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, December 21, 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.

x

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 [16] 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[17].

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 [11]:

  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.

Task Name	September	October	November	December	January
Definition and planning	Definitions				
Objectives		Objectives			
State of art		State of art			
Design		Design			
Implementation			Implementation		
Assessment and improvements			Evaluation of results		
Exploring additional objectives			Additional objectives		
Writing				Thesis writing	
Final project defense					
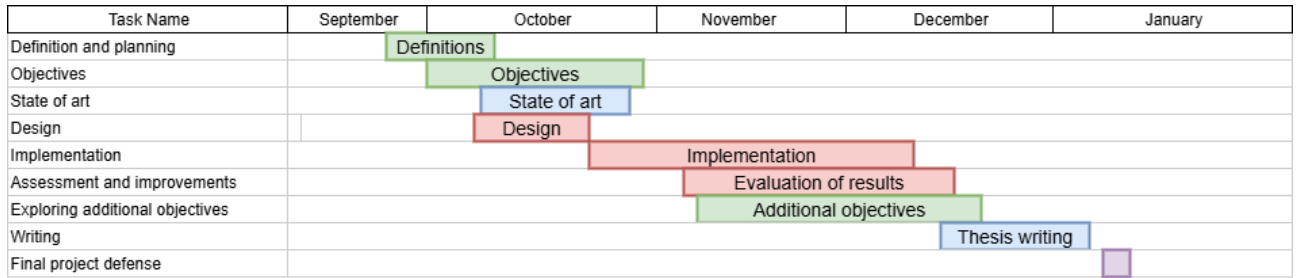
Figure 1.1: Stages of the project.

- 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 2.15 and clusters 2.16 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 [17] 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 [17]. A paradigmatic example is the Virgo cluster with an estimate mass of $\sim 1.2 10^{15} M_\odot$ and having M87 as most massive and dynamically dominant central galaxy.
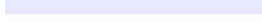
Figure 2.1: Future group image
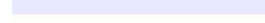
Figure 2.2: Future cluster image

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
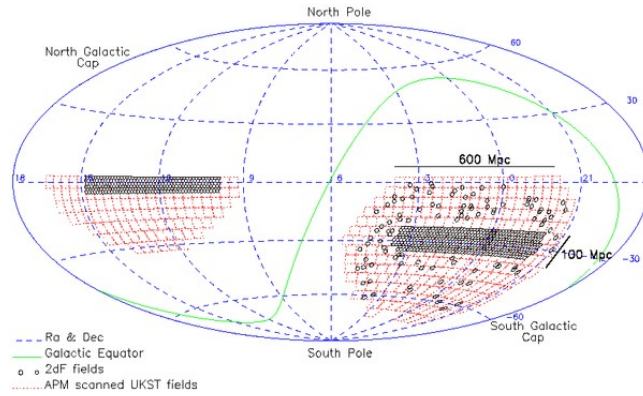
Figure 2.3: 2dfGRS sky coverage
Source [8].

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) [23] 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 [12].
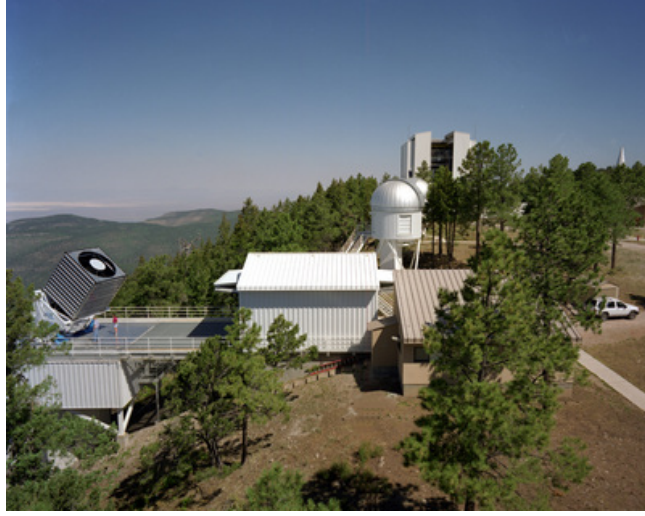
Figure 2.5: SDSS Data release 7 sky coverage
Source [5].
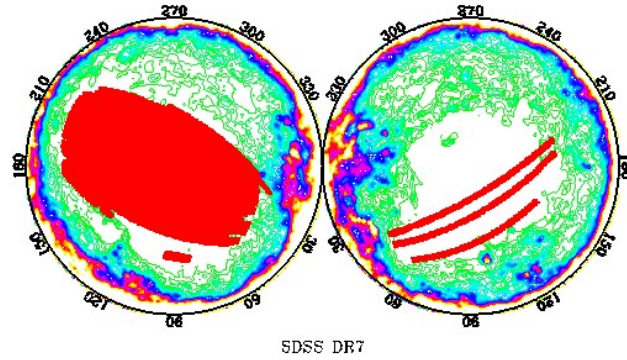
Figure 2.6: SDSS Data release 7 sky coverage
Source [12].

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 [12]. 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 [12]) 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 to 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]:

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

.

3.1   Real space galaxy catalog

Redshift surveys suffer distortions that ate primarily categorized into two distinct sets:

1. FOG (Finger Of God): reduction in the correlation produced on small scales by the virialized motion of galaxies within the dark matter halos.

2. Kaiser effect: boost of the correlation in large scales induced by the infall motion of galaxies towards overdensity regions.

Both categories of distortions manifest within the correlation function of the galaxy distribution and the global matter distribution.

Several accurate methodologies are available to mitigate the effects of both issues, at least partially. This discussion will focus on the approach described by Shi et al. in [24], where the authors propose a method to correct these distortions as follows:

The observed redshift (z_{obs}) of a galaxy is a combination of the cosmological redshift (z_{cos}) due to the Hubble flow and the Doppler shift caused by the galaxy's peculiar velocity (v_{pec}).4 The total relationship is expressed as:

$$z_{obs} = z_{cos} + z_{pec} = z_{cos} + \frac{v_{pec}}{c}(1 + z_{cos}). \tag{2.4}$$

.

And then for v_{pec} we have:

$$v_{pec} = v_{cen} + v_\sigma. \tag{2.5}$$

.

Where v_{cen} is the line-of-sight component of the coherent velocity of the halo center, representing the infall motion that causes the large-scale Kaiser effect. v_σ is the line-of-sight component of the internal velocity dispersion of the galaxy with respect to its halo center, representing the random motion that causes the small-scale Fingers-of-God (FoG) effect.

The resulting effective redshifts corresponding to these two distinct effects are formally defined as follows:

$$z_{FOG} = z_{cos} + \frac{v_\sigma}{cen}(1 + z_{cos}). \qquad (2.6)$$

.

Roughly speaking v_{cen} is the manifestation of z_{Kaiser}, whereas FOG effect is depicted by v_σ.

