be discovered by DBSCAN. Hence, the predicted labels corresponding to the test set were extracted from the whole predicted label list in order to be evaluated. The final performance of the model on unseen (test) data was 91.8%.

B. Hierarchical clustering

In the configuration of Agglomerative clustering algorithm, the total number of clusters (n clusters) parameter was assigned a value of 'None', enabling the algorithm to find the optimal number of clusters corresponding to a varying distance threshold value. The method to calculate the distance between galaxies, affinity, was provided by relation (3.1). Furthermore, the linkage criterion, responsible for determining the proximity between sets of observations, was set to single, so that the algorithm to use the minimum of the pairwise distances between galaxies in two clusters, as the cluster distance.

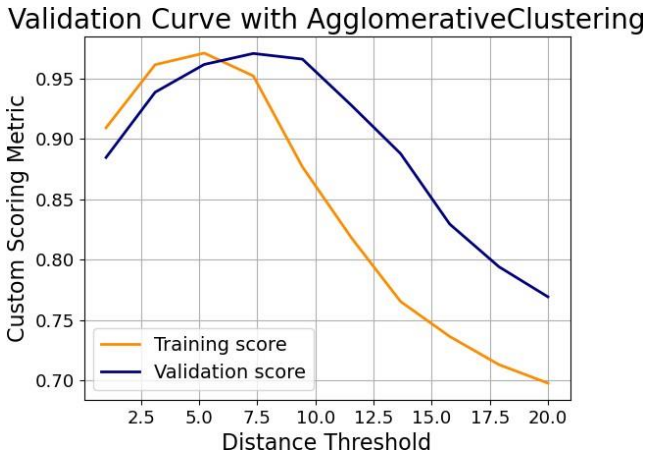

Figure 6. Validation curve of AgglomerativeClustering.

A cross-validation procedure was performed in order to find the optimal distance threshold value that maximized the custom scoring function. The train set was split into 5 folds and the result of the process was that the best distance threshold value was around 6. Then the best model was acquired and fit on the whole labeled data for the predicted labels of all galaxies to be computed. 769 cluster labels were obtained for all labeled galaxies, way higher than ground truth. As observed on the validation curve, a distance threshold which resulted a number of predicted cluster labels $\sim$ 190, was be around 12.5. But its crystal clear that for values higher than 6 the model is overfitting.

The predicted labels of the test galaxy set, were extracted, in order to evaluate the best model acquired by RandomizedSearchCV process. The evaluation resulted in 94.1%.

C. Sequential method

For the sequential approach, DBSCAN and Agglomerative Clustering were trained on distinct datasets. The initial dataset was first divided into 80%-20% train test samples, and the train set was further split into 40%-40%.

The first 40% part underwent training with the DBSCAN algorithm, resulting an epsilon value of 2. Then the best DBSCAN model was applied on the second 40%, by fitting it and predicting its galaxy cluster labels. In that way, galaxies that were assigned as outliers were dis-

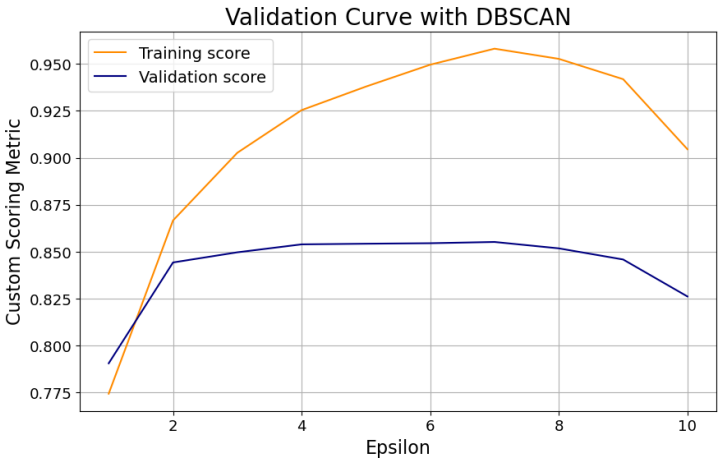

Figure 7. Validation curve of DBSCAN on sequential method.

