UNIVERSITY OF CRETE

Identification of clusters in observational and mock galaxy catalogs

by

Giorgos Savathrakis

Diploma Thesis for the Degree
of Master of Science in Physics

in the
University of Crete
Department of Physics

October 2021

UNIVERSITY OF CRETE

Abstract

University of Crete
Department of Physics

by Giorgos Savathrakis

Galaxy clusters are large gravitationally bound structures that consist of a wide range of galaxies. An important objective when studying galaxy clusters is to determine their shape and number of galaxies. In this work, we propose a numerical method to define the boundaries of possible clusters based on a variation of the k Nearest Neighbors (kNN) algorithm and Monte Carlo sampling methods. We obtained the data on which our analysis was conducted, from the Heraklion Extragalactic Catalogue (HECATE), which provides the names and coordinates of specific galaxies, as well as from the Extended Virgo Cluster Catalogue (EVCC) which provides the coordinates of galaxies that belong to the Virgo cluster, and were used as a validation of our method. We conclude that our method successfully identifies the locations of clusters and determines their shape.

Acknowledgements

I would like to thank my supervisor, professor Andreas Zezas, for his invaluable help and guidance in this project without which its completion would not have been possible.

Contents

List of Figures

List of Tables

Abbreviations

GCs	Galaxy Clusters
kNN	k Nearest Neighbors
MC	Monte Carlo
HECATE	Heraklion Extragalactic Catalogue
EVCC	Extended Virgo Cluster Catalogue
VCC	Virgo Cluster Catalogue
SDSS	Sloan Digital Sky Survey

Symbols

RA	Right Ascension
Dec	Declination
D	Distance
d_k	Distance of kth galaxy
ρ	Galactic numerical density
σ	standard deviation
R_{200}	Characteristic cluster radius (density $< 200\Omega_m^{-1}\times$mean galaxy density)
H_o	Hubble's constant
α	Statistical Significance

Chapter 1

Introduction

1.1 Galaxy Clusters

1.1.1 Sizes and shapes

Galaxy clusters (GCs) are the largest structures in the Universe, held together by the gravitational pull of the galaxies that belong to them. Typical GCs consist of some hundreds to several thousands of galaxies. These are categorized as rich clusters. On the other hand, clusters that contain a few dozen galaxies are categorized as poor clusters. GCs are also classified, in terms of their shape, as regular or irregular depending on their spherical symmetry [1].

1.1.2 GC boundaries

There are several methods by which GC boundaries can be determined. For example, R_{vir} is defined as the radius within which the gas reaches thermal equilibrium at approximately 10^7K [2]. It can be detected observationally as the point beyond which the X-ray profile of the cluster vanishes behind the background emission. Another method is to define the point beyond which galaxies follow the cosmic expansion and are therefore outside the gravitational well of the cluster. It can be calculated using observational data for its members' redshifts. Galaxies with velocities within the velocity dispersion range, are considered as members of the cluster. Such methods however, are not automated. We therefore rely on numerical methods which detect overdensities within a distribution of galaxies. kNN algorithm is one of these methods.

FIGURE 1.1: The Coma cluster which is an example of a regular rich cluster
https://earthsky.org/clusters-nebulae-galaxies/the-coma-berenices-galaxy-cluster/

1.2 kNN

In machine learning classification tasks, k Nearest Neighbors (kNN) algorithm is used to determine the class of a sample, depending on the class in which the majority of its k closest samples in the parameter space belong to. A variation of this algorithm [3] can be used as a clustering method to determine the existence of overdense structures in galactic data. Assuming a distribution of galaxies in 3-dimensional space, each one of them is considered as the center of a spherical volume, enclosed within the distance of kth closest galaxy to the center. A 2-dimensional representation of this assumption is shown in figure 1.2.

Subsequently, we can obtain the galactic numerical density in each point by calculating the distance of the kth nearest galaxy to that point from equation 1.1 where index 1 refers to the galaxy in the center of the volume, and index 2 refers to the galaxy at the edge of the volume. d is the radial distance from the observer and RA and Dec are the right ascension and declination of the galaxy respectively.

$$\mathrm{d_k} = \sqrt{\mathrm{d_1^2} + \mathrm{d_2^2} - 2\mathrm{d_1 d_2}(\cos(\mathrm{Dec_1})\cos(\mathrm{Dec_2})\cos(\mathrm{RA_1} - \mathrm{RA_2}) + \sin(\mathrm{Dec_1})\sin(\mathrm{Dec_2}))} \tag{1.1}$$

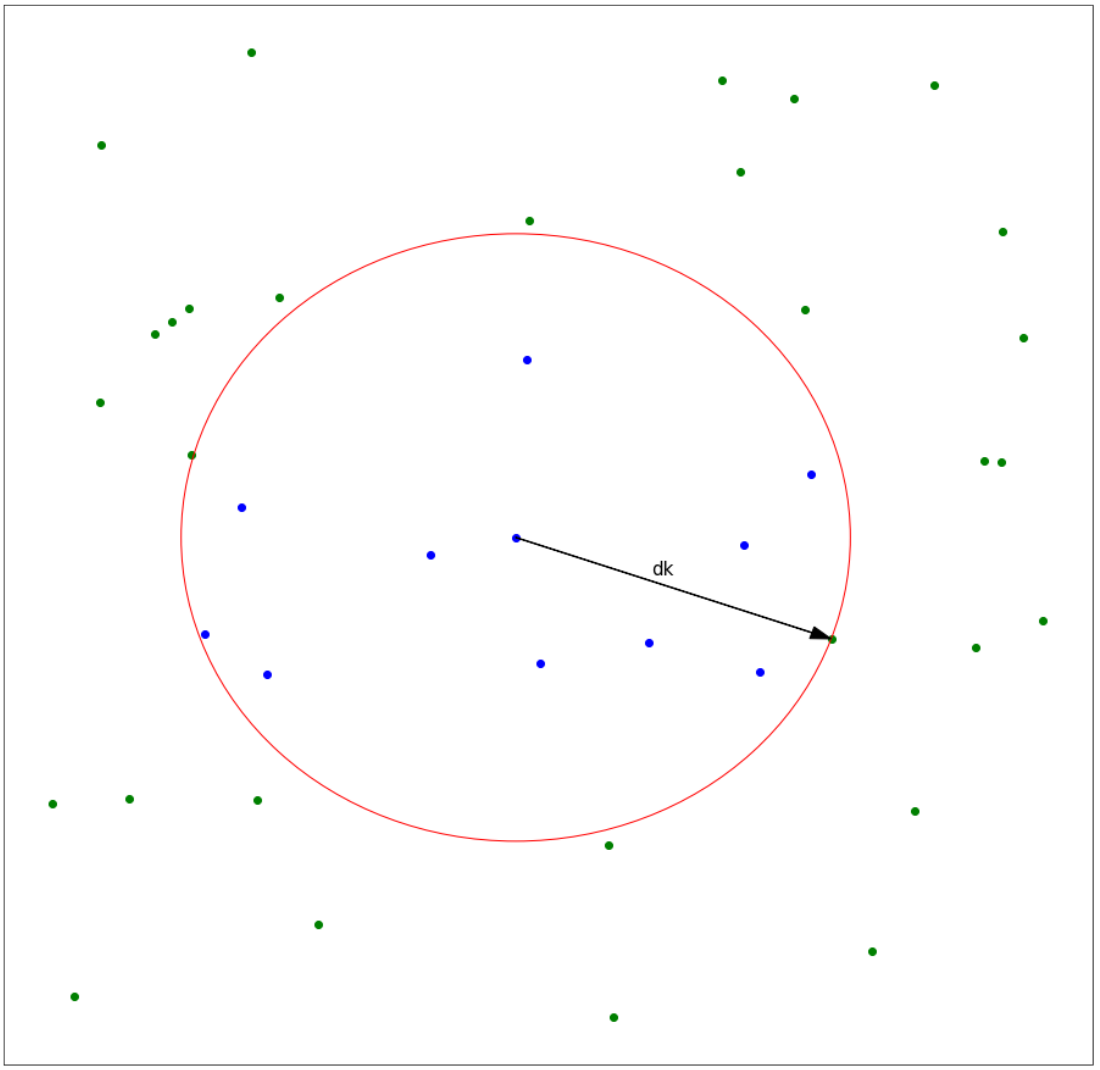

FIGURE 1.2: Representation of 2d kNN density calculation where the blue points are the k samples enclosed by the volume defined by the red circle whose radius is the kth farthest sample. Green points represent the samples outside that radius.

We can now calculate the numerical density from equation 1.2

$$\rho = \frac{k}{(4/3)\pi d_k^3} \tag{1.2}$$

Each galaxy represents a point in our grid, where we calculate the numerical density of galaxies. The algorithm's sensitivity in wide or narrow overdense regions depends on the value of k. Larger values of k correspond to a much broader volume for each point. Therefore, the algorithm will be able to detect overdense structures that span across a wider area. On the other hand, smaller values of k will lead to the detection of smaller local overdensities.

However, in order to determine the boundaries of a candidate cluster, we need to calculate the significance of possible overdensities compared to a large number of uniform galactic distributions.

1.3 Approach

Our approach is to create a density map of our galactic data from HECATE [4] using the kNN algorithm explained in section 1.2. Our null hypothesis is that there is no cluster in our distribution. Therefore, we create N $>>$ 1 Monte Carlo (MC) samples to calculate the statistical significance of any overdense structures. Each sample consists of M points, where M equals the number of galaxies in the catalogue. These points are drawn from a uniform density distribution in the same parameter space as our data. We create a density map for every MC sample. Since the densities are calculated in the location of each galaxy, a point-to-point comparison between the densities of our data and the MC samples is impossible. To overcome this we create 2-dimensional bins, where we sum the numerical densities of the galaxies within them. For each MC sample in particular, we first calculate the summed densities in each bin and subsequently, we calculate the mean density and standard deviation σ of all samples in each bin. Determining whether a region is overdense depends on the deviation in terms of σ of the numerical density of our data, from the uniform numerical density of the MC samples in the same bins. The density maps are generated using different values of k. We then calculate the total statistical significance within the boundaries that surround the cluster, which consists of multiple bins, by considering the sums of the pixel densities from our original data, the pixel densities of the MC samples and the standard deviations. The optimal value of k depends on the total significance of each structure. Subsequently, we evaluate the algorithm's performance, by calculating the number of galaxies from ground truth data, that lie within the boundaries of the cluster for the optimal value of k. Finally, we conduct the same analysis on mock galaxy catalogues that are created from dark matter halo simulations.

