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, October 27, 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 obtaining large catalogs of galaxies. One objective of these studies is to map vast areas of the universe. This work relies on data comming from two such examples: 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 [12] to form the so-called cosmic web [3].

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 work better and its possible causes.
- Detect possible existence of outliers and patterns.
- Use validation methods to obtain a hyper-parameter tunning in order to optimize galaxy 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 on computational efficiency directly translates into lower energy consumption, this is the most tangible sustainability implication in this project.

[1] For exemple: modify the distance in order to mitigate the already known Redshift-Space distortions along the line of sight[13].

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 [9]:

  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 contains model building througth application of the algorithms and validation of the outcomes.

- Visualization: graphical view of the results.

[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.

The third point is the longest; in fact, it is an iterative process dedicated to improving the results obtained from the models. All models will consist of unsupervised algorithms, particularly density-based ones.

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-^{0}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 project development. Excluding the final project defense, the stages have been grouped on three blocks:

- The Planning stage (shown in green) involves gathering resources and defining the project's objectives.
- The technical development stage (shown in red) includes design, data processing, method application and outcomes assessment.
- Research and writing stages (shown in blue).

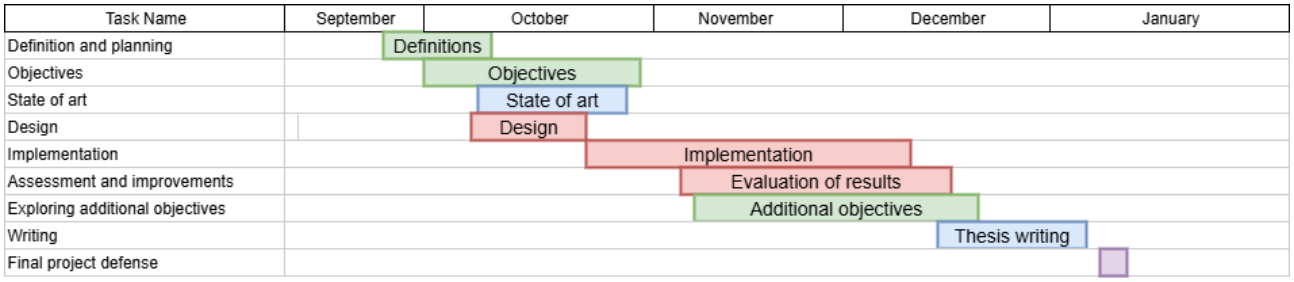

Figure 1.1: Stages of the project.

An iterative and continuous review of the results is 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 be rearranged and redefined. This is why the 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 focusing on two distinct components: first, the inherent challenges associated with survey-collected data; and second, an overview of the Machine Learning techniques that will be applied throughout this study.

1 Spectroscopic Surveys

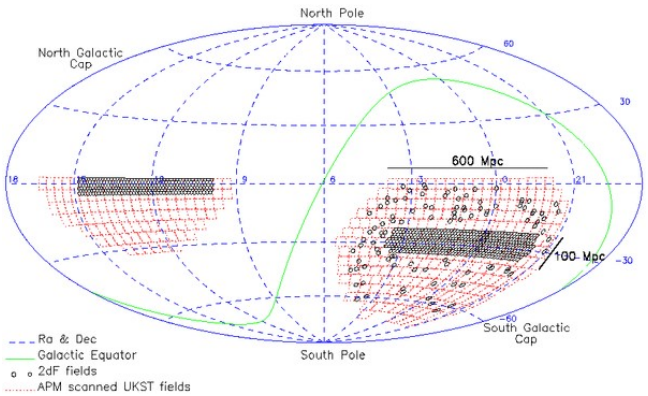

Figure 2.1: 2dfGRS sky coverage obtained from [6]

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:

- 2dfGRS: Contains 245591 objects, of which 221414 are considered reliable galaxy data. The final data release was published in 2003, the survey leveraged the unique capabilities of the 2dF (2-degree Field) facility built by the Anglo-Australian Observatory in the southern hemisphere. A view of 2dfGRS coverage is shown in 2.1

- SDSS modelC petrosian magnitude data realease 7 (DR7) 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 2.2. Groups constructed are drawn up from [14].

