UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

Exploration of unsupervised machine learning methods to study galaxy clustering

Ana Sofia Chagas Carvalho

Mestrado em Física
Especialização em Astrofísica e Cosmologia

Dissertação orientada por:
António da Silva, Alberto Krone-Martins

2019

Resumo

Enxames de galáxias são objetos essenciais para a compreensão da evolução de galáxias, mas também são fundamentais para questões sobre o setor escuro do universo. Todavia, o estudo de enxames assenta na correta identificação das galáxias que lhe pertencem. A missão espacial Euclid tem como objetivo explorar o setor escuro do universo, identificando assinaturas da taxa de expansão do universo e da evolução de estruturas cósmicas, observando o universo até redshift $z = 2$. Para isso, vão ser estudados e medidos efeitos de lentes gravitacionais em galáxias e também propriedades de agrupamentos de galáxias. No entanto, a quantidade de dados que a missão Euclid irá coletar (e também os já coletados por outras missões cosmológicas como SDSS, DES, LSST, etc.) é demasiado grande, impedindo a aquisição de informação espectroscópica detalhada para todas as galáxias detetadas que é necessária para identificar as galáxias membro de enxames e a sua distribuição, que são essenciais para derivar as propriedades destes. Portanto, o desenvolvimento de técnicas de análise de dados que permite o estudo de enxames diretamente de dados astrométricos e fotométricos, usando o mínimo de informação espetroscópica possível, tem um grande valor para a extração de informação cosmológica.

Este projecto tem como objectivo o estudo de novos métodos para a identificação de membros de enxames de galáxias de forma não supervisionada, sendo que métodos já existentes serão também adotados e modificados. O primeiro capítulo introduz a noção de aglomerados de galáxias, a sua definição inicial (mais do que 50 membros ligados gravitacionalmente, dentro de um diametro de cerca $1.5h^{-1}$ Mpc ou maior). Depois, são referidas algumas propriedades observáveis (no ótico, raios-X, etc), como a luminosidade, a riqueza, cor, contagem de membros, entre outras, havendo uma especial atenção para a luminosidade nos raios-x e para a luminosidade observada devido ao efeito Sunyaev-Zel'dovich. Estas propriedades são estudadas e usadas para o desenvolvimento, confirmação e comparação de aspectos teóricos de cosmologia, em partícular, sobre a matéria escura. Esta matéria escura foi primeiro deduzida por Zwicky in 1933. Tendo em conta modelos dinâmicos e teoria Virial, estimou-se que a massa total de enxames de galáxias é bastante maior da estimada quando se estuda a luz proveniente de objectos luminosos (a maioria sendo estrelas e gás) que constituem as galáxias de um enxame. Esta última é cerca de 3% a 5% da massa total estimada a partir dos modelos dinâmicos. Ao excesso dessa massa deu-se o nome de matéria escura e a partir daí a evidência da existência de um tipo de massa que não é observada com a tecnologia de hoje constituiu um desafio para a cosmologia. Enxames de galáxias não constituem apenas sondas para a matéria escura, mas para estudar o desenvolvimento da Estrutura de Grande Escala, estrutura filamental (de matéria, como galáxias, grupos e enxames) que resultou de perturbações do campo de inflação que foram amplificadas pela gravidade, e também como sondas para estudar a energia escura e teorias de gravidade modificada. Existem diversas missoes e sondas que procuram observar e/ou detectar enxames de galáxias precisamente para estudar o setor escuro do universo. É o caso da missão Planck, que esteve activa durante 30 meses. Com a missão Planck foi possível combinar enxames de galáxias num catálogo, o PLANCKSZ2 (Planck 2nd Sunyaev-Zeldovich Source), do DES (Dark Energy Survey) cujo objectivo científico é estudar a origem do universo acelerado e também da matéria escura, do SDSS (Sloan Digital Sky Survey) com a qual foi possível construir diversos catálogos, como o eBOSS, que contem quasars e galaxias. As sondas que operam nos raios-x (ROSAT e XMM-Newton) são também bastante importantes para a detecção de enxames de galáxias, uma vez que o gás contido nestes objectos (o intracluster medium ou ICM) emite

radiação devido ao efeito de bremmstrahlung. Neste capítulo são também mencionados métodos de identificação de enxames de galáxias, virados para sondas no ótico, como é o caso da futura missão Euclid. Neste capítulo são também brevemente explicados alguns conceitos necessários à elaboração do trabalho mencionado nesta dissertação, como os conceitos relacionados com Machine Learning e algumas ferramentas matemáticas e estatísticas.

No segundo capítulo é explicado o método utilizado nesta dissertação, o método UPMASK. Este é um método não supervisionado e desenhado para utilizar a mínima informação possivel sobre os dados fotométricos e astrométricos, sem realizar nenhuma suposição dependente de modelos dos objectos que se está a estudar. Para além de utilizar o UPMASK, tal como é, em enxames de galáxias, foram também realizadas modificações a este método com o objectivo de o aprimorar - particularmente no seu tempo de execução. Para esse efeito, são usadas ferramentas como a Tesselação de Voronoi e o teste estatístico Anderson-Darling, e funções de regressão.

De seguida, no terceiro capítulo, são aplicadas todas as versões modificadas do método UPMASK, bem como a versão original, a dados simulados que foram gerados tendo em conta também outras simulações para a Estrutura de Grande Escala. É definida uma pureza e completude e com estes parâmetros, vão ser realizados estudos dos parâmetros internos ao método, bem como estudos acerca do seu tempo de execução. Neste capítulo são também estudadas as diferenças entre usar um sistema de filtros idêntico ao do DES e um sistema de filtros idêntico ao do Euclid e para que redshifts de enxames de galáxias a utilização deste filtros é optima.

No quarto capítulo o método UPMASK e todas as suas versões são aplicadas ao enxame de galáxias mais estudado, o enxame Coma. Este enxame é um dos mais famosos, pois é numeroso um dos mais próximos. Foi também com este enxame que Zwicky demonstrou que existe uma fração de massa dinâmica que não emite luz. Retiraram-se os objectos do catálogo Pan-STARRS (Panoramic Survey Telescope and Rapid Response System) dentro de um campo que se sabe em que o enxame Cluster está situado. Depois de separar estrelas de galáxias, foi então aplicado o método aos dados, desta vez sem informação sobre pureza e completude, uma vez que o catálogo utilizado nao fornece informações sobre os membros pertencentes. Assim desta forma no quinto capítulo, aplicou-se novamente o método, mas desta vez com o objectivo de re-encontrar os enxames de galáxia que foram identificados pela sonda Planck (utilizando o catálogo PLANCKSZ2), utilizando o catálogo Pan-STARRS. Assim, para cada enxame, foram retirados objectos num campo que é compatível com as coordenadas fornecidas pelo PLANCKSZ2, e separadas também as estrelas das galáxias, utilizando a mesma metodologia adoptada no capítulo anterior. Executando os testes, teve-se especial atenção aos enxames que não foram confirmados por uma fonte externa ao Planck - pois as propriedades destes "novos enxames" poderão contribuir para a determinação e a aprimoração de ou outros avanços na área da cosmologia.

Finalmente, no sexto capítulo apresento os resultados e conclusões obtidos ao longo deste trabalho de dissertação, a importância que este trabalho tem para o conhecimento científico, pois fornece uma ferramenta que procura tornar eficiente a seleção e análise de enxames de galáxias, numa era de "Big Science", onde é humanamente impossível todos os dados serem analisados por mãos humanas. Neste capítulo procura-se também discutir oportunidades para trabalho futuro, desde a implementação de possíveis ferramentas mais simples e computacionalmente mais rápidas até à possível observação dos enxames de galáxia contidos no PLANCKSZ2, sem uma validação externa, mas que foram identificados pelo UPMASK.

Palavras-chave: métodos: análise de dados, métodos: estatísticos, galáxias: enxames: geral.

Abstract

Galaxy clusters are essential objects to understand galaxy evolution. Moreover, they are fundamental in the quest to unravel the Dark Sector of the Universe. Nevertheless, their study relies on the correct identification of whether galaxies are members of the cluster or not. The space survey Euclid, has as one of its goals to probe the Dark Sector of the Universe by detecting signatures of the expansion rate of the Universe and the growth of cosmic structures. For this purpose two main probes will be used: gravitational lensing effects on galaxies and the properties of galaxy clusters. However, the amount of data that will be collected by Euclid and by existing and future large cosmological surveys as SDSS, DES, LSST, etc., is big enough to prevent gathering detailed spectroscopic information for all the detected galaxies, and thus to obtain the membership that is essential to derive the properties of these clusters. Accordingly, the development of data analysis techniques that enable the study of clusters directly from the astrometric and photometric data, using a minimum amount of spectroscopy, is highly valuable for extraction of cosmological information in this era of large surveys and precision cosmology. This project has the study of new methods of unsupervised membership assignment in galaxy clusters as its center, while also adopting and modifying existing models. I studied and modified the UPMASK method, whose development is described in this dissertation. This method and its modifications were validated using simulated data of MICECAT, and an extensive study of parameters of the test was done. Later, UPMASK was applied to the Coma Cluster, using observations and measurements from the Pan-STARRS survey catalogue. Finally, the method was used to rediscover the galaxy clusters of the PLANCKSZ2 catalogue within the Pan-STARRS survey catalogue - taking particular attention to the ones that were not validaded by an external source to the PLANCKSZ2.

Keywords: methods: data analysis, methods: statistical, galaxies: clusters: general.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Galaxy clusters are essential objects to understand galaxy evolution and the formation of the cosmic structure on large scales. Moreover, they are fundamental in the quest to unravel the dynamics of the so called "Dark Sector" (eg. Dark Energy and Dark Matter) of The Universe. In section 1.1 I introduce the main objectives and the motivation behind this dissertation. In section 1.2 I discuss Galaxy Clusters, by presenting their properties and importance for cosmology, as well as the main cluster surveys that can be used for that purpose. section 1.3 shortly reviews some methods that identify Galaxy Clusters in optical surveys and in section 1.4 I briefly describe the statistical tools and the main concepts underlying the methods I developed.

1.1 Motivation

Galaxy Clusters are fundamental objects to study Cosmology, Large-Scale Structure (LSS) and to understand galaxy evolution as a function of environment Nevertheless, their study using optical surveys relies on the correct identification of whether galaxies are members of the cluster or not. This is commonly known as the membership assignment problem. If individual galaxy distances were perfectly known this would be a trivial problem to solve. However in the vast majority of the cases distance estimations are much worse than the cluster size plaguing any naif solution for the identification problem.

The Euclid space survey, aims to probe the Dark Sector of the Universe by detecting signatures of the expansion rate of the Universe and the growth of cosmic structures. For this purpose two main probes will be used: gravitational lensing effects on galaxies and the properties of galaxy clustering. However, the amount of data that will be collected by Euclid and by existing and future large cosmological surveys as SDSS, DES, LSST, etc., will be very large and will lack detailed spectroscopic information for all the detected galaxies, and thus to obtain galaxy membership that are essential to derive the optical properties of these clusters. Accordingly, the development of data analysis techniques that enable the study of clusters directly from the astrometric and photometric data, using a minimum amount of spectroscopic data is of utmost importance to extract valuable cosmological information in this era of large galaxy surveys and precision cosmology.

During this dissertation I present a new cluster identification method based on the UPMASK algorithm [Krone-Martins and Moitinho, 2014, 2015] and validate it on simulations of astrometry and photometry for the ESA/Euclid space mission and also on other relevant ground-based surveys that will provide additional information for the Euclid Survey. In the final part of

the dissertation I apply the developed method to real data from existing sky surveys trying to detect new clusters and to study previously known galaxy clusters. The work presented in this dissertation is most relevant for extragalactic astronomy studies involving galaxy clusters, and uses modern statistical learning methods, thus opening a broad range of academic and industrial opportunities and contributing to important tasks of the Portuguese participation in the ESA/Euclid space mission survey.

1.2 Galaxy Clusters

1.2.1 Cluster Properties

Galaxy clusters and galaxy groups are collapsed structures that are historically defined in the optical by the number of gravitationally bound galaxies. A group typically consists of less than 50 members in a diameter smaller than $1.5\,h^{-1}\,\mathrm{Mpc}$, while galaxy clusters are more massive gravitational structures, consisting of more than 50 members within a diameter of about $1.5\,h^{-1}\,\mathrm{Mpc}$ and larger [Bower and Balogh, 2004, Mamon, 1996], with $h = H_0/100\,km\,s^{-1}\,\mathrm{Mpc}^{-1}$ being the dimensionless Hubble parameter. The typical population of a galaxy cluster contains elliptical galaxies in the central region with the remaining population being composed by spirals or irregulars type galaxies [Dressler, 1980]. However not all clusters follow this pattern, leading to galaxy cluster classifications ruled by cluster population. Galaxy clusters also have different classifications such as, for example:

- Abell classification [Abell, 1965]: a cluster is regular if it is circularly symmetric with predominantly Elliptical (E) and/or lenticular galaxies (S0) populating the center and irregular if the cluster has more spirals or a less defined structure;
- Bautz-Morgan classification [Bautz and Morgan, 1970]: cluster is of type I if it has a center supergiant elliptical (cD) galaxy, type II if the central galaxy is between a cD or a giant elliptical galaxy, and type III if the cluster has no dominant central galaxy;
- Oemler classification [Oemler, 1974]: a cluster is a cD cluster if it shows one or two dominant cD galaxies, a spiral rich if it has a proportion of E(Elliptical): S0(Lenticular): S(Spiral) of 1:2:3, and a spiral poor if it shows no dominant cD and has a proportion of E:S0:S of 1:2:1.

In optical surveys, besides studying the morphology of the galaxy members of a cluster, the main observables are the luminosity, color and richness. The richness is the number of galaxies associated to the cluster. It is not an easy determination since it is difficult to determine with precision the cluster membership. There are some known ways to define the richness. This parameter was first defined by Abell [Abell, 1958] as the number of galaxies within 2 magnitudes from the third most luminous galaxy and within 3 Mpc in radius. On the other hand, Zwicky [Zwicky et al., 1961] also defined the richness of a cluster as the number of galaxies inside the isopleth (a contour at which the cluster surface brightness is twice the local background magnitude).

Galaxy Clusters are important probes of dark matter. Models for these objects predict a total mass that is higher than the mass estimated from the light of the stars within the member galaxies [e.g. Smith, 1936, Zwicky, 1933] – $\sim 3-5\%$ from the total predicted mass. The excess

Figure 1.1: Abell 1835 ($z = 0.25$) in different wavelenghts [Allen et al., 2011]. Left - X-ray (Chandra X-ray Observatory/A. Mantz). Center - Optical (Canada France Hawaii Telescope/A. von der Linden et al.). Right - SZ (Sunyaev Zel'dovich Array/D. Marrone.)

of "unseen" mass was then called dark matter, or "*dunkle Materie*", by F. Zwicky [Zwicky, 1933]. Due to the high mass of these objects, space-time is bent, distorting the shapes of the objects in the sky for the observer. From images of gravitational lenses caused by a galaxy cluster, it is possible to estimate the total mass of a cluster, which has been found to be greater than the mass from the stars and gas of the cluster. Moreover the evidence for dark matter is not only highlighted by the kinematics of galaxy clusters, but also by the Intracluster Medium (ICM) that permeates the cluster. This gas is a hot primordial gas (H and He) that contributes about $\sim 13\%$ for the total binding mass. The particles of this gas emit X-ray radiation, mainly due to Bremmsstrahlung process [Peterson and Fabian, 2006] and they also produce Cosmic Microwave Background (CMB) spectral distortions via SZ (Sunyaev-Zel'dovich) effect [Sunyaev and Zeldovich, 1970]. Because this gas is gravitationally bounded to the halo and not to the individual galaxies of the cluster, the particles of the gas are dispersed in and around the halo. Taking advantage of the X-ray radiation from the emitting gas particles, the observation of X-rays in galaxy clusters can unveil the existence of a dark matter halo (figure 1.1). The X-ray Luminosity (L_X) of the ICM gas that is going generated mainly through Bremmsstrahlung process can be expressed as [Borgani, 2008]:

$$L_x = \int_V \left(\frac{\rho_{gas}}{\mu m_p} \right)^2 \Lambda(T) dV \tag{1.1}$$

where ρ_{gas} is the gas density, μ is the mean molecular weight of the gas, m_p is the proton mass and Λ is the plasma cooling function [Sutherland and Dopita, 1993], which for the Bremmsstrahlung emission, $\Lambda(T) \propto T^{1/2}$. One can also define a X-ray surface brightness along a line of sight for a given energy band ($b_X(E)$) as [Birkinshaw, 1999]:

$$b_X(E) = \frac{1}{4\pi(1+z)^3} \int n_e(\mathbf{r})^2 \Lambda(E, T_e) dl \tag{1.2}$$

where n_e is the number density of the electrons and here, $\Lambda(E, T_e)$ is the spectral emissivity of the gas, observed at energy E.

The SZ, or the Sunyaev-Zel'dovich effect [Sunyaev and Zeldovich, 1970] is the scattering of the CMB photons by electrons from the hot gas. Since galaxy clusters are also composed by the

hot ICM, they are a source of this effect. In fact, the CMB spectrum is distorted by a thermal and a kinetic SZ effect. The former is associated with the thermal motion of electrons in the gas and the latter arises due to the bulk motion of the gas cloud. Such spectral distortions are given by:

$$\Delta I_{th} = I_0 g(x) y \qquad y = \frac{k_B \sigma_T}{m_e c^2} \int T_e n_e dl \tag{1.3}$$

$$\Delta I_k = -I_0 h(x) \frac{v_r}{c} \tau \qquad \tau = \int \sigma_\tau n_e(l) dl \tag{1.4}$$

where I_{th} and I_k are the spectral distortions caused by the thermal and kinetic SZ effect, respectivally. I_0 is a constant and defined as $I_0 = 2k_B^3 T^3/h^2c^2$, $g(x)$ and $h(x)$ are the SZ frequency dependence functions, v_r is the line of sight velocity of the center of mas of the gas cloud, y is the line-of-sight Compton SZ parameter, τ is the optical depth, T_e is the electron density, σ_T is the Thompson cross section and n_e is the number density of the electrons. For typical cluster velocities and optical depth, the SZ thermal effect becomes dominant. The total SZ signal inside the anproximated angular size of a cluster can be approximated by:

$$Y = \int y d\Omega = d_A^{-2} \int y dA = \frac{k_B \sigma_T}{mc^2} d_A^{-2} \int_V T_e n_e dV \tag{1.5}$$

where d_A is the angular diameter distance. Y is the volume integrated SZ Y-flux or also called Y-luminosity. The presence of a galaxy cluster does not only affect the CMB with the SZ effect, but it also induces a polarization signature in the CMB due the scattering of CMB photons by the gas. This polarization includes effects as the CMB quadrupole induced polarization in clusters and carries information about the ICM. Apart from the quadrupole induced polarization, there is also kinetic polarization effects that are generated due to the bulk motion of the gas cloud. Another polarization effect is also the double scattering induced polarization, due to CMB double scattering events by the ICM. If those effects can be observed the polarization from clusters in the CMB can be used as a powerful probe of structure formation and cosmology. [Avelino et al., 2016]

1.2.2 Clusters as Cosmological Probes

Observationaly, in small scales the matter distribution is not homogeneous. This is believed to be the result of perturbations in the inflation field which were amplified by gravity, sculpting galaxies, groups and clusters, and the filamentary structure that permeates the entire Universe [Bahcall, 1988, Springel et al., 2006]. Therefore, the study of these objects and structures and their properties is essential to understand the Universe. In particular, galaxy clusters are recognized as fundamental probes to study dark matter and dark energy/modified gravity [Allen et al., 2011, Roos, 2012]. Optical surveys of galaxy clusters unveiled the Large-Scale Structure. From the spatial distribution of galaxy clusters, it has become clear that the universe is connected by filaments and voids, forming the "cosmic web" [Bahcall, 1988, Springel et al., 2006]. Therefore, galaxy clusters can be used to trace the LSS [Bahcall, 1988, Springel et al., 2006]. This has already been simulated by N-body simulations that show the evolution of the cosmic structure and the distribution of galaxy clusters [Boylan-Kolchin et al., 2009].

The number density of galaxy clusters is very sensitive to the power spectrum and the growth rate of density perturbations and also to the cosmological background parameters, since they

are among the latest bound structures that were formed. Their number density is given by the cluster mass function which is a function of the mass M and the redshift z of their formation. The Press-Schechter (f_{PS}) [Press and Schechter, 1974] mass function can be expressed as:

$$\frac{dn}{dM} = -\sqrt{\frac{2}{\pi}}\frac{\rho_{m0}}{M}\frac{\delta_c(z)}{\sigma(M,z)}\frac{d\ln\sigma(M,z)}{dM}\exp\left(-\frac{\delta_c(z)^2}{2\sigma(M,z)^2}\right) \tag{1.6}$$

where ρ_{m0} is the matter mean density of the universe in the present, δ_c is the threshold of linearly extrapolated density of structure collapse and $\sigma(M,z)$ is the variance of the linear density field [Nunes et al., 2006]. This is expressed as:

$$\sigma^2(M,z) = \frac{D^2(z)}{2\pi^2}\int_0^\infty k^2 P(k) W^2(k,M) dk \tag{1.7}$$

where $D(z)$ is the growth rate of linear perturbations, $P(k)$ is the linear density power spectrum and $W(k,M)$ is the Fourier transform of the smoothing kernel in the real space.

Surveys study the redshift distribution of galaxy cluster abundancies and compare them with the model predictions. Galaxy cluster number counts can be computed as:

$$\frac{d^2N}{dzd\Omega} = \frac{d^2V}{dzd\Omega}\int_0^\infty \frac{dn(M,z)}{dM} f_{survey}(M,z) dM \tag{1.8}$$

Where $d^2N/dzd\Omega$ is the number of galaxy clusters per unit redshift and per unit solid angle, $d^2V/dzd\Omega$ is the volume element of the model and $f_{survey}(M,z)$ is the selection function of the cluster survey. The selection function depends on the noise and the sky coverage of the survey. In most surveys (including the Euclid Survey), this integral cannot be done in M (since it is not a direct observable quantity). It is therefore necessary to make a change in the integration variable to:

$$\int_0^\infty \frac{dn}{dM}\frac{dM}{dS} f[M(S),z]\, dS \tag{1.9}$$

where $S(M)$ is the survey observable (for example, the X-ray luminosity $L_X(M)$ or the cluster richness $\lambda(M)$). These functions are called the cluster scaling relations. Assuming that the dominating force contributing to the formation and evolution of structures is the gravity and that clusters are virialized, cluster thermodynamic properties should be related to their total mass. Assuming that the gravitational collapse is scale-free, the mass of clusters can be scaled. This property is known as the self-similar approximation for galaxy cluster. Similarly, the thermodynamical properties should also be scaled. The scaling relations can be parametrized through the following equation:

$$Y = A \quad E(z)^{\beta_{ss}}(X/X_0)^\alpha \tag{1.10}$$

where Y and X are cluster properties, A a normalization of Y at $X = X_0$ and usually a function of redshift, β_{ss} is the power law index of the redshift scaling, α is the power law index of the independent X property and $E(Z) = H(z)/H_0 = [\Omega_m a^{-3} + (1-\Omega_m-\Omega_\Lambda)a^{-2} + \Omega_\Lambda]^{1/2}$, where Ω_m, Ω_Λ and a are: the mass density parameter, the dark energy (or cosmological constant) density parameter and the expansion factor, respectively. For example, the scaling function between the SZ integrated flux, Y and the mass M ($Y-M$ scaling) is written with $\alpha = 5/3$ and $\beta_{ss} = 2/3$.

As also mentioned, N-body simulations showed that galaxy clusters can be used as a probe of cosmology, in particular the galaxy cluster profiles. A result from simulations is that the dark matter in clusters follows the following expression, known as the Navarro Frenk and White (NFW) density profile:

$$\rho_{NFW}(r) = \frac{4\rho_s}{x(1+x)^2} \tag{1.11}$$

where x defined as $x = r/r_s$, r_s is the scale radius and ρ_s is the density at $r = r_s$. Taking into account the hydrodynamic N-body simulations it was also possible to obtain radial pressure profile, taking into account the cluster gas. From SZ simulations and observations of resolved clusters, the profile takes the following form, known as the generalized NFW profile of SZ signal in clusters:

$$\frac{P(r)}{P_{500}} = \frac{P_0}{x^\gamma(1+x^\alpha)^{(\beta-\gamma)/\alpha}} \tag{1.12}$$

Surveys that observe galaxy clusters, in particular the SZ survey of Planck, compare the distribution of clusters observed with the model prediction by using equation 1.8 in order to constrain cosmological and cluster parameters. Future large surveys such as the LSST (Large Synoptic Survey Telescope), the WFIRST (Wide Field Infrared Survey Telescope) and Euclid are expected to provide better constrains for cosmology, using galaxy clusters as probes. In the next section, I will briefly summarize the main galaxy cluster surveys that have and will be used for cosmology. Some of these will also be used in this dissertation to assess cluster identification by our new algorithm.

1.2.3 Cluster Surveys

ESA/Euclid Space Mission Survey

The Euclid Space mission will investigate the evolution of cosmological structures and the distance-redshift relationship by measuring shapes and redshift of galaxies (and galaxy clusters) [Laureijs et al., 2011]. The mission will also include other working group science projects, such as the cluster of galaxies, CMB Euclid galaxy survey cross-correlations, strong lensing statistics, cool brown dwarfs, large streams and merger history of galaxies, galaxy evolution, stellar populations studies, high-z Lyman break galaxies, supernovae and transients and exo-planets [Laureijs et al., 2011]. The mission is expected to be launched in 2022 and it will have a nominal duration of 6 years. The satellite will orbit the Second Sun-Earth Lagrangian point (L2) and it will complete two surveys: a Euclid Wide Survey and Three Euclid Deep Fields. The Wide Survey will cover 15.000 square degrees of the sky, in an area where the sky is free of contamination from the Galaxy and the Solar System. This survey will measure weak lensing, redshift space distortion and baryonic acoustic oscilations. The three Deep Field surveys will cover about 40 square degrees in total and will be 2 magnitudes deeper than the Wide survey. The Euclid Deep Fields survey will be used primarily for calibrations of the wide survey and also for science of quasars, AGNs and high-z faint galaxies. The satellite will have on board a $1.2\,m$ telescope that has a focal length of $24.5\,m$. The payload consists of a visible imager (VIS) [Cropper et al., 2016] and a Near Infrared Spectrometer and Photometer (NISP) [Maciaszek et al., 2016]. The VIS instrument will image all galaxies of the Euclid survey to measure their shapes and to investigate lensing effects on background galaxies. This instrument has a single broad band filter that covers wavelengths from $500\,nm$ to $900\,mn$. The other on board

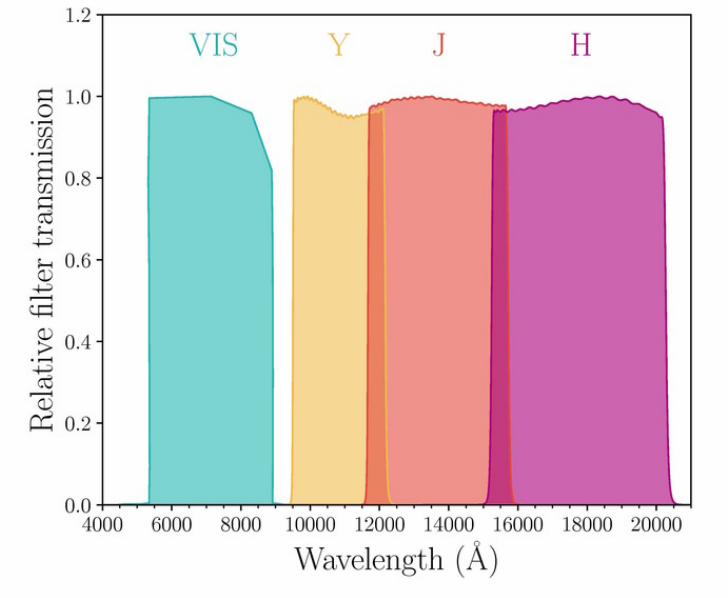


Figure 1.2: Normalized filter transmission of VIS (blue) and NISP filters - Y (yellow), J (red), H (purple). [Inserra, C. et al., 2018]

instrument, NISP, will provide photometry of all the galaxies observed with VIS and will also near infrared low resolution spectra for millions of galaxies. This instrument will be primarily used to estimate the distribution and clustering of galaxies and their evolution out to redshift $z \sim 2$. The photometric channel of the NISP instrument will have three filters: Y ($900\,nm$–$1192\,nm$), J ($1192\,nm$–$1544\,nm$) and H ($1544\,nm$–$2000\,nm$). The coverage of the VIS instrument as well as the filters Y, J, H are represented in figure 1.2. The spectroscopic channel of NISP will be equipped with four low resolution near infrared grisms: three red grisms ($1250\,nm$ – $1850\,nm$) and one blue grism ($920\,nm$ – $1250\,nm$). The VIS and NISP instruments will then be used to measure properties of galaxies and galaxy clusters.

SDSS

The Sloan Digital Sky Survey (SDSS) [York et al., 2000] is a survey that currently covers one quarter of the sky, measuring the positions and brightness of celestial objects. In particular, for galaxies and quasars, it will also measure distances. The telescope of the survey [Gunn et al., 2006] is a $2.5\,m$ telescope that can image a 3° field of view, without distortions. The SDSS has performed several surveys over the year. The SDSS is in its forth project (SDSS-IV) [Blanton et al., 2017] and until now has produced a total of 17 data releases. One of the most recent surveys from SDSS is the eBOSS [Dawson et al., 2016] cosmological survey, that contains quasars and galaxies.

DES

The DES (Dark Energy Survey) [Abbott et al., 2018] is a survey which its goal is to probe the origin of the accelerating universe and to help uncover the nature of dark energy. Since the project launch, in 2013, the DES has imaged about 3000 million galaxies in the southern sky. The telescope of DES is a $4\,m$ telescope, located at the Cerro Tololo Inter-American Observatory. This telescope is also equipped with the DECam (Dark Energy Camera) [Honscheid et al., 2008].

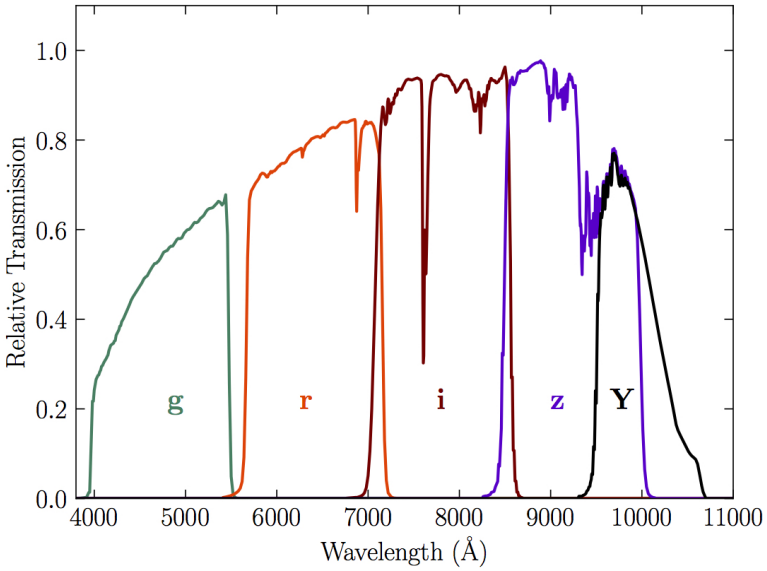


Figure 1.3: Relative Transmission of the DES filters - g (green), r (orange), i (red), z (blue) and y (black) [Abbott et al., 2018]

The system of filters that is implemented on DES is representend in figure 1.3.

Planck Survey

The Planck Mission [Planck Collaboration, 2005] was launched on 14 May 2009 and it was designed to image the temperature and polarization anisotropies of the CMB. It performed five full sky surveys in 30 months. The PLANCKSZ2 (Planck 2nd Sunyaev-Zel'dovich Source) Catalogue [Planck Collaboration et al., 2016], used in this dissertation, contains SZ detections of galaxy clusters, obtained during 29 months of observations. It has a total of 1653 objects, from which 1203 are confirmed clusters.

X-Ray Surveys

- **ROSAT**
  The ROSAT (ROentgen SATellite) is an telescope that performed observations in the X-Rays. With its observations, it was possible to combine the objects in an all-sky survey catalog [Voges et al., 1999]. The telescope imaged a variety of different objects, from clusters of galaxies to comets and neutron stars.

- **XMM-Newton** The XMM-Newton (X-ray Multi-Mirror Mission) [Jansen et al., 2001] is a mission of the European Space Agency, launched in 1999. The scientific purpose of the mission is to observe X-ray objects, to entail their X-ray emission distributuions, spectral and their temporal variability. With the XMM, it was possible to conjugate a catalogue of groups of galaxies and galaxies cluster, the XCS, that was aimed to constrain cosmological parameters and the scaling relations for galaxy clusters, amongst other scientific objectives.

1.3 Cluster Identification in Optical Surveys

Following the scope of the Euclid Mission, I present below a brief overview of methods that were already studied and tested with the Euclid mock galaxy catalogue [Euclid Collaboration et al., 2019], as a preparation for the upcoming mission in the framework of a Cluster Finder challenge. The mock catalogue has essential information about photometric parameters (redshift, magnitudes), their probability distribution functions and the sky coordinates.

1.3.1 AMASCFI

The Adami, MAzure & Sarron Cluster FInder (AMASCFI) [Sarron et al., 2018] method detects a galaxy cluster using the sky coordinates (α, δ) and photometric redshifts. It divides the catalogue in photometric redshift overlapping slices, according to the photometric redshift error. Then, for each slice, galaxy density maps are built based on an adaptative kernel. The density maps will be analysed using a structure detection algorithm (SExtractor). The initial structures will be joined into larger ones, using a FoF (friends of friends) algorithm. The structures will be merged if they are less than 1Mpc apart and with a difference in redshift below 0.05. Then, the candidate cluster will take the sky coordinates and redshift as the mean of all the individual galaxies contained in the merge, weighted by its galaxy number density. Then, AMASCI will count the galaxies with $m_H < m_H^* + 2.5$, where m_H is the H filter magnitude and m_H^* is the knee magnitude of the luminosity function (LF) which was calibrated using the Coma Cluster, inside a cylinder of radius $R_{det} = 1$ Mpc h^{-1}. The method will then rescale the detection radius until it converges. This method uses assumptions about the typical size of a cluster and also the m_H^*.

1.3.2 AMICO

The Adaptative Matched Identifier of Clustered Objects (AMICO) method [Bellagamba et al., 2018] takes the sky coordinates (α, δ), the photometric redshifts and magnitudes. A filter is redshift dependent and defined taking into account a cluster and noise model. This filter will amplify the S/N contrast. The noise is modeled assuming a spatially uniform LF while the cluster model is the combinantion between a galaxy density profile and a cluster galaxy LF. The convolution of the galaxy distribution and this filter generates a 3D amplitude map, where the peaks are assumed to be detections. AMICO also performs a membership probability for each galaxy to belong to a given detection. From the position of the peaks, the method will output the sky coordinates and redshift of the cluster candidate. This method uses assumptions about a density profile for galaxies and also the luminosity function.

1.3.3 HCFA

The Hierarchical Cluster Finder Algorithm (HCFA) method (Díaz-Sánchez, in prep. Please see [Euclid Collaboration et al., 2019]) will work with the sky coordinates (α, δ) and photometric redshifts to search for overdensities of galaxies, using different angular scales, hierachically. Similarly to the AMACSFI, the HCFA method will also slice the catalogue in redshift, with an overlapping redshift bins of size ($\Delta z = 0.05$). A galaxy is labeled with its local density, taking into account the neighbor galaxies and an angular scale (the primary angular scale is set equal to 0.2 Mpc). A critical density (n_{gc}) is then defined as the density above $3\sigma_{n_g}$ to the mean local density $\langle n_g \rangle$: $n_{gc} = 3\sigma_{n_g} + \langle n_g \rangle$, where $3\sigma_{n_g}$ is the standard deviation of the local galaxy density

field. Galaxies with densities below n_{gc} are removed from the sample. The remaining galaxies are then merged using a FoF algorithm, in which the angular scale is the same as the primary scale. Local densities are recomputed again, and the process is repeated in the following steps, while increasing the angular linking scale. The alorithm will then iterate until groups do not merge anymore or the linking scale reaches 0.6Mpc. Centroids are computed taking into account the galaxy distribution of each cluster candidate. The cluster redshift is then determined as the mean redshift of the galaxies inside the centroid. This method makes assumptions about the typical size of a cluster.

1.3.4 PZWav

The PZWav method (Gonzalez, in prep. Please see [Euclid Collaboration et al., 2019]) searches for overdensities on fixed physical scales. This method is a wavelet-type algorithm and requires information of the sky coordinates, photometric redshift and magnitudes of the galaxies of the catalogue. Each galaxy has a probability distribuition that depends on the redshift. From the catalogue, only galaxies that are brighter than $m_H < m_H^* + 2$ are selected. After this step , the remaining galaxies are distributed into redshift bins (that were generated by the algorithm), weighted according to their probability distribution. The density maps are convolved using a difference-of-Gaussians of a fixed physical size - this size should match the physical size of cluster cores. Different density maps are also computed, but using a redshift probability distribuition that was randomly shuffled relative to the positional information, in order to calculate a uniform noise threshold. Galaxy cluster candidates are identified in each redshift slice and merged across bins and their centroids are taken from the peaks of the overdensities in the density maps. The candidate redshift is the mean redshift from all galaxies that lie within 30" of the centroid and are within $\Delta z = 0.12$ of the bin. This method uses assumptions about the typical size and the m_H^* evolution.

1.3.5 sFOF

The sFoF method [Farrens et al., 2011] uses a friend-of-friends algorithm. In order to detect galaxy clusters, it is necessary to specify the sky coordinates and redshifts (either specroscopic or photometric redshifts) of the catalogue. The transverse linking and line-of-sight linking lenght will determine the number of cluster candidates. Each galaxy of a FoF group is classified as a cluster member. The cluster candidate sky coordinates are computed by taking the median of all the coordinates of the galaxy members. This method depends only on the position of each object and as such, it makes no assumptions that are model-dependent.