Chapter 2

Data

2.1 HECATE

The Heraklion Extragalactic Catalogue (HECATE) [4] is an all-sky galaxy catalogue which includes approximately 200000 galaxies within a distance of D$\lesssim$200Mpc. It incorporates galaxies from different databases namely from NED [5] and HyperLEDA [6]. The sample selection occurred from the careful identification of identical galaxies inside

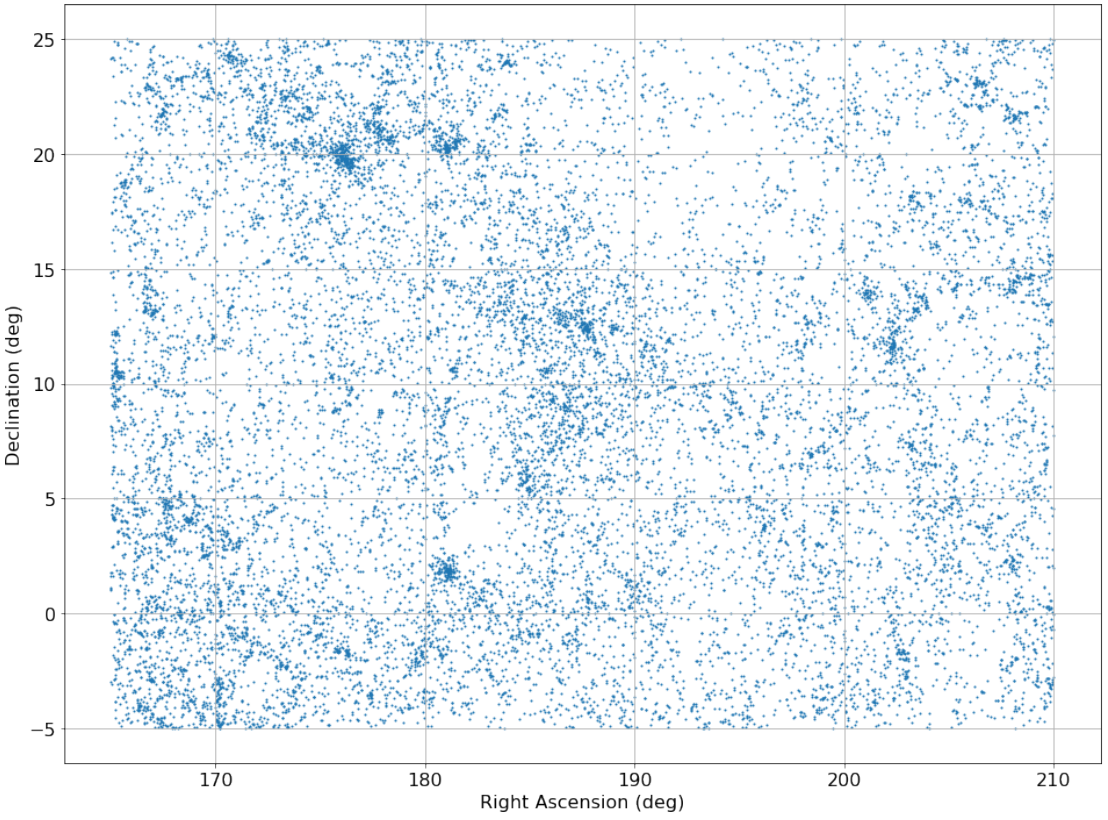

FIGURE 2.1: Locations of HECATE galaxies within the subset parameter space

these catalogues, as well as from the exclusion of objects that contain multiple galaxies (e.g. groups, clusters), while including their members. Additionally, redshift dependent distances have been obtained for the majority of the galaxy without available redshift independent distances. The catalogue also includes photometric data as well as information regarding the galaxies' inclination, size and coordinates. In our analysis we include a subset of 16962 galaxies. The RA of the members ranges from 165 to 210 degrees and their declination from -5 to 25 degrees as shown in figure 2.1, covering the area around the Virgo cluster centered at RA=187° 42' 21.35" and Dec=+12° 23' 28.0439".

2.2 EVCC

As a ground truth sample we adopt the Extended Virgo Cluster Catalogue (EVCC) [7] which is based on a spectroscopic survey by the Sloan Digital Sky Survey (SDSS) and provides data for galaxies within a radius of $3.5R_{vir}$ from the center of the Virgo cluster.

FIGURE 2.2: Spatial distribution of EVCC galaxies

EVCC consists of a total 1589 galaxies making it 5.2 times larger than its previous version the VCC. Apart from the velocity and the locations of the galaxy members, EVCC provides information regarding each sample's photometry, declination and morphological

classification. As shown in figure 2.2, the region covered by the distribution, ranges from RA=175° to 200° and from Dec = -4° to 25°. We intend to use the data from EVCC as validation for our clustering method on the galaxies from HECATE. EVCC does not explicitly provide the distances of its galaxies but we can obtain them by dividing the velocities with the Hubble constant. We assume its value to be H_o=70km/s/Mpc.

Chapter 3

Methods

3.1 Creation of density maps

Starting with the galaxies from HECATE, we run the kNN algorithm to calculate the galaxy number density at the location of each galaxy in the catalogue. In figure 3.1

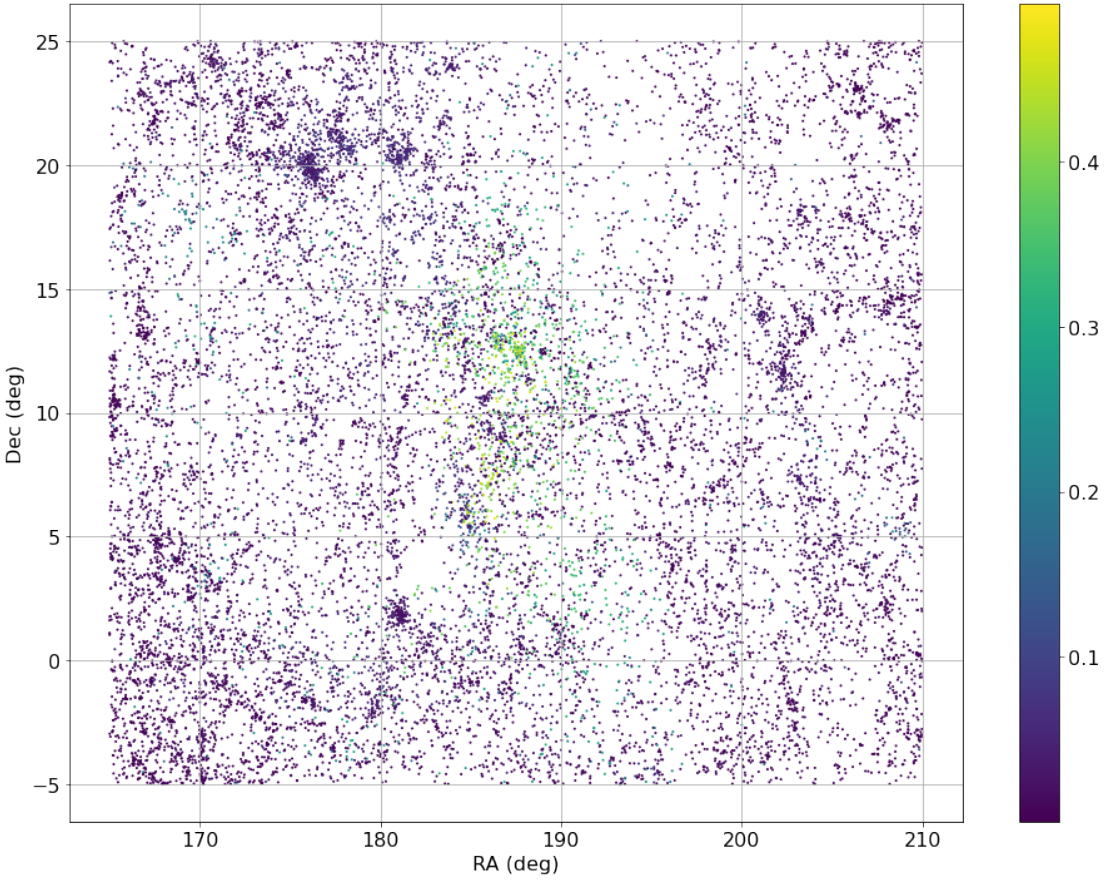

FIGURE 3.1: Density map for k=1500. The color scale depicts the density in Mpc^{-3} at each point

we present the density map created for k=1500. We can see that there are certain regions, located mostly in the center, where the density is larger compared to other regions. This provides an indication that clusters may be found in these regions. The quantitative analysis for the presence of a cluster and the determination of its outline however, is done by comparing the calculated densities of our original data to the Monte Carlo simulations.

3.2 Binning of the parameter space

As explained in section 1.3, we separate our parameter space into bins. This is done in order to effectively calculate the statistical significance of possible overdensities. We set the size of these bins to be 0.2deg×0.2 deg which provides enough resolution, while containing enough objects for the assessment of the uncertainties. The density of each bin is the sum of the densities calculated in every point within the bin. In figure 3.2 we present the binned parameter space in 2 dimensions for the galaxies in HECATE as well as the density of each bin in logarithmic scale. We see that the overdense regions are more apparent, which is expected due to the summation of the densities. Specifically, apart from the central region where we know that the Virgo cluster is located, we see another apparent overdensity in the upper left part of the map at RA$\simeq$176 degrees and Dec$\simeq$20 degrees. These coordinates are close to the center of the Leo cluster. This means that the algorithm is sensitive to clusters of varying sizes and shapes which may render it suitable for the detection of multiple clusters simultaneously.

3.3 Monte Carlo samples

In order to assess the features identified by the kNN method we perform a set of simulations where we draw the locations of each object from a uniform distribution. This is equivalent to assuming that all objects in the area are uniformly distributed and there are no overdensities. We then run the kNN analysis for each set of draws. We create 100 MC samples for the calculation of the statistical significance of overdensities in the HECATE catalogue, each consisting of 16962 members. The kNN algorithm is then run to create the density map of every sample. Given the large number of objects, 100 Monte Carlo draws are adequate in order to obtain a picture of the statistical fluctuations of the kNN analysis.

