# Central Lancashire Online Knowledge (CLoK)

CrossMark

# A 3D Voronoi+Gapper Galaxy Cluster Finder in Redshift Space to $z \sim 0.2$ I: an Algorithm Optimized for the 2dFGRS

Sebastián Pereira[1,4], Luis E. Campusano[1,5], Nancy Hitschfeld-Kahler[1,4], Daniel Pizarro[1,4], Christopher P. Haines[1,6], Roger G. Clowes[2], Gabriel Marinello[2], and Ilona K. Söchting[3]

[1] Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile; luis@das.uchile.cl
[2] Jeremiah Horrocks Institute, University of Central Lancashire, Preston PR1 2HE, UK
[3] University of Oxford, Astrophysics, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK

## Abstract

This paper is the first in a series, presenting a new galaxy cluster finder based on a three-dimensional Voronoi Tesselation plus a maximum likelihood estimator, followed by gapping-filtering in radial velocity(VoML+G). The scientific aim of the series is a reassessment of the diversity of optical clusters in the local universe. A mock galaxy database mimicking the southern strip of the magnitude(blue)-limited 2dF Galaxy Redshift Survey (2dFGRS), for the redshift range $0.009 < z < 0.22$, is built on the basis of the Millennium Simulation of the LCDM cosmology and a reference catalog of "Millennium clusters," spannning across the $1.0 \times 10^{12}$–$1.0 \times 10^{15} M_\odot\, h^{-1}$ dark matter (DM) halo mass range, is recorded. The validation of VoML+G is performed through its application to the mock data and the ensuing determination of the completeness and purity of the cluster detections by comparison with the reference catalog. The execution of VoML+G over the 2dFGRS mock data identified 1614 clusters, 22% with $N_g \geqslant 10$, 64 percent with $10 > N_g \geqslant 5$, and 14% with $N_g < 5$. The ensemble of VoML+G clusters has a ~59% completeness and a ~66% purity, whereas the subsample with $N_g \geqslant 10$, to $z \sim 0.14$, has greatly improved mean rates of ~75% and ~90%, respectively. The VoML+G cluster velocity dispersions are found to be compatible with those corresponding to "Millennium clusters" over the 300–1000 km s$^{-1}$ interval, i.e., for cluster halo masses in excess of ~$3.0 \times 10^{13} M_\odot\, h^{-1}$.

*Key words:* cosmology: observations – dark matter – galaxies: clusters: general – large-scale structure of universe – methods: data analysis – surveys

## 1. Introduction

Clusters of galaxies, because of their large masses and detectability over a large fraction of the age of the universe, provide test ground for cosmology (Henry et al. 2009; Vikhlinin et al. 2009; Mantz et al. 2010; Rozo et al. 2010; Clerc et al. 2012; Benson et al. 2013), for the history of the agglomeration of matter (Jing et al. 1998; Peacock & Smith 2000; Seljak 2000; Scoccimarro et al. 2001; Berlind & Weinberg 2002; Zheng et al. 2005, 2009), as well as for the description of galaxy evolution (e.g., Dressler 1980; Lewis et al. 2002; Boselli & Gavazzi 2006; Haines et al. 2015). In addition, galaxy clusters are one of the key probes of dark energy for ongoing and upcoming major photometric surveys (e.g., DES: Flaugher 2005; LSST Science Collaboration et al. 2009)

Arguably, Messier (Messier 1781) was the first to detect a cluster of galaxies (Virgo), albeit in terms of inhomogeneities on the distribution of nebulae over large areas of the sky. Since 1925, optical clusters have been defined by the non-random spatial distribution of galaxies (see Way et al. 2011). Several decades later, the first large catalogs of galaxy clusters were created using photographic plates (e.g., Abell 1958; Zwicky et al. 1968; Abell et al. 1989), whose listed clusters, remarkably, are still among the most cited and studied ones. More recently, the advent of optical multi-band and spectroscopic surveys has led to the development of a large number of cluster-finding

algorithms. Cluster-finders are applied onto either photometric or redshift surveys, and use methods ranging from simple matched filters to applications of Voronoi tessellations that may or may not utilize the cluster red sequence (CRS). Remarkably, few of the new generation optical catalogs of clusters have been utilized for cosmological parameter constraints (e.g., Rozo et al. 2010; Mana et al. 2013) or for the re-evaluation of the diversity of galaxy constituency in clusters (e.g., Ascaso et al. 2015).

A recent example of the use of an optical catalog of massive clusters for constraining cosmology was made possible partly by the development of a new generation red sequence cluster-finder (redMaPPer) by Rykoff et al. (2014) and its application to the Sloan Digital Sky Survey (SDSS, Strauss et al. 2002) DR8 Photometric Galaxy Catalog over the 0.08-0.55 redshift range. It refers to the dark matter *halo assembly bias* (Gao et al. 2005; Croton et al. 2007), whose inference requires the identification of isolated halo samples. Supporting evidence has been reported by Miyatake et al. (2016) in massive clusters, and later expanded by More et al. (2016).

In 1974, Oemler (Oemler 1974) described some diversity in the Galaxy constituency in Abell clusters, which included a "spiral-rich" category (for a compilation, see Nilo Castellón et al. 2014). Further evidence for an expanded variety started to emerge about a decade ago. Goto et al. (2002) developed the cut-and-enhance method. From its application to the SDSS, they produced a "CE cluster catalog" that included different galaxy density profiles, luminosity functions, and clusters dominated by either spirals or faint ellipticals. Additionally, Donahue et al. (2002) have suggested that the CRS could be characteristic of massive clusters only (e.g., Aguerri et al. 2007; Ascaso et al. 2012).

---

The C4 algorithm developed by Miller et al. ([2005](#)) does not use the CRS. It identifies clusters as overdensities in a seven-dimensional positional and color space, which allows the possible detection of clusters dominated by blue galaxies. The "C4 catalog," with its 748 clusters, was found to contain ∼90% of the Abell clusters and ∼98% of the known X-ray clusters to $0.03 \leqslant z \leqslant 0.12$ within the surveyed area. However, a large number (555) of their low-$z$ detections by Miller et al. ([2005](#)) had no counterparts in the literature, calling into question their properties and eventual comparison with archetypal clusters. More recently, the realization that the most massive clusters do not always have a luminous red galaxy at their centers (Hoshino et al. [2015](#)) is also due to progress on cluster finders

Automatic cluster finders with fewer assumptions and parameters offer the possibility to identify a wider range of galaxy clusters, and to contribute to a better understanding of cluster formation and evolution. Because the Voronoi Tessellation (VT) is a powerful nonparametric technique for exploring the subtleties of the Galaxy distribution (e.g., Ramella et al. [2001](#)), we chose it as the basis of a new cluster finder, the VoML+G algorithm (VOronoi-based, with a maximum likelihood estimator—MLE—plus a Gapper velocity filter), designed for the robust detection of high-multiplicity (i.e., high number of member galaxies) clusters in galaxy samples within the local universe. The principal motive for developing VoML +G is to further the study of cluster diversity and explore new ways to probe the dependence of halo clustering on assembly history.

In this paper, we describe the VoML+G algorithm and its validation through a mock two-degree Field Galaxy Redshift Survey (Colless et al. [2001](#), hereafter 2dFGRS), derived from the Millennium cosmological simulation (Springel et al. [2005](#); Lemson & Virgo Consortium [2006](#)). This redshift survey was chosen because of its depth ($z \sim 0.2$) and width, its blue magnitude limit, and the large body of literature on clusters in the survey region. The focus of our study is on relatively high-multiplicity massive clusters, and thus the accompanying papers will present the cluster catalog resulting from its application to the actual 2dFGRS data, reassessment of the Abell-like galaxy clusters, discussion of the newly detected clusters, characterization of the cluster galaxy populations as a function of mass, and their spatial clustering properties (Campusano et al., Paper II). Finally, we will present a detailed study of the richest VoML+G clusters, both new and previously known, including a determination of their luminosity functions (Campusano et al., Paper III).

The VoML+G code has the same initial stage as the Pizarro et al. ([2006](#)) algorithm, but the gapper stage is modified to have cluster galaxies selected[7] within $R_{200}$ for the determination of more robust velocity dispersions. An important aspect of the code is that the clusters are not modeled, and the Galaxy members are chosen only on the basis of their 3D positions. In the first stage, the density peaks in a 3D galaxy distribution are identified using a Voronoi Tessellation (VT) followed by a maximum likelihood estimation (MLE) of galaxy structures ($N_g \geqslant 2$). The second stage uses the centroids of the above structures to define the cluster members through a "gapping filter" equivalent to a cylinder, oriented along the line of sight, of given radius ($R_{200}$ or $0.5\,h^{-1}\,\mathrm{Mpc}$) and with a length determined by a specific velocity-gap.

To enhance the algorithm's selection of dynamically simple non-interacting clusters, we incorporate two constraints: an upper limit of two Abell Radii ($R_A$[8]) for the size of the cluster identified in the initial stage of the algorithm, and the rejection of subsequent clusters if they contain galaxies in common with members of previously selected ones. The former is to limit sizes to the ones of the Abell-like clusters and the latter is to strengthen the selection of isolated systems.

The structure of the paper is as follows. We present in Section [2](#) the cluster detector. Section [3](#) describes the simulated 2dFGRS data set, based on the Millennium Simulation. In Section [4](#), we determine the optimization of the parameters of the algorithm. Section [5](#) presents the resulting VoML+G cluster catalog, its characteristics, and a comparison with the Millennium reference clusters. In Section [6](#), we present two-dimensional displays for a few VoML+G clusters, including false-positive detections. Section [7](#) describes the evaluation of the completeness and purity of the VoML+G cluster catalog. In Section [8](#), we discuss the overall performance of the algorithm, impact of the borders of the survey area, masking effects by bright stars, comparison with efficiency values from other cluster finders, and future work. Section [9](#) summarizes our conclusions.
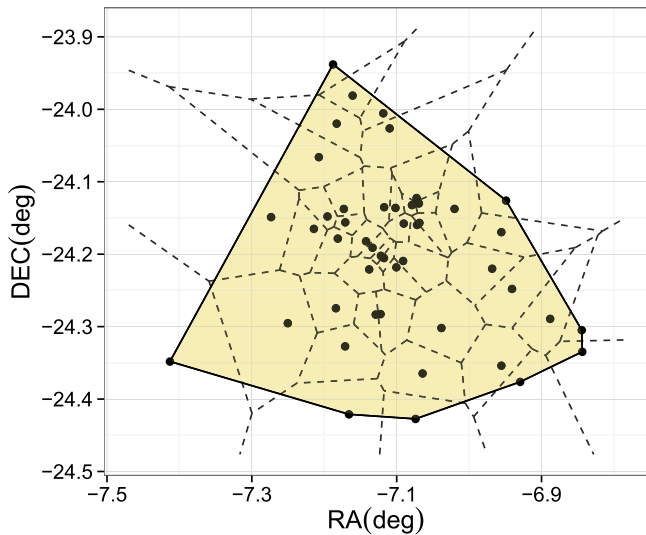
A cosmology with $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, $\Omega_M = 0.3$, and $\Omega_\Lambda = 0.7$ is adopted throughout this paper.

## 2. The Algorithm

The Pizarro et al. ([2006](#)) cluster-finding algorithm is based partly on an extension to 3D of the VT-MLE method, introduced into astronomy by Söchting et al. ([2002](#), [2004](#)) to locate low-$z$ galaxy clusters in color ($B_J$-$R$) slices. In fact, they also incorporated into their algorithm the approach of (De Propris et al. [2002](#), hereafter DP02) who, using the 2dFGRS database, reproduced semi-automatically a collation of clusters from the following catalogs: Abell (Abell [1958](#), Abell), the Automated Plate Measuring (Dalton et al. [1997](#), APM), and the Edinburgh-Durham (Lumsden et al. [1992](#), EDCC). De Propris et al. ([2002](#)) used a "double-gapping filter" technique (Zabludoff et al. [1990](#), hereafter the ZHG technique) to reassess the Abell/APM/EDCC clusters by isolating galaxies within one Abell radius from their cataloged centers. Given that the velocity-gap used by DP02 for the successful recovery of Abell-like clusters was $1000\,\mathrm{km\,s^{-1}}$, it is possibly the appropriate value to use for the application of VoML+G on the "real 2dFGRS" database. This value is nevertheless validated through runs of VoML+G on the mock 2dFGRS (Section [4](#)).

Both the Pizarro et al. ([2006](#)) algorithm and our refined version of it, VoML+G, are based on the above mentioned approaches. These algorithms operate in a pseudo-3D space over a galaxy distribution determined from the Galaxy angular positions and redshifts (hereafter the "redshift-space"). Pizarro et al. ([2006](#)) tested their code with the "real 2dFGRS" data set and the known clusters therein (i.e., not simulations, but real data) and automatically detected, using a velocity gap of $1000\,\mathrm{km\,s^{-1}}$, ∼67% of the Abell-like clusters in the list of DP02 and generally with similar properties. As mentioned, VoML+G improves the Pizarro et al. ([2006](#)) implementation by adopting a more sophisticated gapper stage and by using cosmological simulations to confirm the cluster detections. Its first stage employs a 3D extension of the VT-MLE method,

---

[7]  $R_{200}$ = Radius of a sphere whose mean density is 200 times the mean density of the universe.

[8]  $R_A = 1.7\ \mathrm{arcmin}/z \sim 2.14\ h_{70}^{-1}\ \mathrm{Mpc}$.