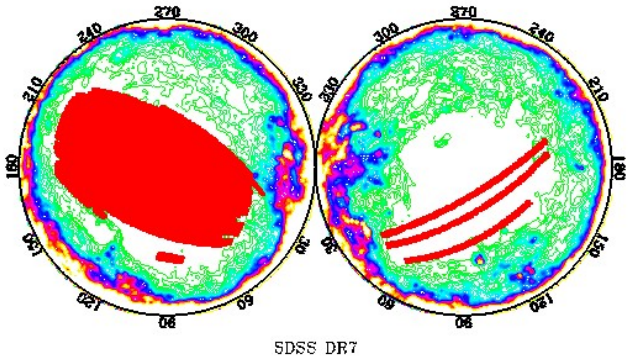

Figure 2.2: SDSS Data release 7 sky coverage obtained from [14].

All redshift survey are subject to several kind of challenges, among others:

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 [2]. 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 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.cThis relationship is described by the equation:

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

The equation 2.1 is strictly valid on small redshifts $z \ll 1$ [5] (which means nearby objects), for higher redshifst it is necessary to use a full cosmological model address the redshift-distance relation. According with the most recently theories [5]:

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

.

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 with the chosen cosmological model. In this work we assume the values of the ΛCDM, according with [5]:

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

.

3 Machine Learning applyied to cosmology

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

3.1 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 $F(X_i)$

In this context, 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.

3.2 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: work by in distance and simillarity patterns, they can be divided as

- Hierarquical: work by create sucessive partition of data and hierarquical tree creation called dendogram. Examples agglomertative clustering.

- Partitional: an initial set of clusters must be set in advance, the set is improved on an iterartive proccess. Example k-means.

- Model-Based: Algorithms that assume the data is generated by a mixture of underlying probability distributions (e.g., Gaussian Mixture Models, GMM).
- Density-based: define 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.

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.

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.

3.3 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 regading a core-point g is the maximum of the two: core-distance and euclidean distance, in other words:

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

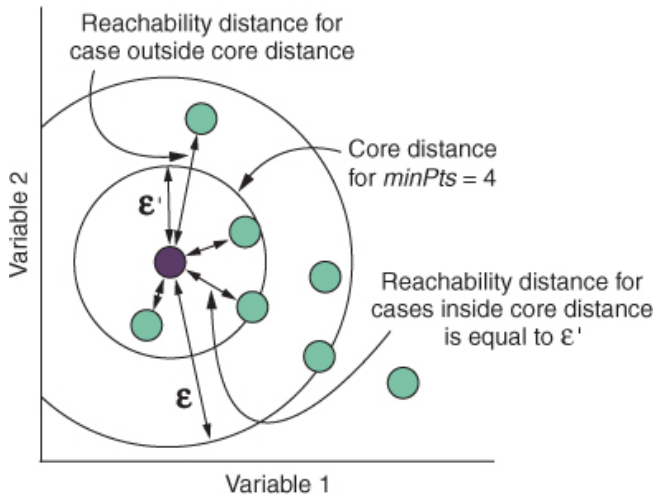

Figure 2.3: Core and reachability distances obtained from [4].

Note that reachability-distance is only defined respect to a core-point. We can see an illustrative example of both core-distance and reachability-distance in the figure 2.3.

OPTICS work by seting 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.

For example figure 2.4 shows a random-generated set points arround four fixed points within [0,1]x[0,1] square. OPTICS is then applied to this set, resulting reachability plot is shown in figure 2.5.