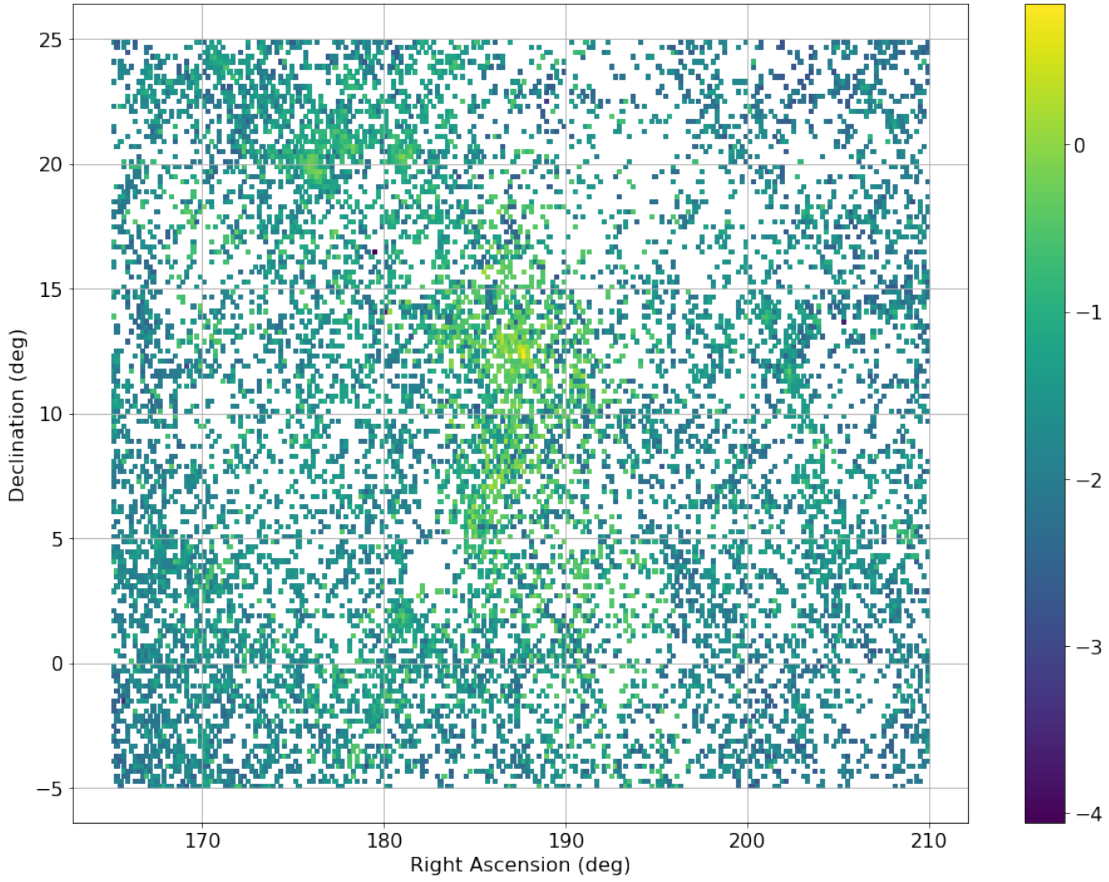

FIGURE 3.2: Density map for the 2-dimensional binned parameter space of HECATE. The colorbar represents the $\log_{10}$ of the bins' densities

3.4 Calculation of statistical significance

Having created the density maps for each sample (both HECATE and dark matter halos), as well as the spatial grid in which our data are distributed, we can calculate the statistical significance of each bin by comparing its density with the distribution of densities estimated by the MC samples. The significance of the overdensity of each pixel is measured as the deviation in terms of σ of the density measured from the actual data from the average density in that bin estimated from the MC simulations [8]. In order to determine the boundaries of the clusters for each k, we identify the pixels (bins) that exceed a given significance level. The total significance of a given cluster is calculated by combining the significance of the pixels within its boundary from equation 3.1

$$\alpha_{\rm tot} = \frac{\sum_{\rm i} {\rm N_i} - \sum_{\rm i} {\rm B_i}}{\sqrt{\sum_{\rm i} \sigma_{\rm i}^2}} \tag{3.1}$$

where

$\sum_{\rm i} {\rm N_i}$: the sum of the densities within the cluster for our original data

$\sum_i B_i$: the sum of the densities within the cluster for the Monte Carlo samples. It is important to note that the boundaries have been created from the actual data and not the MC draws
$\sum_i \sigma_i^2$: the sum of the squared standard deviations within the cluster from the MC draws
α_{tot}: the total statistical significance of the prevalent cluster

The optimal k is the one that maximizes the total significance. We create density maps with k=[100,200,500,1000,1500,2000] for the HECATE galaxies.

3.5 Halo simulations

As a final validation of our model's efficiency at identifying and delineating possible clusters, we include data provided by dark matter halo simulations. These simulations are provided by the IllustrisTNG project [9] [10] [11] [12] [13] which creates simulated galaxies considering a variety of processes that lead to galaxy formation. The data

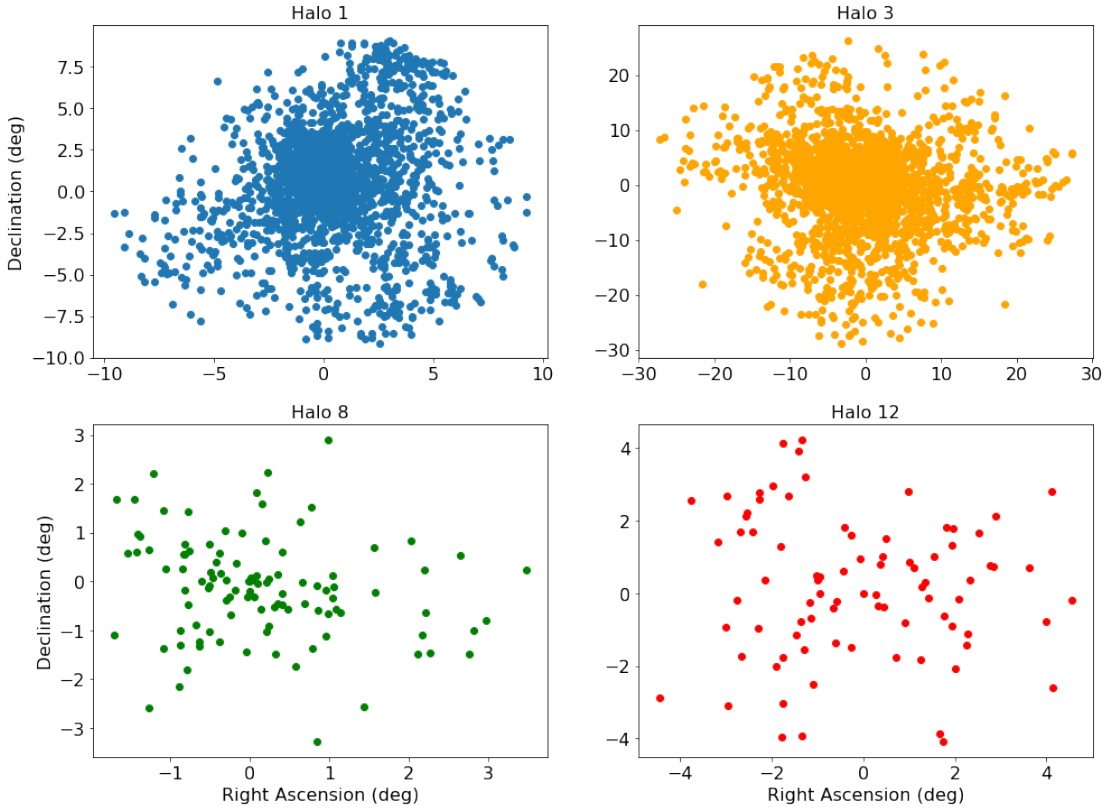

FIGURE 3.3: Spatial distributions of galaxies in selected dark matter halo simulations

releases cover 3 sets of side lengths each, 50, 100 and 300 Mpc. For our analysis, we use the last set because, due to its size, it is the most appropriate for cluster detection. The masses of the halo members M_{200}, range from 10^{10} to 10^{15} $M_\odot$ and their distribution

reaches a radius of $2R_{200}$. The locations of the dark matter halo members are given in cartesian coordinates, at a reference frame in which the center is assumed to be at the center of the halo. By assuming a random location in the sky for each halo center, we can transform the cartesian coordinates from the center of the halo, to equatorial coordinates. In total we have 14 dark matter halo simulations. The number of objects in each halo ranges from 70 to 2000 and we present the spatial distribution of 4 representational halos in figure 3.3. When conducting the kNN analysis for the dark matter halos we create 10000 MC samples because the number of objects in each halo is much smaller than the galaxies in HECATE. The values for k that we use for the dark matter halos are k=[10,20,50,100,200,500].

3.6 Dark matter halos' coordinates transformation

The transformation of the dark matter halo coordinates, mentioned in paragraph 3.5, is done by firstly assuming an arbitrary location of the halo center on the sky. For simplicity we choose the RA and Dec of the center to be (0,0) and the radial distance D can range between 15 and 200 Mpc in order to be consistent with the parameter space of HECATE. The cartesian coordinates of the halo center in a reference frame centered at the observer, can be calculated as:

$x_{cen} = Dcos(Dec)cos(RA) = D$

$y_{cen} = Dcos(Dec)sin(RA) = 0$

$z_{cen} = Dsin(Dec) = 0$

Therefore, the halo members' coordinates in the observer's reference frame are:

$x_i = x_{cen} + x_i'$

$y_i = y_{cen} + y_i'$

$z_i = z_{cen} + z_i'$

where the $'$ represents the coordinates in the halo frame and i is the index of each halo member. Transforming the cartesian to equatorial coordinates in the observer's reference frame is done as follows:

$RA_i = \arctan(\frac{y_i}{x_i})$

$Dec_i = \arcsin(\frac{y_i}{\sqrt{x_i^2+y_i^2+z_i^2}})$

$D_i = \sqrt{x_i^2 + y_i^2 + z_i^2}$

Chapter 4

Results

In this chapter we present the results obtained from the kNN clustering algorithm, implemented on both the HECATE catalogue and dark matter halo simulations.

4.1 HECATE density map for k=1500

Starting with HECATE, we first create the density map for k=1500 and then we overplot the data from EVCC to estimate the number of galaxies that lie in the bins with the highest statistical significance. This is done as a preliminary evaluation of the algorithm's performance. In figure 4.1 we present the binned density map for the galaxies in HECATE and for k=1500. The bins are categorized according to their statistical significance. Regions with large statistical significance have a higher probability of belonging into overdense structures like galaxy clusters. If the statistical significance of a region is considerably small (i.e. $\alpha < \sigma$), then the region is underdense. The scaling of the levels of statistical significance ranges from regions where $\alpha < 0$ up to regions where the statistical significance reaches its maximum value. The intermediate levels have been selected in order to better visualize the resulting density map according to our own estimate of the significance levels required to determine an overdensity. We see that in the middle of the map there is such an extensive overdense region centered at RA $\simeq$ 187deg and Dec $\simeq$ 12deg. This region corresponds to the actual location of the Virgo cluster. There are several more overdense regions, one near the upper left and another near the
lower left corner of the map. Specifically, for the region corresponding to the Virgo cluster, we can compare the locations of the densest bins in that region, to the locations of the galaxies in the EVCC.