1.3.6 WaZP

The Wavelet Z-Photometric cluster finder (WaZP)(Benoiust, Dietrich et al., in prep. Please see [Euclid Collaboration et al., 2019] and [Dietrich et al., 2014]) requires the sky coordinates, photometric redshift and magnitudes of the galaxy catalogue. The catalogue is sliced in photometric redshift, with overlapping bins. The overlapping is given by the scatter of $P(z)$ and the galaxies are weighted by the probability distribution function of them lying on a given bin. This galaxy distribution is pixelized into a grid with a step of size $1/16\,\mathrm{Mpc}$ and then filtered using a waveleth. This will select and identify structures . This method uses assumptions about the typical size and the m_H^* evolution.

1.3.7 RedGOLD

The RedGOLD [Licitra et al., 2016a,b] is a modified version of algorithms like RedMaPPer [Rykoff et al., 2014] and takes into account galaxy morphology and color cuts on clusters at high redshifts. The algorithm will select overdensities of galaxies in a color-color plane. Galaxies at $z < 1.5$ show a red sequence, and as such at this redshift, the method will select overdensities of red passive galaxies.

1.3.8 Voronoi

The Voronoi diagram can be a handy tool to identify clusters. The Voronoi diagram is briefly explained in subsection 1.4.6, since this tool is also applied in this dissertation. This mathematical tool was applied to the Euclid Mock Catalogue [Euclid Collaboration et al., 2019] according to the following steps. The catalogue will be divided in overlapping redshift slices. In each slice, a Voronoi tesselation will be applied to the sky coordinates. For each galaxy it will be computed the area covered by the first and second order Voronoi-Delaunay neighbours. The areas are sorted, and a distribution is computed. Galaxies whose areas are below 1.5σ the mean value are kept as "cluster seeds". The first order neighbours will be "attached" to the seeds, and the assemble will grow outwards, adding the first order neighbours of the galaxies that were lastly "attached". More galaxies will be linked to the seeds as long as the second order neighbours area are smaller than a pre-defined cut-off and at least 10 new members are added in each growth step. Thereafter, the results from different redshift slices will be merged together (using the information from the sky coordinates and from the photometric redshift). Each cluster is defined by a center, computed using the member galaxies sky coordinates and photometric redshift values from each merged seed. The algorithm computes cluster areas, by asumming the Voronoy-Delauney areas as well as the observed richness of clusters (after removing the estimated background galaxy density).

The algorithms above, with exception to Voronoi and sFoF, take strong assumptions of what a galaxy cluster is and therefore, these algorithms tend to detect the clusters that match our standard interpretation. However, the most interesting galaxy clusters are the ones that challenge the knowledge of science, with less common or unknown properties. As such, it is important to implement a method that aims to identify cluster candidates based on the least assumptions possible.

1.4 Statistical Tools and Astrophysical Concepts

In this section we briefly introduce the main statistical and computational methods and techniques being used in this dissertation, as well as review a few astrophysical concepts that are necessary to understand the application of these techniques. The UPMASK method (described in the next chapter in section 2.1), which is a method that was created in order to characterize stellar clusters, will be applied, validated and also modified in this dissertation. Principal Component Analysis, K-Means and Kernel Density Estimation are concepts used by UPMASK in order to characterize clusters. In the modification of the method, the concepts of Voronoi Tesselation and Anderson-Darling test were used. In order to perform a simple separation between

stars and galaxies of a catalogue, as explained in section 4.1.1, the concepts of Point Spread Function and Kron Magnitudes were needed to be reviewed.

1.4.1 Supervised Machine Learning

Just like Unsupervised Learning, Supervised Learning is a class of Machine Learning Algorithms. This type of Machine Learning tries to learn a function that can map a labeled data, based on values of expected outputs for our input variables [Kotsiantis, 2007]. To achieve that, supervised learning methods train the machine using a labeled data. This training set consists of a pair, the measurements and the label or known outcome. Then, a new set of data is analysed by the method, which has previously trained the relation between the labels, and uses the calculated relation to produce an outcome. This way, we can predict output variables from new input data.

Classic examples of supervised learning are classification and regression methods. In this dissertation I use regression methods, which are described in section 2.6 and subsection 2.5.

1.4.2 Unsupervised Machine Learning

Just like Supervised Learning, Unsupervised Learning is also a class of Machine Learning Algorithms. The goal of this type of learning is to teach machines how to handle data, with no supervision in form of labelling data [Celebi and Aydin, 2016], meaning the input data will be dealt without any given output variables. Several solutions are based in mathematical approaches in order to interpret patterns or information from the data. The goal of Unsupervised Learning algorithms is to find patterns in the data. The input patterns have an underlying probability distribution which can be estimated. From the estimated density, statistical properties and information will be extracted from the inputs. This way, the algorithm will learn about the inputs in an unsupervised way.

The classic examples of unsupervised learning are clustering and data reduction. In this report I use K-means clustering algorithm and the data reduction algorithm Principal Component Analysis. Both are briefly defined in subsections 1.4.3 and 1.4.4.

Machine learning is a tool that astronomy, astrophysics and cosmology are using more as time passes. In an era where the amount of information is overwhelming to analyse each piece of data, the humanity is resorting on machines to do the "heavy" work. A recent article [Baron, 2019] overviews some of the tools also described in this disseration while also giving examples of works where the statistical tools were applied, in the context of astronomy. For example, in the context of Unsupervised Learning, K-means (subsection 1.4.4) was used to study the spectra of stars and galaxies [Sánchez Almeida et al., 2010, Sánchez Almeida and Allende Prieto, 2013], as well as PCA was used to, for example, estimate physical parameters from spectra [Zhang et al., 2006].

1.4.3 Principal Component Analysis

Principal Component Analysis (PCA) [Hotelling, 1933] is a dimensionality reduction technique and an unsupervised learning algorithm. It applies an orthogonal transformation to the variables of the data. This orthogonal transformation can be thought as an ellipsoid of n dimensions that is fit to the data. The axes of the ellipsoid correspond to the unit eigenvectors of the

covarience matrix. By doing this, an orthogonal linear transformation is being applied to the data, transforming it into a new coordinate system. The first component is such that has the maximum variance and the other components are ordered with decreasing variances. This tool is useful for multi-dimensional features of our data, since the idea behind using PCA in data analysis is that although many features of the data live in multi-dimensional space, many of the features might not be equally interesting. As such, the PCs tranformation will re-organize our data in the way that has the most variance as possible.

1.4.4 K-Means

K-means [Forgy, 1965] is an unsupervised learning and clustering algorithm. It divides the data in clusters of objects with similar characteristics and chooses k random points in the data space. These k points will be the cluster centers. Then, it will assign data points to the closest center, thus forming k groups. It will move the centers to the means positions from each respective group. This process will iterate until the mean positions from each group converges.

1.4.5 Kernel Density Estimation

Kernel density estimation (KDE) [Parzen, 1962, Rosenblatt, 1956] is an approach to estimate the probability density function of a variable. It is also known as the Parzen-Rosenblatt window method. The shape of the density of the data can be discovered through KDE, offering a solid and strong alternative to density estimations through histograms. The latter are not smooth and depend on the width and end points of the bins, in contrast with the KDE that is smooth and has no fixed structure.

The Kernel Estimator of a function of unknown density f from which it was drawn an independent sample $(x_1, x_2, ..., x_n)$ is

$$\widehat{f_h}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1.13}$$

where K is the kernel, n the sample size and h is the bandwidth. The kernel functions can have various forms, such as uniform, tophat, exponential, linear, Epanechnikov and Gaussian functions, and others. The bandwidth is a free parameter however if the underlying density of the data is a Gaussian, the optimal choice for h is given by the so called Silverman rule of thumb [Silverman, 1986]:

$$h = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}} \tag{1.14}$$

where h is the bandwidth, σ is the standard deviation of the sample and n the size of the sample.

1.4.6 Voronoi Tesselation

For a better understanding of this concept, I will only speak of the $\mathbb{R}^2$ case. Let us take a distribution of k points in a plane. The Voronoi diagram is a partition of this plane into k convex polygons such that each polygon contains only one point. Each polygon region encloses all locations that are closer to the point that originated the poligon than to the other points. The Voronoi Tesselation [Voronoi, 1908] can be defined to $\mathbb{R}^N$. Given a set of points $x_{i\,i=1}^{k} \subset \Omega$,

in which Ω is an open bounded domain $\Omega \in \mathbb{R}^2$, the Voronoi region V_i corresponding to the point x_i is defined by

$$V_i = \{x \in \Omega | \|x - x_i\| < \|x - x_j\| \text{ for } j = 1, ..., k, j \neq i\} \tag{1.15}$$

In this way, we will have a tesselation of space. The Voronoi diagrams are used in a large range of fields, such as medicine, biology, informatics and many others.

1.4.7 Anderson-Darling Test

The Anderson-Darling test [Anderson and Darling, 1952] is a goodness-of-fit statistic that tests the hypothesis that a sample $X_1, ..., X_m$, with an empirical distribution function (EDF) $F_m(x)$ comes from a continuous population with a completely specified distribution function $F_0(x)$. The Anderson-Darling test can be expressed as

$$A_m^2 = m \int_{-\infty}^{\infty} \frac{(F_m(x) - F_0(x))^2}{F_0(x)(1 - F_0(x)} dF_0(x) \tag{1.16}$$

In this project a two-sample Anderson-Darling test is used to test the hypothesis that two samples have the same distribution function. The two sample test can be written as

$$A_{mn}^2 = \frac{mn}{N} \int_{-\infty}^{\infty} \frac{(F_m(x) - G_n(x))^2}{H_N(x)(1 - H_N(x))} dH_N(x) \tag{1.17}$$

where $G_n(x)$ is the empirical distribution of the second independent sample $Y_1, ..., Y_n$, that comes from a continuous population with distribution function $G(x)$. $H_N(x) = (mF_m(x) + nG_n(x))/N$, with $N = m + n$, is the empirical distribution function of the combined sample. A k-sample Anderson-Darling test is also described in Scholz and Stephens [1987].

Lastly, to understand underlying concepts of magnitudes and their measurement, it is important to review the indispensable concepts of the Point Spread Function and Kron Magnitude.

1.4.8 Kron Magnitude

Kron expressed a luminosity-weighted radius R_1, that defines the first moment of the surface brightness light profile [Kron, 1980]. This radius is defined as:

$$R_1(R) = \frac{2\pi \int_0^R I(x) x^2 dx}{2\pi \int_0^R I(x) x dx} \tag{1.18}$$

where x is the radius and $I(x)$ the intensity profile. Kron claimed that using an aperture radius twice the size of R_1, one can obtain $> 90\%$ of the galaxy flux. The author shows that this approach can be used to galaxies of different morphology, such as elliptical and spiral galaxies. Using Sérsic intensity profiles [Sérsic, 1963], it is possible to obtain an aperture radius that collects almost all the incoming light from a galaxy, optimized to their intensity profile. The flux measured through an aperture defined by the Kron radius is then a Kron Magnitude.

1.4.9 Point Spread Function

When we take a picture of a point object, the image will not appear as a mathematical point, the light of this point will be spread out in order to form a finite area in the image. This is the Point Spread Function (PSF). Images of bigger objects are a convolution of the light source and the Point Spread Function. The resulting image depends on the resolution of the optical system capturing the image as well as on atmospherical or light scattering effects that varie with the wavelenght of the incoming light. The PSF for a perfect optical system can be described as an "Airy Pattern" [Airy, 1835] which has the following expression:

$$I(R) = \frac{1}{(1-\epsilon^2)^2}\left[\frac{2J_i(x)}{x} - \epsilon^2\frac{2J_1(\epsilon x)}{\epsilon x}\right]^2 \tag{1.19}$$

where $I(R)$ is the surface brightness in the focal plane, R is the radial distance in the focal plane from the optical axis and x is defined by R as $x = ka\sin\theta$, with k the wavenumber, a the aperture radius and θ the observation angle, ϵ is the fractional radius of the central obscuration of the primary aperture and the aperture diameter (also known as annular aperture obscuration ratio), and J_1 is the Bessel function of the first kind of order 1. The PSF and its determination are particularly important in astronomy and astrophysics since stars and quasars can be approximated to a point object, due to its size distance and light intensity. As such, a PSF magnitude is the light flux measured, assuming that the intensity profile results from a PSF.

Chapter 2

Modifying UPMASK

To detect and study galaxy clusters using individual galaxies, one needs to adopt some membership assignment method. Thus, in this chapter, I start by studying the viability of applying UPMASK, which is a method created for stars, to identify galaxy groups and clusters. This is the first time this method is being used with galaxies and therefore its necessary to validate its original version. I then explore modifications of the original method, looking to optimize it. I explore a Voronoi Tesselation scheme, then a grid scheme and a fitting function procedure to speed up UPMASK. I start by describing the method and follow by presenting my modifications.

2.1 UPMASK

UPMASK, or Unsupervised Photometric Membership Assignment in Stellar Clusters [Krone-Martins and Moitinho, 2014, 2015] is a method designed to characterize stellar clusters, considering minimal photometric and astrometric data, but it can use other data features. The definition of the authors of a stellar cluster is that 'a stellar cluster is a spatial over-density of stars with a common origin' and that 'cluster members will be clustered in most spaces, including positional space' - this accounts for the fact that due to the resolution of the mapping and/or image that is taken of a cluster, the cluster members might not appear clustered in the positional space however, the members will have similarities in their properties since they have a common origin. In this dissertation, the same assumptions are made for galaxy clusters. The standard UPMASK method uses the following statistical methods: Principal Component Analysis (PCA), clustering algorithms, Kernel Density Estimation (KDE), data re-sampling and iterative processes. A diagram of the UPMASK Kernel is found in figure 2.1.

UPMASK first samples a simulated data-set from the original data and its error distributions. Then it applies PCA [Hotelling, 1933] to the photometric data, and then groups the data in the space defined by the Principal Components with K-means [Forgy, 1965], in order to group objects with similar photometric properties. Afterwards, each such mathematical cluster is analyzed in the astrometric space with Kernel Density Estimators and then compared with the KDEs [Parzen, 1962, Rosenblatt, 1956] of a uniform random distribution; the method eliminates those that are compatible with the KDE of a uniform distribution, and iterates. Having computed

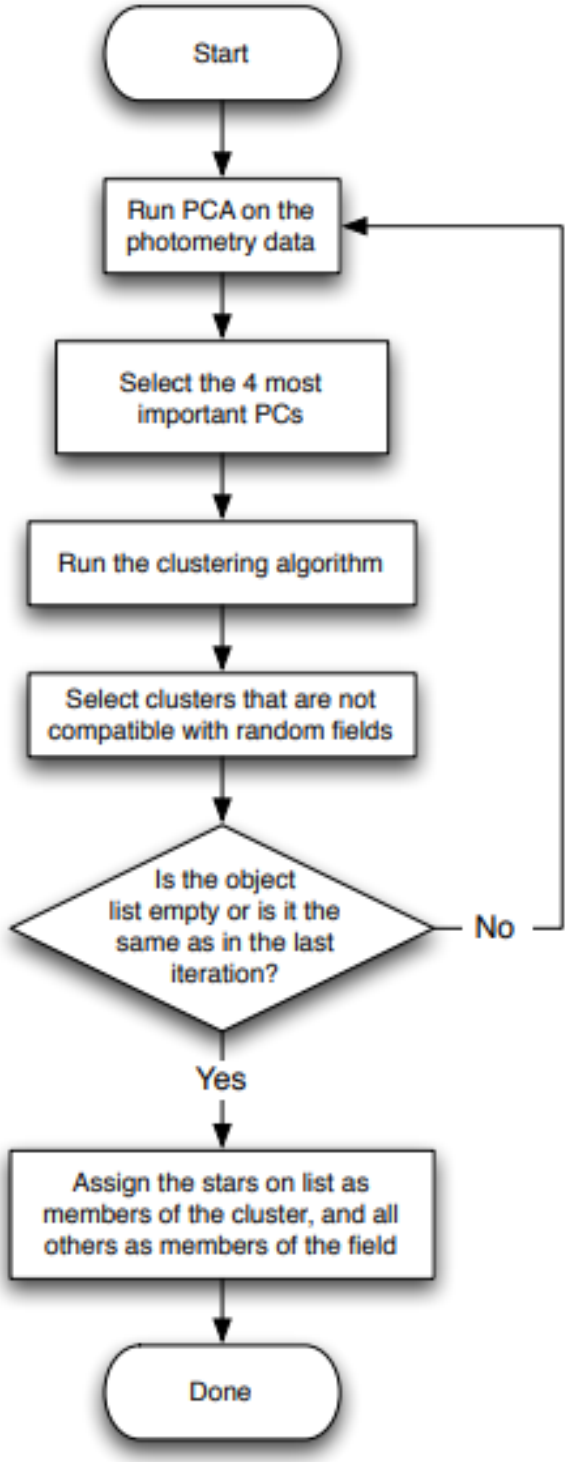


Figure 2.1: Flowchart of UPMASK Kernel. [Krone-Martins and Moitinho, 2014]

the KDEs for a K-means cluster, we calculate

$$D(\phi) = \frac{\max(\phi) - \langle\phi\rangle}{\sigma_\phi} \tag{2.1}$$

where ϕ are the densities estimated in the KDE step and $D(\phi)$ measures the difference between the maximum value and the mean of ϕ, scaled by the standard deviation of ϕ. This quantity is also computed for the sets (Φ) resulting from estimating KDE densities for a number of random uniform realizations, all with the same area and number of points. This results in a vector of Φ_i

$$D(\Phi_i) = \frac{\max(\Phi_i) - \langle\Phi_i\rangle}{\sigma_{\Phi_i}} \tag{2.2}$$

The set that resulted in Φ will be classified as a group if

$$D(\phi) \geq \langle D_\Phi\rangle + T \times \sigma_{D_\Phi} \tag{2.3}$$

where D_Φ is a set of $D(\Phi_i)$, $\langle D_\Phi\rangle$ its mean and σ_{D_Φ} its standard deviation. T is a level parameter that by default, is set to 1. The right member of the equation defines the threshold value.

Applying this routine in an iterative way, UPMASK discerns clustered from field objects (objects that are in the field of view, not bound to the cluster) by purifying the cluster distribution at each iteration. At the end of each iteration, each point will be classified in a binary way: 0 as not bound and 1 as a bound object. By running the entire routine an ammout of times (this is called the *nruns* UPMASK parameter) that can be set by the user, UPMASK naturally takes into account all observational errors and provides a score that is a frequentist membership probability for each object in the analysed data-set.

2.2 Modification I: Voronoi selection procedure with Anderson-Darling Test

The kernel density estimation step is computationally time consuming, and thus the application of UPMASK in large regions of the sky requires significant computational resources. Therefore, I studied an alternative approach to replace the KDE - on the flowchart in figure 2.1, this corresponds to the fourth and fifth and sixth boxes. Here I apply a Voronoi Tesselation [Voronoi, 1908] procedure to the K-means mathematical clusters in the positional space and then compare the polygon areas distributions of the K-means clusters with those obtained from a random uniform point distribution with the same number of points and sky area. In figure 2.2 are represented the Voronoi Tesselations and the distributions of sky areas of the polygons of a random uniform point distribution and a random normal point distribution, for illustration of the method. Note that the normal random realization is naturally clustered at the center of the upper left panel, whereas the uniform random realization, naturally shows a typical uniform distribution of points. To perform the comparison between the area distributions, I adopt an Anderson-Darling Test [Anderson and Darling, 1952]. This test is applied twice, and will give us p-values, which will be used to discern between two tests: the first, compares the distribution of Voronoi areas of our data against the distribution of Voronoi areas of a 2D random uniform realization, within the same area and number of points of our data; and the second, compares two distributions of Voronoi areas from different 2D random uniform realizations - the later test

will be performed 100 times maximum, to the same kind of generated Voronoi areas, so that in the end it is possible to take the mean and the standard deviation of the p-value. Therefore, a K-means group is classified as a clustered group if

$$p_{kmeans-uniform} \leq \langle p_{uniform-uniform} \rangle - T \times \sigma_{uniform-uniform} \tag{2.4}$$

where $p_{kmeans-uniform}$ is the p-value of the Anderson-Darling test of the distribution of Voronoi areas of our data against the distribution of Voronoi areas of a 2D random uniform realization, $p_{uniform-uniform}$ is the mean p-value of the Anderson-Darling test of two distributions of Voronoi areas from different 2D random uniform realizations, $\sigma_{uniform-uniform}$ is the standard deviation of p-values from the Anderson-Darling test of two distributions of Voronoi areas from different 2D random uniform realization and T is a threshold level.

To increase the speed of the computational process I created a lookup table that stores the number of objects of the K-means classes and the mean Voronoi areas of a uniform distribution with the same number of objects. If the method already passed through a K-means class with a similar number of objects, it will take the Voronoi areas of the uniform distribution from the Lookup Table instead of regenerating one.

2.3 Modification II: Voronoi selection procedure with simple statistic comparison

As we will see in chapter 3, the Anderson-Darling Test can be strict. This statistical test compares a whole distribution of areas and therefore the areas distribution of a cluster that do not have a clear difference will be eliminated. Therefore I have implemented another alternative that comes from a comparison of a parameter calculated with the mean of the distributions of the Voronoi Polygons areas. For a K-means group, the Voronoi Tesselation is performed and the mean of the density of the polygons (inverse of the area) is taken. A random uniform point distribution is drawn 100 times maximum, with the same number of points and area of the K-means group. Then, I apply the Voronoi Tesselation to the realizations and take the mean of the density of the polygon areas. From this set, I am able to calculate a final mean and its standard deviation. The group is assigned as an identified cluster if:

$$\langle N_{kmeans} \rangle > \langle N_{uniform} \rangle + T \times \sigma_{uniform} \tag{2.5}$$

in which $\langle N_{kmeans} \rangle$ is the mean of the density of the polygons from the Voronoi Tesselation applied to the K-means group, $\langle N_{uniform} \rangle$ is the mean of all the uniform realizations mean density of the polygons from the Voronoi Tesselation, $\sigma_{uniform}$ is the standard deviation of all the uniform realizations mean density of the polygons from the Voronoi Tesselation and T is a threshold level that can be adjusted by the user. This T is set to be equal to 3 by default.

2.4 Modification III: Grid selection procedure

I also studied another approach to replace the kernel density estimation procedure in UPMASK. Instead of using a method that estimates the density function with time consuming statistical computations, I used a "contingency table" approach to select clustered groups in the

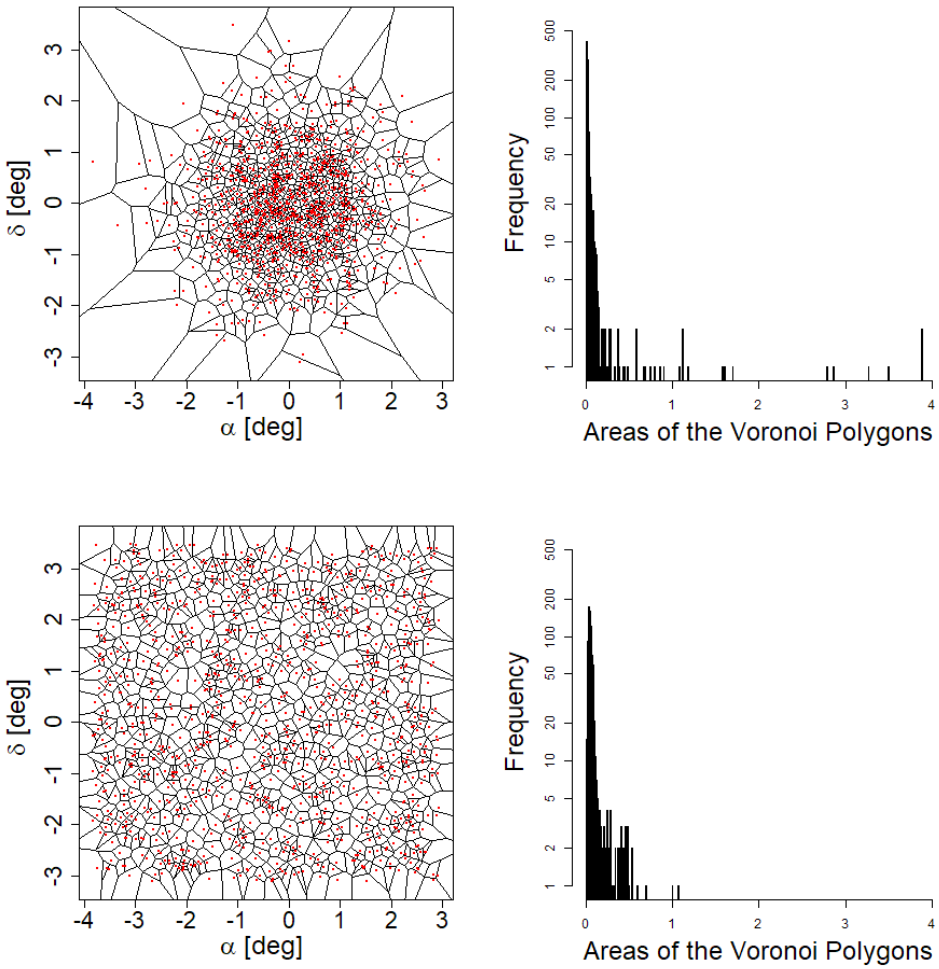


Figure 2.2: Top: Left - Voronoi Tesselation (black lines) of a random normal distribution (red points). Right - Histogram of the distribution of polygon areas that resulted from the Voronoi Tesselation on the left. Bottom: Left - Voronoi Tesselation (black lines) of a random uniform distribution (red points). Right - Histogram of the distribution of polygon areas that resulted from the Voronoi Tesselation on the left.

positional space. The method works as follows. Our variables will be the two coordinates of the positional space. The challenge will then be to choose the number and size of the intervals to grid the data. For that, I used the Silverman's rule of thumb (equation 1.14). This rule gives the optimal choice for the bandwidth of the kernel if the underlying density is Gaussian. I choose this bandwidth since the goal is to replace the kernel density estimation part of UPMASK - where this rule is also being used.

We apply the Silverman's rule to each coordinate of the 2 dimensional dataset. We then take the grid bandwidth to be equal to the smaller of the 2 bandwidths as it will lead to more divisions on the 2D sample. This bandwidth is applied in both directions to build the contingency table. After the bandwidth is calculated, the 2d space will be divided in intervals whose size are given by the bandwidth. The points are distributed according to the intervals in which they fall in. In the end, we will have a 2D matrix that, in each entry, it will contain the total number of the points falling in each 2D bin. We will then have a frequency of the points placed in this 2D matrix. With this information, the maximum of the frequency is taken, the mean is subtracted and the result will be divided by the standard deviation of the frequencies, similarly to the computation made in equation 2.1 for the KDE described in section 2.1. The calculated parameter will be compared with a mean of the same parameter calculated from multiple random uniform realizations of the same number of points, generated within the same area and with the grid performed with the same number of intervals. This comparison is performed similarly to the one for the KDE in section 2.1, using also a threshold level parameter.

2.5 Modification IV: Grid selection procedure with fast thresholding

To reduce iteration times, I have opted to replace the part where the method compares original data points with a random uniform realizations (with the same area and number of points) with a regression. For this, I have performed several tests using random uniform realizations with different number of points and areas and for each of these realizations, obtained the D parameter the method uses to compare distributions (1000 realizations for a certain area and number points), obtaining as such a 3D matrix with the following dimensions: number of points, area and D parameter (see equation 2.2). I have performed this test firstly for the number of points ranging from 1 to 100 in steps of 1. Then from 100 to 1000 in steps of 50. This choice is simply due to the fact that we expect a galaxy cluster to have more than 50 members, but since the clustering algorithm will divide the data in subsamples of galaxies, the number of objects per group is expected to be around 1-100. The other sequence was created to account for possible groups samples that fall outside that boundary. I have noticed that for the same number of points, the D parameter is independent of the area. Thus, we average the computed parameter over the area for each number of points (since we performed 1000 realizations) and obtain a threshold (see the right handed side of equation 2.3). This is the threshold this version of method uses to compare the input data.

The values of threshold generated are represented in figure 2.3, that depend on the number of objects per cluster of K-means. With the number of points generated, we adjusted a curve in an interval that covers all the cases of interest of the threshold value, in terms of number of objects per cluster of k-means. As such, UPMASK will compare the parameter for the comparison with

		a_0	1.149227	a_1	0.200402	a_2	-0.001189		
b_0	1.351×10	b_1	-7.293×10^{-2}	b_2	4.462×10^{-4}	b_3	-1.443×10^{-6}
b_4	2.529×10^{-9}	b_5	-2.422×10^{-12}	b_6	1.187×10^{-15}	b_7	-2.315×10^{-19}

Table 2.1: Fitting function coefficients of equation 2.6 giving the thresholds of the Grid method

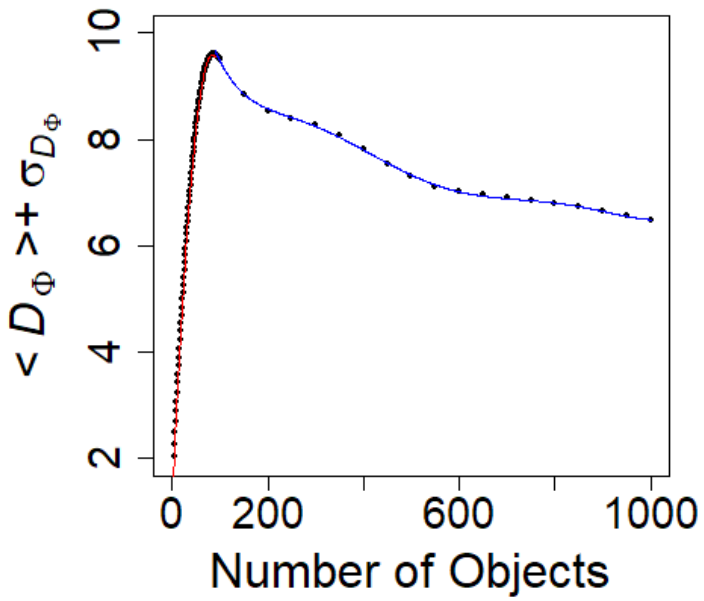


Figure 2.3: Fit Function for the Grid method. The red line corresponds to the first branch of the function and the blue line corresponds to the second branch

a random uniform realization by using this function instead of performing several test for that distribution.

The regression function has two branches of the following expression:

$$f(x) = \begin{cases} a_0 + a_1 x + a_2 x^2 & n < 90 \\ b_0 + b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4 + b_5 x^5 + b_6 x^6 + b_7 x^7 & n \geq 90 \end{cases} \tag{2.6}$$

in which the coefficient values are presented in table 2.1. It is woth to add that the creation of this regression function, as well as the points that it generated, was performed outside the UPMASK method. The method simply uses the function that was found.

2.6 Modification V: KDE selection procedure with a fitting function

Following the same reasoning as in 2.5, I have replaced the part where the method compares our data points with a random uniform realization with a regression. The regression function

c_0	3.0532634	c_1	-0.0381186	c_2	0.0011377	c_3	-0.0000115
d_0	2.648	d_1	-7.205×10^{-4}	d_2	3.781×10^{-7}	d_3	-7.897×10^{-11}

Table 2.2: Fitting function coefficients of equation 2.7 giving the thresholds for the thresholds of the KDE method

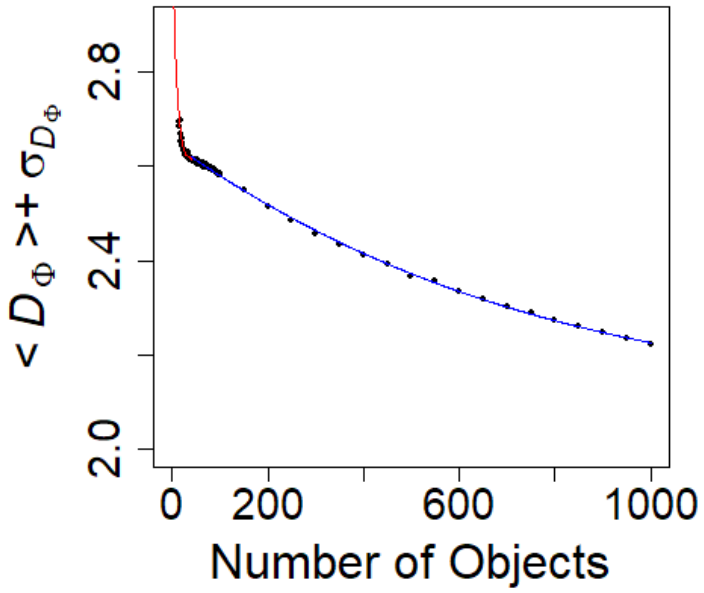


Figure 2.4: Fit Function for the KDE method. The red line corresponds to the first branch of the function and the blue line corresponds to the second branch

has also two branches, with the following expression:

$$f(x) = \begin{cases} c_0 + c_1 x + c_2 x^2 + c_3 x^3 & n < 40 \\ d_0 + d_1 x + d_2 x^2 + d_3 x^3 & n \geq 40 \end{cases} \tag{2.7}$$

in which the coefficient values are presented in table 2.2. The final function is represented in figure 2.4

To validade the UPMASK and all proposed modifications of the method, I will use state-of-the-art galaxy catalogues from N-body simulations that contain photometric information. This study is presented in 3.

Chapter 3

Validation with Simulated data

This chapter is dedicated to the study of the viability of the UPMASK method for the detection of galaxy groups and clusters in photometric galaxy surveys. The validation process was carried out using state-of-the-art N-body simulations containing photometric and galaxy-dark-matter halo membership information that allows to determine which galaxies are part of true cluster halos or are field galaxies. These simulations are described in section 3.1. Section 3.2 presents a detailed description of the tests and method yields for all version of UPMASK under investigation, followed by a discussion of how the different modifications work and compare when varying method parameters.

3.1 Simulated data description

To test the application of UPMASK in galaxy clusters I used large-scale structure simulations that contain photometric galaxy luminosities for the DES [Honscheid et al., 2008] and Euclid survey [Laureijs et al., 2011] bands. The data I used is in fact a small subset of the second version of the MICE galaxy mock catalogue [Crocce et al., 2015, Fosalba et al., 2015a,b] and it was obtained from the CosmoHub portal[1] [Carretero et al., 2017]. These data has been used in several studies by the Euclid Consortium[2] to simulate (and assess the performance) of Euclid observations.

The galaxy mock catalogue was constructed from the N–body MICE-GC simulation, using a hybrid Halo Occupation Distribution (HOD) and Halo Abundance Matching(HAM) prescription to populate Friends-of-Friends (FoF) dark matter halo from the MICE-GC simulation. The cosmological parameters used to generate the simulations are $\Omega_m = 0.25$, $\sigma_8 = 0.8$, $n_s = 0.95$, $\Omega_b = 0.044$, $\Omega_\Lambda = 0.75$ and $h = 0.7$.

The full MICE catalogue contains 499,609,997 mock galaxies. It is complete for the DES survey down to a i band magnitude of 24, up to $z \sim 1.4$, in a sky area with $\delta > 30^\circ$ and $\delta < 30^\circ \wedge 30^\circ < \alpha < 60^\circ$. For the Euclid survey the data is complete in the H band down to H~ 24 up to $z \sim 0.45$, to H~ 23.5 up to $z \sim 0.9$, and to H~ 23.0 up to $z \sim 1.4$.

From the full MICE catalogue I took all objects inside an area of one square degree located at $30.5^\circ \leq \delta \leq 31.5^\circ$, $0.5^\circ \leq \alpha \leq 1.5^\circ$. This field was chosen to guarantee the maximum possible completeness for the DES (and Euclid) bands. This area contains 82533 galaxies, for which I kept the following galaxy properties: DES [Honscheid et al., 2008] filters magnitudes g, r, i, z,

[1] https://cosmohub.pic.es
[2] http://www.euclid-ec.org

y (represented in figure 1.3), Euclid [Laureijs et al., 2011] filter magnitudes RIZ, Y, J and H (represented in figure 1.2), astrometry information (α, δ), redshift, halo mass, halo identifier and galaxy identifier. From this subset, I selected a smaller subset around the most massive cluster that I will use throughout this chapter to test UPMASK and its modified versions. This resulted in a galaxy catalogue with a total of 8774 galaxies, 171 of them belonging to the most massive cluster in the field, which lies at $z = 1.09$.

All runs performed for this dissertation were carried out on a regular desktop computer with the following hardware and software specifications: Intel Core i5-4460 CPU at 3.20GHz; 16Gb of non-ecc memory at 2400MHz; Intel/ASUSTeK integrated graphics controller; Ubuntu 18.04 LTS (bionic) operating system and CRAN R version 3.4.4.

3.2 UPMASK detection results: Purity and Completeness

In this section we present all the results and tests performed with the MICE simulated data. The discussions of the results are presented in subsection 3.3. First we applied the original UPMASK method to the selected galaxy mock catalogue. We start by using the DES filters g, r, i, z and y and the respective color combinations. Then, we adopted Euclid filters RIZ, Y, J and H and also their respective colors. To compare performances between different method versions, I choose to set *nruns*= 10 in all versions. This choice is a good compromise between having relatively short CPU running times and a fair membership frequentist probability, with 10% resolution for all method versions.

The first plot of Figure 3.1 represents the original data, while the middle left and right plots of Figure 3.1 represent the results of the unmodified method applied to DES and Euclid simulated observations. The white circle Figure 3.1 indicates the size and position of the cluster we are searching for. We have estimated a center taking into account all the galaxies in the field and the radius by taking the distance of the most further galaxy of the cluster.

To perform a more quantitative comparison, we compute the completeness and purity metrics using the points contained in the estimated halo size using the following expressions:

$$\text{Purity} = \frac{N_{UPMASK,cluster}}{N_{UPMASK,detec}}, \tag{3.1}$$

$$\text{Completeness} = \frac{N_{UPMASK,cluster}}{N_{cluster}}, \tag{3.2}$$

where $N_{UPMASK,cluster}$ is the number of known objects belonging to the simulated cluster above a certain UPMASK probability threshold; $N_{cluster}$ is the number of known objects belonging to the simulated cluster and $N_{UPMASK,detec}$ is the number of objects above a certain UPMASK probability threshold. We present the results from DES and Euclid simulated observations in the bottom left and right plots of Figure 3.1, respectively.

