Open University of Catalonia (UOC) Master's Degree in Data Science

MASTER'S THESIS

Area: 4: Data Science

Applying Density-Based Algorithms to Galaxy Cluster Catalogs

Unveiling Galaxy Structure with Unsupervised Clustering

Author: Carlos Toro Peñas

Tutor: Laura Ruiz Dern

Professor: David Masip Rodo

Madrid, November 6, 2025

Copyright

FINAL PROJECT RECORD

Title of the project:	Applying Density-Based Algorithms to Galaxy Cluster Catalogs
Author's name:	Carlos Toro Peñas
Collaborating teacher's name:	Laura Ruiz Dern
PRA's name:	David Masip Rodo
Delivery date (mm/yyyy):	MM/YYYY
Degree or program:	Master's degree in Data Sicience
Final Project area:	4: Data Science
Language of the project:	English
Keywords:	clustering, galaxy clusters, cosmology

Dedication/Quote

Acknowledgements

Abstract

This work primary focuses on apply density-based algorithms to datasets from major surveys, including the Two-degree Field Galaxy Redshift Survey (2dFGRS) and the Sloan Digital Sky Survey (SDSS). The application will be followed by hyperparameter tuning and a performance assessment to identify the algorithms' strengths and weaknesses in actual galactic group detection. In the future, these methods may be applied to new surveys and other celestial regions.

Keywords:clustering, galaxy clusters, cosmology.

Resumen

Este trabajo tiene como tema central la aplicación de diferentes algoritmos basados en densidad a juegos de datos obtenidos de estudios como Two-degree Field Galaxy Redshift Survey (2dFGRS) y Sloan Digital Sky Survey (SDSS). Como resultado de esa aplicación, se hará un ajuste de híper-parámetros así como una evaluación del desempeño de tales algoritmos para analizar sus fortalezas y debilidades en su habilidad para la detección de cúmulos galácticos catalogados. Futuramente, se podrán realizar aplicaciones de estos algoritmos a nuevos estudios y otras regiones del firmamento.

Palabras clave: clustering, cúmulos de galaxias, cosmología.

Contents

List of Figures

Chapter 1

Introduction

1 Context and motivation

When studying the Universe at medium and large scales, we enter the field of galaxy surveys, which rely on dedicated telescopes to obtain large catalogs of galaxies. One objective of these studies is to map vast areas of the Universe. This work relies on data coming from two of these surveys: the Two-degree Field Galaxy Redshift Survey (2dFGRS) and the Sloan Digital Sky Survey (SDSS).

The datasets generated by these surveys are highly suitable for analysis through Machine Learning (ML) methods. Specifically, the redshift feature can be interpreted as a measure of distance by applying cosmological models, as will be detailed in Section **??**.

The matter distribution in space is far from random; instead, galaxy groups represent the next fundamental structure in the Universe above the level of individual galaxies. At an even higher structural level we find clusters, which are more numerous aggregations of galaxies composed of hundreds or thousands of members. These structures are shaped by dark matter into filaments and voids [15] to form the so-called cosmic web [2].

Determining the structure of groups and clusters is therefore crucial for understanding the distribution of matter in the Universe. This is where clustering algorithms come into play. As will be discussed in section **??**, the density-based algorithms are the most appropriate for this study.

1.1 Personal motivation

Given my personal background and strong interest in astrophysics and cosmology, upon entering the field of Data Science, one can find the vast potential for applying the multiple Machine Learning (ML) techniques to these scientific domains. The study of the Universe's large-scale

structure, in particular, stands out as a critical area where ML methods can yield substantial scientific advancements.

2 Goals

There are two list of goals we considered to address separatedly:

2.1 Main goals

- Apply density-based algorithms to galaxy datasets adquired from 2dFGRS and the Sloan Digital Sky Survey SDSS, in order to obtain a validated model that can effectively approximate the observed structure of groups and clusters.
- Determine which of the applied algorithms perform best and investigate the reasons for their effectiveness..
- Detect potential outliers and patterns.
- Use validation methods to obtain a hyper-parameter tunning in order to optimize galaxy group/cluster detection.

2.2 Secondary goals

- Generate a visualization map of the data used in this study.
- Detect methods to improve this study in future works in following areas: Data Enhancement, Algorithmic Refinement[1].

3 Sustainability, diversity, and ethical/social challenges

Cosmological findings fundamentally change our understanding of humanity's place in the Universe. Discoveries related to dark matter, dark energy, or the vastness of the cosmos can have profound philosophical implications.

Sustainability The most direct social responsibility implication lies in the immense power required to process and store astronomical data. This work while purely theoretical, relies on an infrastructure that carries a heavy sustainability burden. That is why focus

[1]For exemple: modify the distance in order to mitigate the already known Redshift-Space distortions along the line of sight[20].

on computational efficiency directly translates into lower energy consumption, this is the most tangible sustainability implication in this project.

Ethical behaviour and social responsibility The major impact on this matter concerns the use of resources. This project mitigates resource impact by using only shared, openly licensed libraries and datasets whose usage terms are fully respected by the authors. Of course, all references to the utilized datasets and other previous works are properly cited, given that this project relies fundamentally upon them.

Communicating the outcomes clearly and accurately is also a commitment from the authors of this work.

Diversity, gender and human rights Astronomy and cosmology, like many sciences, have historically struggled with issues of diversity and inclusion gender and human rights matters[2],

The authors are committed to respecting these questions throughout this work. More generally, the further advancement in science benefits society by better equipping it to address issues on these matters.