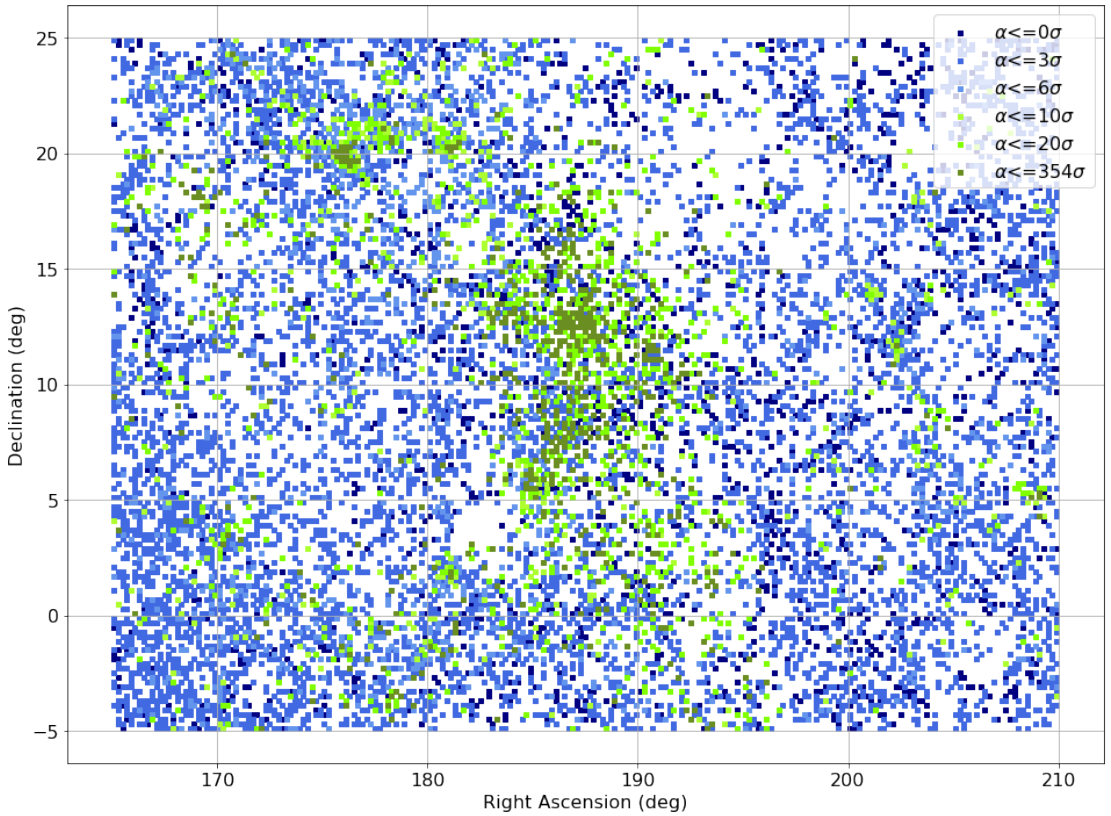

FIGURE 4.1: Density map for galaxies in HECATE for k=1500. The color scale represents the statistical significance of each bin. Dark green bins are overdense regions and dark blue bins are underdense regions.

4.2 Comparison with EVCC data

The equatorial coordinates of the galaxies in EVCC can be overplotted on the density map of figure 4.1 to examine if the bins on which they lie are statistically significant enough to be parts of a cluster. The result is shown in figure 4.2, where we see that the majority of the EVCC galaxies are in the bins corresponing to the highest levels of statistical significance. It is noticeable that the two largest subclumps that characterize the Virgo cluster are also detected by the algorithm. These are the positions where the two main groups of the EVCC galaxies are located, both of which have large statistical significance.

4.3 Cluster boundaries for HECATE

Subsequently, we create contour plots for the different values of k and for different confidence thresholds. In each case we calculate the statistical significance within the cluster boundaries that are spatially correlated with the Virgo cluster and depending on the configuration that maximizes the statistical significance, we determine the optimal

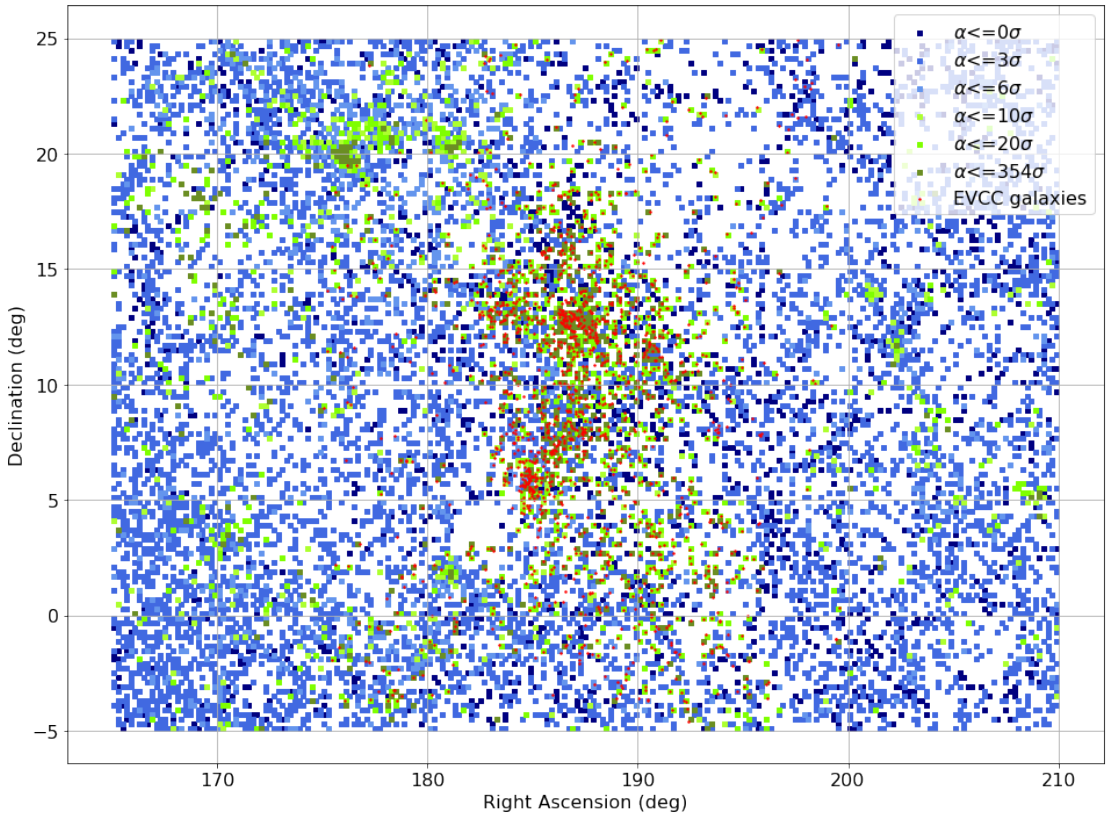

FIGURE 4.2: Overplot of HECATE density map for k=1500 and EVCC galaxies. The objects in red correspond to the galaxies in EVCC

value of k and σ_{thres}. The same analysis is then done on the dark matter halos for the values of k considered most likely to yield reliable results. Now that we have seen that the density maps, created with the kNN algorithm, yield reliable results, we can determine the clusters' boundaries. To achieve this, we create density maps for a set of k values that we consider likely to yield stable solutions for the cluster boundaries, and we draw contours surrounding the regions that exceed a certain level of statistical significance α. Due to the large number of bins, and the close proximity of bins with large and small statistical significance in some regions, we smooth the contour using a gaussian filter. A common significance threshold for determining whether a result refutes the null hypothesis is 3 standard deviations. We define the contour levels to be at this value of α and we create the density maps for the values of k mentioned in paragraph 3.4 for the galaxies in HECATE.

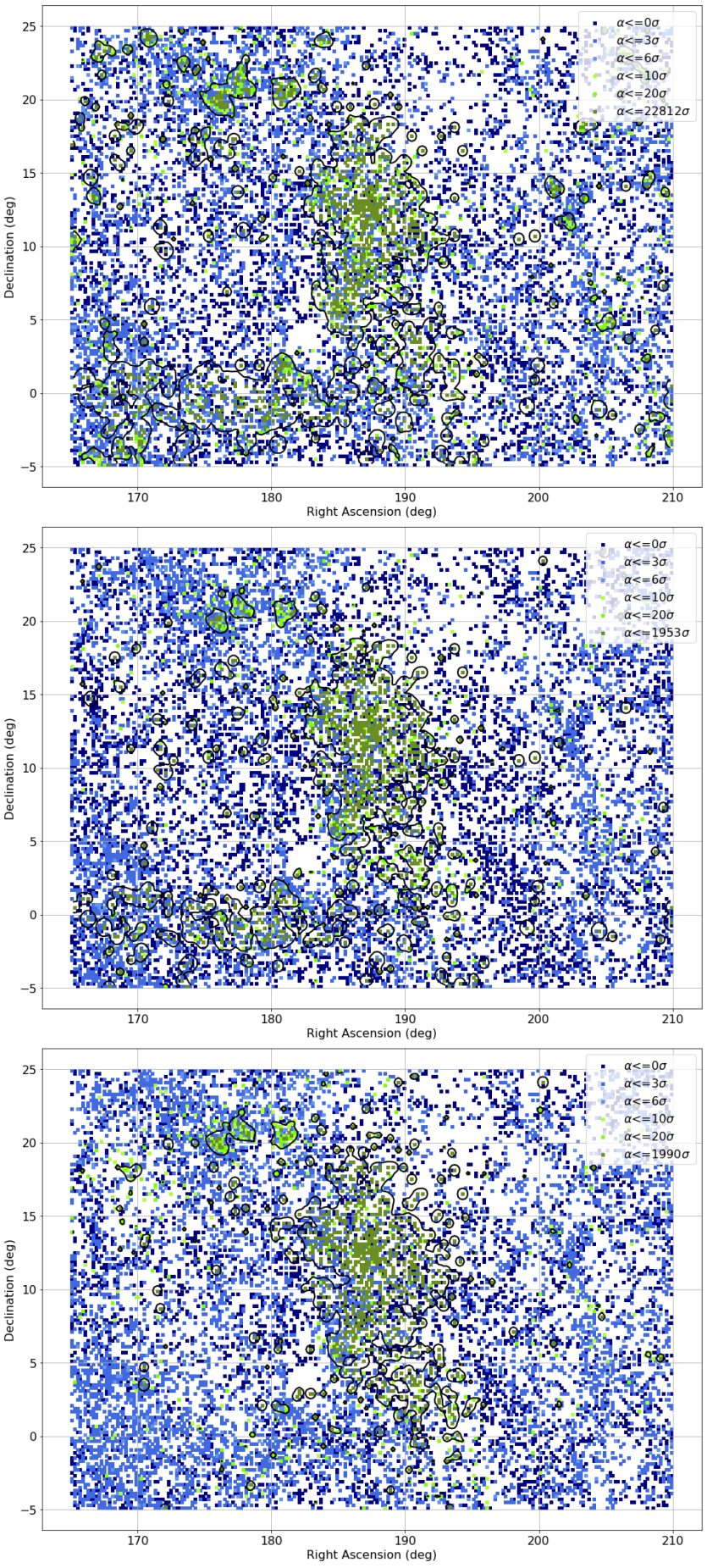

FIGURE 4.3: Density maps with significance threshold 3σ. Top: k=100, middle: k=200, bottom: k=500. The black lines are the boundaries that surround the overdense regions

