UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

THE VOCLUDET GALAXY CLUSTER FINDER: OPTIMIZATION, VALIDATION AND VISUALIZATION

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS, MENCIÓN COMPUTACIÓN

SEBASTIÁN ALFREDO PEREIRA GALLARDO

PROFESOR GUÍA:
NANCY HITSCHFELD KAHLER

PROFESOR COGUÍA:
LUIS CAMPUSANO BROWN

MIEMBROS DE LA COMISIÓN:
ALEXANDRE BERGEL
FRANCISCO FORSTER BURON
JUAN MARÍN CAIHUAN

SANTIAGO DE CHILE
2016

Resumen

Actualmente la astronomía se enfrenta al reto de manejar y analizar el gran volumen de información que se genera cada día. Muchas de las preguntas sin respuesta que existen en la actualidad requieren medidas de alta precisión de los diversos componentes del universo. Por esta razón, se necesitan algoritmos de procesamiento de datos, que sean capaces de tomar la cantidad masiva de datos y procesarla para obtener catálogos enriquecidos de información que sean de fácil acceso.

Los cúmulos de galaxias, siendo altamente masivos, nos permiten trazar las regiones de alta densidad de distribución de la materia, por lo que una muestra completa proporciona entre otros, una descripción de la formación de estructuras en el universo temprano. Vocludet es un algoritmo de detección de cúmulos de galaxias, que detecta estos objetos utilizando las propiedades geométricas y astrofísicas de las galaxias. Este trabajo describe el análisis, validación y optimización de Vocludet, a través del uso de los datos artificiales, obtenidos a partir de una simulación de distribución de materia. Para la validación, los resultados del algoritmo se comparan con el catálogo de datos simulados, en términos de tasa de recuperación y pureza, es decir, qué fracción del catálogo de referencia se recupera y qué fracción de los grupos detectados son reales, respectivamente. La simulación de datos utilizada consiste en un catálogo artificial del Milennium Run, una gran simulación del universo. Este catálogo contiene información acerca del agrupamiento de galaxias que puede ser usado para comparar con los resultados obtenidos por Vocludet. Además, una herramienta de visualización se desarrolla para mostrar de forma interactiva los grupos en cualquier plataforma que posea un navegador de Internet moderno. Esto último con el propósito de realizar debugging, así como también presentar el resultado final.

Los resultados finales indican que Vocludet tiene una tasa de recuperación de $\sim 59\%$ en general y $\sim 66\%$ de pureza. Sin embargo, cuando se restringe el análisis sólo a los cúmulos con más de 10 galaxias, las tasas de recuperación y pureza son $\sim 75\%$ y $\sim 90\%$ respectivamente. Además, otras propiedades de interés de los cúmulos tales como dispersiones de velocidad presentan un estrecho acuerdo con los valores correspondientes para los cúmulos de referencia, lo que refuerza aún más la evidencia de Vocludet como un detector de clúster fiable.

Abstract

Nowadays, astronomy faces the challenge of handling and analyzing the vast amount of information being generated every day. Many of the unanswered questions that exist today require high precision measures of the diverse components of the universe. For this reason, data processing algorithms are needed, which are able to take the massive amount of data and process it to obtain rich information catalogs that are easily accessible.

Galaxy clusters, being highly massive, allow us to trace the high density regions of matter distribution, so a complete sample provides amongst other, a description of the structure formation in the early universe. Vocludet is a galaxy cluster detection algorithm, that detects these objects using geometrical and astrophysical properties of galaxies. This work describes the analysis, validation and optimization of the Vocludet galaxy cluster detection algorithm through the use of mock data from a matter distribution simulation. For the validation, the results of the algorithm are compared with the catalog of simulated data, in terms of recovery rate and purity, that is, how much of the reference catalog is recovered and what fraction of the detected clusters are real ones, respectively. The simulated data used consists of a mock catalog of the Millennium Run, a very large simulation of the universe. This catalog contains information about galaxy clustering that can be compared against the obtained results. Additionally, a visualization tool is developed to interactively display the clusters on any platform having a modern Internet browser with the purpose of debugging as well as presenting the final outcome.

The final results indicate that Vocludet has an overall $\sim 59\%$ recovery rate and a $\sim 66\%$ purity. However, while restricting the analysis only to clusters with more than 10 galaxies, the rates of recovery and purity are $\sim 75\%$ and $\sim 90\%$ respectively. Furthermore, other properties of interest of the clusters such as velocity dispersions agree closely with the corresponding values for the reference clusters, which further strengthens the confirmation of Vocludet as a reliable cluster detector. Similar works have reported overall completeness rates close to 55% while overall purities reported are $\sim 90\%$, thus making Vocludet a competitive algorithm in the field.

Una dedicatoria corta. Por ejemplo, A los creadores de U-Campus

Contents

List of Tables

List of Figures

Introduction

At present, astronomy faces the challenge to manage and analyse the vast amount of information being generated every day. This challenge becomes even more relevant with the pass of time, since there are future projects in which the amount of data required to be processed is even greater. An example of this is the Large Synoptic Survey Telescope project (LSST) [1], which will be able to create a detailed and quick mapping of the whole-sky in just a couple of nights, generating data in the order of Terabytes each day.

Many of the questions that currently exist in astronomy need high-precision measurements of multiple components of the universe to be able to find an answer. This is why processing algorithms are needed to take the vast amount of data available and classify it in order to obtain information-rich catalogs and allow scientists to access them in a simple and coherent manner. In particular, these kind of catalogs are essential for the study of cosmology.

Galaxy clusters are systems which contain from a few to thousands of galaxies, dominated by elliptical galaxies and with extensions of the order of millions of light years. Each galaxy, in turn, typically comprises hundreds of millions of stars, many of them much more massive than the sun. These objects, being so massive, serve as a probe for the mapping of high-density regions in matter distribution. As a result, a complete sample provides, among other results, a description of the formation of structures in the early universe. This is why the study of the distribution of clusters of galaxies is fundamental to answer many of the questions in cosmology, besides testing existing theories or formulate new ones based on the available data.

Definitions

Since this work is strongly tied to astronomy, it features a multitude of terms related to said field with which the reader might not be familiar. Some of these terms are briefly described next.

Galaxy cluster

A galaxy cluster, or cluster of galaxies, is a structure that consists of anywhere from hundreds to thousands of galaxies that are bound together by gravity with typical masses ranging from $10^{14}-10^{15}$ solar masses. They are the largest known gravitationally bound structures in the universe.

Parsec

Is a unit of length used to measure large distances to objects outside the solar system. One parsec is the distance at which one astronomical unit subtends an angle of one arcsecond. Its symbol is pc, but it is usually found as Mpc — a megaparsec — equal to 10^6 parsec.

Redshift

Redshift is defined as the change in the wavelength of the light divided by the wavelength that the light would have if the source was not moving — called the rest wavelength. Due to the accelerated expansion of the universe, objects further from the observer present higher redshift values, which means it can be used as a proxy to measure distances. It is represented as z.

Velocity dispersion

In astronomy, the velocity dispersion (σ) is the statistical dispersion of velocities about the mean velocity for a group of objects, such as an galaxy cluster.

Hubble constant

The Hubble Constant is the unit of measurement used to describe the expansion of the universe.

Abell radius

It is defined as $1.72/z$ arcminutes, where z is the redshift of the cluster, which is equivalent to $1.5h^{-1}Mpc$, where h is the dimensionless Hubble constant in units of $100kms^{-1}Mpc^{-1}$.

Voronoi Tessellation

Let X be a metric space with distance function d. Let K be a set of indices and let $(P_k)_{k\in K}$ be a tuple (ordered collection) of nonempty subsets (the sites) in the space X. The Voronoi cell, or Voronoi region, R_k, associated with the site P_k is the set of all points in X whose distance to P_k is not greater than their distance to the other sites P_j, where j is any index different from k. In other words, if $\mathrm{d}(x,\, A) = \inf\{\mathrm{d}(x,\, a) \mid a \in A\}$ denotes the distance between the point X and the subset A, then

$$R_k = \{x \in X \mid \mathrm{d}(x, P_k) \leq \mathrm{d}(x, P_j) \text{ for all } j \neq k\}$$

The Voronoi diagram is simply the tuple of cells $(R_k)_{k\in K}$.

Delaunay triangulation

The Delaunay triangulation (also known as Delaunay Mesh) of a discrete point set P in general position corresponds to the dual graph of the Voronoi tessellation for P. The dual graph of a plane graph G is a graph that has a vertex for each face of G.

Abell Radius (R_A)

The Abell radius of a cluster, used to measure its compactness and extent, is the distance out to which cluster members are counted. It is defined as $1.72/z$ arcminutes, where z is the mean redshift of the cluster.

Motivation

In 2006, the master's thesis entitled "Galaxy Cluster Nonparametric Detection using Maximum Likelihood Estimation of Features in Voronoi tessellations" [2] is developed. In it a galaxy cluster detection algorithm (Vocludet) is described. The algorithm is designed to detect multiple clusters in three dimensional space by using the Voronoi tessellation to detect high-density regions in space, in addition to using astrophysical properties to determine the components of each cluster. One feature to be improved in this algorithm is the galaxy purity in the resulting clusters, that is, to reduce the amount of false positive galaxies in the final clusters. This is discussed in more detail in chapter 2. Another pending aspect of the previous work that remains undone is the validation of the algorithm. Besides this, creating a data visualization software is invaluable to the study of the algorithm and its results.

Part of the pending work was done as part of a bachelor's thesis during the first half of 2014, by the author of this thesis. In that work, part of the validation was made, along with a wall display visualization to study the broad results of the algorithm. General performance statistics were obtained, indicating results comparable to other cluster detectors in the literature. However, a more in depth analysis is required in order to understand the full capabilities of Vocludet. In particular, an independent assessment of the performance of each of the two stages is needed. Therefore, this thesis poses an extension to the previous work, and seeks to deepen the analysis of the algorithm, and generate optimizations both in efficiency and quality of the results.

Objectives

The general and specific objectives are next described.

General objective

The general objective consists in the optimization and validation of the Vocludet algorithm and the creation of an interactive visualization tool to help with the study of the results.

Specific objectives

- Optimize the algorithm. This objective considers the optimization not only of the free parameters of the algorithm, but also a possible modification of the algorithm itself.
- Validate the algorithm by making use of a simulated catalog.
- Obtain measures of the algorithm performance so it can be compared to other procedures.

- Design and implement a visualization tool which allows to study different runs of the algorithm, using multiple sources of data.

Methodology

The methodology adopted to accomplish the objectives is next described.

For the algorithm validation:

- Study and examine existing literature: Vocludet software thesis, documentation and related bibliography. This includes the study of the individual characteristics and distribution of galaxy clusters, as well as the physics applied in related fields such as cosmology, astrophysics and others.

- Acquire and analyse the data to be used for the validation, select an appropriate sample and do the necessary processing required for the algorithm to read the data.

- Evaluate the results of the execution of the algorithm. Implement statistical analysis scripts in order to compare and evaluate the quality of the results. In this stage, the variation of the algorithm's parameters must be taken into consideration.

- Improve the quality of the detection, in case the results are not the expected and/or they can be enhanced.

For the visualization tool:

- Determine the best technology available that suits the needs of the study.

- Design and implement the 3D visualization of the clusters and the galaxies close to them. The application must be able to support multiple runs of the algorithm with varying parameters and catalogs of galaxies.

- Implement the user interaction with the application. In this stage it must be considered that the user has to be able to visualize the clusters from different perspectives so a simple interface is needed.

Thesis contents

This thesis is organized as follows: Chapter 1 describes related content based on the literature of galaxy cluster finder algorithms and visualization techniques, as well as the theoretical basis of this work. Chapter 2 covers the algorithm itself: the previous version and the improved one. Chapter 3 describes the tools developed to accomplish the objectives of this work. Chapter 4 explains the optimization process. Chapter 5 describes in detail the analysis of the results and the validation of the algorithm. And finally, the last chapter summarizes the thesis and proposes future lines of work.

Contributions of this research

The contributions of this research are:

- The algorithm: Validated and optimized through the use of simulated galaxy data, which allows it to be used on existing real galaxy catalogs, with the purpose of enriching them with new information and to study the galaxy cluster distribution in the universe.

- The visualization tool: The software facilitates the analysis of the clusters but it can also be applied to the visualization of other 3D point distribution with a similar tendency of clustering.

- Publication: The work describing the validation and optimization of the algorithm is material fit for scientific publication. "Pereira S., Hitschfeld-Kahler N., Campusano L. et al., A 3D Voronoi+Gapper Galaxy Cluster Finder in Redshift Space to $z \sim 0.2$ I: Algorithm and Validation, to be submitted to the Astrophysical journal July 2016"

Chapter 1

Literature review

1.1 Algorithms

There are multiple strategies utilized to detect galaxy clusters. One of them consists in the search for overdensities in the galaxy distribution. For this, different approaches are used, such as friend-of-friends [3] algorithms, density maps [4] or Voronoi tessellations [5] [6].

Other methods make use of astrophysical properties of galaxy clusters. One example of this is the red sequence technique [7], which utilises the fact that galaxy clusters posses a highly regular population of elliptical and lenticular galaxies. An elliptical galaxy is a type of galaxy having an approximately ellipsoidal shape and a smooth, nearly featureless brightness profile, while a lenticular galaxy is an intermediate between an elliptical galaxy and a spiral galaxy in terms of morphological classification [8] [9] [10].

Alternatively, other methods are based on galaxy luminosity profiles, such as [11] [12] [13].

Each one of these approaches makes assumptions about the general properties of the clusters and therefore the generated catalogs are constituted mainly by clusters reflecting these assumptions.

1.2 Theoretical basis

The first stage of the algorithm presented in this work is heavily based in geometry and statistics. The stage is called VT-MLE (Voronoi Tessellation - Maximum Likelihood Estimator) due to the use of the Voronoi Tessellation and the statistical technique MLE. The Voronoi Tessellation is a partitioning of a plane into regions based on distance to points in a specific subset of the plane. That set of points is specified beforehand, and for each seed there is a corresponding region consisting of all points closer to that seed than to any other. These regions are called Voronoi cells. The Voronoi diagram of a set of points is dual to its Delaunay triangulation. Each face of a Voronoi cell is defined by a convex polygon whose

normal vector is parallel to the edge that connects a point-galaxy and its neighbour point-galaxy. The Voronoi tessellation into convex cells (convex polyhedra) provides a natural way to measure the packing of the objects. The volume of each cell is inversely proportional to the packing efficiency of its point; a large cell volume indicates that its point-galaxy is comparatively isolated. An example of this can be seen in figure 1.1.

When the Voronoi Tessellation is applied on a set of galaxies, the volume of each cell directly determines the density distribution because $\rho_{\rm i} = V_{\rm i}^{-1}$, where $\rho_{\rm i}$ and $V_{\rm i}$ are respectively the galaxy number density and the volume of a cell associated with an object i. Thanks to this fact, a ranking of cells can be produced, in which each cell is ordered by decreasing density, i.e. high density cells are high in the ranking. The simplest approach to locate the density peaks is to define a density contrast with respect to the background. The density contrast, $\delta_{\rm i}$, at the position of the ith object is defined as:

$$\delta_{\rm i} = (\rho_{\rm i} - \bar{\rho})/\bar{\rho} \tag{1.1}$$

$$\bar{\rho} = \frac{1}{n}\sum_{\rm i=1}^{n}\frac{1}{V_{\rm i}} \tag{1.2}$$

where $V_{\rm i}$ is the volume of the Voronoi cell around object i, and n is the overall number of objects. In order to be locally adaptive, in each iteration $\bar{\rho}$ is re-defined over a restricted domain over which the local mean density is computed and the MLE determined. The MLE is evaluated for structures of point-galaxies superimposed on "noise" (i.e. unrelated to the clusters) and produces a mixture of two random samples: (i) structures characterized by the probability p (the mixture parameter) and the total volume of the cells in the cluster (support) $V \subset K$, where K is the overall volume of investigation, i.e., the convex hull of the total set of points; and (ii) unrelated points with complementary probability $1-p$ and support K. The mathematical framework presented below was originally proposed and applied for a 2D space [14], and later extended to 3D.

The density associated with a point $x \in K$ is

$$f(x) = \frac{p}{|V|}\mathbf{1}_V(x) + 1 - p \tag{1.3}$$

where $\mathbf{1}_V(x) = 1$ if $x \in V$ and $\mathbf{1}_V(x) = 0$ otherwise, and $|V|$ denotes the Lebesgue measure of V, which is the normalized volume (ratio between cell and domain volume). The likelihood that a particular ensemble of points is a structure, is

$$\mathcal{L}(x; V, p) = \prod_{\rm i=1}^{n}[f(\mathbf{x}_{\rm i})] = (\frac{p}{|V|} + 1 - p)^{N_V}(1-p)^{n-N_V} \tag{1.4}$$

where N_V is the number of objects in V and n is the number of objects in K. Because the mixture parameter p and V are not known, the above have to be reduced to a partial maximized (profile) likelihood. For a fixed V, the MLE of p is

$$\hat{p} = (N_V - |V|n)/(n - |V|n) \tag{1.5}$$

and the partial maximised likelihood becomes

$$L(x; V) = \left(\frac{1}{n}\right)^n \left(\frac{N_V}{|V|}\right)^{N_V} \left(\frac{n - N_V}{1 - |V|}\right)^{n - N_V} \tag{1.6}$$

which can be more conveniently expressed as a log-likelihood

$$l(x; V) = -n \ln n + N_V \ln\left(\frac{N_V}{|V|}\right) + (n - N_V) \ln\left(\frac{n - N_V}{1 - |V|}\right)$$

V is constructed from Voronoi cells, ensuring that any constraints are defined by the data points themselves.

The following approach was adopted to find the estimator $\hat{V}$ for V. To compute $\hat{V}$: first, $\hat{V}$ is initialized as the empty set, and then, after adding the starting cell, new cells are merged into $\hat{V}$ one at a time in ascending volume order. At each stage, *posterior* to the starting step, each cell of the outside border is considered a candidate for merging, and the variation in likelihood of the closure after it is merged is computed, and the merging corresponding to the maximum likelihood closure is ultimately selected. After all cells are merged into $\hat{V}$, the stage at which the log likelihood is maximized is chosen as the approximate MLE.

1.3 Mock galaxy data

A significant amount of data is required for the validations. The use of real data is discarded since it contains an intrinsic error due to the limitations of the measuring tools, the errors introduced by the data processing pipeline and the fact that the distance to an object in space in only calculated through a proxy measurement (usually the redshift). A first alternative of synthetic data is the custom generation of mock galaxy catalogs. This approach, however, is disadvantageous due to the complexity of the task and the difficulty of reproducing the intricate interactions of the vast amount of objects in space. A second, more practical approach, is to use the available data of already calculated simulations. Catalogs obtained from this kind of sources have already been validated by the scientific community and they usually provide easy access to a wide range of options.

One of these cases is the Millennium Simulation. The simulation contains 2160^3 particles of mass $8.6 \times 10^8 h^{-1} M_\odot$ within a comoving box of size $500 h^{-1}$ Mpc on a side. The Millennium database provides positions and velocities of all simulated particles which are stored in 63 snapshots, spaced logarithmically from $z = 20$ to $z = 0$, where dark matter halos are identified using a standard friends-of-friends (FOF) algorithm with a linking length of 0.2 units of the mean particle separation.

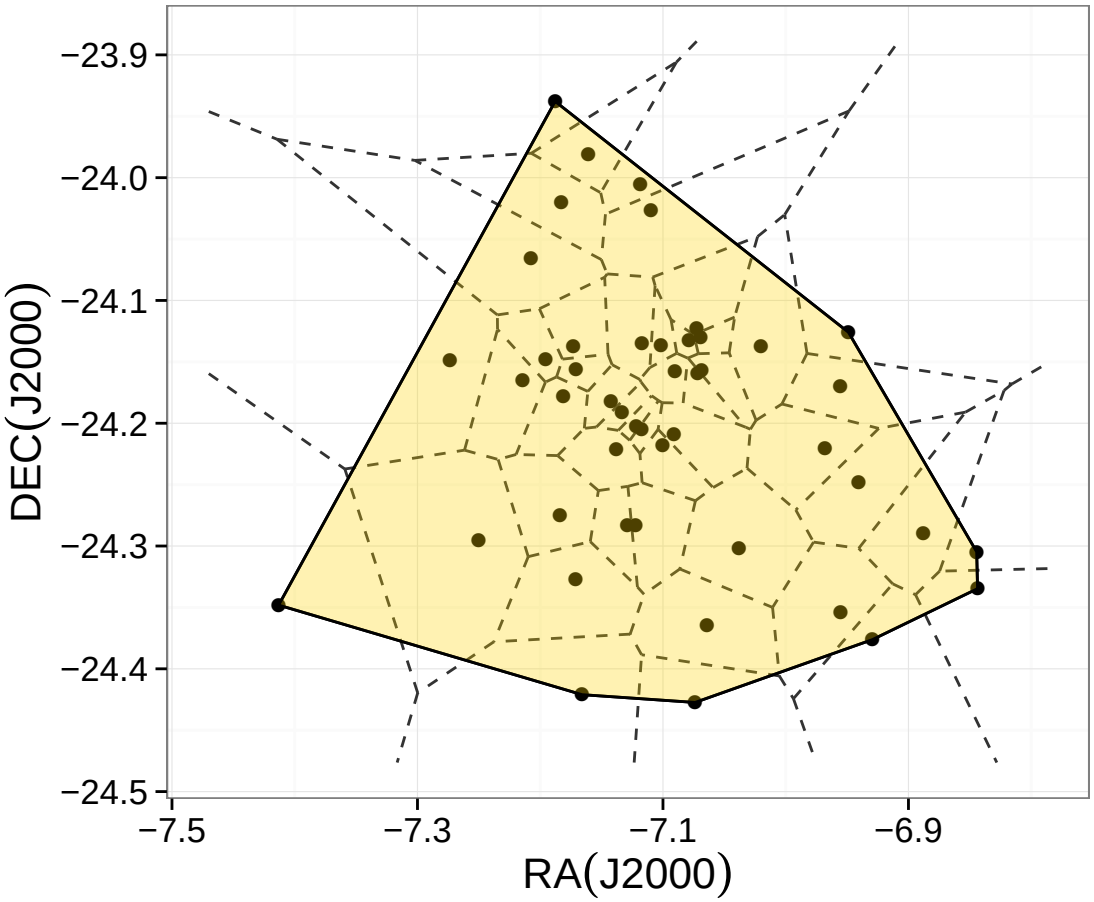

Figure 1.1 Example of a 2D Voronoi tessellation applied on a previously detected VTMLE cluster. Each point represents a galaxy and the dashed lines the Voronoi tessellation. The polygon is the convex hull of the cluster. In this figure it can be seen the relation between cell size and local density, since the cells get smaller the closer to the denser central region they are. The actual algorithm uses a 3D tessellation.

1.4 Visualizations

Visualizations play a crucial role in astronomy, due to the large amount of data and the complexity of the subjects. Most of the visualizations developed consist of a variety of static images and plots. Some of those plots are standard, such as the wedge plot, that allows to overview the spatial distribution in a catalog of celestial objects. Other kinds of plots are generated on a case-to-case basis, depending on the objective of the study and the type of data to be used.

For more complex data, and usually as a way of helping with science outreach, short animations or movies are created. An example of this is the Millennium Simulation movie of the large-scale structure in the Universe [15] [1]. The movie shows the dark matter distribution in the universe at the present time, based on the Millennium Simulation, the largest N-body simulation carried out at the time. By zooming in on a massive cluster of galaxies, the movie highlights the morphology of the structure on different scales. Another example of this approach is the video created for the study of the Laniakea supercluster of galaxies [16]. The video illustrates the observed local distribution of galaxies, the observed departures from the expansion of the universe of the fraction of the galaxies with distance measurements, the inferred three-dimensional flow pattern of the local galaxies, and the inferred underlying distribution of matter causing these flows.