Following the methodology proposed in [24], the authors define several coordinate spaces by utilizing Equations 2.4, 2.5, and 2.6 to recalculate the positions (Equation 2.2) in each space.

For the present study, we will utilize the coordinate system referred to as the Real space which when reconstructted based on correcting redshift space distortions is called **Re-Real space** [24]. This space is particularly relevant as it incorporates the effects of both the Kaiser and FOG distortions.

Employing this approach yields a modified catalog where the new redshift values effectively transform the coordinate space, as illustrated in Figure 2.7 . We will use this Re-Real space dataset to thoroughly investigate how this transformation influences the resulting clustering of galaxies.
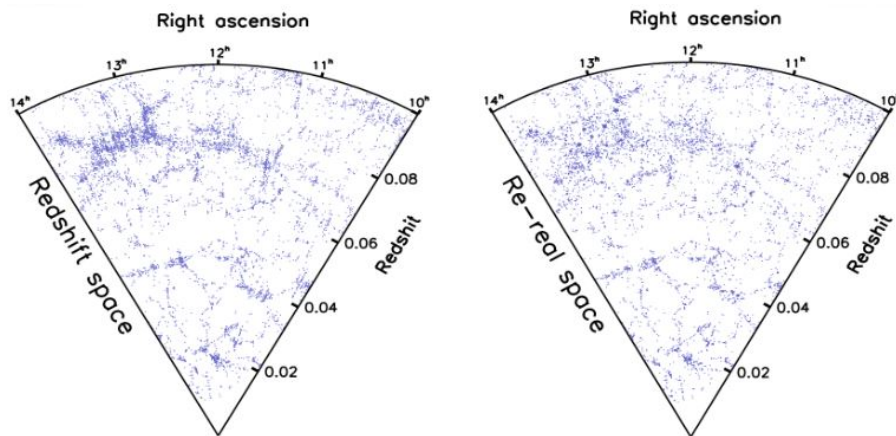

Figure 2.7: SDSS-DR7: In the left-hand side a sample of SDSS in the redshift space. The right-hand side shows SDSS in Re-Real space
Source [24].

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:

$$reachability - distance(p,q) = max\{core - distance(p), dist(p,q)\}$$
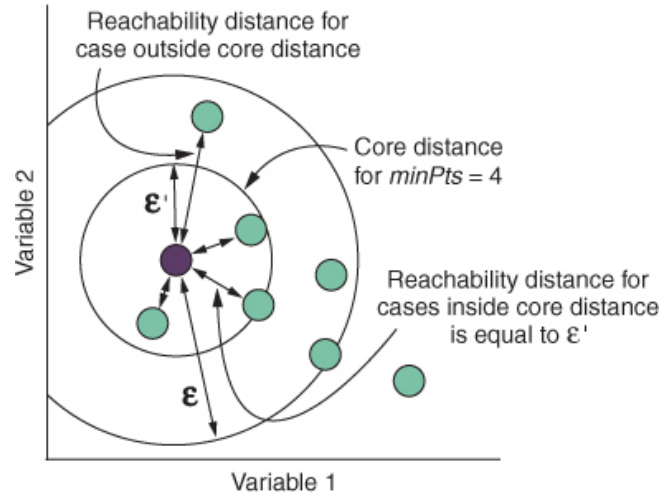
Figure 2.8: Core and reachability distances
Source [3].

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 2.8.

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.
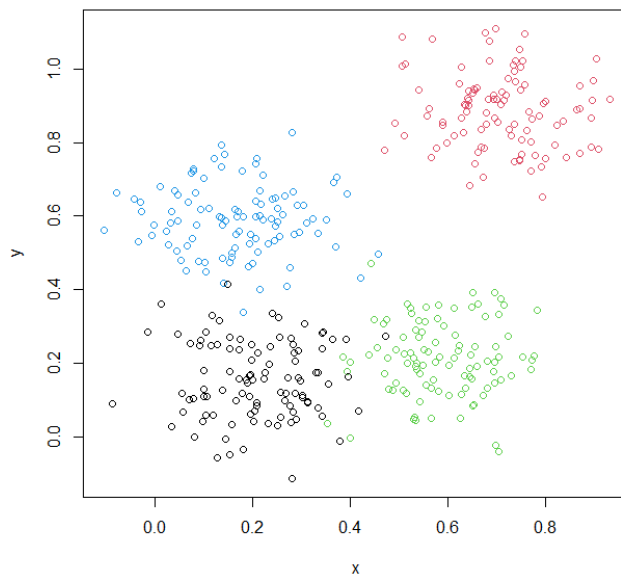
Figure 2.9: An example of data set in plane $\mathbb{R}^2$.

For example, figure 2.9 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 2.10.

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
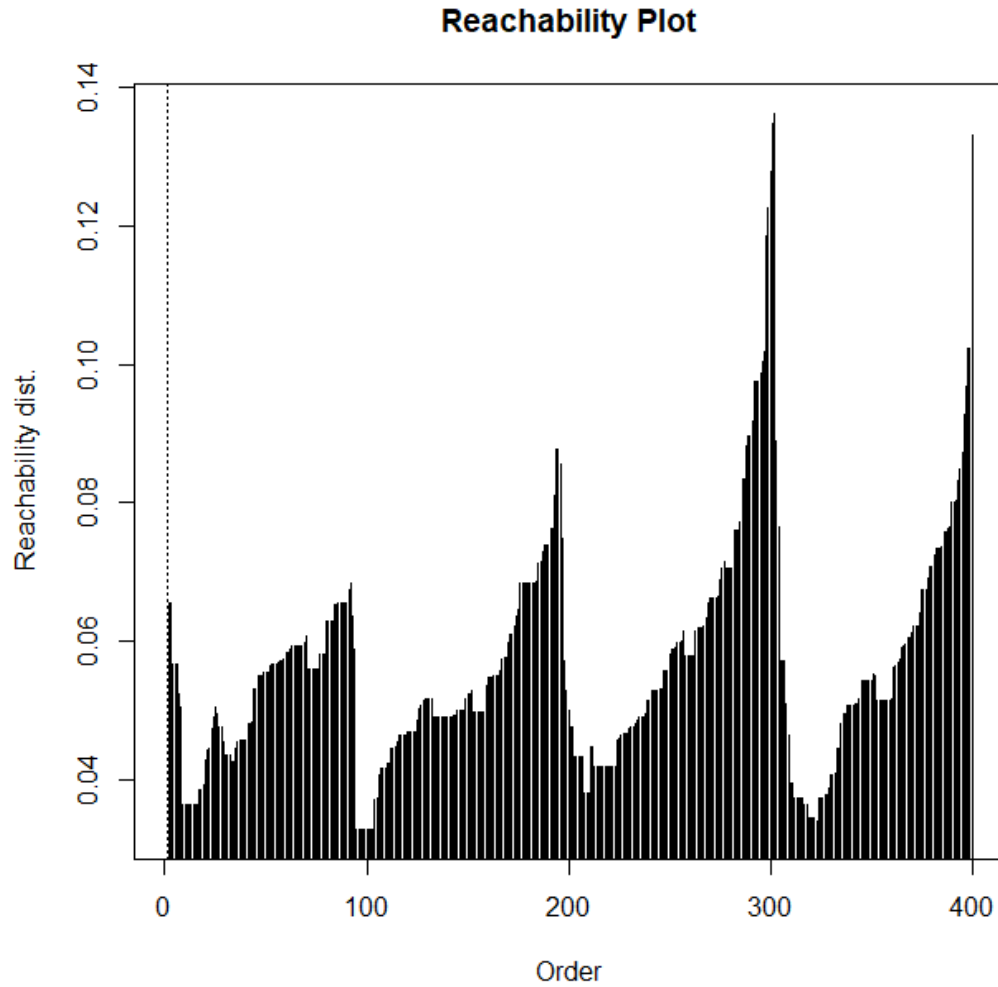
Reachability Plot