Afterwards, we tested the five modified versions of UPMASK. We run all versions with the same internal parameters - 4 PCA principle components and a mean of 50 objects per K-mean cluster. These versions are: the original UPMASK (see section 2.1), KDE with a fitting function (modification V, see section 2.6), Voronoi tesselation selection with the Anderson-Darling test (modification I, see section 2.2), Voronoi tesselation selection with simple mean comparison (modification II, see subsection 2.3), Grid selection procedure (modification III, see section 2.4) and finally Grid selection procedure with a fitting function fast thresholding (modification IV

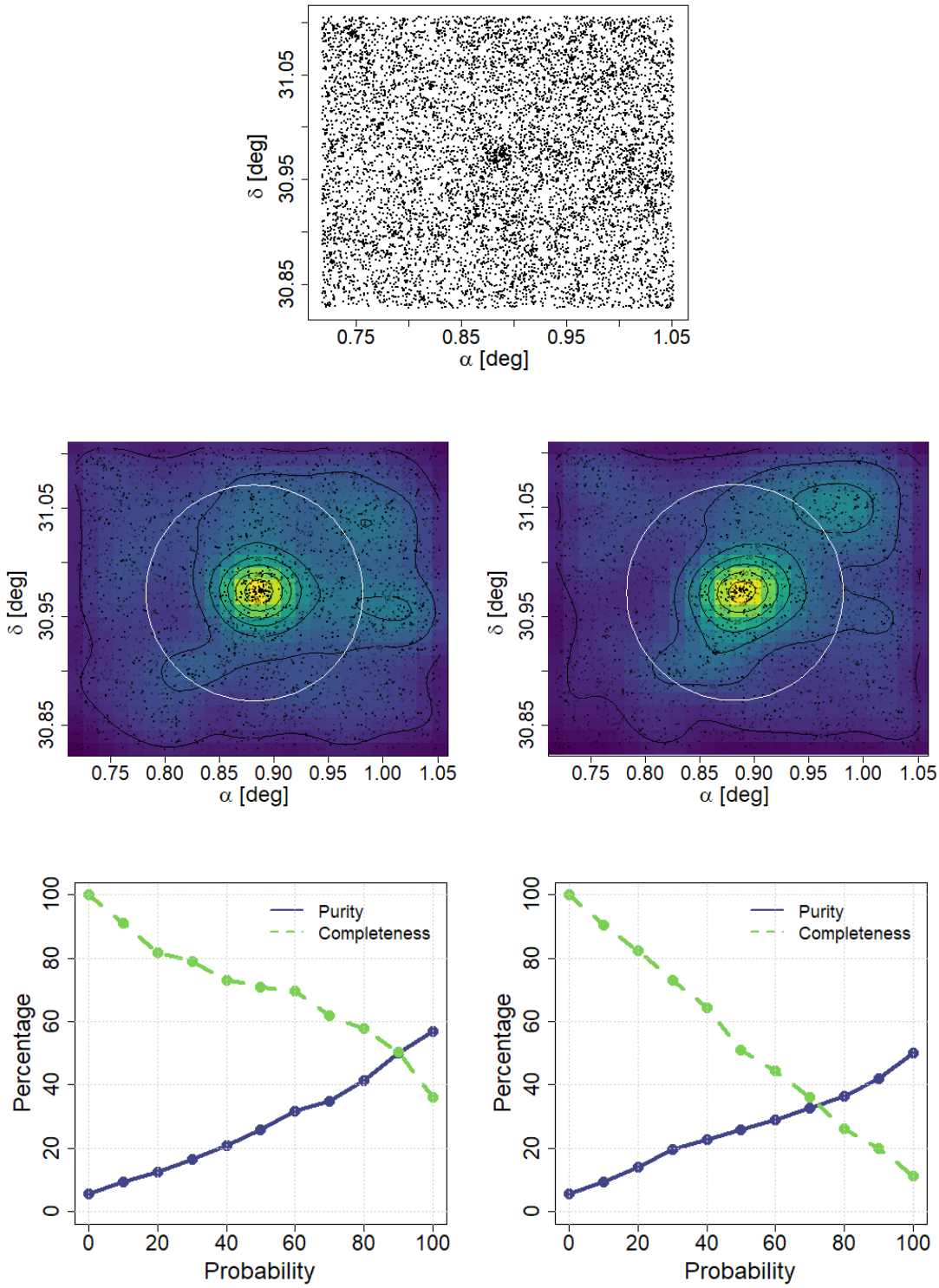


Figure 3.1: This figure includes images and detection yields for the original version of UPMASK. The top panel shows all galaxies in the selected field, as extracted from the original MICE catalogue. The middle left and right panels panels show KDE density maps of UPMASK detected galaxies using DES (g, r, i, z, y) and Euclid (RIZ, Y, J and H) filters, respectively. The color scale was set the same in both panels, with a color scheme where lower (higher) densities are in blue (yellow). The white circles indicate the cluster dark-matter halo radius. The bottom panels show the Completeness (green dashed) and Purity (blue solid) functions obtained from the unsupervised UPMASK classification: on left - results from DES; on the right - results from Euclid filters.

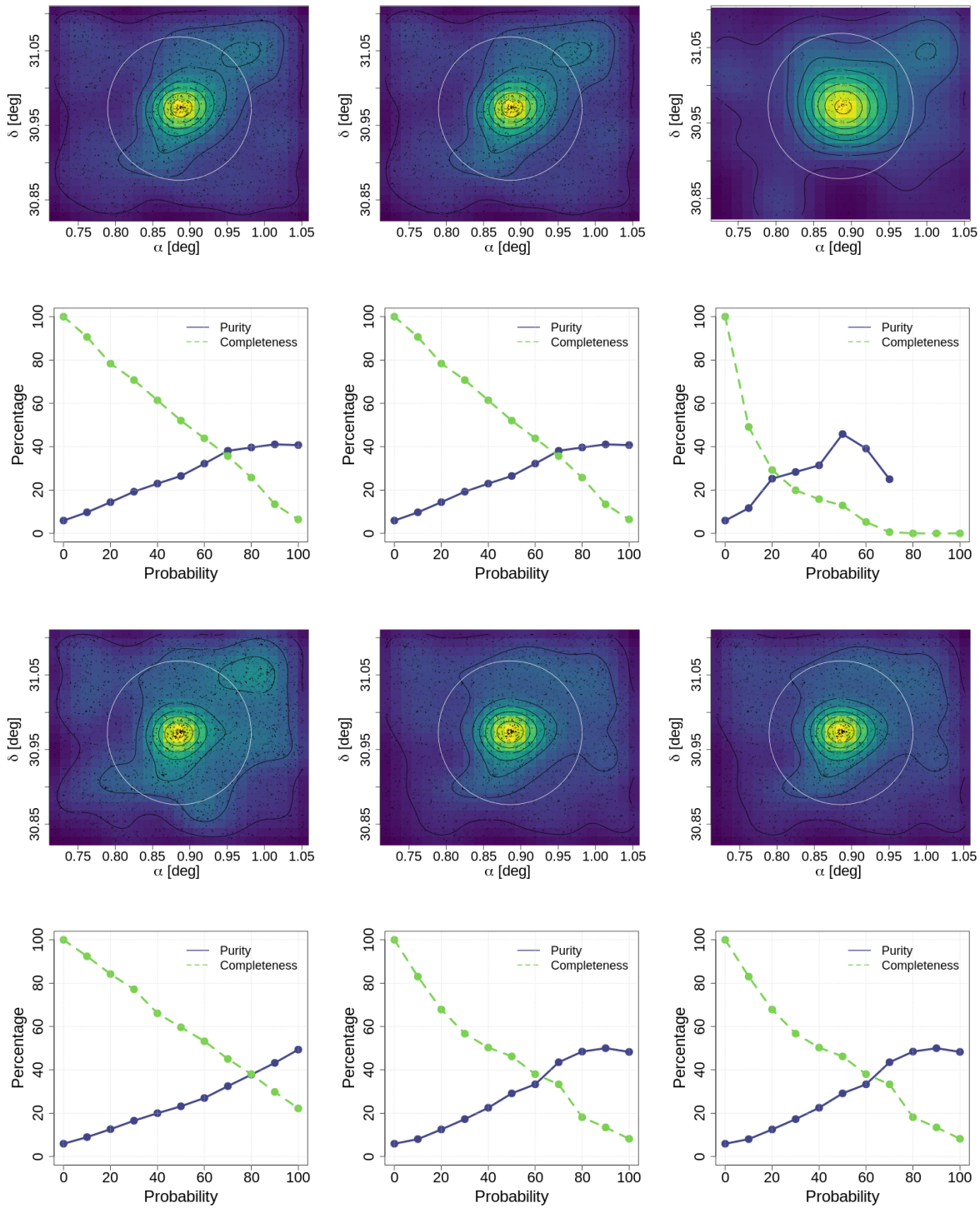


Figure 3.2: The detection resuts for the original and all modified UPMASK version. The results are displayed in groups of two rows, with each column corresponding to a version. The first and third row show the density maps of UPMASK detected galaxies for the following versions: KDE, KDE + Fitting Function, Voronoi + Anderson-Darling Test, Voronoi + Mean Comparison, Grid and Grid + Fitting Function. The color scale was set the same for all the field panels, with a color scheme where lower (higher) densities are in blue (yellow). The white circles indicate the cluster dark-matter halo radius. The second and fourth rows show the Completeness (green dashed) and Purity (blue solid) functions obtained from the unsupervised UPMASK classification, corresponding to the field above each case.

UPMASK Version	CPU running time (s)	%
KDE	274	100
KDE + Fitting Function	262	96
Voronoi + Anderson-Darling Test	190	69
Voronoi + Mean Comparison	148	54
Grid	303	110
Grid + Fitting Function	289	105

Table 3.1: CPU running times for the different UPMASK versions. The last column shows the relative CPU running time percentage of each version with respect to the original UPMASK method.

see section 2.5).

All runs performed here were created with the same random seed (a numerical value used to generate a well defined sequence of random numbers) for all versions of the method. Since we are comparing different method versions, and some of the versions require different amounts of random numbers we decided to perform only one run for each method version. We note however that the K-Means function used in UPMASK is unable to preserve random number sequences (it uses system randomized seeds!) and therefore small "fluctuations" may appear while making different runs realizations with the same seed and parameters. Note that this introduces an intrinsic "sample variance" that can be stabilized by performing several runs for the same method version. I checked that these fluctuations have a small impact on the purity and completeness functions for a wide range of membership probabilities, especially those derived from a larger number of objects, i.e. at lower membership probabilities. This means that only the highest membership probabilities are more sensitive to run-to-run fluctuations. The results obtained from running the five UPMASK modifications plus the original version (see section 2.1)) are presented in Figure 3.2 for comparison. The completeness and purity curves were computed according to equations 3.1 and 3.2) for all method versions. The CPU running times taken by each version of UPMASK are also shown in table 3.1, as well as the CPU running time percentage of each version with respect to the original UPMASK method.

I end this section with a study of the impact of the UPMASK parameters on our results. I used the same field data and run the different code versions for the Euclid filters. For each version, it was adopted $nruns = 10$ and $nruns = 100$ (with the exception of study for the change of the number of principle components, where it was adopted $nruns = 100$), resulting in a resolution of 10% and 1% in the frequentist cluster membership probability, respectively. In the following subsections I present our findings for how the number of principal components and the number objects per cluster affects the purity and completeness of the sample and also the computational time.

3.2.1 Dependence on Principal Components

Principal Components are ordered according to the variance of the original data projected in the PCA basis (see subsection1.4.3). Accordingly, by selecting only the first few n principal components, one can capture the most important part of the original data. UPMASK calculates each PC from the input photometry. For this cluster field, we have used 11 quantities for the PC determination, namely: the Euclid filters magnitudes (RIZ, Y, J and H), all the possible color combinations for these magnitudes (RIZ-Y, RIZ-J, RIZ-H, Y-J, Y-H and J-H), and the sum of of the 4 magnitudes (RIZ+Y+J+H). The standard deviation of each principal component is

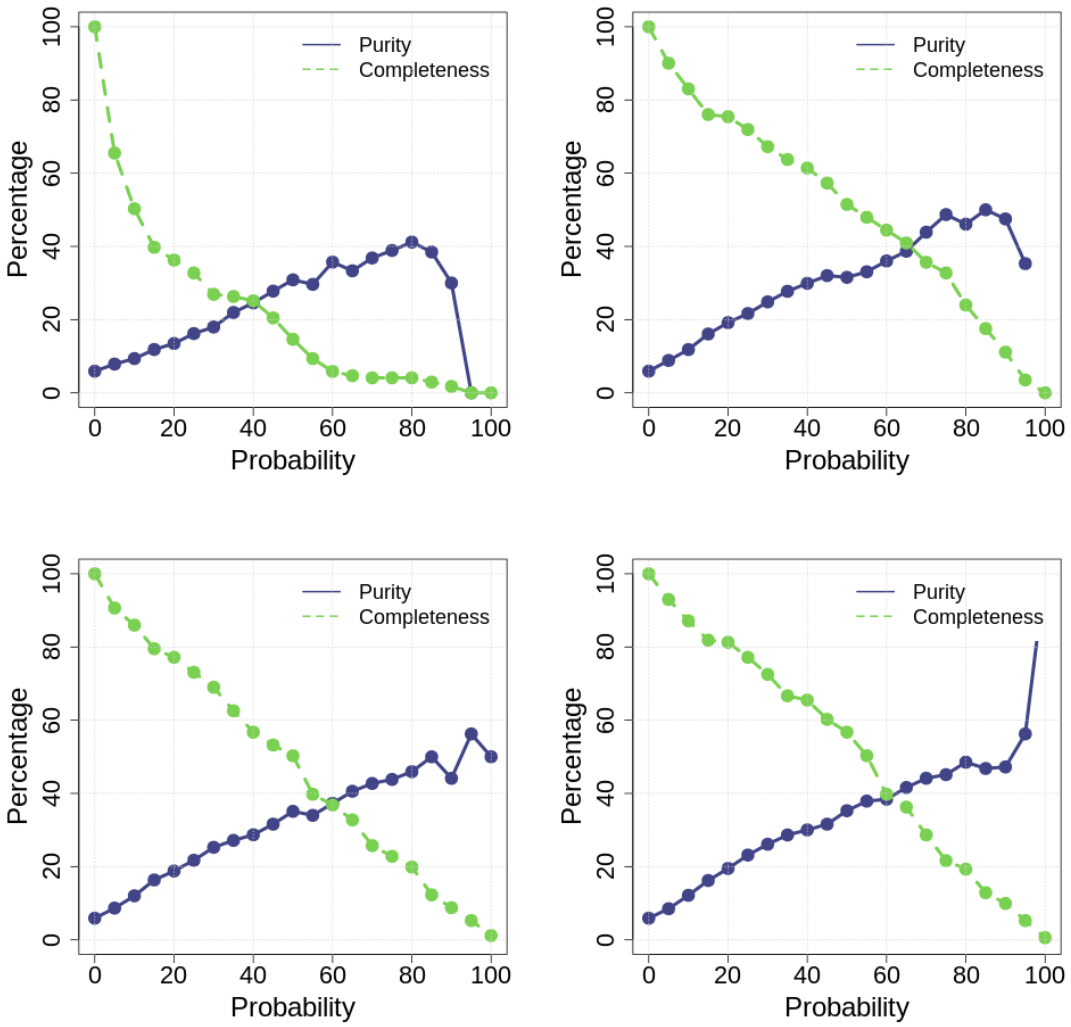


Figure 3.3: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK classification (applied to the data of section 3.1) for different number of PCs being used in the method. Top - Left: Completeness and Purity of UPMASK when using 2 Principal Components, Right: Completeness and Purity of UPMASK when using 3 Principal Components; Bottom - Left: Completeness and Purity of UPMASK when using 4 Principal Components, Right: Completeness and Purity of UPMASK when using 5 Principal Components.

Principal Component	Standard Deviation
PC_1	2.604481
PC_2	1.792770
PC_3	9.108574×10^{-1}
PC_4	3.631163×10^{-1}
PC_5	2.028331×10^{-1}
PC_6	4.036392×10^{-15}
PC_7	2.559386×10^{-15}
PC_8	2.356076×10^{-15}
PC_9	1.541463×10^{-15}
PC_{10}	1.157850×10^{-15}
PC_{11}	7.081957×10^{-16}

Table 3.2: Principal Components of the photometric data described in section 3.1 and the Standard Deviation for each PC

PC	Time (s)	PC	Time (s)
2	1073	3	1193
4	1224	5	1249

Table 3.3: CPU running time according to the number of Principal Components used.

shown in table 3.2. Its possible to see that the first five principal components are those that most contribute to the data representation. As such, I have applied the UPMASK method to the data when only taking into account 2, 3, 4 or 5 principal components. I have used the KDE method, and a number of objects per K-means cluster equal to 50. The curves of completeness and purity for each case are represented in figure 3.3. The CPU running times for each of these scenario are also represented in table 3.3.

Taking into account the completeness and purity curves of figure 3.3 and the computational time that it took to perform each of the tests (Table 3.3), I have chosen to use the first 4 principal components for the applications below. Using the first 5 principal components would include more information about our data, however the difference between these and the test runs with 4 principal components does not justify the additional run time.

3.2.2 Parameters Impacting KDE Versions

KDE - Original UPMASK Method

In this section I study the way internal UPMASK parameters (also known as hiper-parameters) impact the completeness, purity and the CPU time results that were obtained with the original UPMASK method (described in section 2.1). The first parameter I consider here is the number of objects per K-means cluster (i.e, the mean number of objects in a cluster determined by the clustering algorithm). To do so I run the original version of UPMASK with 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 objects per K-means cluster. For each of these runs we set *nruns*= 100. Figure A.1 of the Appendix A shows the results obtained for each of these cases. I then repeated the runs (i.e. with the same set of number of objects per K-mean cluster) for the case of *nruns*= 10. These results are represented in figure A.2 of the Appendix A. Here, I selected to re-display the cases 50, 95 and 125 objects per K-mean clusters in figure 3.4. The completeness and purity curves of these three cases are represented in the top row for *nruns*= 100

$\#_{objects}$	KDE Time_{100} (s)	KDE Time_{10} (s)	KDE Fit Time_{100} (s)	KDE Fit Time_{10} (s)
20	1962	254	1960	242
35	1467	256	1463	245
50	1299	274	1298	262
65	1235	321	1236	309
80	1200	354	1199	342
95	1158	422	1157	405
110	1190	498	1193	477
125	1254	544	1255	521
140	1387	659	1386	630
165	1566	752	1565	749
180	1693	878	1693	871

Table 3.4: CPU running time of UPMASK using KDE, for different number of objects per cluster. KDE Time_{100} is the CPU running time for the tests realized with the original UPMASK method, made with *nruns*= 100, KDE Time_{10} is the CPU running time for the tests realized with the original UPMASK method, made with *nruns*= 10, KDE Fit Time_{100} is the CPU running time for the tests realized with UPMASK with the KDE and a fitting function implemented (section 2.6), made with *nruns*= 100, and KDE Fit Time_{10} is the CPU running time for the tests realized with UPMASK with the KDE and a fitting function implemented (section 2.6), made with *nruns*= 10.

and in the third row for *nruns*= 10. Table 3.4 presents the CPU running times for all of these runs. The columns KDE Time_{100} and KDE Time_{10} are the CPU times for *nruns*= 100 and *nruns*= 10, respectively.

KDE with a Fitting Function (modification V)

In this subsection is shown the time tests for the UPMASK method with the KDE and a fitting function implemented, as described in section 2.6 for different number of objects per K-means cluster. This study is presented in figures A.3 and A.4 of the Appendix A for the following number of objects per K-means cluster: 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 objects. In table 3.4 is shown how much time this version of UPMASK (KDE Fit Time_{100} for *nruns*= 100 and KDE Fit Time_{10} for *nruns*= 10) took to run for each number of objects per K-means cluster. In this chapter, it was selected only some of the tests to show (50, 95 and 125). The completeness and purity curves of these three cases are represented in the second row of figure 3.4 for *nruns*= 100 and in the fourth row for *nruns*= 10.

Dependence on the threshold level

Next we study the impact of changing the threshold level of UPMASK, equation 2.3 in section 2.1. The default value is $T = 1$, but that does not mean that is an optimal value to use with KDE for all cases of interest. As such, we have performed test runs for $T = \{0.5, 1, 1.5, 2, 2.5, 3\}$ using the KDE, a number of objects per cluster equal to 50 and with *nruns*= 100. The completeness and purity curves for each case are represented in figure 3.5.

3.2.3 Parameters Impacting Voronoi Versions

Voronoi + Anderson-Darling Test (modification I)

In this subsection is shown the time tests for the UPMASK method with the Voronoi and the Anderson-Darling test implemented, as described in section 2.2, for different number of objects

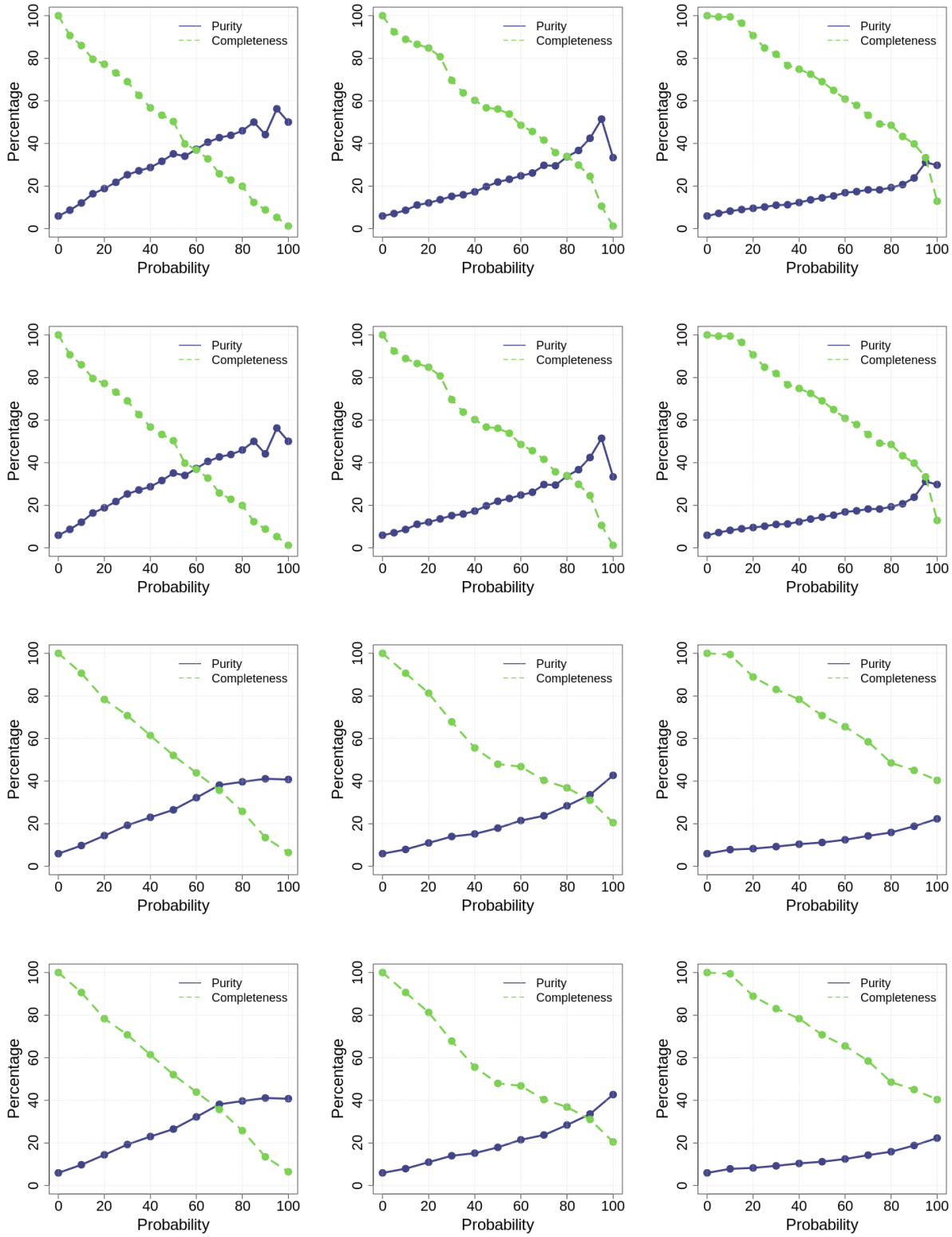


Figure 3.4: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK classification with the KDE implementation (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). The left, middle and right column is the test performed for 50, 95 and 165 objects per cluster, respectively. In the top row are the tests for the original UPMASK method (*nruns*= 100). In the second row are the tests for the UPMASK method with the KDE and a fitting function implemented (*nruns*= 100). In the third row are the tests for the original UPMASK method (*nruns*= 10). In the fourth row are the tests for the UPMASK method with the KDE and a fitting function implemented (*nruns*= 10)

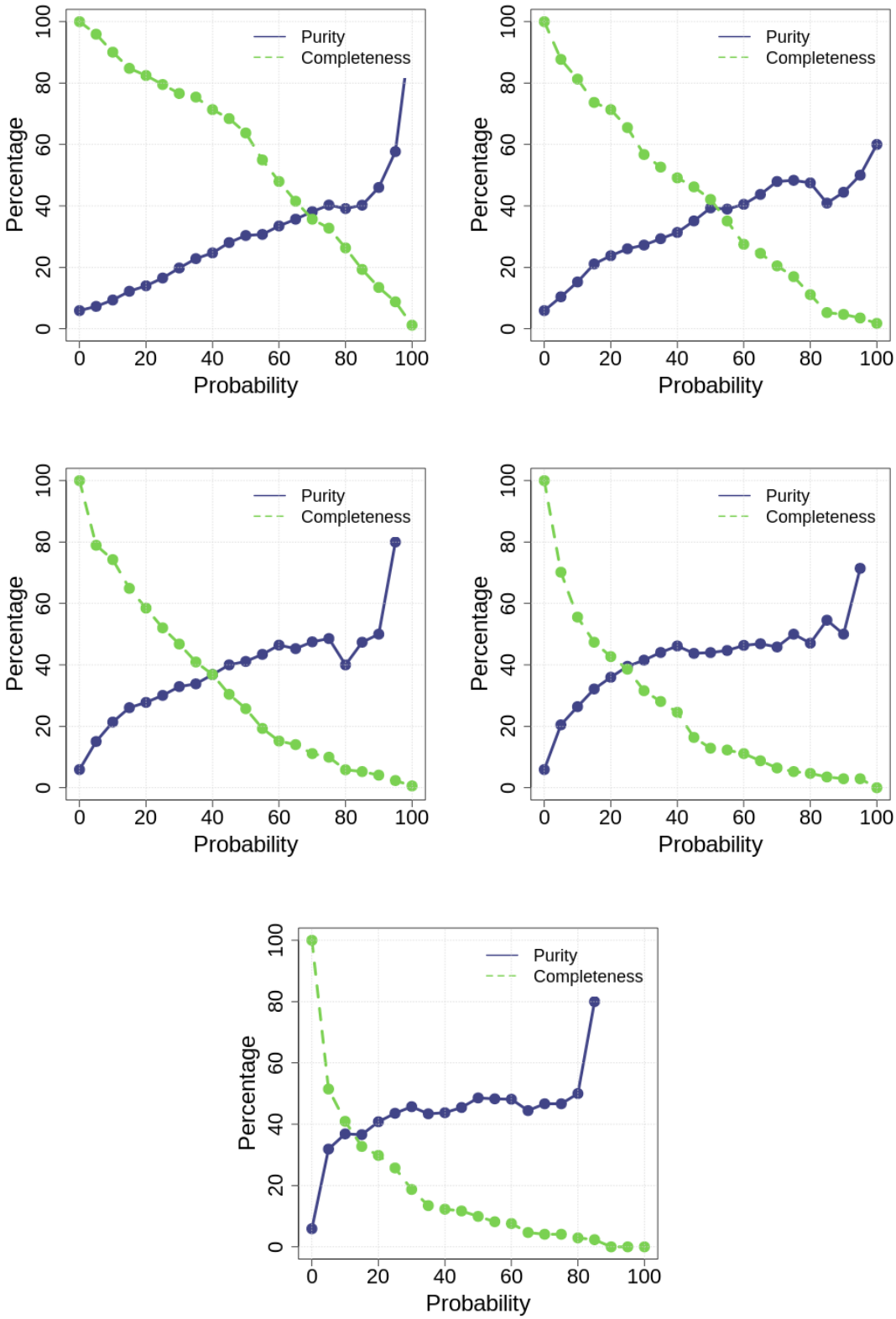


Figure 3.5: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK classification with the KDE implementation (applied to the data of section 3.1) for different values of threshold level, T (equation 2.3). From left to right, top to bottom, the tests where perform with $T = \{0.5, 1, 1.5, 2, 2.5, 3\}$, while keeping the other parameters equal (4 PCs, *nruns*= 100, 50 objects per cluster from the K-means)

$\#_{objects}$	Voronoi$_{AD}$ Time$_{100}$ (s)	Voronoi$_{AD}$ Time$_{10}$ (s)	Voronoi$_{mean}$ Time$_{100}$ (s)	Voronoi$_{mean}$ Time$_{10}$ (s)
20	2811	295	1861	187
35	2002	215	1350	146
50	1758	190	1228	148
65	1396	159	1091	137
80	1256	144	1102	142
95	1059	126	977	144
110	1026	129	972	153
125	996	124	1014	164
140	863	115	1034	185
165	866	117	997	217
180	836	117	982	204

Table 3.5: CPU running time of UPMASK using Voronoi, for different number of objects per cluster. Voronoi$_{AD}$ Time$_{100}$ is the CPU running time for the tests realized with the UPMASK Voronoi and Anderson-Darling Test implemented (section 2.2), made with *nruns*= 100, Voronoi$_{AD}$ Time$_{10}$ is the CPU running time for the tests realized with the UPMASK Voronoi and Anderson-Darling Test implemented (section 2.2), made with *nruns*= 10, Voronoi$_{mean}$ Time$_{100}$ is the CPU running time for the tests realized with UPMASK with the Voronoi and a comparison of means (subsection 2.3), made with *nruns*= 100, and KDE Fit Time$_{10}$ is the CPU running time for the tests realized with UPMASK with the Voronoi and a comparison of means implemented (subsection 2.3), made with *nruns*= 10.

per K-means cluster. This study is presented in figures A.5 and A.6 of the Appendix A for the following number of objects per K-means cluster: 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 objects. In table 3.5 is shown how much time this version of UPMASK (Voronoi$_{AD}$ Time$_{100}$ for *nruns*= 100 and Voronoi$_{AD}$ Time$_{10}$ for *nruns*= 10) took to run each of objects per K-means cluster. Here we re-display only the cases of 50, 95 and 125 objects per k-means clusters. The completeness and purity curves of these three cases are represented in the first row of 3.6 for *nruns*= 100 and in the third row for *nruns*= 10.

- **Dependence on the threshold level**
  The Voronoi with the Anderson-Darling test implementation also has a threshold level T as one if its parameters, as described in equation 2.4 in section 2.2. The default value is $T = 1$ but here I also perform a study of the variation of this parameter T. I have also adopted the following values of $T = \{0.5, 1, 1.5, 2, 2.5, 3\}$ using the Voronoi and the Anderson-Darling test. The results for the completeness and purity curves for each case are represented in figure 3.7.

Voronoi Mean (modification II)

This section shows impact of changing internal parameters of the UPMASK method with the Voronoi and a mean comparison implementation (modification II), see subsection 2.3. This study is presented in figures A.7 (*nruns*= 100) and A.8 (*nruns*= 10) of the Appendix A for the following number of objects per K-means cluster: 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 objects. In table 3.5 is shown how much time this version of UPMASK (Voronoi$_{mean}$ Time$_{mean}$ for *nruns*= 100 and Voronoi$_{mean}$ Time$_{mean}$ for *nruns*= 10) took to run for each number of objects per K-mean cluster. Here, I again re-display only the cases 50, 95 and 125 objects

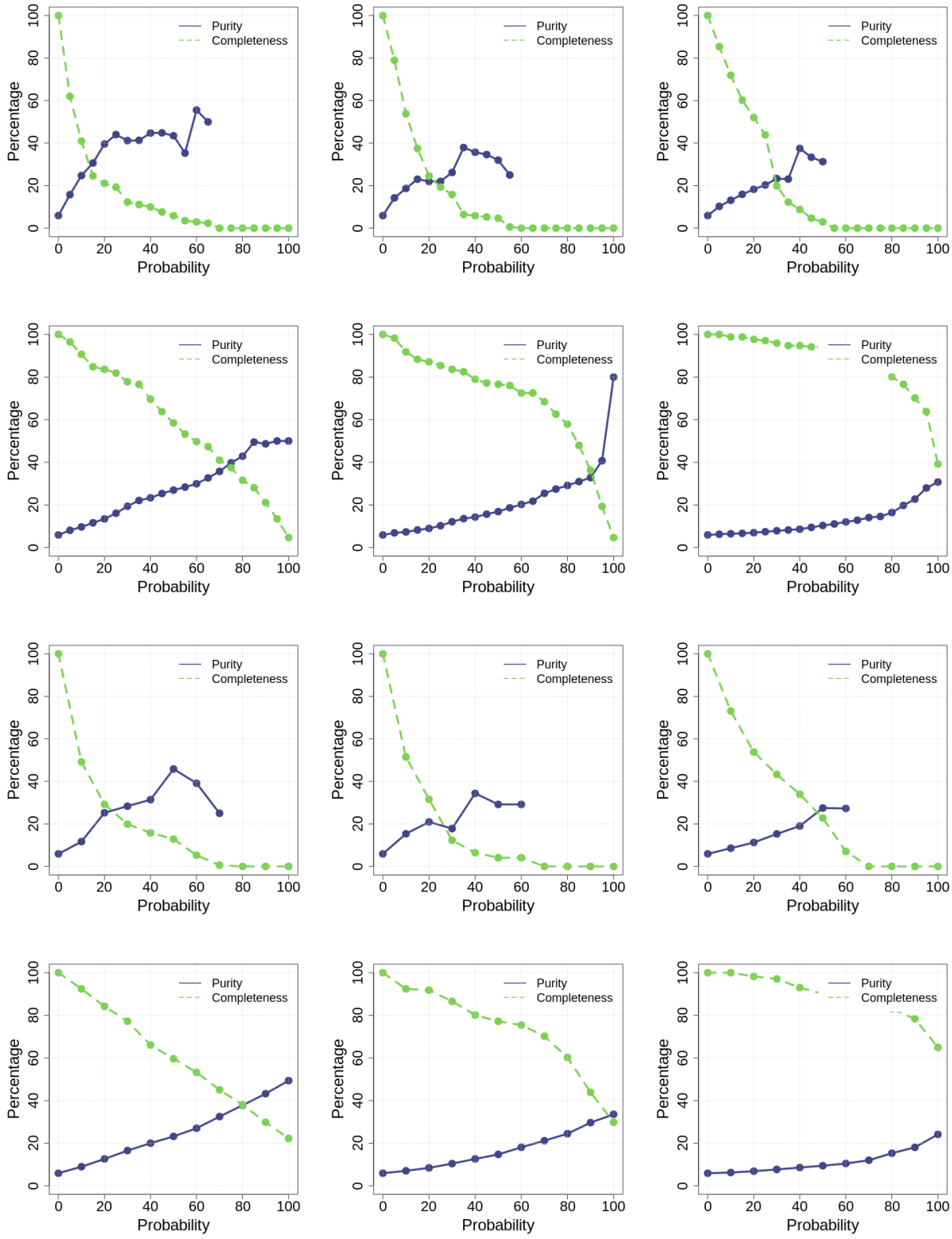


Figure 3.6: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK classification with the Voronoi implementation (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). The left, middle and right column is the test performed for 50, 95 and 165 objects per cluster, respectively. In the top row are the tests for the UPMASK method with the Voronoi and Anderson-Darling implemented ($nruns$= 100). In the second row are the tests for the UPMASK method with the Voronoi and a mean comparison implemented ($nruns$= 100). In the third row are the tests for the UPMASK method with the Voronoi and Anderson-Darling implemented ($nruns$= 10). In the fourth row are the tests for the UPMASK method with theVoronoi and a mean comparison implemented ($nruns$= 10)

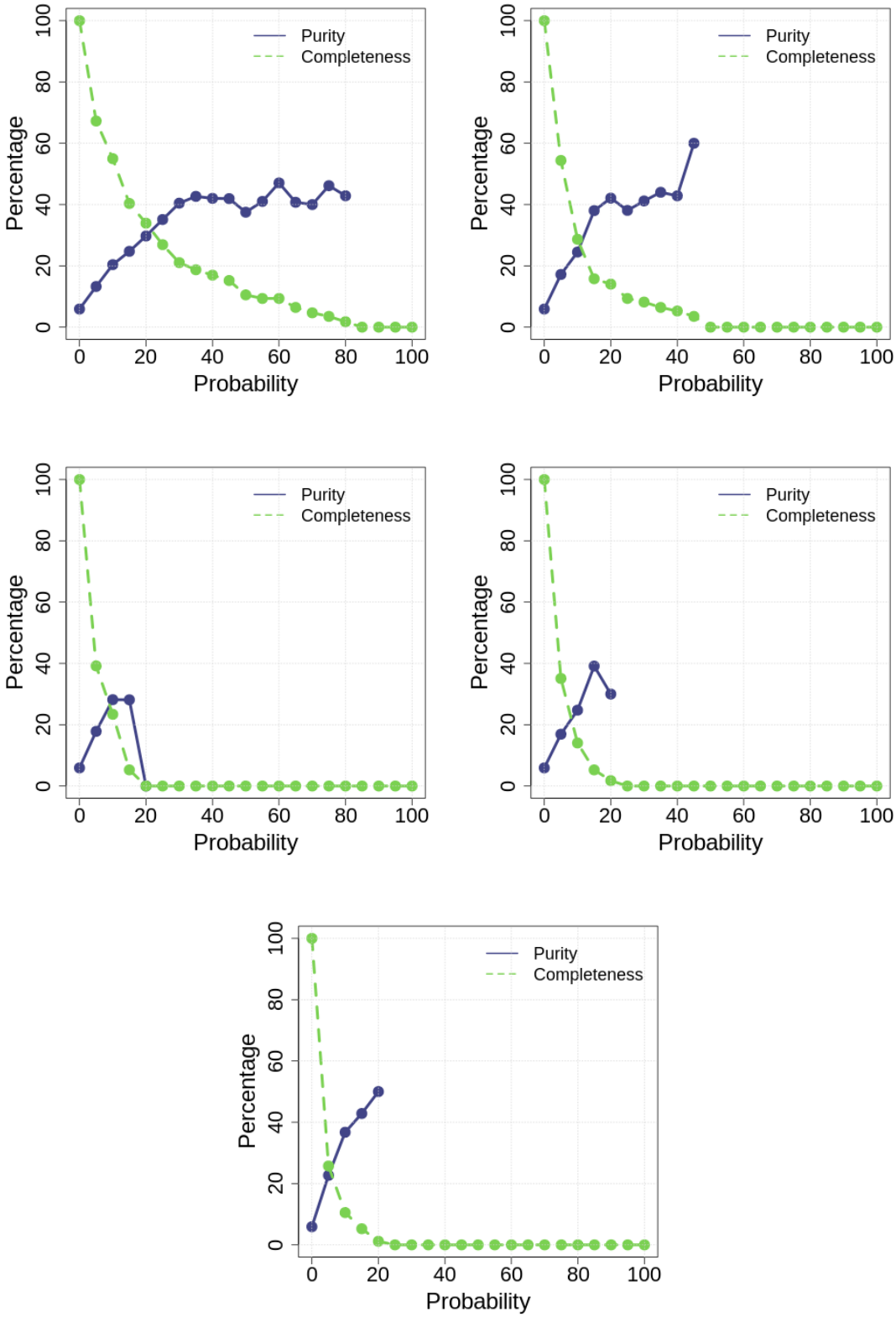


Figure 3.7: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK classification with the Voronoi + Anderson-Darling Test implementation (applied to the data of section 3.1) for different values of threshold level, T (equation 2.3). From left to right, top to bottom, the tests where perform with $T = \{0.5, 1, 1.5, 2, 2.5, 3\}$, while keeping the other parameters equal (4 PCs, *nruns*= 100, 50 objects per cluster from the K-means)

$\#_{objects}$	Grid Time_{100} (s)	Grid Time_{10} (s)	Grid Fit Time_{100} (s)	Grid Fit Time_{10} (s)
20	2603	333	2594	319
35	1761	301	1750	290
50	1522	303	1521	289
65	1410	367	1406	353
80	1308	370	1309	355
95	1286	398	1285	379
110	1279	446	1280	428
125	1355	496	1352	479
140	1313	522	1310	502
165	1359	596	1357	596
180	1283	602	1289	607

Table 3.6: CPU running time of UPMASK using Grid, for different number of objects per cluster. Grid Time_{100} is the CPU running time for the tests realized with the UPMASK method with a Grid implemented (2.4), made with *nruns*= 100, Grid Time_{10} is the CPU running time for the tests realized with the original UPMASK method with a Grid implemented (2.4), made with *nruns*= 10, Grid Fit Time_{100} is the CPU running time for the tests realized with UPMASK with the Grid and a fitting function implemented (section 2.5), made with *nruns*= 100, and Grid Fit Time_{10} is the CPU running time for the tests realized with UPMASK with the Grid and a fitting function implemented (section 2.5), made with *nruns*= 10.

per K-means cluster. The completeness and purity curves of these three cases are represented in the second row of 3.6 for *nruns*= 100 and in the fourth row for *nruns*= 10.