Finally, and much less common are interactive visualization tools developed specifically for a study. The rare occurrence of these cases is explained by the high development cost of the software in terms of time and resources. Additionally these tools are usually very specialized, which decreases the chances of reusability in other studies. An example of this is the GyVe tool [17], an interactive visualization tool for understanding structure in sparse three-dimensional (3D) point data. The scientific goal driving the tool's development was to determine the presence of filaments and voids as defined by inferred 3D galaxy positions within the Horologium-Reticulum supercluster (HRS). GyVe provides visualization techniques tailored to examine structures defined by the intercluster galaxies. Specific techniques include: interactive user control to move between a global overview and local viewpoints, labelled axes and curved drop lines to indicate positions in the astronomical RA-DEC-cz coordinate system, torsional rocking and stereo to enhance 3D perception, and geometrically distinct glyphs to show potential correlation between intercluster galaxies and known clusters.

[1] `https://wwwmpa.mpa-garching.mpg.de/galform/virgo/millennium/`

Chapter 2

Background

2.1 The original Vocludet Algorithm

The Vocludet algorithm, originally described in [18, 19], was developed through the integration of two separate steps, each one a proven cluster finder taken from the literature. The first one, based on the geometry of the distribution of galaxies, is an extension of the Voronoi-Tessellation-based Maximum Likelihood Estimator [20, 21] and it is used to determine points in space that mark regions of greater galaxy density. The second one is the ZHG technique [22], which makes use of the observed separation of cluster galaxies from a background population in velocity space to determine galaxy membership for clusters, based on potential cluster positions. Both of these stages are described in detail in the next section.

2.1.1 Description

The VT-MLE algorithm (First stage)

The input for the algorithm is a set P of n points in a 3D space where each point can represent a galaxy. Before applying the cluster detection algorithm over P, the Voronoi tessellation T and the Delaunay mesh D of the set of points P are computed using *qhull* [23], a free software for computing the convex hull, the Delaunay triangulation, and the Voronoi diagram of a set of points. The Voronoi tessellation T gives an estimation of the local point density associated with each point $p_{\mathrm{i}} \in P$ as the reciprocal volume of the Voronoi cell $t_{\mathrm{i}} \in T$. Conversely, the vertices of the corresponding Delaunay mesh D, form an adjacency graph for the cells of T, in a way that for each cell $t_{\mathrm{i}} \in T$ associated with the point p_{i}, its neighbouring set $N(p_{\mathrm{i}})$ can be easily obtained. The banned seeds are those that have already been incorporated into a previous cluster.

The algorithm for the detection of a single cluster is shown in Algorithm 1, and the algorithm for the detection of multiple clusters (sequentially) is shown in Algorithm 2.

Algorithm 1 FindSingleClusterVTMLE(P, T, A, B, p_k)

Input: $P \leftarrow \{p_0, p_1, ..., p_n\}$, *points representing galaxies of the search domain.*
Input: $T \leftarrow \{t_0, t_1, ..., t_n\}$, *Voronoi Tessellation of* P.
Input: $A \subset P$, *set of already clustered points.*
Input: $B \subset P$, *set of banned seeds.*
Input: p_k, *seed point of this cluster.*
Output: a set of points corresponding to a cluster

```
R ←, Function. Computes the radius of a cluster
R_A ←, Function. Computes the Abell radius of a cluster
Cluster ← {p_k}
L_max ← L(Cluster) ,  Compute the partial maximized likelihood.
for all t* neighbour cell of Cluster, and its associated point p* do
  if L(Cluster ∪ p*) > L_max then
    if p* ∈ A then
      B ← B ∪ p*
      continue
    end if
    Cluster ← Cluster ∪ p*
    L_max ← L(Cluster)
  end if
  if R(Cluster) > 2R_A(Cluster) then
    break
  end if
end for
return Cluster
```

At the start, the algorithm selects the smallest volume cell in the whole Voronoi Tessellation. This cell is called a cluster "seed". The partial maximised likelihood (L) is first computed for this seed, and then subsequently calculated for the union of this seed and each of its neighbors; if it is possible to increment the value of the MLE by adding any of its neighbors to the cluster seed, then it is done. The process continues trying to increment the MLE value by adding more neighbors to the cluster until one of the following situations is encountered. (1) It is impossible to increment the MLE value for the actual structure by adding any of its neighbors; the structure information is stored.

(2) Every neighbor considered to be a potential member of this current structure is detected to be already a member of a previously-detected structure. Then, the current structure growth process is stopped and the seed and galaxies belonging to the growing structure are released (set available to be a part of other structures), because continuing on this structure would lead to an overlap. (3) The extent of the structure equals or exceeds 2 R_A; the process is then stopped and the structure information is stored.

Once the above process is completed, the algorithm starts again and selects the next cluster seed, that is, the smallest volume cell which is not part of any other previously detected cluster. The algorithm detects as many clusters as possible, without overlap, always proceeding from the smallest seed volume available, and finishes when all the available seeds

Algorithm 2 FindMultipleClustersVTMLE (P,T)

Input: $P = \{p_1, ..., p_n\}$, *points representing galaxies of the search domain.*
Input: $T = \{t_1, ..., t_n\}$, *Voronoi Tessellation of* P.
Output: a series of sets of points , where each set is a cluster

```
A ← {} , set of already clustered cells
B ← {} , set of banned seed cells
Clusters_i ← {} , series of found clusters
i ← 1
while (T ≠ A ∪ B) do
    Find the smallest volume cell t_k ∈ (T − (A ∪ B))
    Cluster_i ← FindSingleClusterVTMLE(P, T, A, B, t_k)
    A ← A ∪ Cluster_i
    ++i;
end while
return Cluster
```

have been considered. Note that since the algorithm proceeds sequentially over all potential cluster seeds, all the galaxies that comply to the rules will end up in a group.

ZHG algorithm (Second stage)

The input for the ZHG algorithm is a starting cluster position $\overrightarrow{d}$ in the sky, with a corresponding estimated redshift (z_{est}). The candidate member galaxies are the ones located within a cone corresponding to the Abell radius evaluated for the estimated redshift, $R_A(z = z_{est})$, and applied at the estimated cluster center. Once the final member galaxies are selected by ZHG technique and the final cluster redshift is computed. VOCLUDET does not impose a lower size limit and thus can detect a wide range of cluster/group extents.

There are two specific parameters in Algorithm 3, adopted according to the convention used by [24]: z_{gap}=0.00333 (1,000 km s^{-1}) and a maximum cluster size of 2 Abell Radius, R_A. The values of these parameters are astrophysically motivated and therefore they make possible a connection between the clusters selected by the algorithm and the ones existing in standard catalogs.

The ZHG member-identification algorithm is a fairly simple process, and its effectiveness depends on two factors. First, the good quality of the detection of the initial position $\overrightarrow{d}$ of the cluster center to be processed. The second factor is actually the membership detection algorithm itself: ZHG is a well-established technique to identify the cluster members in redshift space.

Algorithm 3 $ZHG(\mathrm{d}, P, A, z_{gap})$

Input: $\overrightarrow{\mathrm{d}}$ *, Vector pointing from origin to the center of a cluster* C*.*
Input: set $P \subset \mathbb{R}^3$ where each point represents a galaxy of the survey.
Input: A *, set of galaxies that already belong to another cluster.*
Input: z_{gap} *, redshift interval size used by the Excise function to filter out galaxies according to their depth*
Output: a set of points , where each point represents the position of a galaxy identified as member of the cluster.

$C' \leftarrow \{x \in P \mid \angle(\overrightarrow{x}, \overrightarrow{\mathrm{d}}) < R_A(\overrightarrow{\mathrm{d}})\}$
if $C' \cap A \neq \emptyset$ **then**
  $R' \leftarrow min\{\angle(\overrightarrow{\mathrm{d}}, \overrightarrow{a}), a \in A\}$ *, there was a colission, use a smaller radius.*
  $C' \leftarrow \{x \in P \mid \angle(\overrightarrow{x}, \overrightarrow{\mathrm{d}}) < R')\}$
end if
$E \leftarrow Excise(C', z_{gap})$
$\sigma_E \leftarrow$ standard deviation of redshifts in E
return $Excise(E, \sigma_E)$

Algorithm 4 function $Excise(A, gap)$

Input: $A \subset \mathbb{R}^3$ *, set of* (ra, dec, z) *points*
Input: $gap \in \mathbb{R}$
Output: A selection of points based on the redshift they represent: the set of successive points in the redshift space confined by a intervals of redshift with no galaxies. The minimum size of each interval (lower and upper) is *gap*.

$A' \leftarrow$ {the series of $a_{\mathrm{i}} \in A$, sorted by redshift value from lowest to highest.}
$a_K \leftarrow$ {the point of A' with redshift closest to the mean of all redshifts in A'.}
$a_j.z \leftarrow$ {the redshift of a_j} $\forall j$
$RESULT \leftarrow \{a_j\}$
$K \leftarrow$ *, Index of median z value of the points in* A'
for $\mathrm{i} = K$ to $n - 1$ **do**
  if $(a_{\mathrm{i}+1}.z - a_{\mathrm{i}}.z) < gap$ **then**
    $RESULT \leftarrow RESULT \cup a_{\mathrm{i}+1}$
  else
    break, *filter out upper interval*
  end if
end for
for $\mathrm{i} = K$ to 2 **do**
  if $(a_{\mathrm{i}}.z - a_{\mathrm{i}-1}.z) < gap$ **then**
    $RESULT \leftarrow RESULT \cup a_{\mathrm{i}-1}$
  else
    break, *filter out lower interval*
  end if
end for
return $RESULT$

2.2 The improved algorithm

2.2.1 Problem identification

A previous study of Vocludet [25] showed suboptimal results for the determination of the velocity dispersions of the clusters. This problem arises in the second stage of Vocludet, ZHG, and it is caused due to the addition of extra galaxies not belonging to the real clusters. The addition of extra galaxies, usually with velocity values which deviate significantly from the actual mean, causes an increase in the velocity dispersions of the detected clusters. Since velocity dispersions are further used to estimate cluster's masses, it is crucial to obtain the best results possible.

To correct the encountered errors and improve these results, the second stage of the algorithm is modified. Since the addition of extra galaxies is mostly caused by the fixed angle of search in the ZHG stage ($1R_A$), this approach is replaced with one where the angular size of the clusters is determined by their core galaxy distribution. The radius of the cluster core is then defined in terms of its radius R_{200}. The radius R_{200} is defined as the distance from the center up to where the mean interior density is 200 times the critical density of the universe, and is expected to contain the bulk of the cluster mass. The exact formula for the R_{200} radius is the following [26]:

$$R_{200} = \frac{2.02 \cdot \sigma_v}{1000 kms^{-1}} \frac{h_{70}^{-1} Mpc}{\sqrt{\Omega_\Lambda + \Omega_0 * (1+z)^3)}}$$

Where h_{70} is the normalized Hubble constant and Ω_Λ and Ω_0 are 0.7 and 0.3 respectively, for the standard cosmology, and σ_v is the velocity dispersion of the initial distribution of galaxies. The initial distribution of galaxies to use to calculate the R_{200} is taken from all the galaxies within $0.5h^{-}1Mpc$ ($\sim 0.3R_A$) in projection space.

The new algorithm is named GapperR200 and it is described in the next section.

2.2.2 GapperR200 Algorithm

The process starts by considering the centroids of the VT-MLE start-up structures, considered in order of increasing seed volume (decreasing Voronoi cell density). The complete set of steps is the following:

1. The first step in the algorithm attempts to correct possible deviations from the "true" cluster center as detected by the previous stage. A center adjustment process takes place: starting from the center provided by VT-MLE, all galaxies within $0.5h^{-1}Mpc$ in projected distance and separated by 1000 kms^{-1} gaps are selected. The (RA, DEC) centroid of these galaxies is then calculated and defined as the new center to be used. This step is repeated 3 times to avoid an excessive diversion.

2. Taking the new center (RA, Dec, z) provided by the previous step, a fixed redshift cut is applied, removing all objects more than 4500 kms^{-1} from the initial cluster redshift estimate.
3. All galaxies within $0.5h^{-1}Mpc$ and 4500 kms^{-1} of the cluster center are selected.
4. The selected galaxies are ordered by their velocity and any gap between them of width 1000 kms^{-1} or larger is identified. Galaxies separated by these gaps from the main system are excised. The resulting group of galaxies is called the initial core.
5. The velocity dispersion σ_v and the central redshift C_{BI} are calculated using the biweight estimator following [26]. Similarly to step 3, galaxies separated by gaps of width σ_v from the main system are excised.
6. The maximum and minimum redshifts of the ensemble of remaining galaxies are recorded (z_{min} and z_{max}). These values provide the fixed redshift limits for the remainder of the process.
7. The velocity dispersion and central redshift (σ_v, C_{BI}) are recalculated, along with the cluster radius R_{200}.
8. Considering all galaxies within R_{200} and between z_{min} and z_{max}, the velocity dispersion σ_v is recalculated. R_{200} and σ_v are iteratively recalculated until they stop changing. If a σ_v value is repeated, the algorithm would get into a loop. To avoid this, instead of taking the repeated value, we take the mean of the previously found σ_v values and continue as before.
9. If the number of resulting galaxies is less than two, when the initial core had at least 5 members, then the initial core is selected to be the final cluster. We call clusters found this way, type II clusters.
10. Finally, if the resulting cluster has no galaxies in common with a previously detected one, then it is recorded as a new valid cluster.

GapperR200 this produces two types of cluster detections: *type I*, where member galaxies are defined within a section of radius R_{200} and *type II*, where member galaxies are defined within the initial section of radius $0.5h^{-1}Mpc$. The GapperR200 galaxy cluster member identification algorithm is also presented as Algorithm 5, which uses the functions defined in Algorithms 7 and 6.

One downside of the new approach is the increase in the importance of the initial cluster position accuracy. The new initial angular radius is roughly a third of the previous value, so whereas with ZHG a cluster could easily be recovered even if the initial position had an offset of $0.5R_A$ from the actual cluster center, the new algorithm would be unable to recover it. To counteract this negative aspect, an adjustment step is introduced in the algorithm (step 1 or algorithm 6), which aims to correct slight errors produced in the VT-MLE stage by adjusting the cluster position, moving it closer to the overall highest density.

Considering N to be the number of galaxies in the catalog each step of the GapperR200 algorithm is either $O(N)$ or $O(1)$, with the exception of the sorting stage which is $O(NlogN)$. The iteration process could in theory happen as much as N times, with an initial group containing one galaxy and each iteration adding a new one. The final complexity of the algorithm is therefore $O(N^2)$. However, in practice, the number of iterations is very low, with the clusters quickly converging to a stable value. Since the first stage of Vocludet,

Algorithm 5 $Gapper R_{200}(\mathrm{d}, P, A)$

Input: $\overrightarrow{\mathrm{d_i}}$ *, Vector pointing from origin to the center of a cluster* C*.*
Input: set $P \subset \mathbb{R}^3$ where each point represents a galaxy of the survey.
Input: A *, set of galaxies that already belong to another cluster.*
Output: a set of points , where each point represents the position of a galaxy identified as member of a cluster.

$\mathrm{d}_f \leftarrow Recenter(\mathrm{d_i})$
$C' \leftarrow \{x \in P \mid abs(\mathrm{d}_f.z - x.z) < 4500 \ \wedge \ \angle(x, \mathrm{d}_f) < 0.5h^{-1}Mpc\}$
$CORE \leftarrow Excise(C', 1000)$
$\sigma_{CORE} \leftarrow Dispersion(CORE)$
$CORE' \leftarrow Excise(CORE, \sigma_{CORE})$
$\sigma_{CORE'} \leftarrow Dispersion(CORE')$
$z_{min}, z_{max} \leftarrow CORE'$ *, redshift limits*
$R_{200} \leftarrow R_{200}(\sigma_{CORE'})$
$RES \leftarrow \{x \in P \mid x.z \in \{z_{min}, z_{max}\} \ \& \ \angle(x, \mathrm{d}_f) < R_{200}$
$SIGMA \leftarrow Dispersion(RES)$
while TRUE **do**
  $R'_{200} \leftarrow R_{200}(SIGMA)$
  $RES' \leftarrow \{x \in P \mid x.z \in [z_{min}, z_{max}] \ \& \ \angle(x, \mathrm{d}_f) < R'_{200}\}$
  $SIGMA' \leftarrow Dispersion(RES)$
  if $SIGMA = SIGMA' \ \wedge \ R_{200} = R'_{200}$ **then**
    break;
  end if
end while

Algorithm 6 $Recenter(\mathrm{d}, P)$

Input: $\overrightarrow{\mathrm{d_i}}$ *, Vector pointing from origin to the center of a cluster* C*.*
Input: set $P \subset \mathbb{R}^3$ where each point represents a galaxy of the survey.
Output: Corrected cluster center

$\mathrm{i} \leftarrow 0$
repeat
  $C' \leftarrow \{x \in P \mid abs(\mathrm{d}.z - x.z) < 4500 \ \& \ \angle(x, \mathrm{d}) < 0.5h^{-1}Mpc\}$
  $CORE \leftarrow Excise(C', 1000)$
  $\mathrm{d} \leftarrow Centroid(CORE)$
  $\mathrm{i} \leftarrow \mathrm{i} + 1$
until $\mathrm{i} = 3$
return d

Algorithm 7 function $Excise(A, gap)$

Input: $A \subset \mathbb{R}^3$, set of (ra, dec, z) points
Input: $gap \in \mathbb{R}$
Output: A selection of points based on the redshift they represent: the set of sucessive points in the redshift space confined by a void of *gap*.

$A' \leftarrow$ {the series of $a_\mathrm{i} \in A$, sorted by redshift value from lowest to highest.}
$a_K \leftarrow$ {the point of A' with redshift closest to the mean of all redshifts in A'.}
$a_j.z \leftarrow$ {the redshift of a_j} $\forall j$
$RESULT \leftarrow \{a_j\}$
for i $= K$ to $n - 1$ **do**
  if $(a_{\mathrm{i}+1}.z - a_\mathrm{i}.z) < gap$ **then**
    $RESULT \leftarrow RESULT \cup a_{\mathrm{i}+1}$
  else
    break;
  end if
end for
for i $= K$ to 2 **do**
  if $(a_\mathrm{i}.z - a_{\mathrm{i}-1}.z) < gap$ **then**
    $RESULT \leftarrow RESULT \cup a_{\mathrm{i}-1}$
  else
    break;
  end if
end for
return $RESULT$

VT-MLE, is also $O(N^2)$, the complexity of the complete algorithm is $O(N^2)$. In terms of memory, the space complexity of GapperR200 is N, since it only stores the full galaxy catalog and indices for each cluster.

In chapter 5.2 a detailed analysis of the algorithm is described.

Chapter 3

Analysis and Visualization software

3.1 Analysis

Part of the analysis of the algorithm started with the previous work by the same author of this thesis, in which the overall performance of the cluster detector was calculated. This section describes the extension of the previous work in terms of the analysing tools and the software used to create a variety of plots.

3.1.1 Results module and graphing library

Besides generating the final output of the algorithm, Vocludet includes modules which produce intermediate and derived results. The original software produced results for each stage of the algorithm. This was further extended to the GapperR200 stage, where new relevant information is created. This includes information about the close and distant environment of the clusters. After a final cluster is produced by GapperR200, two additional clusters are created: both use the previous version of this stage (ZHG) with a radius of 1 and 3 R_{Abell}, but limited to the radial size determined by GapperR200. The first of these two clusters is then called the extended cluster and represents an attempt at recovering the whole cluster instead of being limited to R_{200}, whereas the second one is only used for creating plots showing the vicinity of the structures. Furthermore, basic properties of the GapperR200 stage are computed and included in summary tables, such as the type of each cluster, velocity dispersions, angular size, etc.

3.1.2 Graphing library

The generation of plots and figures is streamlined through the use of R scripts. R is a programming language and software environment for statistical computing which also supports multiple plotting and data visualization libraries. Among them is ggplot2, which is an im-

plementation of Leland Wilkinson's Grammar of Graphics [27] —a general scheme for data visualization which breaks up graphs into semantic components such as scales and layers. The use of the grammar of graphics approach allows the creation of highly modifiable plots, which can also be partly reused to generate further figures.

The library is composed of multiple R scripts, with most of them taking charge of a particular subset of plots based on the research interest, such as characterizing not detected clusters or describing the clusters obtained in each stage of the algorithm. Examples of these plots can be found throughout this document. Furthermore, the library contains scripts which generate summarized tables of results necessary to publish the outcome of this research.

3.2 Visualization Software

The visualization software developed has as its main objective to assist in the analysis of the resulting clusters of the Vocludet algorithm. The application uses web browser technology WebGL to display 3D graphics of the detected cluster's galaxies and complementing information. WebGL (Web Graphics Library) is a JavaScript API for rendering interactive 3D computer graphics and 2D graphics within any compatible web browser without the use of plug-ins. WebGL is integrated completely into all the web standards of the browser allowing GPU accelerated usage of physics, image processing and effects as part of the web page. WebGL elements can be mixed with other HTML elements and composed with other parts of the page or page background, which allows extensive functionality ideal for a visualization software.

The use of a web browser as the medium through which the visualization is displayed offers advantages such as the no need of installing additional software and the ubiquity of the application with the use of a web server. A web server allows the centralization of information and equalization of the analysis conditions which are frequent in collaborations among astronomers and other scientists, such as international research teams.

3.2.1 Software architecture

The software is divided into 2 layers: the backend and the frontend. The backend implements the API to access the algorithm results. The frontend utilises the data provided by the backend API and displays it accordingly. Both backend and frontend are implemented using Node.js which is an open-source, cross-platform runtime environment for developing server-side Web applications, using JavaScript. A simplified version of the overall architecture, including the interaction with the output of the main program is shown in figure 3.2.