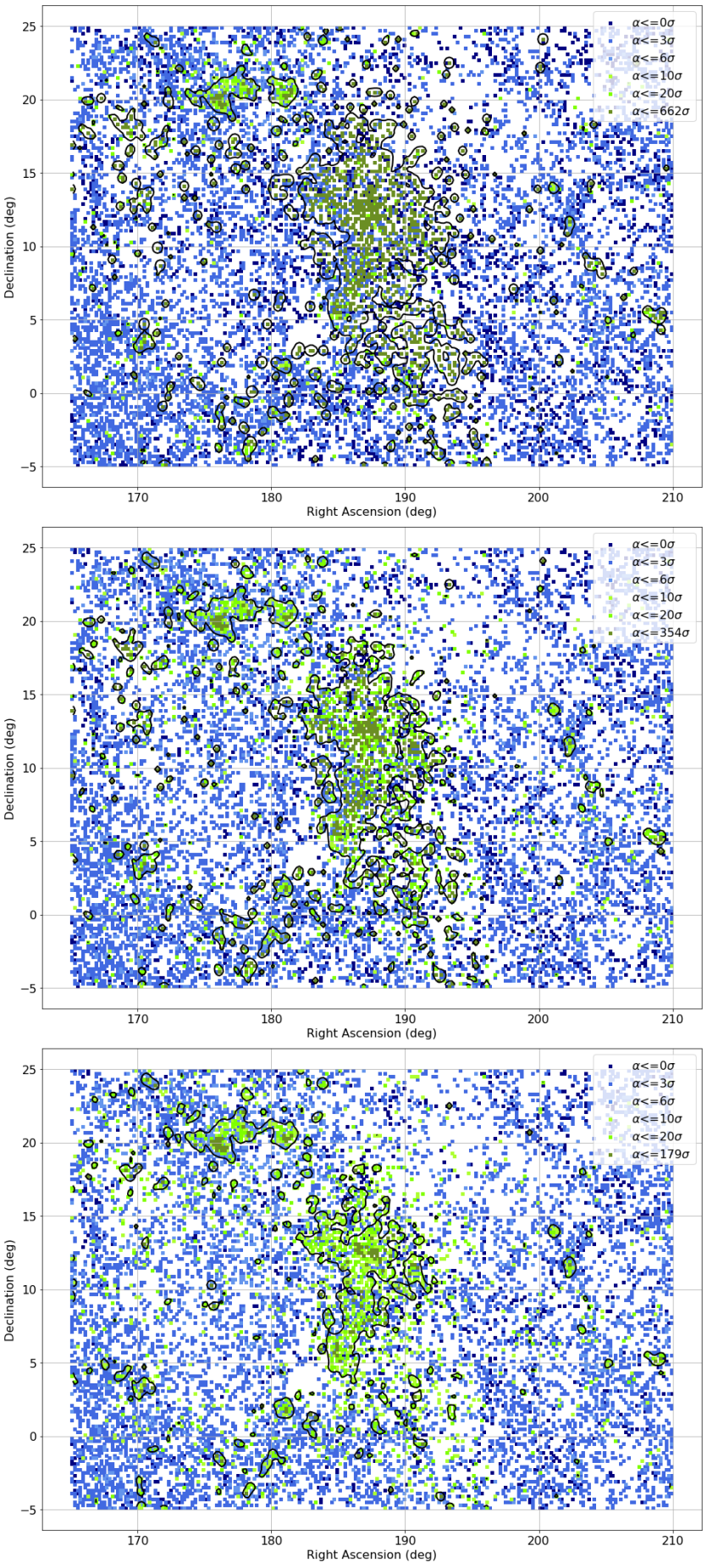

FIGURE 4.4: Density maps with significance threshold 3σ. Top: k=1000, middle: k=1500, bottom: k=2000. The black lines are the boundaries that surround the over-dense regions

In figures 4.3 and 4.4 we see that for the smaller values of k, the resulting boundaries surrounding the Virgo cluster are quite stable. However, when k becomes larger (i.e. k=2000) the Virgo boundary begins to shrink. This happens because for regions that are further away from the center of the cluster, the kNN algorithm will have to calculate the density around each point for a much larger radius that will have to include galaxies in much larger distances and therefore the density calculated will become smaller.

Apart from the Virgo cluster, we can see that there are additional overdense regions detected by the algorithm and we know that one of these is the Leo cluster centered at the coordinates mentioned in paragraph 3.2. This proves that the algorithm is sensitive enough to identify multiple clusters at the same time, as well as defining their boundaries to a very good similarity with the real shape of the respective clusters.

4.4 Total significance of Virgo region

We calculate the total significance α of the region inside the contour corresponding to the Virgo cluster, from equation 3.1. This calculation is done for each k, to find the k that maximizes the total significance α. In figure 4.5 we plot the total significance as a

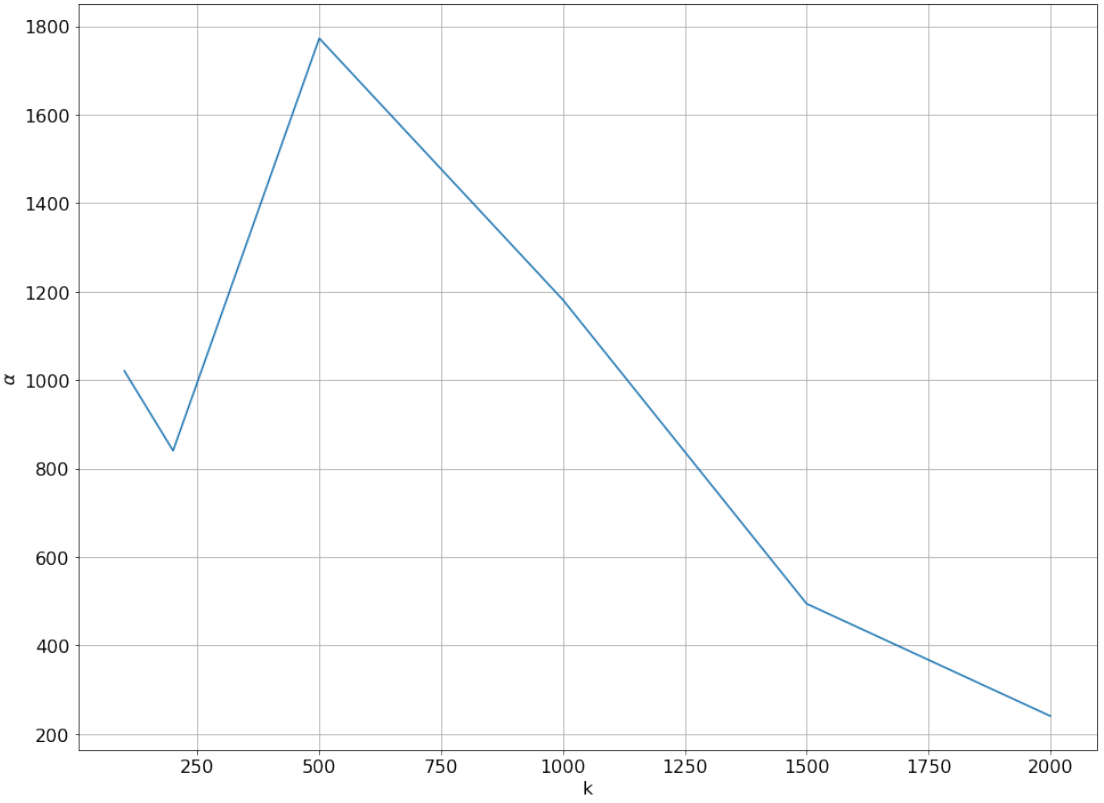

FIGURE 4.5: Statistical significance α vs k for the Virgo region

function of k, for the contour enclosing the region of the Virgo cluster. We see that the maximum statistical significance is $\alpha_{\max}$=1763.76 and it is achieved for k=500. In order to obtain a more quantitative evaluation of the algorithm's performance in calculating the Virgo boundaries, we count the number of EVCC galaxies enclosed by the contour for the optimal k=500. The results are shown in figure 4.6. The EVCC galaxies included

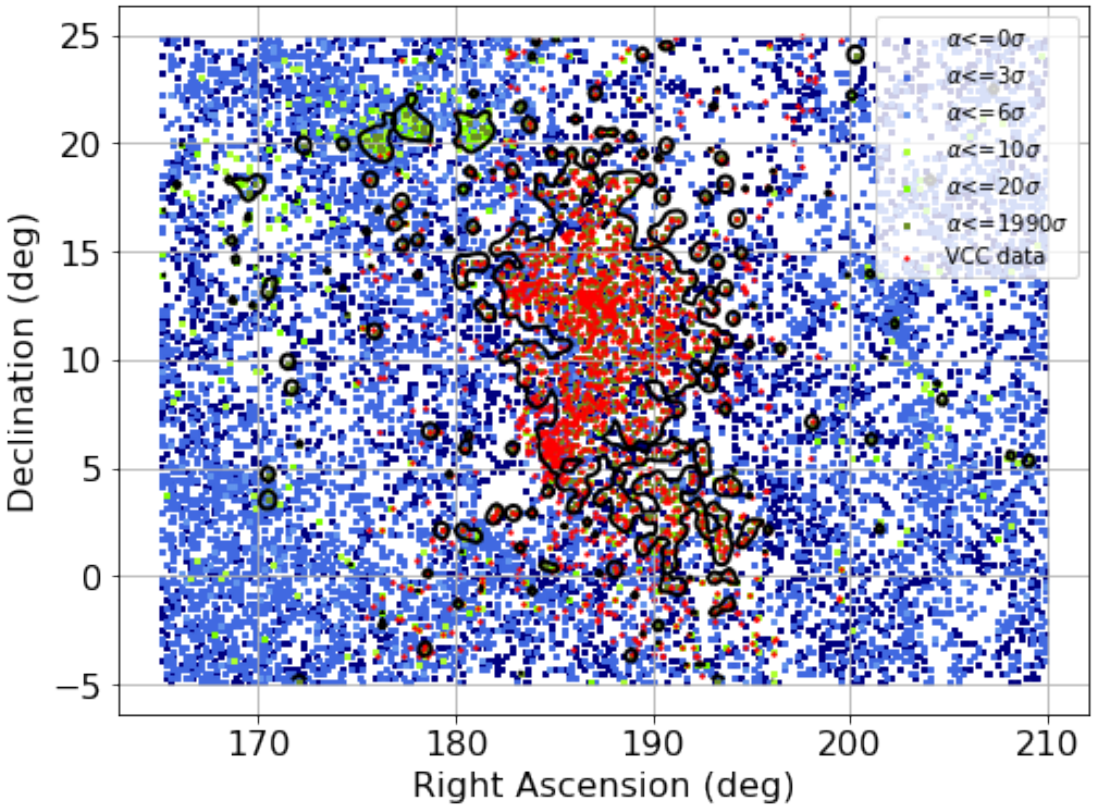

FIGURE 4.6: EVCC galaxies inside the contour enclosing the Virgo region in HECATE for k=500

in the cluster are 60% of the total catalogue. The reason for this is that the other overdense structures, namely the ones immediately below the Virgo cluster in figure 4.6, also contain a significant number of galaxies from EVCC. These regions could actually be parts of the Virgo but due to the 2-dimensional binning and contour creation they appear as seperate regions. However in general, the Virgo boundaries, defined based on our method, have a striking resemblance to the actual shape of the Virgo cluster. In table 4.1 we present the total statistical significance within the predicted Virgo cluster region for different significance thresholds for the pixels in the outskirts of the cluster, and which are used to define its boundary. The trend remains the same, since by setting the boundaries at larger confidence intervals, the area of the cluster shrinks, containing the regions with the highest possible statistical significance.