- **Dependence on the threshold level**
  The UPMASK version with the Voronoi and a mean comparison also has a threshold level T, as described in equation 2.5 in subsection 2.3. This threshold level is $T = 3$ by default, but I have studied different values of T such as $T = \{0.5, 1, 1.5, 2, 2.5, 3\}$. The results for the purity and completeness curves for each value of T are represented in figure 3.8.

3.2.4 Parameters Impacting Grid Versions

Grid (modification III)

In this subsection is shown the time tests for the modified UPMASK method using the contingency table (Grid), as described in section 2.4, for different number of objects per cluster, i.e, the mean number of objects in a cluster determined by the clustering algorithm. This study is presented in figures A.9 (*nruns*= 100) and A.10 (*nruns*= 10) of the Appendix A for the following number of objects per K-means cluster: 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 objects. In table 3.6 is shown how much time this version of UPMASK (Grid Time_{100} for *nruns*= 100 and Grid Time_{10} for *nruns*= 10) took to run for each number of objects per K-mean cluster. Here, I again re-display only the cases 50, 95 and 125 objects per K-means cluster. The completeness and purity curves of these three cases are represented in the second row of 3.9 for *nruns*= 100 and in the fourth row for *nruns*= 10.

Grid with a Fitting Function (modification IV)

In this subsection is shown the time tests for the modified UPMASK method using the contingency table (Grid) and a fitting function, as described in section 2.4, for different number of objects per K-means cluster. This study is presented in figures A.11 (*nruns*= 100) and A.12

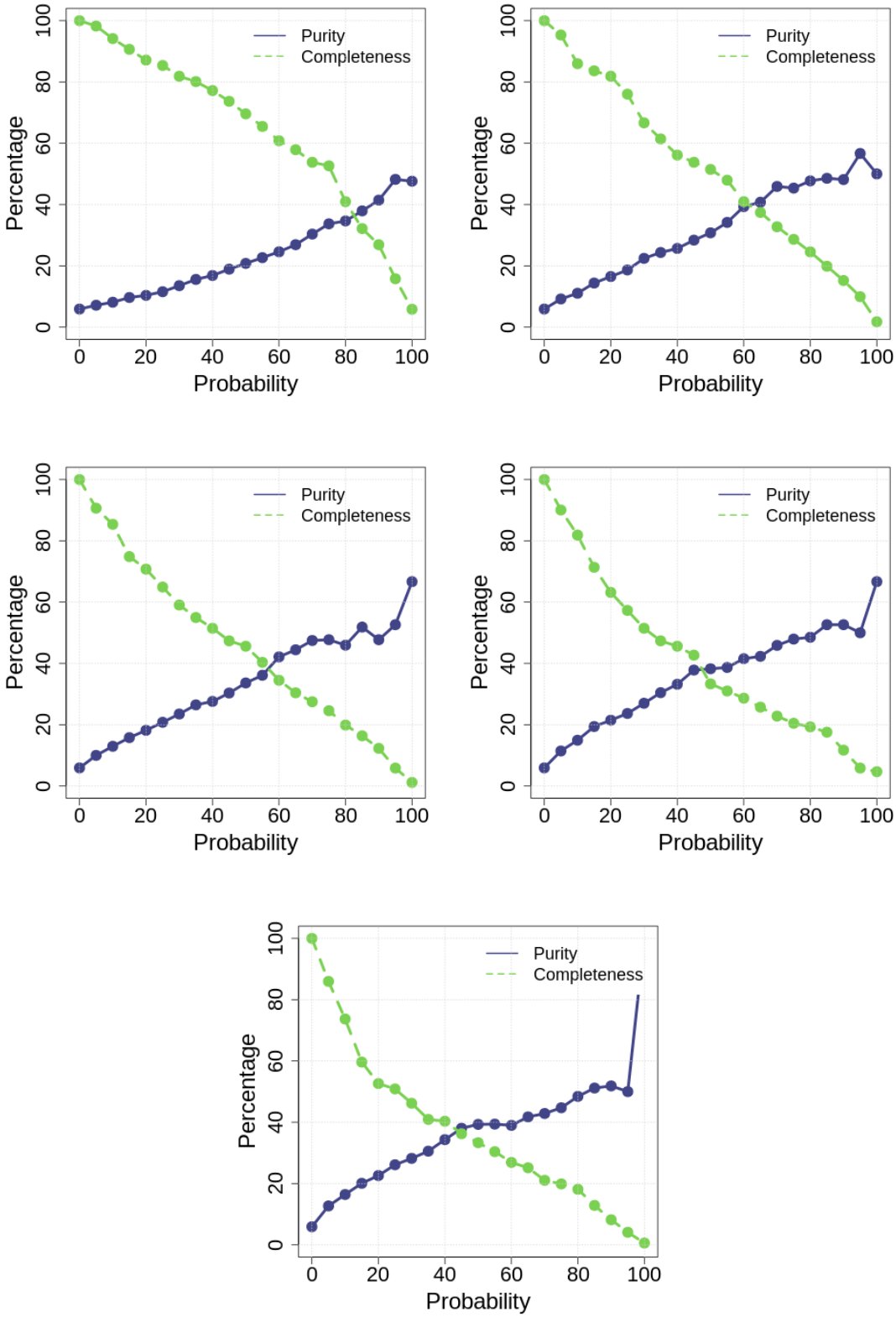


Figure 3.8: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK classification with the the Voronoi and a mean comparison implementation (applied to the data of section 3.1) for different values of threshold level, T (equation 2.3). From left to right, top to bottom, the tests where perform with $T = \{0.5, 1, 1.5, 2, 2.5, 3\}$, while keeping the other parameters equal (4 PCs, *nruns*= 100, 50 objects per cluster from the K-means)

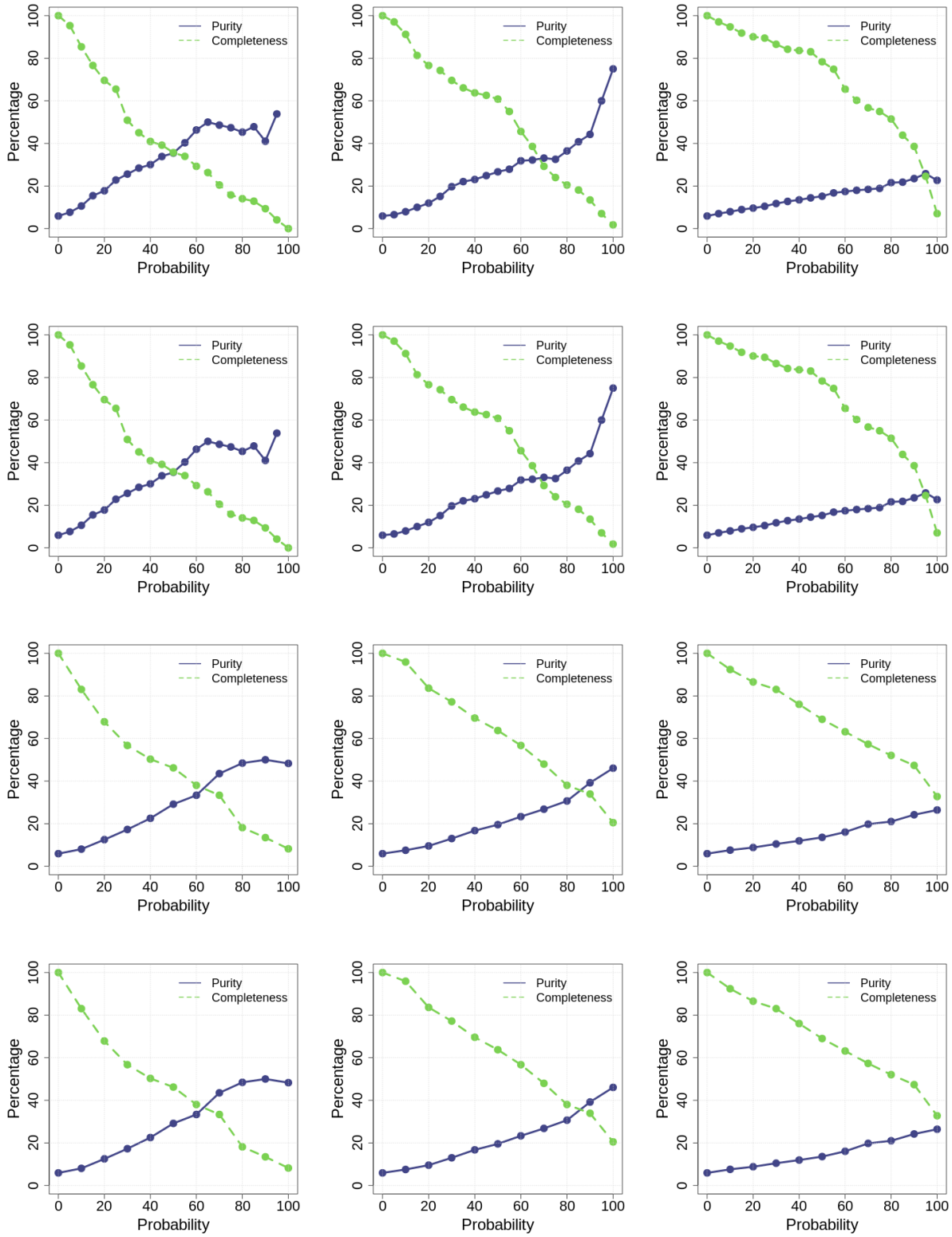


Figure 3.9: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK classification with the Grid implementation (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). The left, middle and right column is the test performed for 50, 95 and 165 objects per cluster, respectively. In the top row are the tests for the UPMASK method with the Grid (*nruns*= 100). In the second row are the tests for the UPMASK method with the Grid and a fitting function implemented (*nruns*= 100). In the third row are the tests for the UPMASK method with the Grid (*nruns*= 10). In the fourth row are the tests for the UPMASK method with the Grid and a fitting function implemented (*nruns*= 10)

(*nruns*= 10) of the Appendix A for the following number of objects per K-means cluster: 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 objects. In table 3.6 is shown how much time this version of UPMASK (Grid Fit Time_{100} for *nruns*= 100 and Grid Fit Time_{10} for *nruns*= 10) took to run for each number of objects per K-means clusters. Here, I again re-display only the cases 50, 95 and 125 objects per K-means cluster. The completeness and purity curves of these three cases are represented in the second row of 3.9 for *nruns*= 100 and in the third row for *nruns*= 10.

Dependence on the threshold level

Similarly to the KDE implementation, the Grid implementation in UPMASK is also dependent of a threshold value T, that is analogous to the one described in equation 2.3. The tests above applied a $T = 1$ but a study needs to be performed in order to understand the impact of this parameter. As such, I have performed the test for $T = \{0.5, 1, 1.5, 2, 2.5, 3\}$ using the Grid with a fitting function, a number of objects per K-means cluster equal to 50 and with *nruns*= 100. The completeness and purity curves for each case are represented in figure 3.10.

3.3 Results

Considering the results presented in figure 3.1, the adoption of the UPMASK unsupervised method using DES bands is more efficient for lower-z ($z < 1$) than using the Euclid bands. This results from the position of one of the main features of galaxy spectra (the rest-frame 4000$\mathring{A}$ break) in the observed spectra. We expect this tendency to change for higher redshift clusters, as this is mainly driven by the difference in the wavelength range of both filters systems, as seen in figures 1.3 and 1.2. For the analysis of the studied cluster ($z = 1.09$), the DES filters attain a slightly higher purity, however they do result in a significant improvement in the completeness at most probability thresholds. We plan in the future to use simulated cluster populations spanning over a wider range of redshifts to better characterize and understand the behaviour of the method when applied to DES and Euclid filter systems.

The modification with the fitting function (either with KDE or Grid implementation), give the same result. Moreover, for low number of runs, the UPMASK is improved in processing time, although of only about 5%. However, when applied to more fields, this 5% improvement will be essential. The modified Voronoi+Anderson-Darling-based implementation of the method resulted in a significant CPU time improvement in comparison to the original KDE-based implementation ($\sim 30\%$). However, comparing the top left and top right plots of Figure 3.2, it is noticeable that the KDE version results in a sample of probable members that are more spatially concentrated than the Voronoi+AD version. Also, the left and right plots in the second row of Figure 3.2, indicate that the Voronoi+AD implementation results into a more incomplete sample with respect to the KDE-based implementation. This is due the fact that the Anderson-Darling test is a stricter test (it compares the two whole sample distribution) while the KDE comparison only depends on simple statistical calculations, such as the mean and the standard deviation. When the Anderson-Darling test is replaced by the comparison of the mean of the Voronoi areas distribution, the completeness is improved with respect to the previous version (figure 3.2). The Grid modification seems to result in a slightly purer sample when comparing to the KDE implementation, but overall they look very similar. It is unclear why the Grid implementation

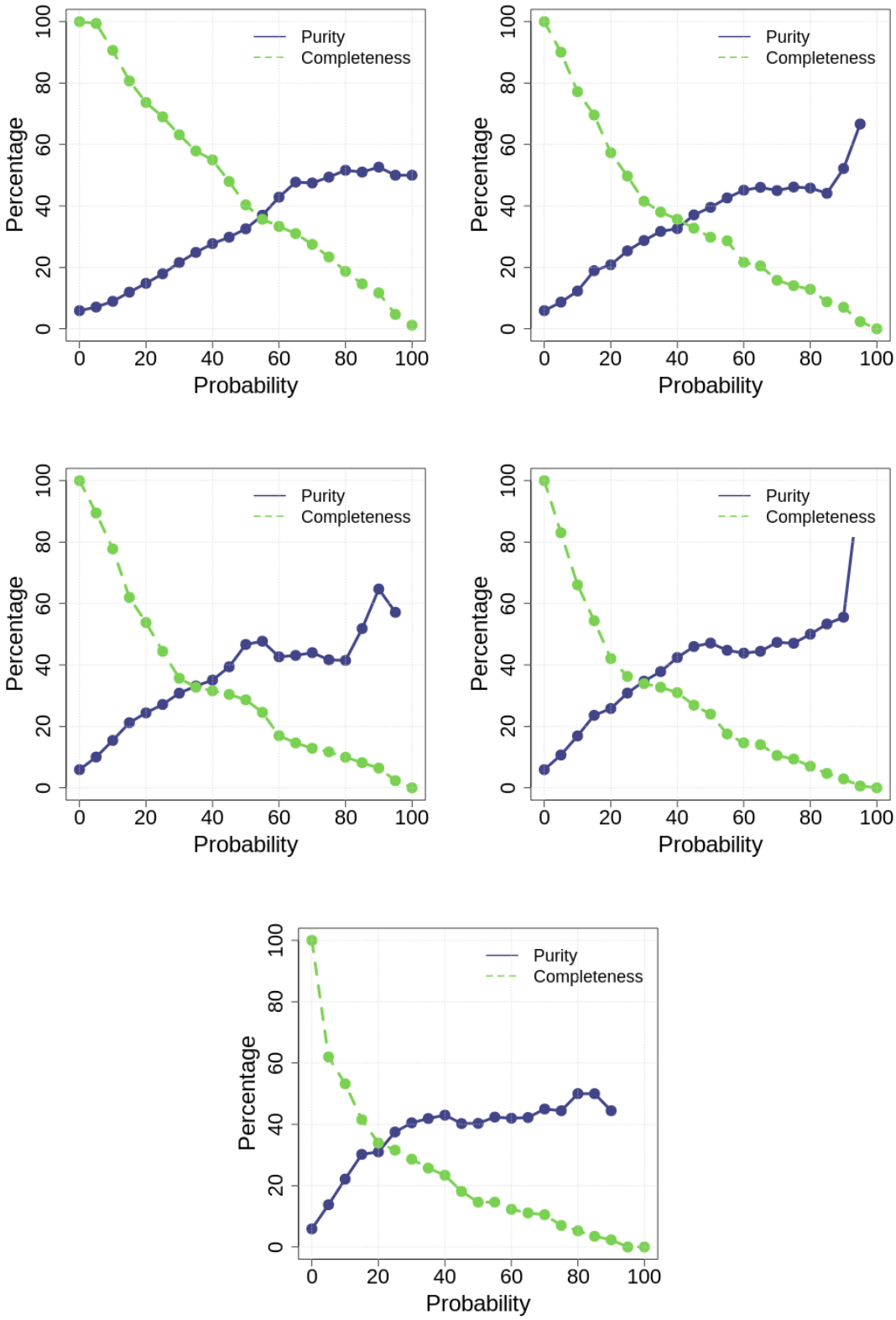


Figure 3.10: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK classification with the Grid implementation (applied to the data of section 3.1) for different values of threshold level, T (equation 2.3). From left to right, top to bottom, the tests where perform with $T = \{0.5, 1, 1.5, 2, 2.5, 3\}$, while keeping the other parameters equal (4 PCs, *nruns*= 100, 50 objects per cluster from the K-means)

takes more time to run than the KDE implementation, as the later needs to perform complex computations for the density estimations while the Grid tasks are more simple. When using the Grid outside the method, the CPU running time seems to be faster than when using it together with UPMASK. This should be a challenge of compiling functions and packages and that lies outside the scope of this dissertation.

The analysis about the principal components of the photometric data of Euclid magnitudes was mentioned in section 3.2.1, where it was computed the standard deviation for each principal component. The sixth component and beyond have a very small standard deviations and therefore they can be disregarded. From the results in table 3.3, where it is shown the times it takes for UPMASK to run with 2, 3, 4 and 5 principal components, and the results in figure 3.3, I decided to adopt 4 principal components for the rest of the dissertation, since the difference in completeness and purity between the test that uses 4 principal components and the one that uses 5 does not justify taking more time for UPMASK to run.

The times in the columns KDE Time_{100} and KDE Fit Time_{100} of table 3.4 start to decrease, and then they increase again. This behaviour is also present in the columns Grid Time_{100} and Grid Fit Time_{100} of table 3.6. The first part can be explained by the fact that by dividing the data in small groups, it will mean many more groups to be analyzed individually, and by raising the number of objects by groups (and therefore reducing the number of groups), the CPU running time will decrease. The increase in time in the second part might be caused by taking more time for the method to converge. The higher the number of points in the same area, the closer this area might look to an uniform distribution of points, and as such, in one iteration this area is classified as belonging to a bound object, in the next one, this area might be classified as an uniform one, making it harder for the method to converge - and as a result, the method performs more internal iterations.

Changing the number of objects per cluster of the K-means, not only impacts the CPU running time, but also the completeness and purity. Overall, comparing the plots in figures 3.4, 3.6 and 3.9 (this is more clear when also looking at the figures in appendix A), the higher the numbers being grouped in the same cluster of the K-means, the higher the completeness gets, but the purity also decreases. The higher number of objects in a group, more contaminated this group is, and therefore if this group will be classified as a group that belongs to a cluster, the purity will decrease. Since there are more objects that are not ignored in the method, the completeness will logically increase. It is not possible to say what is the optimal number of objects per K-mean clusters, because it depends on the type of study that the user aims to do. If the goal is to achieve a pure sample, then using a small number of objects per K-means is ideal and on the other hand, to achieve a more complete sample, then a larger number of objects per K-means should be used.

Comparing the tables of time between the KDE and the KDE with a fitting function, the time difference of the application with $nruns = 100$ is not much different. This is due to the fact that the method itself already has a lookup table implemented, and as such, the fit version for bigger number of runs and the one without the fit are equivalent.

Still comparing the figures of the method as it is (the original UPMASK version) with its version with a fitting function, see figures 3.4 and 3.9, it is evident that the fitting function produces the same result (the same seed was used, however if a seed is not set, then it is expected that the method returns compatible results with small fluctuations. This is important for the case where small number of runs is used, because it accelerates the method.

From the figures of the threshold level variation tests (figures 3.5, 3.7, 3.8 and 3.10), the smaller the threshold, the higher the completeness. The purity also seems to be affected, although not as clear as the completeness. The trend is that for a higher threshold level, the purity also increases. Increasing the value of the threshold level, it increases the value the total threshold of the method needs to achieve so that a group from the K-means is classified as belonging to a cluster. By increasing the threshold, there are more groups being classified as following an uniform distribution, and therefore the completeness will decrease. Following the same logic, groups from the K-means that are without a doubt clustered, will be classified as such and be kept in the UPMASK kernel.

Finally, although there is certainly more room for accelerating and tailoring the UPMASK methodology to be adopted for cluster finding (the Grid implementation has potential to be improved), the original implementation of the method seems to provide a promising unsupervised methodological alternative for membership determination in galaxy cluster studies.

Chapter 4

Coma Cluster

Now that we have validated the UPMASK method and my modifications in simulated data, I am going to apply it to real data of the Coma cluster. The Coma Cluster is one of the most well known galaxies clusters and one of the most studied. The more than 10^3 galaxies living in this cluster are typically elliptical and lenticular galaxies. The cluster, located at $\alpha = 12^h59^m48.7^s$ (α_{J2000}), $\delta = 27°58'50''$ (δ_{J2000}), has an angular size between 1 to 2 square degrees [Omer et al., 1965] and a mean redshift $z = 0.023$ [Mahdavi and Geller, 2001]. As mentioned in section 1.2, F. Zwicky showed [Zwicky, 1933] that the Coma cluster contains a huge fraction of non visible mass that he, historically, named "dark matter". Since the Coma cluster is one of the most well known clusters, with its galaxy members being investigated in different studies [e.g. Fossati et al., 2013, Hammer et al., 2010, Yagi et al., 2016, Zwicky, 1951], one can, in principle, use Coma as a "test bed" for the validation and bench-marking of the UPMASK method, through the computation of purity and completeness functions using real galaxies. Here we choose not to compute such quantities because different studies provide different numbers of member galaxies and used different observational data sets from the one we adopt in this dissertation - the Pan-STARRS survey (see section 4.1). Here we will focus on a more qualitative validation of the different versions of the method to detect and image clusters and its substructures (see section 4.2 and 4.3).

4.1 Optical Study of the galaxies of the Coma Cluster

In this dissertation we decided to use recent observations of the Pan-STARRS (Panoramic Survey Telescope and Rapid Response System) survey [Chambers et al., 2016] as input for our different versions of the UPMASK method. Here I used the second data release of the Pan-STARRS catalogue [Flewelling et al., 2016], published in 28th January 2019, and extracted astrometric and photometric information for all galaxies around a squared region that contained the Coma cluster (see text below).

The Pan-STARRS survey has two main surveys: the 3π survey and the Medium Deep Survey [Chambers et al., 2016]. The 3π covers 3π steradians ($\approx 30 \times 10^3$ deg^2) of the sky in 5 filters ($grizy_{P1}$) and has a depth of 21-23 mag. The Medium Deep Survey covers 10 fields of 7 deg^2 each and has a depth of 24-26 mag. The telescope of the survey is located in Haleakala, Hawaii. This telescope has a diameter of 1.8 m and a field of view diameter of 3 degrees, making a field of view of 7 square degrees. The camera of the telescope is the Gigapixel Camera #1 (GPC1). This camera consists of an 8×8 array of orthogonal transfer array (OTA) CCDs, and each OTA

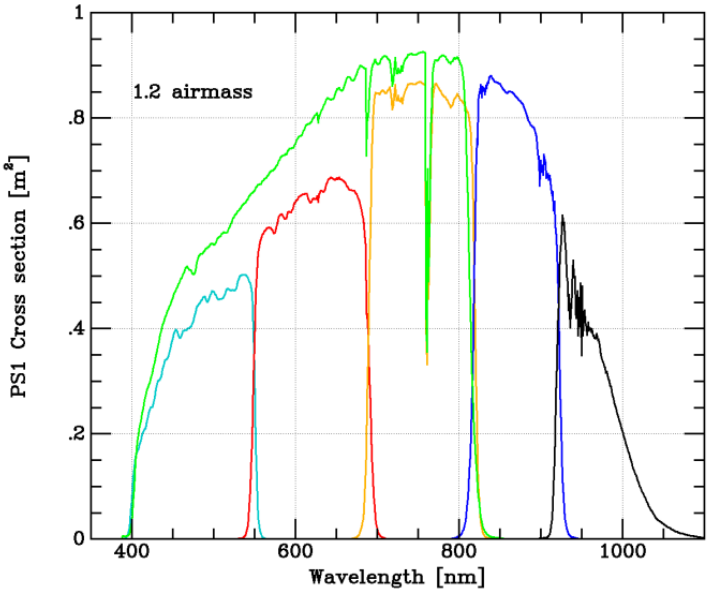


Figure 4.1: Cross section of the PanStarrs Filters: g (light blue), r (red), i (yellow), z (dark blue), y (black) and open (green) [Chambers et al., 2016].

is divided in an 8×8 array of cells, each cell being an independent 590×598 $10\,\mu$m pixel CCD. Pan-STARRS uses the $grizyw_{P1}$ filter system. The notation $P1$ is used to not be confused with other photometric systems. The filters $grizy_{P1}$ cover the 3π sky area but this survey is not covered by the w_{P1} filter, since the latter is mostly used for closer objects, and as such the data related to this filter is not used in this dissertation. The Pan-STARRS filters are represented in figure 4.1. The mean filter wavelengths are: g - $486.6\,nm$, r - $621.5\,nm$, i - $754.5\,nm$, z - $867.9\,nm$ and y - $963.3\,nm$.

As mentioned earlier, the data used is from the second data release of Pan-STARRS [Flewelling et al., 2016]. The following parameters were extracted from this catalogue: object identifier, astrometry (α, δ), the number of detections, and for each filter (g, r, i, z, y), the following parameters were also extracted - number of single epoch detection in the filter, the mean PSF (Point Spread Function) magnitude from the filter detection and its standard deviation, the mean Kron magnitude [Kron, 1980] from the filter detection and its standard deviation. From this catalogue we selected all the objects that are within a radius of 2 degrees around the center of the Coma cluster. The selected set had 3604017 objects.

4.1.1 Separating stars from galaxies

The subset obtained from Pan-STARRS contains mainly stars and galaxies, but there is also other types of objects (for example, asteroids, planets) present. For the purpose of the dissertation, a separation between galaxies and other types of objects is necessary. To perform this separation, I used the difference between the PSF magnitude of a filter with its Kron magnitude, see [Farrow et al., 2014].

Stars are usually point sources, and as such, its PSF magnitude matches the magnitude obtained from its Kron profile. On the other hand, galaxies have different morphologies, are dispersed, and therefore the PSF is not the best profile to use to measure their magnitudes. As such, the PSF magnitudes measured for galaxies are in average, higher than the Kron magnitudes. This method was applied to Pan-STARRS data [Farrow et al., 2014], where the authors

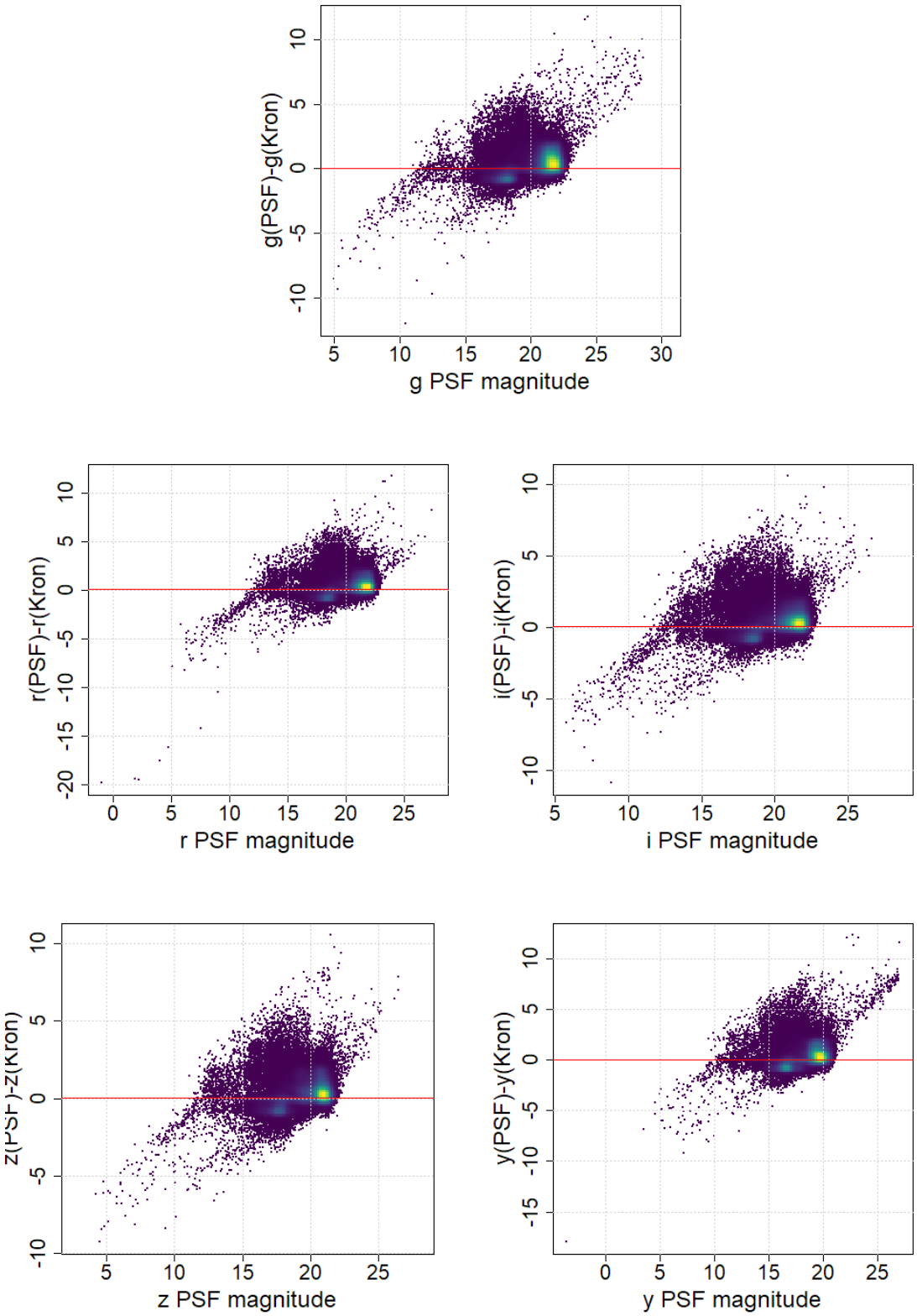


Figure 4.2: Star-Galaxy separation for each filter of the objects in the Coma Cluster. From left to right, top to bottom, is plotted the difference between the PSF magnitudes and the Kron magnitudes (vertical axis) against the PSF magnitude for the Pan-SATRRS g, r, i, z and y bands. The red line represents $PSF_{mag} - Kron_{mag} = 0.05$. Objects above this line are considered as galaxies. The colors indicate the number density of objects (brighter colors correspond to higher densities) in the panels.

suggested a cut of of magnitude differences of $PSF_{mag} - Kron_{mag} = 0.05$ as a way to separate galaxies from stars, i.e. sources above (below) this cut are classified as galaxies (stars). According to the authors, a cut $PSF_{mag} - Kron_{mag} = 0$ is to be avoided because the Kron magnitudes need a correction to the magnitudes. I therefore applied this simple separation rule, calculated the difference for each of the filters, and classified an object as a galaxy if the object had $PSF_{mag} - Kron_{mag} > 0.05$ for every filter. This resulted in a set of 37281 galaxies, about 1% of the objects of the initial data-set.

figure 4.2 shows the color magnitude difference $PSF_{mag} - Kron_{mag}$ for all objects in each filter. The final separation between galaxies and stars was achieved by gathering the objects (galaxies) that fall above the red line, defined by $PSF_{mag} - Kron_{mag} = 0.05$ in all panels of the figure.

The star-galaxy separation could be performed using other methods, for example, using unsupervised clustering methods such as applying the K-means, since this method should be able to discern between two classes of objects in the photometric space, for example. However, the purpose of this dissertation is to identify galaxy clusters and therefore, the star-galaxy separation is assumed as already performed. For that reason, I have chosen the simplest separation recommended by the Pan-STARRS.

After applying the separation, the possible colors of the PSF magnitudes of the Pan-STARRS filter system were calculated, as well as their correspondent uncertainty (computed using simple uncertainty propagation).

4.2 Applying UPMASK

figure 4.3 shows images of the coma cluster obtained by running the original UPMASK method and modified versions (described in chapter 2) with the selected galaxy data-set. Each panel correspond to an UPMASK version. Starting at the top to the bottom, from left to right we have the following sequence of versions: UPMASK with KDE (original version), with KDE+Fitting Function, with Voronoi+Anderson-Darling Test, with Voronoi+Mean Comparison, with Grid, and with Grid+Fitting Function. For each case, the input variables were the galaxy sky coordinates (α, δ), the mean PSF magnitudes (of the available filters - g, r, i, z and y), their uncertainty, the colors calculated as described in the previous section and their corresponding uncertainties. The number of Principle Components (PCs) was set equal to 4 PCs in all versions of the method. This is justified in section 3.2.1, where it is explained that these 4 components contain the majority of information about the data (these are the PCs with the biggest standard deviations in descending order of importance) and keep individual CPU running times manageable for all version methods. The number of objects per K-means cluster was set equal to 50 because, although the Coma Cluster is populated by more than 10^3, the authors of UPMASK recommend a usage of number of objects per K-means cluster that is able to divide the data in smaller samples [Krone-Martins and Moitinho, 2014]. To achieve a resolution of 1% in the frequentist membership probabilities given by UPMASK, the number of runs was chosen equal to 100. As explained at the beginning of this chapter we decided not to address the purity and completeness of our detected objects, because the membership of real cluster galaxies from Pan-STARSS and its relation to existing studies of galaxy membership in Coma is not part of the subject of this dissertation and, on the other hand, because it is very hard or practically impossible to determine from observations which galaxies are true members

Probability	KDE	KDE Fit	Voronoi$_{AD}$	Voronoi$_{mean}$	Grid	Grid Fit
$\geqslant$50%	2537	2537	741	6093	567	567
$\geqslant$60%	1947	1947	443	4407	382	382
$\geqslant$70%	1452	1452	286	3010	222	222
$\geqslant$80%	1076	1076	214	2057	144	144
$\geqslant$90%	744	744	124	1136	81	81
$\geqslant$100%	294	294	16	-	20	20

Table 4.1: UPMASK Galaxy Detection for the Coma Cluster, for all the applied versions. Here, is represented the number of galaxies, for a certain detection, *i.e.*, above a certain threshold UPMASK membership probability, when the KDE, KDE+Fit, Voronoi+Anderson-Darling Test, Voronoi+Mean Comparsion, Grid, and Grid+Fit were used when studying the objects of the Coma Cluster field.

of a given cluster. Nevertheless a possible extension of the work in this chapter could be to use studies with available galaxy membership (and photometric) information to compute purity and completness functions for Coma with the different versions of UPMASK.

Table 4.1 lists the number of objects that, for a given UPMASK membership probability, were classified as belonging to a cluster, for all the versions used (see the next section).