Figure 2.10: Example of OPTICS reachability plot

- 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 $1 \le i \le n$.

- 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.

The figure 2.11 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*.
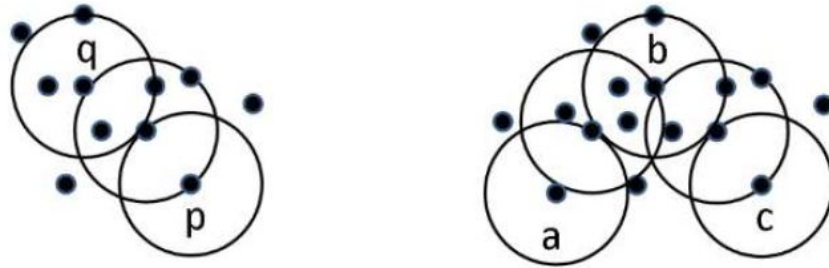
Figure 2.11: Left: density reachablability. Right: density connectivity
Source [3].

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.

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 $C \leftarrow 0$ // Cluster counter

2   **for** *each point P in D* **do**

3   **if** *P is unvisited* **then**

4   mark P as visited;

5 $NE \leftarrow \texttt{RegionQuery}(D, P, \epsilon)$;

6   **if** $|NE| < MinPts$ **then**

7   mark P as **Noise**;

8   **end**

9   **else**

10 $C \leftarrow C + 1$;

11 $\texttt{ExpandCluster}(D, P, NE, C, \epsilon, \text{MinPts})$;

12   **end**

13   **end**

14 **end**

Algorithm 2: ExpandCluster and RegionQuery functions from DBSCAN Algorithm

---

1 Function `ExpandCluster`$(D, P, NE, C, \epsilon, MinPts)$

2 $\quad$ assign P to cluster C;

3 $\quad$ **for** *each point P' in NE* **do**

4 $\quad\quad$ **if** *P' is unvisited* **then**

5 $\quad\quad\quad$ mark P' as visited;

6 $\quad\quad\quad$ $NE' \leftarrow$ `RegionQuery`(D, P', ϵ);

7 $\quad\quad\quad$ **if** $|NE'| \geq MinPts$ **then**

8 $\quad\quad\quad\quad$ $NE \leftarrow NE \cup NE'$;

9 $\quad\quad\quad$ **end**

10 $\quad\quad$ **end**

11 $\quad\quad$ **if** *P' is not yet assigned to a cluster* **then**

12 $\quad\quad\quad$ assign P' to cluster C;

13 $\quad\quad$ **end**

14 $\quad$ **end**

15 end

16 Function `RegionQuery`(D, P, ϵ)

17 $\quad$ **return** *all points $P' \in D$ such that $distance(P, P') \leq \epsilon$*

18 end

As mentioned, the algorithm takes an unvisited point p and evaluates its eps-neighborhood throught 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 throught the *ExpandCluster* function.

4.5 sLOS: sOPTICS Modifying the distance

There is a direct application from [17] which works by modifing the distance along de line of sight as shown at 2.12. This distorsion is the same referred as *FOG* in the section 3.1

This algorithm works with a normal OPTICS but the distance has been modified by a new one wich work as follows: Given two point u and v the usual Euclidean distance is calculated as

$$D^2(u, v) = \sum_{i=1}^{3}(u_i - v_i)^2. \tag{2.7}$$

. Instead, a new version of distance concept is created by calculate the so called Ellongated

Euclidean Distance as

$$D^2_{Elongated}(u, v, sLos) = d^2_{traverse}(u, v) + d^2_{sLOS}(u, v, sLos) \tag{2.8}$$

. Where

$$d_{sLOS}(u, v, sLos) = sLOS * \frac{\sum_{i=1}^3 (u_i - v_i)u_i}{\sqrt{\sum_{i=1}^3 u_i^2}} \tag{2.9}$$

.

Since the sLOS factor must be calculated, the final consideration addresses the distance metric. To ensure a symmetric distance concept—and thereby guarantee the stability of the core-point definitions—the distance chosen is defined as:

$$d^{sym}_{sLOS}(u, v, sLos) = sLOS * \frac{d_{sLOS}(u, v, sLos) + d_{sLOS}(v, u, sLos)}{2} \tag{2.10}$$

.

Figure 2.12 illustrates the elongated Euclidean distance's effect on the OPTICS clustering results. Clusters are computed in an elongated way along the line of sight in order to cure the redshift space distortion.
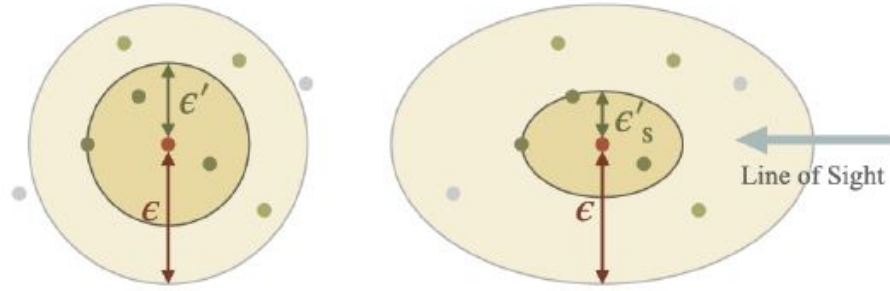

Figure 2.12: Illustration of how sLos algorithm works elongating the line of sigth
Source [17].