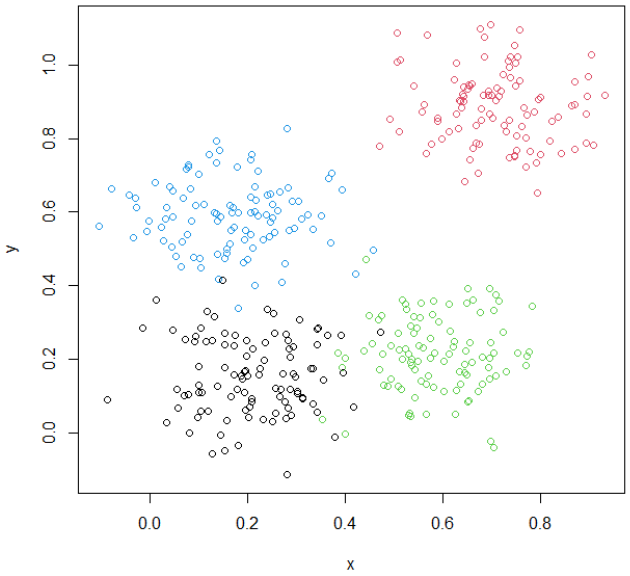

Figure 2.4: An example of data set in plane $\mathbb{R}^2$.

3.4 DBSCAN

DBSCAN 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 such that:

  1. $p_1 = p$ and $p_n = q$.
  2. p_{i+1} is directly density-reachable from p_i for all $1 \leq i < n$.

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

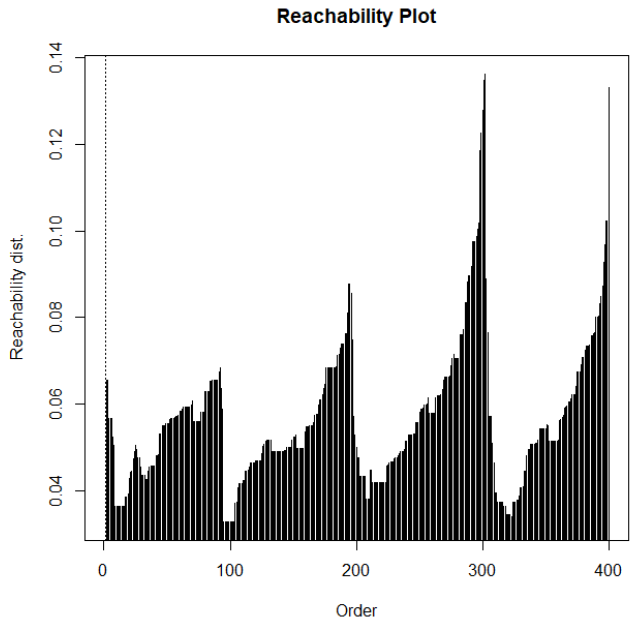

Figure 2.5: Example of OPTICS reachability plot

- $\forall p, q, \in C$ 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*.

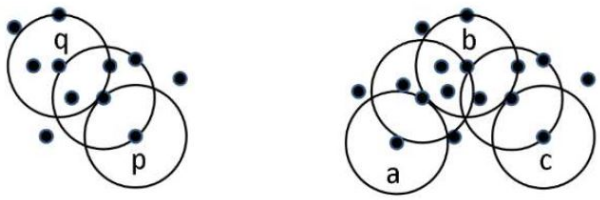

Figure 2.6: Left: density reachablability. Right: density connectivity, from [4].

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 ← 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 said, 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.

3.5 HDBSCAN

3.6 Previous machine learning applications in galaxy clustering

This section briefly reviews several Machine Learning (ML) applications in cosmology, with particular emphasis on clustering techniques.

In 1937, Erik Holmberg in Lund Observatory in Sweden published an article [1] in which he investigated the clustering tendencies in the metagalactic system. The article also includes a catalog of 827 galaxy groups.

It is easy to find works us supervised methods, for example, Thomas et al. [8] generate predictive regression models based on the MACSIS simulation to predict cluster features from specific observables. On the other hand, Sadikov et al. [7] 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. [10] 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" [11].

The most significant application of density-based algorithms to galaxy distribution is a recent article (dated 2025) by Hai-Xia-Ma et al.[13]. 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 used a modified version of OPTICS called sOPTICS to mitigate the redshift space distortion along line-of-sight caused by galaxies' peculiar velocities.

Bibliography

[1] Holmberg E. (1937). A study of double and multiple galaxies. *arXiv e-prints*.

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

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

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

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

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

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

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

[9] 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*.

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

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

[12] 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*.

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

[14] Xiaohu et al.(2007) Yang. Galaxy groups in the sdss dr4: I. the catalogue and basic properties. *ArchivX*.