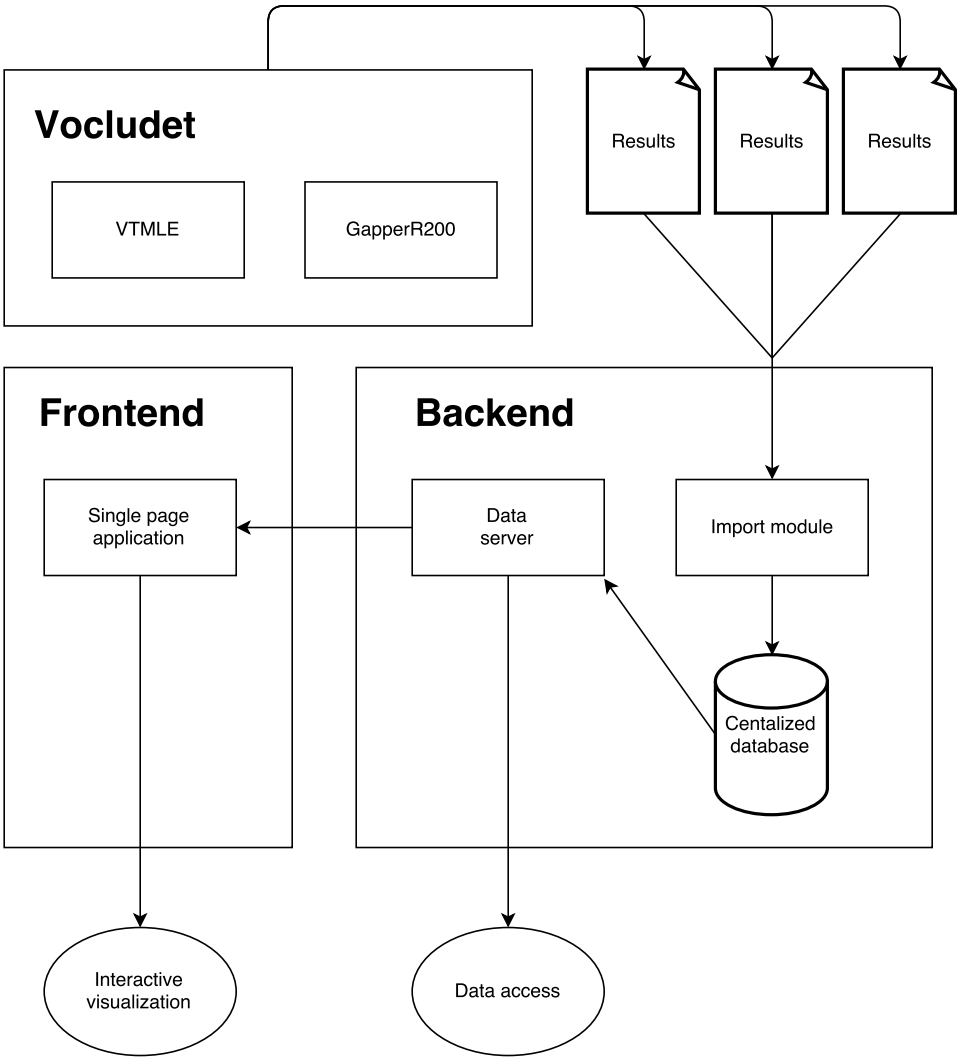

Figure 3.1 Visualization software overall architecture diagram

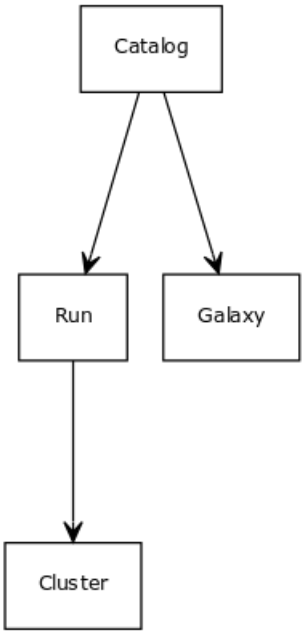

Figure 3.2 Visualization software backend data structure

3.2.2 Backend

The backend consists primarily of a server and the multiple routes that form the API. The data structure is as shown in figure 3.2. Each catalog comprises a list of Galaxies and a multiple Runs. Each Galaxy element contains information about a single galaxy within the catalog, such as the position and redshift. The Runs are individual executions of the algorithm using a corresponding list of Galaxies, usually with different parameters and/or implementations of the algorithm. Keeping track of each Run allows the user to compare different executions of the algorithm and detect possible problems or determine the best set of parameters. Within each run is the list of resulting clusters. Finally, each cluster contains information about the position, members, multiplicity and ID. The information is stored in MongoDB, a cross-platform document-oriented database.

3.2.3 Server

The server is the backbone of the visualization tool. It allows users to connect to the application by delivering the code which is then executed in the users' browser. It also implements the API to communicate with the database containing the algorithm information.

The different routes implemented by the server are the following:

/catalog
: Returns a list of available (loaded) catalogs.

/catalog/:catalogId
: Returns a summary of the selected catalog, with information such as total number of

galaxies, source and limits.

/catalog/:catalogId/galaxy
Return a list of positions of the catalog's galaxies.

/catalog/:catalogId/galaxy/:galaxyId
Returns information about a particular galaxy, such as position, magnitude and redshift.

/catalog/:catalogId/run
Returns a list of available runs.

/catalog/:catalogId/run/:runId
Returns information about a particular run, such as the parameters used.

/catalog/:catalogId/run/:runId/cluster
Returns the list of cluster positions resulting from the particular run.

/catalog/:catalogId/run/:runId/cluster/:clusterId
Returns information about a particular clusters, such as the list of galaxies and velocity dispersion.

3.2.4 Frontend

The frontend is implemented as a single page application, that is, a web app that loads a single HTML page and dynamically updates that page as the user interacts with it. The application is divided into three regions: the main display, the control panel and the additional information panel. It communicates with the server via the API, so the user is able to select which run of the algorithm to display and within that run select individual clusters.

Main display

This is the region where the clusters of galaxies are actually displayed. The user can rotate the scene in a 3D sphere around the currently selected cluster by clicking and dragging. This allows the user to observe the clusters in different angles, while being able to maintain the focus on the object of interest.

Galaxies belonging to the current cluster are shown as bright spheres with its brightness corresponding to the galaxy magnitude. This is accomplished by overlaying a texture simulating a light flare (see figure 3.7) and modifying its opacity to control how bright it appears. Since it is important to study the surroundings of each cluster, neighboring galaxies are also displayed, though as smaller and opaque spheres to allow the focus to be on the cluster's galaxies.

Additionally, two cones are displayed. An inner red cone indicating the size of the selected

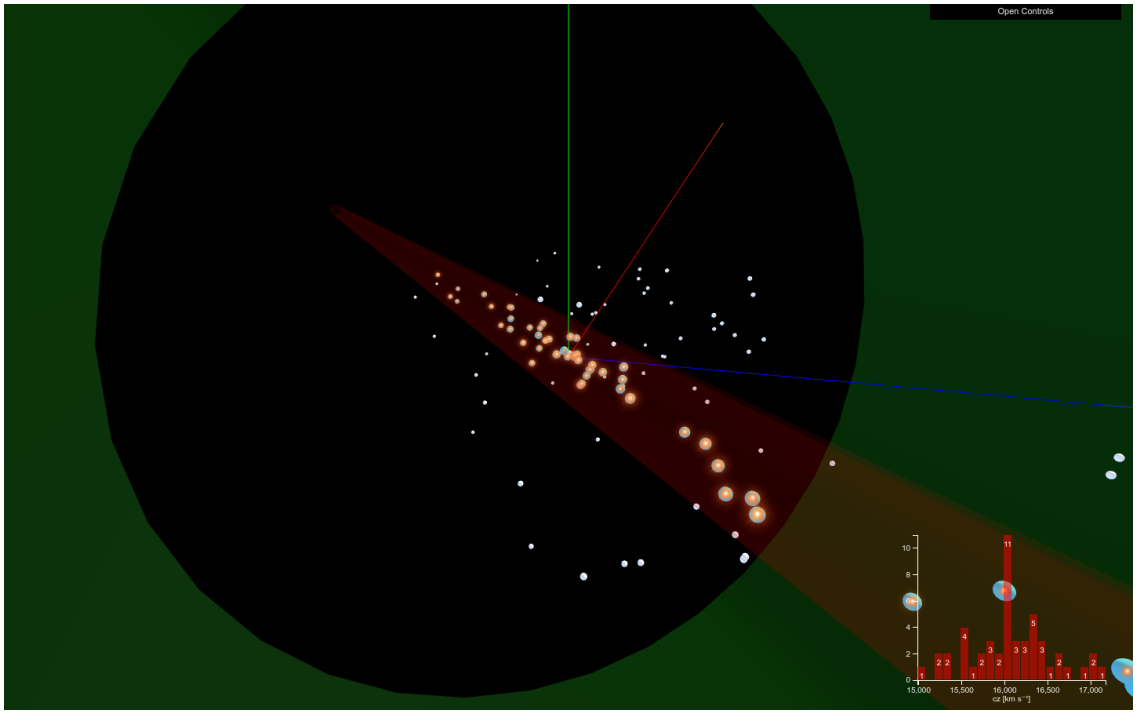

Figure 3.3 Visualization software screen capture. It shows the view from inside the 1 Mpc radius cone where a combination of radial and angular distribution can be seen.

cluster and an outer green one with a fixed size of 1 Mpc. These two objects are used to get a sense of scale without the need to add a complex spherical grid system. They also help with the orientation, since the apex of the cones point to the origin of the cartesian system, which in the case of a real catalog would be the Earth. Figure 3.3 shows the view from inside the green outer cone, with black background color. The clusters appear elongated due to the *finger of god effect*, which is a redshift-space distortion caused by local radial velocities acting as a Doppler effect. Examples of other functionalities can be seen in figures 3.4, 3.5 and 3.6.

Control panel

The control panel allows the user to control various aspects of the visualization. They are implemented as a series of slider controllers and a list selector to choose which cluster to display. The controls are the following:

Inner cone radius
: Controls the size of the inner cone. When a cluster is initially selected to be displayed, this value corresponds to its angular radius.

Environment radius
: Controls the size of a virtual (not displayed) cone inside which the environment galaxies are rendered. Environment galaxies outside this virtual cone are not shown to avoid cluttering the view. Initially has a value of 1 Mpc.

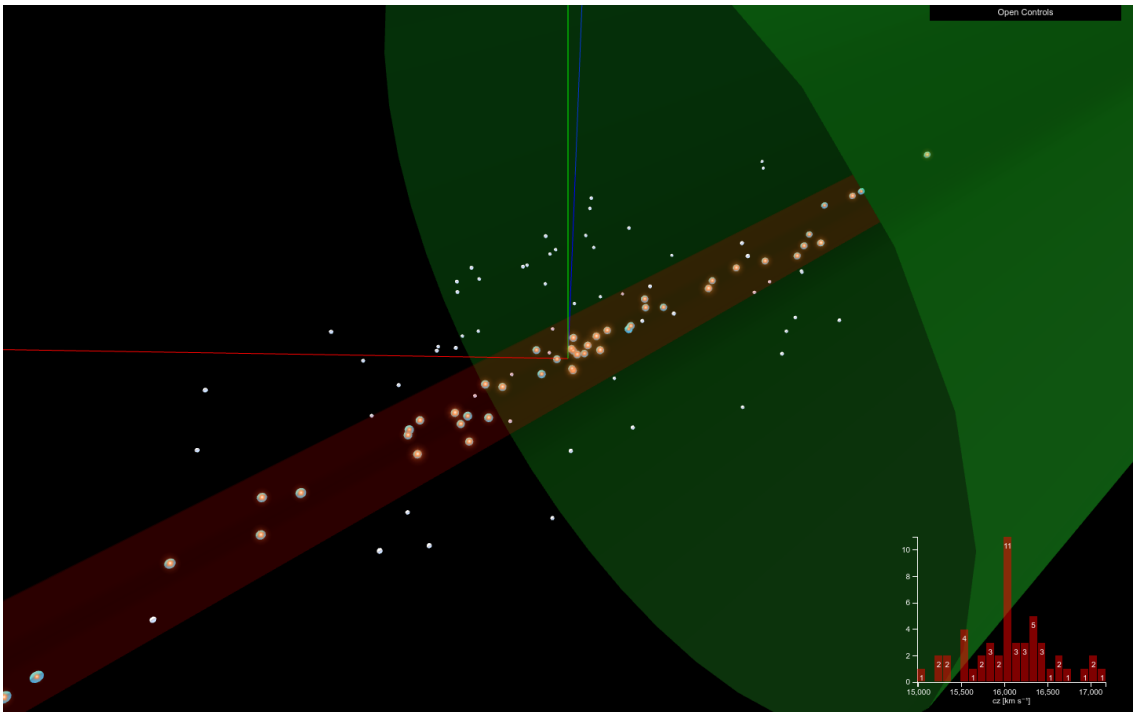

Figure 3.4 Visualization software screen capture. It shows the view from outside the 1 Mpc radius cone where the radial distribution is better displayed.

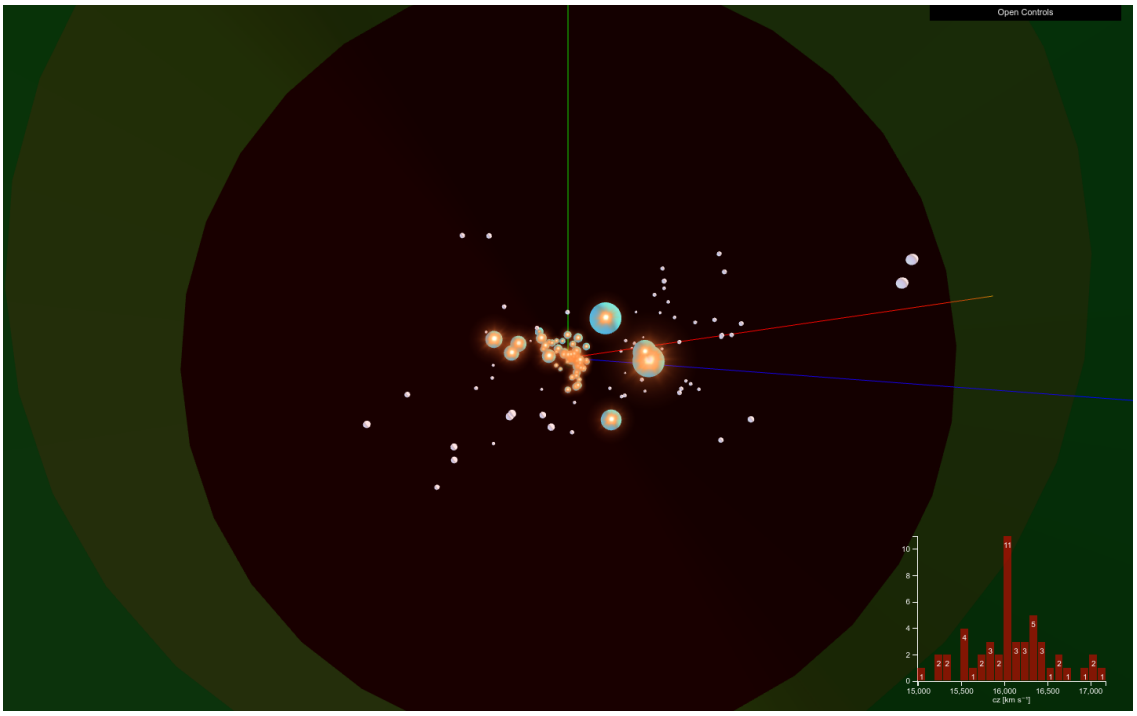

Figure 3.5 Visualization software screen capture. It shows the view from the center of the cluster cones where the angular distribution is best appreciated.

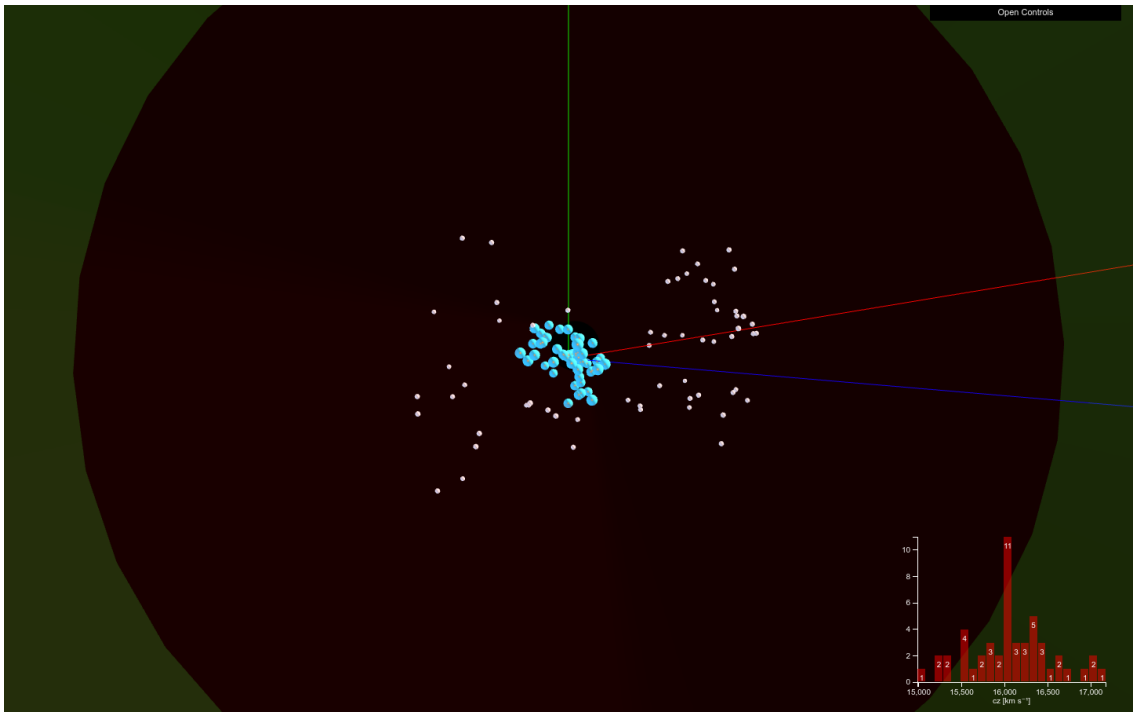

Figure 3.6 Visualization software screen capture. It shows the view from outside the 1 Mpc radius cone using a higher field of view than figure 3.5, which helps separate the cluster from the nearby galaxies.

Figure 3.7 Lens flare texture used in the visualization software

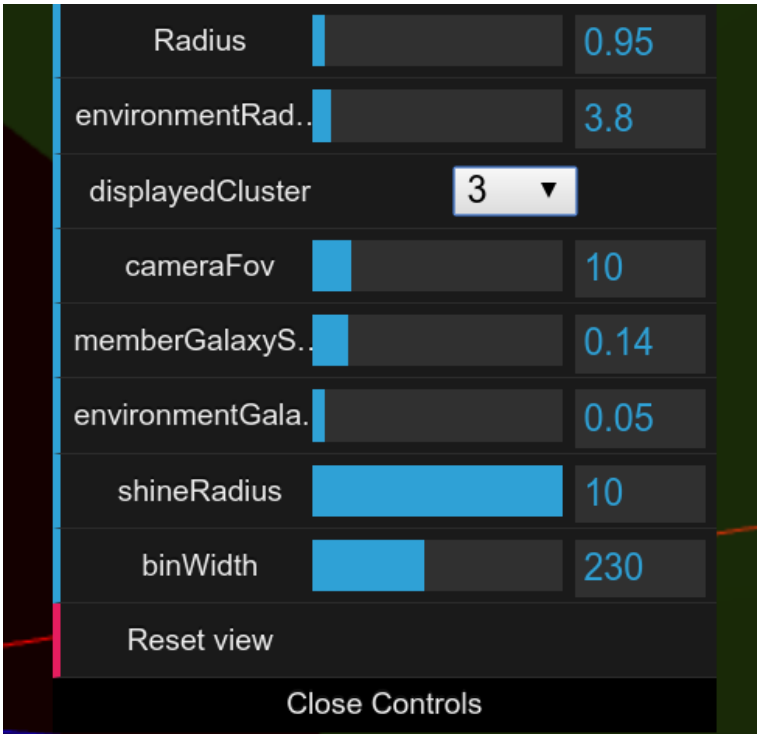

Figure 3.8 Visualization software control panel

Displayed cluster
Allows the selection of the cluster display. Initially set to the first cluster in the catalog.

Camera field of view
Controls the FOV, which is the extent of the observable scene that is seen at any given moment. Its initial value is 60 deg.

Member galaxy size
Controls the size of each member galaxy. Initially 0.1 Mpc.

Environment galaxy size
Controls the size of each environment galaxy. Initially 0.05 Mpc.

Shine radius
Controls the size of the shine halo around each member galaxy. Initially 2 times the member galaxy size.

Bin width (Additional information panel)
Returns information about a particular clusters, such as the list of galaxies and velocity dispersion.

A screen capture of the control panel can be seen in figure

Additional information panel

In this panel, the velocity distribution of the cluster's galaxies is displayed. This is shown as a histogram with variable bin size in the lower right corner of the application. This is information is essential in identifying false positive clusters, since real clusters usually present Gaussian distribution. However, lack of unimodality does not necessarily mean that a cluster is a false positive, because these objects might contain substructures which disturb the smooth velocity distribution, making it appear multimodal. Figures 3.9 and 3.10 show a sample distribution, using two different bin sizes. In the histogram with narrower bin width, a hint of the presence of substructures with peaks at 16,000 kms^{-1} and 16,300 kms^{-1} can be seen, which is not visible using the wider bin sizes.

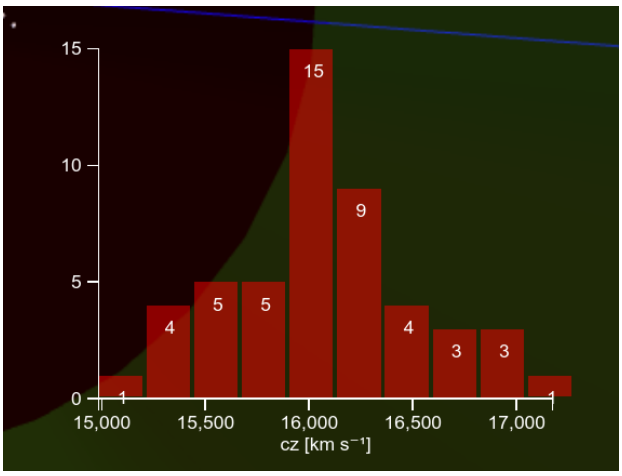

Figure 3.9 Sample histogram using bin width $\sim 50kms^{-1}$

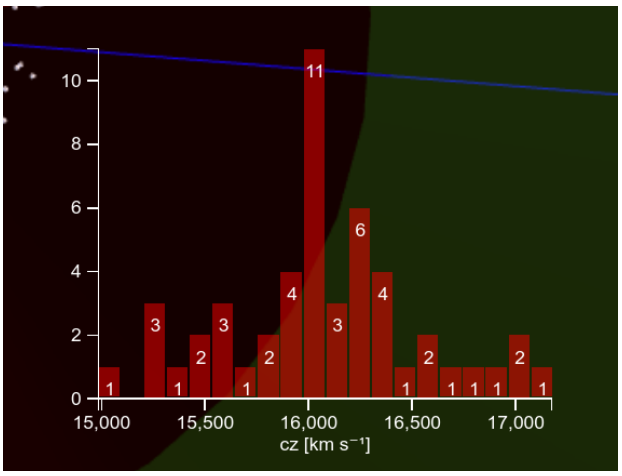

Figure 3.10 Sample histogram using bin width $\sim 100kms^{-1}$

3.2.5 Potential uses of the tool

The first and most natural use of the application is the analysis of the results of the algorithm. The possibility of managing multiple runs, each one of those using different parameters, allows the easy comparison of the effects they have over the resulting structures. With the ability to rotate the structures in 3 dimensions comes the possibility of identifying galaxy outliers in both the radial and angular coordinates.

Besides the analysis of results, since the visualization is designed to work on any modern browser, it is also embeddable on any web page. This facilitates the communication of the results to the rest of the scientific community because besides just sharing static images, there is the possibility of presenting the results in an interactive manner. The same principle applies to the data accessibility. Instead of providing a single static file containing all the results from a particular run (although this is also possible), the user is able to query just the data he or she is interested on, such as a particular cluster mentioned in a publication.

Chapter 4

Optimization

4.1 Mock galaxy data

Data from the Millennium Simulation of the LCM cosmology is then used to produce a mock database for use in the optimization and validation process of the algorithm.

The subset of galaxies used in this paper was taken from the all-sky mock catalog of [28], which is limited at an apparent AB magnitude of 18 in the r-band filter from SDSS and they include apparent magnitudes in the 8 optical filters from both SDSS and 2MASS [29]. A strip of sky is extracted from the first catalog (table Blaizot2006_AllSky_RT_1 in the simulations) covering 1350 deg^2, and spanning 90 deg $\times$ 15 deg. The size of the strip is defined so it should be precisely set to match the Southern strip of the 2dF galaxy redshift survey, which will be used at a later stage to test real data. A wedge diagram that shows the spatial distribution of galaxies is shown in figure 4.1. A lower limit of $z = 0.009$ is adopted for the selected dataset since at smaller redshifts most galaxies would belong to the Local Supercluster. Additionally, an upper limit of $z \sim 0.3$ is set for the galaxies in the sample.