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:

$$mreach(p, q) = maxcore - dist(p), core - dist(q), dist(p, q)$$
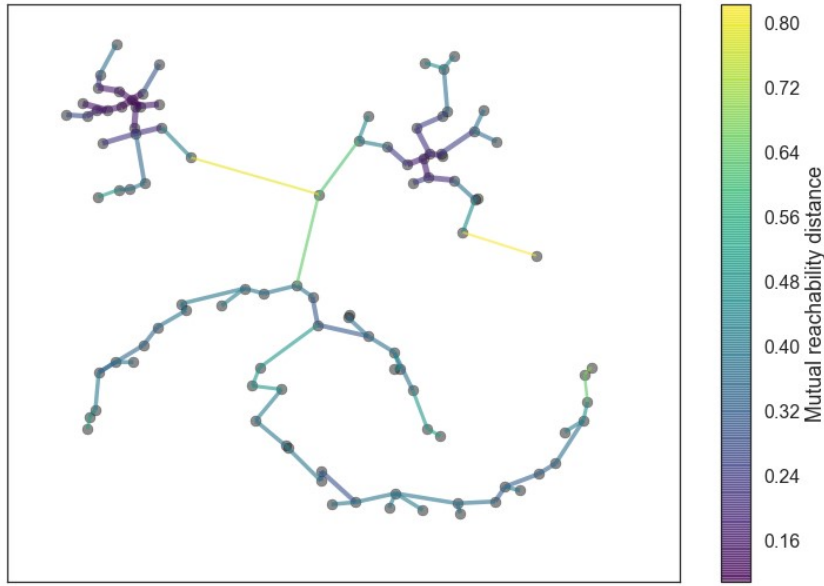
Figure 2.13: Minimium Spanning Tree (MSP
Source [19].

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* [13], 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 2.17.

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 2.14 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.
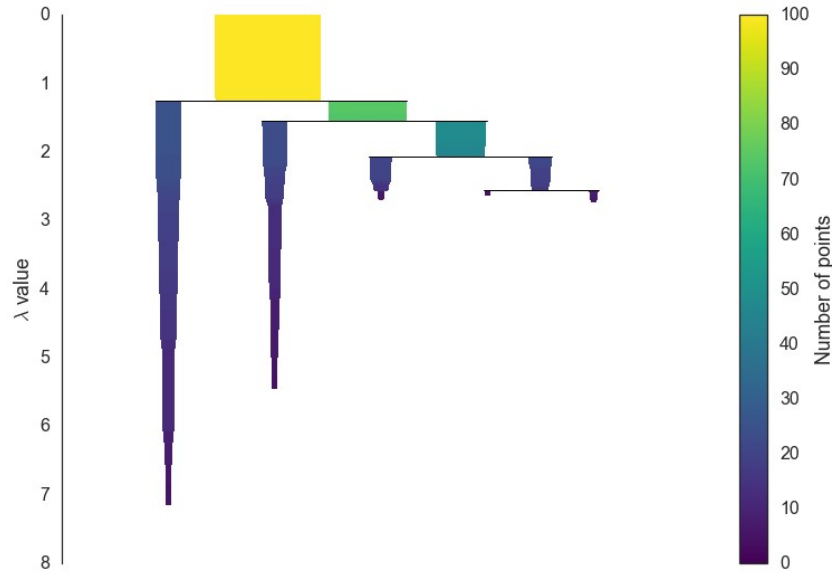
Figure 2.14: Condensed tree from HDBSCAN
Source [19].

Employs Stability for Final Selection: Instead of relying on a cutoff distance, HDBSCAN calculates the cluster persistence (often called "Excess of Mass")[13] 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.
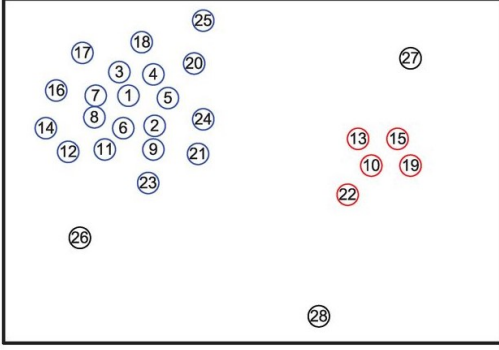
Figure 2.15: Different colors correspond to different clusters.
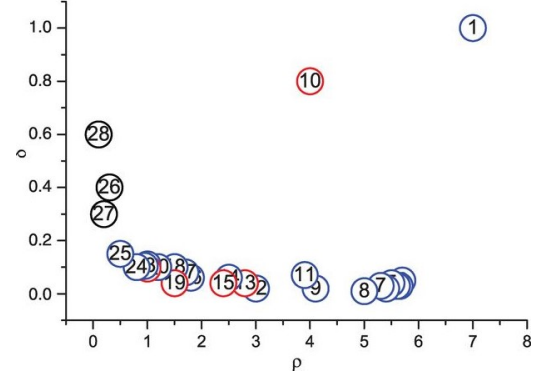
Figure 2.16: (B) Decision graph for data Source [22].

4.7 Density Peaks Clustering (DPC)

Density Peaks Clustering (DPC) is a newer (proposed in 2014) density-based algorithm, By analyzing the Decision Diagram (ρ vs δ), we expect to be able to isolate cluster cores from the background field population.

The Density Peak Clustering (DPC) algorithm operates under the fundamental assumption that cluster centers correspond to points where the local density attains its maximum. A critical distinguishing characteristic is that each center is separated by a considerable distance from all other centers, while being immediately surrounded by points of comparatively lower local density. Therefore, two quantities are calculated for each point, p_i, in the dataset:

1. Local density (ρ_i). Which is calculated as $\rho_i = \sum_j \chi(d_{ij} - d_c)$, where d_c is a cutoff distances and $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$ if $x \geq 0$. In other words, ρ_i is the number of points that are closer than d_c to the point i.

2. Distance from next point δ_i with higher density. This parameter is measured by computing: $\delta_i = min_{j:\rho_j > \rho_i}(d_{ij})$. and for the point with highest density $\delta_i = max_j(d_{ij})$. Therefore δ_i is much larger than other typical values for those points where density reaches its local or global maximum.

Following the computation of both parameters, a diagram as presented in figure 2.17 can be utilized to establish appropriate thresholds for ρ and δ in order to select the points which represent centers for each clustersδ. This process allows for the identification of points representing the center of each cluster. DPC excels at cluster center detection because these selected points inherently correspond to locations where the local density reaches its maximum.