4.3 Results

Although the galaxy separation performed in section 4.1.1 is a rough separation, it is clear that the adopted cut in magnitude differences (the red line in figure 4.2) is able to successfully reveal the internal structure of the Coma cluster (see figure 4.3). This clearly demonstrates that all UPMASK versions can in fact be applied to detect the potential locations of galaxy clusters. In fact, all versions show evidences for the existence of two peaks in the selected field of view of figure 4.3. To further analyse this feature I constructed color – magnitude diagrams such as those represented in figure. 4.4. The left panel of this figure shows all object in the field before applying the cut in magnitude differences to identify the galaxies. The panel on the right shows a zoom into the panel on the left, but considering only the selected galaxies (i.e. the object obtained after the cut). By observing the left plot of the figure, it seems that there are two types of object distributions. One, that goes along a line with a low slope, are stars, while the vertical distribution of objects are the galaxies. The Star-Galaxy separation performed seems to be able to separate galaxies from stars, but some of the galaxies may be cut due to the separation. It is also possible that at the intersection of the two types of distribution there is some contamination of stars. However, due to the differences in photometric properties, the stars should be eliminated by the UPMASK method, or kept if they belong to a stellar cluster. Even so, stellar clusters can be cleared from the galaxy clusters when analysing their color and magnitude distribution differences. If the galaxies belong to a cluster, their color-magnitude distribution should spread along a vertical (or a high slope) line. This is due to the fact that galaxies from the same cluster are approximately at the same redshitft and therefore suffer the same reddening. The scatter around this line should increase with the increasing magnitude (in this case, r magnitude) because of the higher observational errors at higher magnitudes (bottom of the panels). This effect provides an overall understanding of the distribution of points on the right panel of figure 4.4. As mentioned earlier, all version of UPMASK reveal an overdensity in the center of the fields, that corresponds to the center of the Coma Cluster. As expected by the results in section 3.3, different methods will eliminate different amount of objects. The results

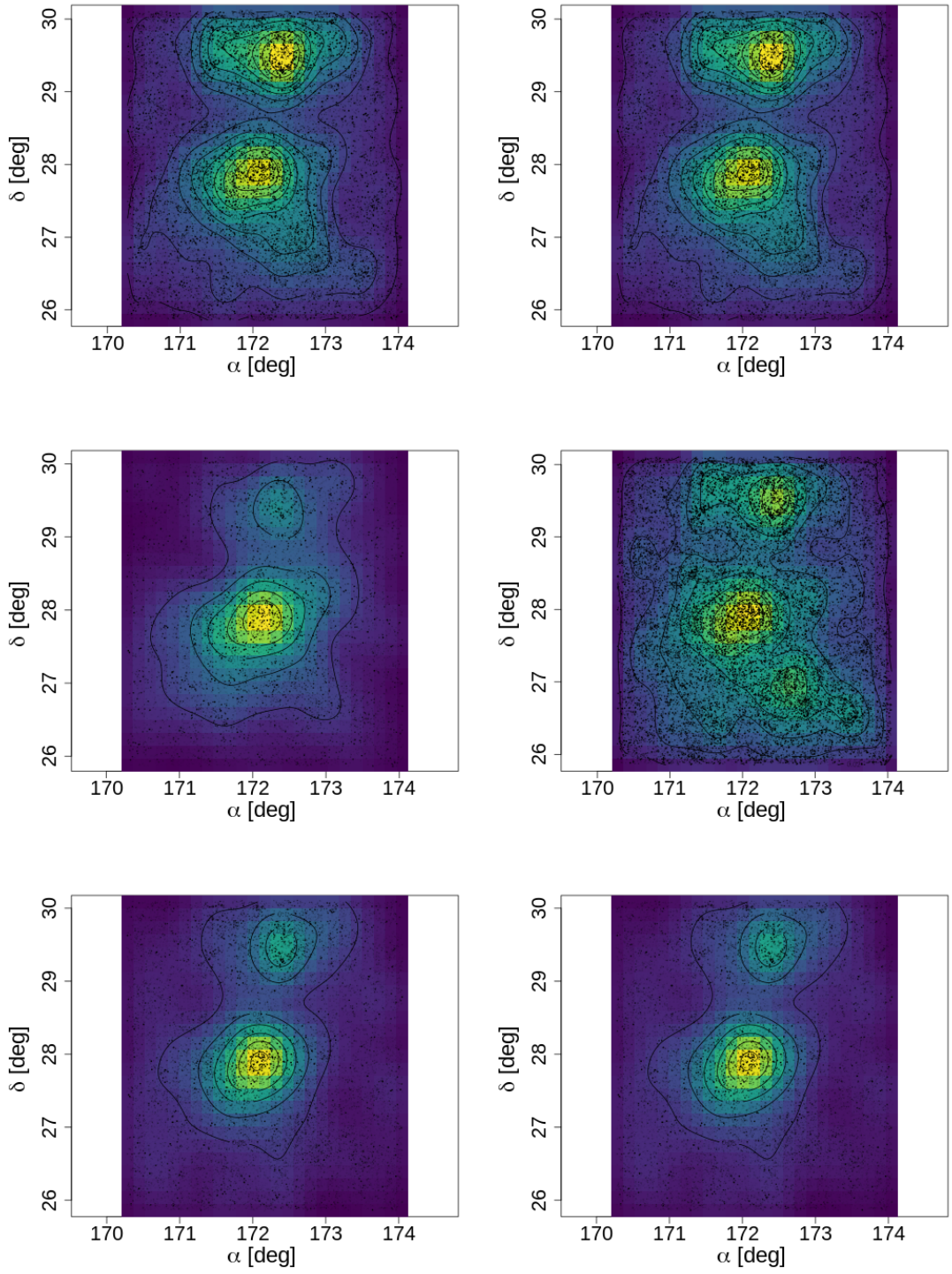


Figure 4.3: Application of UPMASK on Coma. The figures represent KDEs (color map: the brighter, the denser) and iso-contours of the most likely members (Each KDE image was computed for objects above a membership probability of 50%) of the astrometric space (α, δ). The point transparency corresponds to a membership probability that the object belongs to the cluster. Top - Left: results from the original UPMASK version, Right: results from the KDE + fitting function version; Middle - Left: results from the Voronoi + Anderson-Darling test version, Right - results from the Voronoi + comparison of means version; Bottom - Left: results from the Grid version, Right: results from the Grid + fitting function version.

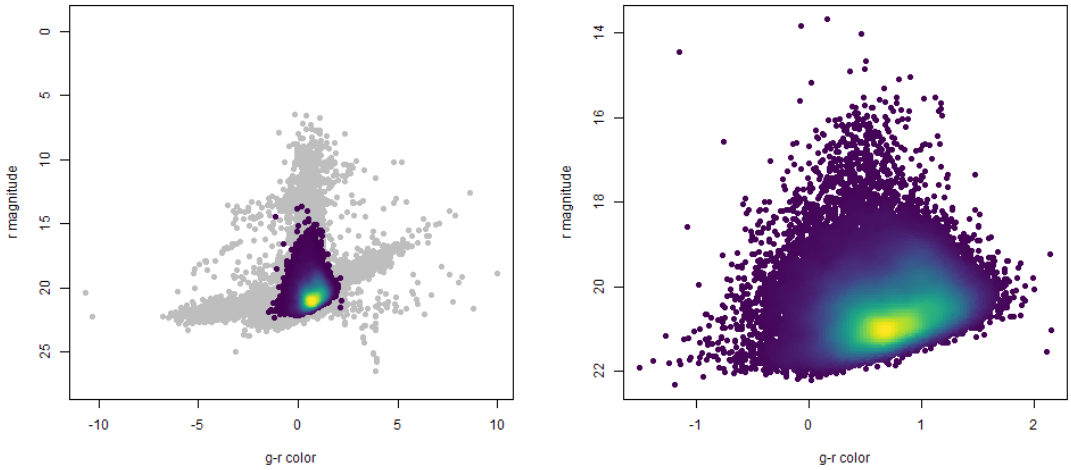


Figure 4.4: Color (g-r) vs Magnitude (r) of the objects of the Coma Cluster field. Left: For all the objects. The gray points are the objects classified as stars and the colored points are the objects classified as galaxies. Right: For the objects classified as galaxies. The colors of the points are based on the local density (the brighter, the denser).

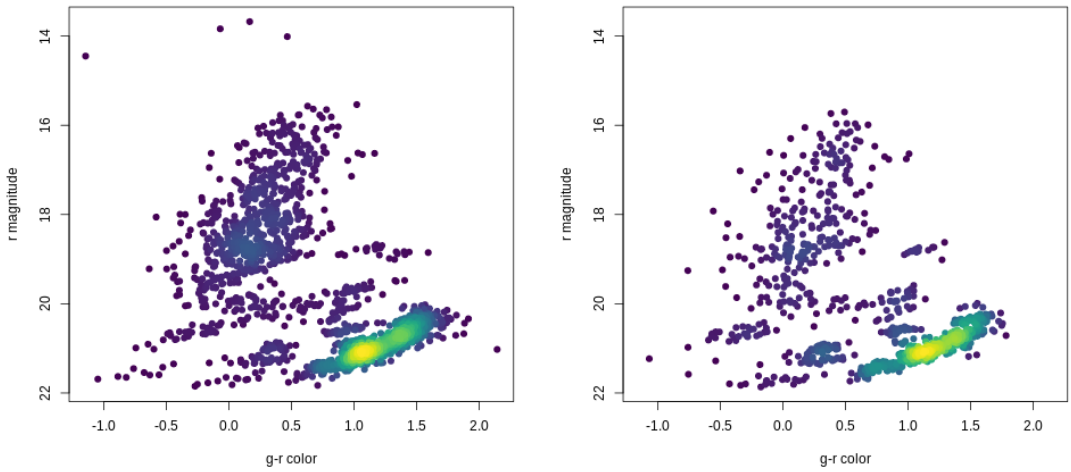


Figure 4.5: Color (g-r) vs Magnitude (r) for the two over-densities of the Coma Cluster field. A cut of $\delta = 28.5°$ was made in order to separate the two over-densities present in figure 4.3. The points are the resulting galaxies from the KDE version of UPMASK, that are above an UPMASK probability of 50%. The colors of the points are based on the local density (the brighter, the denser). In the left plot are the objects in $\delta \geqslant 28.5°$ and in the right plot are the objects in $\delta < 28.5°$.

using the methods implemented with KDE and Grid are very similar when comparing with the version with its correspondent fitting function implemented - in fact, according to table 4.1, they give the same results because the methods use the same seed. The version with the Voronoi and the mean comparison is the one that keeps more objects, and as such there is a higher probability that the results are not as pure. However, since it is also one of the fastest versions, it can be used on a first approach to identify galaxy clusters in photometric surveys. For a more extensive study of a certain cluster, versions that have a more strict comparison criteria should be applied, such as is the case of the Grid or even the Voronoi + Anderson-Darling test version.

It seems that there is another over-density, as already mentioned, at $\alpha \approx 172.5°, \delta \approx 29.5°$ detected in all of the UPMASK versions. In order to compare the two overdensities, it was made a cut in the field of $\delta = 28.5°$, and studied its color and magnitude plots, represented in figure 4.5, where only the points above a membership probability of 50% in the application of the KDE version was used. To discern between the two structures, we plot in figure 4.5 the color-magnitude diagram for the top (right panel) and bottom (left panel) objects separated by the cut in $\delta = 28.5°$. We show this plot only for the original UPMASK version, but we confirmed that all versions show the same typical behaviour of this version. As we can see in the right plot of figure 4.5, the peak in color (the overdensity) seems to be in a slightly redder zone, than the visible peak in left plot of the figure. However, the errors in the r magnitude at values the range of the peaks are typically large and therefore it is not possible to draw a definite conclusion about the difference of reddening of these structures. Nonetheless, when observing this region of the sky with Aladin Sky Atlas [Boch and Fernique, 2014, Bonnarel et al., 2000] (an interactive sky atlas that contains astronomical catalogues and information from the Simbad database [Wenger et al., 2000]) , we confirm that there is an unnamed galaxy cluster in approximately the same region as the top one in figure 4.3, at the bottom one corresponds to the location of the Coma cluster.

Chapter 5

Looking for the lost clusters of the Planck survey

There are many clusters that are detected via the SZ effect that are still today lacking optical counterpart. After we tested UPMASK and my modifications in simulations and in real data, here we search for the optical counterparts of Planck detected clusters. We use the Pan-STARRS catalogue and the Sunyaev-Zel'dovich detections of PLANCKSZ2 [Planck Collaboration et al., 2016]. We are specially interested in clusters that were not confirmed by other surveys. We start by describing the catalogues and the data selection and cross-matching process in sections 5.1 and 5.2. We then apply the UPMASK method to the galaxy data in section 5.3 and present our main findings in section 5.4.

5.1 Planck Clusters

The catalogue PLANCKSZ2 - Planck 2nd Sunyaev-Zel'dovich Source Catalogue [Planck Collaboration et al., 2016] contains detections of galaxy clusters, obtained during 29 months of observations. It has a total of 1653 objects, from which 1203 ($\approx 73\%$) are confirmed clusters. Since the release of the PLANCKSZ2 catalogue, some of the unconfirmed clusters (at the time of the realease of the catalogue) already have optical counter parts. Recent works took images from SDSS, WISE and Pan-STARRS to find optical counterparts [Zohren et al., 2019]. There is also projects to validate the unidentified PSZ2 sources in the nothern sky with the Roque de los Muchachos Observatory [Aguado-Barahona et al., 2019, Streblyanska et al., 2019], to validate the unidentified sources whose area overlap with the SDSS DR12 [Streblyanska et al., 2018], imaging the unidentified sources with the Mayall telescope from the Kitt Peak National Observatory [Boada et al., 2019], taking measures of galaxy clusters from the Russian-Rurkish telescope, the Sayan Observatory, the Calar Alto telescope and the SAO RAS telescope [Zaznobin et al., 2019], and even using non-professional telescopes [Boucher et al., 2018].

The clusters of the PLANCKSZ2 catalogue are represented in a sky projection in figure 5.1. From the PLANCKSZ2, we worked with a catalogue that was constructed by the union of three pipelines adopted to detect galaxy clusters [Planck Collaboration et al., 2016]. In this work, we used information about the sky coordinates (α, δ), position uncertainty, integrated SZ intensity distortion (Y_{5R500}), and the object identifier.

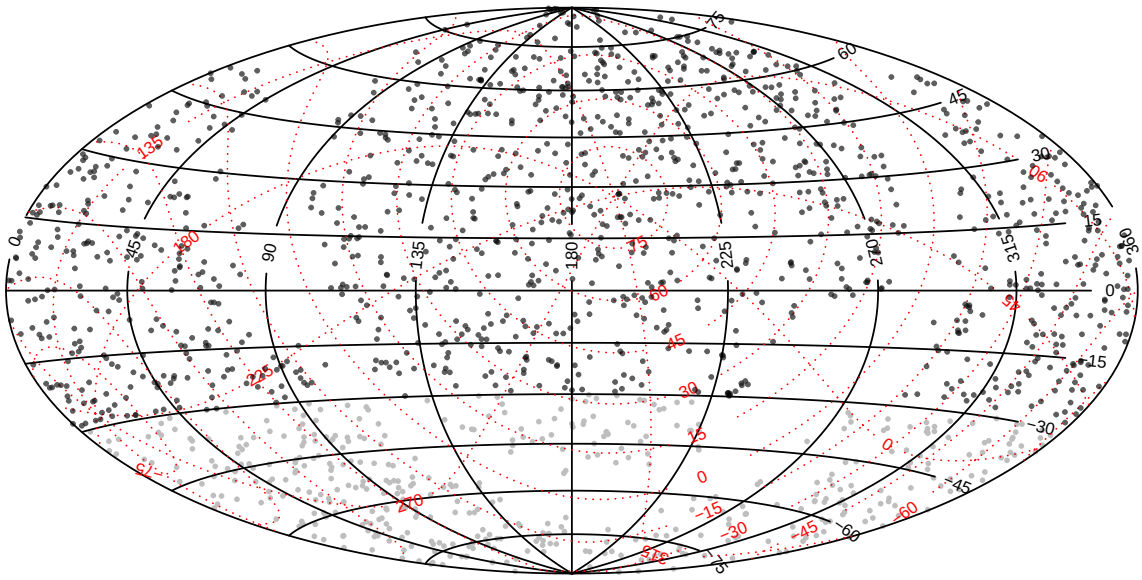


Figure 5.1: Projection of the distribution of Planck Clusters in the sky. The gray points are all galaxy clusters from the PLANCKSZ2 catalogue and the black points are the galaxy clusters from the PLANCKSZ2 catalogue within the Pan-STARRS sky coverage. The red lines are the Galactic coordinate lines.

5.2 Planck Clusters in Pan-STARRS Catalogue

Using the coordinates of each entry in the catalogue of 5.1, I searched for the ones whose positions were covered by second data release of the Pan-STARRS catalogue [Chambers et al., 2016, Flewelling et al., 2016] (section 4.1). The 1212 Planck clusters in the sky coverage of Pan-STARRS are represented in figure 5.1 as black dots, of which 25% (301 clusters) had no external validation (e.g. in the optical or X-rays) that confirmed them as real clusters. These Planck unconfirmed cluster candidates may not correspond to real clusters because the signal detected by Plank may result from unresolved sources with similar spectral signatures or intervening gas clouds along the line-of-sight. The data was obtained through an automatic process, using the Pan-STARRS API `https://catalogs.mast.stsci.edu/api/v0.1/panstarrs/` server, that allows searches through the data parameter space using the coordinates of the galaxy clusters of PLANCKSZ2 and an estimated radius. In this catalogue, it was selected a radius of $3\sigma_{pos}$ in which σ_{pos} is the positional uncertainty of the planck detection. The position uncertainties are typically large for this type of objects, due to the Planck low spatial resolution [Planck Collaboration et al., 2014] (which for the main fraction of the sources is around 1 arcmin, and this position error can reach 5 arcmin) and also because of clusters being such large objects that it is difficult to measure its core without further studies. Therefore, the uncertainty can also be used as a proxy to the radius of a cluster. For every field, a maximum of objects is 50000, since if contained many more, the file would be too heavy and the connection to the server would be lost. From the Pan-STARRS catalogue, the same parameters as in section 4.1 were used, *i.e.*, information about the object identifier, astrometry (α, δ), the number of detection and the number of single epoch detection, mean PSF magnitude, standard deviation of the PSF magnitude, mean Kron magnitude and standard deviation of the Kron magnitude for the bands of the Pan-STARRS survey catalogue (g, r, i, z, y) .

After having the objects that are contained in Pan-STARRS, the star-galaxy separation

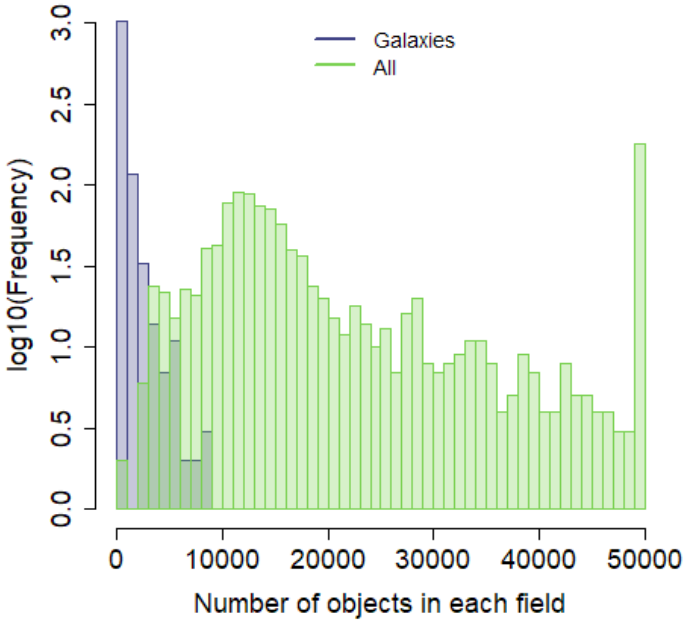


Figure 5.2: Histogram of number of objects in each 1212 Pan-STARRS fields. Here is represented the distribution of objects in each field (including stars, galaxies and other types) (Green) and the distribution of objects that were classified as galaxies, when performing the Star-Galaxy separation (Blue). The vertical axis is in logarithmic scale of the frequency, and each bin is 1000 objects wide.

was performed through the same procedure as the one applied in section 4.1.1, where it is taken advantage of the difference between the magnitudes resulting from a PSF and the Kron magnitudes obtained from a Kron profile. Figure 5.2 shows the distribution resulting objects. In this graph is represented the total number of objects that lie in each field obtained from the Pan-STARRS catalogue, as well as the distribution of galaxies that were classified as such when performing the star-galaxy separation. The distribution has a mean of 21611 and a standard deviation of 14819 objects, as well as a minimum of 161 and a maximum of 50000 objects. This maximum also corresponds to a peak in the histogram since this was the maximum number of objects gathered from the server for a given area. The stars and galaxies are then separated, so the possible colors between the PSF magnitudes of the bands are calculated, together with their uncertainty (through Propagation of Error computation). After this, the data is finally ready to be studied with the UPMASK method.

5.3 Applying UPMASK

For each data set that was selected around each cluster, it was applied the UPMASK method, using the coordinates (α,δ) and the photometric data (mean PSF magnitudes of each Pan-STARRS band and their correspondent uncertainty as well as the colors and their correspondent uncertainty) of each object. It was used 4 PCs, as this is the optimal amount of PCs (section 3.2.1). Some of the fields have very little number of objects after the separation star-galaxy and there is also a lot of individual fields to be studied, and therefore, it was used 10 objects per clusters for all the fields. As we are interested in only detect the Planck clusters, with optical parameters, a resolution of 10% in the UPMASK frequentist probability seems to be a reasonable

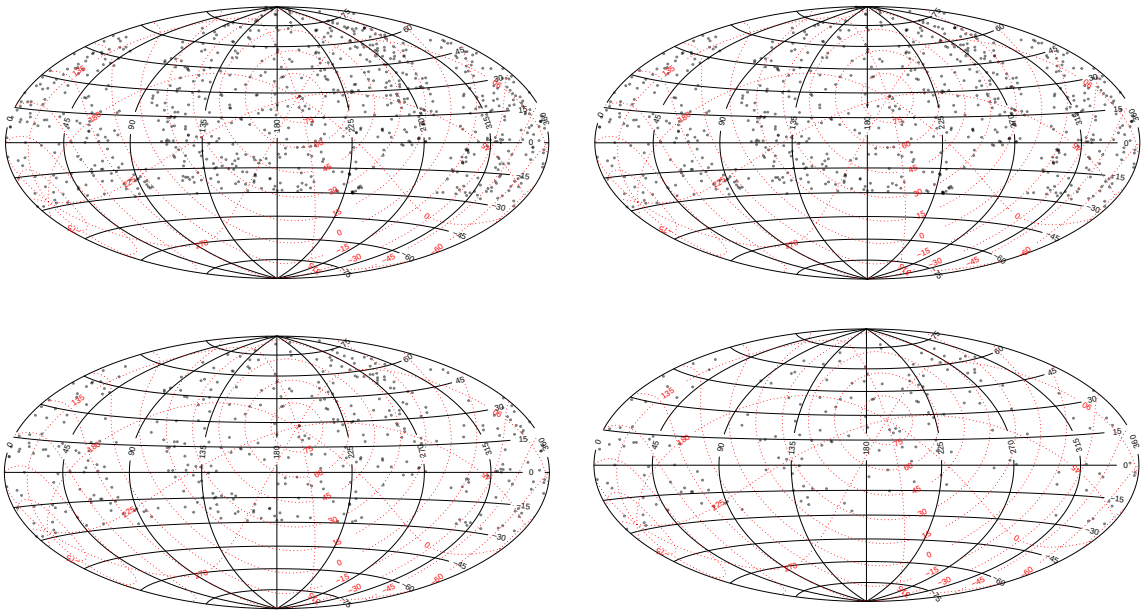


Figure 5.3: Projection of the distribution of Planck Clusters in the sky detected by UPMASK for Clusters that have more than 10 objects with probability above of 50, 70, 90 and 100 % (Left to right, top to bottom). Each point indicates one galaxy cluster from the PLANCKSZ2 catalogue. The red lines are the Galactic coordinate lines.

one. For that, a number of 10 runs was used. Our classification of a cluster (and objects that belong to one) will depend on the probability membership. As such, each field was divided on how many objects are above a certain probability. This is represented in figure 5.3, where it is plotted the projected distribution of clusters that have more than 10 objects above 50%, 70%, 90% and 100%. In table 5.1 is a more detailed division, where the objects are grouped in class of number of objects above a certain probability.

It is also particularly interesting to look just for the galaxy clusters that were captured in the PLANCKSZ2 survey catalogue that were not confirmed by an external counterpart (at the moment of the release of the PLANCKSZ2 catalogue). The result of UPMASK for originally unconfirmed SZ clusters is shown in table 5.2.

The first plot of figure 5.4 shows the dependency of the measured SZ flux with the redshift (for the sources that have available redshift). This is a visual representation of the selection function of Planck. This figure shows three distributions, all PSZ2 sources (grey dots), the PSZ2 sources that fall in the Pan-STARR sky-coverage (blue open circles), and lastly, the UPMASK detections (blue dots). These latter data points were defined as the sources that have more than 10 objects above 50% of UPMASK membership probability. With this plot, one can conclude that the detections do not seem to depend on the redshift (rather, the dependency is only present because of the Planck selection function). The second plot shows the counts and redshift for the three distributions described previously. The third plot shows the ratio of the detected objects over the total number of PSZ2 sources in the Pan-STARRS sky coverage. The ratio seems to have a smooth behavior until redshift 0.5. Above this redshift, there is not many objects, and therefore the ratio fluctuates. However, a follow up is needed to find and compute a redshift of the detected clusters in the Pan-STARRS and to compare with the redshift of the PSZ2 sources.

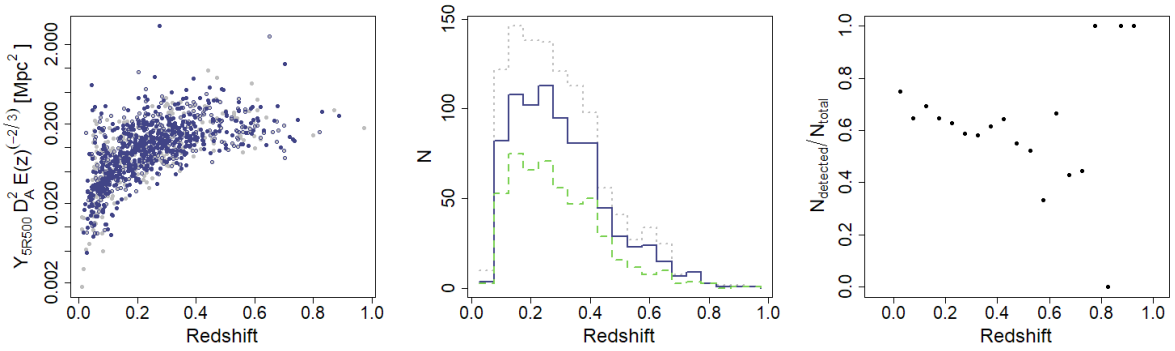


Figure 5.4: Variation of PSZ2 sources properties with redshift. Left - Dependency of the measured Planck SZ flux with redshift. The grey points represent all the PSZ2 sources, the blue open one represent the PSZ2 sources that are included in the Pan-STARRS sky coverage and the blue cloud represent the sources that UPMASK detected. (Here, we stabilished a detection criteria as the fields that contain more than 10 members above 50% membership probability. This detection criteria is the same in all panels of this figure.) Center - Variation of PS2 source counts with redshift, for all sources (dotted grey), for sources included in the Pan-STARRS sky coverage (solid blue) and for the UPMASK detections (dashed green). Right - Fraction of the detected sources as a function redshift.

Probability	0	]0,10]	]10,50]	]50,100]	]100,150]	]150,200]	]200,∞[
P(⩾10)	278	9	179	175	97	83	391
P(⩾20)	314	6	211	193	114	86	288
P(⩾30)	332	11	265	189	151	121	143
P(⩾40)	356	12	302	240	213	62	27
P(⩾50)	378	30	336	367	87	11	3
P(⩾70)	456	59	584	111	1	1	-
P(⩾90)	605	209	293	5	-	-	-
P(=100)	787	243	181	1	-	-	-

Table 5.1: UPMASK results of Planck Clusters in Pan-STARRS catalogue. The table entries are the number of clusters that have a certain amount of objects above a certain membership probability.

Probability	0	]0,10]	]10,50]	]50,100]	]100,150]	]150,200]	]200,∞]
P(⩾10)	42	1	35	31	22	16	154
P(⩾20)	48	1	41	41	21	21	128
P(⩾30)	53	1	51	39	44	43	70
P(⩾40)	58	2	56	62	79	25	19
P(⩾50)	61	11	70	111	39	6	3
P(⩾70)	77	12	174	36	1	1	-
P(⩾90)	112	78	110	1	-	-	-
P(=100)	174	81	46	-	-	-	-

Table 5.2: UPMASK results of originally unconfirmed SZ Planck Clusters in Pan-STARRS catalogue. The table entries are the number of clusters that have a certain amount of objects above a certain membership probability.

	Streblyanska et al. False	Streblyanska et al. True
UPMASK False	182 (40%)	39 (9%)
UPMASK True	81 (18%)	148 (33%)

Table 5.3: Comparison table between the results of [Aguado-Barahona et al., 2019, Streblyanska et al., 2019] and UPMASK. The "True" and "False" fields are the unconfirmed PSZ2 sources that were found or not by the respective methods. The results are normalized to the total number of unconfirmed sources in the Pan-STARRS sky coverage.

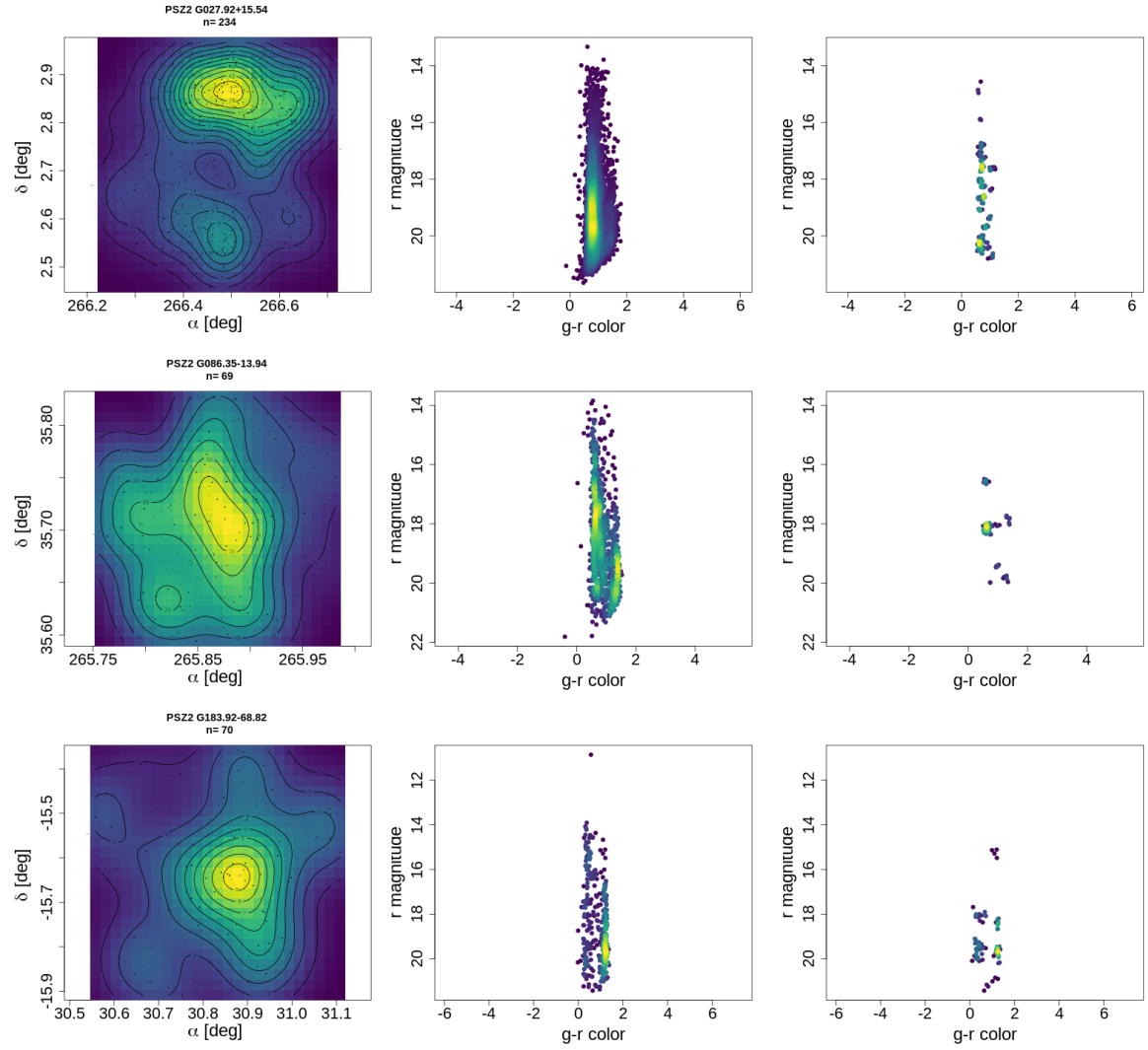


Figure 5.5: Examples of typical UPMASK detections in Pan-STARRS fields around PSZ2 sources. Rows show results for the cluster fields PSZ2 G027.92+15.54, PSZ2 G086.35-13.94 and PSZ2 G183.92-68.82, respectively. Left - KDE image representation with iso-contours (color map: the brighter, the denser; see text for details) obtained from the distribution of galaxies detected by UPMASK for the respective cluster field. Center - Color(g-r) *versus* Magnitude (r) for the galaxies contained in this field. Right - Color(g-r) Magnitude (r) for the galaxies with probability above 50% contained in this field. The colors of the points of the last two graphs are based on the local density (the brighter, the denser).

5.4 Results

If we classify a cluster by saying that all the bounded objects should have a UPMASK membership probability higher than 50% and that a galaxy cluster should have at least 50 members, by taking into account the results from table 5.1, we can conclude that UPMASK was able to discover optical counterparts for around 40% of the 1212 galaxy clusters of the PLANCKSZ2 catalogue (inside the Pan-STARRS sky coverage). This percentage rises to about 62% if the membership probability is chosen to be 10% (while keeping the same number of bound objects). Although a 40% percentage may be considered a relatively low value, one needs to keep in mind that we are trying to detect the galaxy clusters in the optical and infrared wavelengths. These wavelengths carry no SZ effect information about the clusters. On the other hand, the method makes no assumption about the galaxy cluster model (as it is usually assumed in other cluster identification methods, and simply separates objects that have similar photometric properties that appear clustered in the sky.

Following the same criteria to classify a cluster, when looking at table 5.2, we were able to rediscover about 52% (159 out of 301) of the galaxy clusters that had no external validation, *i.e.*, clusters that were also identified in other observations (like for example SDSS) - at the time the catalogue was published. However, the interesting objects are those that have a high membership percentage. For example, those that have more than 10 objects with probability equal to 90%, as they give good candidates for galaxy clusters. We have compared our results with other works in the literature. Table 5.3 shows the comparison of our results with the ones obtained in [Aguado-Barahona et al., 2019, Streblyanska et al., 2019], whose catalogue was kindly provided by the authors. Here, we compare both results (Streblyanska et al. and UPMASK detections) with all the PSZ2 sources without an optical counterpart (at the time of release of the PLANCKSZ2 catalogue). The authors report that there is some spurious detections (ie, no obvious cluster observed in the photometric data) and therefore we have not included these detections in the table. The "False" and "True" fields refer to the fields that were not found in the list of PSZ2 sources without an optical counterpart and those that are a detection, respectively. Overall, UPMASK is 73% in accordance with the results in the aforementioned articles. Our method shows a tendency to return more positive results (18%, which corresponds to 81 sources) when comparing with the "False" results of the paper than negative (9%) when comparing with the "True" results of the paper. This "optimistic" behaviour of UPMASK may be easily understood given the nature of the method that makes no assumptions about the cluster model nor uses direct redshift information about the galaxies. In these conditions more fortuitous detections may be returned by the method. This is not necessarily a drawback given that all sources detected by UPMASK would need to be confirmed, anyway, to discard false positive results.

Figure 5.5 shows information for three detected clusters, each in one row. The panels on the left are KDE images with iso-contours obtained from the spatial distribution of galaxies that have membership probabilities higher than 50%. The panels on the right show the corresponding color *versus* magnitude diagrams for the same probability. The middle panels show the color-magnitude diagrams for all galaxies contained in the field. This is not the usual way to display color–magnitude diagrams for galaxies, however since that axis system was the one applied to the Star-Galaxy separation in chapter 4, we decided to maintain the same axis representation. The galaxies before the cut in probability present a distribution that corresponds to a red sequence. The ones after the cut also follow the same tendency. The figures in appendix B show this

same type of plots for all 46 (unconfirmed) clusters that have between 10 and 50 objects with a membership probability equal to 100%. The majority of the cases in this appendix also present a well defined distribution of a red sequence. All cluster images in this appendix include an estimated radius and a center for each cluster. The former is represented by a white circle. These two quantities were estimated using an R function, Mclust [Scrucca et al., 2016], that is a clustering and/or classification tool, based on Gaussian mixture models. We have forced the tool to search each field, with two mixture components. This tool gives, for each mixture component, the mean of the points that belong to them, as well as a variance that describes a centroid. With this, we have selected the centroid that is closer to the peak of the KDE for the field in question, and took the maximum variance as the radius of the cluster, and the centroid position as the center of the cluster. This way of estimating a radius seem to work better with clusters having an approximately circular shape, or just one over-density peak (as one can see in the colored images obtained from the KDE method).

Chapter 6

Conclusion

A complete physical characterization and study of galaxy clusters requires their detection and the identification of member galaxies. In this dissertation a previously existing method designed to perform detection and membership of stars in stellar clusters, UPMASK, was modified and then applied to galaxy clusters for the first time. We show here that the method is also effective to be used tho detect this type of objects, using only photometric and astrometric data. Moreover, our study indicates that the performance of the method seems to be redshift independent.

