We present a comparison between two optical cluster finding methods: a matched filter algorithm using galaxy angular coordinates and magnitudes, and a percolation algorithm using also redshift information. We test the algorithms on two mock catalogues. The first mock catalogue is built by adding clusters to a Poissonian background, while the other is derived from N-body simulations. Choosing the physically most sensible parameters for each method, we carry out a detailed comparison and investigate advantages and limits of each algorithm, showing the possible biases on final results. We show that, combining the two methods, we are able to detect a large part of the structures, thus pointing out the need to search for clusters in different ways in order to build complete and unbiased samples of clusters, to be used for statistical and cosmological studies. In addition, our results show the importance of testing cluster finding algorithms on different kinds of mock catalogues to have a complete assessment of their behaviour. Galaxies: clusters: general – Cosmology: large-scale structure of Universe

Comparison of two optical cluster finding algorithms

Large and unbiased samples of clusters of galaxies are invaluable tools for investigating cosmology and the large scale structure of the Universe.

Since the compilation of the first optical samples (Abell Abell1958, Zwicky et al. Zwicky1961), it was apparent that selection effects in such catalogues are more difficult to understand and quantify than those in galaxy catalogues. Indeed, although the detection is done on the basis of a galaxy overdensity, the spatial scale and the magnitude of the overdensity vary with the (unknown a priori) redshift. Therefore, other properties of the cluster galaxy population as the morphology (presence of giant ellipticals) and photometric properties have taken an important role in detections.

Another important problem in detecting optical clusters is the presence of a significant background of field galaxies, which reduces the significance of a detected overdensity, especially at high redshift. A pioneering study on the detection of clusters and its dependence on various selection effects, applying a simple detection algorithm on simulated catalogues, was done by Cappi et al. (Cappi1989); see also van Haarlem et al. (vanHaarlem1997) and Reblinsky & Bartelmann (Reblinsky1999). Until recently, samples of high redshift clusters were selected almost only in the X-ray band, where this “background pollution” is far less important than in the optical band. However, optical cluster samples are still important because the objects are selected on the basis of the stellar light of the galaxy population, thus giving complementary information with respect to the hot gas-X-ray selected clusters. The complementarity of optical and X-ray based searches for clusters has been further reassessed by Donahue et al. (Donahue2001, Donahue2002), who showed that these searches sample different cluster populations, only partially overlapping. See also Holden et al. (Holden1999), Adami et al. (Adami2000).

The first automated and objective searches of optical clusters (Dalton Dalton1992, Dalton1994, Lumsden et al. Lumsden1992) started when large field galaxies catalogues became available (APM and COSMOS). These searches produced catalogues of nearby clusters.

Recently, more refined statistical techniques, as e.g. the matched filter algorithm (Postman et al. Postman1996) and its refinement EISily (Lobo et al. Lobo2000) were applied to deep imaging surveys like the EIS (Nonino et al. Nonino1999, Scodeggio et al. Scodeggio1999) in order to detect clusters at higher redshift. At the same time, algorithms based on different techniques have been developed, but all based on the detection of some kind of overdensity. It can be an overdensity (or better a sequence) in a colour-magnitude
plot, like in the red sequence method (Gladders & Yee Gladders2000) or an overdensity of photons in the unresolved background, like in the background fluctuations method (Dalcanton Dalcanton1996, Zaritsky et al. Zaritsky1997), which has been used for the Las Campanas Distant Cluster Survey (Gonzalez et al. Gonzalez2001, Gonzalez2002). The availability of multiband photometric surveys has encouraged the development of methods making use of colour information, like the already cited red sequence method and the “cut & enhance” method (Goto et al. Goto2002).

Quite surprisingly, little work has been done in order to estimate the relative efficiency and power of different methods in terms of completeness and spurious detections as a function of redshift (see Olsen et al. Olsen2001, Kim et al. Kim2002).

The new generation of redshift surveys, having a high degree of completeness on a wide volume, will permit for the first time the detection of clusters as three dimensional (α, δ and redshift) overdensities, overcoming in part the problem of the high background pollution, and new detection methods have been developed to take advantage of redshift information (e.g. Marinoni et al. Marinoni2002). Indeed, in these cases the main problem is the decrease of the total number of galaxies as a function of redshift, which could be taken into account with the selection function.

A growing number of such surveys is already available or will be started soon, e.g. CNOC2 (Yee et al. Yee2000), SDSS (York et al. York2000), 2dF (Colless et al. Colless2001), VVDS (Le Fèvre et al. LeFevre2001), DEEP II (Davis et al. Davis2001).

In order to avoid biases in subsequent studies, a key information is the selection function of the catalogues produced by the algorithm – that is, the fraction of detected objects with respect to the total population as a function of richness, redshift and other parameters.

A relatively simple way to find out such information is to create a mock catalogue of galaxies with known characteristics, thus having a complete a priori knowledge of the sample of objects we want to investigate. Generally speaking, the simplest way to set up such a catalogue is to build a background of galaxies on which a number of clusters with known richnesses and density profiles are superimposed. Using a simple mock catalogue with known parameters represents a first test for the algorithms, in order to identify the main biases without ambiguity.

On the other hand, the real Universe cannot be simply thought as a superposition of clusters and background galaxies, but includes complex large scale structures such as filaments, “walls” and superclusters. Moreover, clusters of galaxies show a huge variety of shapes, profiles and substructures, while a mock catalogue can usually reproduce only a limited range of these parameters. Therefore, more refined tests need more realistic catalogues, as those generated with N-body simulations (see White & Kochanek White2002, Kochanek et al. Kochanek2003 for an application to cluster finding algorithms). Such catalogues come remarkably close to what the real Universe is, as can be seen by computing basic properties such as number counts and angular/spatial correlation function. They also offer a complete knowledge of the galaxy sample, without the additional worries (incompleteness, measuring errors, star/galaxy discrimination) brought about by real surveys. On the other hand, we cannot decide a priori the positions and features of the clusters in the sample. The cluster sample can instead be reconstructed a posteriori, starting from a quantitative definition of cluster.