Then we can select $\rho = \rho_0$ and $\delta =_0$ to obtains centers for clusters:
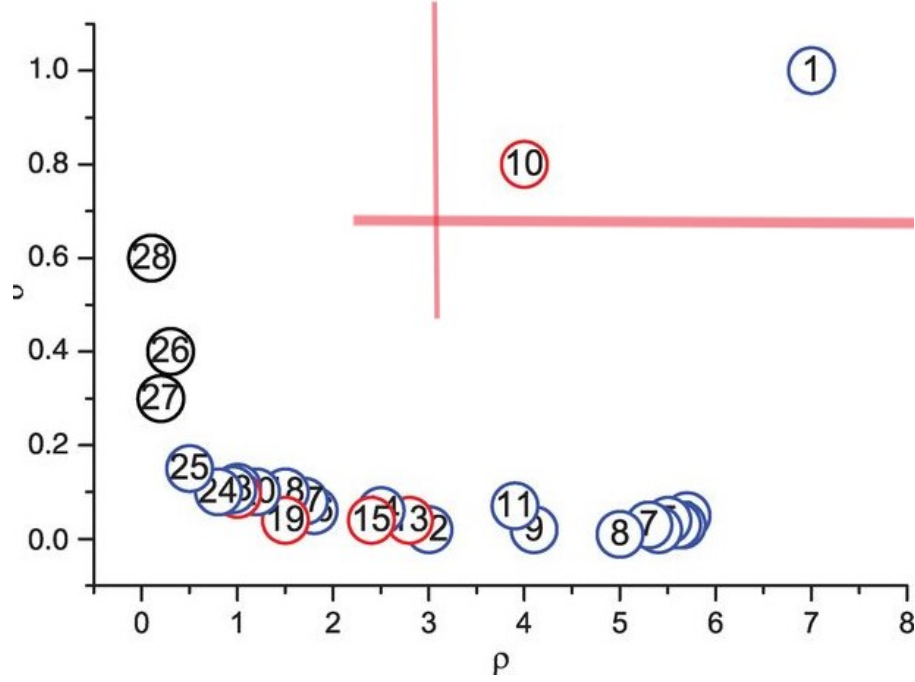
Figure 2.17: $\rho = \rho_0$ and $\delta = \delta_0$ inside decision graph: Points with a high local density (located on the right side of the x-axis). Points that are far from any other point with a higher density (located at the top of the y-axis).

Source [22].

5   The two-point correlation function: a statistical point of view in density analysis

An alternative approach to investigating the cosmic galaxy density field is through the two-point correlation function (2PCF). This statistic provides a robust measure for quantifying the spatial clustering of a galaxy distribution by determining the excess probability of finding a pair of objects at a given separation scale, r, relative to a random (Poisson) distribution.

The two-point correlation function $\xi(r)$ and the power spectrum, $P(k)$, form a Fourier transform pair, representing the clustering signal in real and spectral space, respectively. While $\xi(r)$ provides an intuitive measure of spatial separations, the power spectrum is the standard framework for characterizing the primordial density fluctuations observed in the Cosmic Microwave Background (CMB).x

According to [21], if one pick a galaxy at random, the probability (dP) of finding another galaxy at a distance r within a small volume dV is given by:

$$dP = \bar{n}^2[1 + \xi(r)]dV_1 dV_2$$

Where $\bar{n}$ is the mean number density of galaxies in the survey and $\xi(r)$ is the correlation function.

The value of $\xi(r)$ tells us how the galaxies "feel" each other at different scales [21]:

1. $\xi(r) > 0$ (Clustering): You are more likely to find a pair of galaxies at that distance than you would by pure chance. On small scales, gravity pulls galaxies together, making this value high.

2. $\xi(r) = 0$ (Randomness): The distribution is perfectly random (like static on a TV).

3. $\xi(r) < 0$ (Anti-clustering): Galaxies are "repelling" each other or are more spread out than a random distribution.

5.1 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 [20].

In the 1920s Edwin Hubble proved that spiral and elliptical nebulae were extragalactic systems (galaxies) far outside the Milky Way [20].

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 [18].

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. [10] generate predictive regression models based on the MACSIS simulation to predict cluster features from specific observables. On the other hand, Sadikov et al. [9] 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. [14] 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" [15].

The most significant application of density-based algorithms to galaxy distribution is a recent article (dated 2025) by Hai-Xia-Ma et al.[17]. 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.

Chapter 3

Implementation

1 ETL processing of datasets

This section describes the Data Engineering Pipeline that converts the astronomical raw data into the machine learning-ready format you used for your 2dFGRS analysis.

Our Python framework acts as a bridge between the raw observational catalog and the unsupervised learning models. By producing a sanitized CSV with pre-calculated, scaled Cartesian coordinates to operate with maximum efficiency and physical accuracy which data definition is shown in table 3.1.

The final objective of this pipeline is to associate individual galaxies with their respective Dark Matter Halos or larger structures. To achieve this, we follow three basic steps to merge the algorithmic output with the physical catalog:

1. Format the galaxy catalog to a CSV file.

2. Format the group catalog to a CSV file.

3. Merge galaxy and group catalog and transform coordinates and distances.

The synthesis results in a unified dataset formatted for computational efficiency. The header of this processed file, which contains the spatial and environmental metadata for each galaxy, is displayed in Figure 3.1.

Due to the varying structures of the survey catalogs, the data acquisition and preparation phase is divided into distinct modules to ensure inter-survey compatibility.

1.1 2dF Galaxy Redshift Survey (2dfGRS)

1. *2dfGRS.dat*: Which comprises 245,591 individual galaxy entries. To ensure high-fidelity measurements and minimize redshift uncertainty, we applied a quality constraint of $Q \geq 3$,

Field-Name	FDescription	Field-Type
GAL_ID	ID of galaxy en each catalog	numerical
ra	Right ascension coordinate	decimal
dec	Declination coordinate	decimal
x	X cartesian coordinate	decimal
y	Y cartesian coordinate	decimal
z	Z cartesian coordinate	decimal
redshift	cell8	decimal
dist	Raw distance value	decimal
GROUP_ID	id-group galaxy belongs to	numerical

Table 3.1: Datasheet metadata.

excluding objects with poorly determined spectral features or low signal-to-noise ratios

2. *group_members*: a supplementary group-membership file consisting of 104,913 galaxies.

1.2 Sloan Digital Sky Survey (SDSS)

Among the diverse datasets provided by the SDSS archive, the imodelC_1 file was identified as the most suitable for this analysis. It contains the required astrometric and photometric parameters to accurately cross-match with our random catalogs, allowing for a robust calculation of the large-scale structure signal.

1. *SDSS7*: Galaxy catalog of the survey.

2. *imodelC_1* Comprises 245,591 entries for each galaxy. (again we applied a quality constraint of $Q \geq 3$.)

Our pipeline performs a multi-source integration of the raw data files, executing the necessary joins and quality filters to produce a unified CSV file optimized for clustering analysis. A representative sample of this finalized data product is provided in Figure 3.1, demonstrating the successful synthesis of spatial and environmental metadata.

1.3 Real Space Galaxy Catalogue

Finally, we incorporated the SDSS 'Real Space Galaxy Catalogue', which provides galaxy coordinates reconstructed to account for redshift distortions as we explained in sections 4.5 and 3.1 . This enables a robust comparison between observed clustering and theoretical predictions by isolating the isotropic real-space correlation function.