TABLE 4.1: Total statistical significance in Virgo region for different values of k and different confidence intervals

	α_{tot} in Virgo region					
	k=100	k=200	k=500	k=1000	k=1500	k=2000
$\alpha \geq 2\sigma$	1001.06	808.79	1695.22	1149.35	503.50	259.26
$\alpha \geq 3\sigma$	1020.80	840.42	1772.65	1181.39	494.52	240.64
$\alpha \geq 5\sigma$	1083.06	867.13	1828.75	1167.23	500.28	210.24
$\alpha \geq 10\sigma$	1156.47	896.82	1902.91	1203.79	371.57	166.36
$\alpha \geq 20\sigma$	1324.01	923.22	1995.95	1151.43	372.00	148.54

4.5 Density maps for dark matter halos

The second test of the algorithm is done on dark matter halo simulations. Since, for the observational data from HECATE the algorithm's performance is very promising, we consider it eligible for evaluation on mock observations. As in the previous part, the first step is the creation of the density maps. We do this for each of the 14 halos included in the subset for the same values of k mentioned in paragraph 3.5. In figures 4.7, 4.8 and 4.9 we present the density maps for the first 3 halos in the subset, along with the boundaries of the detected clusters.

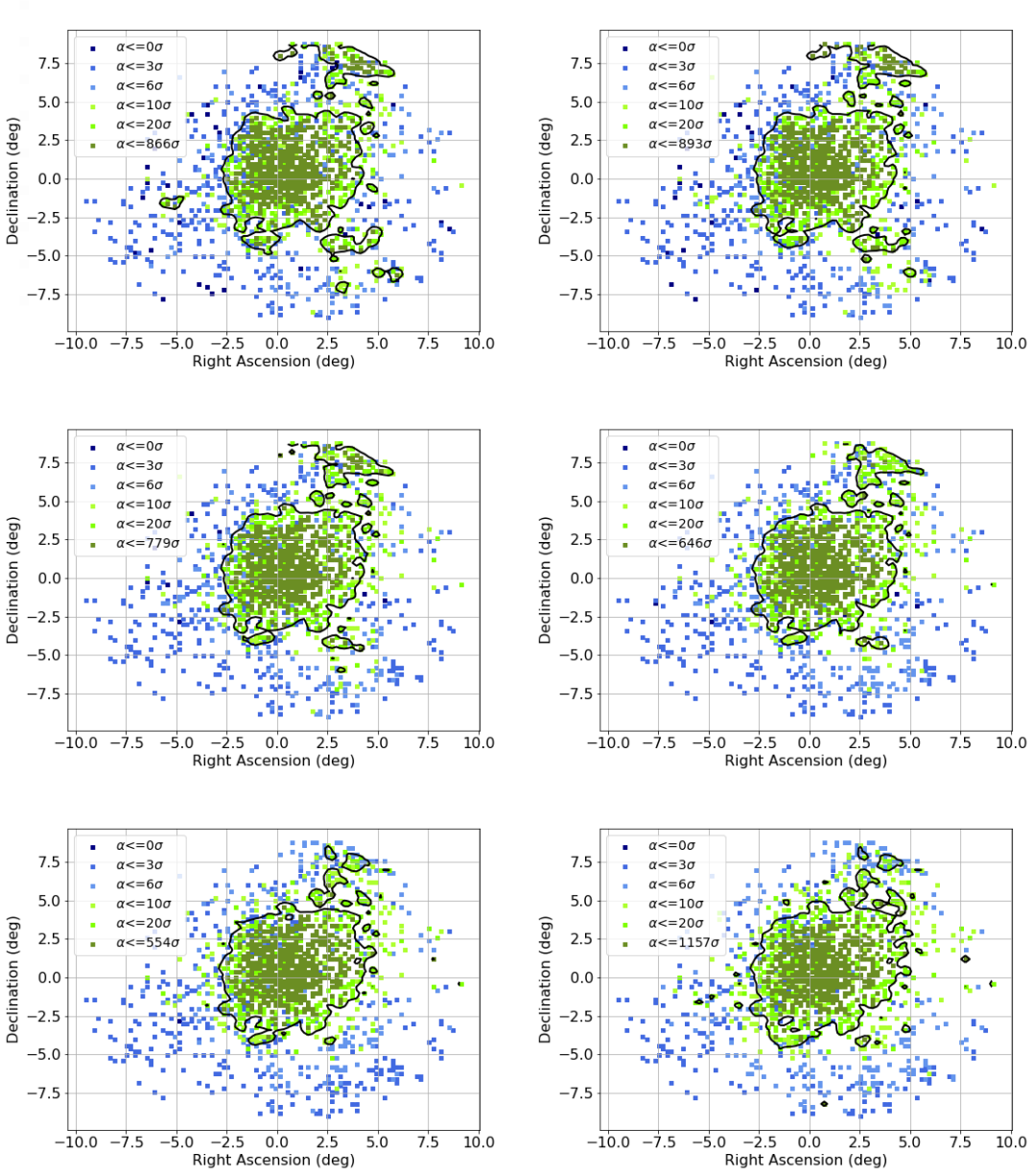

FIGURE 4.7: Density maps for halo 1 showing the regions with $\alpha > 3\sigma$ for each k. Upper left:k=10, Upper right:k=20, Middle left:k=50, Middle right:k=100, Bottom left:k=200, Bottom right:k=500

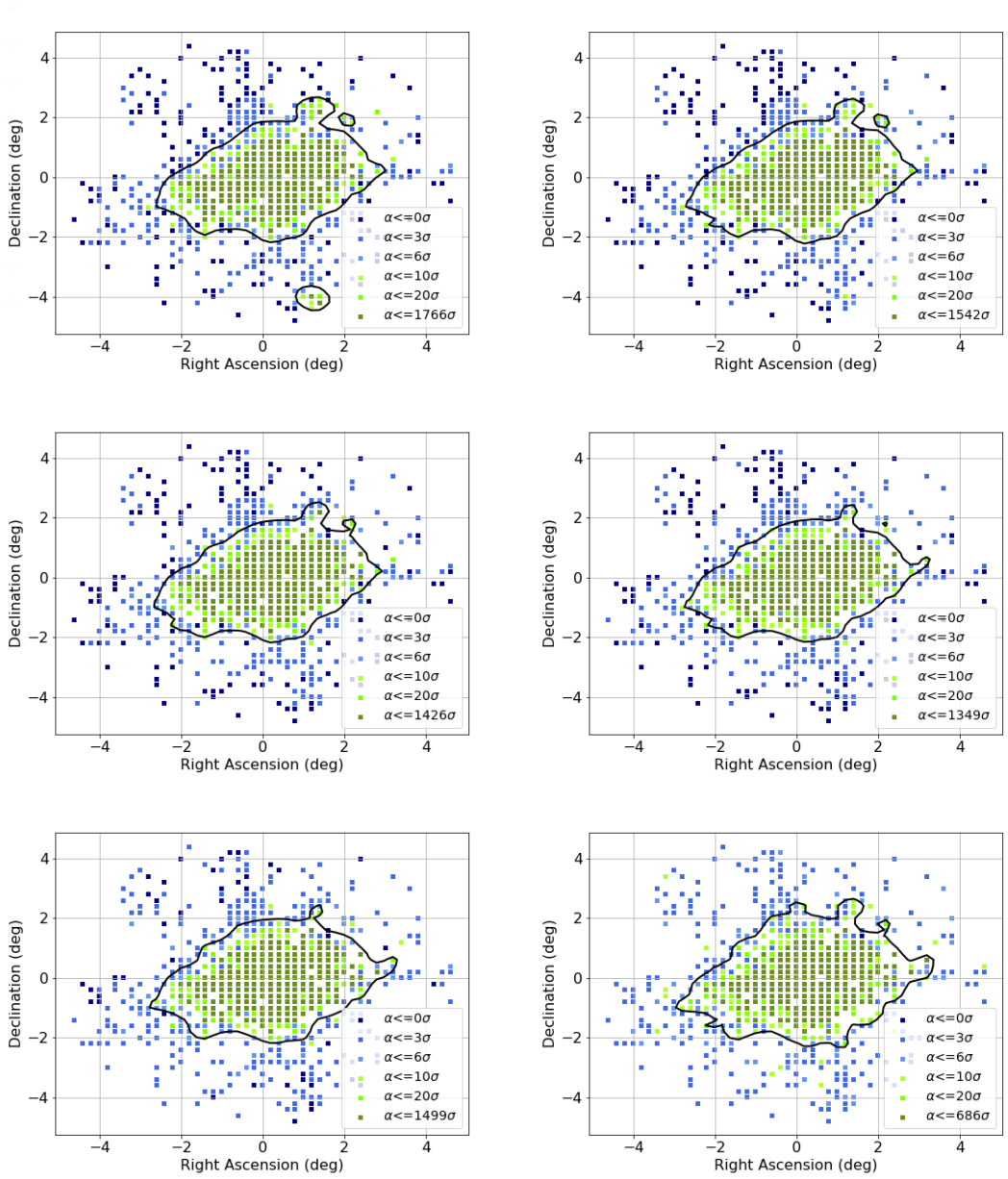

FIGURE 4.8: Density maps for halo 2 showing the regions with $\alpha > 3\sigma$ for each k. Upper left:k=10, Upper right:k=20, Middle left:k=50, Middle right:k=100, Bottom left:k=200, Bottom right:k=500