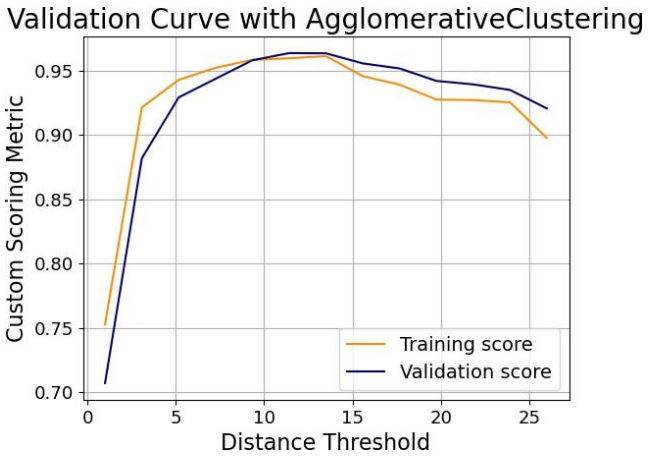

Figure 8. Validation curve of Agglomerative clustering on sequential method.

carded from this set. The new 'cleaned' galaxy set was used to train the Agglomerative clustering algorithm, resulting a distance threshold of 5. The final step was to apply this best agglomerative clustering model on the whole labeled set to obtain the predicted galaxy cluster labels, resulting 1028 different cluster labels. As mentioned before, the evaluation was implemented on the test set that was initially left outside of the whole training procedure. By extracting the test sets predicted galaxy labels, the performance according to the custom scoring function achieved a score of 97.4%.

The best hyperparameters of each clustering method is presented on Table II. To assess the performance of clustering algorithms in classifying predicted galaxies as part of ground truth clusters, a comprehensive examination was conducted focusing on five well known galaxy clusters: Virgo, Coma, Hydra, Abell 2197, and Abell 2199 (Fig. 11-14). This 3D representation aimed to discern the efficacy of the algorithms in accurately categorizing the predicted galaxies and discerning whether they truly belong to the aforementioned clusters, serving a robust measure of the algorithm's precision and reliability in the task of astronomical clustering. The analysis concentrated on these well-established clusters to ensure the validity and reliability of the clustering results, thereby providing insights into the appropriateness of the applied algorithms in the context of astronomical clustering.

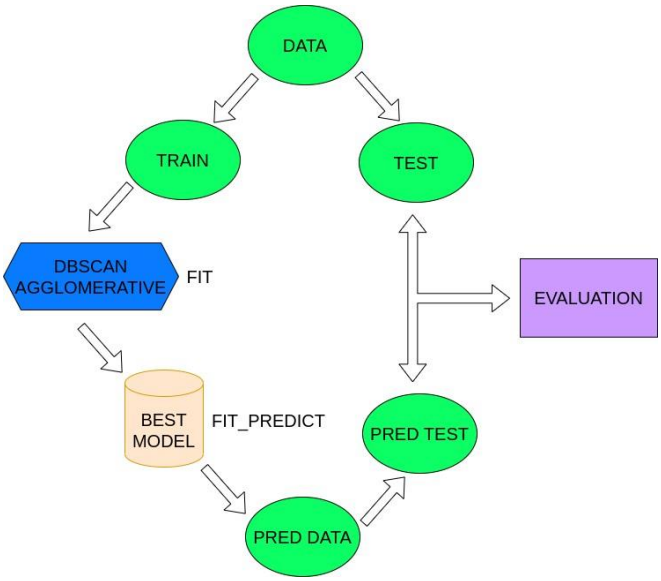

Figure 9. Summary of DBSCAN and Agglomerative clustering processes.