	GAL_ID	ra	dec	x	y	z	redshift	dist	GROUP_ID
0	2	3.627292	-32.966861	0.099988	0.006339	-0.064981	0.1229	0.119417	12097
1	3	3.586292	-32.388000	0.085393	0.005352	-0.054273	0.1038	0.101322	542
2	5	3.608000	-32.711528	0.084924	0.005355	-0.054652	0.1036	0.101132	4846
3	6	3.612542	-32.862444	0.106340	0.006714	-0.068832	0.1308	0.126850	4847
4	7	3.613417	-33.013278	0.089646	0.005661	-0.058362	0.1099	0.107120	1462

Figure 3.1:   Final format of dataset

2 The two-point correlation function (2pcf)

Drawing on the comparative analysis provided by [7], we implement four distinct estimators to quantify the clustering signal. This selection allows for a rigorous cross-examination of the statistical bias and variance inherent in each method when applied to large-scale galaxy catalogs.

Natural

$$\widehat{\xi}_N = \frac{DD}{RR} - 1 \tag{3.1}$$

.

Davids and Peebels:

$$\widehat{\xi}_{DP} = \frac{DD}{DR} - 1 \tag{3.2}$$

.

Hamilton:

$$\widehat{\xi}_{Ha} = \frac{DD\,RR}{DR^2} \tag{3.3}$$

.

Landy and Szalay:

$$\widehat{\xi}_{LS} = \frac{DD - 2DR + RR}{RR} \tag{3.4}$$

.

Upon applying these estimators to the 2dFGRS dataset, as we will se in 5 we successfully recovered a distinct clustering feature at approximately 100 h^{-1}Mpc (see figures **??** and **??**. This signal is statistically consistent with the predicted scale of Baryon Acoustic Oscillations (BAOs), representing a detection of the primordial acoustic horizon in the late-time galaxy distribution.
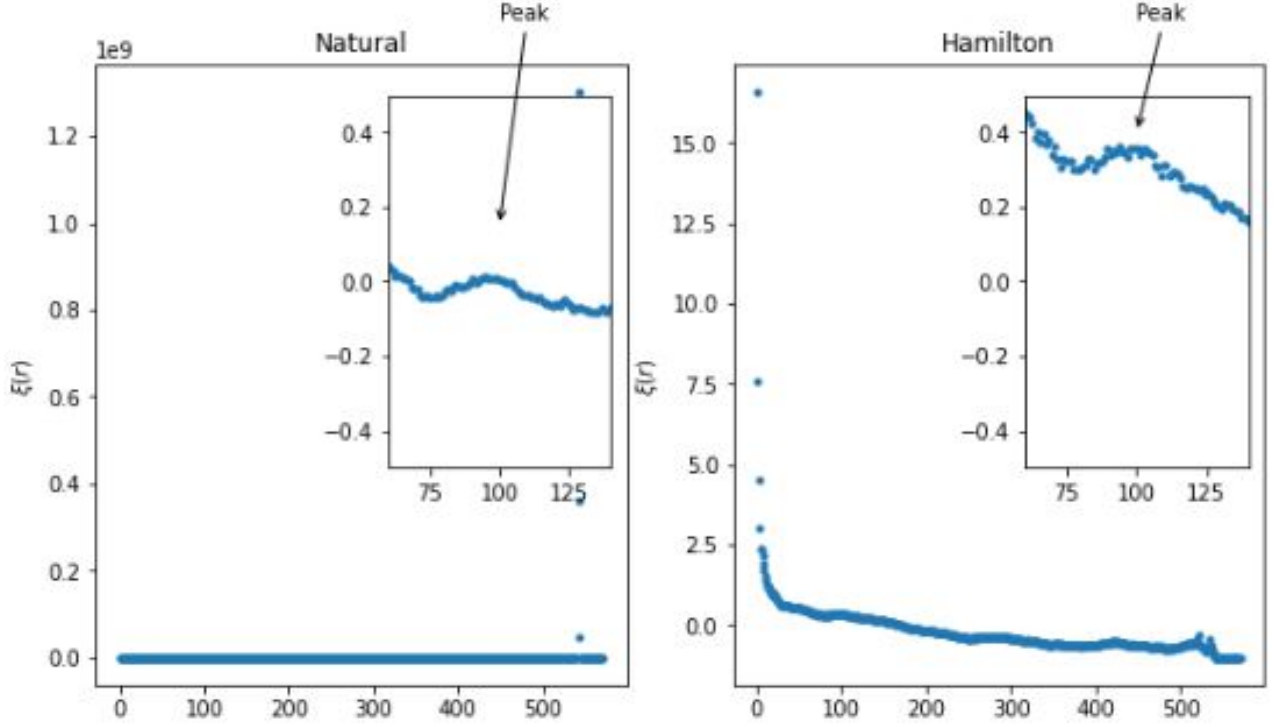
Figure 3.2:   Natural and Hamilton estimators measured on the 2dFGRS sample.

3   Application of density-based algorithms to datasets

We performed a comparative evaluation of several density-based clustering frameworks to assess their capability in recovering the physical halo distribution. The algorithms were selected based on their distinct approaches to density reachability, hierarchical extraction, and noise handling

We tested several algorithms in order to obtain a model of density clustering both for non-scaled and scaled data to ensure that the density metrics are isotropic and not biased by the differing scales of the coordinate axes. to prevent .

- OPTICS: Utilized to generate a reachability plot, providing a visualization of the hierarchical density structure and identifying the spatial ordering of galaxies.

- OPTICSXi: An extension of OPTICS used to extract clusters in a hierarchical mode by identifying steep density gradients (the ξ parameter), allowing for the detection of clusters with varying densities.

- DBSCAN: Implemented as a baseline density-based method to identify clusters as density-connected components based on a fixed global proximity threshold (ϵ).