Conceptually, a galaxy cluster consists of a gravitationally bound system. Following the approach as [30], a cluster in the Millennium Simulation is defined as a collection of galaxies that belong to the same parent halo, with at least 5 members and where membership is defined by examining the Millennium Simulation catalogs. These clusters may contain a virialized population of galaxies, as well as galaxies bouncing in and out, and others falling in for the first time. Thus, the masses and radii of the clusters are obtained. The resulting mass distribution for the Reference Cluster Catalog is shown in figure 4.2 whose corresponding median is $2.220 \cdot 10^{13}$ solar masses.

The final dataset contains $156,494$ galaxies and 1850 clusters, hereafter, the Reference Galaxy Dataset and the Reference Cluster Catalog respectively. A histogram of the masses of the reference catalog clusters is shown in figure 4.2 where it can be seen that the mass range represented is 10^{12} to 10^{15} solar masses.

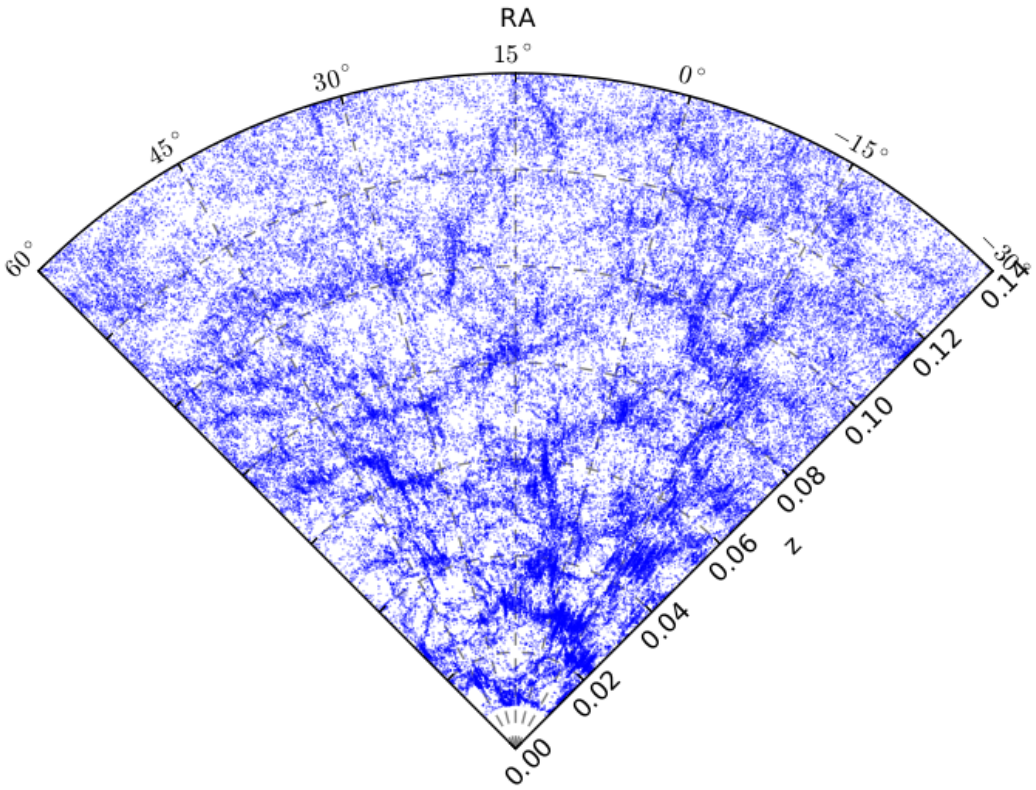

Figure 4.1 Wedge diagram showing the mock 2dF database, restricted to $z < 0.14$, based on the Millennium Simulation

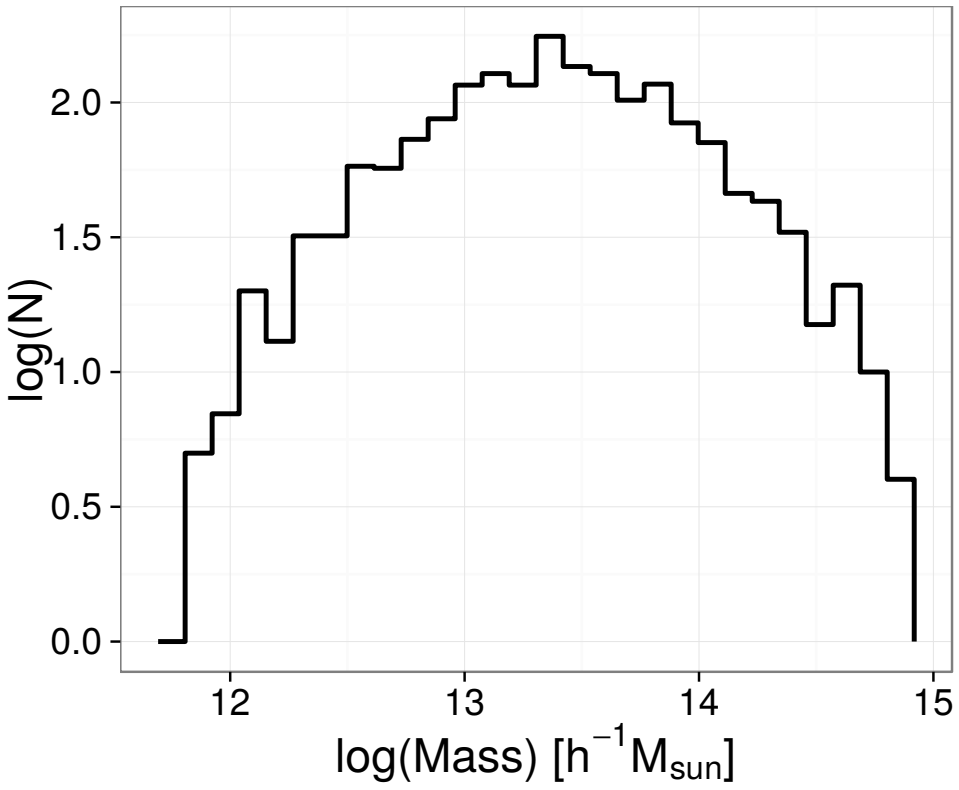

Figure 4.2 Reference catalog cluster mass distribution. N indicates the number of clusters per bin.

4.2 Parameter optimization

Before proceeding with the algorithm validation, an optimization of the algorithm takes place to ensure that the free parameters are adequate to the particular dataset to be used. The validation and optimization are done using the mock dataset previously described. The effectiveness of a cluster finder is expressed through the completeness and purity of the output catalog in comparison with the reference cluster catalog. In our case, the completeness is defined as the fraction of clusters in the reference catalog that has a counterpart in the Vocludet catalog, and, purity as the fraction of Vocludet clusters that have a correspondence on the reference catalog.

As mentioned in section 1, Vocludet has three parameters that affect the output of the algorithm. The domain size used to calculate the local density for each seed in the VT-MLE stage is originally set to be 50 h_{70}^{-1} Mpc. This chosen radius is large enough to average over the neighboring large scale structure features and small enough to be sensitive to extended trends, but it is still subject to optimization.

The maximum allowed cluster diameter (2 R_A) was found to have little impact on the results of the algorithm, since most seeds stop growing because of the MLE condition, and do not reach the maximum allowed cluster size.

The final free parameter is the velocity gap in the GapperR200 stage. The value of this parameter will affect the output in multiple ways. Since the cross-correlation between Vocludet clusters and the reference catalog requires the adoption of a minimum percentage of galaxies in common, the velocity gap value influences the recovery rates and purity. Besides that, it is desirable to obtain values of velocity dispersion as close to the reference catalog as possible.

4.3 Domain size

The first stage of the algorithm sets the maximum recovery rate for the whole process. A cluster which is not close enough to a seed cannot be later discovered by GapperR200. For this reason the focus of VTMLE's optimization is on the recovery rate, rather than the purity.

The initial value chosen for the domain size is $\rho_{Radius} = 50Mpc$, and is justified by the known large scale distribution properties. To pick the optimal size, several values close to the original domain size are selected. For each of this values VTMLE is run and the recovery rates are calculated in terms of the amounts of galaxies in the reference clusters. A reference cluster is considered a valid detection when a VTMLE seed is situated at an angular distance no greater than $0.5R_{Abell}$ and a redshift difference of no more than 0.003. The results are shown in figure 4.3.

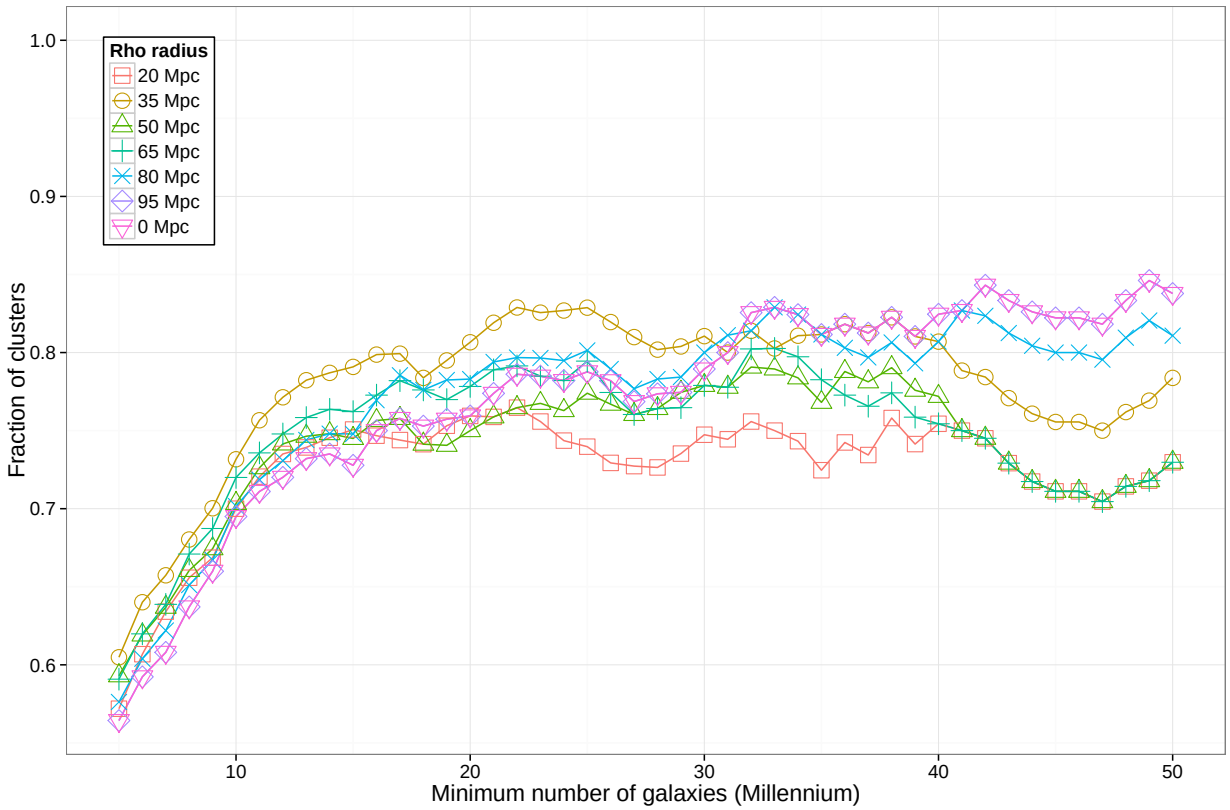

Figure 4.3 VTMLE recovery rate by domain size. It can be seen that for clusters up to 30 galaxies, the best results are obtained with a domain size of $35Mpc$

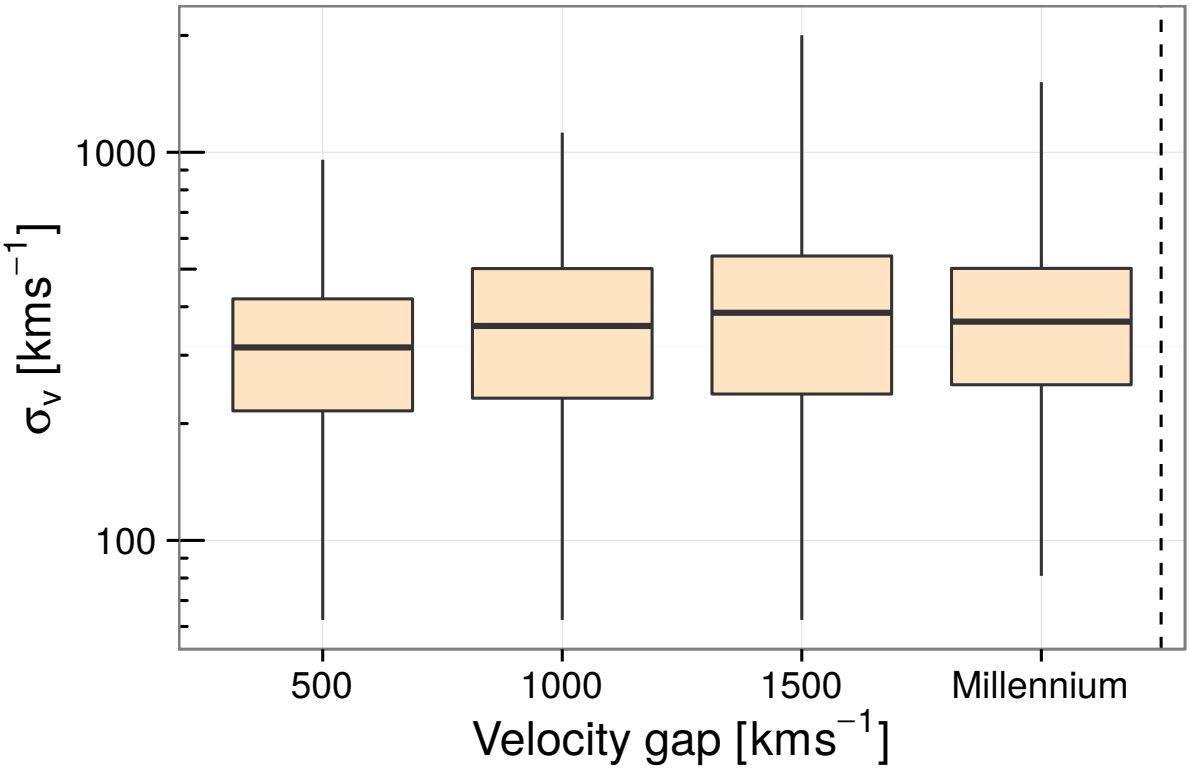

Figure 4.4 Velocity dispersion box plot of multiple values of velocity gap, including the Millennium reference catalog. The bottom and top of the boxes represent the first and third quartiles, and the band inside the boxes indicates the second quartile (the median). The bottom and top ends of the vertical lines indicate the minimum and maximum values, respectively.

4.4 Velocity gap

To find the optimum velocity gap value, a series of runs of the algorithm are executed, narrowing the range of values according to the similarity of the output with the reference catalog. Figure 4.4 shows a box plot graph as an initial analysis of the impact of the velocity gap. It can be seen that the values do not vary significantly. The distribution with a velocity gap of $1000 kms^{-1}$ is the closest to the Millennium dataset, by a small margin. The velocity dispersions for all clusters are calculated using the biweight estimator.

Regarding the recovery rate, the results are dependent on the multiplicity of the clusters, as is shown in figure 4.5. For lower multiplicity clusters (with a number of galaxies $Ng < 15$), the velocity gap of $500 kms^{-1}$ produces lower recovery rates than the other two. For higher multiplicities, the results are the opposite. It is important to note that clusters with more than 30 galaxies are much fewer than clusters with $Ng < 15$.

The purity rates show a behavior similar to the recovery rates, as can be seen in figure

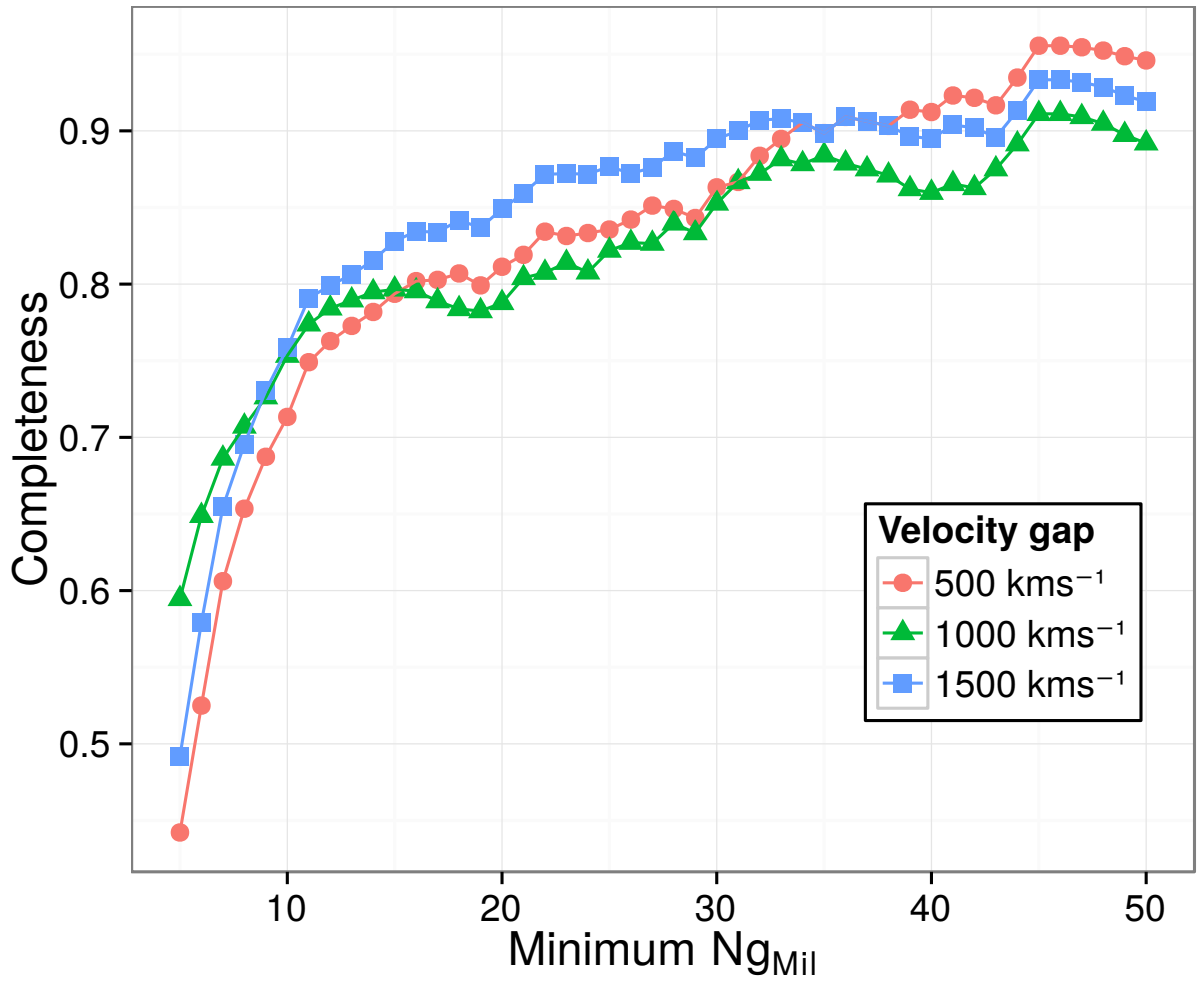

Figure 4.5 Vocludet completeness rate (fraction of reference clusters recovered) by minimum number of galaxies in the Millennium cluster (Ng_{mil}). To consider a match, the galaxy overlap must be of at least 25%.

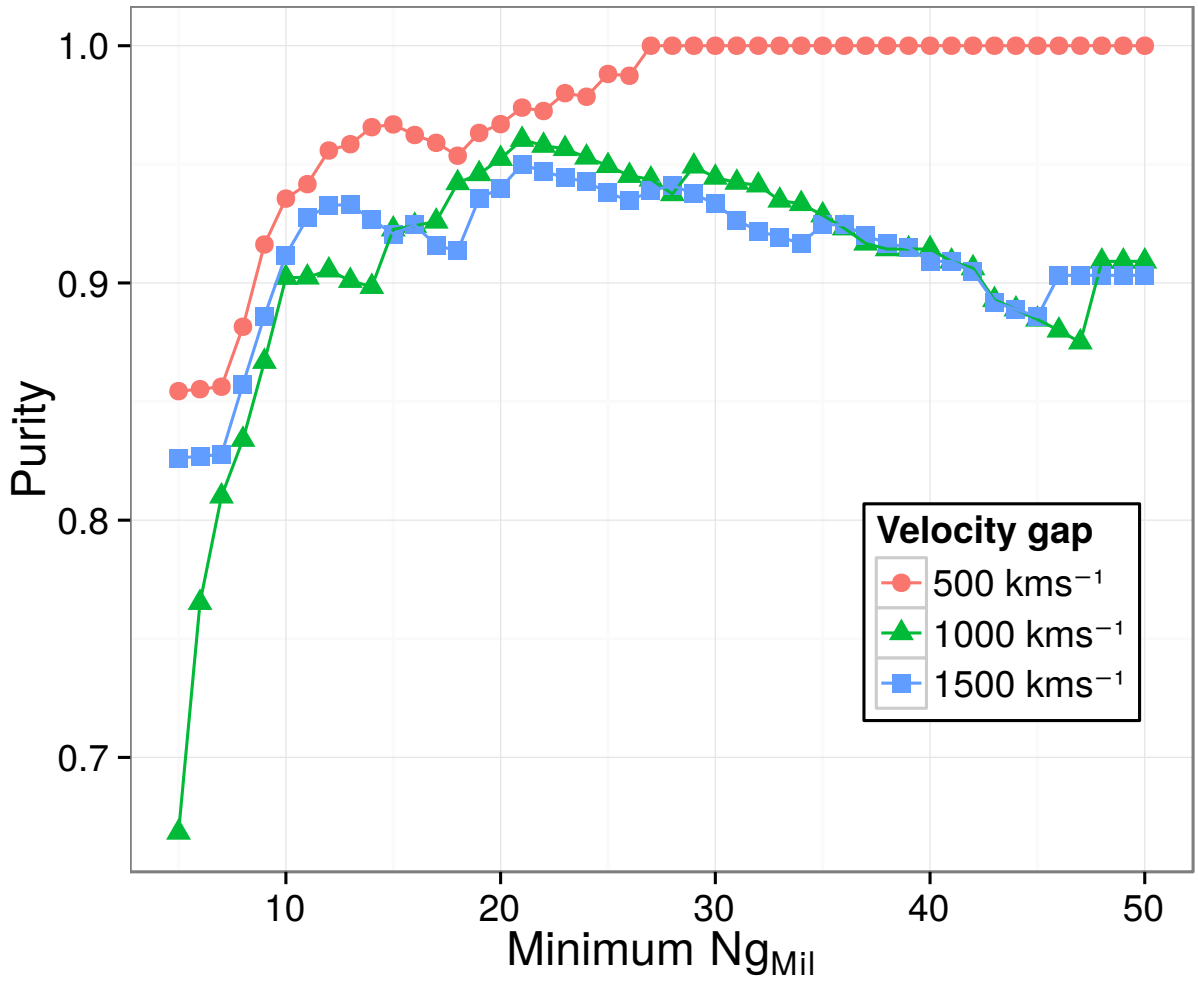

Figure 4.6 Vocludet purity rate (fraction of detected clusters which have a match in the reference catalog) by minimum number of galaxies in the Millennium cluster. To consider a match, the galaxy overlap must be of at least 25%.

4.6. The $500 kms^{-1}$ value is considerably better for clusters with $Ng > 25$, for which the purity rate reaches $> 95\%$.