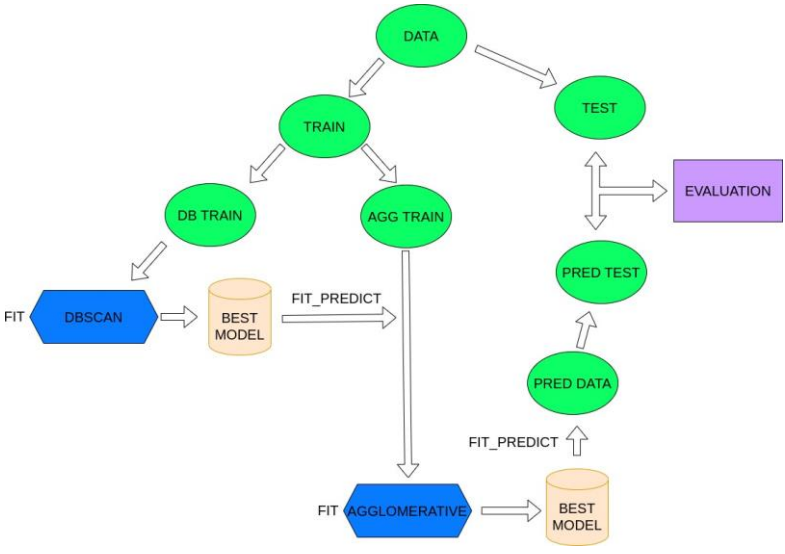

Figure 10. Summary of DBSCAN and Agglomerative clustering processes on sequential approach.

VI. DISCUSSION

Studying the results of the DBSCAN method, as observed in its validation curve, a more suitable choice would be an epsilon value of$\sim$ 6. That epsilon value besides giving the highest validation score also results a total number of 155 clusters, when its fit on the whole labeled set. Moreover, the level of overfitting is higher at that specific value, because the gap between validation and training scores are greater over the value of 2. Since the presence of a plateau, for the validation score, for these two values, the choice of adopting an epsilon value of 2, not only provided the same number of predicted clusters as the ground truth but also achieved a high performance when it was evaluated on the test galaxy set.

As it is seen on figures 11-14, the DBSCAN model performs well in identifying the predicted clusters as structures. Also it can classify projected galaxies as parts of a different clusters (misclassifications) and not as part of the ground truth clusters provided by the known catalogs. Galaxies that are randomly aligned along our line of sight and the cluster being investigated, are correctly identified as misclassifications as well. Regarding Coma cluster's evolutionary state, its known to be in a equilibrium state. This is justified from its spherical shape, but also from galaxies identified as projections at higher redshift. For the case of Abell 2197 and 2199, which appear to be on a merging process, the algorithm cannot distinguish their separation since the galaxies are relatively close to each other.

Agglomerative clustering method, as observed in its validation curve predicts the same number of ground truth clusters at a distance threshold value of 12.5. The problem is that the model at that point clearly overfits. So the choice of a distance threshold value of 6, was more appropriate despite predicting a large number of clusters.

The predicted clusters are well defined as structures also in this case. A problem occurred in the case of Coma and Leo clusters, as observed on the first figure, where both clusters are predicted as merged. Lastly, it is clear that the algorithm cannot identify as much projection galaxies as DBSCAN.

The sequential approach, where the two clustering methods were combined, seems to be a more appropriate method. Predicted clusters are of course well defined (Fig. 11-14). DBSCAN tends to misclassify a larger number of galaxies in clusters that cover a broad redshift range, whereas Agglomerative clustering does not display this behavior to the same level. This difference is especially notable in the case of the Coma Cluster, as illustrated on figures 11-12. This inconsistency can be attributed to the inherent strictness of DBSCAN in excluding more galaxies as projected, contrasted with a more tolerant approach of Agglomerative clustering concerning the redshift dimension. Additionally, DBSCAN seems to exclude galaxies that reside the edges of the extended Virgo cluster and Hydra cluster (Fig. 11-13). However, employing a sequential approach appears to offer a balanced agreement between the tendencies of the two clustering models, bridging the gap and yielding more consistent results.

Given also the fact that the sequential approach achieved a higher scoring of 97.4%, when evaluated on the unseen galaxy set, highlights that is a better approach to be applied on such galaxy data.

VII. CONCLUSION

In this thesis, three different clustering approaches were discussed for the purpose of clustering the whole HECATE catalog. Given the advantage of having a small labeled HECATE sample, led to the semi-supervised procedures proposed.

All of these clustering techniques work quite well with the labeled galaxy set, since they can identify the structures of the clusters and possible galaxy alignments and projections. DBSCAN alone predicts the total number of ground truth clusters and does a better job in identifying misclassifications compared to Agglomerative clustering. Moreover, DBSCAN can be quite strict, on excluding galaxies out of clusters. Agglomerative clustering, on the other hand seems to better detect under structures, i.e subgroups, on the ground truth cluster set, since the predicted number of clusters was higher.