**Figure 1.** Example of a 2D Voronoi tessellation applied on a detected VT-MLE structure. Each point represents a member galaxy and the dashed lines the Voronoi tessellation. The polygon is the convex hull of the cluster. This figure shows the relation between cell size and local density, because the cells get smaller in the denser central region. The actual algorithm uses a 3D tessellation.

developed by Allard & Fraley (1997) for a mixture of uniform point processes, for the detection of galaxy groupings in the Galaxy distribution; its output provides only starting values of cluster centers. The second stage (Gapper$R_{200}$), inspired on the ZHG technique, selects member galaxies applying a specific velocity gap. All galaxies within $0.5\,h^{-1}$ Mpc and $4500\,\mathrm{km\,s^{-1}}$ of a starting cluster center are the initial member galaxies, while an iterative process computes $R_{200}$ and the velocity dispersion through the expressions of Finn et al. (2008) and Beers et al. (1990) respectively, until convergence is found. Otherwise, it is the starting membership that is adopted.

The first stage of VoML+G is purely geometrical. It starts by dividing the whole redshift space into Voronoi cells. A Voronoi cell is a uniquely defined convex region of space that contains all points in space closer to a particular galaxy position than to any other galaxy position. Each Voronoi cell is associated with a galaxy position, with the smaller cells lying in the regions of highest point density. Next, the resulting cells are ordered according to their volume, from smallest to largest, producing a master list that establishes the "order of priority" of the galaxies to be used by the code. Schematically, the formation of structures starts with the selection of the first cell in the master list and identification of all its surrounding cells. The cells are considered consecutively, in an order determined by their priority, to test through MLE if they are to be joined with the starting cell to build a "structure." The construction of the first structure is completed when the MLE refuses the addition of a cell. The code then proceeds to build another structure by selecting the free cell of the highest priority in the master list. We classify the initial cell (or corresponding galaxy) of any completed structure as the "seed cell" (or "seed galaxy"). The output structures are saved in a list, in the order of their completion. As an example, Figure 1 illustrates in 2D the VT of one of the structures built in the VT-MLE stage of the algorithm, where the polygon is the convex hull[9] of the

---

[9] The convex hull of a set of points S is the intersection of all convex sets that contain S. It is represented by a convex polygon (2D) or polyhedron (3D).

cluster. In our 3D case, points in the convex hull or very close to the border have associated cells that span beyond the convex hull, even with infinite volume. The VT-MLE algorithm discards those galaxies in the database that are associated with cells that have any portion outside the convex hull, because the volumes of these cells provide poor estimations of the densities in the vicinity of these points.

The centroids of the structures generated in the VT-MLE stage constitute the input for Gapper$R_{200}$. The number of structures detected in the first stage is the upper limit to the number of clusters that Gapper$R_{200}$, and thus the algorithm, can identify. The process starts with the selection of the first centroid in the list, followed by the selection of the initial members of the cluster through a "gapping velocity filter" over the galaxies within a cylinder oriented along the line of sight, $0.5\,h^{-1}$ Mpc in radius, the axis of which passes through the input centroid. An iterative process ends in either of two cluster detection cases: *case a*, where member galaxies are defined within a section of radius $R_{200}$; or *case b*, where member galaxies are defined within the initial cylinder with a section of radius $0.5\,h^{-1}$ Mpc.

VoML+G contains three parameters. One of them belongs to the VT-MLE stage: (i) the domain size around a Voronoi cluster seed, over which a mean density is computed, which is a necessary value for the MLE calculation. The other two belong to the Gapper$R_{200}$ stage: (ii) the value of the velocity gap; and (iii) the adopted radius ($0.5\,h^{-1}$ Mpc) for the section of the cylinder over which the selection of cluster galaxies is performed. Section 4 describes the procedure used to adopt the above parameters.

The algorithm, as applied to the mock 2dFGRS, also has two built-in constraints. One limits the growth of the clusters in the VT-MLE stage to a maximum size of $2\,R_A$, and the other prohibits galaxies from belonging to more than one cluster. In actuality, the MLE condition generally stops cluster growth before a size of $2\,R_A$ is attained. Section 2.2 evaluates how the second constraint affects cluster detections in the Gapper$R_{200}$ stage.

Notice that no minimum size limit is imposed for the cluster detections. Thus, small groups can also be detected; however, their analysis is outside the scientific scope of this work.

### 2.1. The VT-MLE Stage—The Determination of Start-up Cluster Centroids

Allard & Fraley (1997) implemented the VT-MLE basic method for a space that is of isometric nature, and therefore without preferred directions. This is not true for the redshift space we use, where distortions along the line of sight, like the Finger-of-God effect, are known to exist. Nevertheless, we do not introduce corrections to counteract distortions because they would imply some cluster parameterization. As a consequence, some fragmentation of the "mock clusters" are expected in this stage of our algorithm.

A feature of the MLE structure detection is its sensitivity to the diminishing galaxy number density, in the sample, with increasing redshift. This happens when the mean galaxy density, required in the determination of galaxy overdensities in the MLE, is computed over a "local sphere" centered on the seed galaxy of the structure under construction. The radius of the "local sphere" is adopted in the optimization stage (Section 4).

### 2.1.1. The MLE Applied to the 3D Voronoi Tessellation

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 neighbor 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.

The volumes of the Voronoi cells directly determine the density distribution because $\rho_i = V_i^{-1}$, where $\rho$ and $V_i$ are respectively the Galaxy number density and the volume of a cell associated with an object $i$. The simplest approach to locate the density peaks is to define a density contrast with respect to the background. The density contrast, $\delta_i$, at the position of the $i$th object is defined as

$$\delta_i = (\rho_i - \bar{\rho})/\bar{\rho} \qquad (1)$$

$$\bar{\rho} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_i}, \qquad (2)$$

where $V_i$ 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 has been proposed and applied for a 2D space by Allard & Fraley (1997) and extended to 3D by Pizarro et al. (2006). The density associated with a point $x \in K$ is

$$f(x) = \frac{p}{|V|}1_V(x) + 1 - p, \qquad (3)$$

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

$$\mathcal{L}(x; V, p) = \prod_{i=1}^{n}[f(x_i)] = \left(\frac{p}{|V|} + 1 - p\right)^{N_V}(1 - p)^{n - N_V}, \qquad (4)$$

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 expression has to be reduced to a partial maximized (profile) likelihood. For a fixed $V$, the MLE of $p$ is

$$\hat{p} = (N_V - |V|n)/(n - |V|n) \qquad (5)$$

and the partial maximized likelihood becomes

$$L(x; V) = \left(\frac{1}{n}\right)^n\left(\frac{N_V}{|V|}\right)^{N_V}\left(\frac{n - N_V}{1 - |V|}\right)^{n - N_V}, \qquad (6)$$

which can be more conveniently expressed as a log-likelihood:

$$l(x; V) = -n\ln n + N_V\ln\left(\frac{N_V}{|V|}\right)$$
$$+ (n - N_V)\ln\left(\frac{n - N_V}{1 - |V|}\right), \qquad (7)$$

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

Next, the estimator $\hat{V}$ for $V$ is computed; $\hat{V}$ 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 we compute the variation in likelihood of the closure after it is merged. Ultimately, the merging corresponding to the maximum likelihood closure is selected. After all cells are merged into $\hat{V}$, the stage at which the log likelihood is maximized is chosen as the approximate MLE.

### 2.1.2. The Generation of Structures

The input for the algorithm is a set $P$ of $n$ points in a 3D space, where each point represents a galaxy. Before applying the structure detection algorithm over $P$, the Voronoi tessellation $T$ and the Delaunay mesh $D$[10] of the set of points $P$ are computed using *qhull* (Barber et al. 1996), a public software for computing the convex hull, the Delaunay triangulation, and the Voronoi diagram of the set of points. The Voronoi tessellation $T$ gives an estimation of the local point density associated with each point $p_i \in P$ as the reciprocal volume of the Voronoi cell $t_i \in T$. Conversely, the vertices of the corresponding Delaunay mesh $D$ form an adjacency graph for the cells of $T$, in such a way that for each cell $t_i \in T$ associated with the point $p_i$, its neighboring set $N(p_i)$ can be easily obtained.

At the start, the algorithm selects the smallest volume cell in the *whole* Voronoi Tessellation. The partial maximized likelihood ($L$) is first computed for this seed, and then subsequently calculated for the union of this seed and each of its neighbors; while the value of the MLE value increases by the addition of a neighbor, the merging into the structure is executed. The process continues until one of the following situations is encountered: the MLE value does not increase when a neighboring cell is tested for inclusion, in which case the information on the structure is stored; alternately, if a neighboring cell that passes the test for inclusion into a structure happens to be already a member of a previously detected structure, the build-up of the structure is aborted because of the non-overlap condition, and the cells involved are liberated for later consideration by the algorithm (i.e., they could potentially become part of another cluster). The third possible scenario is that the extent of the structure equals or exceeds 2 $R_A$; then, the process is stopped and the structure information is stored.

Once the above process is completed, the algorithm starts a new iteration by selecting the next "free cell" in the master list. The algorithm detects as many structures as possible, *without overlap*, and finishes when all the available seeds have been considered. Note that, because the algorithm proceeds

---

[10] The Delaunay tessellation is a dual graph of the Voronoi tessellation. It is produced when each pair of input points is connected by a straight line.

sequentially over all potential seeds, all the galaxies that comply with the rules will end up in a structure.

### 2.2. The GapperR$_{200}$ Stage

The process starts by considering the centroids of the structures found in the VT-MLE stage in the same order they were generated.

The complete set of steps is the following:

1. A starting centroid value is considered; all galaxies within $0.5\,h^{-1}$ Mpc (this is one of the parameters of the algorithm) in projected distance and separated from each other by less than $1000\,\mathrm{km\,s^{-1}}$ (velocity-gap parameter, see Section 4.2) are selected, and their centroid determined. For best results, by comparison with the mock clusters, the process is repeated two more times and the resulting centroid is adopted as the cluster center for use in the following steps.

2. Using the cluster center (R.A., Decl., $z$), a fixed redshift cut is applied, removing all objects more than $4500\,\mathrm{km\,s^{-1}}$ from the cluster redshift estimate.

3. Following (Biviano et al. 2006, hereafter BI), all galaxies within $0.5\,h^{-1}$ Mpc and $4500\,\mathrm{km\,s^{-1}}$ of the cluster starting center are selected. This is the starting membership.

4. The member galaxies are ordered by their velocity and any gap between them of width $1000\,\mathrm{km\,s^{-1}}$ or larger is identified. Galaxies separated by these gaps from the main system are excised. These enclosed galaxies are defined as the starting cluster core.

5. The velocity dispersion $S_{\mathrm{BI}}$ and central redshift $C_{\mathrm{BI}}$ are calculated using the biweight estimator, following Beers et al. (1990). Similarly to step 4, we excise galaxies separated from the main system by gaps of width $S_{\mathrm{BI}}$ or larger.

6. The maximum and minimum redshifts of the ensemble of remaining galaxies are recorded ($z_{\mathrm{min}}$ and $z_{\mathrm{max}}$). We adopt these values to be the cluster members redshift limits for the remainder of the process.

7. The velocity dispersion and central redshift ($S_{\mathrm{BI}}$, $C_{\mathrm{BI}}$) are recalculated, along with the cluster radius $R_{200}$ following Finn et al. (2008).

$$R_{200} = \frac{2.02 \times S_{\mathrm{BI}}}{1000\,\mathrm{km\,s^{-1}}} \frac{h_{70}^{-1}\mathrm{Mpc}}{\sqrt{\Omega_\Lambda + \Omega_M \times (1+z)^3}} \qquad (8)$$

8. All the galaxies within $R_{200}$ and between $z_{\mathrm{min}}$ and $z_{\mathrm{max}}$ are selected for the recalculation of the velocity dispersion $S_{\mathrm{BI}}$ and $R_{200}$.

9. The previous step is repeated until $S_{\mathrm{BI}}$ and $R_{200}$ stop changing. In the case that successive $S_{\mathrm{BI}}$ are equal, the mean value of the previous $S_{\mathrm{BI}}$s is used instead, in order to avoid an algorithm loop. Because the number of possible sets is finite and we are avoiding loops, this step is guaranteed to converge.

10. A resulting cluster with two or more member galaxies is a *case a*—or "$R_{200}$"—detection. A resulting cluster with a single galaxy is discarded, unless its starting core (step 4) had at least five galaxies, in which case it is a *case b*—or "$0.5\,h^{-1}$ Mpc"—detection.

11. The detection, either *case a* or *case b*, is checked to determine whether it has galaxies in common with a previously registered cluster. A new valid cluster, i.e., its member galaxies, is recorded if the non-overlap condition is verified, and the final center is computed.

12. The next centroid from the VT-MLE output is loaded for consideration in step 1.

Both *case a* and *case b* detections constitute valid VoML+G clusters; thus, the completeness and purity rates will be evaluated for the ensemble of detections, and future applications will not discriminate between them. Nevertheless, in some of the analyses that follow, we decompose the properties and performance of the ensemble into the two detection cases. After all, they represent different solutions within the GapperR$_{200}$ stage; only *case a* detections are confined within $R_{200}$, and their properties should be contrasted.

## 3. Mock 2dFGRS Data

Data from the Millennium Simulation of the LCDM cosmogony is used to produce a mock 2dF database that is employed for the optimization and validation process of the algorithm. The Millennium Simulation contains $2160^3$ dark matter (DM) 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 in 63 snapshots, spaced logarithmically from $z = 20$ to $z = 0$.

Cluster parent DM halo have been identified by the friends-of-friends algorithm (with a linking length of 0.2 units of the mean particle separation), and so it is neither necessarily spherical nor precisely of radius $R_{200}$. In addition to a virialized population of galaxies, the parent DM halos may contain galaxies moving in and out, as well as others falling in for the first time. This particular link between galaxies and DM halos corresponding to clusters is quite different from the observational definition of a cluster based on an actual galaxy redshift survey; therefore, very tight relations between halo and galaxy cluster properties are not expected.