To conclude this section, this work uses powerful analytical methods derived from ML techniques, all tools could be adapted for surveillance, military intelligence, or other uses that might infringe on human rights or privacy. Scientists must be mindful of how their methods and code are shared.

4 Approach and methodology

We will apply classical phases drawn from the data life-cycle, which cover:

- Collection: download datasets drawn from surveys such as the SDSS and 2DFGRS to generate galaxy clustering models. These datasets are available at [10]:

  https://gax.sjtu.edu.cn/data/Group.html

- Storage: keep donwnloaded data set in csv files.

- Preprocessing: stage containinig the tasks of cleaning, filtering, sampling, and fusion.

- Analysis stage: which includes model building through the application of the algorithms and validation of the outcomes.

[2] An example in gender matter can be seen in the eighth chapter of the documentary television series Cosmos: A Spacetime Odyssey, titled "Sisters of the Sun," hosted by Neil deGrasse Tyson.

- Visualization: graphical view of the results.

The Analysis Stage will utilize an iterative methodology dedicated to enhancing the robustness and accuracy of the model outputs. All models employed in this stage will consist of unsupervised algorithms, with a particular focus on density-based clustering techniques to identify structures and outliers within the data.

To evaluate the performance of these models, the following criteria will be followed:

- Detected Clusters: clusters successfully classified (often referred to as True Positives at the group level).
- Undetected Clusters: clusters not found or not identified in the output-clusters set (equivalent to True Negatives at the group level).
- Cluster Purity Ratio: proportion of members in a output-cluster that actually belong to the underlying cluster/group structure.
- Cluster Completeness Ratio: proportion of members of a true underlying group/cluster that are successfully included within the detected output-cluster.
- Misclassified Members: individual data points (galaxies) belonging to an actual structure but classified outside of any detected output-cluster (often referred to as False Negatives at the individual member level).
- External Data Classified as Members: individual data points (galaxies) not belonging to any actual group but erroneously classified inside a detected output-cluster often referred to as False Positives at the individual member level).

The computational work for this study will primarily utilize Python, with supplementary analysis performed using R.

5 Schedule

A Gantt diagram in figure 1.1 shows the different stages of the project development. Excluding the final project defense, the stages have been grouped in three blocks:

- Planning (shown in green) involves gathering resources and defining the project's objectives.
- Technical development (shown in red) includes design, data processing, method application and outcomes assessment.

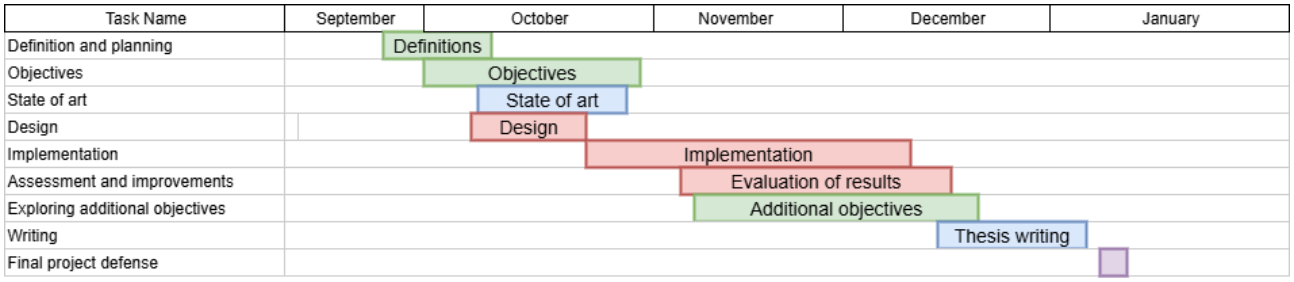

Figure 1.1: Stages of the project.

- Research and writing (shown in blue).

An iterative and continuous review of the results will be performed throughout the analysis process due to several causes: issues stemming from the algorithms, data processing, and the workflow itself. As a result, initial objectives might be rearranged and redefined. This is why aditional objetives stage is necessary.

Chapter 2

State of the art

This chapter serves to establish the current academic context for the research area addressed by this work. This is achieved by first defining the target objects—galaxy clusters and groups—and subsequently providing a comprehensive overview of the redshift surveys that furnish the requisite data. The chapter concludes with a concise review of the Machine Learning techniques, particularly the unsupervised algorithms, that will be applied throughout this study.

1 Galaxy groups and clusters: Target objects for structure identification

Galaxy groups 2.1 and clusters 2.2 are the largest gravitationally structures in the Universe and are crucial probes of the underlying cosmic dark matter density field. They both consist in dark matter halos containing multiple galaxies, their distinction is typically based on mass and membership:

1. *Galaxy Groups*: These are the most common and lowest-mass virialized systems, typically containing 3 to 50 member galaxies [20] and spanning a total mass range of $\sim 10^{13} - 10^{14} M_{\odot}$. Our own Local Group is a well-known example.

2. *Galaxy Clusters*: These represent the high-mass tail of the halo distribution, typically containing hundreds to thousands of galaxies, with total masses ranging from $\sim 10^{14} - 10^{15} M_{\odot}$. They often host a dominant, massive Brightest Cluster Galaxy (BCG) at their center and are strong emitters of X-rays and which properties dictate cluster formation and evolution [20]. A paradigmatic example is the Virgo cluster with an estimate mass of $\sim 1.21015 M_{\odot}$ and having M87 as most massive and dynamically dominant central galaxy.