The sequential approach was considered as a stacking of the two clustering models to obtain the ultimate prediction. DBSCAN cleaning the galaxy data set from outliers and then Agglomerative clustering performing the clustering process on that cleaned galaxy set. This approach was implemented due to the desire to explore the combined potential of both algorithms, examining how the sequential application could enhance the reliability and accuracy of the clustering results. The high score of 97.4% indicates that its more effective in predicting the already known clusters, than when DBSCAN or Agglomerative clustering are performed separately.

For the case of galaxies classified as projections, a further study on updated catalogs is required. In new surveys, performed with more precise observations, misclassified galaxies can be identified as part of major clusters. This investigation, especially for misclassified galaxies, will enhance the reliability of the clustering results presented in this thesis.

It is important to mention the fact that there is some overfitting on the final results. When best models were acquired, they were applied on the whole labeled set to obtain a galaxy cluster prediction for each labeled HECATE galaxy. In general, it would be more appropriate to be applied on an unseen galaxy set. But given the nature of the data set that was inevitable.

For future applications on the whole HECATE catalog, and not just the labeled set performed in this study, all

three clustering methods can be applied with the hyper-parameters set from this analysis, hoping that the results will still be significant. It should be noted that clustering the whole 204733 galaxy catalog will be a computationally expensive procedure, and that is why this thesis was limited only on the already labeled galaxies of HECATE.

References

[1] Abell, George O., Corwin, Harold G., Olowin, Ronald (1989) A Catalog of Rich Clusters of Galaxies.

[2] Carvalho, Ana Sofia Chagas (2019) Exploration of Unsupervised Machine Learning Methods to Study Galaxy Clustering.

[3] Carroll, Bradley W., Ostlie, Dale A. (2014) An Introduction to Modern Astrophysics.

[4] Deng, Jiahao, Brown, Eli T. (2022) SSDBCODI: Semi-Supervised Density-Based Clustering with Outliers Detection Integrated.

[5] Hashimoto, Yasuhiro, Liu, Cheng-Han (2022) Cluster Membership of Galaxies Using Multi-Layer Perceptron Neural Network.

[6] Hickson, P. (No Year) Systematic Properties of Compact Groups of Galaxies.

[7] Kim, Suk, Rey, Soo-Chang, Jerjen, Helmut, Lisker, Thorsten, Sung, Eon-Chang, Lee, Youngdae, Chung, Jiwon, Pak, Mina, Yi, Wonhyeong, Lee, Woong (2014) The Extended Virgo Cluster Catalog.

[8] Kovlakas, K., Zezas, A., Andrews, J. J., Basu-Zych, A., Fragos, T., Hornschemeier, A., Kouroumpatzakis, K., Lehmer, B., Ptak, A. (2021) The Heraklion Extragalactic Catalogue (HECATE): A Value-Added Galaxy Catalogue for Multi-Messenger Astrophysics.

[9] McConnachie, Alan W., Patton, David R., Ellison, Sara L., Simard, Luc (2008) Compact Groups in Theory and Practice – III. Compact Groups of Galaxies in the Sixth Data Release of the Sloan Digital Sky Survey.

[10] Malavasi, Nicola, Sorce, Jenny G., Dolag, Klaus, Aghanim, Nabila (2023) The Cosmic Web Around the Coma Cluster from Constrained Cosmological Simulations: I. Filaments Connected to Coma at Z = 0.

[11] Sellami, K, Saied, M. A., Ouni, A. (2022) A Hierarchical DBSCAN Method for Extracting Microservices from Monolithic Applications.

[12] Saha, Rohan (2023) Influence of Various Text Embeddings on Clustering Performance in NLP.

[13] Schneider, Peter (No Year) Extragalactic Astronomy and Cosmology.

[14] Tohill, C., Bamford, S. P., Conselice, C. J., Ferreira, L., Harvey, T., Adams, N., Austin, D. (2023) A Robust Study of High-Redshift Galaxies: Unsupervised Machine Learning for Characterising Morphology with JWST up to Z $\sim$8.

[15] Tran, Thanh N., Drab, Klaudia, Daszykowski, Michal (2012) Revised DBSCAN Algorithm to Cluster Data with Dense Adjacent Clusters.