The galaxy data set used in this work was taken from all sky mock catalogs of Blaizot et al. (2005), based on the Millennium Simulation (Springel et al. 2005; Lemson & Virgo Consortium 2006), which are limited at an apparent AB magnitude of 18 in the r-band filter from SDSS, and they include apparent magnitudes in the eight optical filters from both SDSS and 2MASS (Kleinmann et al. 1994). The equivalent $b_J$ magnitudes were determined as $b_J = g + 0.150 + 0.130(g - r)$ (Colless et al. 2003). A strip of sky was extracted from the first catalog (table Blaizot2006_AllSky_RT_1 in the simulations), covering 1194 deg$^2$ spanning 90 deg in right ascension and 15 deg in declination, approximately matching the Southern strip of the 2dF galaxy redshift survey, with a lower limit of $z = 0.009$ (appropriate for the future application of VoML+G to the actual 2dFGRS). Figure 2 shows a wedge diagram corresponding to the selected galaxies where the filaments/voids can be readily appreciated.

Following the approach of Milkeraitis et al. (2010), we define optical clusters in the Millennium Simulation as the collections of at least five galaxies from the Blaizot et al. (2005) mock-2dFGRS catalog that belong to the same DM parent halo. The adopted galaxy data set contains 156,494 galaxies, with 1850 clusters identified to $z \sim 0.22$ (hereafter, the reference galaxy data set and cluster reference catalog,
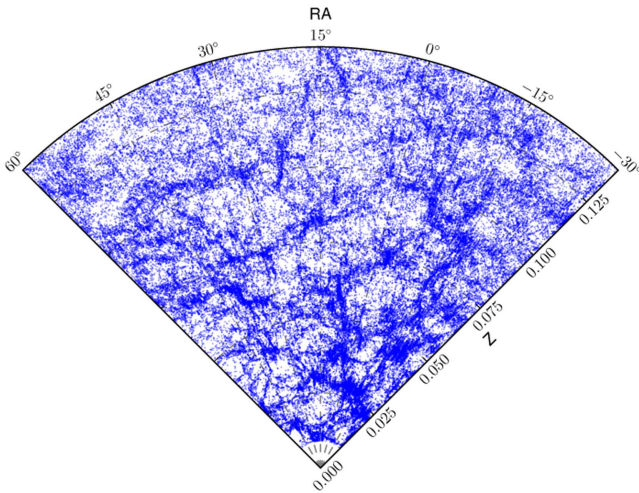
**Figure 2.** Wedge diagram showing the mock 2dF database, restricted to $z < 0.14$, based on the Millennium Simulation

respectively). Halo cluster masses and radii are obtained from the Millennium catalogs for the 1850 clusters.

Using the biweight estimator (Beers et al. 1990), line-of-sight velocity dispersions were calculated for every reference cluster. Figure 3 shows a plot of the velocity dispersions against redshift for clusters with $N_g \geqslant 10$, where crosses and open circles represent, respectively, reference clusters detected and non-detected by the VoML+G algorithm. It can be seen that clusters with $z > 0.10$ have, on average, larger values of velocity dispersion in comparison to $z < 0.10$, and that VoML+G is effective in finding reference clusters to $z \sim 0.15$.

Based on information from the mock cluster catalog, Figure 4 shows a plot of the masses of cluster parent DM halos against the line-of-sight velocity dispersions ($>300 \, \mathrm{km \, s^{-1}}$) of their corresponding optical clusters, where a mass-$\sigma_{cz}$ relationship is apparent. The red line in the plot was calculated from the mass-$\sigma_{cz}$ equation determined by Biviano et al. (2006) in a study of 62 clusters extracted from a LCDM cosmological hydrodynamical simulation

$$ M_{\mathrm{virial}} \equiv A \left( \frac{\sigma_v}{10^3 \, \mathrm{km \, s^{-1}}} \right)^3 \times 10^{14} \, h^{-1} \, M_\odot, \qquad (9) $$

where $A = 1.50 \pm 0.02$. The consistency between the mass-$\sigma_{cz}$ relation determined for the reference Millennium clusters with the equation by Biviano et al. (2006), corresponding to independent samples, is reassuring.

### 3.1. Non-overlap Condition and Detections

We perform a test to investigate the effects, in the GapperR$_{200}$ stage, of not allowing galaxy overlaps between clusters. Instead of using the output centroids from the VT-MLE stage, the first step of GapperR$_{200}$ is loaded with the actual centers of the "mock clusters," i.e., of the Millennium reference clusters. When GapperR$_{200}$ is run on the mock 2dFGRS data using these "ideal conditions," no overlaps between clusters occur. Thus, the Millennium clusters detected by GapperR$_{200}$ are sufficiently separated in velocity space to prevent overlaps from occurring. This result, along with the performance of the VoML+G code to identify the reference clusters (see Section 7), are used later in Section 8 to discuss the consequences of the non-overlap condition.
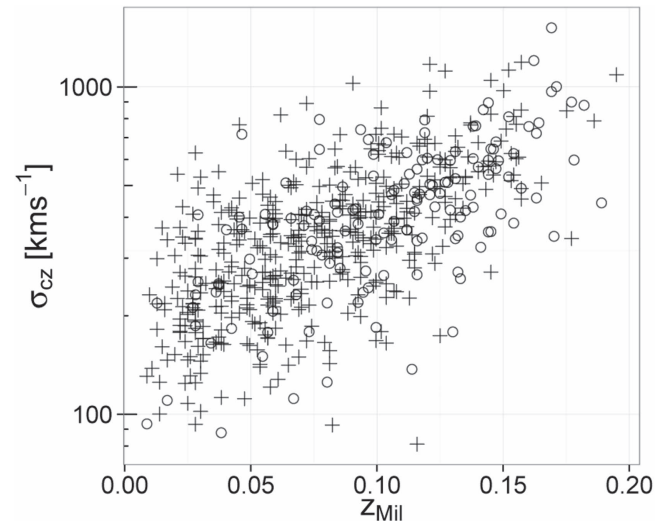


**Figure 3.** $N_g \geqslant 10$ Millennium clusters $z$ vs. velocity dispersion. Information on which of these Millennium clusters are actually detected by VoML+G with its final parameters is coded: empty circles correspond to non-detections, and crosses to detections.
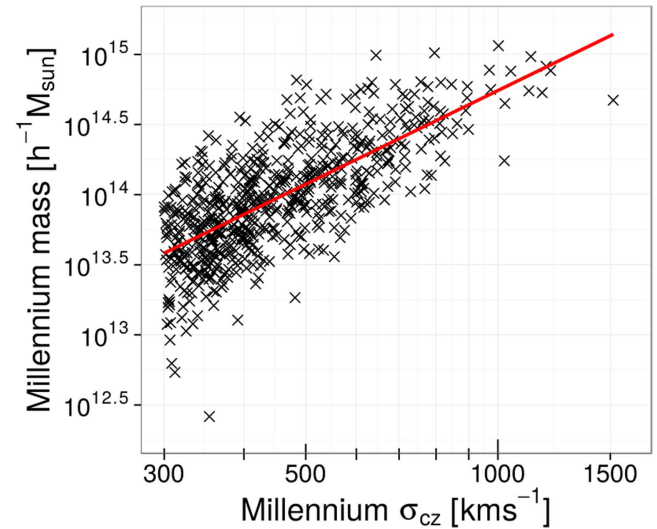


**Figure 4.** Sigma-cz vs. halo mass relation for the Millennium clusters with $\sigma_{cz} > 300$. Black crosses are the data points. The red line corresponds to Equation (9).

### 4. Algorithm Parameters

The radius of the cylinder over which GapperR$_{200}$ selects the initial members of a cluster was already adopted to be $0.5 \, h^{-1}$ Mpc, i.e., equal to $(1/3) \, R_A$. The determination of the two remaining parameters, the domain size and velocity-gap, is presented in this section. VoML+G is run on the mock 2dFGRS data using different sets of parameters and their performance evaluated by comparing their output with the Millennium reference clusters catalog.

Completeness is defined as the percentage of clusters in the reference catalog (i.e., dark matter halos) that have a counterpart in the VoML+G catalog, and purity as the percentage of VoML+G clusters that have a correspondence in the reference catalog. A cluster in one catalog is defined to have a counterpart in another catalog when they have a minimum fraction of galaxy members in common, whose parameter, the minimum fraction of galaxies in common ($f_c$), is equal to 0.25.
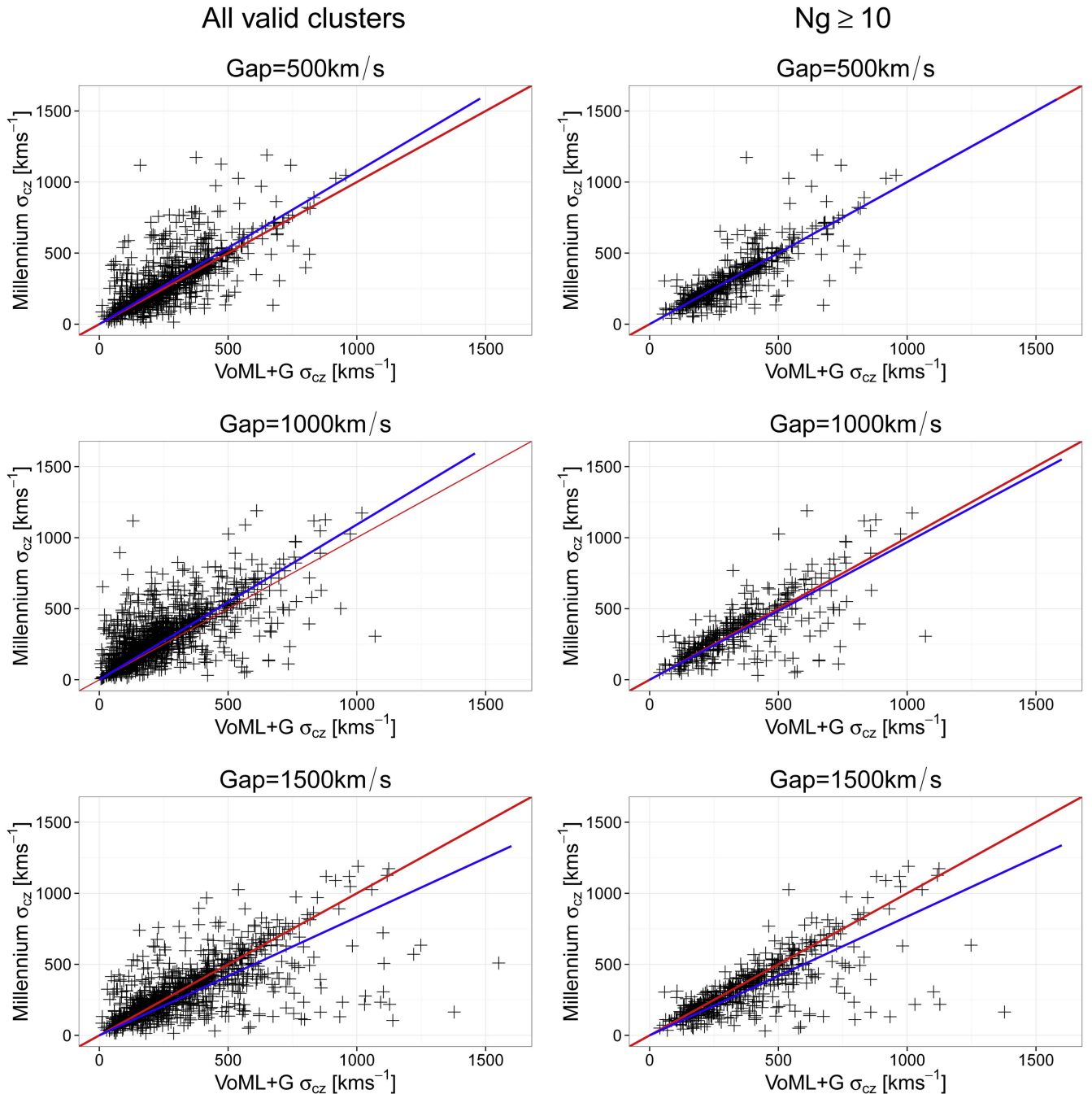
**Figure 5.** VoML+G vs. Millennium reference catalog velocity dispersion for different values of the velocity gap. Top, middle and bottom rows have values of 500, 1000, and 1500 km s$^{-1}$, respectively. The red lines are lines of slope 1 passing through the origin. Blue lines are linear regressions. The left figures include all valid VoML+G clusters with Millennium correspondence whereas those on the right include only the ones with 10 or more galaxies.

In this section, we show that the optimization ensures that VoML +G produces a robust automatic registration of the 2dFGRS mock reference clusters. In particular, we find that the completeness and purity of the detected clusters with $N_g \geqslant 10$ are $\sim$75% and $\sim$90% respectively, and valid across the $1.0 \times 10^{12}$–$1.0 \times 10^{14.5}\, M_\odot\, h^{-1}$ mass range.

### 4.1. Domain Size for Density Calculation

After a series of runs of the algorithm, the optimal domain size for calculating the local density for use in the generation of structures in the VT-MLE stage is found to be 35 $h_{70}^{-1}$ Mpc.

Such a radius is large enough to average over the large-scale structure (LSS) features, but small enough to be sensitive to extended LSS trends—providing some locally adaptive behavior of the algorithm.