Figure 2.1: Future group image

Figure 2.2: Future cluster image

Because both groups and clusters exhibit diverse morphologies and spatial shapes, the subsequent analysis —detailed in Section 4.2— employs unsupervised density-based algorithms. These methods are the most suitable approach for the clustering analysis due to their capacity to effectively fit structures with arbitrary geometries.

2 Spectroscopic Surveys

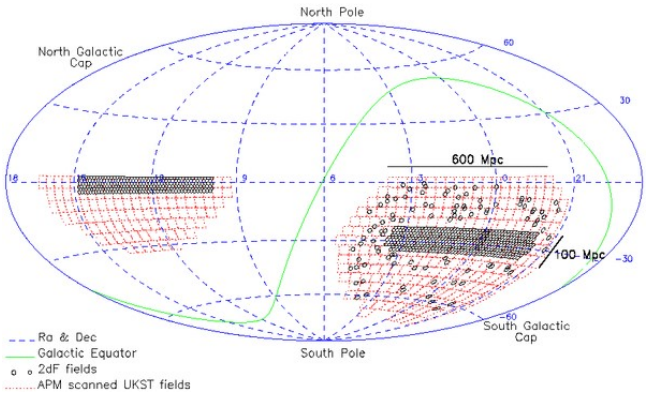

Figure 2.3: 2dfGRS sky coverage
Source [7].

Spectroscopic surveys are fundamental projects in astronomy and cosmology that collect and analyze the spectrum of light from a large number of celestial objects over a wide area of the sky. By splitting the light into its constituent wavelengths, these surveys acquire a vast amount of detailed physical information for each object.

One main purpose of this kind of studies is to obtain a highly precise redshift (z) of each object in order to estimate distances. Therefore it is possible to construct a three-dimensional map of the Universe.

The spectrum serves as well as a unique fingerprint of the source, allowing for the determination of its physical properties. Spectral analysis provides detailed information on the object's chemical composition, temperature, density, and internal motion.

We are fortunate that, nowadays, we have access to data from several major astronomical

surveys, including, but not limited to, the following:

2.1 2dF Galaxy Redshift Survey (2dfGRS)

2dfGRS is a Survey leveraged the unique capabilities of the 2dF (2-degree Field) facility built by the Anglo-Australian Observatory in the southern hemisphere (see in figure 2.4). A view of 2dfGRS coverage is shown in figure 2.3. Data was collected between 1997 and 2004.

Figure 2.4: Australian Astronomical Observatory (AAO)
Source [5].

The data set employed in this analysis originates from the 2003 final data release, which encompasses a total of 245,591 objects. After quality cuts, 221,414 objects were determined to be spectroscopically reliable galaxy data, thus forming the foundation for the subsequent investigation.

2.2 Sloan Digital Sky Survey (SDSS)

The Sloan Digital Sky Survey (SDSS) [19] began collecting data in 2000 and is one of the largest and most influential astronomical surveys ever conducted. The primary goal is to comprehensively map the Universe to expand our understanding of its large-scale structure, the formation of stars and galaxies, and the history of the Milky Way. It uses a dedicated wide-angle optical telescope (the Sloan Foundation 2.5-m Telescope) at Apache Point Observatory in New Mexico (see figure 2.5), and in later phases, also observations in the Southern Hemisphere.

The SDSS has progressed through several phases (SDSS-I, II, III, IV, and the current SDSS-V), with each phase introducing new scientific goals and technological advancements. For this work we will use the modelC petrosian magnitude data realease 7 (DR7) which contains 639359 galaxy entries. This release offers coverage for approximatyely one quarter of the sky sphere, predominantely in the nothern galactic cap as illustrated in figure 2.6. Groups constructed are drawn up from [11].

Figure 2.5: SDSS Data release 7 sky coverage
Source [5].

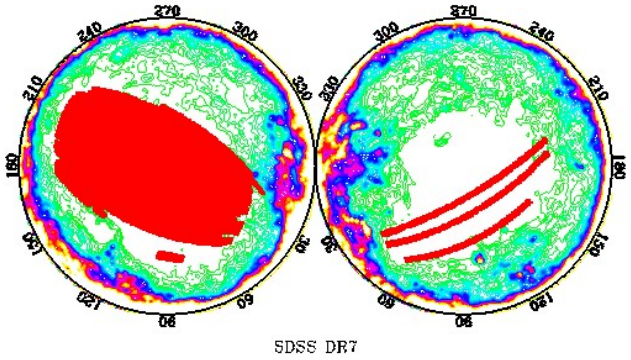

Figure 2.6: SDSS Data release 7 sky coverage
Source [11].

2.3 Inherent challenges associated with survey-collected data

Beyond the substantial logistical and resource constraints associated with constructing and operating large-scale astronomical facilities, all galaxy redshift surveys are subject to a distinct set of fundamental systematic and technical challenges. They can be broadly categorized into those arising from by instrumental and observational limitations, and those derived from subsequent collecting data and analysis. Among these, we can cite the following:

1. Observational and measurement errors: produced by instrumental limitations due long time exposures, observing faint objects in high redshifts where long integration times are required to achieve an adequate signal-to-noise ratio (SNR). Additional noise is introduced by environmental factors, such as atmospheric distortion (seeing) or the inherent difficulty of performing accurate sky subtraction during data processing.