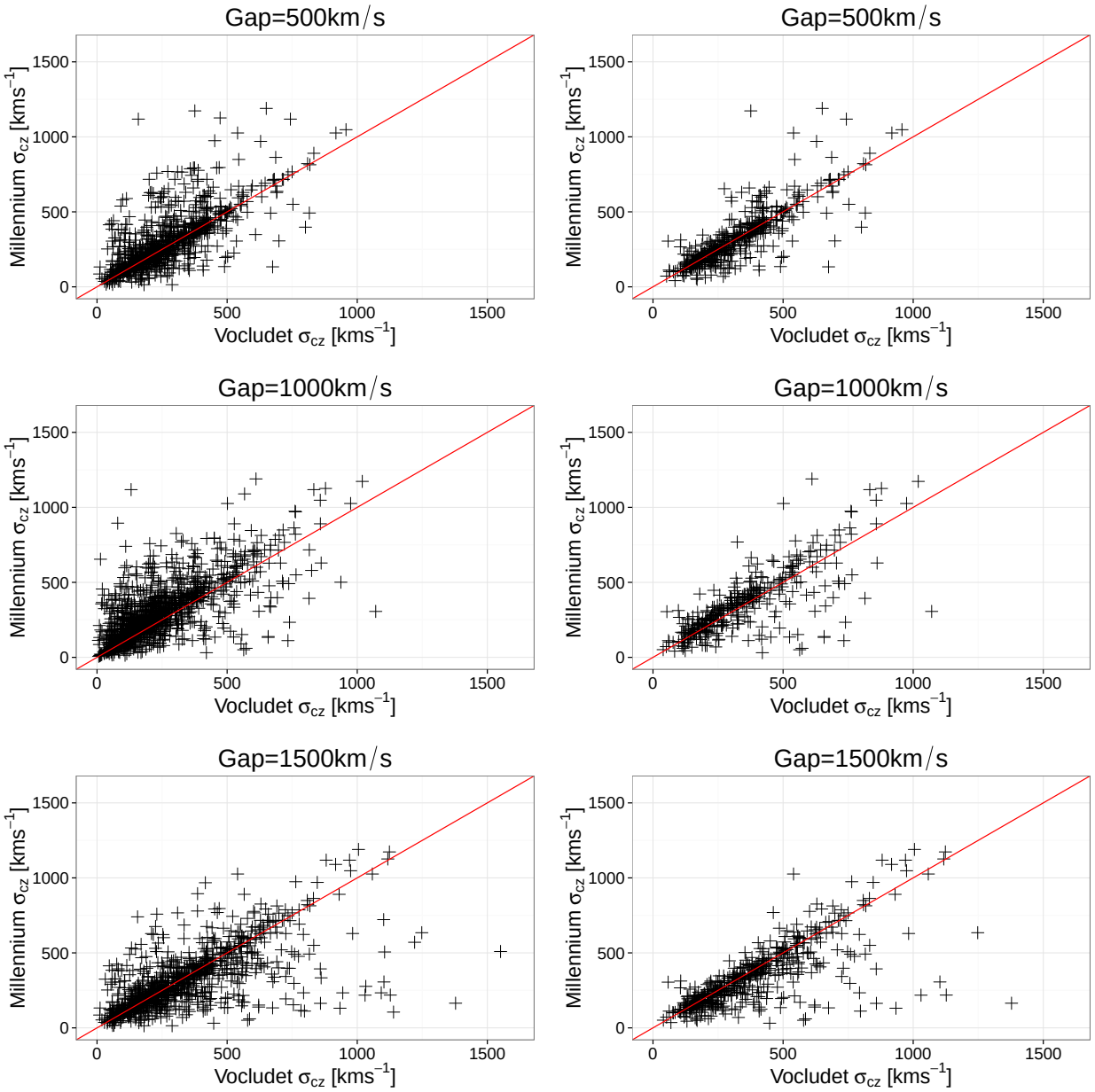

Figure 4.7 Vocludet vs Millennium reference catalog velocity dispersion for different values of velocity gap. The red lines are lines of slope 1 passing through the origin. Figures to the left include all valid vocludet clusters while figure to the right include only the ones with 10 or more galaxies.

Figures in panel 4.7 display scatter plots of the values of velocity dispersion for each pair of valid Vocludet-Millennium cluster, i.e., for pairs of clusters that share at least 25% of their galaxies. As the value for the velocity gap gets higher, the slope of the linear regression decreases. All values of velocity gap stay close to the line of slope 1, but as this value increases, a few outliers start appearing in the region of overestimated velocity dispersion. Finally, table 4.1 displays the summary of the slope values and standard errors for each

Table 4.1. Summary of the Velocity-Gap optimization results

Gap (km/s)	Slope ($Ng \geq 2$)	Std. Error ($Ng \geq 2$)	Slope ($Ng \geq 10$)	Std. Error ($Ng \geq 10$)
500	1.0735	0.0155	1.0233	0.0170
1000	0.9065	0.0140	0.9065	0.0140
1500	0.8033	0.0165	0.8033	0.0165

velocity gap value.

To choose the final value of the velocity gap, all previous analysis are taken into consideration, though giving a greater weight to the recovery rate. Even though the algorithm is fairly stable to the different values of velocity gap, the overall recovery rate is maximized with the value $1500 kms^{-1}$, with $1000 kms^{-1}$ following closely. However, since the $1500 kms^{-1}$ introduces a high number of outliers in the velocity dispersion comparison, it is discarded.

A velocity gap of 500 kms^{-1}, when applied to the mock 2dFGRS, produces a cluster set marginally better than the one corresponding to 1000 kms^{-1}. Given this marginal difference, the 1000 kms^{-1} velocity gap value is chosen, due to this value being shown by DePropris et al. [24] to be optimal for the application of the ZHG technique to a real catalog. It is also important to keep in mind the final goal of the development of the cluster detector which is to allow for the discovery of new structures in the universe, thus making very high purity rates rather counterproductive since they would limit the detection to already known structures.

Chapter 5

Analysis and validation

5.1 Analysis

In this section the VTMLE and GapperR200 stages are analyzed separately. The approach taken is to compare the results generated by each stage to the reference cluster catalog mentioned in section 4.1. The validation of the complete cluster detector will be discussed in section 5.2.

5.1.1 VTMLE: the geometrical step

There are multiple ways to determine if a cluster has a counterpart in the reference catalog. One approach involves counting the number of galaxies in common between two clusters, which we will refer as the intersection. Another strategy is to measure the distance between two clusters; the closer the centroids are, the higher the probability that they correspond to the same structure.

The role of the VTMLE stage in the Vocludet algorithm is to find extended high density regions corresponding to real clusters, so that the second stage (GapperR200) can determine confidently the clusters. Therefore, the percentage of intersecting galaxies in the VTMLE stage is not the most relevant aspect. However, the distance between centroids is crucial. In particular, the angular distance of the clusters is important, since GapperR200 first employs an angular radius of $\sim 0.33 R_{Abell}$.

Figure 5.1 shows a histogram of the angular distance of each one of the Millennium clusters to their closest and intersecting VTMLE detection in terms of the Abell radius. A few intersecting clusters at considerable distances are found (over 1.5 Abell radii). These are rare cases in which the two clusters have a galaxy in common, yet their centroid are considerably distant.

A VTMLE detection is considered a recovery when the centroid of candidate cluster is close

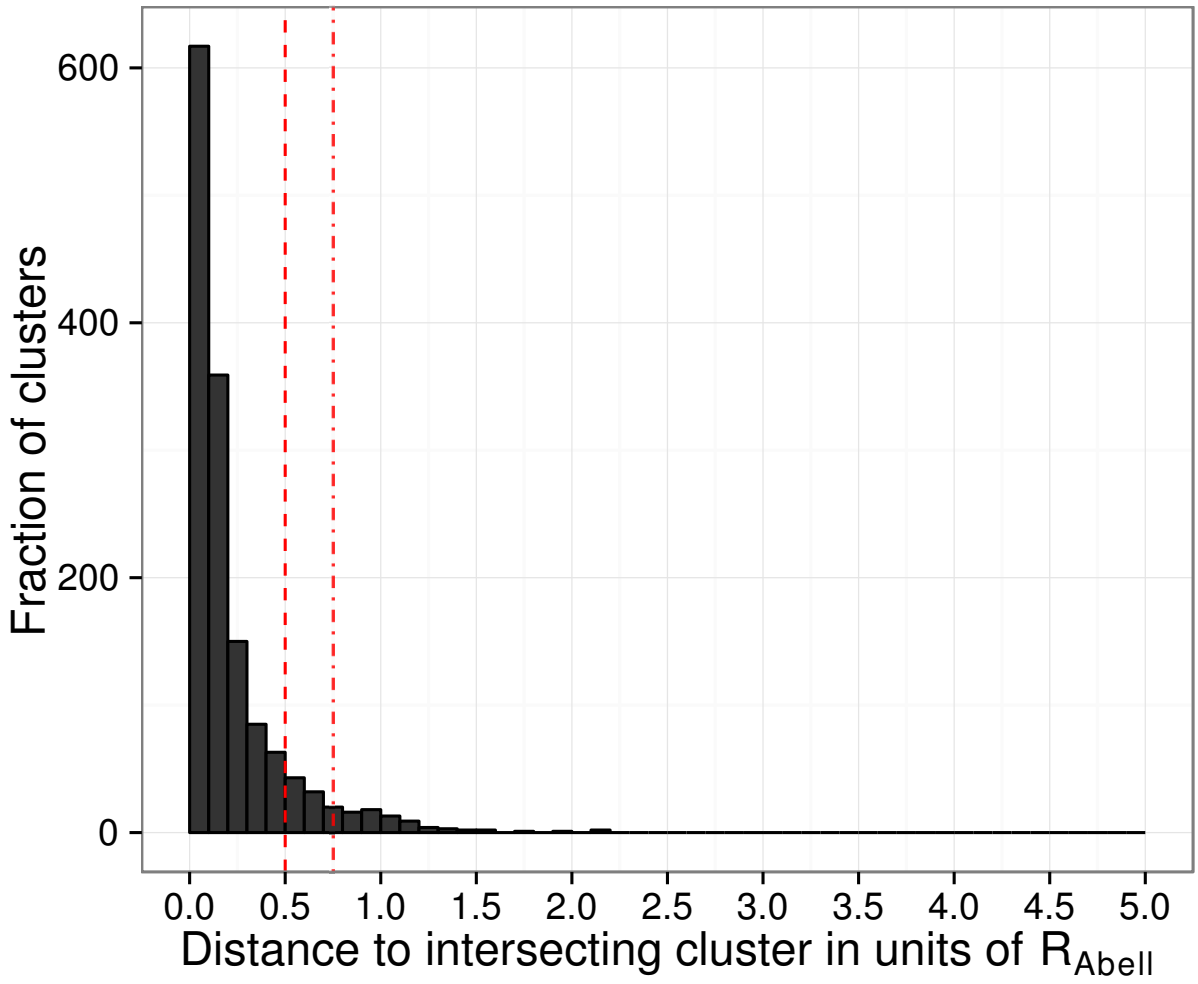

Figure 5.1 VTMLE cluster distance to closest Millennium cluster

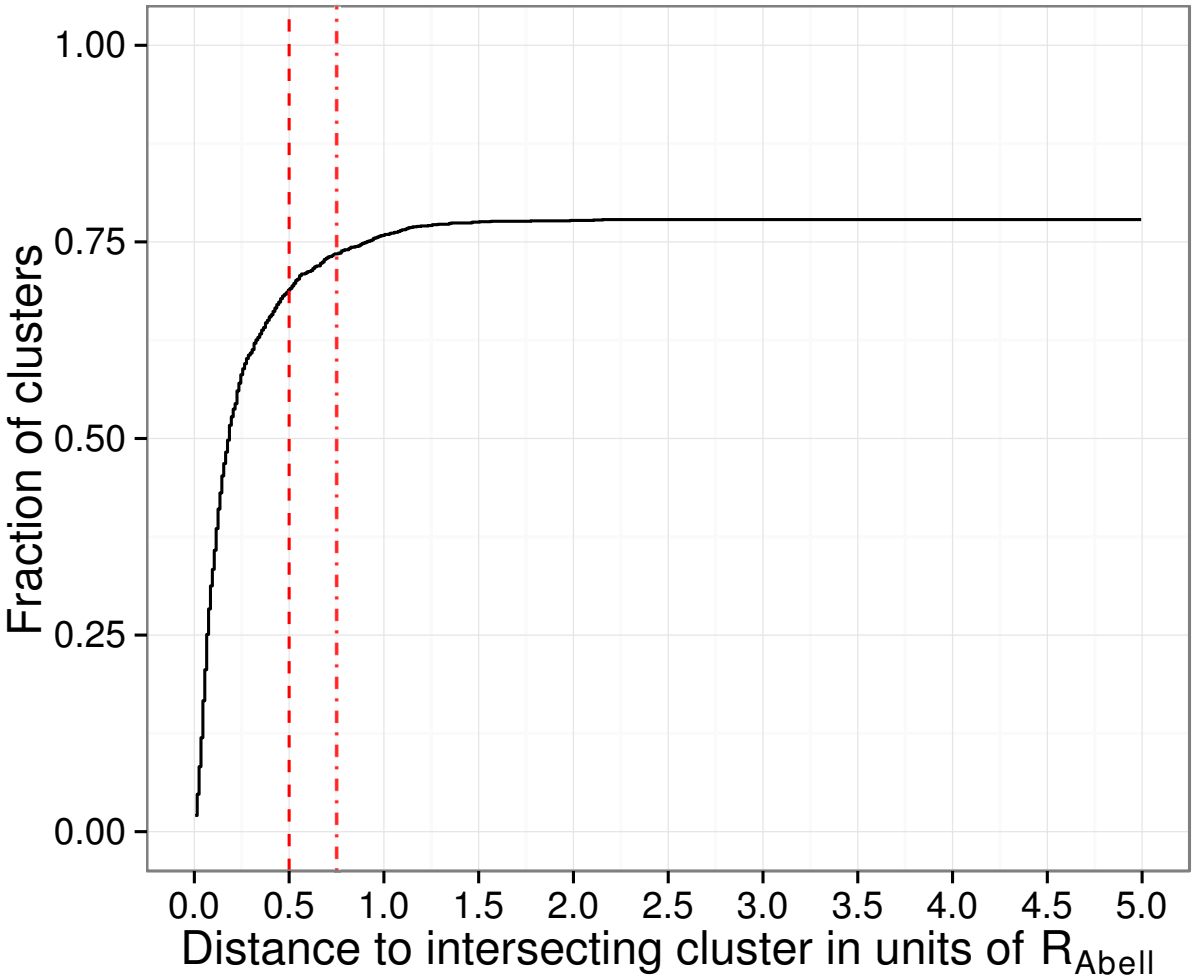

Figure 5.2 Cumulative distribution function of the VTMLE intersecting clusters projected distance, with respect to Millennium reference clusters. Red vertical lines indicate values 0.5 and 0.75, respectively

enough to the reference cluster in angular position and that there is at least some intersection between the two. The maximum angular distance to be allowed between the clusters is connected with the following stage of the algorithm. GapperR200 needs the centroid of the detected candidate cluster to be closer than 0.33 Abell radius to be able to detect a cluster without the recentering step. With the recentering step, the maximum distance considered acceptable is extended to 0.5, but it could be possible to recover farther clusters. Figure 5.2 shows how this distance is distributed among intersecting clusters. It can be seen that most intersecting clusters are no farther than 0.75 Abell radii and a significant fraction is closer than 0.5 Abell radii.

Recovery rate

The recovery rate corresponds to the fraction of the reference clusters detected by the algorithm. The detections are classified according to the angular distance θ to the closest reference cluster (in units of the Abell radius) as follows:

- $0 < \theta < 0.5$: Good
- $0.5 < \theta < 0.75$: Fair
- $0.75 < \theta < 1$: Unlikely
- $1 < \theta$: Not detected

The VTMLE algorithm has an upper limit in recovery of 79% for clusters closer than $1R_{Abell}$. Of these, 6% are of fair quality and 69% of good quality.

5.1.2 Number of galaxies

It is natural to expect greater recovery rates from clusters which are more notorious and denser. One of the features of the clusters that influence in its relevance is the number of galaxies it contains. The reference catalog contains clusters that range from 5 to 300 galaxies.

Figure 5.3 shows a clear relation between the number of galaxies in the reference cluster and the capacity of the algorithm to recover it. The recovery rate gets steadily better up until 15 galaxies, then a slight decrease is observed and then a new rise up until 30 galaxies. This behavior is consistent across the different qualities of the detections and it can be explained by fluctuations caused by the reduced number of clusters with a high number of galaxies. Figure 5.4 shows the amount of clusters detected along with the total number of clusters that meet the minimum number of galaxy members. In this picture the tendency is more clear and the fluctuations can be seen to scale.

5.1.3 Cluster Mass

As with the number of galaxies in the previous section, one could expect the recovery rates to be better for more massive clusters. However, this is not the case, as can be seen in figures 5.5 and 5.6. Figure 5.5 shows the recovery rates when a minimum mass restriction is established. In it the recovery rates appear mostly stable. The "Good" detections have a slightly decreasing slope. This behavior can be explained by the fact that clusters with higher masses tend to have higher velocity dispersions, which causes its members to appear more distant to each other. This affects directly the "Good" detection rate, since it requires the detected centroids to be very close the actual cluster center.

Figure 5.6 shows the recovery rates by mass value instead of minimum mass. In it, the fluctuations with respect to the mean recovery rates are more pronounced towards the higher mass end of the graph. This is because the number of high mass clusters is significantly lower than the low mass one.

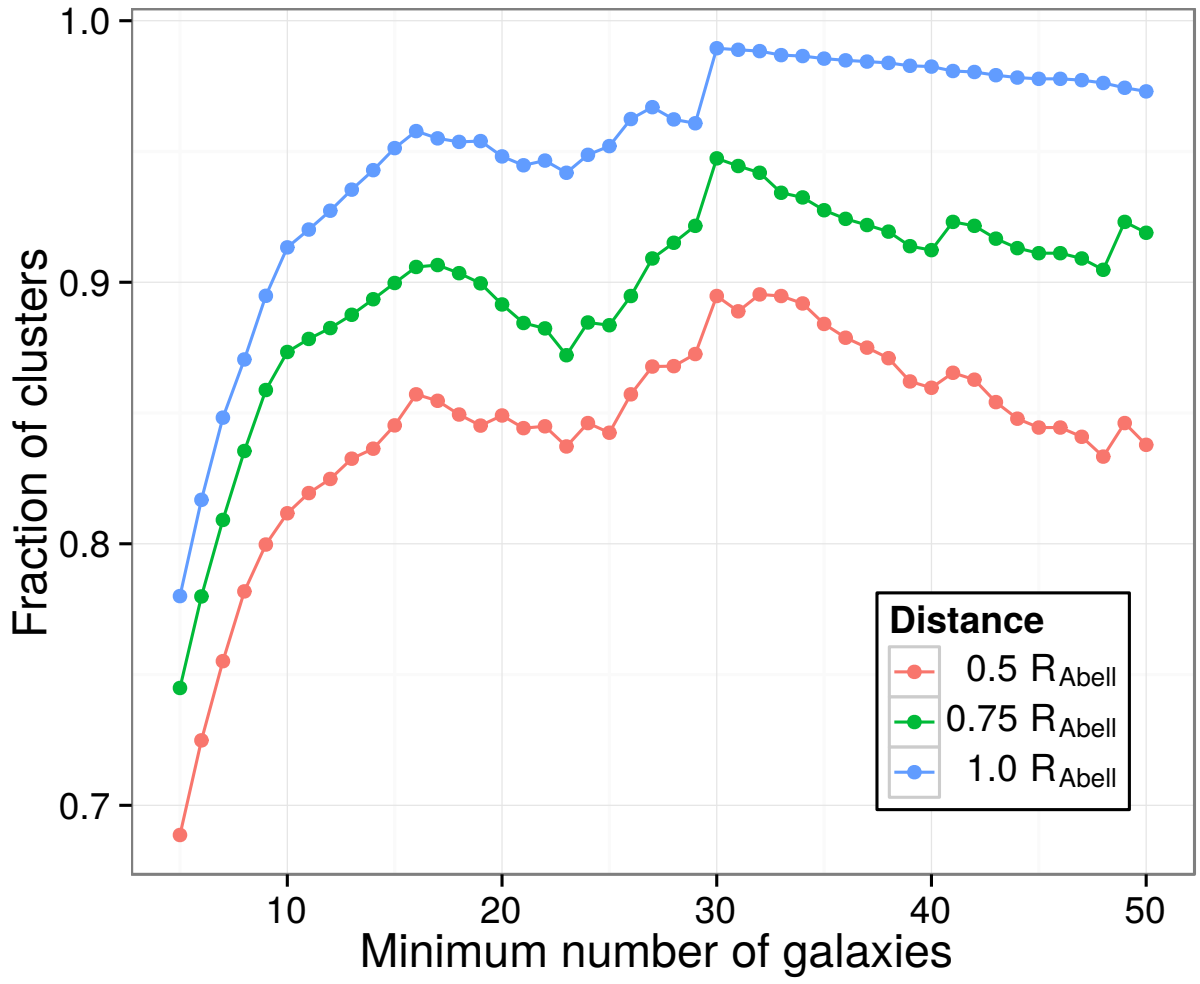

Figure 5.3 VTMLE recovery rates by number of galaxies in Millennium clusters

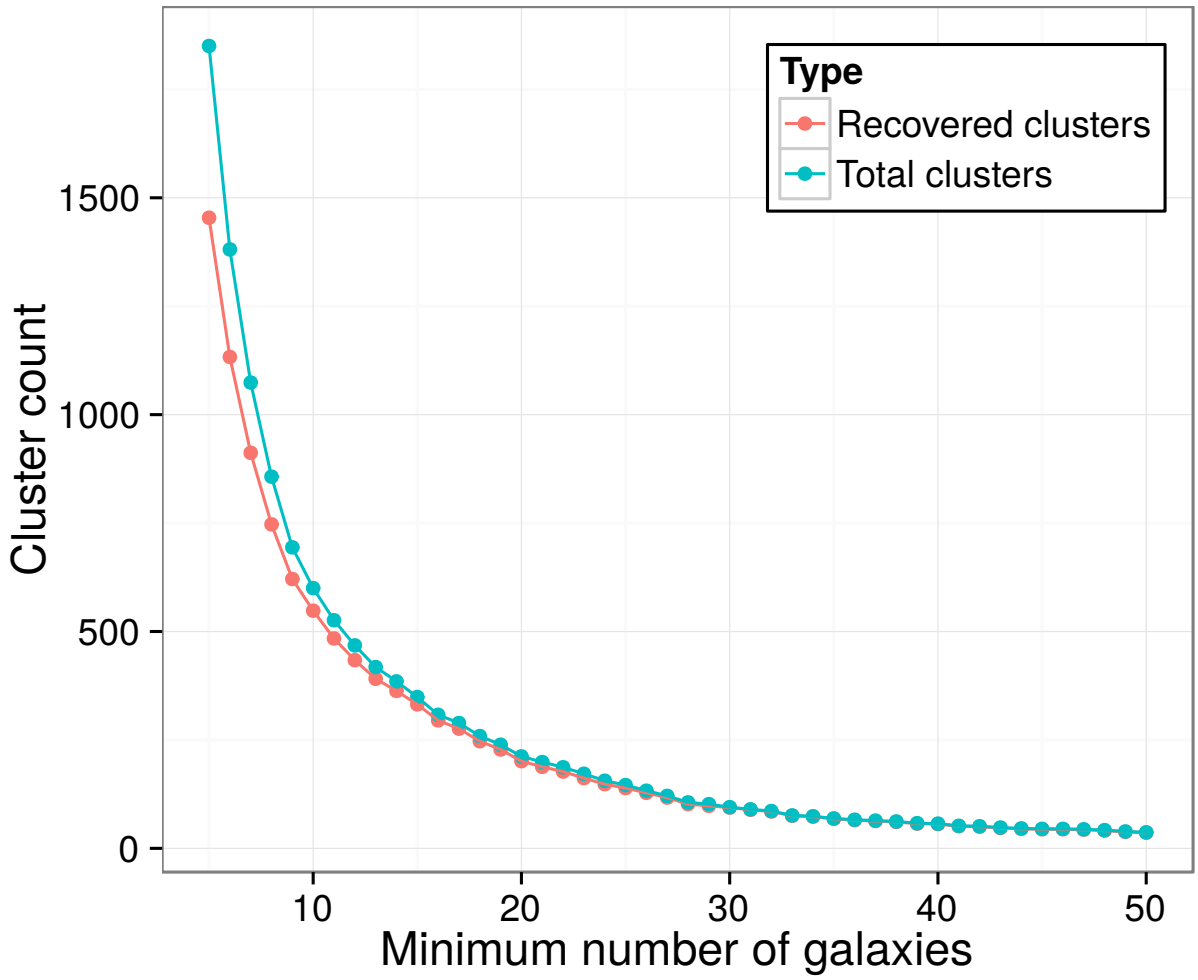

Figure 5.4 VTMLE clusters recovered by number of galaxies in Millennium clusters. Dashed lines represent the mean recovery rates of the respective detection quality