### 4.2. Velocity Gap

Previous work by DP02 and Pizarro et al. (2006) has shown that the adoption of a 1000 km s$^{-1}$ velocity gap interval, when applied to the "real 2dFGRS" galaxy database, reproduces (with good fidelity) the Abell-like clusters contained in the surveyed volume. In this section, we analyze the output of the

**Table 1**
Summary of the Velocity-gap Optimization Results

| Gap (km s$^{-1}$) | Slope (Ng $\geqslant$ 2) | Std. Dev. (km s$^{-1}$) (Ng $\geqslant$ 2) | Slope (Ng $\geqslant$ 10) | Std. Dev. (km s$^{-1}$) (Ng $\geqslant$ 10) |
|---|---|---|---|---|
| 500 | 1.070 | 136 | 1.023 | 111 |
| 1000 | 0.907 | 142 | 0.907 | 148 |
| 1500 | 0.803 | 185 | 0.803 | 178 |

VoML+G code when run with three values of the velocity gap: 500, 1000, and 1500 km s$^{-1}$, and confirm that the 1000 km s$^{-1}$ value produces satisfactory results with the mock 2dFGRS database. Once the pairing between the VoML+G and reference clusters has been established, their performance is evaluated by comparing the velocity dispersions, completeness rates, and purity rates.

Plots of the velocity dispersions resulting from the VoML+G runs versus the values for corresponding reference clusters are shown Figure 5, where the left column includes all cluster pairs and the right one includes only pairs with at least 10 galaxies. The plots, from top to bottom, represent the output using 500, 1000, and 1500 km s$^{-1}$ gap values, where the red and blue lines correspond to slopes of value 1 and linear regressions, respectively. Considering only the right column for our analysis, because the velocity dispersions are very uncertain for clusters with less than 10 galaxies, it is seen that the 500 km s$^{-1}$ gap produces the tighter relation followed by the 1000 km s$^{-1}$ gap. Table 1 summarizes the slope values and standard errors for a slope of 1 for each velocity gap value.

The completeness rates are found to be dependent on the number of galaxies (multiplicity) of the reference clusters. Figure 6 shows the completeness versus the minimum number of galaxies in Millennium clusters where it can be seen that, for $\mathrm{Ng_{Mil}} \lesssim 15$, the lowest performance is for a 500 km s$^{-1}$ gap, although it improves for higher multiplicities. A velocity gap of 1000 km s$^{-1}$ is shown to produce recovery rates larger than 75% for clusters with $\mathrm{Ng_{Mil}} \geqslant 10$, although the gap of 1500 km s$^{-1}$ outperforms it by ~10%. It is important to note that the number of clusters with $\mathrm{Ng_{Mil}} > 30$ is relatively small; therefore, the recovery rates are more uncertain.

The purity rates in Figure 7 are above 90% for $\mathrm{Ng_{Mil}} \geqslant 10$ and all velocity-gaps. Notably, the purity rate is considerably higher (>95%) for clusters with $\mathrm{Ng_{Mil}} > 25$ and a velocity-gap of 500 km s$^{-1}$.

The 1500 km s$^{-1}$ velocity gap value is discarded because it has a considerable larger dispersion (see Table 1) than the dispersion for the other gap values. The completeness rates for the 500 and 1000 km s$^{-1}$ gap values are nearly identical with the exception for $\mathrm{Ng_{Mil}} \geqslant 40$. Although the purity rate for the 1000 km s$^{-1}$ gap is above 90% for $\mathrm{Ng_{Mil}} \geqslant 10$, it is considerably exceeded by the results for the 500 km s$^{-1}$ gap.

We conclude that a velocity-gap of 500 km s$^{-1}$, when applied to the mock 2dFGRS, produces a cluster set marginally better than the one corresponding to 1000 km s$^{-1}$. Given this marginal difference, we have adopted the 1000 km s$^{-1}$ velocity-gap value for the application of VoML+G on the mock 2dFGRS because this value was shown by DP02 to be optimal for the application of the ZHG technique to the "real 2dFGRS."
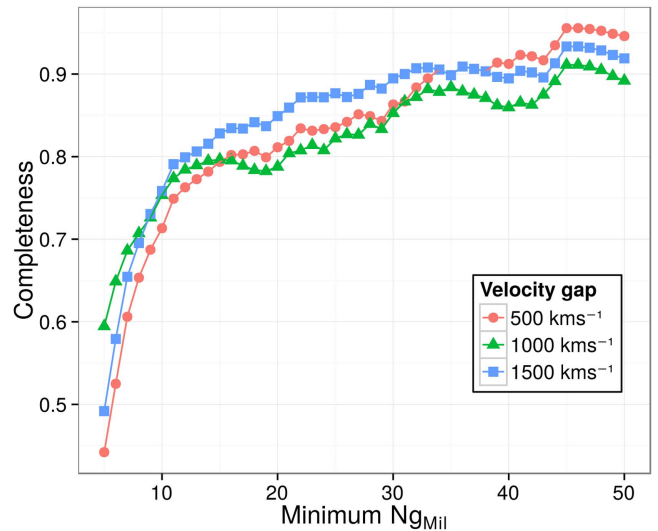


**Figure 6.** VoML+G completeness rate (fraction of reference clusters recovered) by minimum number of galaxies in the Millennium cluster. The galaxy overlap is of at least 25%.
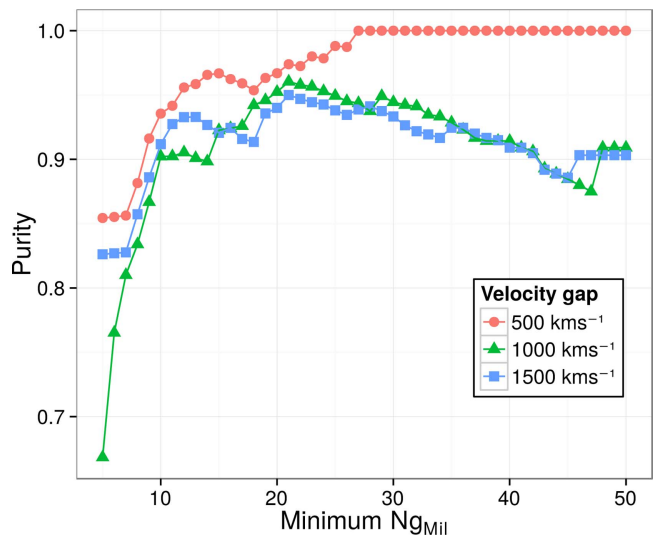


**Figure 7.** VoML+G purity rate (fraction of detected clusters that have a match in the reference catalog) by minimum number of galaxies. The galaxy overlap is at least 25%.

## 5. The VoML+G Cluster Catalog

The execution of VoML+G over the 2dFGRS mock data identified 1614 clusters with $N_g \geqslant 2$ and a median membership of six. Twenty-two percent of the clusters have $N_g \geqslant 10$, and thus are roughly equivalent to what is usually referred as galaxy clusters. Sixty-four percent of the VoML+G detections have intermediate multiplicity ($10 > N_g \geqslant 5$), and the remaining fourteen percent with less than five member galaxies are akin to what is normally referred as small groups of galaxies. Only 22 clusters have $N_g > 50$.

Figure 8 shows the histogram of member galaxies per VoML+G cluster. Solid lines are the total, dotted are *case a* detections, and dashed are *case b* detections. It can be seen that most of the clusters with low multiplicity are *case b* detections, whereas the *case a* detections dominate for galaxy memberships $N_g \geqslant 25$. Out of the 1614 clusters, 681 are *case a* detections and 933 are *case b*.
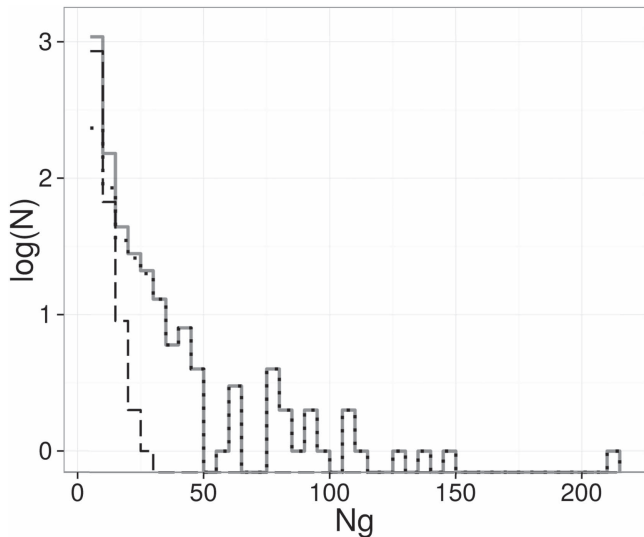
**Figure 8.** Distribution of number of galaxies per VoML+G cluster. Solid lines are the total, dotted are *case a* detections, and dashed are *case b* detections.
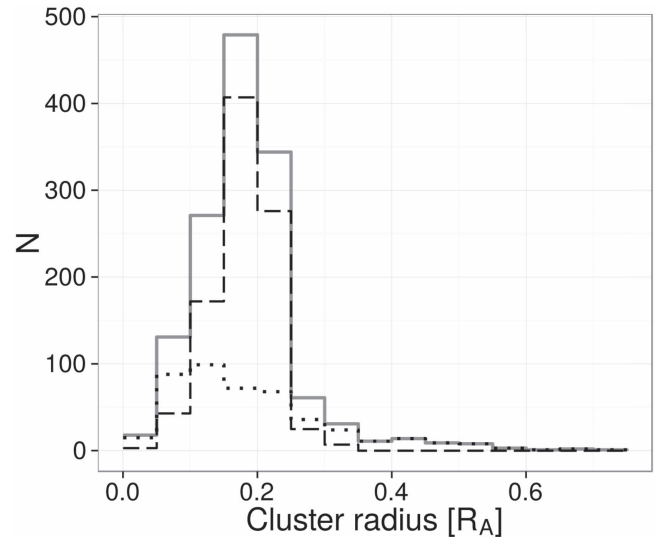


**Figure 9.** Distribution of projected radii of VoML+G clusters. Solid lines are the total, dotted are *case a* detections and dashed are *case b* detections.

The histogram of sizes in Figure 9 has a mean projected radius of 0.3 $R_A$; *case b* clusters reach sizes up to (1/3) $R_A$, consistent with their definition. Interestingly, the Gapper$R_{200}$ stage does not find clusters with sizes larger than 2 $R_A$, a result possibly due to the definition of the clusters within either radii of $R_{200}$ or $0.5\,h^{-1}$ Mpc.

### 5.1. VoML+G Versus Mock Clusters ($N_g \geqslant 10$)

A VoML+G detection is considered a valid identification of a Millennium reference cluster catalog if it contains at least 25% of its galaxies. Out of the 1614 VoML+G clusters, 1072 have counterparts, implying a nominal purity rate of 66% for VoML+G catalog. However, this rather low rate possibly underestimates the actual performance of the algorithm because it is unlikely that a detected cluster with two galaxies (a pair) will be in the reference cluster catalog, because the reference clusters have $N_g \geqslant 5$.

Figure 10 shows the redshift distribution of the Millennium reference clusters (gray line) and of the VoML+G clusters, the latter distinguished by detection type. The solid line is for all detections, the dotted line for *case a* and the dashed line for *case b*. The overall distributions are similar up to $z \sim 0.10$, but the VoML+G distribution drops somewhat after that limit, whereas the reference cluster distribution remains flatter up to $z \sim 0.15$, i.e., it becomes harder for VoML+G to recognize the Millennium clusters for $z \gtrsim 0.10$. Each detection type follows a similar trend up to $z \sim 0.15$; above that limit, the detections are almost exclusively *case b*. Because of this selection effect, and the small number of clusters, beyond $z \gtrsim 0.15$ we conservatively restrict our analyses to cluster samples to within $z < 0.14$.

The comparison between VoML+G detections and the Millennium reference clusters is done by multiplicity ranges, with the $N_g \geqslant 10$ sample specially monitored because of the intended applications of VoML+G. For $N_g \geqslant 10$, Figure 11 compares the multiplicities in log scale of the two cluster samples. Crosses are *case a* clusters and open circles are *case b*. The dispersion around the line of slope 1 is 0.21. The comparison of velocity dispersions for the $N_g \geqslant 10$ subsample
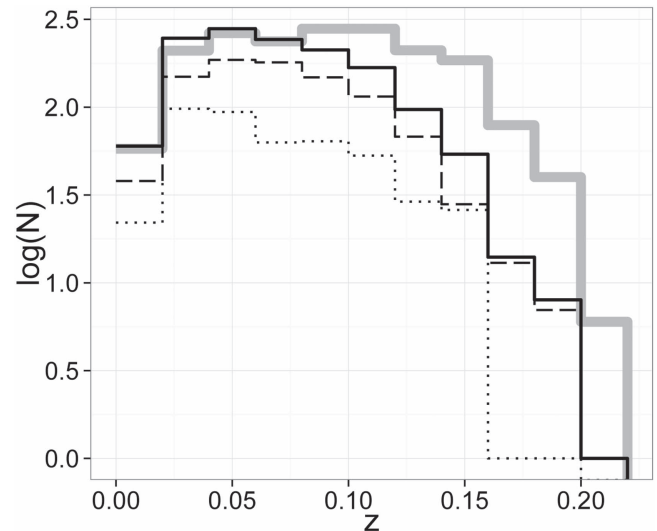


**Figure 10.** Millennium and VoML+G clusters. Millennium distribution in gray. For VoML+G, solid lines are the total, dotted are *case a* clusters and dashed lines are *case b* clusters.

is in Figure 5; the dispersion around the line of slope 1 is $154\,\mathrm{km\,s^{-1}}$. The VoML+G cluster velocity dispersions agree well with the corresponding values for the "true clusters," over the 300–1000 $\mathrm{km\,s^{-1}}$ interval, and can be used as proxies for cluster halo masses in excess of $3.0 \times 10^{13}\,M_\odot\,h^{-1}$.