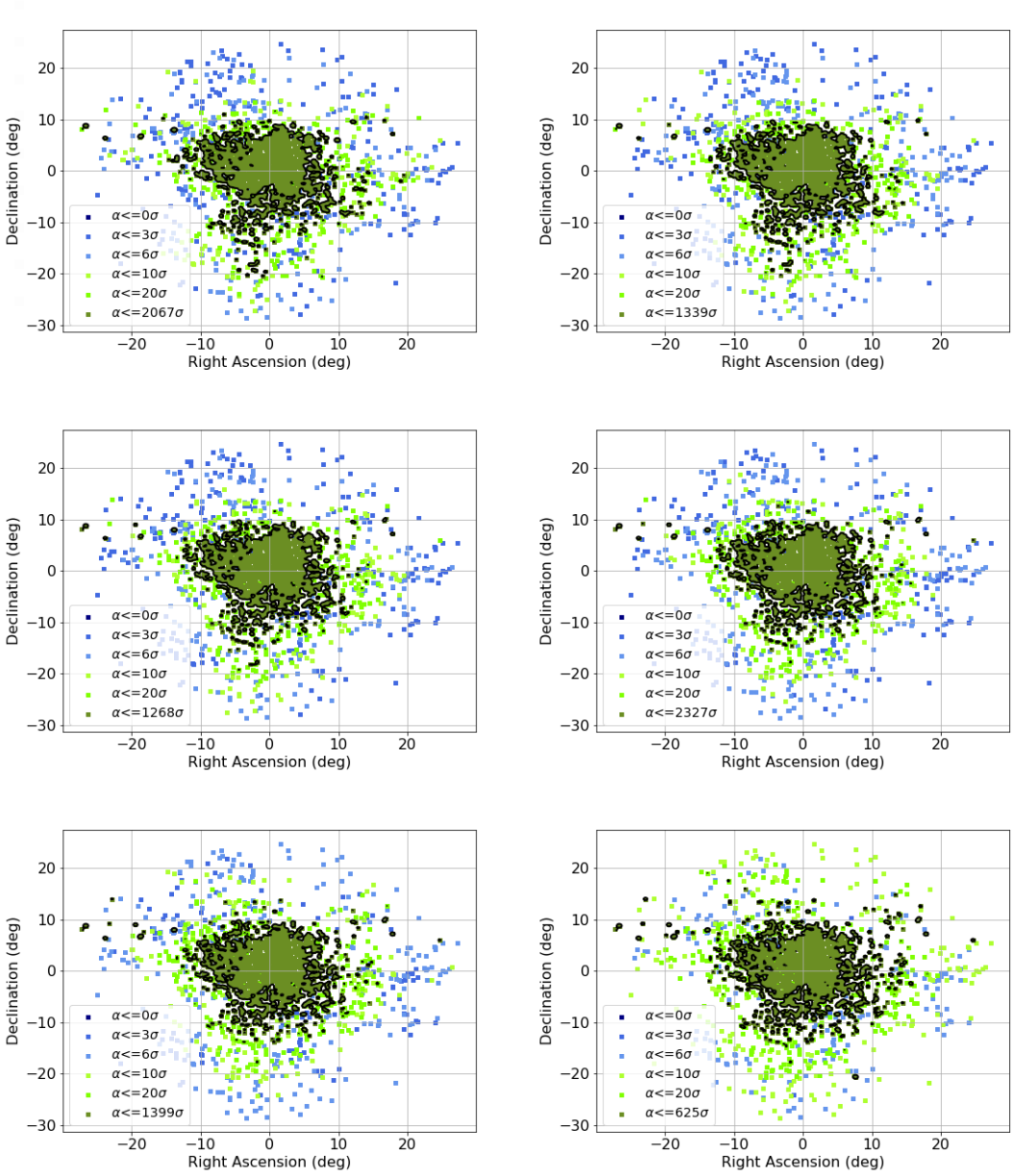

FIGURE 4.9: Density maps for halo 3 showing the regions with $\alpha > 3\sigma$ for each k. Upper left:k=10, Upper right:k=20, Middle left:k=50, Middle right:k=100, Bottom left:k=200, Bottom right:k=500

4.6 Radius covered by cluster boundaries

In the case of the mock halo samples, we can test the algorithm with regard to its efficiency at including objects within a radius of R_{200}. This will give us a picture of the algorithm's behavior in terms of defining the outline of the clusters.

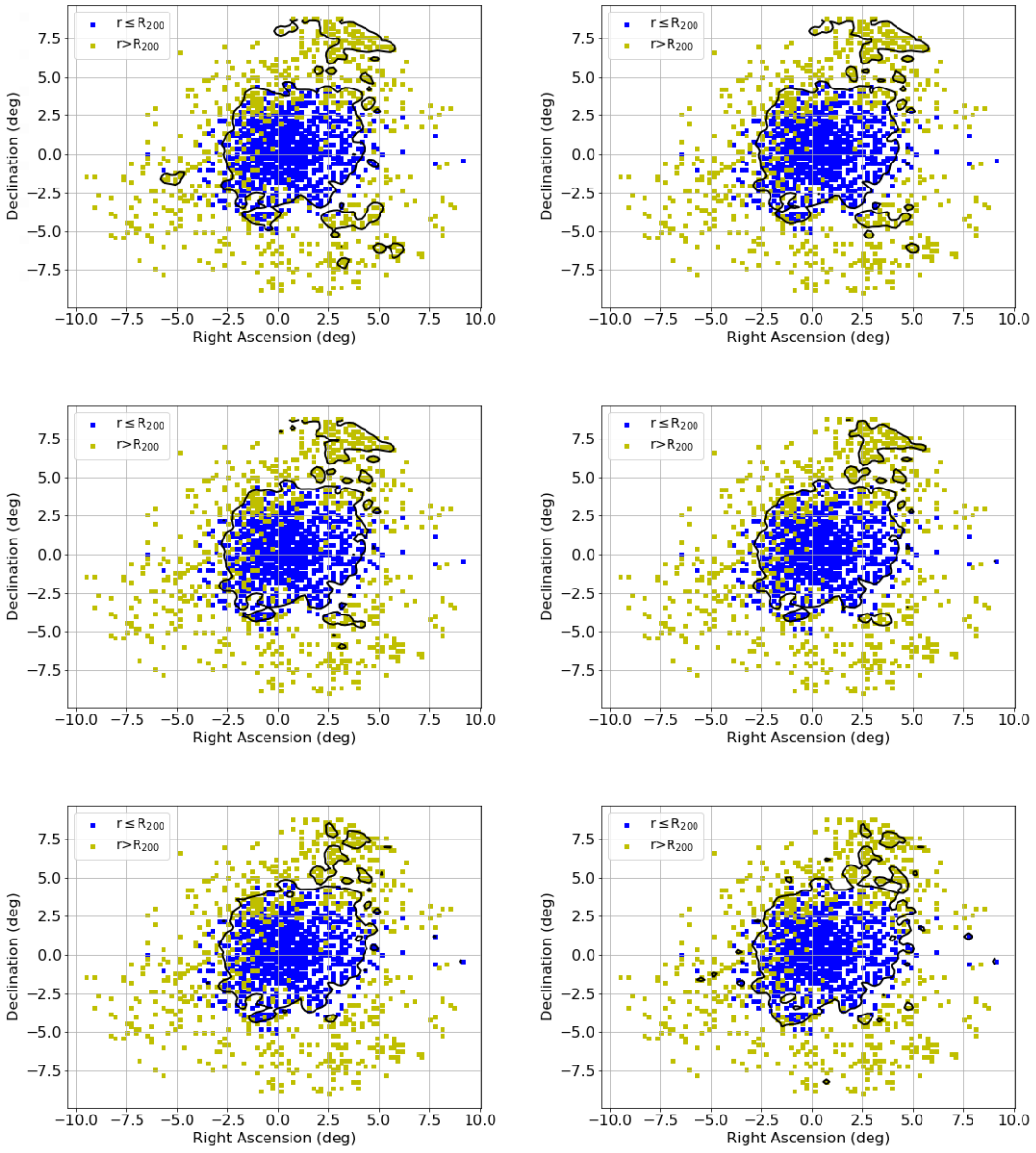

FIGURE 4.10: Map depicting the cluster boundary of halo 1 and the bins colored according to their distance from the center of the halo for each k. Upper left:k=10, Upper right:k=20, Middle left:k=50, Middle right:k=100, Bottom left:k=200, Bottom right:k=500

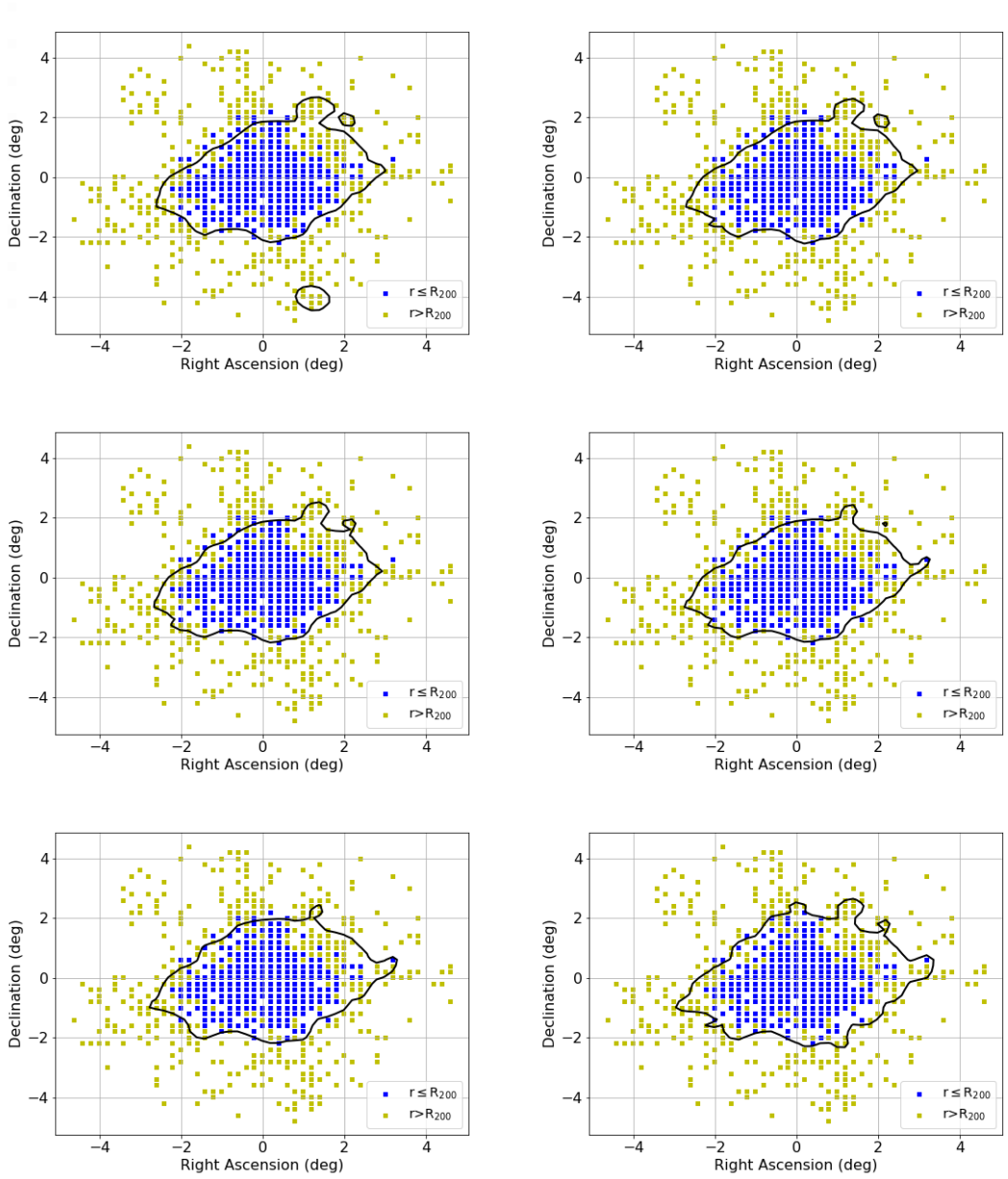

FIGURE 4.11: Map depicting the cluster boundary of halo 2 and the bins colored according to their distance from the center of the halo for each k. Upper left:k=10, Upper right:k=20, Middle left:k=50, Middle right:k=100, Bottom left:k=200, Bottom right:k=500