2. Systematic errors and biases, for example the already mentioned distorsion on redshifts caused by peculiar velocities of galaxies. This effect leads to the apparent elongation of clusters along the line of sight, a well-known phenomenon often referred to as the "Fingers of God" effect [1]. There is also a bias caused by the galaxy type, luminosity or epoch of the Universe can also lead to innacurate outcomes in the survey.

3. Theoretical and modeling uncertainties represent another category of challenges, primarily stemming from the interpretation of observational data. These uncertainties arise both from limitations in the underlying theoretical models and from poorly constrained baryonic effects that impact dark matter simulations.

2.4 Galaxy cluster/group catalog

In order to validate and assess the performance of a constructed galaxy clustering model, it is necessary to compare the results from the target galaxy catalog (the data being clustered) against a well-established, pre-existing group/cluster catalog (the ground truth). For this work, we employ the catalog result obtained by the halo-based group finder developed by in [11]. This method is specifically optimized for grouping galaxies residing in the same dark matter halo and and utilizes an iterative approach based on an adaptive filter modeled after the general properties of dark matter halos (see [11]) for full details).

The halo-based group finder was successfully applied to the galaxy catalogs from both the 2dFGRS and the SDSS (Data Release 7). This valuable group catalog is publicly accessible and can be downloaded from [6]

3 The redshift–distance relation

It is well-established that the Universe is undergoing cosmic expansion. The Hubble–Lemaître Law quantifies this expansion, stipulating that the recessional velocity of a galaxy is directly proportional to its proper distance from the observer. This relationship is described by the equation:

$$V = H_0 D \tag{2.1}$$

Where V, H_0, D are respectively, the velocity, Hubble constant and distance.

The equation 2.1 is strictly valid on small redshifts $z \ll 1$ [4] (which means nearby objects). In higher redshift it is necessary to use a full cosmological model to address the redshift-distance relation. According with the most recent theories [4]:

$$D_p = \frac{1}{H_0} \int_0^z \frac{dz}{\sqrt{\sum_i \Omega_{r0}(1+z)^4 + \Omega_{m0}(1+z)^3 + \Omega_{\Lambda 0}}} \tag{2.2}$$

.

Where $\Omega_{r0}, \_, \Omega_{m0}, \_\Omega_{\Lambda 0}$ represents the density parameters for radiation, mass and dark energy (respectively) in the present epoch $z = 0$.

The specific form of equation 2.2 may vary according to the chosen cosmological model. In this work we assume the values of the ΛCDM, according to [4]:

$$\Omega_{r0} = 0.0001,\ \Omega_{m0} = 0.3,\ \Omega_{\Lambda 0} = 0.7. \tag{2.3}$$

.

4 Machine Learning applied to cosmology

We will present a brief description of Machine Learning algorithms emphasizing those used in this work.

4.1 Supervised methods

Supervised learning focuses on identifying patterns and relationships within labeled datasets. The primary objective of supervised methods is to extract knowledge from the given training data to enable accurate class predictions for new, unseen data. Formally, given a labeled dataset $Z = (X, Y)$, where $X = (X_1, \ldots, X_n)$ are the input features and $Y = (Y_1, \ldots, Y_m)$ are the corresponding labels, the goal is to find a function F such that the relationship $Y = F(X)$ is approximated.

A subset is taken from the original dataset, the so called training data $Z_i = (X_i, Y_i)$. And then the problem is reduced to find the minumum of a loss function, which measures the difference between Y_i and $F(X_i)$.

The input to any supervised algorithm consists of independent variables (or features), and the output comprises the dependent variables (or target variables). Supervised algorithms leverage the information within the training data to learn the intricate relationships between these input and target variables.

However, a detailed discussion of supervised methods is not the scope of this work. We are not interested in making target predictions; instead, our objective is to identify patterns and structure within the data distribution, which will subsequently inform the spatial distribution of matter within a dimensional space.

4.2 Unsupervised methods

Unsupervised learning focuses on analysis and modeling of data that lack output classes or pre-existing labels. This methodology aims to discover intrinsic structure, patterns, and relevant features within the data itself.

Formally, the input consists of a set of observations (or data points) where the feature matrix X is given by $X = (X_1, \ldots, X_n)^T$. The primary objective is to learn the underlying distribution or to find meaningful representations from these input variables without any prior guidance.

From the unsupervised methods set we have: clustering and segmentation. These methods work based on distance and similarity patterns and can be divided as follows:

- *Hierarquical*: this method creates successive partitions of data and a hierarchical tree, called a dendrogram. Examples include agglomerative clustering.
- *Partitional*: an initial set of clusters must be set in advance, the set is improved on an iterative proccess. Example k-means.
- *Model-Based*: these algorithms assume that the data is generated by a mixture of underlying probability distributions (e.g., Gaussian Mixture Models, GMM).
- *Density-based*: this method defines clusters as contiguous regions of high density separated by regions of low density (e.g., DBSCAN).

The key advantage of density-based methods is the fundamental lack of a priori assumptions regarding the underlying data distribution. These algorithms operate by defining clusters as contiguous, dense regions of data points that are separated by sparser areas. This characteristic makes them highly suitable for exploratory data analysis, as they impose no constraints on the shape of the resultant clusters.

A further feature of density-based methods is their intrinsic ability to detect outliers or noise points. These points typically reside in the low-density regions that naturally separate the dense clusters, allowing for robust identification of anomalous observations without a dedicated process.