### 5.2. Non-detections and False-positives

In this section, we examine the properties of the 2dFGRS mock clusters not detected by VoML+G, as well as the false positives, through histograms of number of member galaxies, velocity dispersion, redshift, and mass, which we display in a single figure (Figure 12).

The left column of Figure 12 shows the histograms for the non-detected 2dFGRS reference clusters (dotted line) superposed with the corresponding ones for the complete catalog of reference clusters (solid line). Both populations show roughly the same trend of membership up to $N_g \sim 100$ and the same
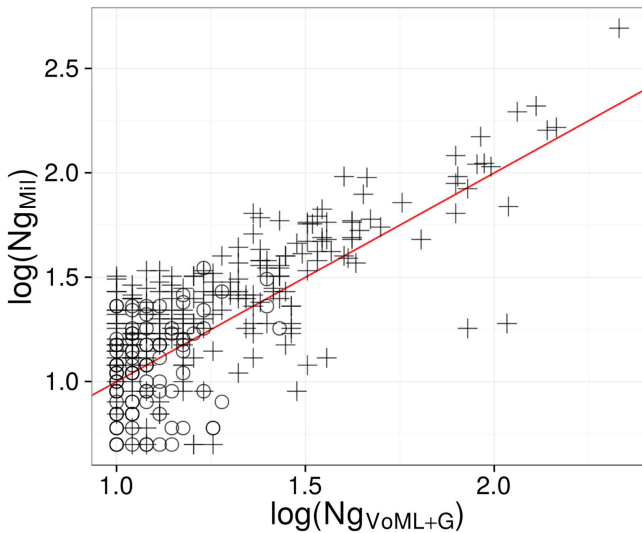
**Figure 11.** $N_g \geqslant 10$ VoML+G-Millennium clusters, multiplicity comparison. Crosses are *case a* clusters and open circles are *case b*. The dispersion around the line of slope 1 is 0.21.

trend of velocity dispersion up to $\sigma_{cz} \sim 1000\,\mathrm{km\,s}^{-1}$. In the redshift distribution, the proportion of missed clusters increases with redshift up to $z \sim 0.15$, principally due to the effects of diminishing galaxy density, survey magnitude limit, and fixed velocity-gap, whereas for $z \gtrsim 0.15$, most of the mock clusters are missed. In relation to the DM halo mass, the histograms for both populations show roughly the same trend up to $1.0 \times 10^{14.7}\,M_\odot\,h^{-1}$. Consequently, it appears that the non-detections by the VoML+G code should not significantly alter the information to be extracted from the sample of VoML+G clusters to $z \sim 0.15$ on multiplicity, velocity dispersion, and mass. The detection efficiency, however, is dependent on redshift, expected for a magnitude limited galaxy sample, and drops markedly at $z \sim 0.15$.

In the right column of Figure 12, the above mentioned histograms are shown for the VoML+G false-positive detections, with *case a* (dotted), *case b* (dashed), and total (solid) number of false-positives. All the quantities are calculated from the VoML+G cluster data only. First, we compare these histograms with the ones corresponding to the complete catalog of reference clusters (left column): (i) multiplicity: the histograms are very different, the false positives will contribute mostly clusters with $N_g \lesssim 10$; (ii) velocity dispersion: although uncertain because of the generally small multiplicity, the false-positives histogram is similar to the one for the reference clusters; (iii) redshift: the false-positive histogram has a negative slope compared to the flatter trend for the reference clusters, with a negligible contribution for $z > 0.15$, and thus will contribute proportionally more low-$z$ ($z \lesssim 0.08$) clusters; (iv) mass: although more uncertain than the velocity dispersions, the false-positive histogram is similar to the one for the reference clusters. Therefore, as a whole, the effect of the false-positives is a relative increase of the $N_g \lesssim 10$ clusters with respect to the $N_g > 10$ clusters, and they occur preferentially at $z < 0.08$. Now, with respect to the relative representation of *case a* and *b* detections on the false-positive clusters, they are dominated by the *case b* detections, i.e., generally by small clusters.

## 6. Display of the VoML+G Clusters

Figure 13 shows three examples of VoML+G cluster detections on the mock 2dFGRS database, where each panel displays up to 1 $R_A$ in extent measured from the cluster centroid, including all the galaxies within the $z$-interval of the VoML+G cluster. One cluster per row is displayed, where the left panel represents the intermediate VT-MLE structure and the right one represents the final VoML+G cluster. The yellow polygons represent the convex hull of a detected structure/cluster. Each point represents a galaxy: large, blue points are galaxies belonging to the detected cluster, and small black ones are galaxies within the defined cluster vicinity. Galaxies (points) belonging to the corresponding Millennium cluster are inside squares. The circles in the right panels have $R_{200}$ radii with centers at the cluster centroids. The top and middle rows are examples of *case a* detections, whereas the bottom row displays a *case b* detection. These three examples, all with $N_g \gtrsim 10$, clearly show a close galaxy match between the VoML+G detection and the respective Millennium cluster.

For VoML+G clusters with $N_g \gtrsim 10$, nearly 10% are false positives. In Figure 14, each of the four panels represents a VoML+G detection, displayed similarly to the previous figure. The yellow polygons represent the convex hull of a detected cluster. Each point represents a galaxy: large blue ones correspond to those belonging to the detected cluster, whereas small black ones are galaxies in the cluster vicinity. The horizontal bars in the bottom right corner of the panels are of size $0.5\,h^{-1}$ Mpc. The upper row corresponds to *case a* detections and the lower one to *case b*.

## 7. Completeness and Purity

We estimate the purity of the sample of detected clusters that could, in principle, be in the mock reference cluster catalog ($N_g \geqslant 5$). Table 2 lists the decomposition of the VoML+G detections by ranges of multiplicity and redshift. The number of detections are provided for two multiplicity groups: $N_g \geqslant 5$ and $N_g \geqslant 10$. The redshifts ranges include the whole detection interval (0.009–0.205), and the sub-intervals 0.009–0.050, 0.050–0.100, and 0.100–0.150. For the whole redshift range, the average purity of the $N_g \geqslant 5$ subgroup is 67%, whereas the purity is $\sim 90\%$ over the whole redshift range for the $N_g \geqslant 10$ subgroup.

The completeness and purity rates are shown in Figure 15 as a function of the minimum number of galaxies (Ng$_{\min}$) that a cluster must possess to be included in a set. Both rates increase with Ng$_{\min}$, up to Ng$_{\min} \sim 30$. The overall mean completeness is 59%, but increases to $\sim 75\%$ when only the $N_g \geqslant 10$ clusters are considered. The nominal overall mean purity is $\sim 66\%$, increasing to $\sim 90\%$ for clusters with $N_g \geqslant 10$.

The dependence of the completeness and purity rates on cluster halo mass is shown in Figure 16. The completeness fluctuates around 70%, with a peak at $\sim 1.0 \times 10^{13}\,M_\odot\,h^{-1}$, decreasing to $\sim 55\%$ for both the least and most massive clusters. The purity rate depends less on mass than the completeness does, fluctuating around 90%. It should be noted that there are few high-mass clusters in the Millennium simulations, and this affects the rate determinations in the upper halo mass range.

The completeness and purity rates for the $N_g \geqslant 10$ subsample as a function of cluster redshift are shown in Figure 17. Both rates are nearly coincident, at $\sim 80\%$, up to $z \sim 0.04$. For $z > 0.04$, the purity is an increasing function, approaching 100% at $z = 0.14$, whereas the completeness

**Non-detected clusters**
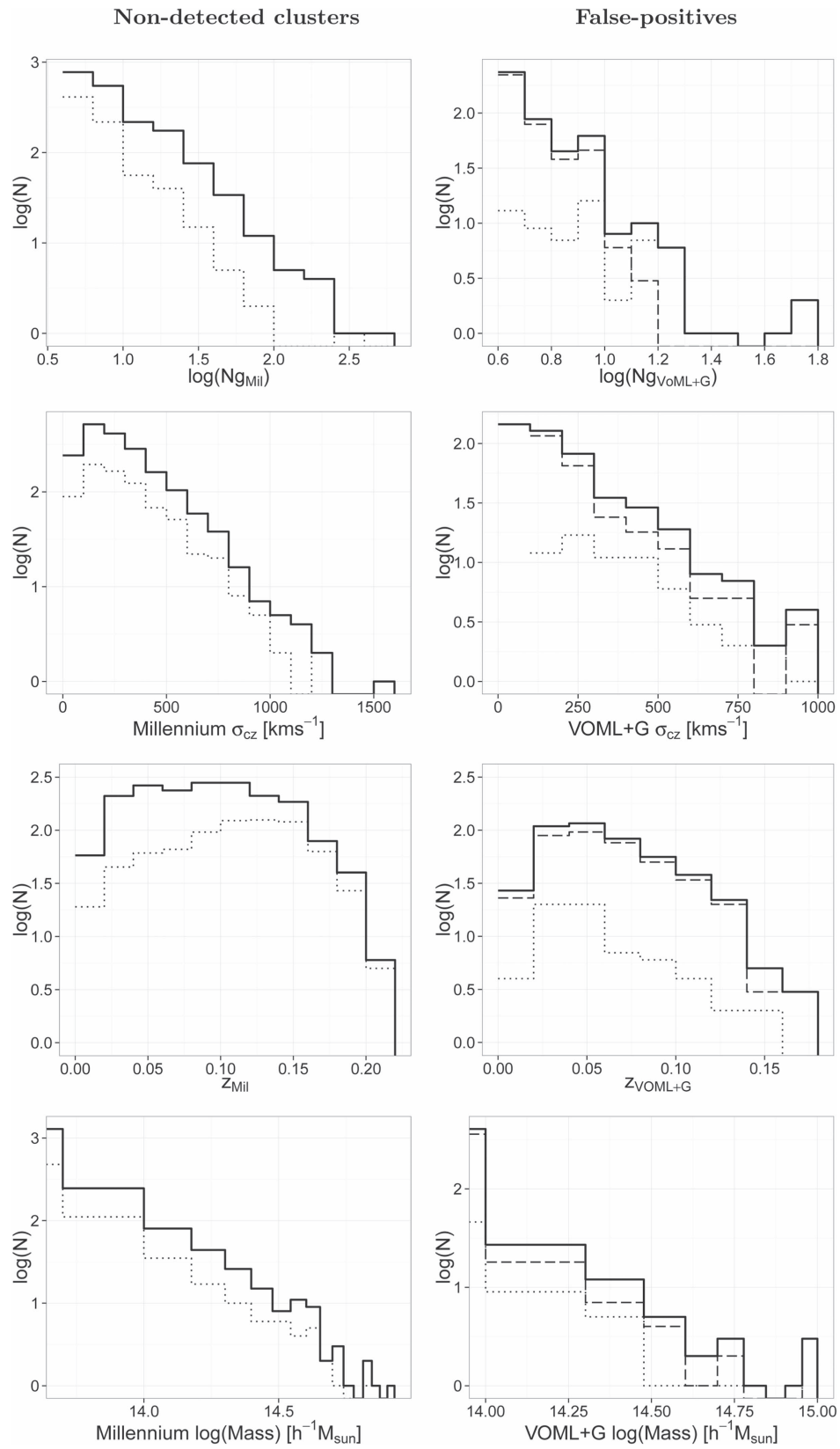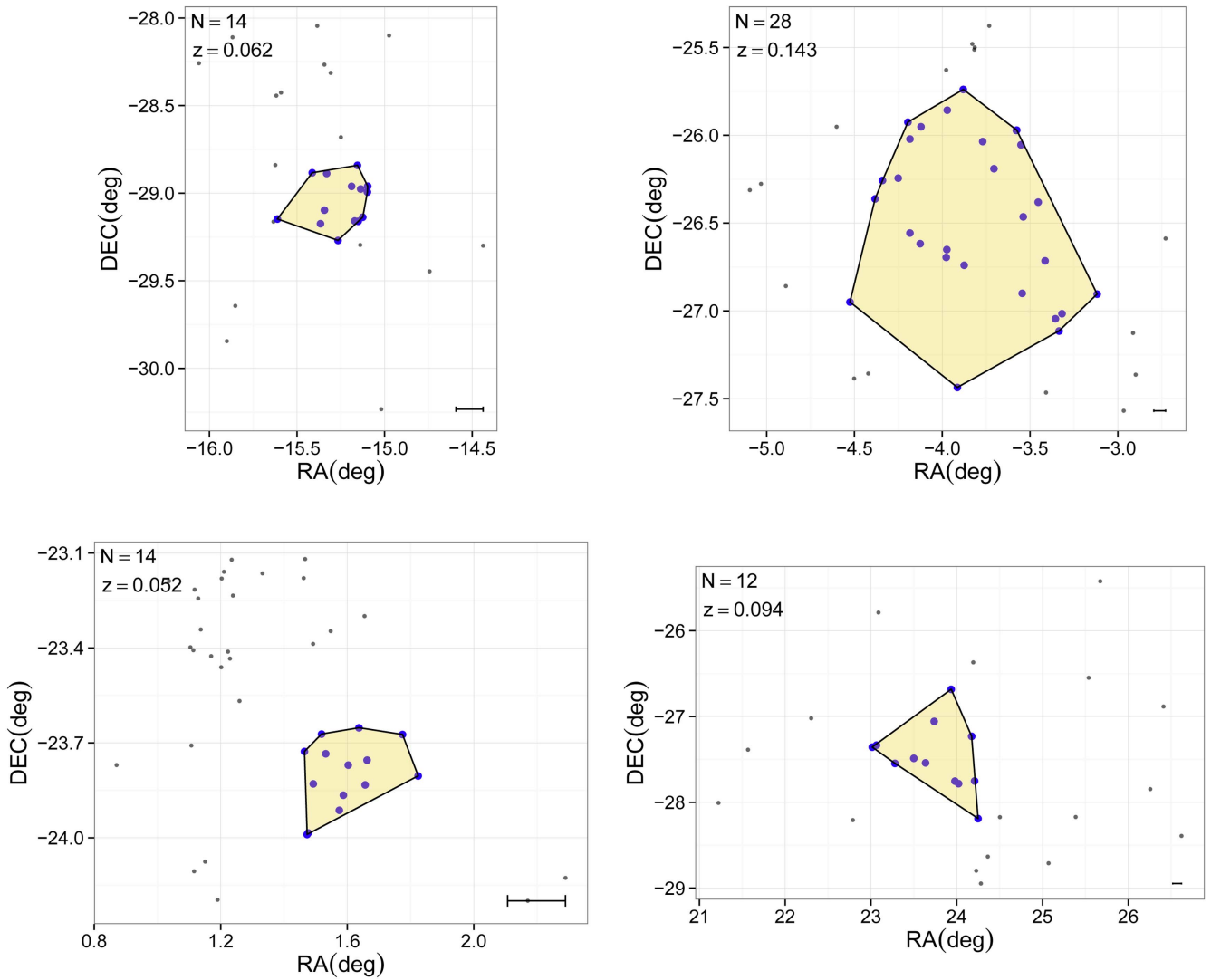
**False-positives**