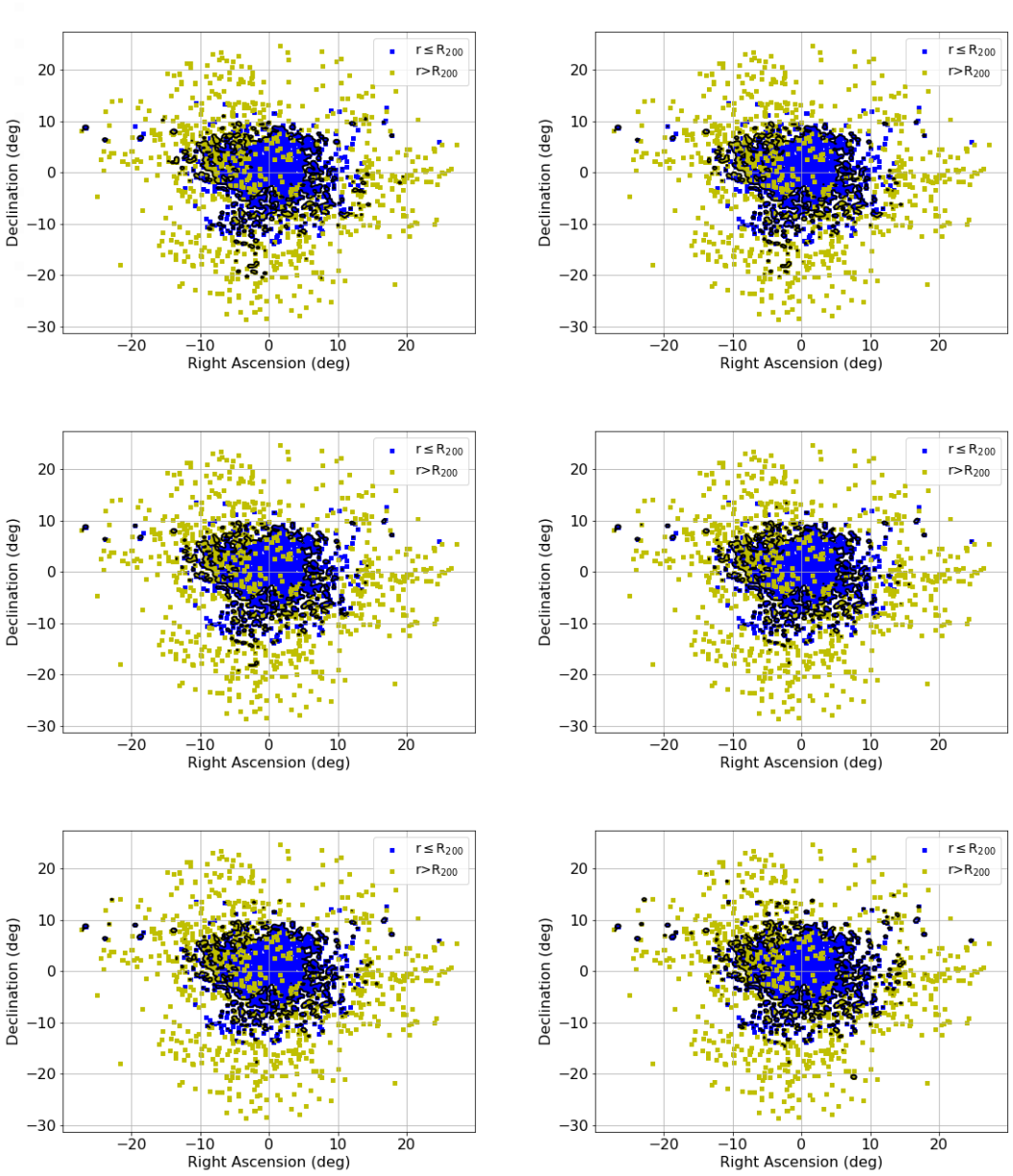

FIGURE 4.12: Map depicting the cluster boundary of halo 3 and the bins colored according to their distance from the center of the halo for each k. Upper left:k=10, Upper right:k=20, Middle left:k=50, Middle right:k=100, Bottom left:k=200, Bottom right:k=500

We can see that for each of the 3 halos presented, the majority of the objects that are located at a distance smaller than R_{200} lie within the contours we defined. However, there are some galaxies for each halo that are found inside the boundary, but whose distance is larger than R_{200}. Their number is not very large though and most of the objects inside the contour are indeed within a distance of R_{200} from the center. The results presented, refer to the first 3 halos that are the most populous (N_{obj}:1900-2000) and the algorithm provides stable results for them. For most of the remaining halos, whose number of objects is between 70 and 200, the algorithm's result is unstable and does not provide a certain cluster detection.

Chapter 5

Conclusions

We have implemented a variation of the kNN algorithm for the numerical detection of clusters in galaxy data in both observational (HECATE) and mock catalogues (TNG project). The resulting density maps, combined with a comparison to a uniform background using Monte Carlo samples, led to the detection of clusters, with boundaries that were determined in order to cover the regions with a statistical significance greater that 3 standard deviations. For the HECATE galaxies, the algorithm exhibited a very good performance at detecting overdensities since, for different values of k, clusters in the same parameter space with different shapes and sizes, were detected simultaneously, namely the Virgo and Leo clusters. Furthermore, the boundaries of the overdensities were similar to the true shape of the corresponding clusters. For example, the Leo cluster has regular shape and in figures 4.3 and 4.4 we see that in these coordinates, the overdensity has a near circular shape in 2-dimensions. On the other hand, the Virgo cluster is less regular with two subclumps, which the algorithm successfully detects as parts of a single cluster. For the dark matter halos our goal was to evaluate the performance of the algorithm at correctly defining the boundaries of the clusters. In order to do that we counted how many objects within a distance of R_{200} are included in the detected cluster. We found that for the first 3 most populous halos, most of the objects are inside the boundary which means that the algorithm is effective at identifying a region close to the R_{200} radius of the cluster. For smaller halos though, due to their small population, the algorithm is not effective at providing a stable result for cluster boundaries.

Bibliography

[1] Bradley W Carroll and Dale A Ostlie. *An introduction to modern astrophysics; 2nd ed.* Addison-Wesley, San Francisco, CA, 2007. URL https://cds.cern.ch/record/1009754.

[2] https://ned.ipac.caltech.edu/level5/Sept11/Norman/Norman4.html.

[3] A. Dressler. Galaxy morphology in rich clusters: implications for the formation and evolution of galaxies. , 236:351–365, March 1980. doi: 10.1086/157753.

[4] K Kovlakas, A Zezas, J J Andrews, A Basu-Zych, T Fragos, A Hornschemeier, K Kouroumpatzakis, B Lehmer, and A Ptak. The Heraklion Extragalactic Catalogue (HECATE): a value-added galaxy catalogue for multimessenger astrophysics. *Monthly Notices of the Royal Astronomical Society*, 506(2):1896–1915, 06 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab1799. URL https://doi.org/10.1093/mnras/stab1799.

[5] G. Helou, B. F. Madore, M. Schmitz, M. D. Bicay, X. Wu, and J. Bennett. *The NASA/IPAC Extragalactic Database*, pages 89–106. Springer Netherlands, Dordrecht, 1991. ISBN 978-94-011-3250-3. doi: 10.1007/978-94-011-3250-3_10. URL https://doi.org/10.1007/978-94-011-3250-3_10.

[6] Dmitry Makarov, Philippe Prugniel, Nataliya Terekhova, Hélène Courtois, and Isabelle Vauglin. HyperLEDA. III. The catalogue of extragalactic distances. , 570: A13, October 2014. doi: 10.1051/0004-6361/201423496.

[7] Suk Kim, Soo-Chang Rey, Helmut Jerjen, Thorsten Lisker, Eon-Chang Sung, Youngdae Lee, Jiwon Chung, Mina Pak, Wonhyeong Yi, and Woong Lee. THE EXTENDED VIRGO CLUSTER CATALOG. *The Astrophysical Journal Supplement Series*, 215(2):22, dec 2014. doi: 10.1088/0067-0049/215/2/22. URL https://doi.org/10.1088/0067-0049/215/2/22.

[8] R. D'Abrusco, G. Fabbiano, and A. Zezas. SPATIAL STRUCTURES IN THE GLOBULAR CLUSTER DISTRIBUTION OF THE 10 BRIGHTEST VIRGO

GALAXIES. *The Astrophysical Journal*, 805(1):26, may 2015. doi: 10.1088/0004-637x/805/1/26. URL `https://doi.org/10.1088/0004-637x/805/1/26`.

[9] Dylan Nelson, Annalisa Pillepich, Volker Springel, Rainer Weinberger, Lars Hernquist, Rüdiger Pakmor, Shy Genel, Paul Torrey, Mark Vogelsberger, Guinevere Kauffmann, Federico Marinacci, and Jill Naiman. First results from the IllustrisTNG simulations: the galaxy colour bimodality. , 475(1):624–647, March 2018. doi: 10.1093/mnras/stx3040.

[10] Annalisa Pillepich, Dylan Nelson, Lars Hernquist, Volker Springel, Rüdiger Pakmor, Paul Torrey, Rainer Weinberger, Shy Genel, Jill P. Naiman, Federico Marinacci, and Mark Vogelsberger. First results from the IllustrisTNG simulations: the stellar mass content of groups and clusters of galaxies. , 475(1):648–675, March 2018. doi: 10.1093/mnras/stx3112.

[11] Volker Springel, Rüdiger Pakmor, Annalisa Pillepich, Rainer Weinberger, Dylan Nelson, Lars Hernquist, Mark Vogelsberger, Shy Genel, Paul Torrey, Federico Marinacci, and Jill Naiman. First results from the IllustrisTNG simulations: matter and galaxy clustering. , 475(1):676–698, March 2018. doi: 10.1093/mnras/stx3304.

[12] Jill P. Naiman, Annalisa Pillepich, Volker Springel, Enrico Ramirez-Ruiz, Paul Torrey, Mark Vogelsberger, Rüdiger Pakmor, Dylan Nelson, Federico Marinacci, Lars Hernquist, Rainer Weinberger, and Shy Genel. First results from the IllustrisTNG simulations: a tale of two elements - chemical evolution of magnesium and europium. , 477(1):1206–1224, June 2018. doi: 10.1093/mnras/sty618.

[13] Federico Marinacci, Mark Vogelsberger, Rüdiger Pakmor, Paul Torrey, Volker Springel, Lars Hernquist, Dylan Nelson, Rainer Weinberger, Annalisa Pillepich, Jill Naiman, and Shy Genel. First results from the IllustrisTNG simulations: radio haloes and magnetic fields. , 480(4):5113–5139, November 2018. doi: 10.1093/mnras/sty2206.