Conversely, many hierarchical and partitional algorithms rely on strong assumptions about the data's structure. For instance, the k-means algorithm requires the number of clusters (k) to be predefined and implicitly assumes that the clusters follow a globular or spherical shape (often analogous to a Gaussian probability distribution). This inherent bias makes them unsuitable for astronomical data, where the spatial distribution of matter is expected to exhibit arbitrary, non-spherical geometries—such as linear filaments, stellar-like distributions, or complex polygonal structures. Thus, these restrictive methods are not appropriate for our analysis.

For this study, we have selected three representative density-based algorithms: OPTICS, DBSCAN, and HDBSCAN. The following section will provide a detailed overview of each method.

4.3 OPTICS

Namely Ordering Points to Identify Cluster Structure: is a density-based, unsupervised algorithm. Its primary mechanism involves ordering the data points based on their reachability distance relative to a specified density threshold.

The output of OPTICS is not a finalized set of clusters but rather a visual tool called the reachability plot (or reachability-distance graph). This plot encodes the density structure of the dataset, from which clusters of varying density and hierarchy can be later extracted.

Let us define the foundational geometric concepts required to understand the OPTICS algorithm.

- *eps-neighborhood* of a point p in S is $NE_\epsilon(p) = \{q \in S : dist(p,q) \leq eps\}$. Then any eps-neighborhood of p is said to be dense if $|NE_\epsilon(p)| \geq minPts$.
- The *core_distance* of a given point p is the minimum ϵ such us $NE_\epsilon(p)$ is dense, in other words:

$$core - distance(p) = min\{\epsilon : |NE_\epsilon(p)| \geq minPts\}$$

- A point is said to be a *core-point* when $NE'_\epsilon(p)$ is dense and $\epsilon' \leq \epsilon$, finally,
- The *reachability-distance* from q regarding a core-point g is the maximum of the two: core-distance and Euclidean distance, in other words:

$$reachability - distance(p,q) = max\{core - distance(p), dist(p,q)\}$$

Note that reachability-distance is only defined with respect to a core-point. We can see an illustrative example of both core-distance and reachability-distance in the figure 2.7.

OPTICS work by setting up two mandatory parameters:

1. Eps (ϵ): The maximum radius to search for neighbors.
2. minPts: The minimum number of neighbors a point needs to have to be considered a core-point.

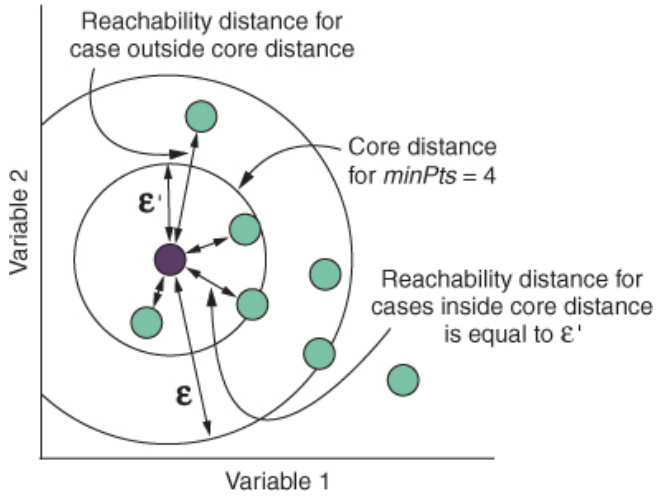

Figure 2.7: Core and reachability distances
Source [3].

For example, figure 2.8 shows a random-generated set of points around four fixed points within [0,1]x[0,1] square. OPTICS is then applied to this, and the resulting reachability plot is shown in figure 2.9.

4.4 DBSCAN

DBSCAN (Density-based Spatial Clustering of Applications with Noise) is another density-based clustering algorithm that leverages several concepts from OPTICS to efficiently extract clusters. However, DBSCAN introduces specific, additional definitions for identifying points and cluster boundaries, which are summarized below.

Given a dataset S, a minimum number of points MinPts, and a neighborhood radius Eps, let p be a core-point of S. Then:

- A point q is defined as *directly density-reachable* from a core-point p if q is within the ϵ-neighborhood of p (i.e., $q \in NE_\epsilon(p)$). This definition is valid only when p satisfies the core-point condition: $|NE_\epsilon(p)| \geq$ MinPts
- A point q is said to be *density-reachable* with respect to Eps and MinPts if there exists an ordered sequence of points $p_1, ..., p_n$ such that:
  1. $p_1 = p$ and $p_n = q$.
  2. p_{i+1} is directly density-reachable from p_i for all $1 \leq i \leq n$.
- The point p is *density-connected* to a point q with respect to Eps and MinPts if there exists third point o such that both p and q are density-reachable from o.

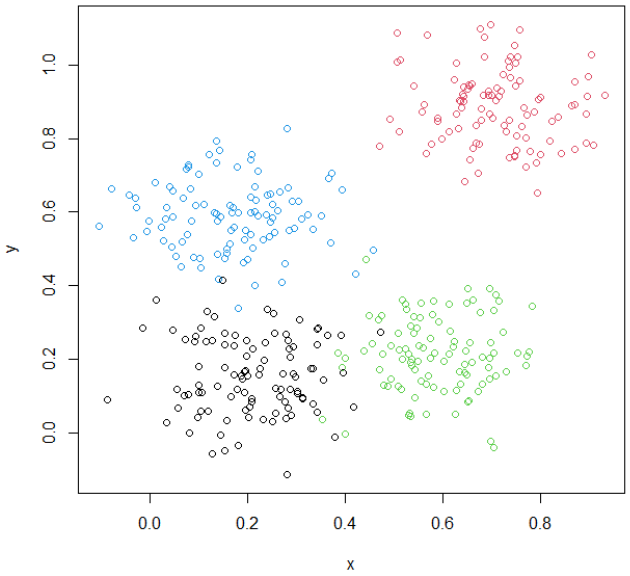

Figure 2.8: An example of data set in plane $\mathbb{R}^2$.