**Figure 12.** Characteristics of the non-detected Millennium clusters and false-positive VoML+G clusters. Each column shows histograms of the number of member galaxies, velocity dispersion, redshift, and mass. The left one is for the non-detected reference clusters (dotted line) superposed with the histograms for the complete catalog of reference clusters (solid line). The right column corresponds to the false-positive VoML+G clusters, with *case a* (dotted), *case b* (dashed), and total (solid) number of false-positives.

**VT-MLE clusters**             **VoML+G clusters**
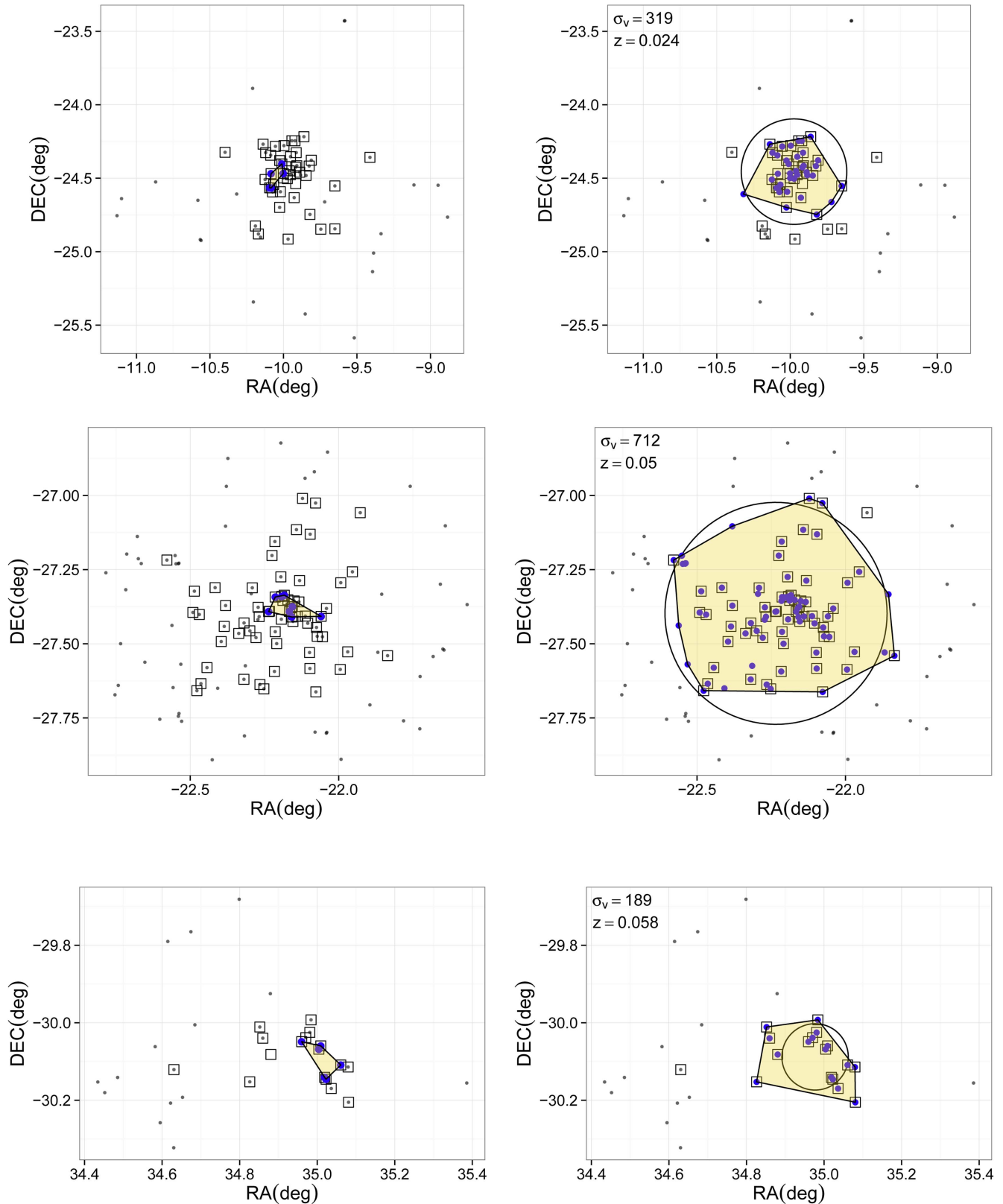


**Figure 13.** VoML+G valid detections: to the left, the VTM-LE cluster; to the right, the final VoML+G cluster. The yellow polygons represent the convex hull of a detected cluster. Each point represents a galaxy: large blue ones belong to a respective detected cluster, whereas small black ones are galaxies in the cluster vicinity. Galaxies that belong to a Millennium cluster are marked with squares. The circles in the right side plots are $R_{200}$ in radii and centered in the geometric centroid of the clusters. Top and middle rows are examples of *case a* detections, whereas the bottom row shows a *case b* detection.

12

**Figure 14.** False-positive VoML+G detections. The yellow polygons represent the convex hull of a detected cluster. Each point represents a galaxy: large blue ones correspond to those belonging to the detected cluster, whereas small black ones are galaxies in the cluster vicinity. Horizontal bars in the bottom right corner of the panels are $0.5\ h^{-1}$ Mpc in size. Top row plots correspond to *case a* detections and the bottom ones to *case b*.

**Table 2**
VoML+G Detections by Ranges of Multiplicity and Redshift

| z range | $N_g \geqslant 5$ | | $N_g \geqslant 10$ | |
|---|---|---|---|---|
| | 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% |

decreases strongly and reaches ~55% at $z = 0.14$. This behavior of the completeness rate reinforces the fact that the detection efficiency of VoML+G is dependent on redshift, as expected for a magnitude-limited redshift sample.
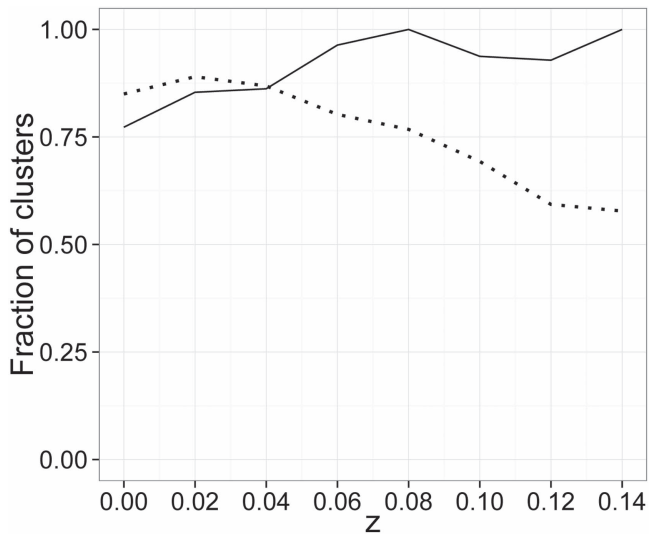
## 8. Discussion

The Millennium reference catalog includes clusters with a minimum number of halo members ($N_g = 5$), which does not
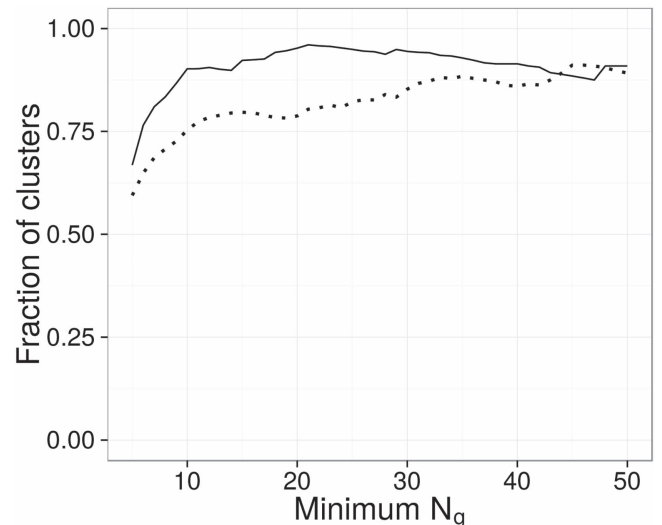
**Figure 15.** Millennium clusters completeness rate (dotted) and purity (solid) rate by cluster sets that have more than $N_{g\,min}$ galaxies.
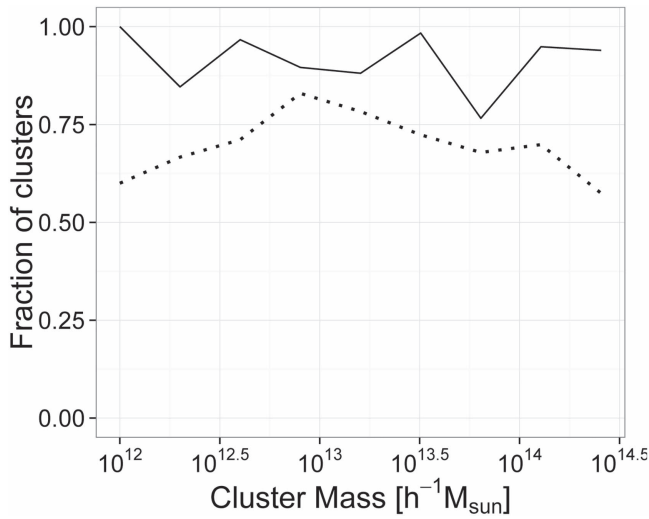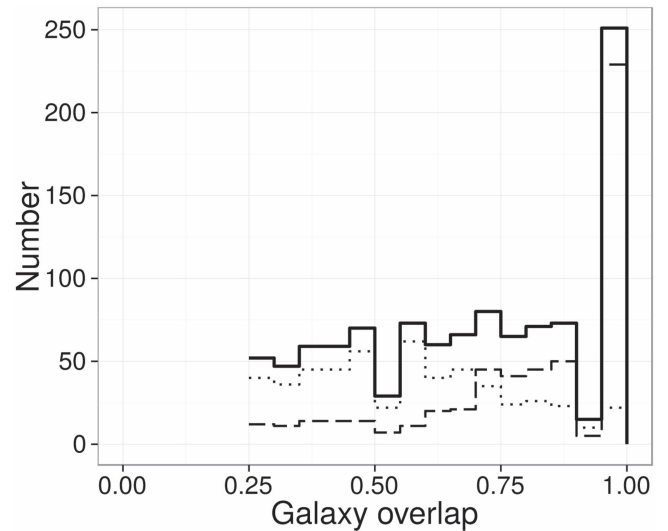
**Figure 16.** Millennium clusters completeness (dotted) and purity (solid) rate by cluster mass range.



**Figure 17.** $N_g \geqslant 10$ VoML+G clusters, completeness (dotted) and purity (solid) rate by cluster redshift range.



**Figure 18.** VoML+G clusters galaxy overlap histogram. Solid lines are the total, dotted are *case a* detections, and dashed are *case b* detections.

correspond to a precise mass cut. We have determined completeness limits in mass for subsamples of the reference catalog defined within $z < 0.0.5$, 0.10, and 0.14. The likelihood that a parent DM halo of a given mass will have five or more member galaxies detected in the mock-2dF catalog depends strongly on redshift, as increasing distance pushes more and more galaxies from the DM halo below the survey magnitude limit. For volume-limited samples, the reference cluster catalogs are 95% complete for DM halos above a minimum mass of $1.4 \times 10^{13} M_\odot\, h^{-1}$ within $z < 0.05$, $3.2 \times 10^{13} M_\odot\, h^{-1}$ for $z < 0.10$ and $7.0 \times 10^{13} M_\odot\, h^{-1}$ for $z < 0.14$ clusters. The distribution of parent halo DM masses corresponding respectively to the reference clusters within $z < 0.0.5$ and 0.14 are shown in Figure 20 as a function of the number of galaxies (Ng$_{Mil}$), where the dashed red lines mark the masses above which the subsamples are 95% complete.