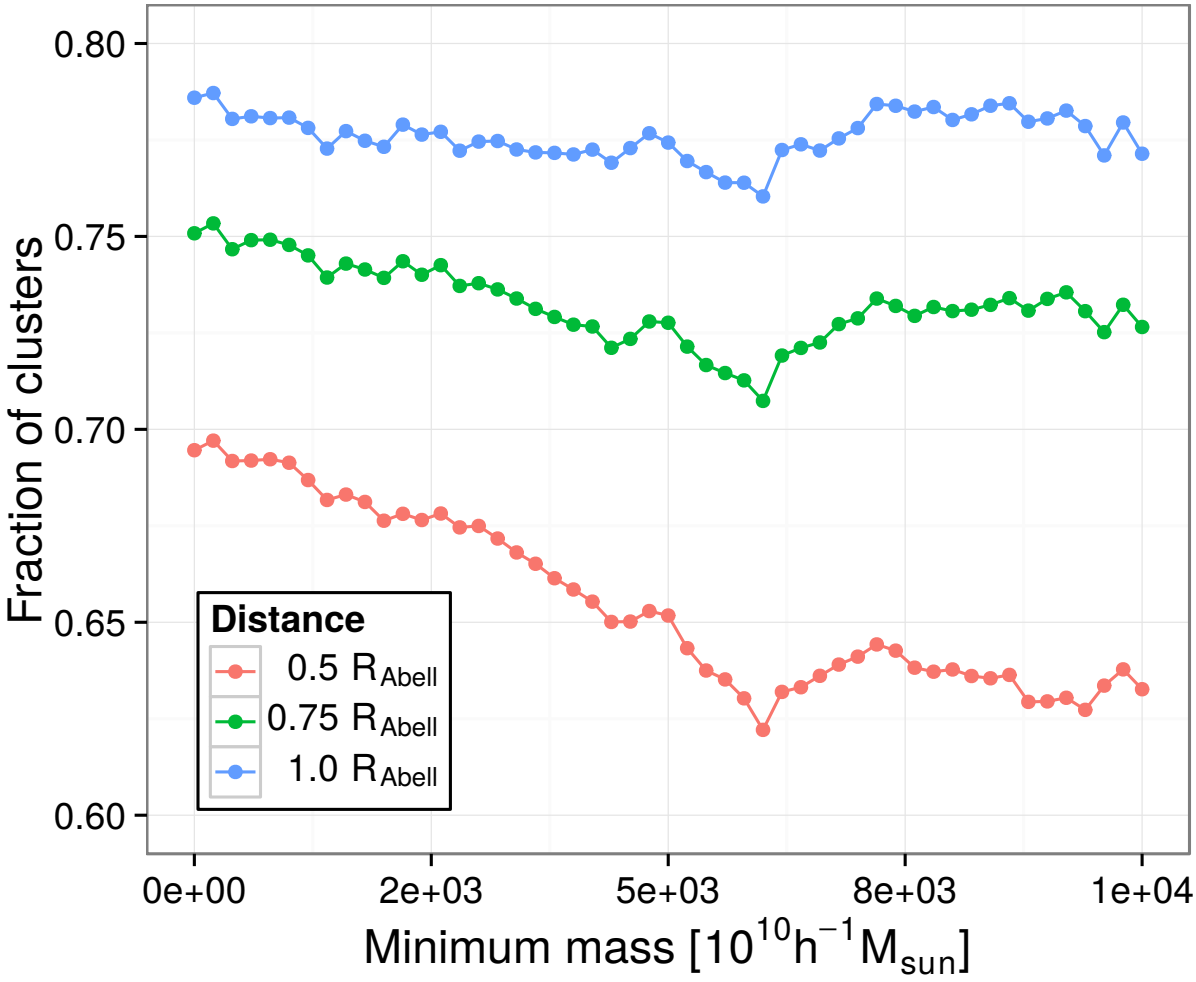

Figure 5.5 VTMLE recovery rates by minimum mass of Millennium clusters

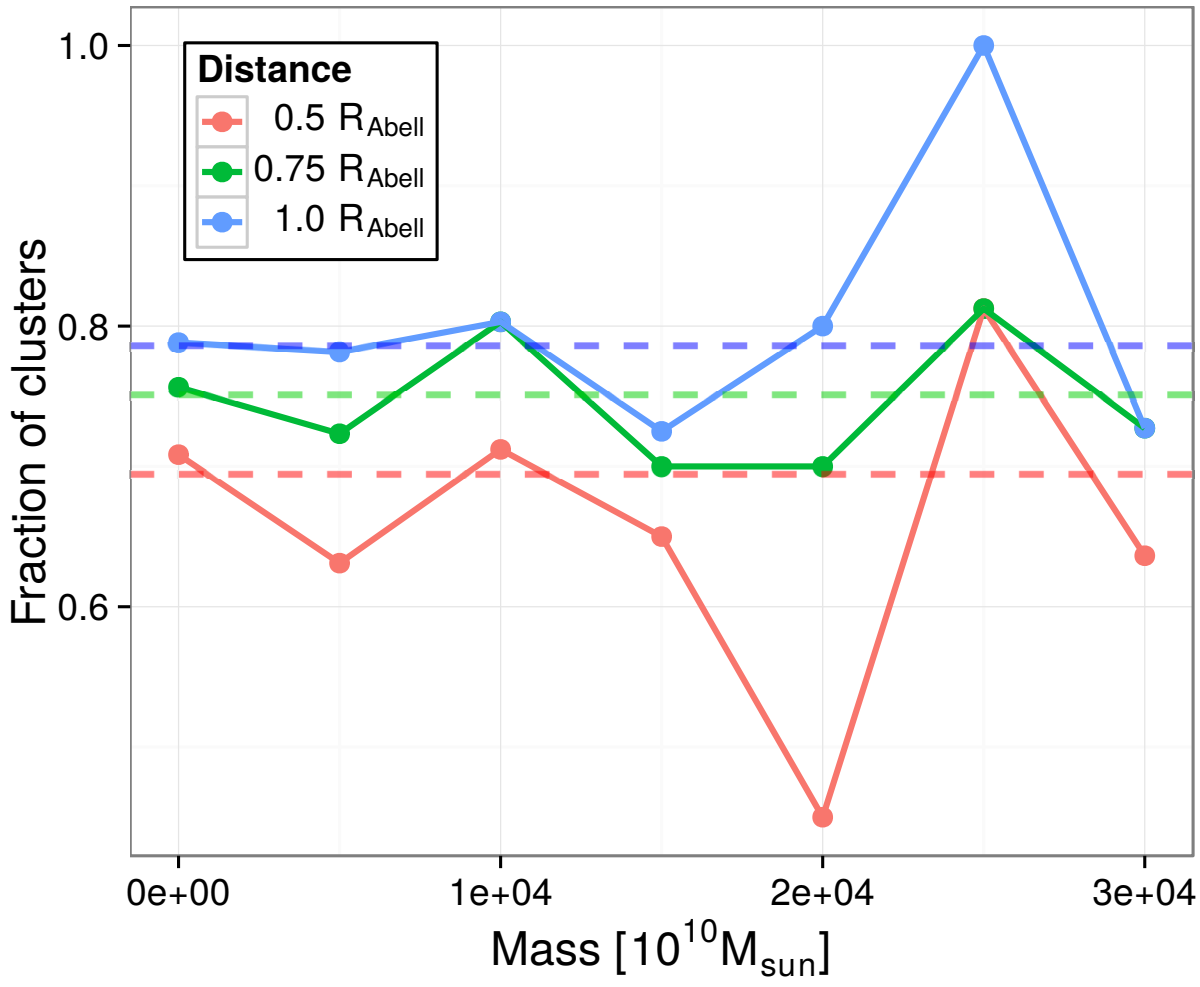

Figure 5.6 VTMLE recovery rates by mass value of Millennium clusters

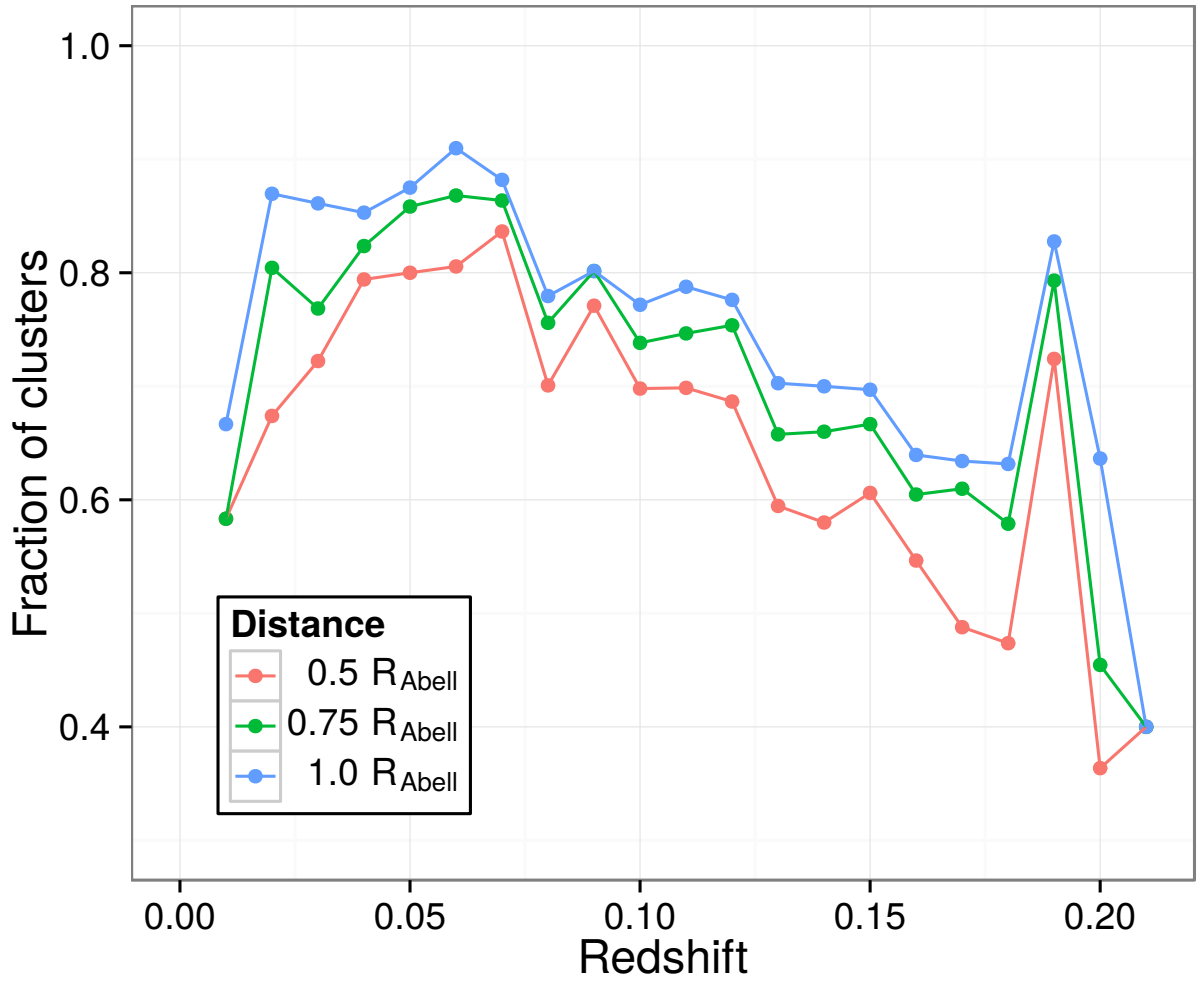

Figure 5.7 VTMLE recovery rates by redshift interval of Millennium clusters

5.1.4 Redshift dependency

One last property of the reference clusters is analysed to study its impact in the recovery rates: the redshift. The apparent redshift of an object depends on the distance of it to the observer, but it is also affected by the peculiar velocity of the object. The further an object is, the more its redshift is dominated by its distance, according to Hubble's law. This means that the redshift displayed by very close objects are primarily dominated by peculiar velocities and the distance determination will have a significant uncertainty. Because of this, the algorithm is expected to perform better at higher redshift. On the other hand, with higher redshift, the apparent brightness of galaxies decreases which means more galaxies are filtered out of the reference catalog because of the brightness limit. This suggests that the recovery rates should decrease with higher redshift.

Figure 5.7 shows the recovery rates by redshift interval. An increase in recovery rates can be seen up to redshift ~ 0.06, then a relatively constant and decrease up to $z = 0.2$. A notable

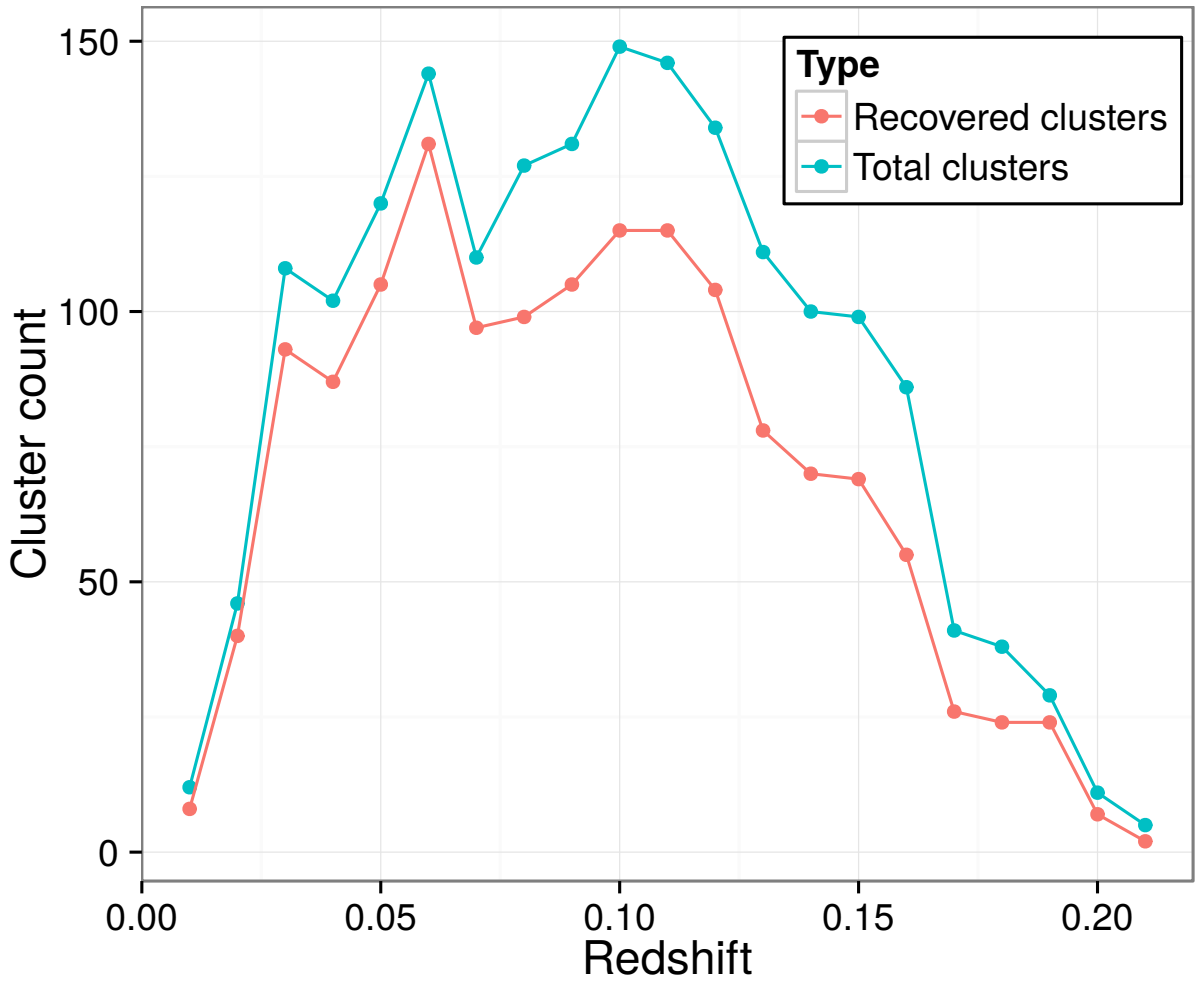

Figure 5.8 VTMLE clusters recovered by redshift interval of Millennium clusters

increase can be seen at redshift 0.19, this could be explained by a region of higher density like a filament or a wall, however figure 5.8 shows that the Millennium reference catalog does not present a higher number of clusters. This is explained instead, by the natural fluctuations in the recovery rates due to the lower number of clusters at this higher redshift values.

The detections are classified according to the angular distance θ to the closest reference cluster (in units of the Abell radius) as follows:

- $0 < \theta < 0.5$: Excellent
- $0.5 < \theta < 0.75$: Good
- $0.75 < \theta < 1$: Fair
- $1 < \theta$: Not detected

Table 5.1 summarizes some of the most important results regarding the recovery rates.

5.1.5 False positive rate

Another measure of the efficiency of the algorithm is the false positive rate. This corresponds to the percentage of clusters detected by the algorithm that do not have a equivalent in the reference catalog.

VTMLE detects a total of 6720 clusters. Out of those, and using the same criteria for detections defined in the previous section, 1600 have a reference cluster counterpart. This means that the overall false positive rate is 74%.

The algorithm generates its output as a series of clusters ordered by the density of the Voronoi cells that originated them. This provides an indication of the quality of the detection, since clusters higher in the ranking have a higher probability of being real ones, and the opposite for clusters with lower priority. Figure 5.9 shows the false positive rate of VTMLE detections filtered by maximum seed priority. A relatively stable region can be seen up to around seed 500, then a consistent increase in the false positive rates, although with a decreasingly positive slope.

Redshift error

When the algorithm is executed with real galaxies, it is inevitable to face uncertainties in the data. Therefore, it is crucial to know how these uncertainties affect the recovery rates of the algorithm. The most important source of error in optical surveys is the redshift. For example, the 2df Galaxy Redshift Survey includes uncertainties of $\sim 85 kms^{-1}$ or about 0.0003 in redshift space (for objects above a minimum quality). A comparable amount of error is hence introduced to the data, altering the redshift of each galaxy by a value taken from a normal distribution of $\sigma = 120 kms^{-1}$ ($\Delta z = 0.0004$).

The algorithm is run with these new values of redshift and the results can be observed in figure 5.10. It can be seen in the figure, that the effect over the distance to the closest