The figure 2.10 illustrates both concepts: density-connectivity and density-reachablability. Then a cluster C is a subset of S satisfying:

- $\forall p, q, \in C$ p is density-connected from q with respect to Eps and MinPts.
- $\forall p, q, \in C$ if q is density reachable from p with respect to Eps and MinPts then $q \in C$. This property is called sometimes as *Maximallity*.

DBSCAN creates then a set of clusters $C_1, ..., C_k$ and all points in S are classified as:

1. *Core-point*: points with a dense neighborhood.
2. *Border-point*: points belonging to a cluster but without a dense neighborhood.
3. *Noise-point*: points do not belonging to any cluster.

The DBSCAN algorithm initiates cluster discovery by selecting an arbitrary, unvisited database point p and retrieving its density-reachable neighborhood (relative to ϵ and MinPts). The subsequent action depends on the nature of p:

1. If p is a core-point: A new cluster is formed containing p and all points density-reachable from p. This process is then iteratively expanded.
2. If p is not a core point: No points are density-reachable from p. DBSCAN assigns p to the noise-point category and proceeds to the next unvisited point.

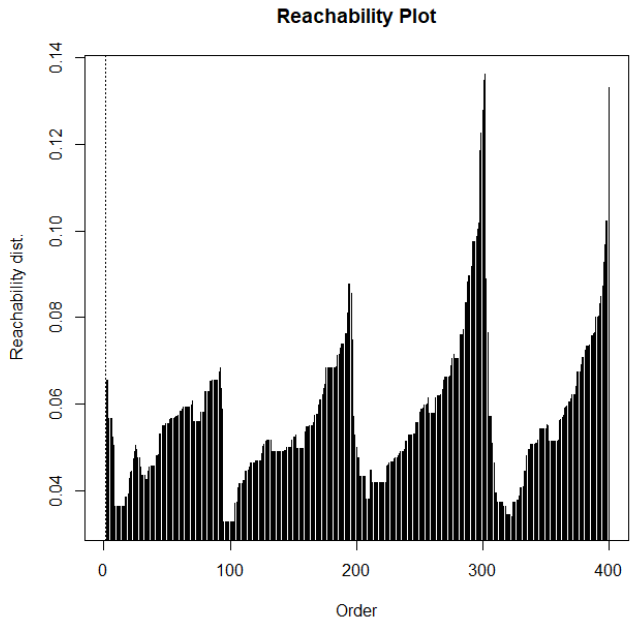

Figure 2.9: Example of OPTICS reachability plot

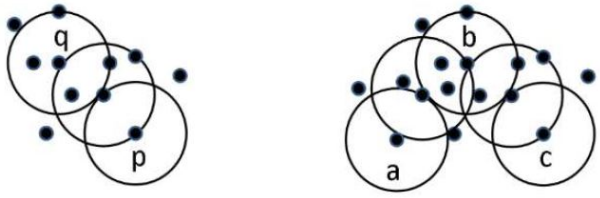

Figure 2.10: Left: density reachablability. Right: density connectivity Source [3].

It is important to note that if p is a border-point of a cluster C, it will eventually be reached during the expansion phase from a core point of C and correctly assigned to the cluster. The algorithm concludes once every point has been processed and assigned either to a cluster or to the noise-point set.

We will present a DBSCAN implementation in following pseudo-code:

Algorithm 1: The DBSCAN Algorithm

Input : Dataset D, ϵ (epsilon), MinPts (minimum points)

Output: Set of clusters C, with noise points unassigned

```
1  C ← 0 // Cluster counter
2  for each point P in D do
3      if P is unvisited then
4          mark P as visited;
5          NE ← RegionQuery(D, P, ε);
6          if |NE| < MinPts then
7              mark P as Noise;
8          end
9          else
10             C ← C + 1;
11             ExpandCluster(D, P, NE, C, ε, MinPts);
12         end
13     end
14 end
```

Algorithm 2: ExpandCluster and RegionQuery functions from DBSCAN Algorithm

```
Function ExpandCluster(D, P, NE, C, ε, MinPts)
    assign P to cluster C;
    for each point P′ in NE do
        if P′ is unvisited then
            mark P′ as visited;
            NE′ ← RegionQuery(D, P′, ε);
            if |NE′| ≥ MinPts then
                NE ← NE ∪ NE′;
            end
        end
        if P′ is not yet assigned to a cluster then
            assign P′ to cluster C;
        end
    end
end
Function RegionQuery(D, P, ε)
    return all points P′ ∈ D such that distance(P, P′) ≤ ε
end
```

As mentioned, the algorithm takes an unvisited point p and evaluates its eps-neighborhood througth the function *RegionQuery*, if it contains fewer than MinPts points p is labeled as noise. Otherwise p is labeled as core point algorithm expand the cluster througth the *ExpandCluster* function.