The execution of VoML+G over the 2df Galaxy Redshift Survey mock data identified 1614 clusters containing two or more members within either $R_{200}$ or $0.5\, h^{-1}$ Mpc. Twenty-two percent of them, to $z \sim 0.14$, have $N_g \geqslant 10$, a subsample

shown to have high completeness (mean of ∼75%) and purity (mean of ∼90%) velocity dispersions that are good mass proxies, and few false positive clusters. These results provide confidence that a similar subsample, to be produced in a future application of VoML+G on the "real 2dFGRS," will mostly consist of physical entities.

Critical for the pairing of clusters between two catalogs is the value adopted for the minimum fraction of galaxies in common ($f_c$), in order to establish counterparts. The adoption of $f_c$ is arbitrary and its usual values are in the 0.3–0.5 range (e.g., Koester et al. 2007). Although we have adopted $f_c = 0.25$, ∼40% of the VoML+G clusters with counterparts in the mock reference catalog have $f_c \geqslant 0.8$ or larger, whereas ∼75% of these VoML+G clusters have $f_c \geqslant 0.5$, as can be appreciated in Figure 18. This matching method is expected to underestimate somewhat the purity of VoML+G because a fraction of the galaxies associated with a DM halo will not be physically within $R_{200}$.

Although the evaluation of the performance of VoML+G does not consider the effect of redshift errors on the mock galaxy sample, their role on the cluster recovery rate needs to be examined. As the 2dFGRS includes uncertainties of ∼85 km s$^{-1}$ or about 0.0003 in redshift (for objects above a minimum quality), we introduced a comparable amount of error on the mock 2dFGRS galaxy data, altering the redshift of each galaxy by a value taken from a normal distribution of $\sigma = 120$ km s$^{-1}$ ($\delta z = 0.0004$). The first stage of the algorithm (VT-MLE), which provides the centers for the cluster definition, was run with the the modified mock data. The detected intermediate clusters were compared with the VT-MLE results, in the absence of redshift errors. Figure 19 displays the cumulative distribution of the detected clusters with Millennium counterparts as a function of the distance between centers, expressed in units of Abell Radius. The cumulative distributions with and without errors are very close to each other. For Millennium cluster counterparts at distances less or equal to one Abell Radius, the effect of redshift error on the recovery fraction is of the order of ∼3%.

Another important issue is how sensitive the algorithm is to border effects. The algorithm filters the galaxies in the border of the Galaxy distribution, thus avoiding numerical divergences
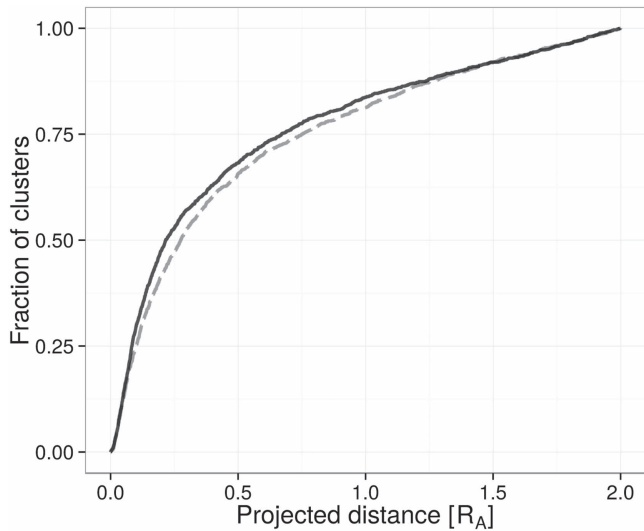
**Figure 19.** Cumulative distribution of the fraction of detected clusters with Millennium counterparts as a function of the distance between centers. The solid line tracks the unmodified mock galaxy data and the dashed line represents the data containing redshift errors.

in the VT-MLE stage. Nevertheless, the VoML+G detection capability of clusters located near the survey limits is inevitably affected. To estimate the incompleteness induced by the proximity to the border, we identified the mock clusters closer than 1 $R_A$ from the boundary. Out of the 158 mock clusters within 1 $R_A$ of the border, 67 of which have $N_g \geqslant 10$, VoML +G recovers 82 (52%) of them, and 47 (70%) of those with $N_g \geqslant 10$. The completeness of the detections in the border is reduced from the general value of 59% to 52%, and from the general value of 75% to 70% for the $N_g \geqslant 10$ clusters. These results are indicative that the border effects on the cluster finder are small.

In view of the anticipated application of the VoML+G algorithm to the "real 2dFGRS" data set, we tested the effect of a survey mask on the recovery of the mock clusters. We applied the 2dFGRS survey geometry mask (Cole et al. 2005, their Figure 3) to the mock galaxy catalog. This mask accounts for the geometry of the tiling of the survey and includes "holes" in the initial source catalog due to effect of bright stars on the UKST photographic plates, as well as plate flaws. The whole area within the main limits of the 2dFGRS was not actually sampled in the "real 2dFGRS," excluding the corners of the rectangle, so a masked rectangular region was used for our analysis with the mock galaxies within $z < 0.14$. The effect of the mask, which filters 2% of the rectangular area, was to lower the number of detections to 84% of the clusters found when the unmasked galaxy sample was used. For mock clusters with 10 or more member galaxies, 88% of the clusters corresponding to the unmasked sample were recovered. This difference in recovery rates suggests that higher multiplicity clusters are more resilient to the suppression of member galaxies by the mask. We conclude that the effect of a realistic mask is to degrade the VoML+G algorithm's detection of the $N_g \geqslant 10$ clusters by a factor $\sim 1/10$.

Cucciati et al. (2010) applied a Voronoi-Delaunay Method (Marinoni et al. 2002) on the VIMOS-VLT Deep-Redshift Survey (VVDS, $z = 0.2$–1.0) to identify galaxy groups. A proper comparison of their results with ours is not possible because of their small search area ($\sim 4\ deg^2$), different redshift

coverage, and low sampling rate. However, their reported completeness and purity rates of $\sim 60\%$ and $\sim 50\%$ provide interesting reference values. A different methodology (3D-MF) by Milkeraitis et al. (2010) employ cluster radial profiles and luminosity functions to search for clusters on the Canada–France–Hawaii Telescope Legacy Survey (photometric redshifts over $z = 0.2$–1.0). Again, no easy comparison is possible with our work, but their reported completeness rate of 54.8% ($M > 1.5 \times 10^{13}\ M_\odot$) and purity rate of 88.8% provide additional reference values. The values attained by VoML+G are 59% and 66%, respectively, for overall completeness and purity rates. They increase to $\sim 70\%$ and $\sim 85\%$ for the $1.0 \times 10^{13}$–$1.0 \times 10^{14}\ M_\odot\, h^{-1}$ mass range (e.g., Figure 16).

Of particular interest is the redshift dependence of the completeness and purity of the $N_g \geqslant 10$ VoML+G cluster subset. We have found that these rates are nearly coincident, at $\sim 80\%$, up to $z \sim 0.04$. However, the completeness for $z > 0.04$ is a decreasing function of redshift and the purity is an increasing one, whereas at $z = 0.14$ these rates approach $\sim 55\%$ and $\sim 100\%$, respectively.

The procedure of identifying cluster members in the algorithm through velocity gaps is supported by the observed sharp separation in velocity space between the clusters and the foreground/background galaxies (Rines & Diaferio 2006). Our optimization tests with the mock 2dFGRS showed that both 500 and 1000 km s$^{-1}$ performed comparably, but we chose 1000 km s$^{-1}$ because of the successful results by DP02 on the real 2dFGRS. A 1000 km s$^{-1}$ value is also supported by the work of Katgert et al. (1996), who concluded from a much smaller sample of galaxies ($\sim 6000$) with comparable redshift uncertainties that it was the optimum value for defining physical systems.

Because of our interest in finding preferentially isolated systems, the algorithm does not allow the registration of a cluster if it contains a galaxy already associated with a previously registered one. The question is whether this condition significantly reduces the fraction of 2dFGRS mock clusters detected by the algorithm. The numerical experiment described in Section 3.1 showed that, when the Gapper$R_{200}$ is forced to find the clusters with starting "true centroids" from the mock cluster catalog itself, no overlaps occur. Although this result suggests that few of the clusters detected automatically by the algorithm would have galaxies in common, the completeness for the $N_g \geqslant 10$ VoML+G cluster subset may be used to infer an upper limit (possibly an overestimation) of $\sim 25\%$ for the loss of mock clusters due to the non-overlap condition. It is also relevant that the exploration of the multiplicities, velocity dispersions, and masses of the non-detected mock clusters (Section 5.2) did not reveal significant biases with respect to the sample of Millennium clusters. The number of non-detections increases with redshift, as expected for a magnitude limited galaxy data set, becoming important at $z \sim 0.10$; for $z \gtrsim 0.15$, most of the mock clusters are lost.

A factor that can affect future applications of VoML+G on the "real 2dFGRS" is the detection of the false-positives, which may potentially alter the validity of the measured properties compared to a "true cluster set." The diagnostic, performed using the 2dFGRS mock catalog, allowed us to anticipate that most of the false-positive clusters would have a small number of members (median of 5) with a few of higher multiplicity. This diagnostic also showed that the VoML+G false positives are preferentially located at $z < 0.08$, a result that prompted us
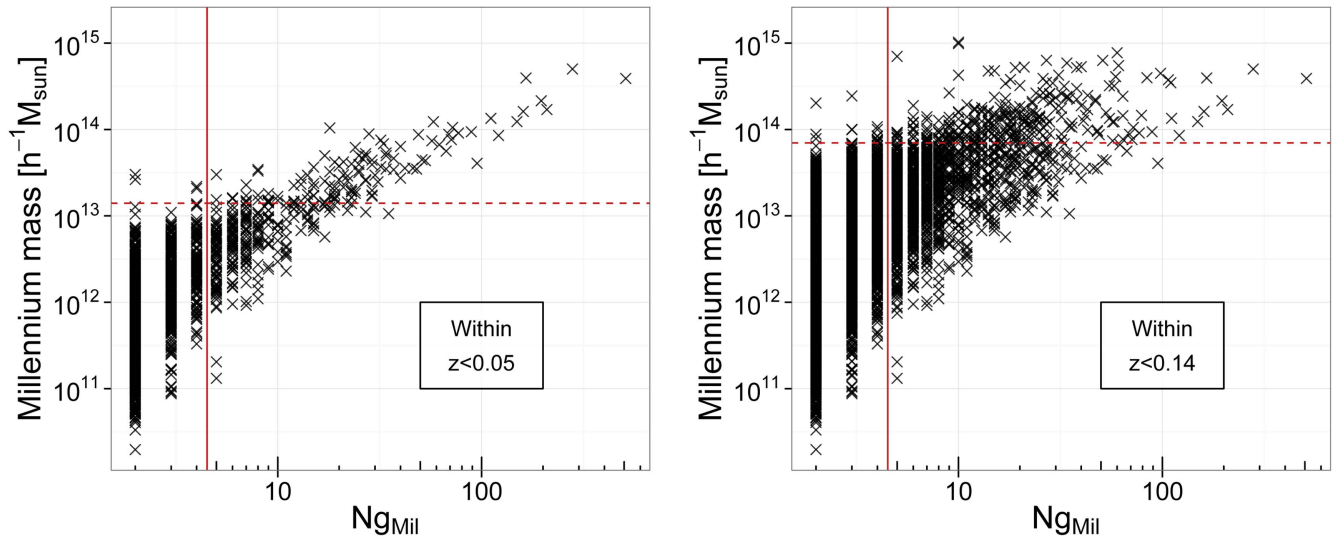
**Figure 20.** Mock clusters' halo DM masses within $z < 0.0.5$ and 0.14 as a function of the number of associated galaxies ($N_g$). The dashed red lines mark the masses above which the subsamples are 95% complete. The vertical red lines indicate the adopted minimum number of galaxies ($N_g = 5$).
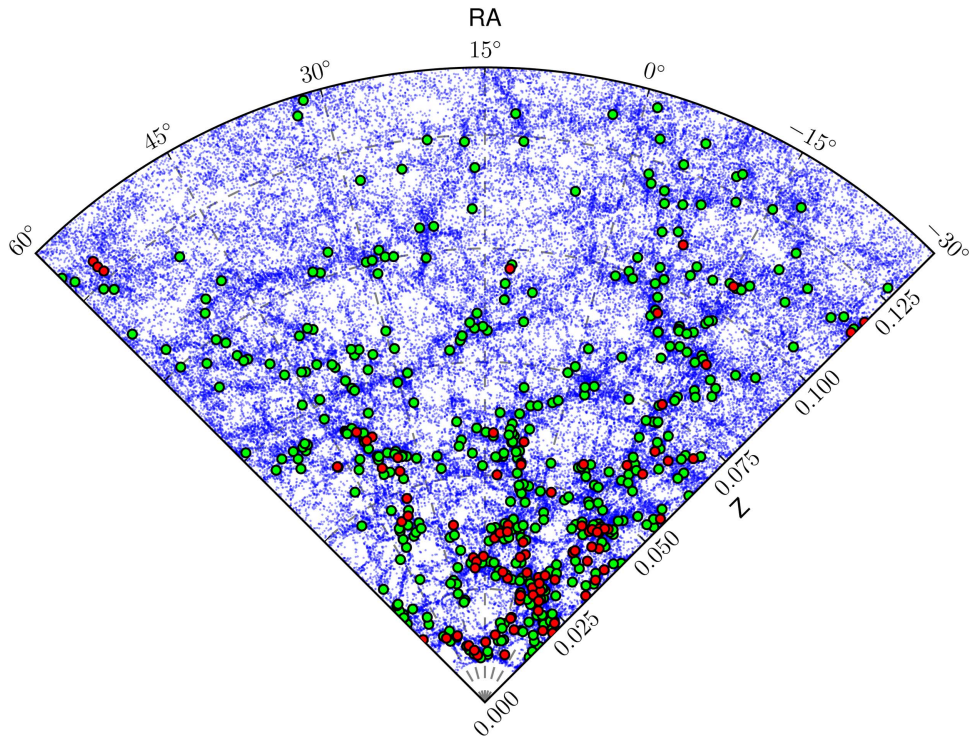


**Figure 21.** Wedge diagram of the VoML+G detections on the mock southern strip of the 2dFGRS. Red circles are false positive detections, and green ones are true positive. Only clusters with more than seven galaxies are included.

to examine, to $z = 0.14$, their spatial distribution superimposed on the Galaxy distribution. In Figure 21, a wedge diagram represents the distribution of the mock 2dFGRS galaxies (blue points), superimposed on the distribution of the true-positive (filled green circles) and false positive (filled red circles) $N_g > 7$ VoML+G clusters. Figure 21 shows that the false-positives are mostly located close to "mock clusters" and, like them, follow the filamentary structure on the galaxy distribution.