Algorithm	Hypeparameters	Range	Value	Score
DBSCAN	*eps*	1-10	2.0	91.8%
	min_samples		4.0	
Agglomerative	*distance_threshold*	1-20	6.0	94.1%
Chain	*eps*	1-10	2.0	97.4%
	min_samples		4.0	
	distance_threshold	1-20	5.0	

Table II. Table of different clustering techniques, with their corresponding best hyperparameter values. Range column refers to the range on which hyperparameters were investigated. Scores were obtained by the evaluation of each model on the labeled test set (20%).

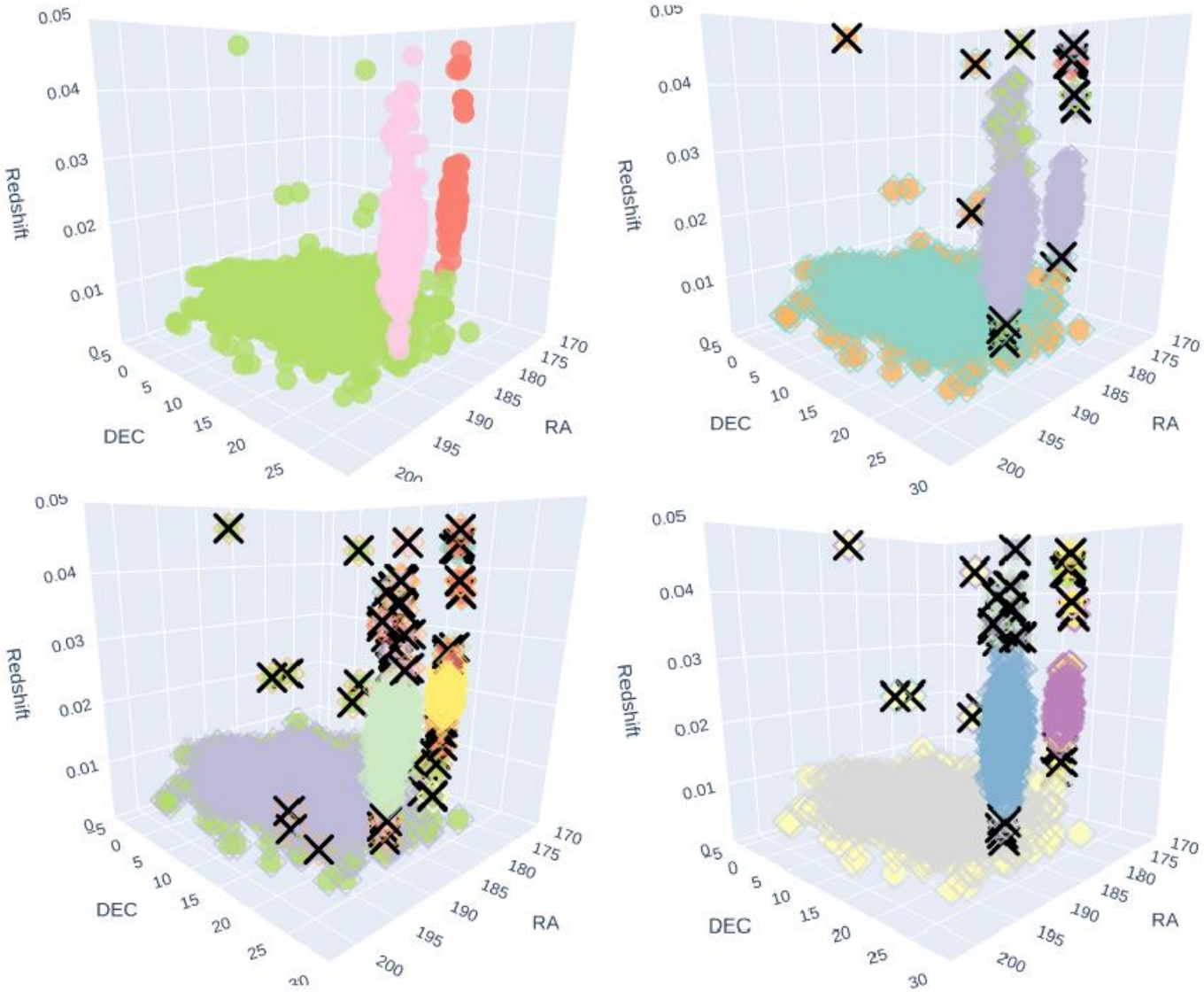

Figure 11. Visualization of galaxy classification and prediction results, where each subfigure corresponds to a different clustering method applied on galaxy clusters. On the top left corner the