- HDBSCAN: A robust hierarchical implementation that constructs a spanning tree to find stable clusters across all possible density scales, making it highly effective for multi-scale
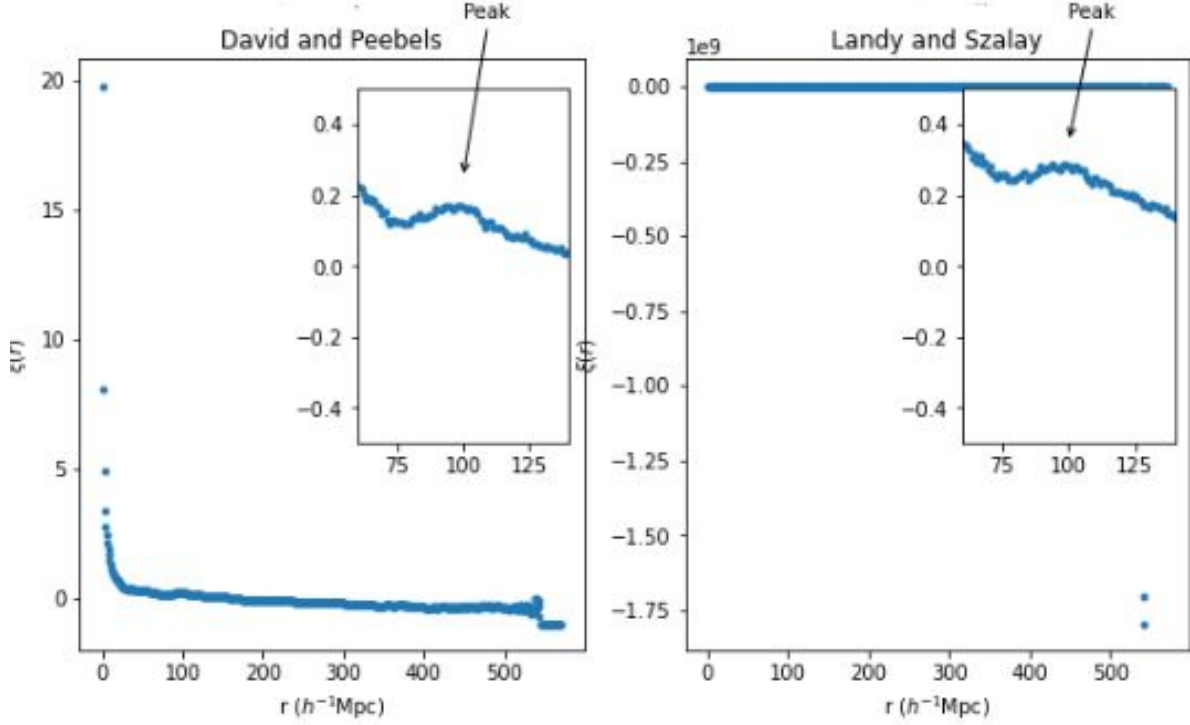
Figure 3.3: David and Peebels and Landy and Szalay estimators for 2dFGRS sample.

cosmological distributions.

- DPC (Density Peaks Clustering): Employed to identify clusters based on the detection of local density maxima and their relative distance from other high-density peaks, which is physically analogous to identifying halo centers.

- sOPTICS and sDBSCAN: These variants account for line-of-sight positional uncertainties due to redshift space distortions with the modified distances as explained at 4.5

It is important to emphasize that this study departs from traditional unsupervised clustering objectives, such as minimizing intra-cluster variance via the Elbow Method. Instead, we treat the Halo-based group distribution as a physical ground truth. Consequently, many standard clustering algorithms —and their default hyperparameter configurations— may fail to yield results consistent with our model of virialized galaxy groups, as they are not intrinsically designed to account for the specific density profiles of dark matter halos.

The performance of each algorithm is evaluated based on its Recovery Rate of known halo members. We prioritize models that maximize the completeness and purity of identified groups

relative to the 2dFGRS/SDSS group catalogs, rather than those that simply minimize global silhouette scores.

Guided by the validation protocols established in [17] we employ the following metrics to assess the topological and member-wise similarity between the density-based models (C) and the halo-based ground truth (H):

We define:

- *total_in_cluster*: number of elements in output-cluster.

- *total_in_cluster_group*: number of elements from orignal group present in an output-cluster.

- *total_in_cluster$_g$roup*: number of elements from orignal group present in an output-cluster.

$$Purity = \mathcal{P} = \frac{total\_in\_cluster\_group}{total\_in\_cluster} \tag{3.5}$$

.

$$Completeness = \mathcal{C} = \frac{total\_in\_cluster\_group}{total\_in\_group} \tag{3.6}$$

.

Purity is also called **Precision** and completeness **Sensitivity** and also **Recall**.

Following the categorical framework of [17] we define the thresholds:

- Purity Threshold ($\mathcal{P} \geq 2/3$): An output-cluster is defined as Pure if at least 66.7

- Completeness Threshold ($\mathcal{C} \geq 1/2$): An output-cluster is defined as Complete if it successfully captures at least 50

With this concepts we evaluate the following ratios:

$$F_p = \frac{N_{pure}}{N_{clusters}} \tag{3.7}$$

.

$$F_c = \frac{N_{complete}}{N_{clusters}} \tag{3.8}$$

.

Algorithm	Hyperparameter	Data Sample	Outcomes Sample	Conclusion
DBSCAN	$\epsilon = 6 \times 10^{-4}$ $minPts = 5$	Non-scaled	$P = 0.65$ $C = 0.87$ $R = 0.42$ $U = 21$	Good in cluster detection
HDBSCAN	-	-	-	Not good in cluster detection.
DPC	$\rho = 8.4 \times 10^{-4}$ $\delta = 0.9985$	Non-scaled	-	Good in cluster center detection
sOPTICS	$\epsilon = 11 \times 10^{-5}$ $minPts = 5$ Best results	Non-scaled	$P = 0.84$ $C = 0.84$ $R = 0.86$ $U = 12$	
DBSCAN	$\epsilon = 6 \times 10^{-4}$ $minPts = 5$	Scaled	$P = 0.72$ $C = 0.81$ $R = 0.41$ $U = 22$	Good in cluster detection
OPTICS	-	-	-	Good in cluster detection.

Table 3.2: Results on 2dFGRS sample.

Fr only for complete and pure output-clusters:

$$F_r = \frac{N_{complete} + N_{pure}}{N_{original\_groups}} \tag{3.9}$$

.

3.1 Application to 2dFGRS catalog

By applying the success-matching protocol derived from [17], we evaluated the performance of each density-based configuration. The table 3.4 summarizes the ability of each algorithm to recover the underlying group or halo distribution within the 2dFGRS survey volume.

3.2 Application to SDSS catalog

We applied the dentisy-based algorithms to the SDSS catalog shown in section 2.2.

The outcomes are showed in table [**?**]. Same results of application in can be shown as well, so sOPTICS is the best fitting algorithm which meets the results showed in [17].

Algorithm	Hyperparameter	Data Sample	Outcomes Sample	Conclusion
DBSCAN	$\epsilon = 6 \times 10^{-4}$ $minPts = 5$	Non-scaled	$P = 0.65$ $C = 0.87$ $R = 0.42$ $U = 21$	Good in cluster detection
HDBSCAN	-	-	-	Not good in cluster detection.
DPC	$\rho = 8.4 \times 10^{-4}$ $\delta = 0.9985$	Non-scaled	-	Good in cluster center detection
sOPTICS	$\epsilon = 11.0 \times 10^{-5}$ $minPts = 5$ Best results	Non-scaled	$P = 0.84$ $C = 0.84$ $R = 0.86$ $U = 12$	
DBSCAN	$\epsilon = 6 \times 10^{-4}$ $minPts = 5$	Scaled	$P = 0.72$ $C = 0.81$ $R = 0.41$ $U = 22$	Good in cluster detection
OPTICS	-	-	-	Good in cluster detection.

Table 3.3: Results on SDSS sample.