Chapter 2 describes the modifications proposed in this dissertation: KDE+Fitting Function; Voronoi+Anderson-Darling test; Voronoi+Mean Comparison; Grid; Grid+Fitting Function. The method and my modifications were validated in chapter 3, by applying them to simulated data of the MICE galaxy mock catalogue. It was seen that the result of this method depends on the filter system of the selected photometric dataset. In this same chapter, it was also studied how the results behave by varying parameters, such as the number of objects per K-means group, the number of Principal Components, and a value for a threshold level. The modification that most improved the CPU running time of the method is the Voronoi+Mean Comparison, a result that holds with respect to all the other versions and also the original one. On the other hand, the modification that provided the highest improvement in purity is the modification that has a Grid implementation. One of the first steps of the UPMASK method is the dimensionality reduction, by using PCA, which in the standard version of the method, the number of principal components is fixed to 4. Therefore, we have studied the optimal number of principal components to be used in the case of galaxies. From the tests performed here we selected four principal components, as additional components were shown to result in similar outcomes, with the penalty of increasing the CPU running time. We have also seen that, using more objects inside each group in the clustering step (which uses the K-means algorithm), allows us to attain a higher completeness, while the purity decreases. Using a smaller threshold level, the completeness decreases, but the purity seems to increase. The modifications and the parameters of the UPMASK method introduced here transform it into a very versatile tool, as the users can adjust the method to their goals: either using this method to find galaxy cluster candidates or, to obtain membership probabilities of clustered galaxies.

One of the main motivations behind this dissertation is a future application to the Euclid space survey. This survey will use observed galaxy cluster properties to detect signatures of the expansion rate of the Universe and the growth rate of cosmic structures. This will also allow for additional constrains on the cosmological parameters that will be derived on the Euclid mission. But firstly, these galaxy clusters need to be detected in order to study their properties which

is something UPMASK can contribute to. It can be used to select regions of interest to look for galaxy clusters, or to search for clusters candidates. However, this is a challenging task, since it is estimated that the Euclid survey will identify more than 60 thousand clusters with a signal-to-noise better than 3, in a redshift range from 0.2 to 2 [Laureijs et al., 2011]. A method to do so could be built by using or adapting UPMASK, to search in different regions of the sky for overdensities of clustered galaxies. This could be a natural extension of this dissertation work.

In chapter 4 UPMASK is applied to real photometric and astrometric data in the direction of the Coma Cluster in the Pan-STARRS catalogue, that contains both stars, galaxies, and other types of astrophysical objects. We separated the galaxies from stars, following the criteria in [Farrow et al., 2014], in which we select all the objects that are above the $PSF_{mag} - Kron_{mag} = 0.05$ line in all of the magnitudes. As expected, stars and galaxies live in different regions of the color-magnitude space, with some overlap. All the versions of UPMASK were then applied to the objects of the field of the Coma Cluster, using the parameters described in section 4.2. The differences found in the results are compatible with the study of the modifications done in chapter 3, with the Voronoi+Mean Comparison classifying more galaxies as members of a cluster at higher probabilities, and thus its results are expected to have a higher completeness, and with the Grid classifying less galaxies as members of a cluster at higher probabilities, the corresponding results are expected to have a higher purity. We have seen, using the methods developed here, that there seem to exist another galaxy cluster in the analysed field - promptly seen in the color density of the panels in section 4.2.

Then, in chapter 5 we look for optical counterparts of Planck SZ clusters. We use data from the Pan-STARRS catalogue and start from a list of candidates from the Planck 2nd Sunyaev-Zel'dovich cluster catalogue. We selected data inside a certain area determined using results from the PLANCKSZ2 catalogue (namely the sky coordinates and the position uncertainty). Similarly to chapter 4, we also performed a Star-Galaxy separation to all the extracted fields, and then the original UPMASK was applied to the fields. The results were analysed according to the number of galaxies with a certain membership probability, as shown in tables 5.1 and 5.2, that highlights the fields of galaxy clusters from the PLANCKSZ2 catalogue that had no previous external validation from a detection in optical wavelengths until this dissertation. From the panels in figure 5.4 we can conclude that UPMASK is able to detect galaxies independently of redshift. We have seen in this dissertation that the method was able to discover optical counterparts for about 40% of the SZ detections. For the fields in table 5.2, we selected those that have more than 10 members with 100% membership probability and estimate a cluster radius using a simple procedure that can be still improved in future works. In the future it can also be interesting to apply the methods developed in this dissertation to search for the new clusters directly from deeper optical surveys as the Large Synoptic Survey Telescope and the Euclid space mission. For the clusters detected in this work, we plan to perform dedicated spectroscopic observation to determine their redshifts and to confirm unambiguously their physical existence.

Finally, thanks to the methods we developed in this dissertation and to the adoption of the most modern optical surveys, we performed the first detection of the optical counterparts to 81 galaxy clusters previously identified by the ESA Planck space mission. This builds up significantly the evidence for the existence of these objects laying at the backbones of the large scale structure of the Universe.

Bibliography

Abbott, T.M.C., Abdalla, F.B., Allam, S. et al (2018). The Dark Energy Survey: Data Release 1. *Astrophysical Journal, Supplement*, 239(2):18.

Abell, G.O. (1958). The Distribution of Rich Clusters of Galaxies. *Astrophysical Journal, Supplement*, 3:211.

Abell, G.O. (1965). Clustering of Galaxies. *Annual Review of Astronomy and Astrophysics*, 3:1.

Aguado-Barahona, A., Barrena, R., Streblyanska, A. et al (2019). Optical validation and characterization of Planck PSZ2 sources at the Canary Islands observatories. II. Second year of LP15 observations. *arXiv e-prints*, page arXiv:1909.06235.

Airy, G.B. (1835). On the Diffraction of an Object-glass with Circular Aperture. *Transactions of the Cambridge Philosophical Society*, 5:283.

Allen, S.W., Evrard, A.E. and Mantz, A.B. (2011). Cosmological Parameters from Observations of Galaxy Clusters. *Annual Review of Astronomy and Astrophysics*, 49:409–470.

Anderson, T.W. and Darling, D.A. (1952). Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes. *Ann. Math. Statist.*, 23(2):193–212.

Avelino, P., Barreiro, T., Carvalho, C.S. et al (2016). Unveiling the Dynamics of the Universe. *arXiv e-prints*, page arXiv:1607.02979.

Bahcall, N.A. (1988). Large-scale structure in the universe indicated by galaxy clusters. *Annual Review of Astronomy and Astrophysics*, 26:631–686.

Baron, D. (2019). Machine Learning in Astronomy: a practical overview. *arXiv e-prints*, page arXiv:1904.07248.

Bautz, L.P. and Morgan, W.W. (1970). On the Classification of the Forms of Clusters of Galaxies. *The Astrophysical journal*, 162:L149.

Bellagamba, F., Roncarelli, M., Maturi, M. et al (2018). AMICO: optimized detection of galaxy clusters in photometric surveys. *Monthly Notices of the RAS*, 473(4):5221–5236.

Birkinshaw, M. (1999). The Sunyaev-Zel'dovich effect. *Physics Reports*, 310(2-3):97–195.

Blanton, M.R., Bershady, M.A., Abolfathi, B. et al (2017). Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. *The Astronomical Journal*, 154(1):28.

Boada, S., Hughes, J.P., Menanteau, F. et al (2019). High Confidence Optical Confirmations among the High Signal-to-noise Planck Cluster Candidates. *The Astrophysical journal*, 871(2):188.

Boch, T. and Fernique, P. (2014). Aladin Lite: Embed your Sky in the Browser. In Manset, N. and Forshay, P., editors, *Astronomical Data Analysis Software and Systems XXIII*, volume 485 of *Astronomical Society of the Pacific Conference Series*, page 277.

Bonnarel, F., Fernique, P., Bienaymé, O. et al (2000). The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources. *Astronomy and Astrophysics, Supplement*, 143:33–40.

Borgani, S. (2008). *Cosmology with Clusters of Galaxies*, volume 740, page 24.

Boucher, V., de Visscher, S. and Ringeval, C. (2018). Optical Follow-up of Planck Cluster Candidates with Small Instruments. *Publications of the ASP*, 130(992):104001.

Bower, R.G. and Balogh, M.L. (2004). The Difference Between Clusters and Groups: A Journey from Cluster Cores to Their Outskirts and Beyond. In Mulchaey, J.S., Dressler, A. and Oemler, A., editors, *Clusters of Galaxies: Probes of Cosmological Structure and Galaxy Evolution*, page 325.

Boylan-Kolchin, M., Springel, V., White, S.D.M. et al (2009). Resolving cosmic structure formation with the Millennium-II Simulation. *Monthly Notices of the RAS*, 398:1150–1164.

Carretero, J., Tallada, P., Casals, J. et al (2017). CosmoHub and SciPIC: Massive cosmological data analysis, distribution and generation using a Big Data platform. In *Proceedings of the European Physical Society Conference on High Energy Physics. 5-12 July*, page 488.

Celebi, M.E. and Aydin, K. (2016). *Unsupervised Learning Algorithms*. Springer Publishing Company, Incorporated, 1st edition.

Chambers, K.C., Magnier, E.A., Metcalfe, N. et al (2016). The Pan-STARRS1 Surveys. *arXiv e-prints*, page arXiv:1612.05560.

Crocce, M., Castander, F.J., Gaztañaga, E. et al (2015). The MICE Grand Challenge lightcone simulation - II. Halo and galaxy catalogues. *Monthly Notices of the RAS*, 453:1513–1530.

Cropper, M., Pottinger, S., Niemi, S. et al (2016). VIS: the visible imager for Euclid. In *Space Telescopes and Instrumentation 2016: Optical, Infrared, and Millimeter Wave*, volume 9904 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 99040Q.

Dawson, K.S., Kneib, J.P., Percival, W.J. et al (2016). The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data. *The Astronomical Journal*, 151(2):44.

Dietrich, J.P., Zhang, Y., Song, J. et al (2014). Orientation bias of optically selected galaxy clusters and its impact on stacked weak-lensing analyses. *Monthly Notices of the RAS*, 443(2):1713–1722.

Dressler, A. (1980). Galaxy morphology in rich clusters: implications for the formation and evolution of galaxies. *The Astrophysical journal*, 236:351–365.

Euclid Collaboration, Adam, R., Vannier, M. et al (2019). Euclid preparation. III. Galaxy cluster detection in the wide photometric survey, performance and algorithm selection. *Astronomy and Astrophysics*, 627:A23.

Farrens, S., Abdalla, F.B., Cypriano, E.S. et al (2011). Friends-of-friends groups and clusters in the 2SLAQ catalogue. *Monthly Notices of the RAS*, 417(2):1402–1416.

Farrow, D.J., Cole, S., Metcalfe, N. et al (2014). Pan-STARRS1: Galaxy clustering in the Small Area Survey 2. *Monthly Notices of the RAS*, 437(1):748–770.

Flewelling, H.A., Magnier, E.A., Chambers, K.C. et al (2016). The Pan-STARRS1 Database and Data Products. *arXiv e-prints*, page arXiv:1612.05243.

Forgy, E. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–780.

Fosalba, P., Crocce, M., Gaztañaga, E. et al (2015a). The MICE grand challenge lightcone simulation - I. Dark matter clustering. *Monthly Notices of the RAS*, 448:2987–3000.

Fosalba, P., Gaztañaga, E., Castander, F.J. et al (2015b). The MICE Grand Challenge light-cone simulation - III. Galaxy lensing mocks from all-sky lensing maps. *Monthly Notices of the RAS*, 447:1319–1332.

Fossati, M., Gavazzi, G., Savorgnan, G. et al (2013). Hα3: an Hα imaging survey of HI selected galaxies from ALFALFA. IV. Structure of galaxies in the Local and Coma superclusters. *Astronomy and Astrophysics*, 553:A91.

Gunn, J.E., Siegmund, W.A., Mannery, E.J. et al (2006). The 2.5 m Telescope of the Sloan Digital Sky Survey. *The Astronomical Journal*, 131(4):2332–2359.

Hammer, D., Hornschemeier, A.E., Mobasher, B. et al (2010). Deep GALEX Observations of the Coma Cluster: Source Catalog and Galaxy Counts. *Astrophysical Journal, Supplement*, 190(1):43–57.

Honscheid, K., DePoy, D.L. and for the DES Collaboration (2008). The Dark Energy Camera (DECam). *arXiv e-prints*, page arXiv:0810.3600.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, 24.

Inserra, C., Nichol, R. C., Scovacricchi, D. et al (2018). Euclid: Superluminous supernovae in the deep survey. *A&A*, 609:A83.

Jansen, F., Lumb, D., Altieri, B. et al (2001). XMM-Newton observatory. I. The spacecraft and operations. *Astronomy and Astrophysics*, 365:L1–L6.

Kotsiantis, S.B. (2007). Supervised machine learning: A review of classification techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Kron, R.G. (1980). Photometry of a complete sample of faint galaxies. *Astrophysical Journal, Supplement*, 43:305–325.

Krone-Martins, A. and Moitinho, A. (2014). UPMASK: unsupervised photometric membership assignment in stellar clusters. *Astronomy and Astrophysics*, 561:A57.

Krone-Martins, A. and Moitinho, A. (2015). UPMASK: Unsupervised Photometric Membership Assignment in Stellar Clusters.

Laureijs, R., Amiaux, J., Arduini, S. et al (2011). Euclid Definition Study Report. *arXiv e-prints*.

Licitra, R., Mei, S., Raichoor, A. et al (2016a). The RedGOLD cluster detection algorithm and its cluster candidate catalogue for the CFHT-LS W1. *Monthly Notices of the RAS*, 455(3):3020–3041.

Licitra, R., Mei, S., Raichoor, A. et al (2016b). The Next Generation Virgo Cluster Survey. XX. RedGOLD Background Galaxy Cluster Detections. *The Astrophysical journal*, 829(1):44.

Maciaszek, T., Ealet, A., Jahnke, K. et al (2016). Euclid Near Infrared Spectrometer and Photometer instrument concept and first test results obtained for different breadboards models at the end of phase C. In *Space Telescopes and Instrumentation 2016: Optical, Infrared, and Millimeter Wave*, volume 9904 of *Proceedings of the SPIE*, page 99040T.

Mahdavi, A. and Geller, M.J. (2001). The L_X-σ Relation for Galaxies and Clusters of Galaxies. *Astrophysical Journal, Letters*, 554(2):L129–L132.

Mamon, G. (1996). The Dynamics of Groups and Clusters of Galaxies and Links to Cosmology. In de Vega, H.J. and Sánchez, N., editors, *Third Paris Cosmology Colloquium*, page 95.

Nunes, N.J., da Silva, A.C. and Aghanim, N. (2006). Number counts in homogeneous and inhomogeneous dark energy models. *Astronomy and Astrophysics*, 450(3):899–907.

Oemler, Augustus, J. (1974). The Systematic Properties of Clusters of Galaxies. Photometry of 15 Clusters. *The Astrophysical journal*, 194:1–20.

Omer, Guy C., J., Page, T.L. and Wilson, A.G. (1965). The coma cluster of galaxies. 1. Size and structure. *The Astronomical Journal*, 70:440.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33(3):1065–1076.

Peterson, J.R. and Fabian, A.C. (2006). X-ray spectroscopy of cooling clusters. *Physics Reports*, 427(1):1–39.

Planck Collaboration (2005). The Scientific Programme of Planck. *ESA publication ESA-SCI(2005)/01*.

Planck Collaboration, Ade, P.A.R., Aghanim, N. et al (2014). Planck 2013 results. XXIX. The Planck catalogue of Sunyaev-Zeldovich sources. *Astronomy and Astrophysics*, 571:A29.

Planck Collaboration, Ade, P.A.R., Aghanim, N. et al (2016). Planck 2015 results. XXVII. The second Planck catalogue of Sunyaev-Zeldovich sources. *Astronomy and Astrophysics*, 594:A27.

Press, W.H. and Schechter, P. (1974). Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *The Astrophysical journal*, 187:425–438.

Roos, M. (2012). Astrophysical and Cosmological Probes of Dark Matter. *Journal of Modern Physics*, 3:1152–1171.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27(3):832–837.

Rykoff, E.S., Rozo, E., Busha, M.T. et al (2014). redMaPPer. I. Algorithm and SDSS DR8 Catalog. *The Astrophysical journal*, 785(2):104.

Sánchez Almeida, J., Aguerri, J.A.L., Muñoz-Tuñón, C. et al (2010). Automatic Unsupervised Classification of All Sloan Digital Sky Survey Data Release 7 Galaxy Spectra. *The Astrophysical journal*, 714(1):487–504.

Sánchez Almeida, J. and Allende Prieto, C. (2013). Automated Unsupervised Classification of the Sloan Digital Sky Survey Stellar Spectra using k-means Clustering. *The Astrophysical journal*, 763(1):50.

Sarron, F., Martinet, N., Durret, F. et al (2018). Evolution of the cluster optical galaxy luminosity function in the CFHTLS: breaking the degeneracy between mass and redshift. *Astronomy and Astrophysics*, 613:A67.

Scholz, F.W. and Stephens, M.A. (1987). K-sample anderson-darling tests. *Journal of the American Statistical Association*, 82(399):918–924.

Scrucca, L., Fop, M., Murphy, T.B. et al (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233.

Sérsic, J.L. (1963). Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Boletin de la Asociacion Argentina de Astronomia La Plata Argentina*, 6:41.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Smith, S. (1936). The Mass of the Virgo Cluster. *The Astrophysical journal*, 83:23.

Springel, V., Frenk, C.S. and White, S.D.M. (2006). The large-scale structure of the Universe. *Nature*, 440:1137.

Streblyanska, A., Aguado-Barahona, A., Ferragamo, A. et al (2019). Optical validation and characterization of Planck PSZ2 sources at the Canary Islands observatories. I. First year of LP15 observations. *Astronomy and Astrophysics*, 628:A13.

Streblyanska, A., Barrena, R., Rubiño-Martín, J.A. et al (2018). Characterization of a subsample of the Planck SZ source cluster catalogues using optical SDSS DR12 data. *Astronomy and Astrophysics*, 617:A71.

Sunyaev, R.A. and Zeldovich, Y.B. (1970). Small-Scale Fluctuations of Relic Radiation. *Astrophysics and Space Science*, 7(1):3–19.

Sutherland, R.S. and Dopita, M.A. (1993). Cooling Functions for Low-Density Astrophysical Plasmas. *Astrophysical Journal, Supplement*, 88:253.

Voges, W., Aschenbach, B., Boller, T. et al (1999). The ROSAT all-sky survey bright source catalogue. *Astronomy and Astrophysics*, 349:389–405.

Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les paralléllοèdres primitifs. *Journal für die reine und angewandte Mathematik*, 134:198–287.

Wenger, M., Ochsenbein, F., Egret, D. et al (2000). The SIMBAD astronomical database. The CDS reference database for astronomical objects. *Astronomy and Astrophysics, Supplement*, 143:9–22.

Yagi, M., Koda, J., Komiyama, Y. et al (2016). Catalog of Ultra-diffuse Galaxies in the Coma Clusters from Subaru Imaging Data. *Astrophysical Journal, Supplement*, 225(1):11.

York, D.G., Adelman, J., Anderson, John E., J. et al (2000). The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120(3):1579–1587.

Zaznobin, I.A., Burenin, R.A., Bikmaev, I.F. et al (2019). Optical Identifications of Galaxy Clusters Among Objects from the Second Planck Catalogue of Sunyaev-Zeldovich Sources. *Astronomy Letters*, 45(2):49–61.

Zhang, J.n., Wu, F.c., Luo, A.l. et al (2006). A non-parametric method of estimating the physical parameters of stellar atmosphere. *Chinese Astronomy and Astrophysics*, 30(2):176–186.

Zohren, H., Schrabback, T., van der Burg, R.F.J. et al (2019). Optical follow-up study of 32 high-redshift galaxy cluster candidates from Planck with the William Herschel Telescope. *Monthly Notices of the RAS*, 488(2):2523–2542.

Zwicky, F. (1933). Die rotverschiebung von extragalaktischen nebeln. *Helvetica Physica Acta*, 6:110–127.

Zwicky, F. (1951). The Coma Cluster of Galaxies. *Publications of the ASP*, 63(371):61.

Zwicky, F., Herzog, E., Wild, P. et al (1961). *Catalogue of galaxies and of clusters of galaxies, Vol. I.*

Appendix A

Completeness and Purity Curves

A.1 KDE

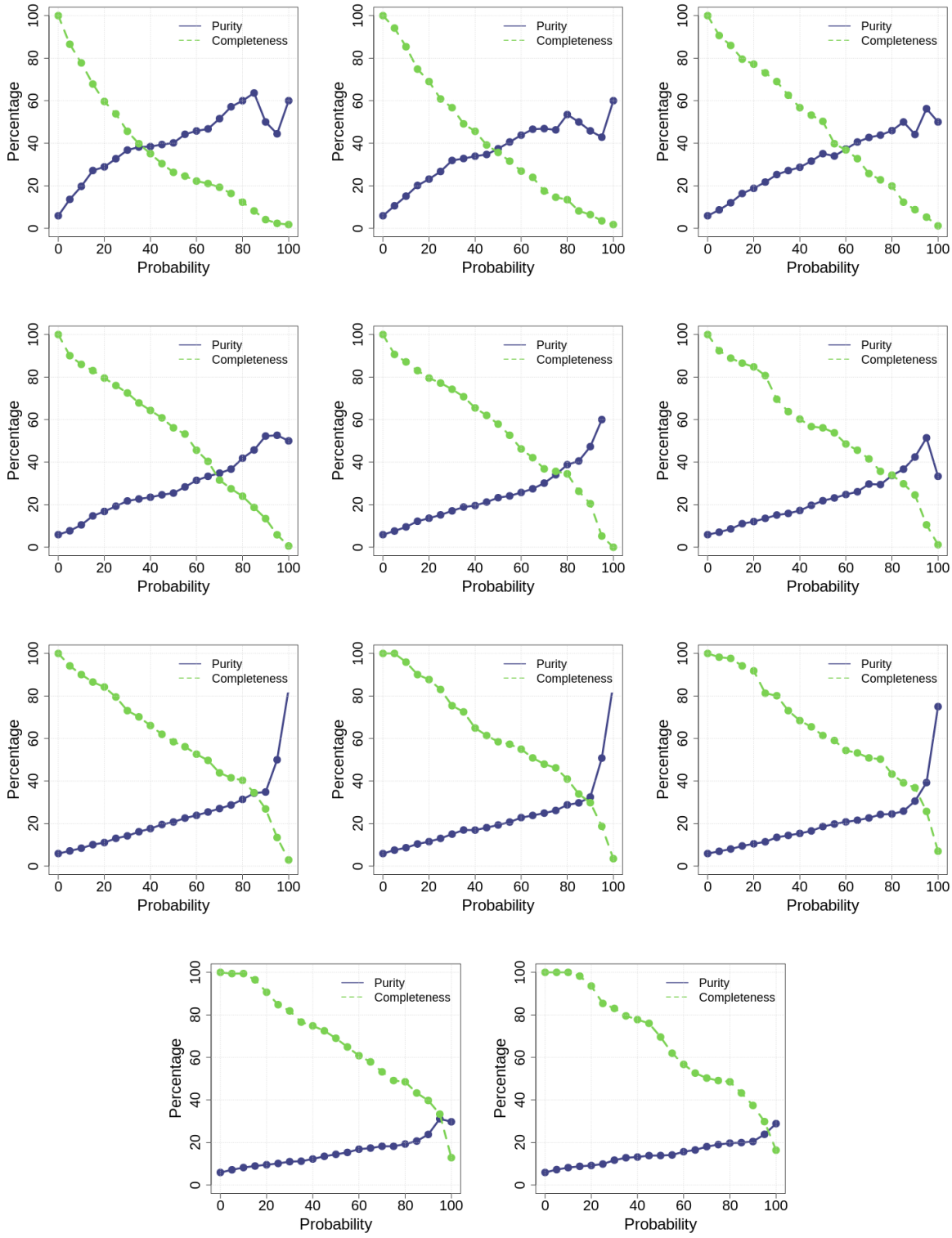


Figure A.1: Completeness (green dashed) and Purity (blue solid) of the original unsupervised UPMASK (with 100 runs) classification (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). From left to right, top to bottom, is the test performed for 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 mean objects per cluster.

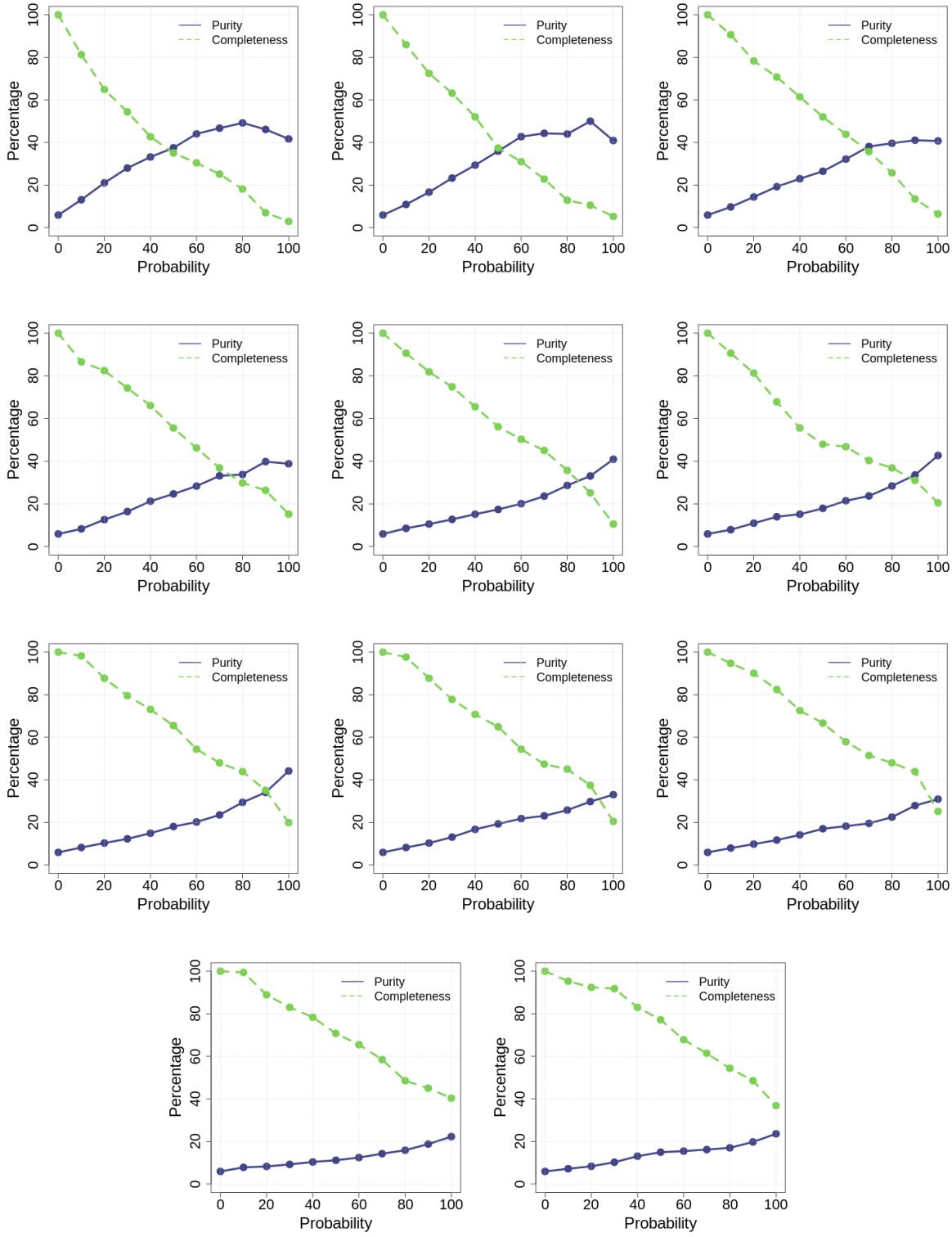


Figure A.2: Completeness (green dashed) and Purity (blue solid) of the original unsupervised UPMASK (with 10 runs) classification (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). From left to right, top to bottom, is the test performed for 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 mean objects per cluster.

A.2 KDE with a fitting Function

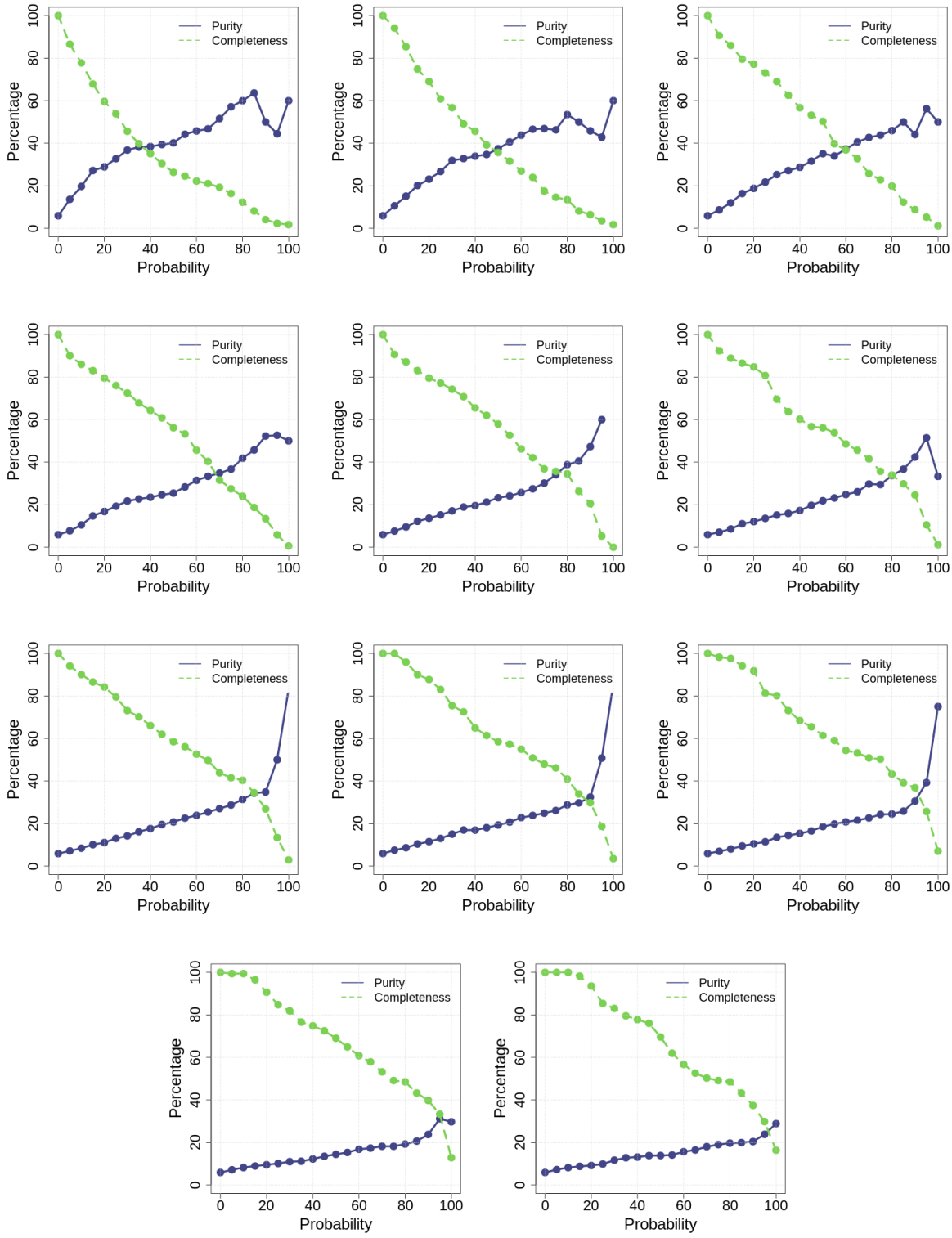


Figure A.3: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK with the KDE and a fitting function version (with 100 runs) classification (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). From left to right, top to bottom, is the test performed for 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 mean objects per cluster.

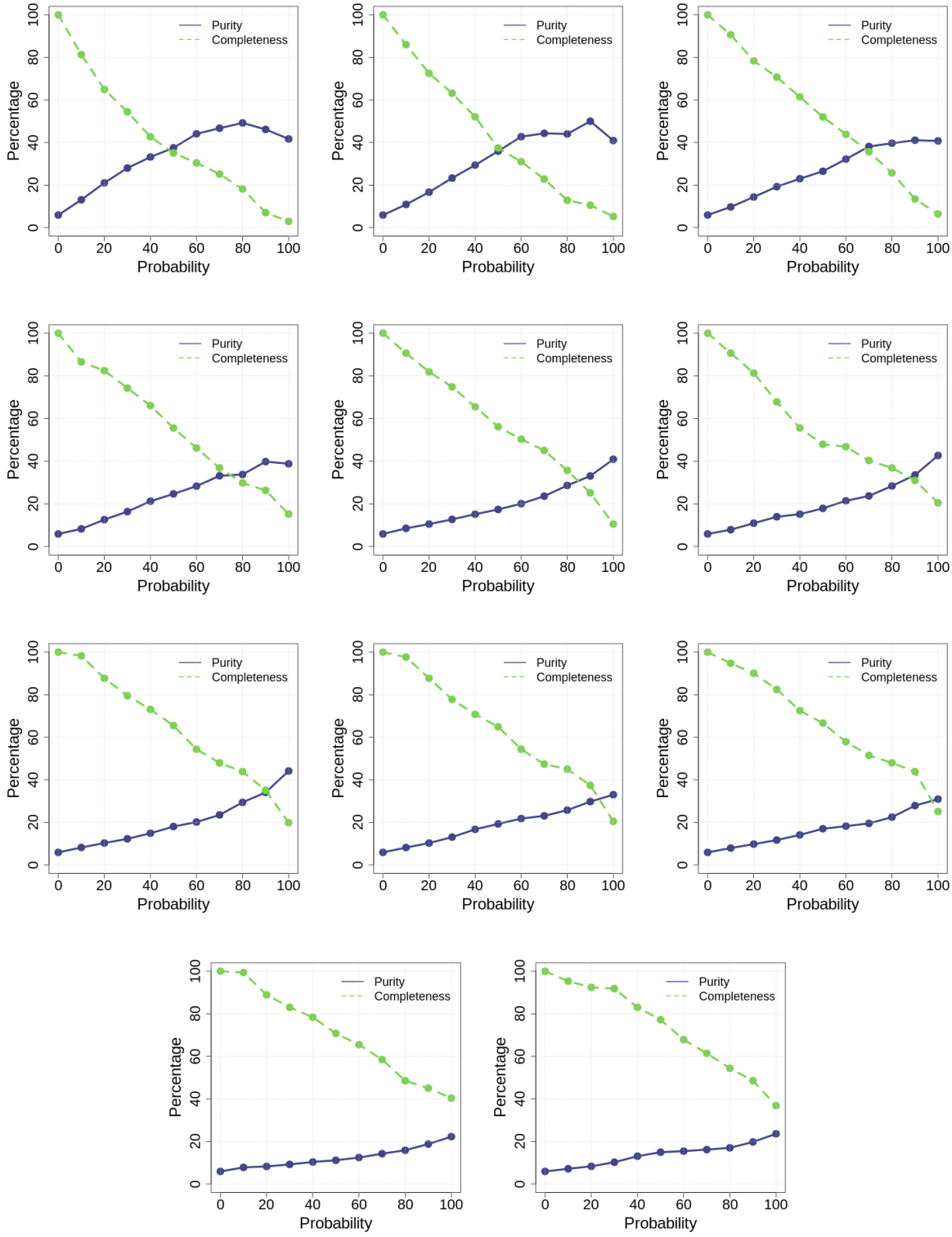


Figure A.4: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK with the KDE and a fitting function version (with 10 runs) classification (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). From left to right, top to bottom, is the test performed for 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 mean objects per cluster.

A.3 Voronoi + Anderson-Darling Test

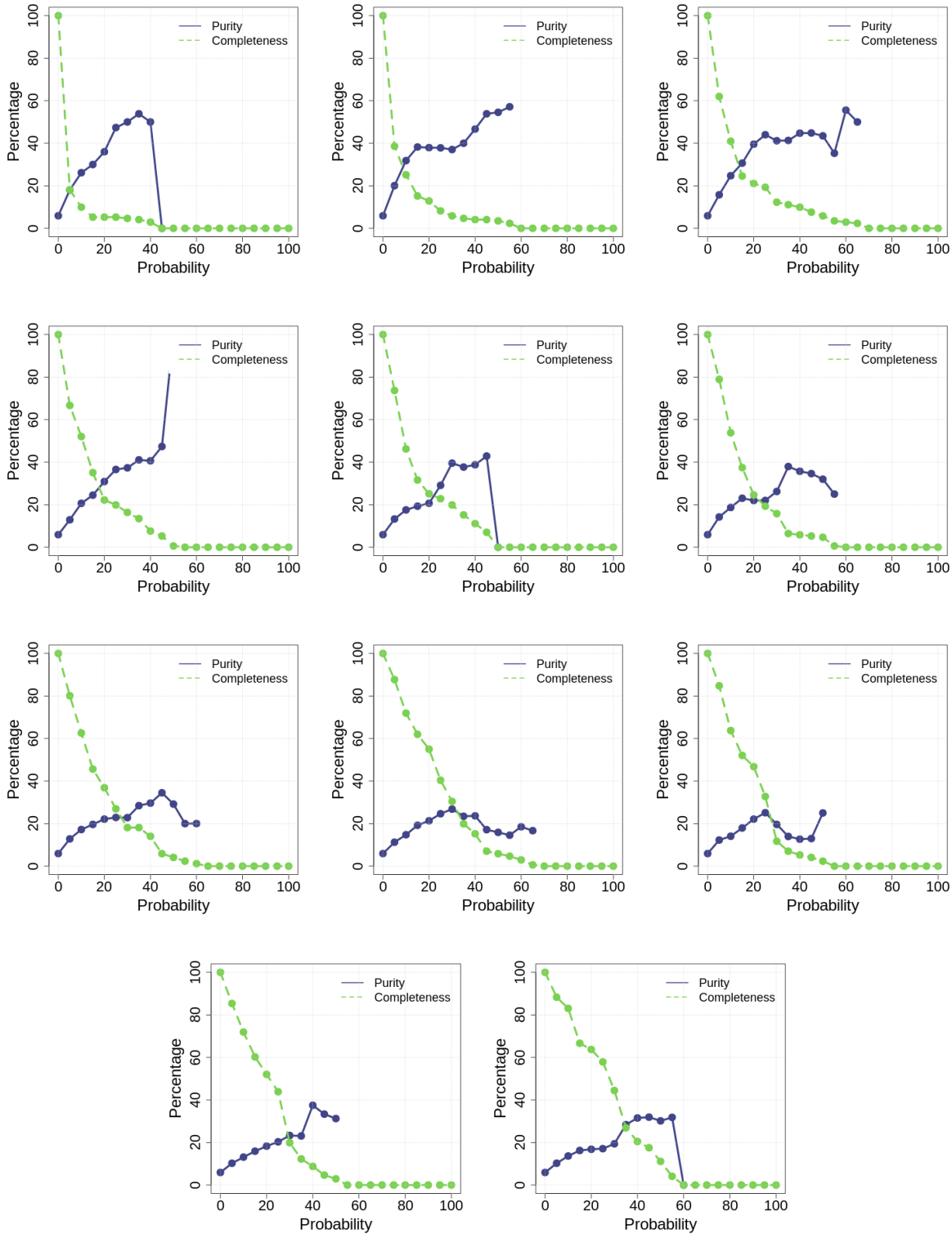


Figure A.5: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK with the Voronoi and Anderson-Darling test version (with 100 runs) classification (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). From left to right, top to bottom, is the test performed for 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 mean objects per cluster.