true galaxies are displayed, serving a reference for accurate classification. Underneath the results of DBSCAN clustering are presented. The top right corner represents Agglomerative clustering results and underneath are the outcomes derived from the sequential method. Actual galaxies are denoted by filled colored circles, whereas predicted galaxies are represented by overlapping empty squares of different color. Misclassified objects are highlighted with an 'X' mark and referring to false positive galaxy measurements. This set illustrates the Extended Virgo Cluster accompanied with Leo (back) and Coma (front) clusters.

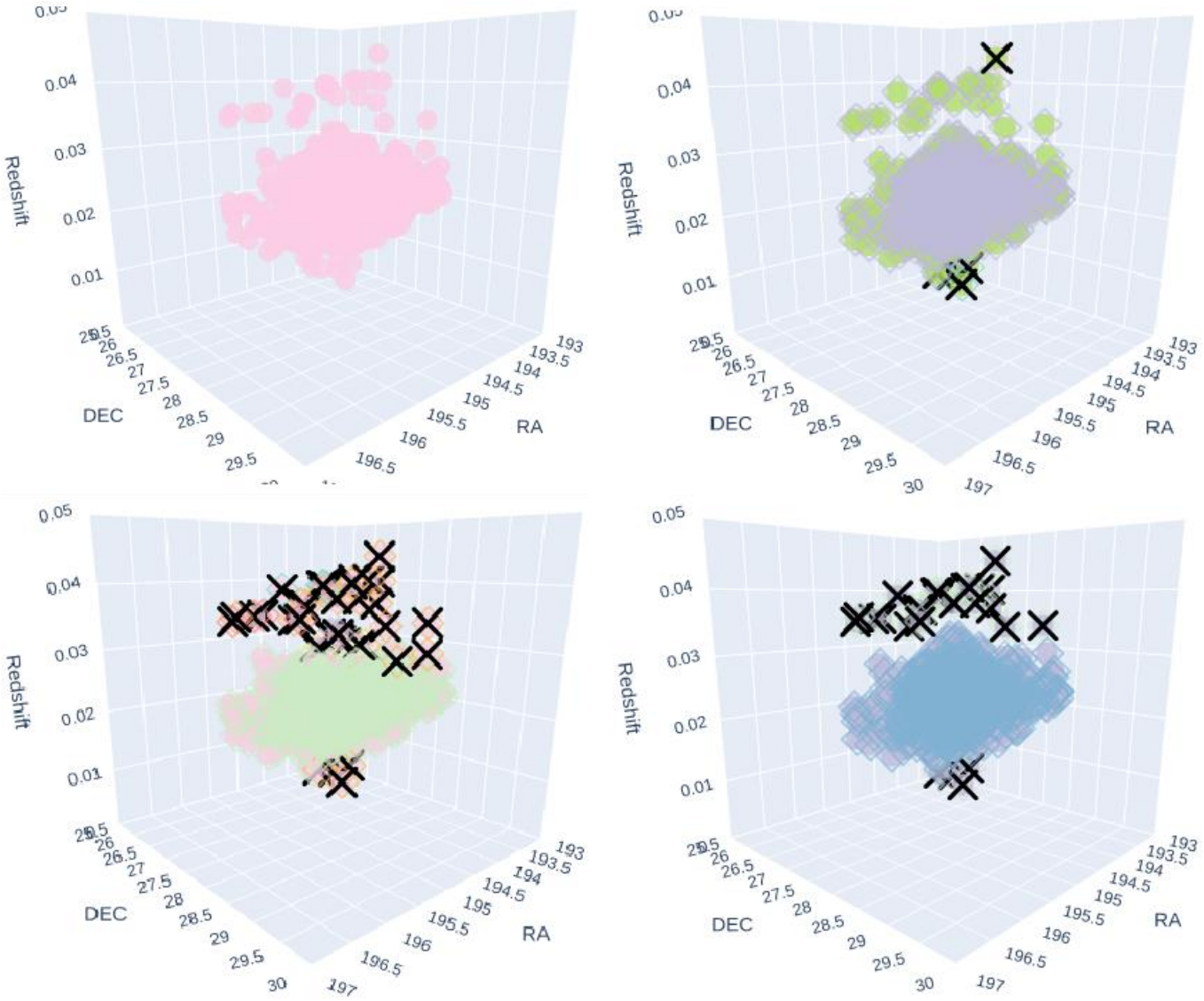

Figure 12. This set illustrates the Coma cluster.