4.5 sLOS: Modifying the distance

To be included if have time. There is a direct application from [20] which works by modifing the distance along de line of sight.

4.6 HDBSCAN

This is other option to perform unsupervised density-based clustering.

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is an extension of DBSCAN that transforms the density-based approach into a hierarchical clustering algorithm. It requires only one mandatory parameter: $min\_cluster\_size$ (which is equivalent to $minPts$ or the minimum size of a dense region).

HDBSCAN introduces a concept of hierarchy of clusters, first it works by estimate the new concept of *mutual reachability distance* between two gien points, p and q:

$$mreach(p,\, q) = maxcore - dist(p),\, core - dist(q),\, dist(p, q)$$

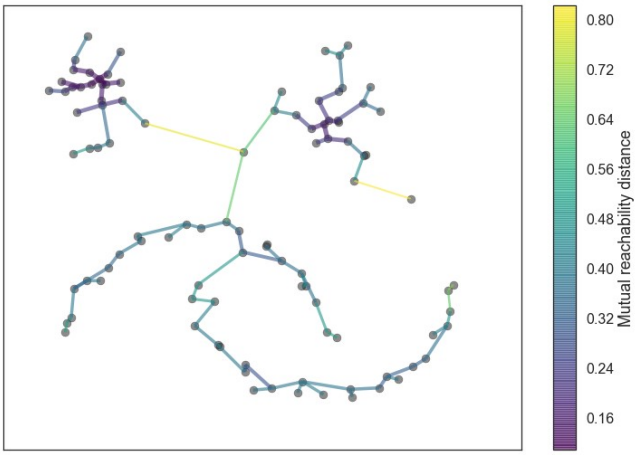

Figure 2.11: Minimium Spanning Tree (MSP
Source [17].

Remember from the 4.3 that within $core - dist$ concept depends directly on the minPts parameter. HDBSCAN uses this new concept of distance to guess dense areas in order to find clusters. First HDBSCAN calculates all mutual reachability distances, then points are placed as nodes in a graph called *Minimum Spanning Tree*, MST [12], and joined by edges representing weights of their mutual reachability distances. This MST is the minimum set of edges that connect points and minimizes the sum of edge weights (in fact the reachability distances). A MST is shown at 2.11.

The Minimum Spanning Tree (MST), constructed using the mutual reachability distance, forms the basis for the hierarchical cluster tree (dendrogram). This hierarchy is generated by iteratively grouping points based on increasing edge weights (mutual reachability distances), where each edge weight represents the density level at which two components become connected. The merged sets at each step constitute the cluster structure across all possible density thresholds (ϵ). The final hierarchy is then simplified through a condensation process based on the user-defined parameter, $minPts$ (or $min\_cluster\_size$). The algorithm traverses then the complete hierarchy. If a cluster splits into two new clusters, and one of the resulting clusters contains fewer than $minPts$ data points, that split is deemed insignificant.

Thus clusters with less than minPts are treated as single clusters, the process is one of re-labeling and pruning to simplify the tree based on persistence:, turning the hierarchy less complex and more interpretable.

The final clusters are extracted from this condensed dendrogram 2.12 by identifying the more stable clusters (most persistent) across varying density thresholds, the clusters are selected by longest lifetime λ, is the inverse of the distance (or density) at which a cluster merges or splits.

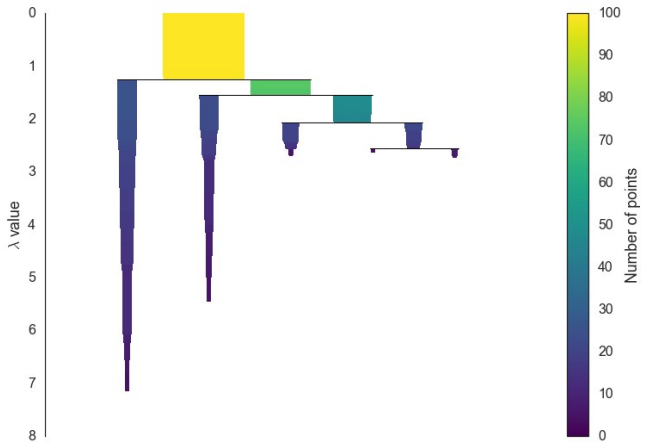

Figure 2.12: Condensed tree from HDBSCAN
Source [17].

Employs Stability for Final Selection: Instead of relying on a cutoff distance, HDBSCAN calculates the cluster persistence (often called "Excess of Mass")[12] for every potential cluster in the hierarchy. It then extracts the clusters that are the most stable (exist over the largest range of density thresholds), which allows it to naturally identify and separate clusters of different local densities."

The strength of HDBSCAN is its adaptability, this can result in a defect because the ability to identify sparse areas as distinct clusters might lead to the spurious detection of minor over-densities that might be categorized as noise in our galaxy catalogs. Despite this potential ambiguity, HDBSCAN is employed in this study because its fundamental mechanism is precisely aligned with the requirements of an unsupervised density-based approach.

4.7 Previous machine learning applications in galaxy clustering

This section briefly reviews several Machine Learning (ML) applications in cosmology, with particular emphasis on clustering techniques, but first we will introduce some historical research in globally galaxy custering.

In 18th centuty Charles Messier and William Herschel noted a concentration of nebulae (we know today that are large galaxies) in Virgo and Coma constellations [18].