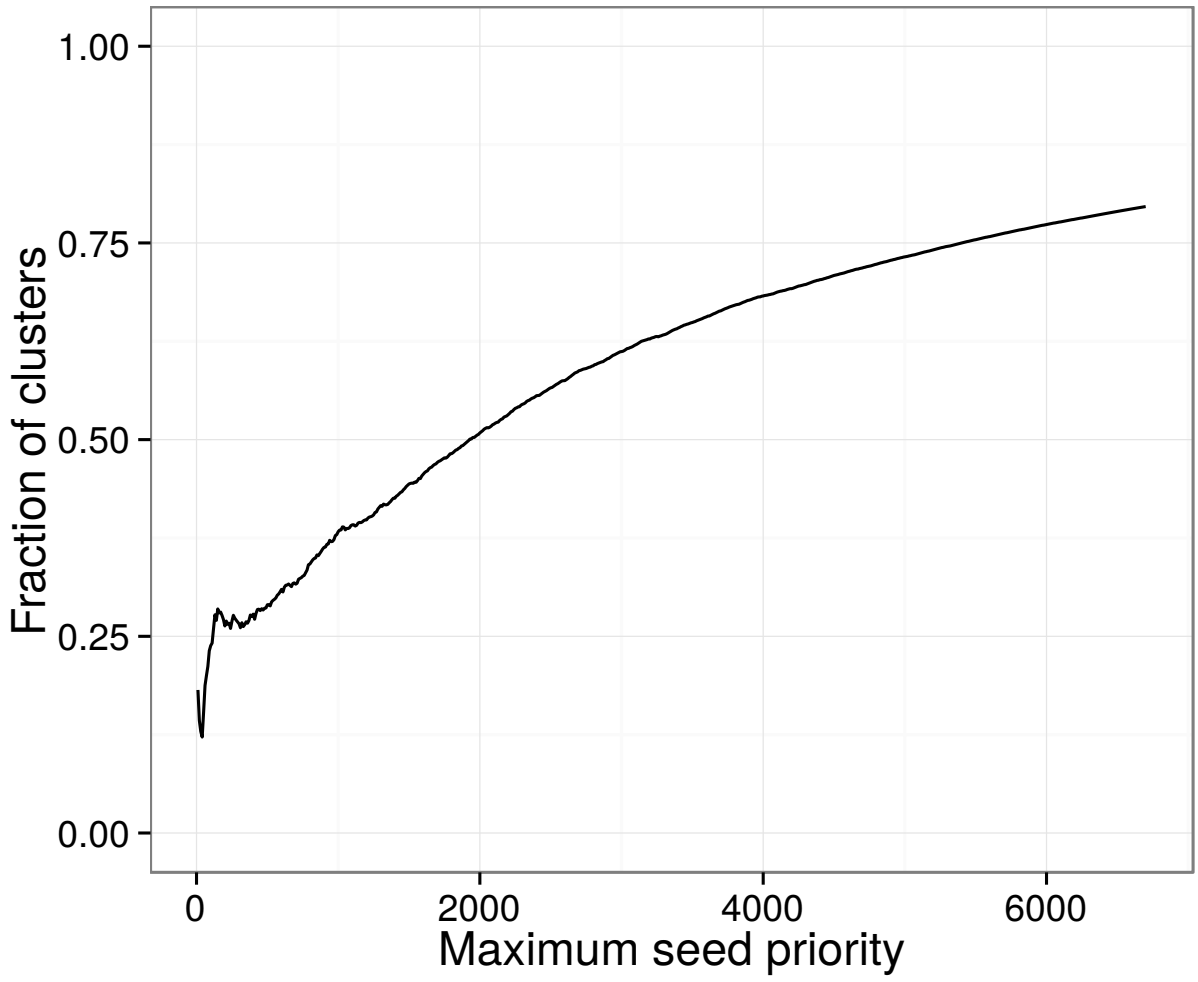

Figure 5.9 VTMLE false positive detections by seed ranking

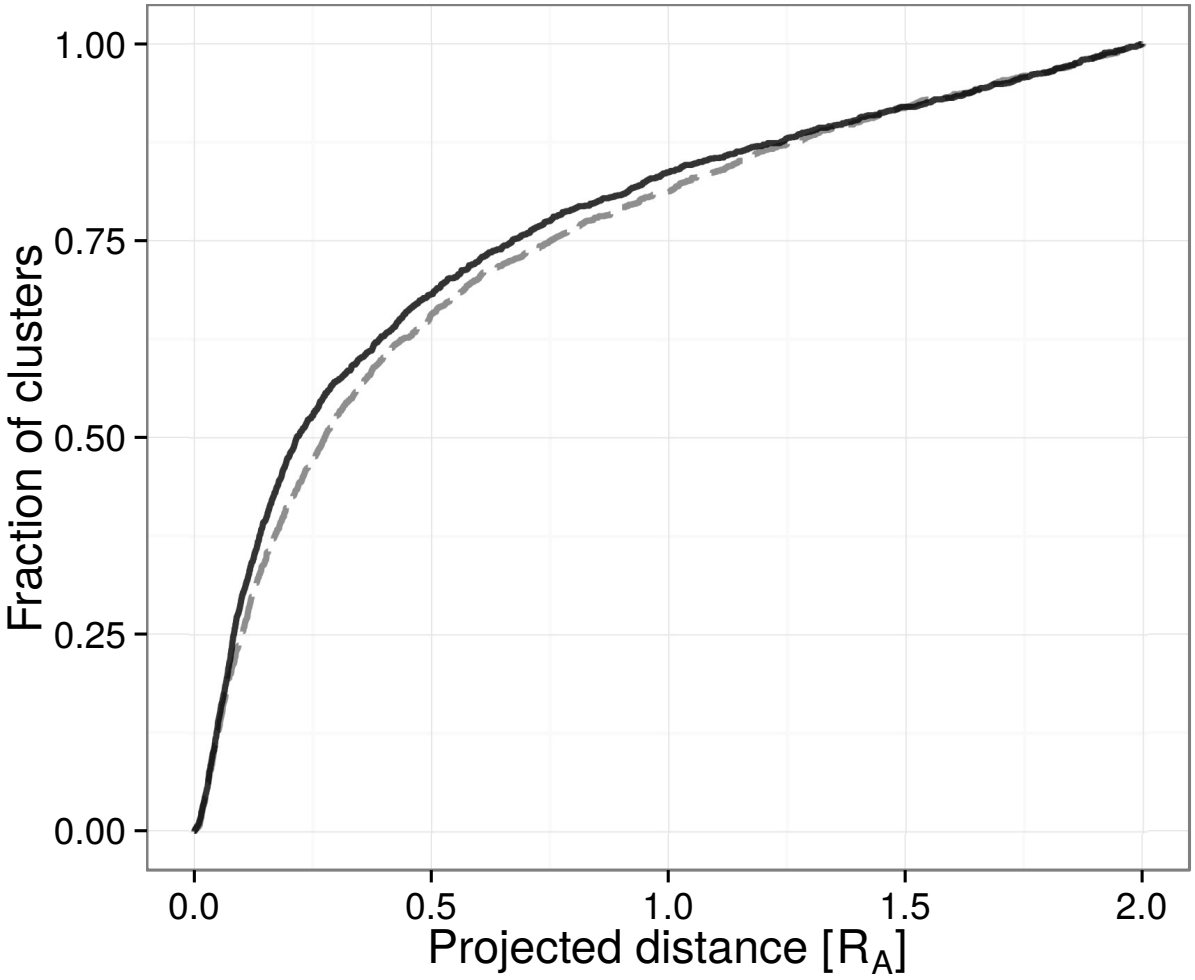

Figure 5.10 Cumulative distribution function of distance between VTMLE detected clusters and the closest reference cluster, including executions with and without error introduced.

Table 5.1. Summary of VTMLE recovery rates

Catalog	Excellent	Good or better	Fair or better
Original	69%	75%	79%
> 15 members	86%	91%	96%
> 30 members	89%	95%	99%
Mass $> 5000\ 10^{10}h^{-1}M_{\odot}$	66%	74%	79%
Mass $> 10000\ 10^{10}h^{-1}M_{\odot}$	64%	74%	78%
$z > 0.06$	67%	73%	78%
$z > 0.15$	53%	62%	66%

reference cluster is minimal. The difference in distances less than 1 Abell Radius are of the order of $\sim 3\%$. Since this is the most important metric for the next stage of the Vocludet algorithm, it is concluded that VTMLE is very robust to errors in redshift space.

5.1.6 The GapperR200 stage

Following the detections made by the VTMLE stage, GapperR200 takes place to determine the galaxy membership of each cluster. It is important to notice that GapperR200 cannot improve the detection rates determined by the previous stage, since it only takes the list of high-density points in space given by VTMLE and converts them to proper clusters. To ensure the best results possible, it is important that the inputs received by this stage are very close to the actual cluster centroids.

This section describes the analysis of the GapperR200 stage independent of VTMLE, though taking into consideration the fact that the VOCLUDET algorithm will run these two together.

Collisions

One important restriction established by the Vocludet algorithm is that no galaxies in common are allowed. This restriction is put into practice the following way: If during the execution of the GapperR200 stage a cluster is found to have a galaxy that was already assigned to another one, the cluster detected in second place gets discarded. The non-overlapping galaxies of the discarded clusters can be assigned to another cluster while processing the rest of the seeds. To understand how this affects the Vocludet recovery rates, we have conducted the following test. We consider an ideal input to the GapperR200 stage, that is, the seeds are located exactly at the Millennium cluster centroids. We find that in these perfect conditions, no collisions occur. We conclude that the Millennium clusters as detected by GapperR200 are not generally close enough to each other for collisions to occur. However, when 2 VTMLE seeds happen to be in close vicinity to each other, only the cluster arising from the first seed (with higher priority) will prevail after the GapperR200 stage . Due to the geometrical nature of the VTMLE stage, a high local density, even if it is in a small region, is enough to

generate a cell of small volume (and thus high priority), which is not always ideal.

5.2 Validation

5.2.1 Resulting catalog

The resulting catalog from the execution of Vocludet over the 2dFGRS mock data is composed of 1614 galaxy groups with a median galaxy membership of $N = 6$. Vocludet produces the clusters in a consecutive fashion, determined by the size of the 3D Voronoi seed cells, so every cluster has a number which indicates its creation number (1 to 1614). Such number correlates qualitatively with descending galaxy number density. The detections are classified into two types (I and II). 681 are Type I and thus have a well defined R_{200} value determined by the GapperR200 algorithm. 933 are Type II, i.e. they could not be assigned a R_{200} value by the GapperR200 algorithm but stil have a significant concentration of galaxies ($>= 5$) in an area of radius $0.5h^{-1}$ and isolated by $1000kms^{-1}$ gaps. 1384 Vocludet detections comply with the condition N$\geq$ 5, that is, the same limit as the Millennium reference clusters.

Figure 5.11 shows the redshift distribution of the Vocludet clusters which shows a drop at $z \sim 0.12$; a similar plot is provided for the Millennium clusters wher it can be seen a similar distribution, though flatter through $z \sim 0.15$.

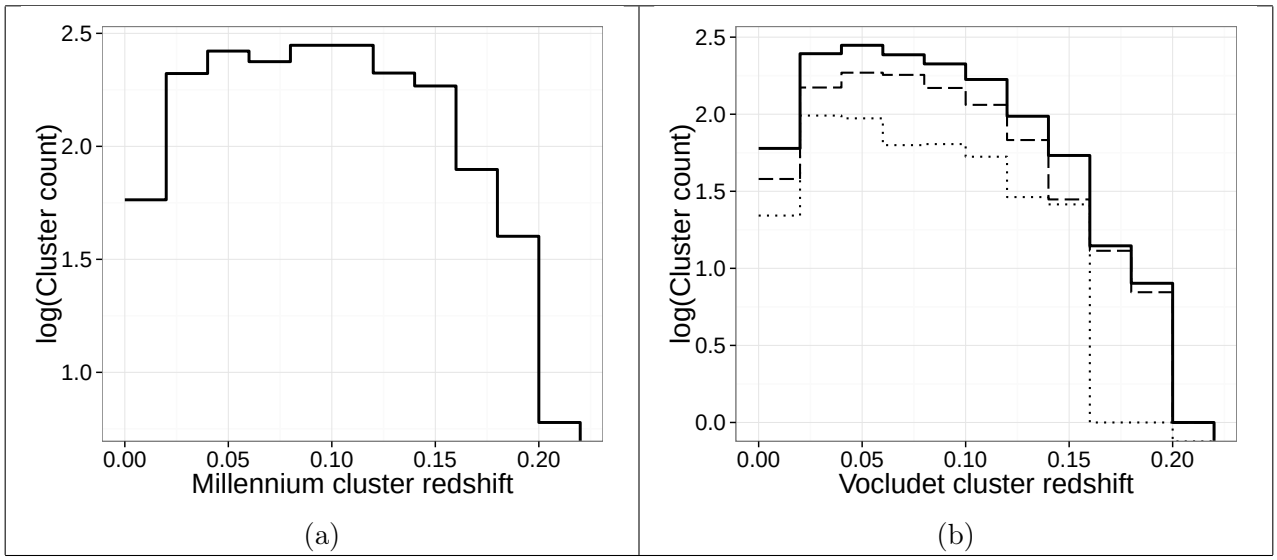

Figure 5.11 Millennium and Vocludet clusters. For Vocludet, solid lines are the total, dotted are Type I clusters and dashed lines are Type II clusters.

A Vocludet detection is considered to be valid if it has a counterpart in the reference catalog, that is, if it contains at least 25% of the galaxies of a Millennium cluster. 925 Vocludet clusters have correspondence in the Millennium catalog, resulting in a completeness of 50%. Table 5.2 lists the decomposition of the Vocludet detections by ranges of multiplicity and redshift, which includes the inferred purity for each of the subsets.

Table 5.2. Breakdown of Vocludet detections (1614) by ranges of multiplicity and redshift

	Ng ≥ 5		Ng ≥ 10	
z range	Number	Purity	Number	Purity
0.009 - 0.205	1384	67%	358	90%
0.009 - 0.050	431	57%	155	83%
0.050 - 0.100	611	66%	144	96%
0.100 - 0.150	298	68%	52	94%

Figure 5.12 shows the histogram of the multiplicities. It can be seen that lower multiplicities are dominated by Type II clusters, while Type I clusters include some clusters with large multiplicities. The ensemble is also dominated by low multiplicity, with the majority of the VOCLUDET clusters having less than 10 galaxies.

Figure 5.13 shows the histogram of sizes; they have a mean projected radius of $0.3R_A$. It can be seen that Type II clusters reach only 0.5 R_{Abell}, while some Type I clusters reach values close to 1 R_{Abell}.

Figure 5.14 shows a comparison with respect to the number of galaxies in corresponding clusters, considering only the Millennium clusters with 10 galaxies or more. Even though most of the corresponding clusters have similar number of galaxies, the Vocludet clusters tend to have slightly less galaxies. The Residual standard error (RSE) with respect to the line of slope 1 is 16.24. In the case of the σ_v, shown in figure 5.15, the dispersion with respect to the slope 1 line is considerable yet the trend is still visible, and its RSE value is 154.49. Finally, the redshift comparison is very close since each valid detected cluster requires having at least 25% of the matching Millennium reference catalog cluster's galaxies, obtaining a RSE of 0.00067.

Vocludet found 933 Type I clusters with a median of 6 galaxies each. Figure 5.12 shows the distribution of the number of galaxies in each cluster. Out of the 933 Type I clusters, 539 are valid detections, having an overall purity of 58%. For Type I clusters of more than 10 galaxies, the purity reaches 88% (101 valid clusters out of 115).

5.2.2 Completeness and purity

The final results for the Vocludet algorithm, using the optimal parameters are the following. The completeness and purity rates are shown in figure 5.17 as a function of multiplicity. It can be seen that both the completeness and purity rates increases with increasing number of galaxies up to $N \sim 30$. The overall recovery rate is 59%, but it increases significantly, to $\sim 75\%$ when we consider only clusters with $N \geq 10$. The overall purity rate, considering all clusters is $\sim 66\%$. Considering clusters with a minimum of 10 galaxies, the purity rate reaches $\sim 90\%$ with little change for higher number of galaxies.

Despite that cross-identification requires a minimum intersection of 25%, most of the

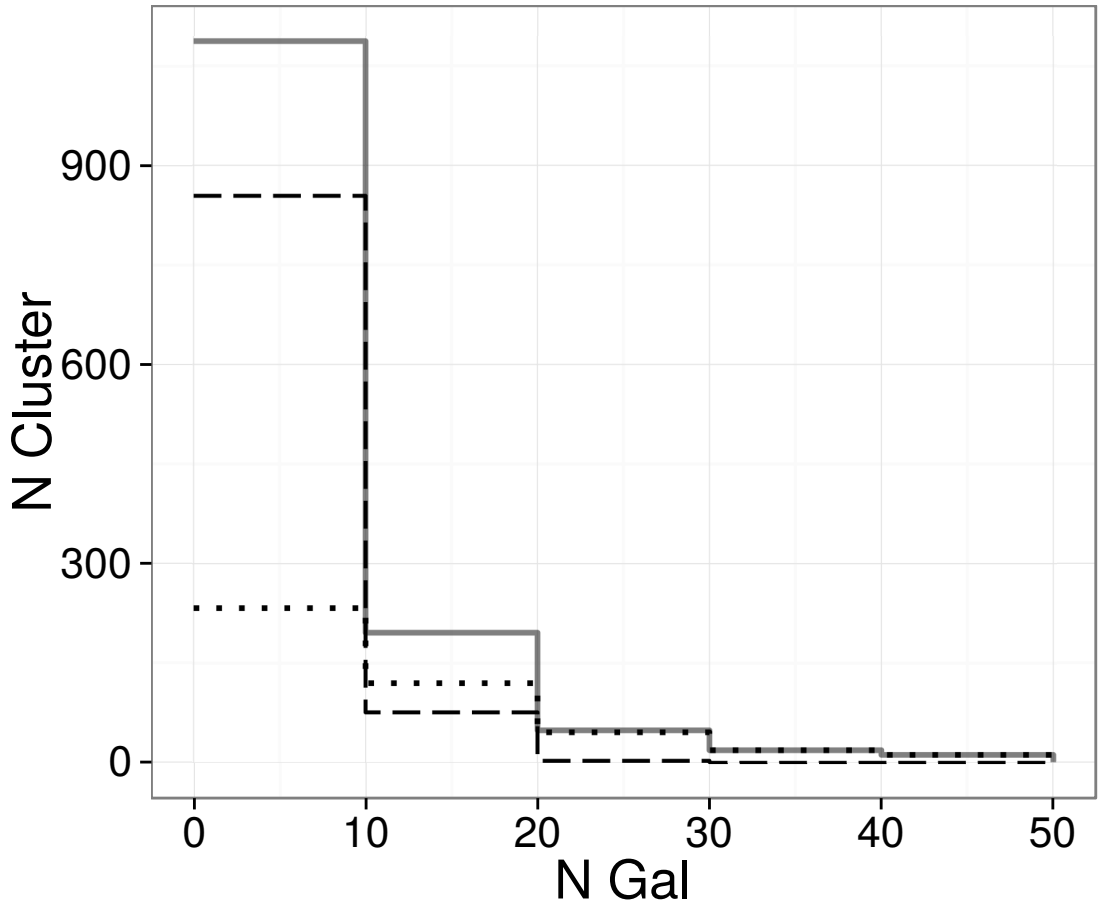

Figure 5.12 Distribution of number of galaxies per Vocludet cluster, up to 50 galaxies. There are 22 clusters with $N_{gal} > 50$ not included in the histogram. Solid lines are the total, dotted are Type I clusters and dashed are Type II clusters.

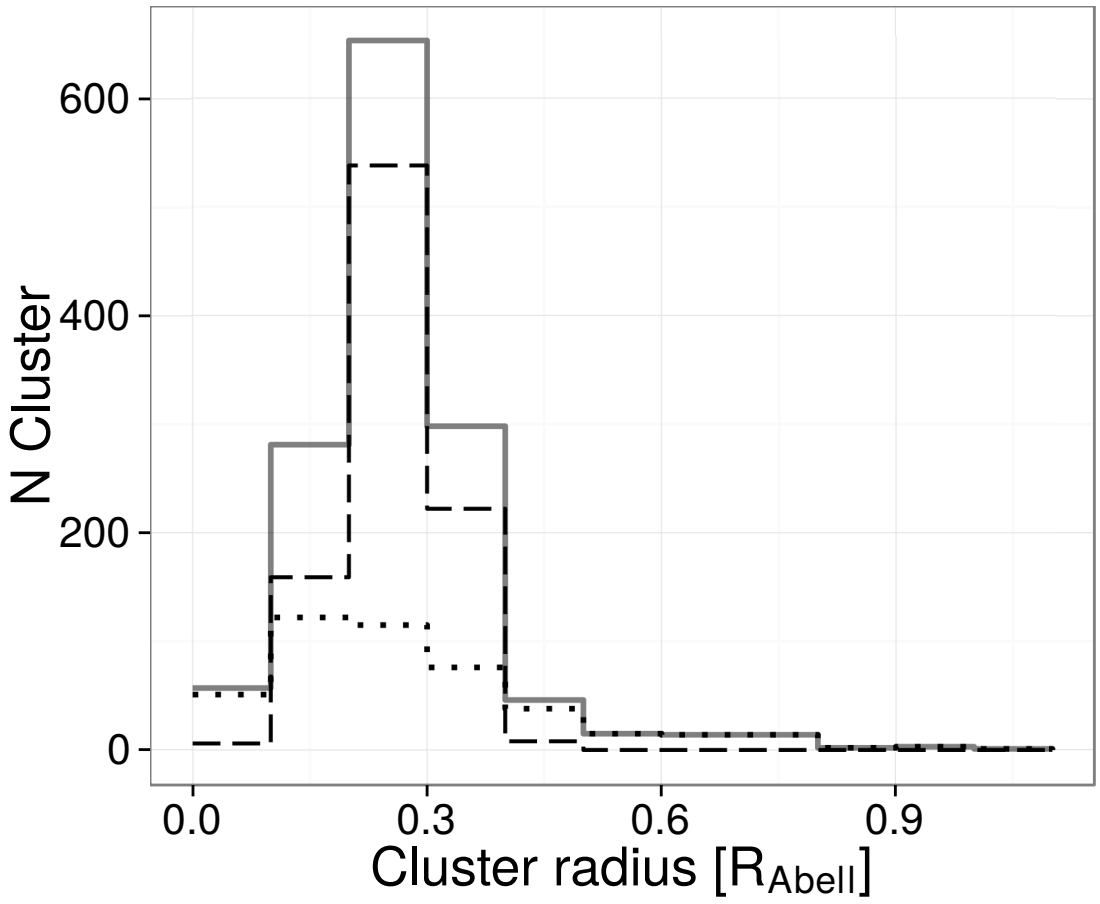

Figure 5.13 Distribution of projected radii of Vocludet clusters. Solid lines are the total, dotted are Type I clusters and dashed are Type II clusters.

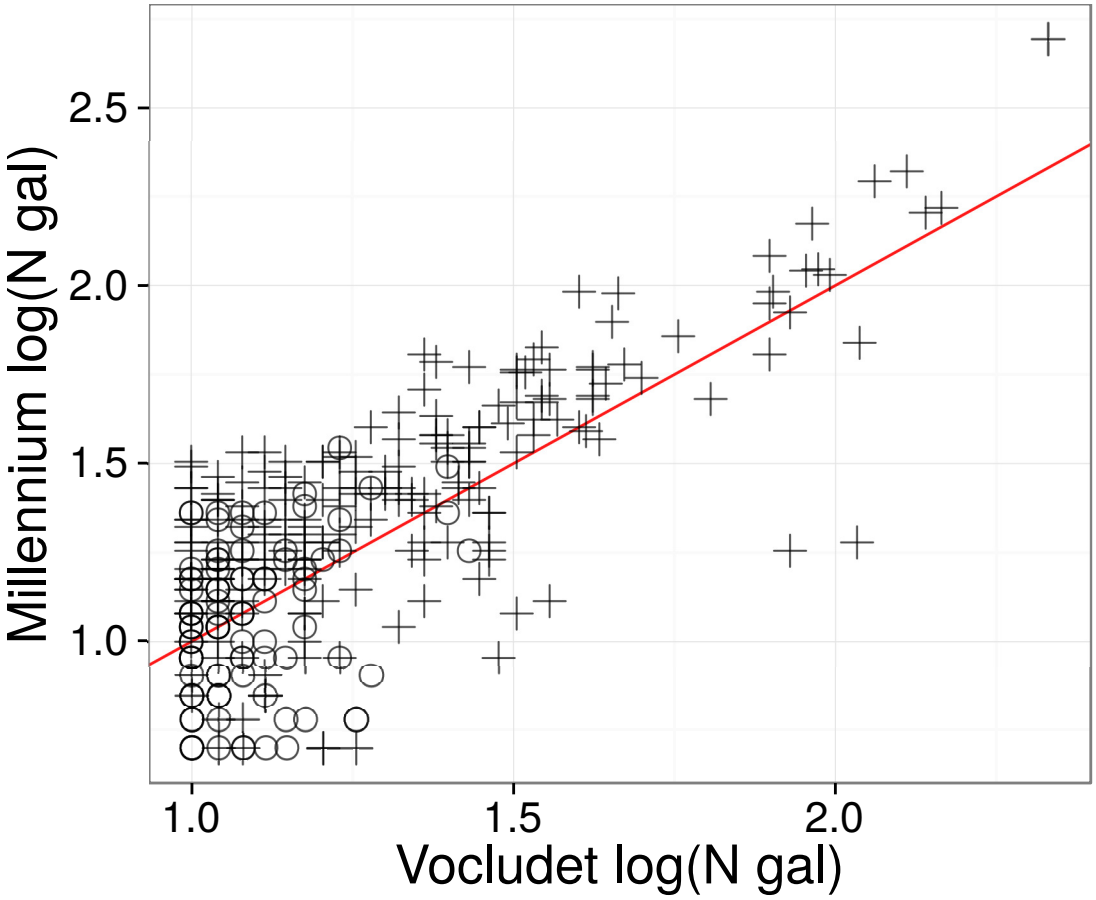

Figure 5.14 $N \geq 10$ Vocludet-Millennium clusters, multiplicity comparison. Crosses are Type I clusters and open circles Type II.