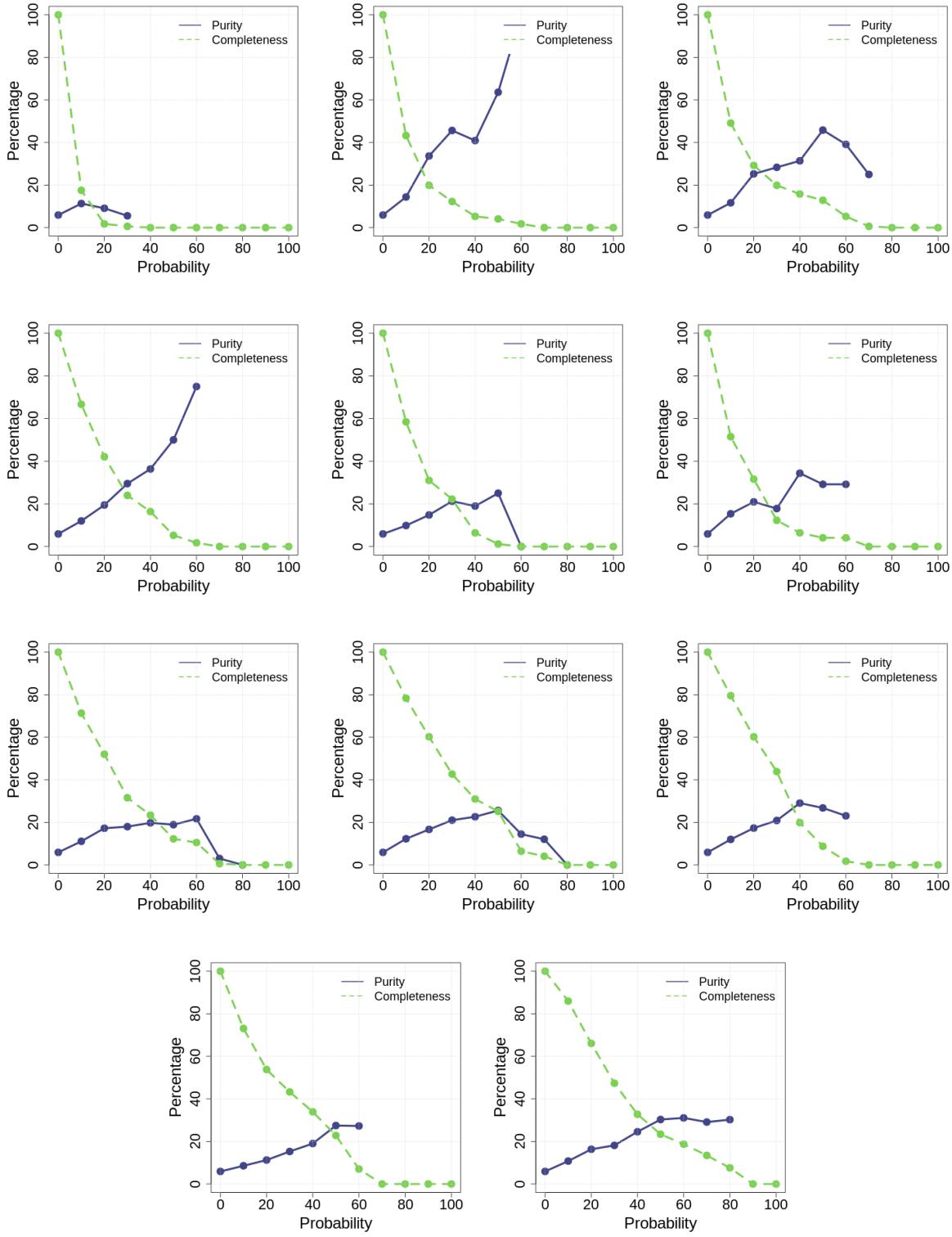


Figure A.6: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK with the Voronoi and Anderson-Darling test version (with 10 runs) classification (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). From left to right, top to bottom, is the test performed for 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 mean objects per cluster.

A.4 Voronoi + Mean Comparison

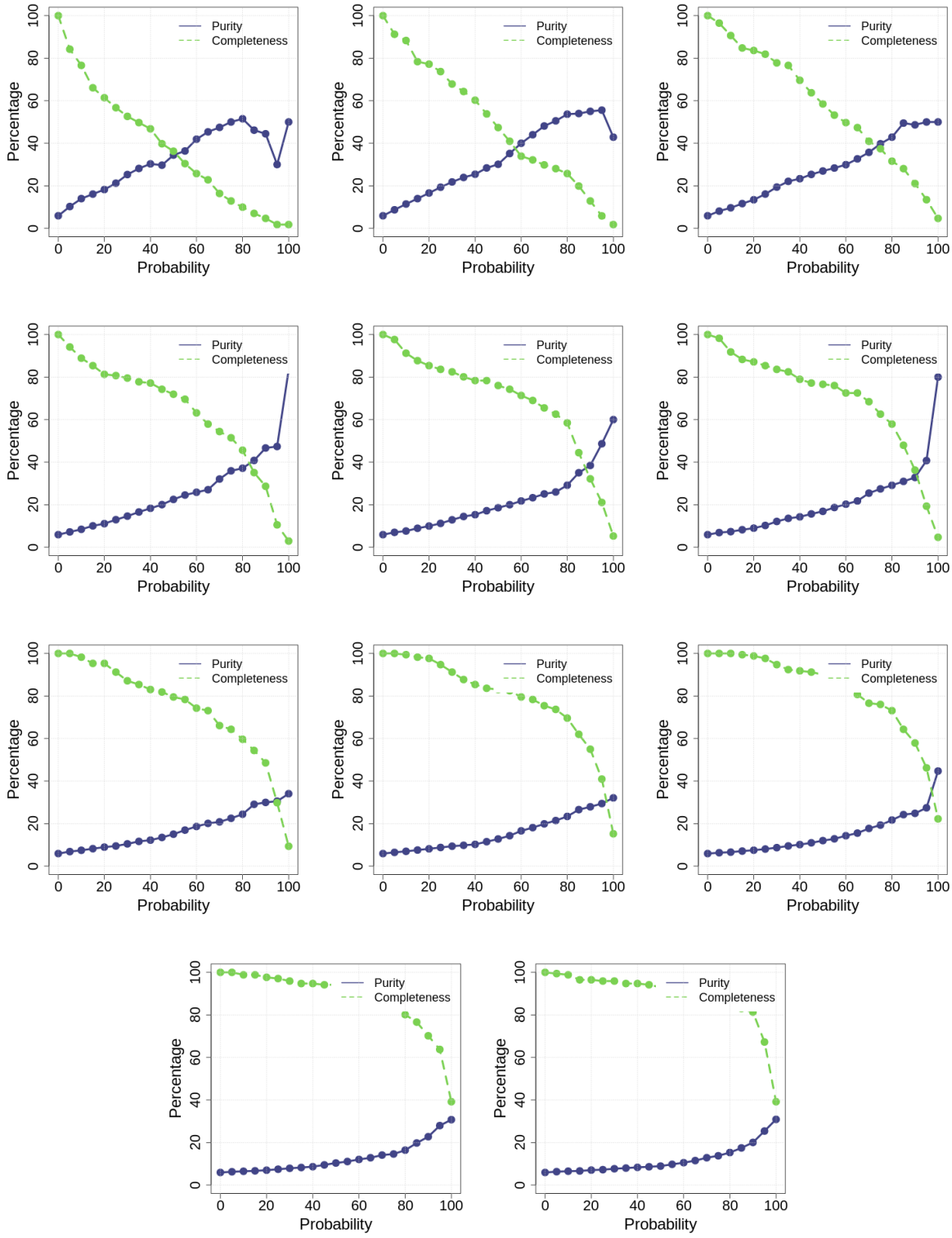


Figure A.7: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK with the Voronoi and a comparison of mean version (with 100 runs) classification (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). From left to right, top to bottom, is the test performed for 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 mean objects per cluster.

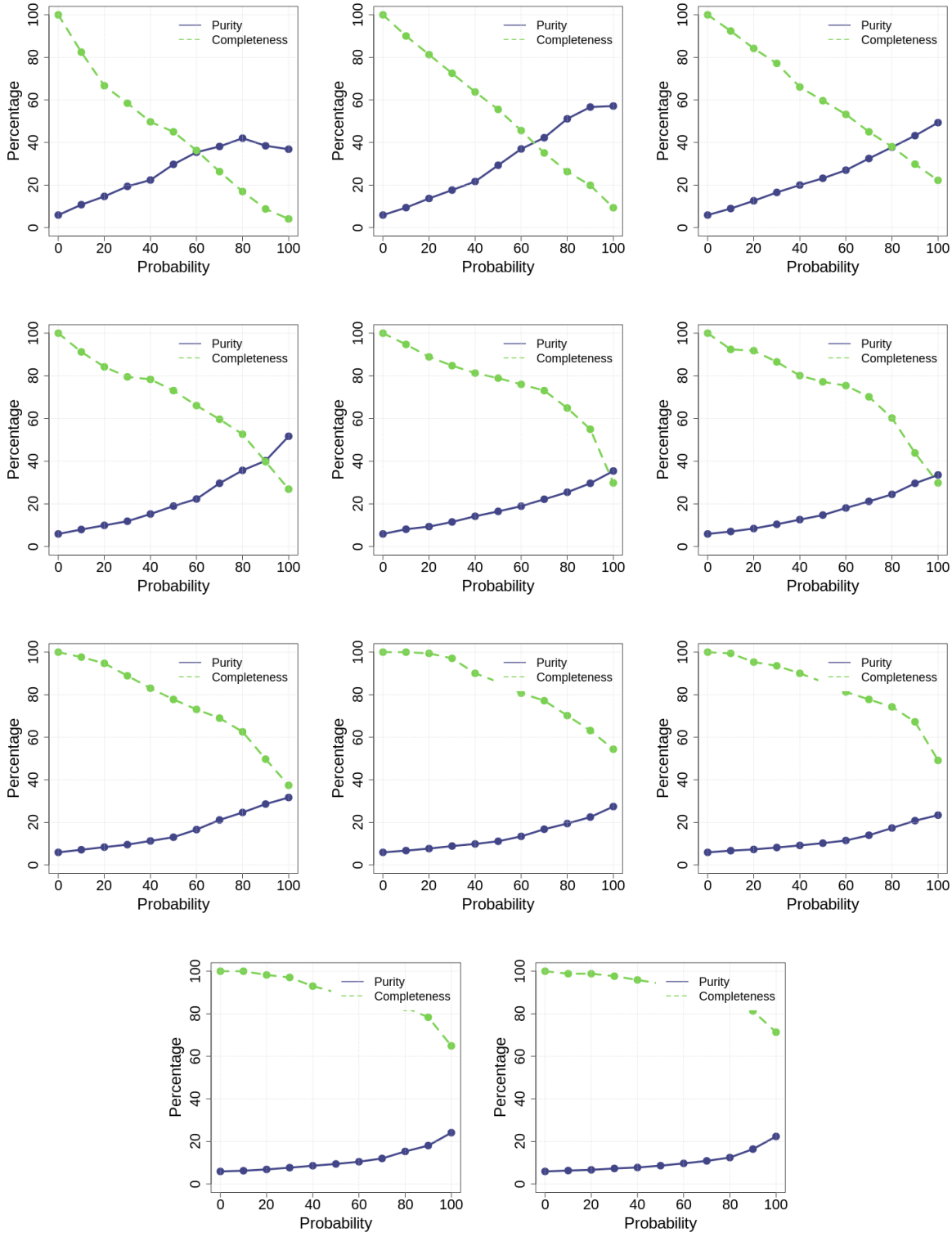


Figure A.8: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK with the Voronoi and a comparison of mean version (with 10 runs) classification (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). From left to right, top to bottom, is the test performed for 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 mean objects per cluster.

A.5 Grid

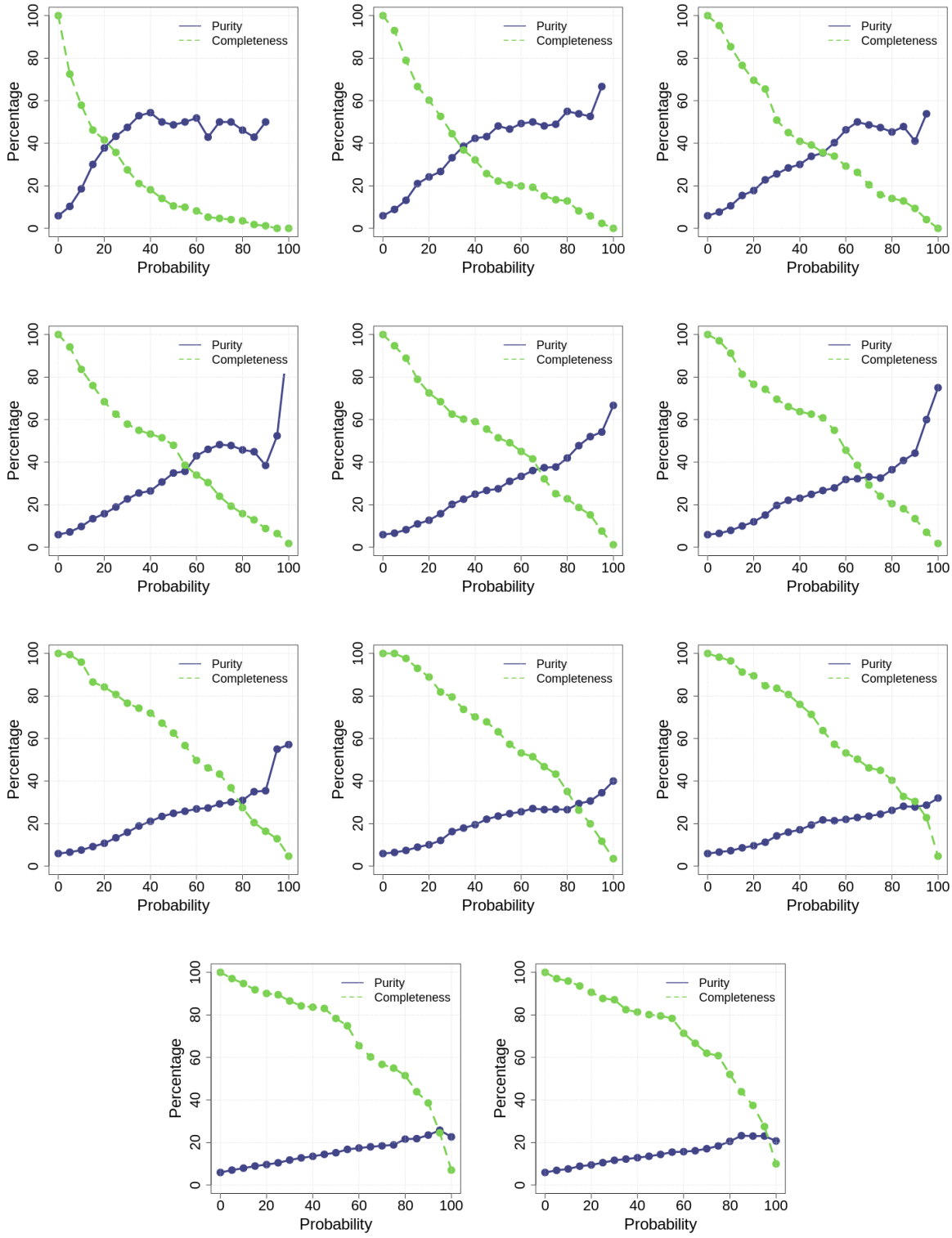


Figure A.9: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK with the Grid version (with 100 runs) classification (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). From left to right, top to bottom, is the test performed for 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 mean objects per cluster.

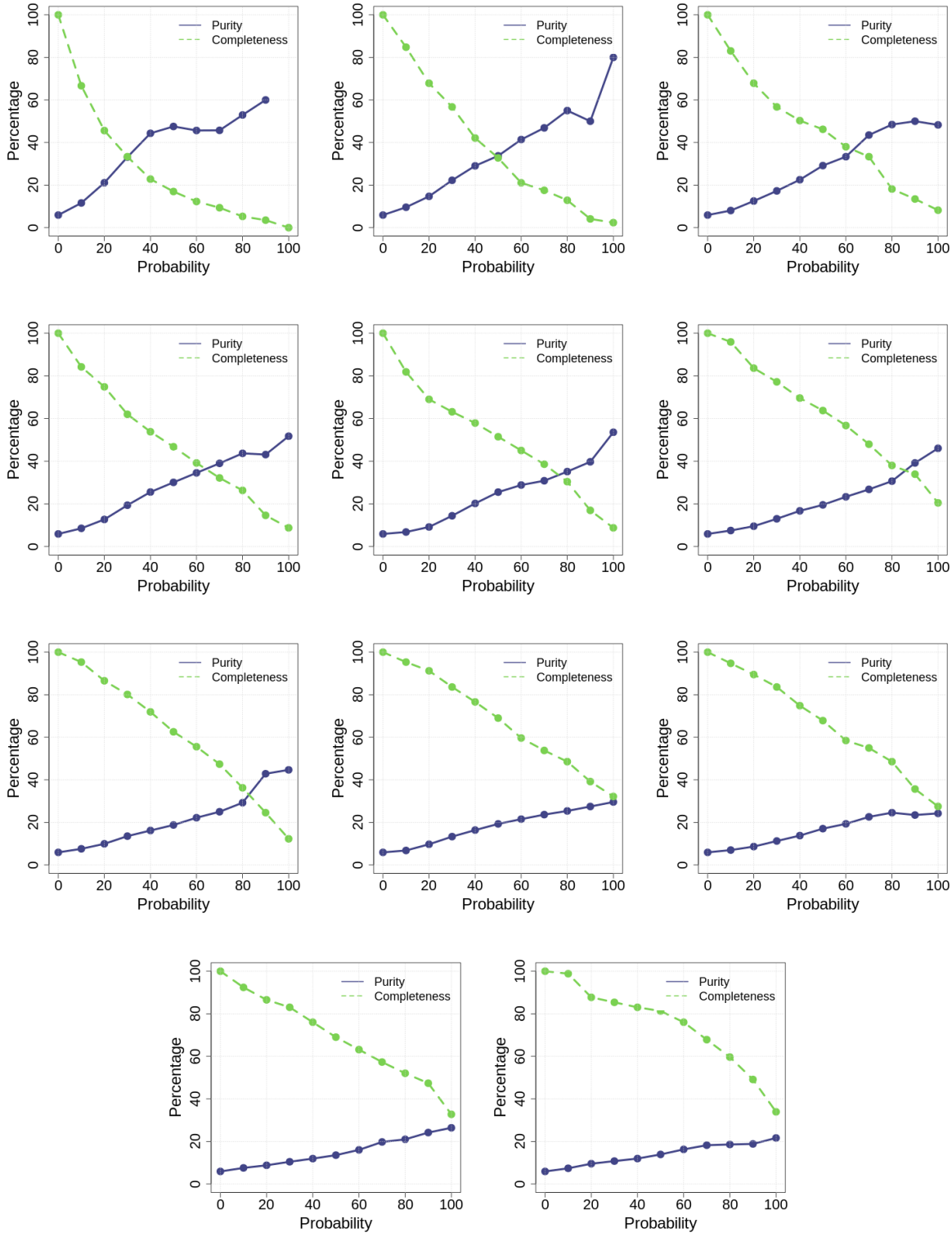


Figure A.10: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK with the Grid version (with 10 runs) classification (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). From left to right, top to bottom, is the test performed for 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 mean objects per cluster.

A.6 Grid with a fitting Function

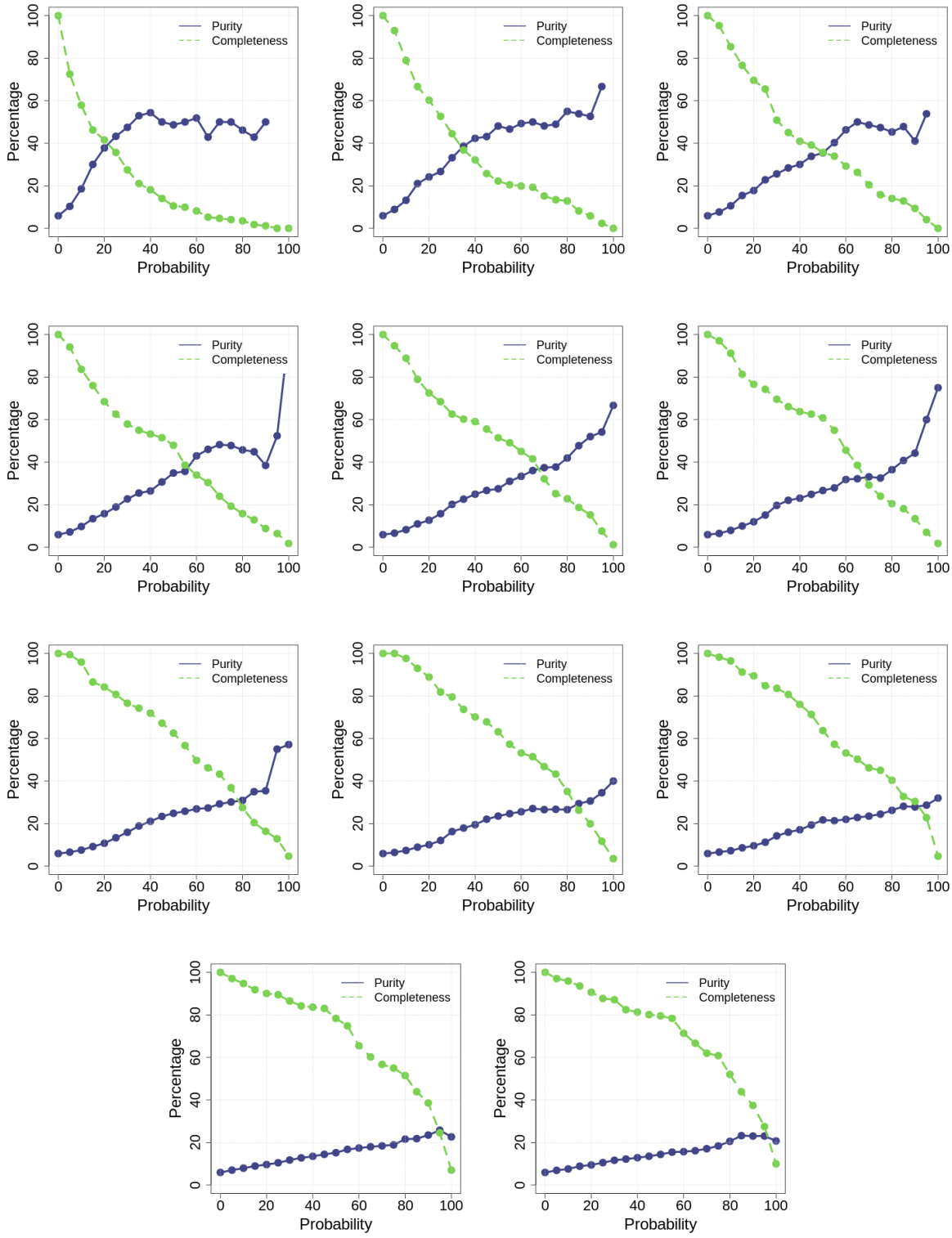


Figure A.11: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK with the Grid and a fitting function version (with 100 runs) classification (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). From left to right, top to bottom, is the test performed for 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 mean objects per cluster.

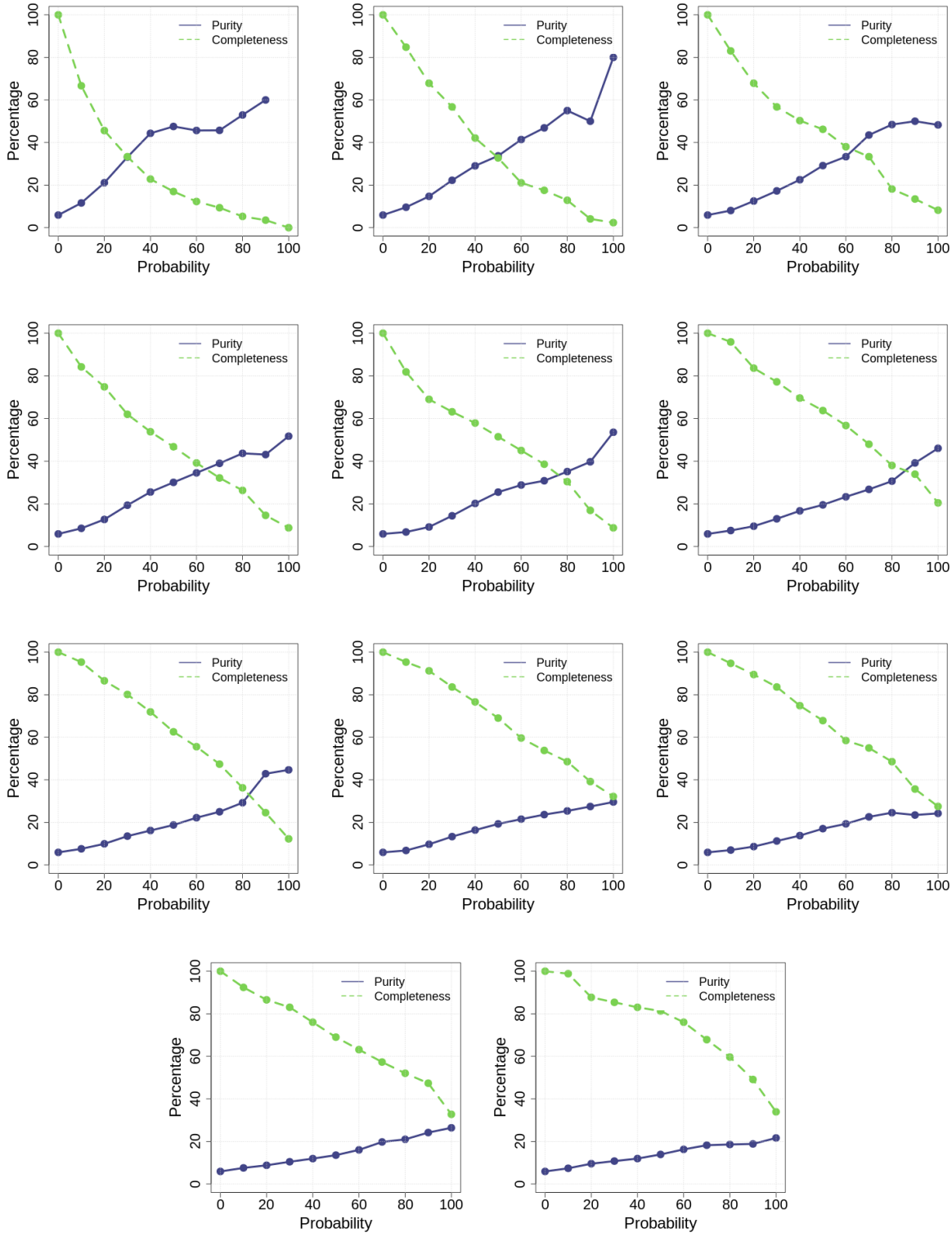
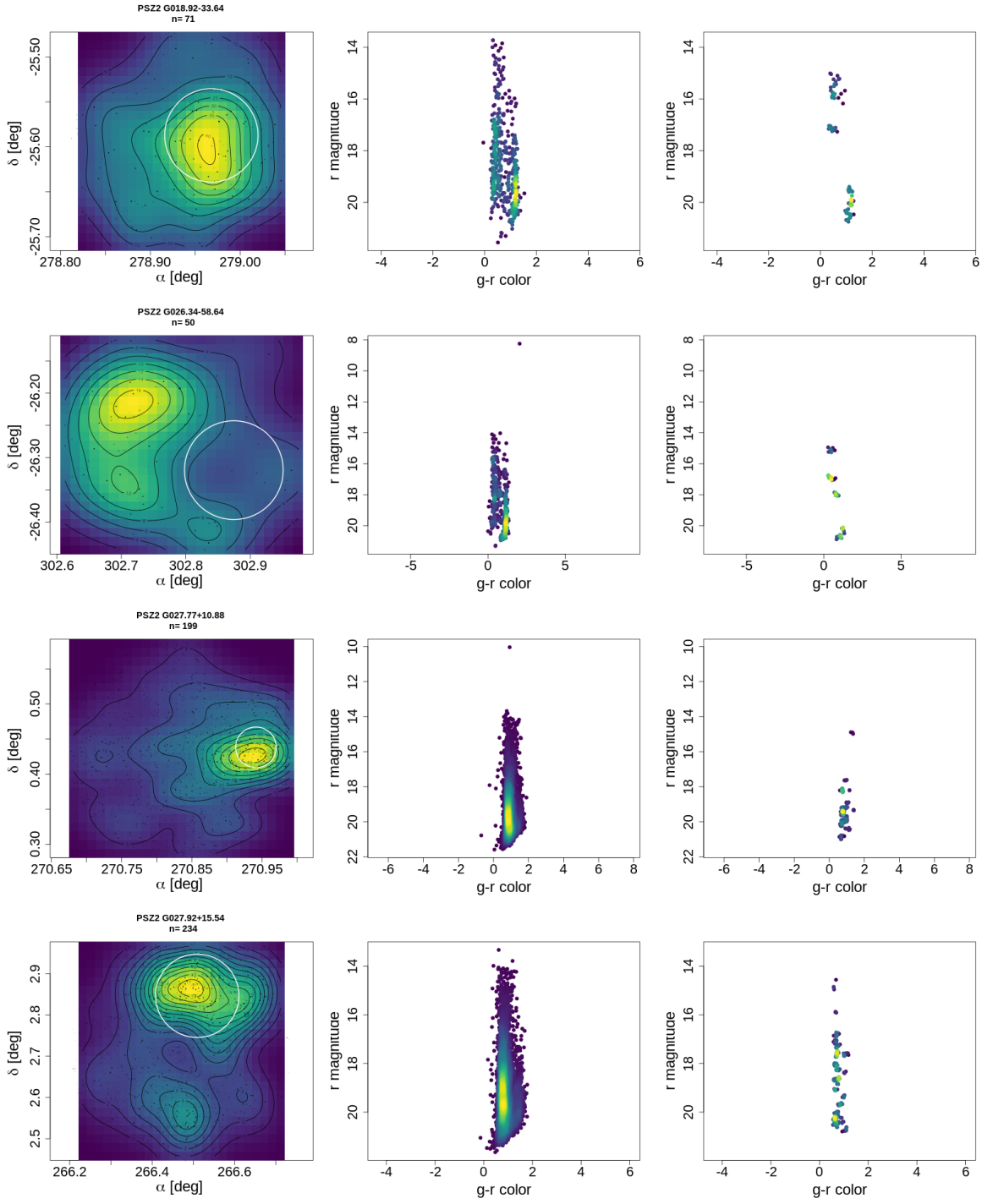


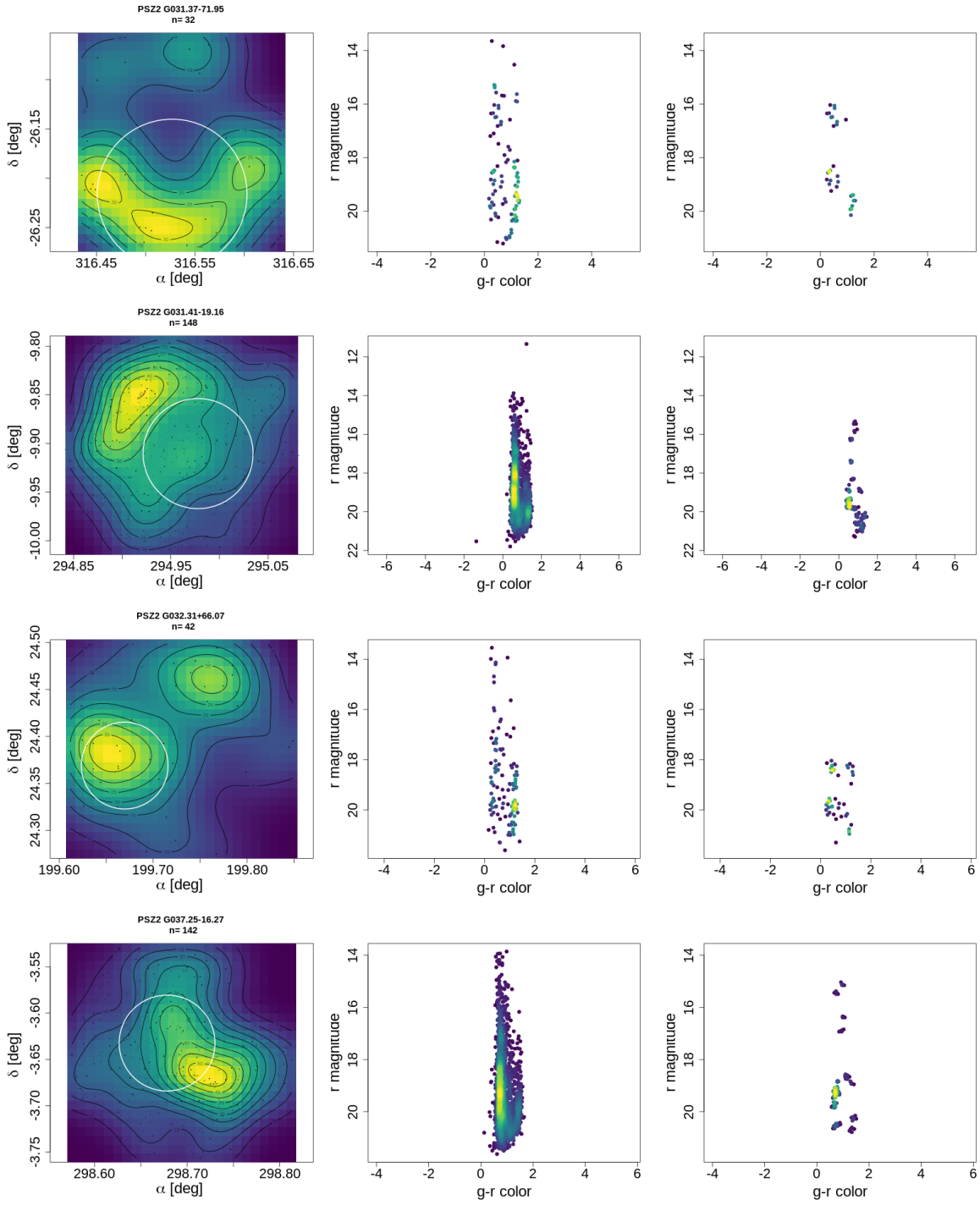
Figure A.12: Completeness (green dashed) and Purity (blue solid) of the unsupervised UPMASK with the Grid and a fitting function version (with 10 runs) classification (applied to the data of section 3.1) for different number of objects per cluster (of the K-means). From left to right, top to bottom, is the test performed for 20, 35, 50, 65, 80, 95, 110, 125, 140, 165 and 180 mean objects per cluster.