The foregoing validations of VoML+G justifies using the code in the future on the "real 2dFGRS," where the focus will be on the $N_g \geqslant 10$ cluster subset to $z \sim 0.14$. Our science program should be achievable because (i) it is deep and wide, unique in the south; (ii) it has a $z \sim 0.2$ redshift limit and a blue

magnitude limit, which favors the detection of low-luminosity late-type galaxies; (iii) it provides each galaxy with a spectroscopic parameter ($\eta$) that is moderately correlated with morphological type (Madgwick et al. 2002); and (iv) it has been extensively searched for clusters that can be used to test the effectiveness of our code to find Abell-like clusters.

We have run VoML+G on the "real 2dFGRS" data set and preliminarily explored the properties of the $N_g \geqslant 10$ output clusters. We have found that known Abell-type clusters studied by De Propris et al. (2002) are accurately rediscovered, that a considerable number of new clusters are revealed, and that nearly half of the detected clusters are dominated, in number, by late-type galaxies (Paper II, in preparation). These first

results appear to open the possibility of comparing the spatial clustering of clusters with different galaxy content, which could be used to further test the predictions of *the halo assembly bias*.

## 9. Conclusions

A new algorithm (VoML+G) for the detection of optical clusters of galaxies, based on the precepts of Pizarro et al. (2006), was created and optimized for the 2dFGRS up to $z \sim 0.2$. The maximum likelihood method that Allard & Fraley (1997) applied to the Voronoi Tessellation of a 2D point distribution, used in the first stage of our algorithm, was shown to be effective when applied to the three-dimensional galaxy distribution to locate regions where galaxy clusters lie. The second stage of our algorithm, which selects the cluster members around the centroids of these regions, based on the technique by Zabludoff et al. (1990) and later tested by De Propris et al. (2002) on the 2dF galaxy redshift survey (Colless et al. 2001), was shown to recover (with high fidelity) the mock reference clusters when a velocity gap of $1000 \, \mathrm{km \, s^{-1}}$ was used to identify galaxies within a cylinder of radius $R_{200}$ or $0.5 \, h^{-1} \, \mathrm{Mpc}$.

A mock 2dFGRS galaxy catalog (redshift range $0.009 < z < 0.22$), based on the Millennium Simulation of the LCDM cosmology, was employed for the optimization, validation, and determination of the selection function of the VoML+G code. The galaxy data set was taken from all sky mock catalogs of Blaizot et al. (2005) and a strip of sky was extracted, covering 1194 deg² spanning 90 deg × 15 deg, approximately matching the Southern strip of the 2dF galaxy redshift survey. The adopted mock galaxy data set contains 156,494 galaxies, and 1850 mock clusters with five or more member galaxies were identified to $z \sim 0.22$. The condition $N_g \geqslant 5$ implies, for volume-limited samples, that the reference cluster catalogs are 95% complete for DM halos above a minimum mass of $1.4 \times 10^{13} \, M_\odot \, h^{-1}$ within $z < 0.05$, $3.2 \times 10^{13} \, M_\odot \, h^{-1}$ for $z < 0.10$, and $7.0 \times 10^{13} \, M_\odot \, h^{-1}$ for $z < 0.14$ clusters.

The border effects on the detection of $N_g \geqslant 10$ clusters by VoML+G do not exceed 7%. Also, for this group of clusters, as diagnosed using the mask of the "real 2dFGRS," the VoML+G detections are reduced only by approximately 10%.

The application of the VoML+G code to the mock 2dFGRS reveals that 22% of the clusters have $N_g \geqslant 10$, 64% have intermediate multiplicity ($10 > N_g \geqslant 5$), and the remaining 14% of the detections containing fewer five member galaxies. The VoML+G detections are found to be minimally affected ($\sim$3%) by the typical galaxy redshift errors such as those present in "the real 2dFGRS." The algorithms built-in conditions of an upper limit for cluster size and exclusivity of the galaxy membership (non-overlap) are estimated to have a minor effect on the recovery of mock clusters, and to be neutral with respect to cluster properties.

Because of differences in survey size, redshift range, and redshift type, a direct comparison of performance of VoML+G with that of existing optical cluster finders should be made with caution. Milkeraitis et al. (2010) have reported an overall completeness rate of 54.8% for clusters more massive than $1.5 \times 10^{13} \, M_\odot$, whereas Cucciati et al. (2010) reported results of an overall completeness rate of $\sim$60%. For the overall purity, they reported rates of 88.8% and $\sim$50%, respectively. The VoML+G cluster finder attains overall rates of 59% and

66%, respectively. The VoML+G rates increase to $\sim$70% and $\sim$85% for the $1.0 \times 10^{13}$–$1.0 \times 10^{14} \, M_\odot \, h^{-1}$ mass range.

Graphical examples of the galaxy membership selection process for some individual $N_g \geqslant 10$ VoML+G clusters, as well as comparisons with mock clusters, provide a visual confirmation of the excellent capacity of the code to reliably recognize this subset of Millennium clusters.

The $N_g \geqslant 10$ VoML+G cluster subsample to $z \sim 0.14$ has: (i) both high mean completeness and purity ($\sim$75% and $\sim$90%, respectively), (ii) velocity dispersions that are halo mass proxies for values in excess of $3.0 \times 10^{13} \, M_\odot \, h^{-1}$, and (iii) $\lesssim$10% of false positives included.

Preliminary results from running the code on the "real 2dFGRS" provide evidence that known Abell-type clusters are accurately rediscovered and that many new clusters are revealed as well. Nearly half of the detected clusters are dominated, in number, by late-type galaxies. These first results indicate that the VOML+G code optimized for the 2dFGRS may be a key to revisit the issue of the diversity of the optical clusters, and to open a new possibility to test the predictions of *the halo assembly bias* through the use of massive galaxy clusters.

## References

Abell, G. O. 1958, ApJS, 3, 211

Abell, G. O., Corwin, H. G., Jr., & Olowin, R. P. 1989, ApJS, 70, 1

Aguerri, J. A. L., Sánchez-Janssen, R., & Muñoz-Tuñón, C. 2007, A&A, 471, 17

Allard, D., & Fraley, C. 1997, J. Am. Stat. Assoc., 92, 1485

Ascaso, B., Benítez, N., Fernández-Soto, A., et al. 2015, MNRAS, 452, 549

Ascaso, B., Wittman, D., & Benítez, N. 2012, MNRAS, 420, 1167

Barber, B., Dobkin, D. P., & Huhdanpaa, H. 1996, in ACM Transactions On Mathematical Software 22, 4

Beers, T. C., Flynn, K., & Gebhardt, K. 1990, AJ, 100, 32

Benson, B. A., de Haan, T., Dudley, J. P., et al. 2013, ApJ, 763, 147

Berlind, A. A., & Weinberg, D. H. 2002, ApJ, 575, 587

Biviano, A., Murante, G., Borgani, S., et al. 2006, A&A, 456, 23

Blaizot, J., Wadadekar, Y., Guiderdoni, B., et al. 2005, MNRAS, 360, 159

Boselli, A., & Gavazzi, G. 2006, PASP, 118, 517

Clerc, N., Pierre, M., Pacaud, F., & Sadibekova, T. 2012, MNRAS, 423, 3545

Cole, S., Percival, W. J., Peacock, J. A., et al. 2005, MNRAS, 362, 505

Colless, M., Peterson, B. A., Jackson, C., et al. 2003, arXiv:astro-ph/0306581

Colless, M., Dalton, G., Maddox, S., et al. 2001, MNRAS, 328, 1039

Croton, D. J., Gao, L., & White, S. D. M. 2007, MNRAS, 374, 1303

Cucciati, O., Marinoni, C., Iovino, A., et al. 2010, A&A, 520, AA42

Dalton, G. B., Maddox, S. J., Sutherland, W. J., & Efstathiou, G. 1997, MNRAS, 289, 263

De Propris, R., Couch, W. J., Colless, M., et al. 2002, MNRAS, 329, 87 (DP02)

Donahue, M., Scharf, C. A., Mack, J., et al. 2002, ApJ, 569, 689

Dressler, A. 1980, ApJ, 236, 351

Finn, R. A., Balogh, M. L., Zaritsky, D., Miller, C. J., & Nichol, R. C. 2008, ApJ, 679, 279

Flaugher, B. 2005, IJMPA, 20, 3121

Gao, L., Springel, V., & White, S. D. M. 2005, MNRAS, 363, L66

Goto, T., Sekiguchi, M., Nichol, R. C., et al. 2002, AJ, 123, 1807

Haines, C. P., Pereira, M. J., Smith, G. P., et al. 2015, ApJ, 806, 101

Henry, J. P., Evrard, A. E., Hoekstra, H., Babul, A., & Mahdavi, A. 2009, ApJ, 691, 1307

Hoshino, H., Leauthaud, A., Lackner, C., et al. 2015, MNRAS, 452, 998

Jing, Y. P., Mo, H. J., & Boërner, G. 1998, ApJ, 494, 1

Katgert, P., Mazure, A., Perea, J., et al. 1996, A&A, 310, 8

Kleinmann, S. G., Lysaght, M. G., Pughe, W. L., et al. 1994, ExA, 3, 65

Koester, B. P., McKay, T. A., Annis, J., et al. 2007, ApJ, 660, 221

Lemson, G., & Virgo Consortium, T. 2006, arXiv:astro-ph/0608019

Lewis, I., Balogh, M., De Propris, R., et al. 2002, MNRAS, 334, 673

LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201

Lumsden, S. L., Nichol, R. C., Collins, C. A., & Guzzo, L. 1992, MNRAS, 258, 1

Madgwick, D. S., Lahav, O., Baldry, I. K., et al. 2002, MNRAS, 333, 133

Mana, A., Giannantonio, T., Weller, J., et al. 2013, MNRAS, 434, 684

Mantz, A., Allen, S. W., Rapetti, D., & Ebeling, H. 2010, MNRAS, 406, 1759

Marinoni, C., Davis, M., Newman, J. A., & Coil, A. L. 2002, ApJ, 580, 122

Messier, C. 1781, Connoissance des Temps pour l'Anne bissextile, 1784, 227

Milkeraitis, M., van Waerbeke, L., Heymans, C., et al. 2010, MNRAS, 406, 673

Miller, C. J., Nichol, R. C., Reichart, D., et al. 2005, AJ, 130, 968

Miyatake, H., More, S., Takada, M., et al. 2016, PhRvL, 116, 041301

More, S., Miyatake, H., Takada, M., et al. 2016, ApJ, 825, 39

Nilo Castellón, J. L., Alonso, M. V., García Lambas, D., et al. 2014, MNRAS, 437, 2607

Oemler, A., Jr. 1974, ApJ, 194, 1

Peacock, J. A., & Smith, R. E. 2000, MNRAS, 318, 1144

Pizarro, D., Campusano, L. E., Clowes, R. G., et al. 2006, Proc. 3rd Int. Symp. on Voronoi Diagrams in Science and Engineering, ed. M. L. Gavrilova (Los Alamitos: IEEE), 112

Ramella, M., Boschin, W., Fadda, D., & Nonino, M. 2001, A&A, 368, 776

Rines, K., & Diaferio, A. 2006, AJ, 132, 1275

Rozo, E., Wechsler, R. H., Rykoff, E. S., et al. 2010, ApJ, 708, 645

Rykoff, E. S., Rozo, E., Busha, M. T., et al. 2014, ApJ, 785, 104

Scoccimarro, R., Sheth, R. K., Hui, L., & Jain, B. 2001, ApJ, 546, 20

Seljak, U. 2000, MNRAS, 318, 203

Söchting, I. K., Clowes, R. G., & Campusano, L. E. 2002, MNRAS, 331, 569

Söchting, I. K., Clowes, R. G., & Campusano, L. E. 2004, MNRAS, 347, 1241

Springel, V., White, S. D. M., Jenkins, A., et al. 2005, Natur, 435, 629

Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, AJ, 124, 1810

Vikhlinin, A., Kravtsov, A. V., Burenin, R. A., et al. 2009, ApJ, 692, 1060

Way, M. J., Gazis, P. R., & Scargle, J. D. 2011, ApJ, 727, 48

Zabludoff, A. I., Huchra, J. P., & Geller, M. J. 1990, ApJS, 74, 1 (ZHG)

Zheng, Z., Berlind, A. A., Weinberg, D. H., et al. 2005, ApJ, 633, 791

Zheng, Z., Zehavi, I., Eisenstein, D. J., Weinberg, D. H., & Jing, Y. P. 2009, ApJ, 707, 554

Zwicky, F., Herzog, E., & Wild, P. 1968, (Pasadena, CA: California Institute of Technology (CIT)), 1961