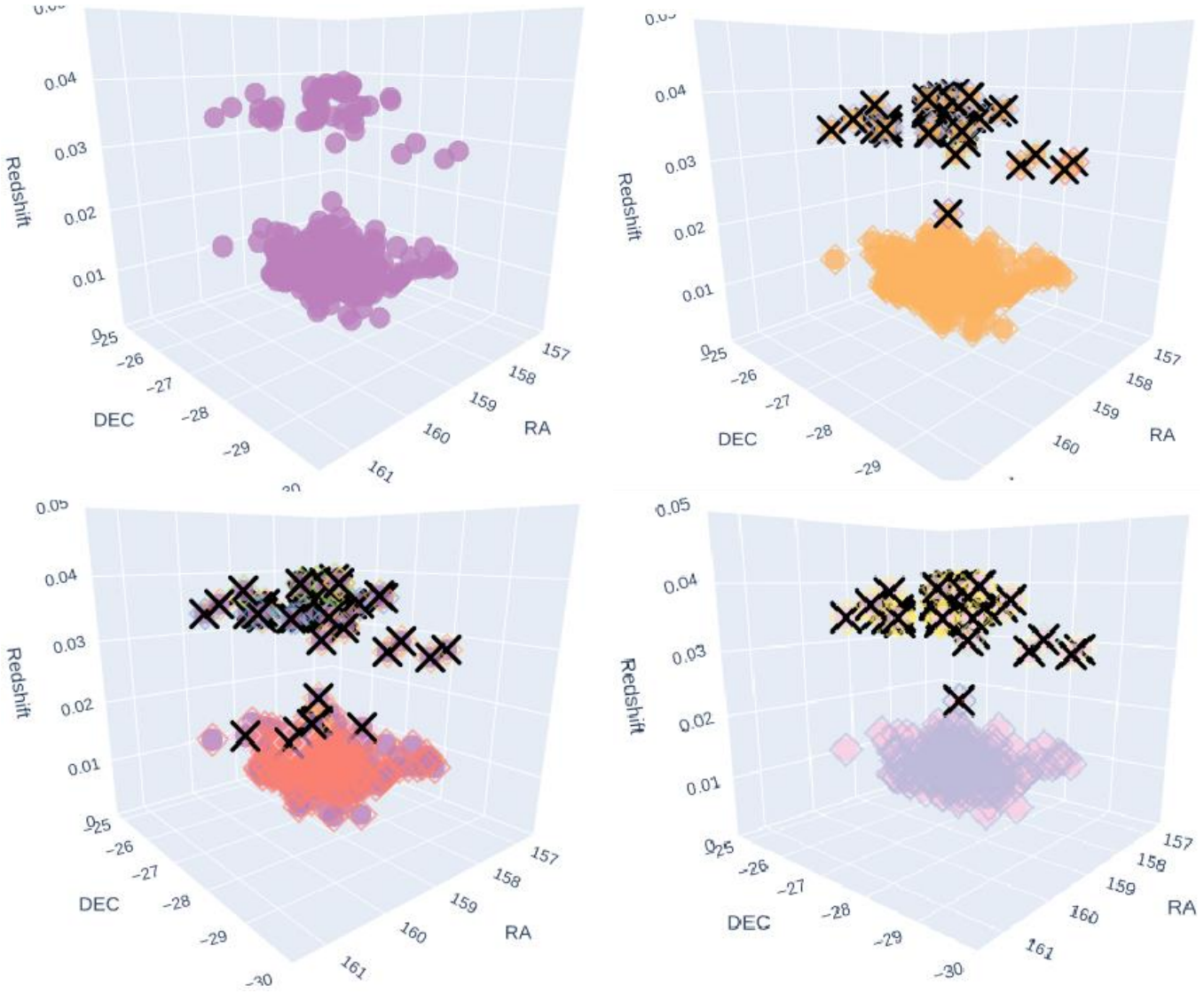

Figure 13. This set illustrates the Hydra cluster.