3.3   Application to SDSS Real Space Galaxy Catalogue

We also applied same density-based algorithms to the SDSS Real Space Galaxy Catalogue as shown article [24].

Once corrected the distorsions we talked in 3.1 all algorithms resulting in a better fitting model detection compared with the ones shown in the 3.2 section which contain space distorsions as showed at 3.1.

3.4   Impact of Standardization (results on scaled data)

While the spatial coordinates were initially defined on a consistent physical, to ensure that our density metrics remained isotropic and independent of the varying scales of the coordinate axes, we implemented a standardization protocol (Z-score normalization).

By transforming the spatial features to have a mean of zero and unit variance—calculated as $z = (x - \mu)/\sigma$—we eliminated the numerical bias inherent in raw coordinate ranges.

This preprocessing step yielded a measurable increase in cluster detection sensitivity across the 2dFGRS, SDSS, and Real-Space Galaxy catalogues as we can see in tables , and . This confirms that enforcing numerical isotropy is essential for correctly identifying groups in the all geometries.

Algorithm	Hyperparameter	Data Sample	Outcomes Sample	Conclusion
DBSCAN	$\epsilon = 3 \times 10^{-4}$ $minPts = 5$	Non-scaled	$P = 0.83$ $C = 0.92$ $R = 0.99$ $U = 6$	Good in cluster detection
HDBSCAN	-	-	-	Not good in cluster detection.
DPC	$\rho = 8.5 \times 10^{-4}$ $\delta = 0.9986$	Non-scaled	-	100% in cluster center detection
DBSCAN	$\epsilon = 2.6 \times 10^{-2}$ $minPts = 5$	Scaled	$P = 0.88$ $C = 0.88$ $R = 0.97$ $U = 7$	Good in cluster detection
OPTICS	-	Scaled	-	Good in cluster detection.

Table 3.4: Results on SDSS Real Space Galaxy Catalogue sample [24].

4   Conclusions of density-based algorithm appllication on different catalogs

The results of our comparative analysis along different catalogs can be summarized in the following key findings:

1. Superiority of Standardized Data: The application of Z-score normalization was the single most influential factor in improving model performance. By ensuring numerical isotropy, the scaled variants (sHDBSCAN, sDBSCAN, and sOPTICS) consistently outperformed their raw-coordinate counterparts across all metrics (P, C, R, and U).

2. Best Performance of sOPTICS (sDBSCAN): Among the tested frameworks, sDBSCAN emerged as the most robust model for recovering the underlying halo distribution. It achieved the highest Valid Match Ratios.

3. Modifying the distance in an elongated way along the Line of Sight improves the detection as sOPTICS algorithm which fit with [17].

4. Elbow method: We observed that traditional geometric optimizations, such as the Elbow Method, are unsuitable for clustering detection. As it was said, our results confirm that local density reachability is a more physically accurate proxy for gravitational binding than global variance minimization.

5 Applying 2PCF Estimators to the 2dFGRS Dataset

This is another mode to analysis the density of galaxies across the Universe: we will move from a discrete classification (deciding which galaxy belongs to which group) to statistical distribution (measuring how galaxies "crowd" together across the entire manifold). While clustering algorithms tell where the groups are, the Two-Point Correlation Function (2PCF) deals with the probability of finding galaxies at specific distances from each other.

We created a python notebook based on the dataset shown in section 1.1. Loaded the 2dFGRS dataset we take a sample to enable us to compare the distribution in the space of galaxies contained in the sample. By contrasting the 2dFGRS sample with a synthetic Poisson we expect this approach may convey information about the geometry of the matter across the Universe.

One interestring point is the use of scipy.spatial.KDTree which allow to improve the time response and calculations by providing an index in a set of k-dimensional space. This indexing strategy was pivotal for both the density-based cluster extraction and the 2PCF pair-counting, reducing the computational complexity from quadratic to logarithmic scales. This ensured that our hyperparameter grid search remained performant even when processing 3D comoving coordinates across 100 Mpc/h scales.

The obtained results are interestring because we can observe a peak over the $100h^{-1}Mp$ which can be explained by Baryonic Acoustic Oscillations, this result provides definitive evidence that our density-based clustering methodology recovers the fundamental 'fingerprint' of the early Universe still visible in the galaxy distribution and consistent with predictions of the $\lambda - CDM$ model.

The BAOs is also one more demonstration of the dark matter presence, sparse in the Universe, shaping it, creating the halos where clusters shown in this study reside.

Finally, our results show that the clusters identified by sDBSCAN and sOPTICS are not random associations, but are the physical manifestations of galaxies residing within the gravitational potential wells of Dark Matter halos, whose distribution was dictated by the sound horizon of the early Universe."

6 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. **BAO (Baryon Acoustic Oscillations)**: Periodic fluctuations in the density of the visible baryonic matter of the universe,

acting as a "standard ruler" for cosmological distances.

Λ**CDM Model**: The current standard model of cosmology, representing a Universe dominated by Dark Energy (Λ) and Cold Dark Matter (CDM).

Dark Matter Halo: A quasi-equilibrium state of dark matter particles. The gravitational "wells" where galaxies and galaxy clusters form and reside. See the introduction section in **??**.

Kaiser effect: distorsion in apparent clustering of galaxies that appeaers caused by motions of galaxies as they fall into large structures, causing a "squishing" or flattening along the line of sight. See [24] for more details.

k-**d Tree (k-dimensional Tree)**: A space-partitioning data structure used to organize points in a k-dimensional space. It allows for high-speed "nearest neighbor" searches, essential for large datasets.

Poisson Distribution: A random spatial distribution where points are placed independently. This serves as the "null hypothesis" against which the 2dFGRS clustering is measured.

Redshift Space Distortions (RSD): An effect where the observed distance of a galaxy is distorted by its peculiar velocity (movement due to gravity), causing clusters to look elongated ("Fingers of God") [24].

Z-score Normalization (Standardization): A preprocessing step that transforms data to have a mean of 0 and a standard deviation of 1, preventing one feature (like survey depth) from dominating the distance calculation.

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] Kerscher M. et al. (2000). A comparison of estimators for the two–point correlation function.

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

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

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

[11] 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.*

[12] 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.*

[13] 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.*

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

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

[16] 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.*

[17] Ma et al.(2025). soptics: A modified density-based algorithm for identifying galaxy groups/clusters and brightest cluster galaxies. *ArchivX.*

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

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

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

[21] P. J. E.(1980) Peebles. The large-scale structure of the universe. *Princeton University Press.*

[22] Alex Rodriguez and Alessandro Laio.(2014). Clustering by fast search and find of density peaks. *Science.*

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

[24] Yang X. et al .(2016) Shi Feng. Mapping the real-space distributions of galaxies in sdss dr7. i. two-point correlation functions. *The Astrophysical Journal, Volume 833, Issue 2, article id. 241, 19 pp. (2016). https://doi.org/10.48550/arXiv.2409.13010.*