The aim of this paper is to investigate the efficiency in detecting clusters and the relative selection effects of the two methods EISily (Lobo et al. Lobo2000) and Spectro (Adami & Mazure Adami2002). The EISily algorithm is a purely bidimensional method which uses both overdensity in number of galaxies and a fit to the luminosity function, while the Spectro method works in the combined bidimensional + velocity space.

We first apply the algorithms to a mock catalogue obtained by adding random clusters to a Poissonian background, then on a mock catalogue by Hatton et al. (Hatton2003) generated by N-body simulations. We define three cases with regard to redshift completeness: 100% completeness down to $I = 24$ (the complete sample), 50% completeness down to $I = 24$ (the deep sample), 33% completeness down to $I = 22.5$ (the shallow sample). The last two cases represent reasonable values for the various recent redshift surveys near completion or already available in the literature.

The methods se:themethods

The EISily algorithm sub:eisily

The EISily algorithm (Lobo et al. Lobo2000) belongs to the matched filter category (Postman et al. Postman1996), which has appeared in several “flavours” since its introduction. Although originally
designed to treat bidimensional data, versions using redshift information have been developed (e.g. Kepner et al. Kepner1999). The characteristics of the code we used for this work is thoroughly described in Lobo et al. (Lobo2000). We will, however, remind its main steps.

EISily is fed with a catalogue of galaxies providing positional data (RA and Dec) and magnitude in one band. Unlike other matched filter algorithms, this one completely separates the spatial part (search for significant galaxy overdensities), which comes first, from the luminosity part (search of a cluster-like luminosity function). Furthermore, no assumptions are made about the physical size and the exact density profile of clusters.

For the spatial overdensity detection, a Gaussian filter is applied to a regular grid of points drawn on the catalogue. The grid spacing is equal to the filter angular size $\sigma_{\text{ang}}$. For each point of the grid, the galaxies within $3\sigma_{\text{ang}}$ are weighted according to their position, and a S/N ratio is computed by subtracting the signal given by the background (measured in a region of radius $6\sigma_{\text{ang}}$ around the point) and dividing by the Poissonian standard deviation. A point is flagged as a possible detection whenever its S/N ratio is greater than the S/N of its eight neighbouring points. This test is not performed on points at the edge of the grid, in order to avoid border effects. The S/N ratio of edge points is used only as a comparison value for inner points.

For each candidate cluster detected in this way, a fine centering is then performed. A local grid is applied around each candidate, with a spacing reduced by one half with respect to the original grid. The S/N ratio is computed on the points of this finer grid and compared with the original value. The point with the highest S/N ratio becomes the new provisional cluster position, and the procedure is repeated, with the grid spacing reduced again by one half. Five iterations are made, so that the local grid becomes up to 32 times finer than the original one.

The filter size is then reduced by a factor $\sqrt{2}$. A new, finer grid is applied on the whole catalogue and the steps described above are performed again. Finally, double detections (the ones with separation less than the average of the scales) are removed, leaving only those with greater signal.

We thus end with a preliminary catalogue of candidate clusters, on which a refined background subtraction is made: the background is computed inside a crown around the cluster region. It may happen that the previously detected overdensity disappears, in which case the candidate is removed. No other rejections are made after this step.

Finally the luminosity part of the algorithm is applied. A Schechter function plus power law is fitted to the galaxies of each candidate to simulate the cluster+background luminosity function, and a maximum likelihood analysis (adapted from Schuecker & Böhringer Schuecker1998) is performed, in order to find the best estimate for $m^*$. The slope $\alpha$ remains fixed throughout the computation: we chose a value of $-1.25$.

The Spectro algorithm

Unlike EISily, this algorithm makes also use of the redshift information. The input parameters are:

- $\text{M inimum and maximum values of Right Ascension, Declination and redshift for the given sample.}$
- $\text{Angular size of the search window on the sky. This size has to be adapted to the kind of structure we want to detect. For clusters, a good compromise is to adapt this size to match a physical size of R}_{200}$ (typically 1.8 Mpc, $H_0 = 65$ km s$^{-1}$ Mpc$^{-1}$). Smaller sizes could be chosen when looking for groups, larger sizes for superclusters or filament searches.
- $\text{Velocity gap for structure analysis, in km s$^{-1}$ (see e.g. Biviano et al. Biviano1997). This is the optimal velocity gap. For example, a good value for rich cluster searches is 600 km/s. This value is deduced by ranking the galaxies of the sample by increasing redshift, making the histogram of the redshift galaxy-galaxy separation and looking for excesses. In other words, this value is close to the minimum velocity dispersion of the structures to be detected.}$
- $\text{Minimum number of objects per elementary structure. This is the number of objects below which an elementary structure is discarded in the first step of the algorithm (see below). A good value for clusters is five.}$
- $\text{Minimum number of objects in the percolated structure. This is the minimum number of objects in the whole structure (the sum of all the objects detected in each elementary structure). Fixing this number to 10 means that only structures with more than 10 galaxies in total will be kept.}$
- $\text{Minimum distance between two percolated elementary structures. A good value for clusters is R}_{200}$. If
two elementary structures (step 1) are closer than this value, these are assumed to belong to the same “global” structure.

Maximum extension in redshift of a structure. A good value is six times the typical velocity dispersion of the structures. This parameter must be greater than the gap value.