In the 1920s Edwin Hubble proved that spiral and elliptical nebulae were extragalactic systems (galaxies) far outside the Milky Way [18].

In 1937 Friz Zwicky published the article *On the Masses of Nebulae and of Clusters of Nebulae*. This was a study about the velocity dispersion of galaxies in the Coma cluster. In this work he showed that the galaxies were moving too fast to be held together by the visible matter, leading to the first evidence and postulation of dark matter to explain the cluster's stability [16].

The first systematic, statistically complete catalog compiled by George Abell in 1958 became the foundation for modern cluster studies, allowing for a rigorous, statistical analysis of galaxy clustering across large volumes.

Severeal works in the literature use supervised methods, for example, Thomas et al. [9] generate predictive regression models based on the MACSIS simulation to predict cluster features from specific observables. On the other hand, Sadikov et al. [8] present an analysis of the X-ray properties of the galaxy cluster population in the $z = 0$ snapshot of the IllustrisTNG simulations, utilizing machine learning to perform clustering and regression tasks.

In contrast, other studies applying Machine Learning (ML) to the galactic Universe directly address the intrinsic properties of galaxies rather than focus on the clustering problem. For example, Dvorkin et al. [13] note that "it has been shown that unknown relations between galaxy properties and parameters describing the composition of the Universe can be easily identified by employing machine learning techniques on top of state-of-the-art hydrodynamic simulations" [14].

The most significant application of density-based algorithms to galaxy distribution is a recent article (dated 2025) by Hai-Xia-Ma et al.[20]. The authors successfully applied density-based algorithms, including a modified version called sOPTICS, to several galaxy catalogs, achieving a notable success in cluster detection. They created the modified version of OPTICS called sOPTICS and used it to mitigate the redshift space distortion along line-of-sight caused by galaxies' peculiar velocities.

As the reader can observe, a gap currently persists in the astronomical literature regarding the widespread application of unsupervised density-based algorithms for the systematic detection of galaxy groups and clusters. This limited exploration of density-based techniques, particularly in validating existing catalogs, underscores the novelty of this work. Furthermore, the large-scale distribution of matter across the Universe presents several fundamental problems that lie at the frontier of modern physics, such as understanding the nature of dark matter and dark energy. By providing robust, objective characterizations of cosmic structures across all scales, this study contributes essential input for constraining cosmological models and addressing these profound mysteries.

5 Glossary

Redshift: Increase in the wavelength of radiation - tipically ligth-. The redshift takes place for several reasons, one of then is when the source of ligth is further away, for example in an expanding Universe, then they speak about cosmic-redshift.

Bibliography

[1] Longair S. Malcom. (1996). *Our Evolving Universe.* Cambridge University press, United Kindom, UK.

[2] Einasto J. (2014). *Dark Matter And Cosmic Web Story.* New Jersey: World Scientific Publishing Co. Pte. Ltd.

[3] Rhys M. (2020). *Machine Learning with R.* Manning publications, United Kindom, UK.

[4] Cepa J. (2023). *Cosmología Física.* Edidiones Akal, Barcelona, ES.

[5] AOO. *The Australian Astronomical Observatory site.* Available at https://aat.anu.edu.au.

[6] Group catalog. *Group Catalogues for 2dFGRS and SDSS.* Available at https://gax.sjtu.edu.cn/data/Group.html.

[7] Colless M. et al. (2001). First results from the 2df galaxy redshift survey. *arXiv e-prints.*

[8] Sadikov M. et al. (2025). Galaxy cluster characterization with machine learning techniques. *arXiv e-prints.*

[9] Thomas J. et al. (2025). An application of machine learning techniques to galaxy cluster mass estimation using the macsis simulations. *arXiv e-prints.*

[10] Blanton M. R. et al.(2005). New york university value-added galaxy catalog: A galaxy catalog based on new public surveys. *New York, NY. Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place. p 2562, available at https://doi.org/10.1086/429803.*

[11] Yang X. et al.(2007). Galaxy groups in the sdss dr4. i. the catalog and basic properties. *The Astrophysical Journal, Volume 671, Issue 1, pp. 153-170.*

[12] Campello R. et al.(2013). Density-based clustering based on hierarchical density estimates. *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) https://doi.org/10.48550/arXiv.2409.13010.*

[13] Dvorkin C. et al.(2022). Machine learning and cosmology. *arXiv e-prints.*

[14] Villaescusa-Navarro J. et al.(2022). Cosmology with one galaxy? *arXiv e-prints.*

[15] Anatole S. et al.(2024). The causal effect of cosmic filaments on dark matter halos. *Lund Observatory, Division of Astrophysics, Department of Physic. Lund, Sweden, p [1-5] available at https://doi.org/10.48550/arXiv.2409.13010.*

[16] Zwcky F.(1937). On the masses of nebulae and of clusters of nebulae. *Astrophysical Journal, vol. 86, p.217.*

[17] The hdbscan Clustering Library. *The hdbscan Clustering Library Site.* Available at https://hdbscan.readthedocs.io/.

[18] Ostriker J. and Mitton S. (2014). *El corazón de las tinieblas.* Pasado y presente.

[19] SDSS. *The Sloan Digital Sky Survey site.* Available at https://www.sdss3.org/.

[20] Yongda Zhu.(2025) Tsutomu T. Takeuchi1, Suchetha Cooray. soptics: A modified density-based algorithm for identifying galaxy groups/clusters and brightest cluster galaxies. *ArchivX.*