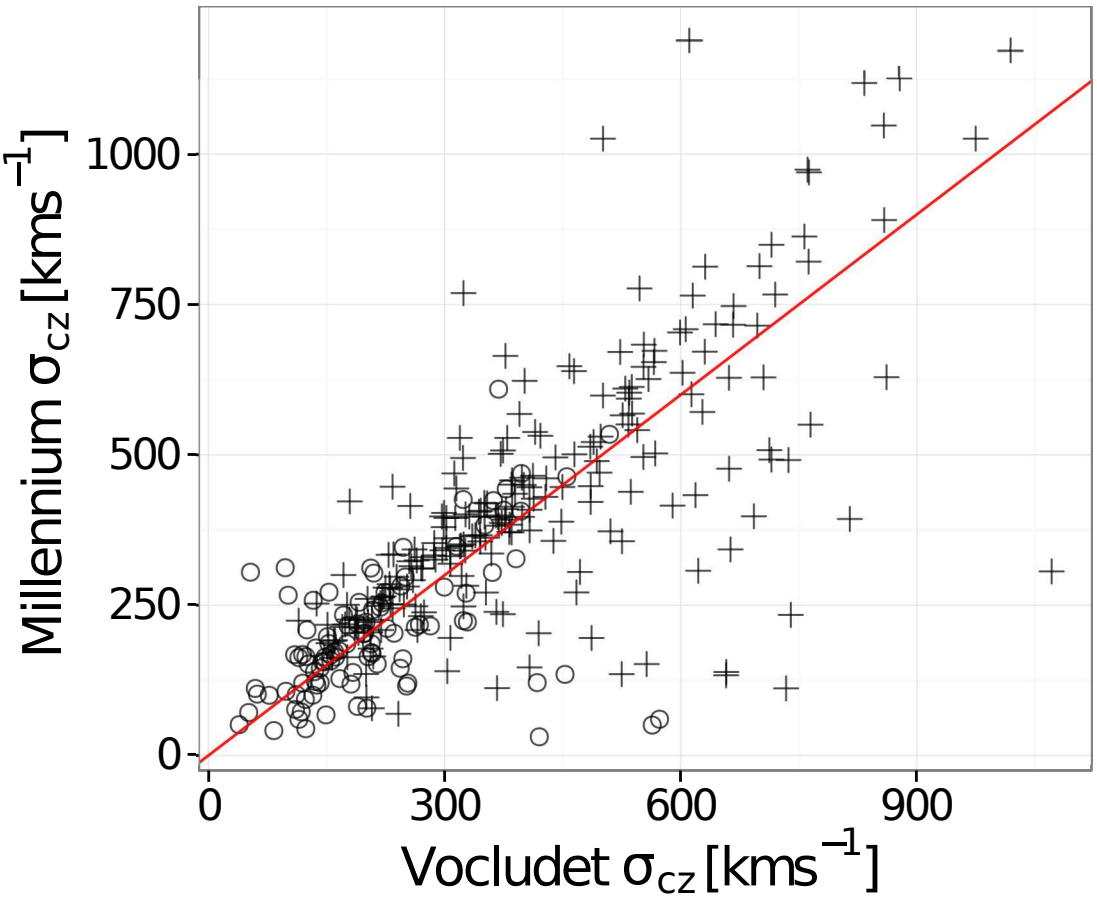

Figure 5.15 $N \geq 10$ Vocludet-Millennium clusters, σ_v comparison. Crosses are Type I clusters and open circles Type II.

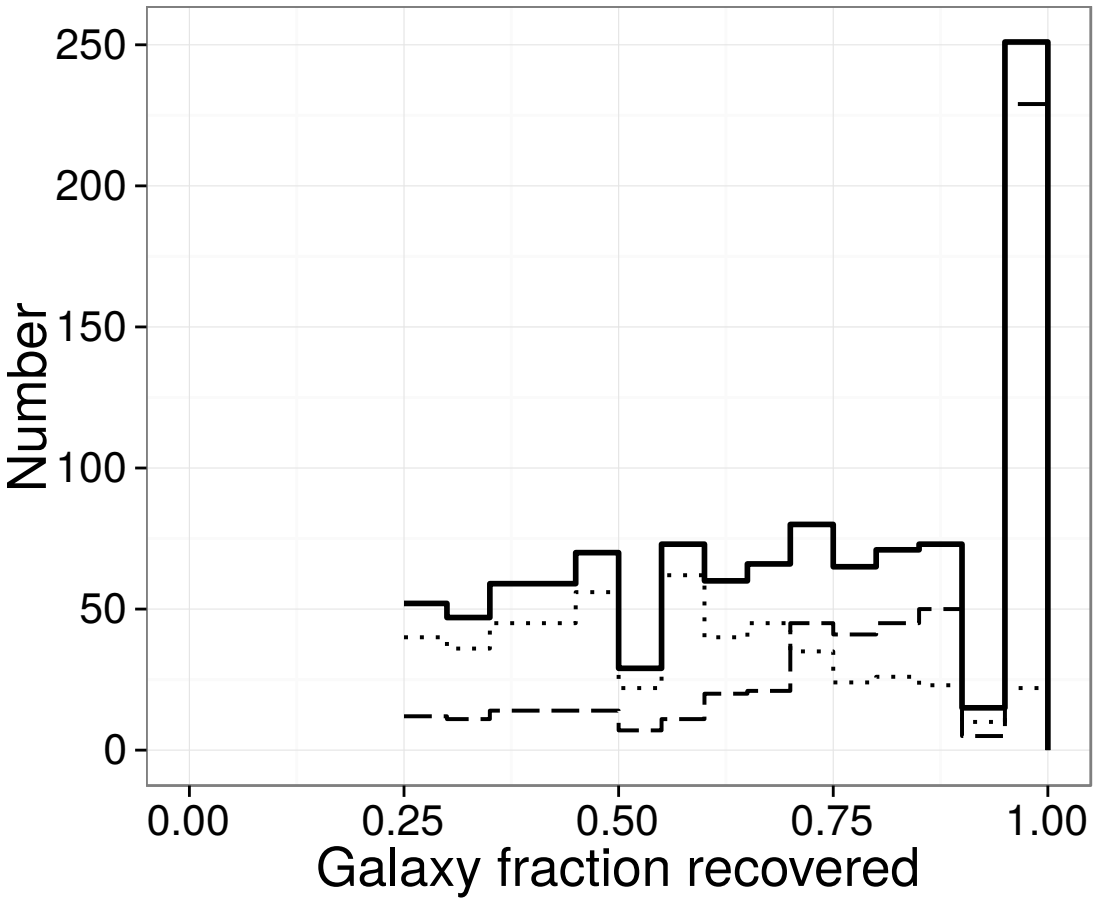

Figure 5.16 Vocludet clusters galaxy completeness histogram

VOCLUDET detected clusters have galaxy completeness values higher than 80% as shown in figure 5.16. 40% of clusters have a purity higher than 80%, and 75% higher than 50%.

Figure 5.18 shows the recovery and purity rates by cluster halo mass. The recovery rate stays around to 70%, with a slight decrease for the least and most massive clusters, while the purity rate oscillates around 90%. High-mass clusters are not numerous in the simulations, which causes higher fluctuations in recovery and purity at these mass domains.

5.2.3 Velocity dispersions

As was mentioned before, the line-of-sight velocity dispersion was calculated using the bi-weight estimator (Beers et al. 1990). As can be seen in figure 5.15, the dispersion with respect to the slope 1 tends to be higher for $\sigma_{cz} > 500 kms^{-1}$. Most of the Type II clusters have $\sigma_{cz} \lesssim 400 kms^{-1}$ and there is a good agreement between the Vocludet and Millennium values.

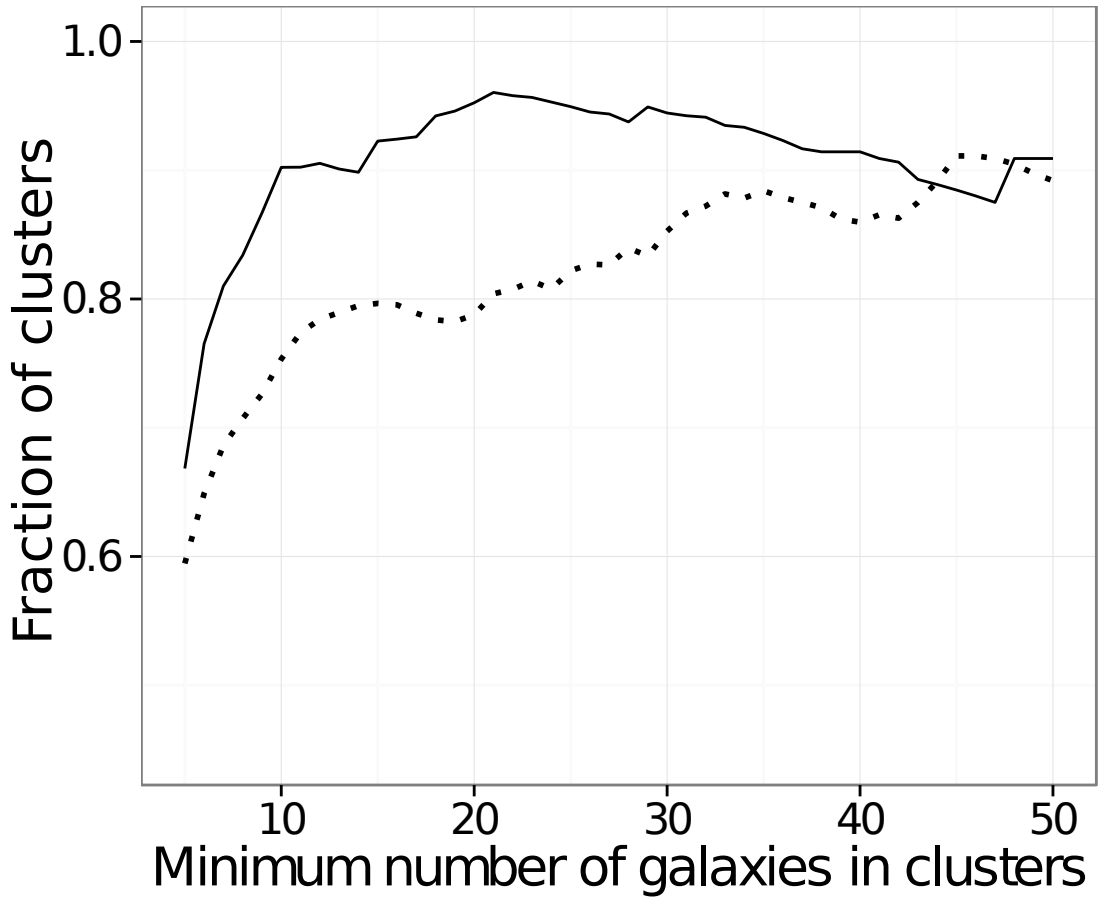

Figure 5.17 Millennium clusters recovery (dotted) and purity (solid) rate by cluster multiplicity range.

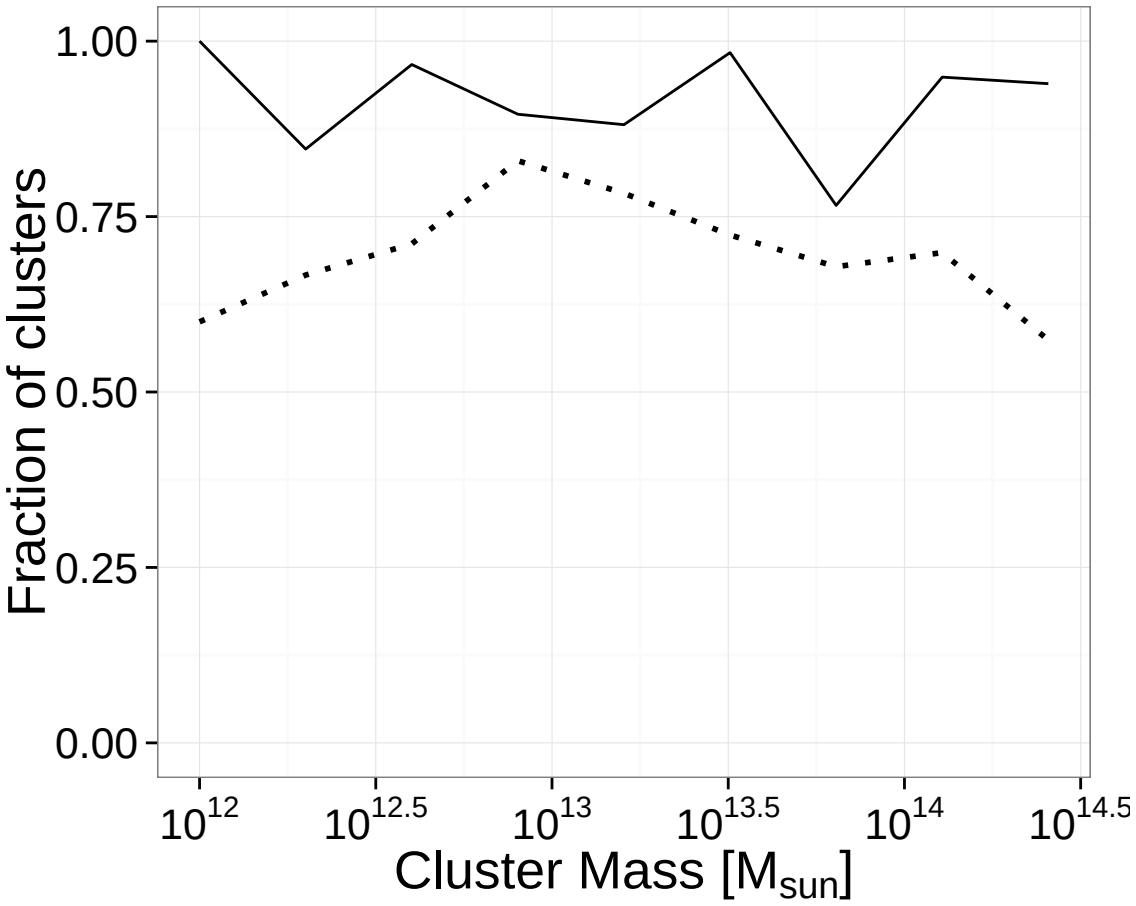

Figure 5.18 Millennium clusters recovery (dotted) and purity (solid) rate by cluster mass range.

5.2.4 Comparison with other works

Many galaxy cluster detection algorithms in the optical already exist in the literature. Some of the recent ones, like the one presented by milkeraitis Milkeraitis et al. report an overall recovery rate of 54.8% for clusters more massive than $1.5 \times 10^{13} M_{\odot}$ and up to 100% for the most massive clusters ($\geq 2.5 \times 10^{13} M_{\odot}$). Their cluster detections are based on galaxy cluster radial profiles, luminosity functions and redshift information. In contrast, Vocludet uses only redshift information and its performance is found to be independent of the mass of the clusters, with recovery rates close to 80% across a wide mass range (see figure 5.18). With respect to the purity rate, Milkeratis et al. reports an overall rate of 84.4% (15.6% false detections). Vocludet has a similar overall purity, though when all detected clusters with less than 10 galaxies are filtered out we attain better performance ($\sim 90\%$ than milkeraitis.

Another algorithm with similar characteristics to Vocludet, is the Voronoi-Delaunay Method marinoni, studied and applied recently to the VVDS survey by cucciati. The algorithm utilises a geometrical method for identifying galaxy groups within flux-limited redshift surveys. They report results of an overall recovery rate of $\sim 60\%$ and a purity of $\sim 50\%$. Additionally, results using an alternative set of parameter to optimize the purity are reported, which consist of a recovery rate of $\sim 45\%$ and a purity of $\sim 75\%$.

Vocludet, as well as both of the previous methods discussed, were tested with data from the Millennium simulations which allowed us to refine, evaluate and compare the different methods. It is important to notice though, that Vocludet was tested with most galaxies in the redshift range $0.009 < z < 0.2$, whereas the other two methods consider the redshift range $0.2 < z < 1$. Overall, the results of the Vocludet algorithm are competitive with other methods, but it surpasses them in the identification of clusters with more than 10 members. The latter is desirable characteristic for the study of the rich clusters.

Conclusions

The reliable detection and study of clusters of galaxies are two important challenges that astronomy faces at present. These vast concentrations of matter are thought to hold the key to reveal some of the biggest mysteries of the universe such as the nature of dark matter and dark energy.

The first objective of this work was to validate and optimize the Vocludet algorithm, so that it can be applied confidently to real galaxy surveys and obtain meaningful results. This implies knowing the strengths and weaknesses of the algorithm, since it is currently impossible to get perfect results. The second objective was to create an interactive visualization tool that helped with the analysis and interpretation of the output of the algorithm.

The algorithm was optimized by first modifying its second stage and later by determining which set of parameters produced the best results over the Millennium Simulation data set. Along with the optimization, the impact the variation of each parameter has over the results was also presented. This is particularly useful for the application to data sets with different characteristics or for the purpose of finding type of cluster with a specific set of properties, such as number of galaxies or extent.

The validation of the algorithm was carried out by the careful analysis of the results of the execution of Vocludet over the Millennium data set, as well as the calculation of multiple statistical values such as the purity and recovery rates, which indicate the capacity of the algorithm to recover only real clusters and the total fraction of them recovered, respectively.

An interactive visualization tool was presented. This tool allows the user to visualize the results of the Vocludet algorithm in a way that is focused both on the algorithm itself and the galaxy clusters. An example of this is the view of each cluster, which is enriched with the rendering of surrounding galaxies located in its vicinity, which helps determine the level of isolation of the structure in space as well as the performance of the algorithm.

The performance achieved by Vocludet is competitive with similar algorithms in the field. The determined completeness and purity of the detected clusters with $N \geq 10$ is of 75% and 90 % respectively, based on a Millennium reference catalog and valid across a wide mass range of these clusters.

Future work

To reveal the full potential of Vocludet, it has to be applied to various sources of real data. This task involves further optimization tailored specifically to the data sets chosen. It could also be revealed the need to keep improving the algorithm via modification of the code itself.

For the visualization tool, there is no limit to the potential features that can be implemented to improve the quality of the analysis provided. The code is developed in a way that it is easily extensible, and the use of web technologies make it portable to essentially any modern platform.

Bibliography

[1] Z. Ivezic, J. A. Tyson, T. Axelrod, D. Burke, C. F. Claver, K. H. Cook, S. M. Kahn, R. H. Lupton, D. G. Monet, P. A. Pinto, M. A. Strauss, C. W. Stubbs, L. Jones, A. Saha, R. Scranton, C. Smith, and LSST Collaboration. LSST: From Science Drivers To Reference Design And Anticipated Data Products. In *American Astronomical Society Meeting Abstracts #213*, volume 41 of *Bulletin of the American Astronomical Society*, January 2009.

[2] D. Pizarro Pizarro. Galaxy cluster detection using nonparametric maximum likelihood estimation of features in voronoi tessellations. Master's thesis, Universidad de Chile, 2007.

[3] I. H. Li and H. K. C. Yee. Finding Galaxy Groups in Photometric-Redshift Space: The Probability Friends-of-Friends Algorithm. *Astronomical Journal*, 135:809–822, March 2008.

[4] C. Adami, F. Durret, C. Benoist, J. Coupon, A. Mazure, B. Meneux, O. Ilbert, J. Blaizot, S. Arnouts, A. Cappi, B. Garilli, L. Guennou, V. Lebrun, O. Lefèvre, S. Maurogordato, H. J. McCracken, Y. Mellier, E. Slezak, L. Tresse, and M. P. Ulmer. Galaxy structure searches by photometric redshifts in the CFHTLS. *Astronomy and Astrophysics*, 509:A81, January 2010.

[5] P. A. A. Lopes, R. R. de Carvalho, R. R. Gal, S. G. Djorgovski, S. C. Odewahn, A. A. Mahabal, and R. J. Brunner. The Northern Sky Optical Cluster Survey. IV. An Intermediate-Redshift Galaxy Cluster Catalog and the Comparison of Two Detection Algorithms. *Astronomical Journal*, 128:1017–1045, September 2004.

[6] C. van Breukelen and L. Clewley. A reliable cluster detection technique using photometric redshifts: introducing the 2TecX algorithm. *Monthly Notices of the Royal Astronomical Society*, 395:1845–1856, June 2009.

[7] M. D. Gladders and H. K. C. Yee. A New Method For Galaxy Cluster Detection. I. The Algorithm. *Astronomical Journal*, 120:2148–2162, October 2000.

[8] M. D. Gladders and H. K. C. Yee. A New Method For Galaxy Cluster Detection. I. The Algorithm. *Astronomical Journal*, 120:2148–2162, October 2000.

[9] T. Kodama, I. Tanaka, M. Kajisawa, J. Kurk, B. Venemans, C. De Breuck, J. Vernet,

and C. Lidman. The first appearance of the red sequence of galaxies in proto-clusters. *Monthly Notices of the Royal Astronomical Society*, 377:1717–1725, June 2007.

[10] K. Thanjavur, J. Willis, and D. Crampton. K2: A New Method for the Detection of Galaxy Clusters Based on Canada-France-Hawaii Telescope Legacy Survey Multicolor Images. *The Astrophysical Journal*, 706:571–591, November 2009.

[11] C. S. Kochanek, M. White, J. Huchra, L. Macri, T. H. Jarrett, S. E. Schneider, and J. Mader. Clusters of Galaxies in the Local Universe. *The Astrophysical Journal*, 585:161–181, March 2003.

[12] L. F. Grove, C. Benoist, and F. Martel. Galaxy clusters in the CFHTLS. II. Matched-filter results in different passbands. *Astronomy and Astrophysics*, 494:845–855, February 2009.

[13] F. Menanteau, J. P. Hughes, R. Jimenez, C. Hernandez-Monteagudo, L. Verde, A. Kosowsky, K. Moodley, L. Infante, and N. Roche. Southern Cosmology Survey. I. Optical Cluster Detections and Predictions for the Southern Common-Area Millimeter-Wave Experiments. *The Astrophysical Journal*, 698:1221–1231, June 2009.

[14] & Fraley C. Allard, D. Journal of the American Statistical Association. In *American Astronomical Society Meeting Abstracts #213*, volume 92 of *Bulletin of the American Astronomical Society*, January 1997.

[15] G. Lemson and t. Virgo Consortium. Halo and Galaxy Formation Histories from the Millennium Simulation: Public release of a VO-oriented and SQL-queryable database for studying the evolution of galaxies in the LambdaCDM cosmogony. *ArXiv Astrophysics e-prints*, August 2006.

[16] R. B. Tully, H. Courtois, Y. Hoffman, and D. Pomarède. The Laniakea supercluster of galaxies. *Nature*, 513:71–73, September 2014.

[17] Fleenor M. Miller J., Quammen C. Interactive visualization of intercluster galaxy structures in the horologium-reticulum supercluster. 2006.

[18] D. Pizarro, L. E. Campusano, R. G. Clowes, P. Virgili, N. Hitschfeld-Kahler, and I. K. Söchting. Clustering of 3d spatial points using maximum likelihood estimator over voronoi tessellations: Study of the galaxy distribution in redshift space isvd. 112, 2006.

[19] D. Pizarro. *Galaxy cluster detection using nonparametric maximum likelihood estimation of features in Voronoi tessellations: Computer Science Engineering Master Thesis: Departamento de Ciencias de la Computación.* Universidad de Chile, 2007.

[20] I. K. Söchting, R. G. Clowes, and L. E. Campusano. Monthly notices of the royal astronomical society. 331:569, 2002.

[21] I. K. Söchting, R. G. Clowes, and L. E. Campusano. Monthly notices of the royal astronomical society. 347:1241, 2004.

[22] Ann I. Zabludoff and Dennis Zaritsky. A collision of subclusters in abell 754. *The Astrophysical Journal Letters*, 447(1):L21, 1995.

[23] C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM TRANSACTIONS ON MATHEMATICAL SOFTWARE*, 22(4):469–483, 1996.

[24] De Propris. R., et al. 2002. *Monthly Notices of the Royal Astronomical Society*, 329:87.

[25] S. Pereira. *Validación de un algoritmo de detección de cúmulos de galaxias (vocludet) y visualización sobre un wall-display.* Universidad de Chile, 2014.

[26] T. C. Beers, K. Flynn, and K. Gebhardt. Measures of location and scale for velocities in clusters of galaxies - A robust approach. , 100:32–46, July 1990.

[27] Leland Wilkinson. *The Grammar of Graphics (Statistics and Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[28] J. Blaizot, Y. Wadadekar, B. Guiderdoni, S. T. Colombi, E. Bertin, F. R. Bouchet, J. E. G. Devriendt, and S. Hatton. MoMaF: the Mock Map Facility. *Monthly Notices of the Royal Astronomical Society*, 360:159–175, June 2005.

[29] S. G. Kleinmann, M. G. Lysaght, and W. L. et al. Pughe. *Experimental Astronomy*, 3:65, 1994.

[30] M. Milkeraitis, L. van Waerbeke, and C. et al. Heymans. *Monthly Notices of the Royal Astronomical Society*, 406:673, 2010.