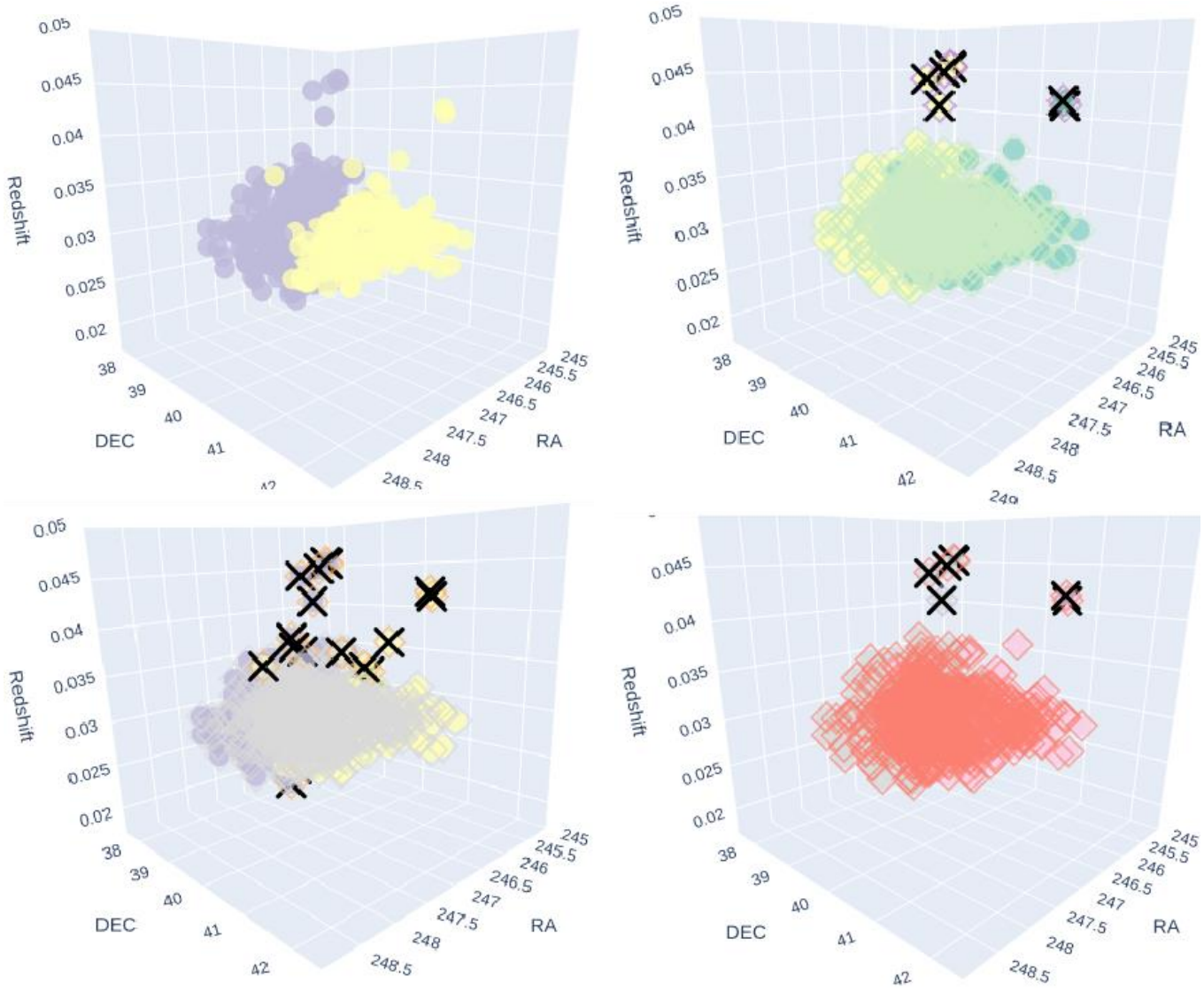

Figure 14. This set illustrates the Abell 2197 and Abell 2199 clusters.