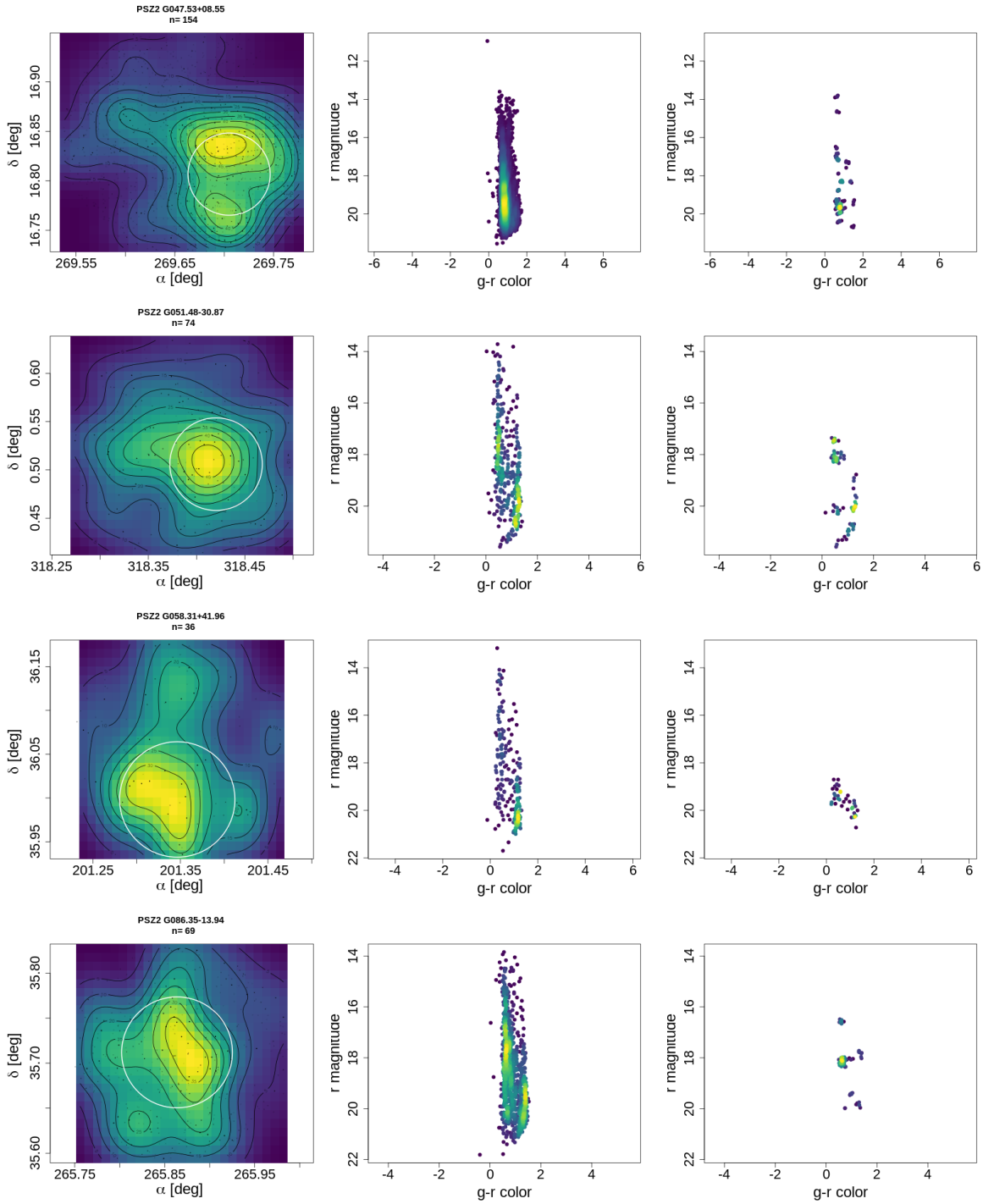
Appendix B

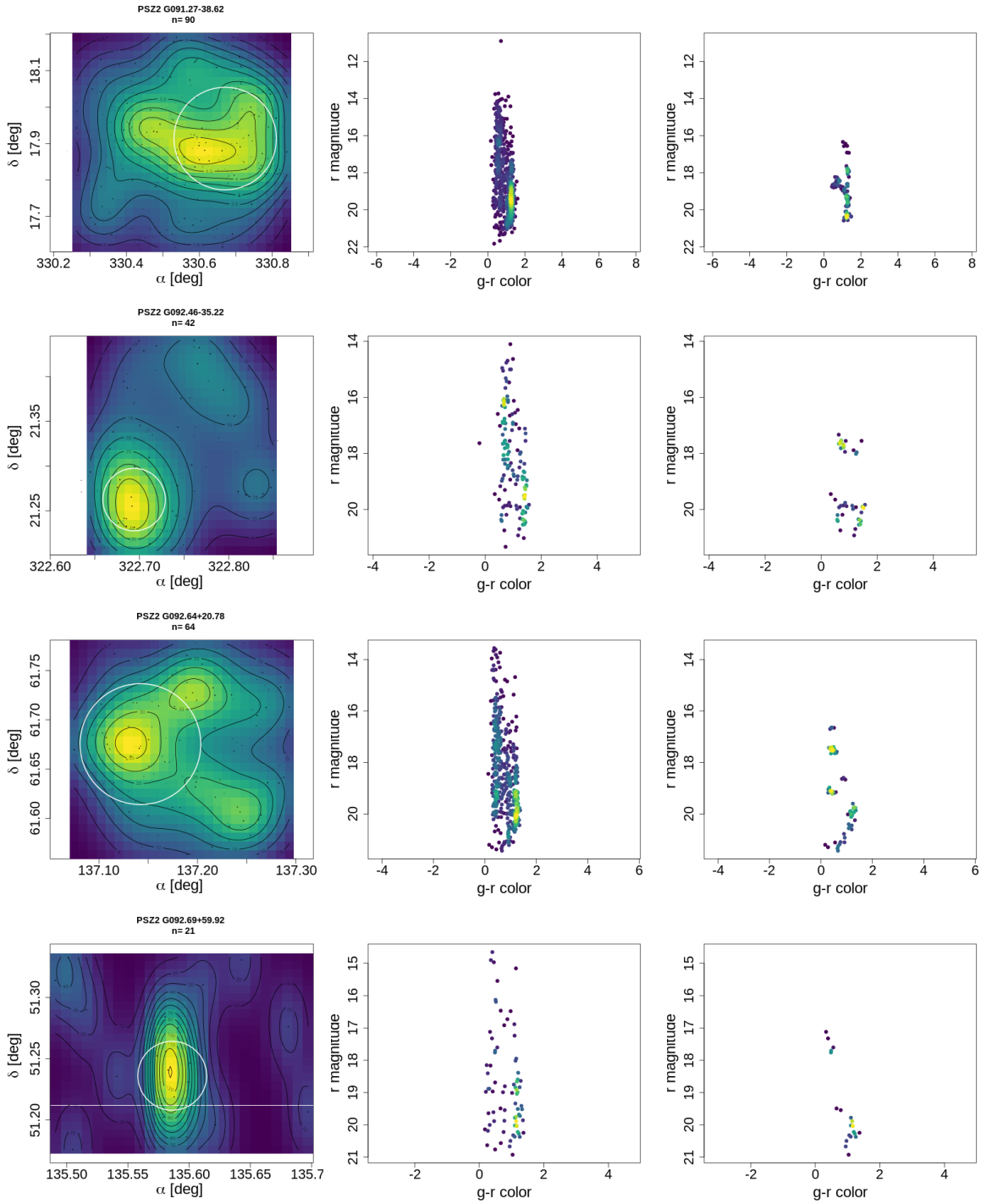
Planck Rediscovered Clusters

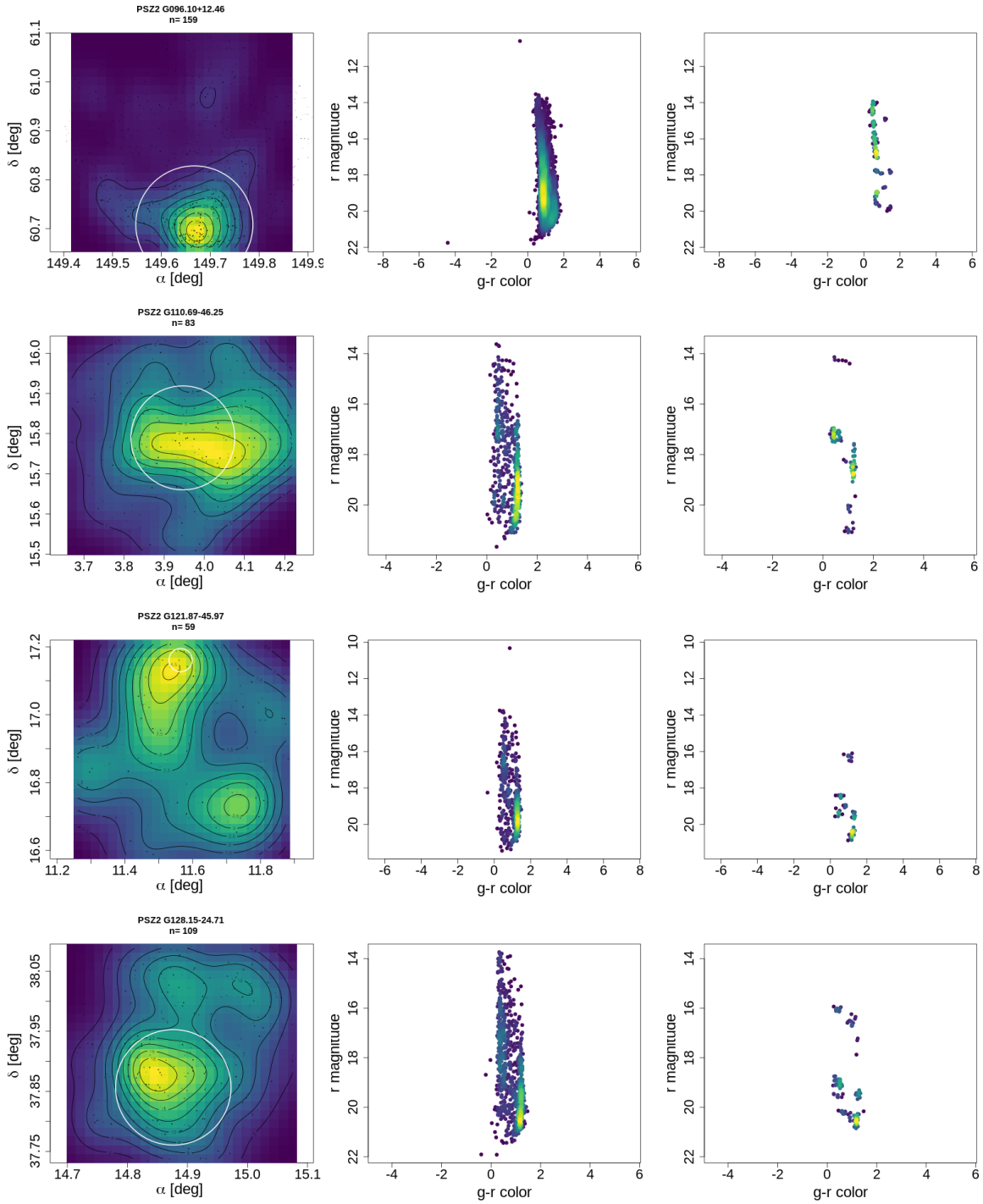
Figure B.1: In this page and the following ones the 46 fields of PLANCKSZ2 sources, with more than 10 objects detected with 100% probability are presented one at each row organized as: Left - A KDE representation (color map: the brighter, the denser), iso-countours and points (in which the opacity corresponds to the probability) of the most likely members (members with probability above 50%). The white circle corresponds to an estimated radius and center of the cluster, of which the procedure is described in chapter 5 and the title is written the name of the PLANCKSZ2 that corresponds to the represented field. Center - Color (g-r) vs Magnitude (r) for the galaxies contained in the field. Right - Color (g-r) vs Magnitude (r) for the galaxies with a probability above 50% contained in this field. The colors of the points of the last two graphs are based on the local density (the brighter, the denser.)

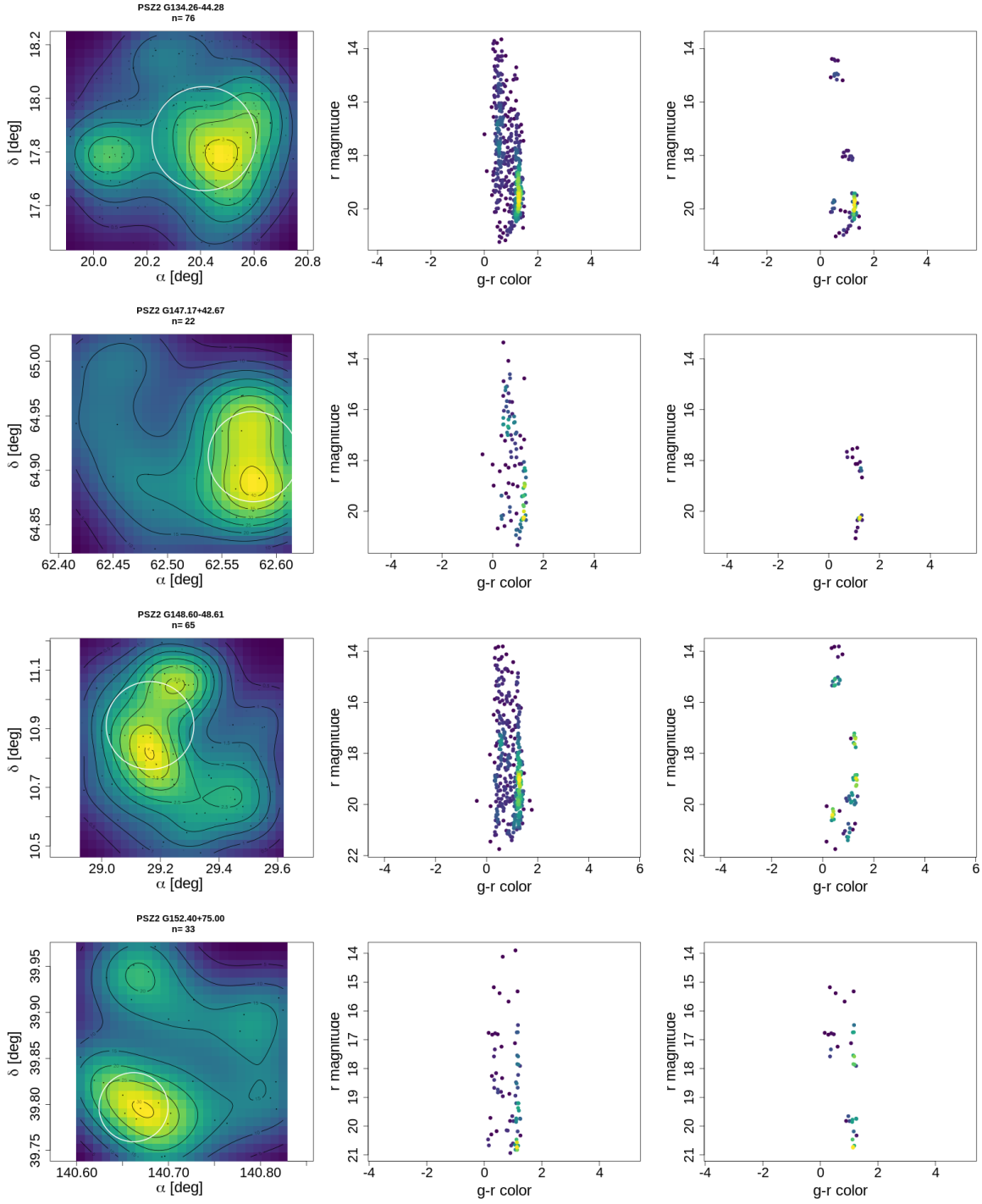


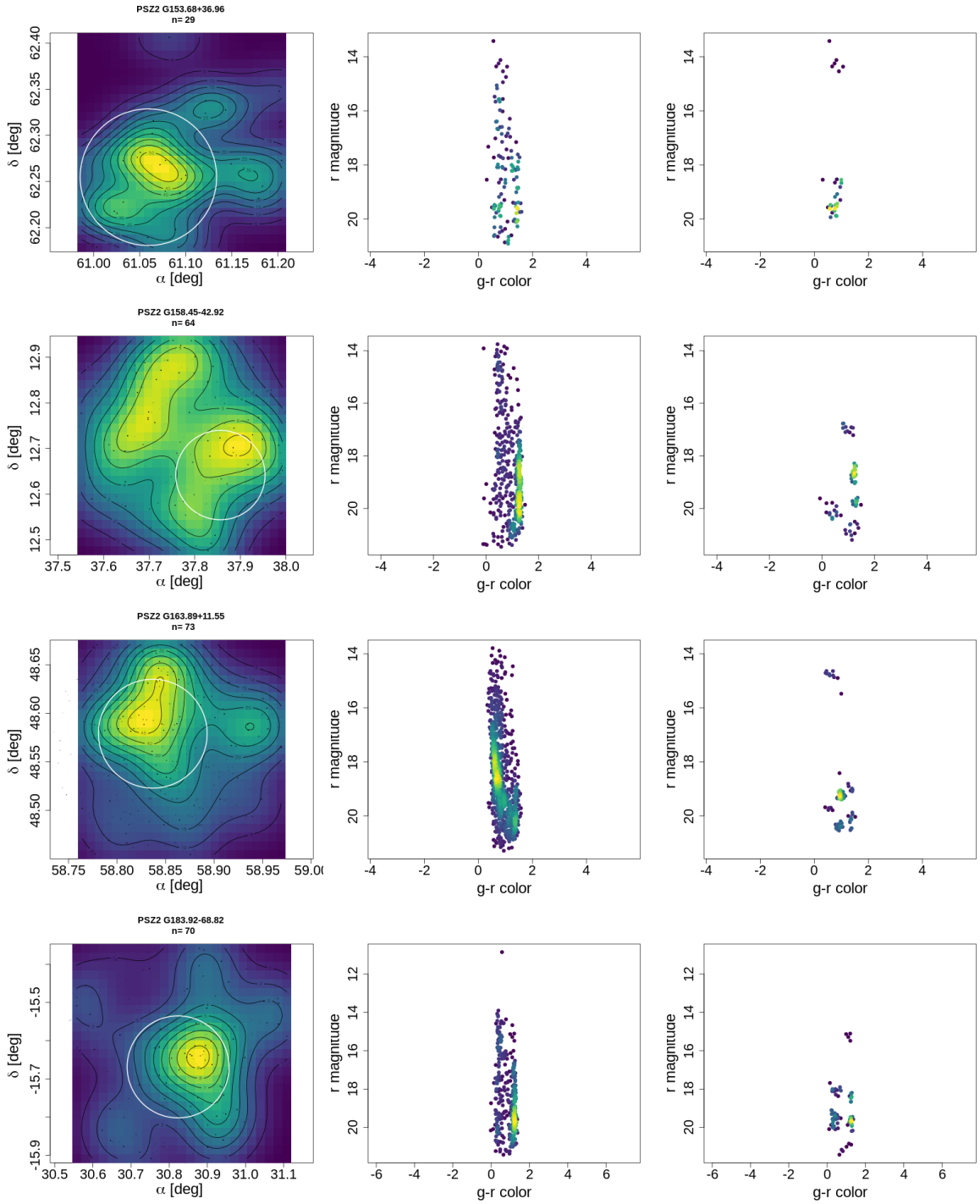

PSZ2 G031.37-71.95
n= 32
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G031.41-19.16
n= 148
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G032.31+66.07
n= 42
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G037.25-16.27
n= 142
δ [deg]
α [deg]
r magnitude
g-r color

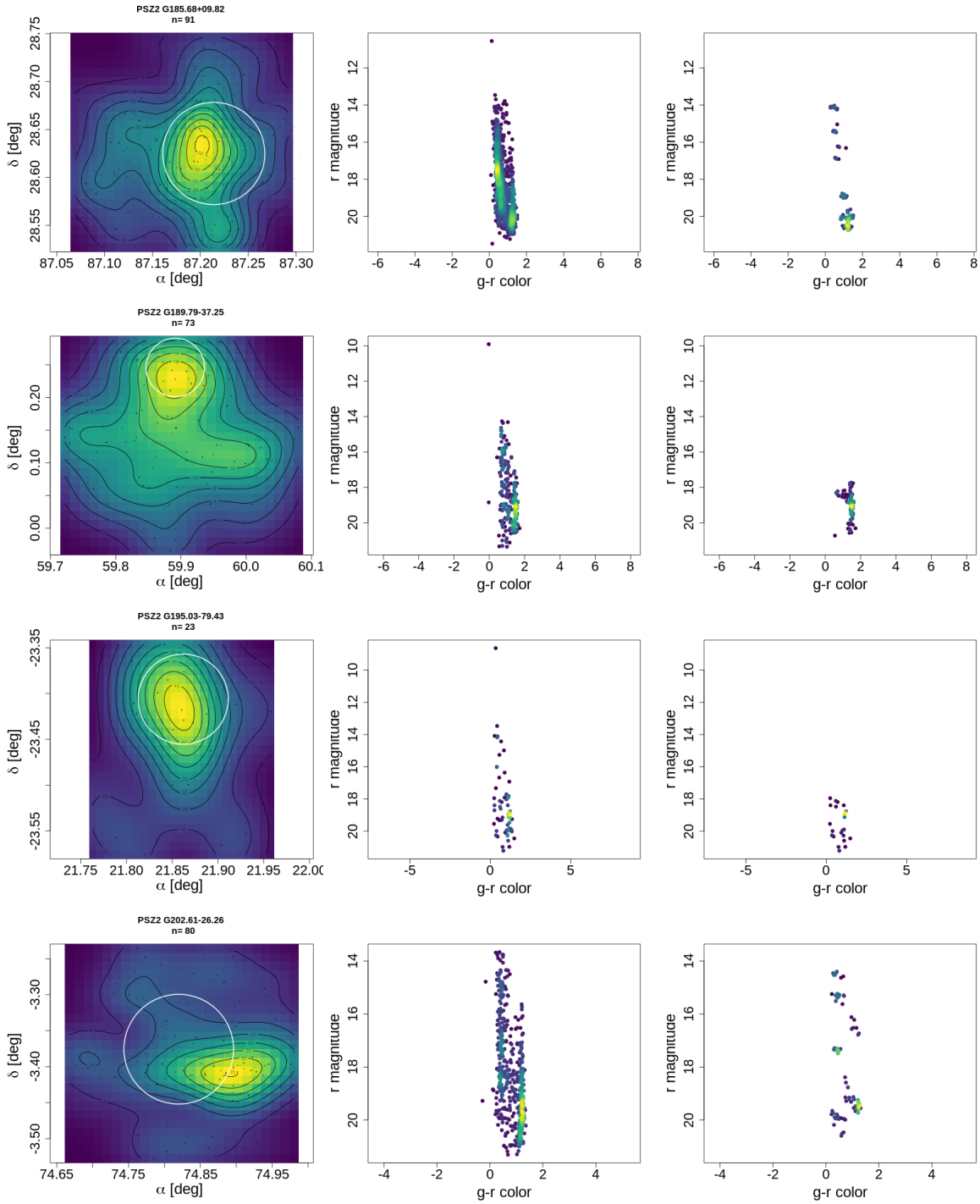

PSZ2 G047.53+08.55
n= 154
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G051.48-30.87
n= 74
PSZ2 G058.31+41.96
n= 36
PSZ2 G086.35-13.94
n= 69

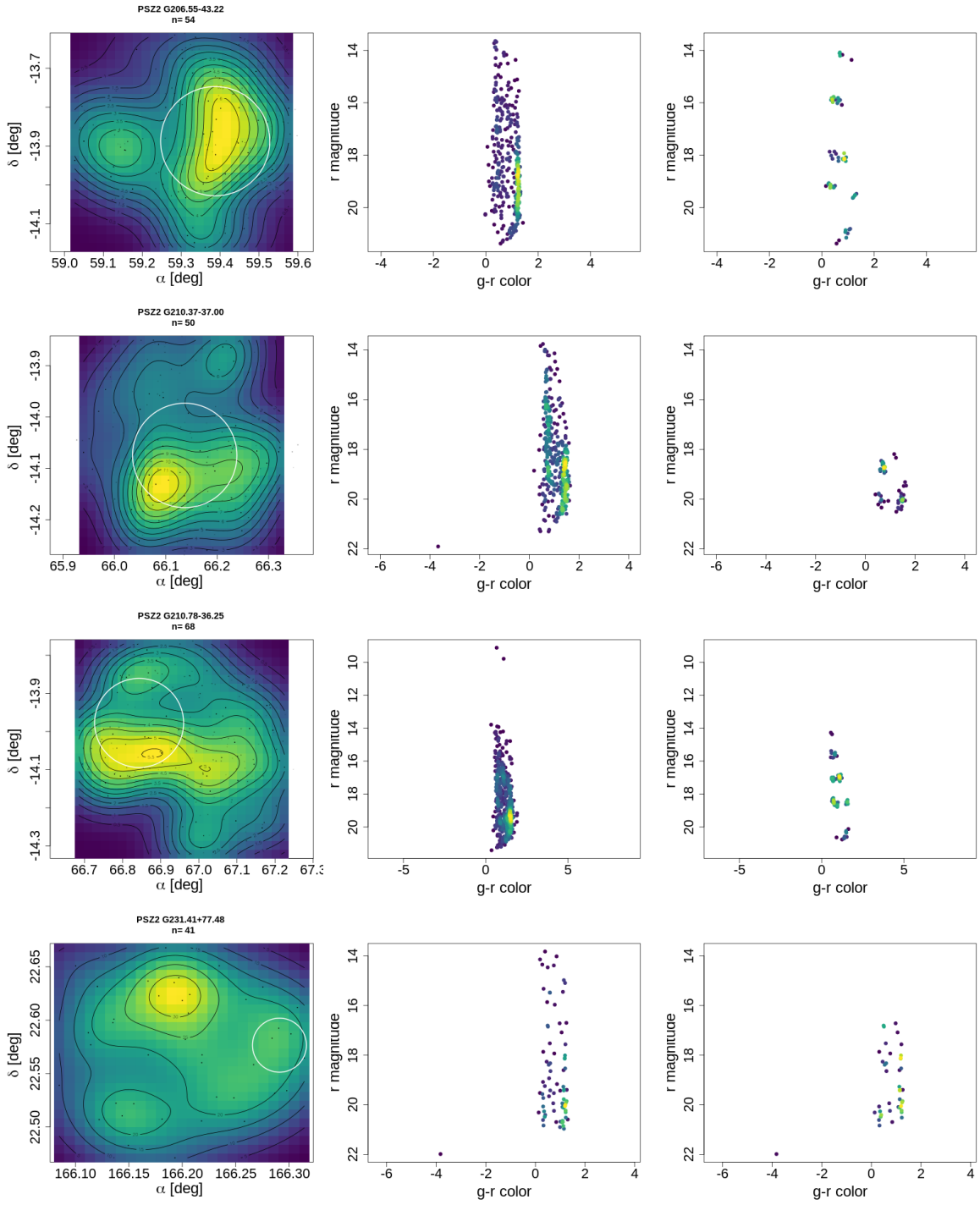

PSZ2 G091.27-38.62
n= 90
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G092.46-35.22
n= 42
PSZ2 G092.64+20.78
n= 64
PSZ2 G092.69+59.92
n= 21

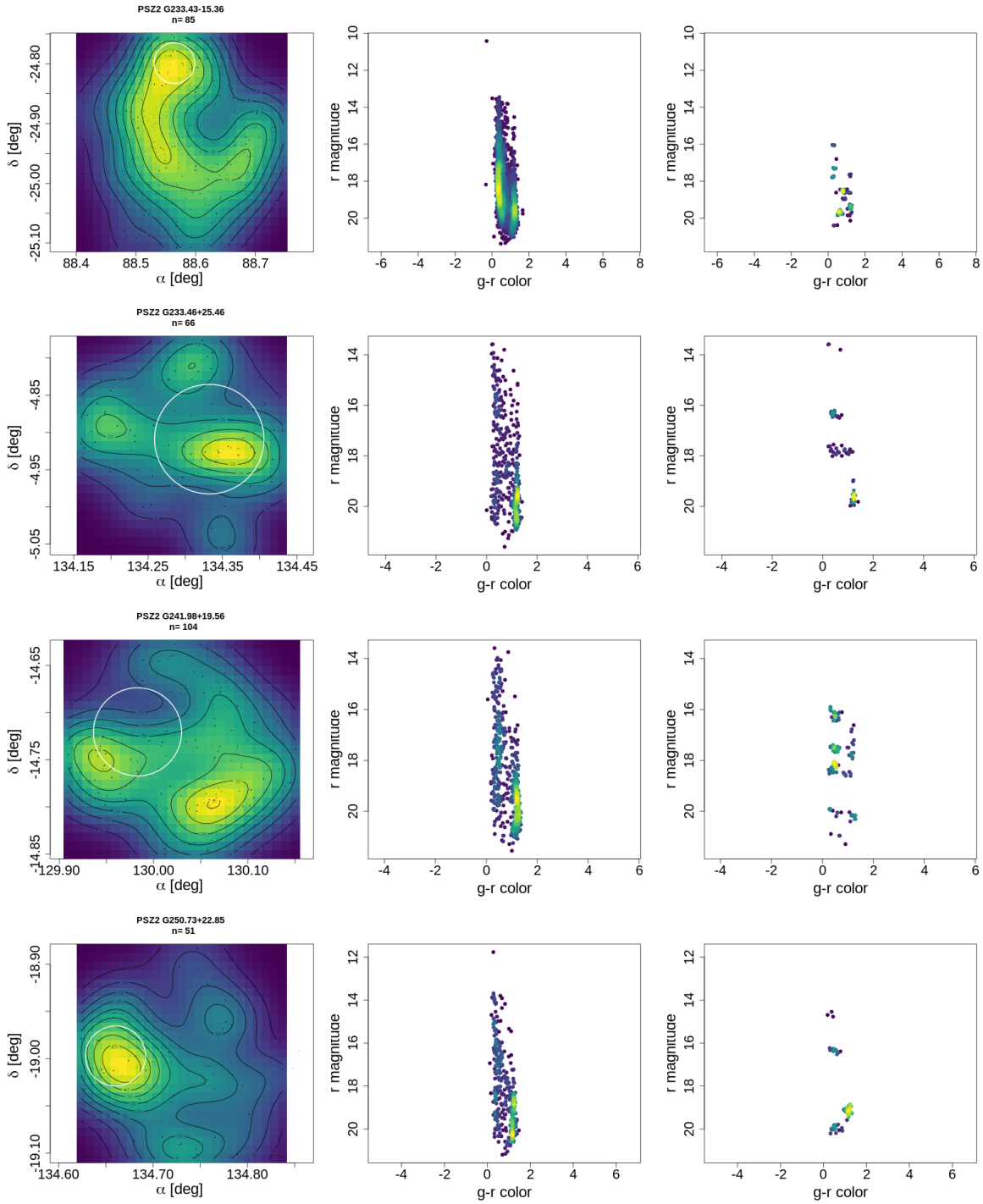

PSZ2 G096.10+12.46
n= 159
PSZ2 G110.69-46.25
n= 83
PSZ2 G121.87-45.97
n= 59
PSZ2 G128.15-24.71
n= 109
δ [deg]
α [deg]
r magnitude
g-r color

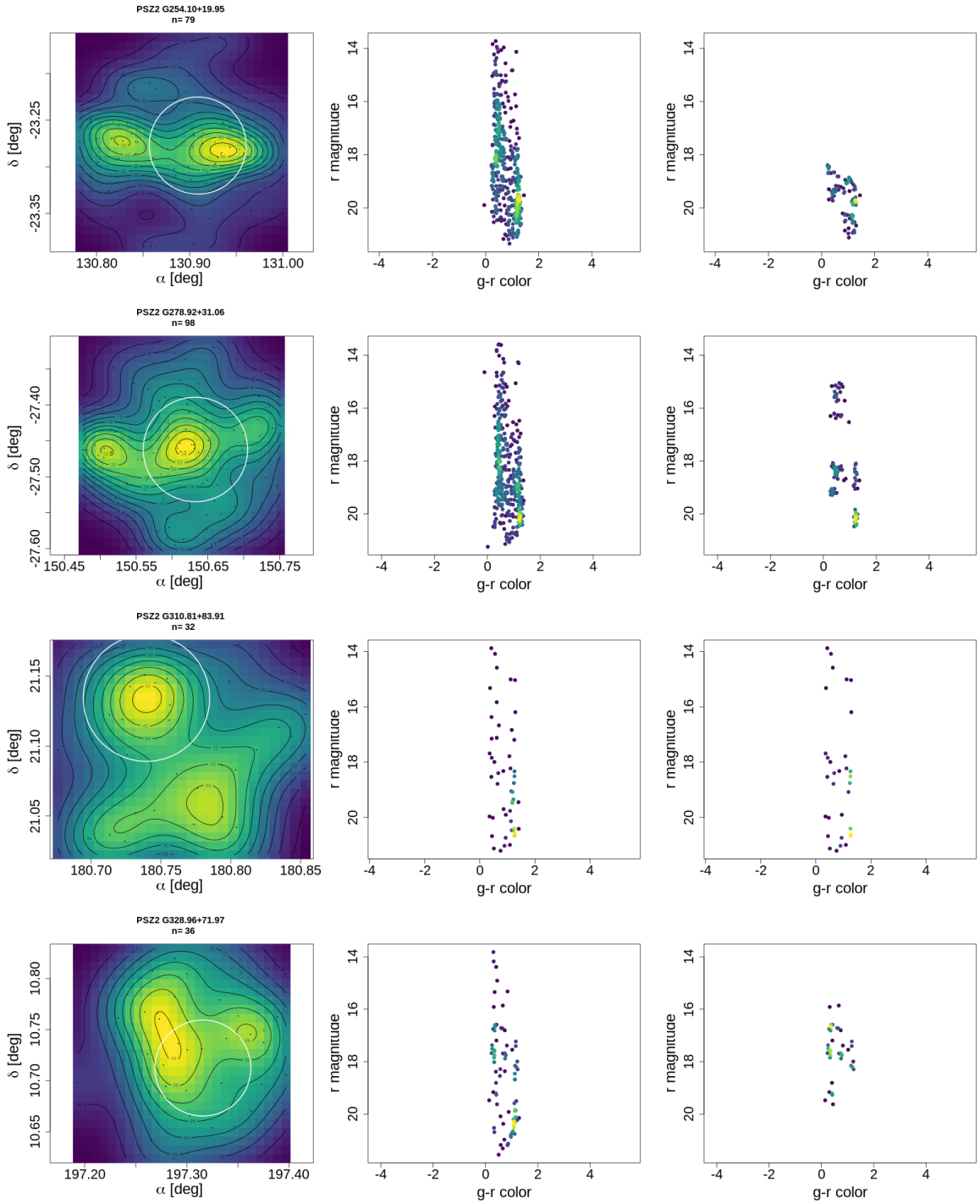

PSZ2 G134.26-44.28
n= 76
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G147.17+42.67
n= 22
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G148.60-48.61
n= 65
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G152.40+75.00
n= 33
δ [deg]
α [deg]
r magnitude
g-r color

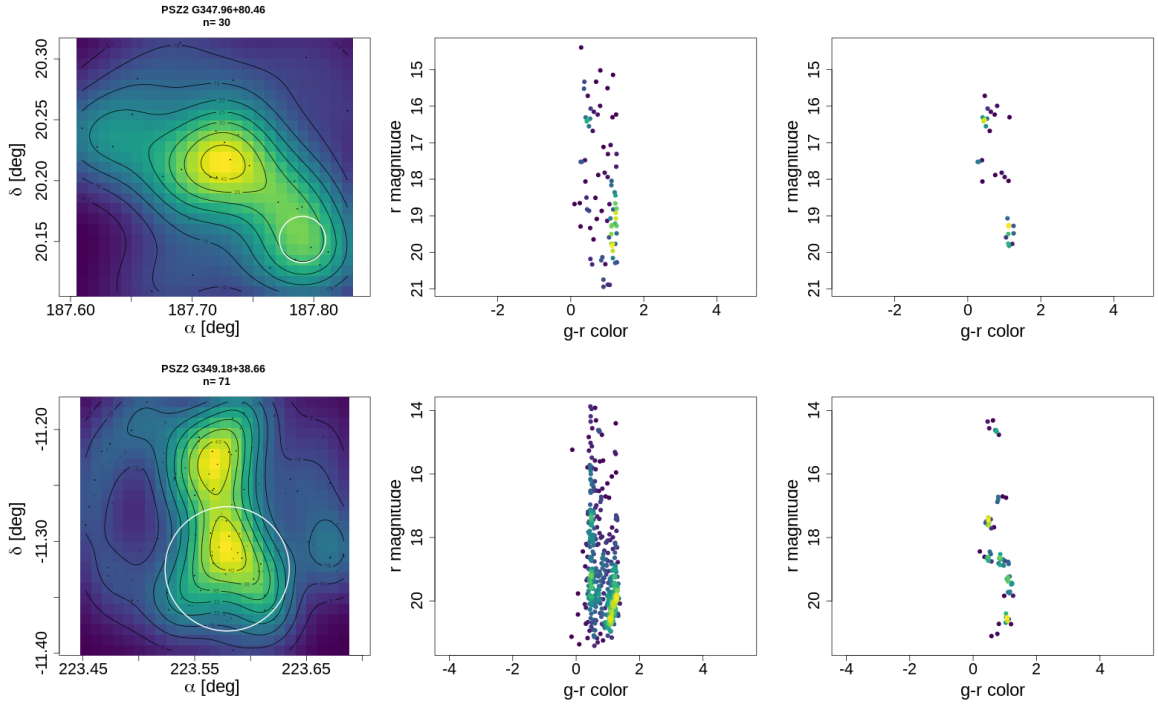

PSZ2 G153.68+36.96
n= 29
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G158.45-42.92
n= 64
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G163.89+11.55
n= 73
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G183.92-68.82
n= 70
δ [deg]
α [deg]
r magnitude
g-r color


PSZ2 G185.68+09.82
n= 91
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G189.79-37.25
n= 73
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G195.03-79.43
n= 23
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G202.61-26.26
n= 80
δ [deg]
α [deg]
r magnitude
g-r color


PSZ2 G206.55-43.22
n= 54
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G210.37-37.00
n= 50
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G210.78-36.25
n= 68
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G231.41+77.48
n= 41
δ [deg]
α [deg]
r magnitude
g-r color


PSZ2 G233.43-15.36
n= 85
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G233.46+25.46
n= 66
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G241.98+19.56
n= 104
δ [deg]
α [deg]
r magnitude
g-r color
PSZ2 G250.73+22.85
n= 51
δ [deg]
α [deg]
r magnitude
g-r color


PSZ2 G254.10+19.95
n= 79
δ [deg]
-23.25
-23.35
130.80
130.90
131.00
α [deg]
r magnitude
14
16
18
20
-4
-2
0
2
4
g-r color
PSZ2 G278.92+31.06
n= 98
-27.40
-27.50
-27.60
150.45
150.55
150.65
150.75
PSZ2 G310.81+83.91
n= 32
21.15
21.10
21.05
180.70
180.75
180.80
180.85
PSZ2 G328.96+71.97
n= 36
10.80
10.75
10.70
10.65
197.20
197.30
197.40


PSZ2 G347.96+80.46
n= 30
δ [deg]
20.15
20.20
20.25
20.30
187.60
187.70
187.80
α [deg]
r magnitude
15
16
17
18
19
20
21
-2
0
2
4
g-r color
r magnitude
15
16
17
18
19
20
21
-2
0
2
4
g-r color
PSZ2 G349.18+38.66
n= 71
δ [deg]
-11.20
-11.30
-11.40
223.45
223.55
223.65
α [deg]
r magnitude
14
16
18
20
-4
-2
0
2
4
g-r color
r magnitude
14
16
18
20
-4
-2
0
2
4
g-r color