

## Large Scale Structures in the Universe. Cluster analysis of faint galaxies in the direction of HERCULES void.\*

**G. Petrov<sup>1</sup>, J.W. Fried<sup>2</sup>, A. Kniazev<sup>3</sup>**

<sup>1</sup>Institute of Astronomy,  
Bulgarian Academy of Sciences, 72 Tsarigradsko chaussee, Sofia 1784,  
Bulgaria

<sup>2</sup>Max Plank Institute of Astronomy, Koenigstuhl 17, Heidelberg, D 69117,  
Germany

<sup>3</sup>European Southern Observatory, Garching bei Muenchen, D-85748, Ger-  
many

### **Abstract.**

Ca. 1850 faint galaxies in the wide range of magnitudes  $13^m \leq B \leq 21^m$  and effective surface brightnesses  $16 \leq \mu_{\text{eff}}(B) \leq 24 \text{ mag arcsec}^{-2}$  are detected in a field of one square degree centered at 1600+18 (1950) (Hercules void). Their coordinates (2000), magnitudes  $m(B)$ , diameters, position angles surface brightness and some morphological parameters are studied using cluster analysis technique. Nearest and Furthest neighbor, Centroid, Median, Group, K\_means and Wards methods were used to determine the substructures in the distribution of faint galaxies. The distance metric in all the cases is squared Euclidean distance. The groups of Low surface brightness galaxies as well as the ones with high Surface Brightness were detected in such manner in the direction of the void. Edge\_on galaxies are not selected because of bias effects of the discrimination between stars and galaxies. Distribution of the concentration index (a parameter describing morphology of the object) was used to separate spiral and elliptical galaxies.

**Keywords** galaxies: fundamental parameters — galaxies: photometry — galaxies: surface photometry — methods: cluster analysis

---

\*This research has made use of the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

## 1 Introduction

The presence of voids in the distribution of galaxies has been discovered in early redshift surveys of galaxies - see e.g. [1,2]. Further studies show that the largest voids are those delineated by rich clusters and superclusters of galaxies [3,4]. The Hercules region has attracted the attention of astronomers since [5] discovered that the Hercules supercluster covers a large area of the sky north of celestial equator in the right ascension range between  $12^h < \alpha < 18^h$ . A collaboration between the Institute of Astronomy of Bulgarian Academy of Sciences and Max Plank Institute for Astronomy (Heidelberg, Germany) is devoted to investigate some known voids [6], including the void in Hercules. We present here the cluster analysis of the distribution of different parameters of faint galaxies in the direction of Hercules void - magnitudes, surface brightness, diameters, position angles, concentration indices etc.

## 2 Observations

The observations were done during the nights of 27 and 28 June 1991. The Ritchie camera of the 2m RCC telescope at National Astronomical Observatory Rozhen (Bulgaria) was used to get two plates centered on RA(2000):  $16^h 01^m 58^s$  and DEC:  $+17^\circ 57' 30''$  (No.1830) and RA:  $16^h 02^m 15^s$  and DEC:  $+17^\circ 51' 43''$  (No.1831). The unvignetted field is  $\sim 50$  sq.arc minutes for each plate and the common unvignetted field is  $\sim 1 \times 1$  sq.deg. The exposure time for each plate was 180 minutes, enough to detect objects fainter than 21 mag. A neutral wedge was exposed for 40 minutes on each plate after the main exposure. The seeing during the observations was very stable and varied only in the range of 1–1.5". ORWO plate ZU21 30x30 sq.cm and Schott filter GG 385 were used to realize the standard B-system. To calibrate these data, CCD frames in B and R were taken with the  $1024 \times 1024$  pixel "Photometrics" camera, attached to the prime focus on the 2m RCC telescope during the night on 28.06.2003 and with new CCD "VersArray:1300B" camera with  $1340 \times 1300$  pixel on 12.04.2005. Usually M92 standard fields [7] were taken twice during the same night.

## 3 Data Reduction

The two plates were scanned in the former Astronomical Institute, University of Muenster on the PDS 2020 GMplus scanning machine with  $25 \times 25 \mu m^2$  square slit and  $20 \mu m$  step, getting  $14400 \times 14400$  *pixel* with a scale of 0.25 arcsec/pixel. The linearization of the images was performed with the photometer wedge and the program INTENAIP from the Astrophysical Institute of Potsdam (AIP) package [8,9] added to the ESO MIDAS package.

We used a package of automatic procedures described in detail in [10] to detect objects above the background, to reject false detections, and to produce astrometry and integrated and surface photometry of detected galaxies. This

software was created to work with Sloan Digital Sky Survey data and was modified for our purpose.

**Detection Algorithm** We selected galaxies with angular sizes  $D \gtrsim 5$  arcsec or 15-18 pixels on our plates. Therefore we have chosen the angular size as the primary criterion to select candidate galaxies. The new programs that we created are based on the Kitt Peak International Spectral Survey [13] reduction package [8, 14] and some programs from the Astrophysical Institute of Potsdam (AIP) package in MIDAS<sup>1</sup> [9, 15, 16] for adaptive filtering and topological operations with masks.

The input data for our programs consist of: (1) – ZU 21 direct plate scanned images for the field in the  $B$  filter, (2) – SDSS\_like tables and (3) – tables with astrometric and photometric coefficients.

The processing steps of our programs can be conveniently described in terms of 8 discrete tasks or modules: (1) Density\_to\_intensity conversion of direct plate images; (2) Filtering of the images; (3) object detection; (4) integrated photometry; (5) creation of surface brightness profiles (SBP hereafter); (6) rejection of false detections; (7) fitting of surface brightness profiles; (8) calculations of total magnitudes.

We briefly outline these tasks below:

**Density\_to\_intensity conversion of direct plate images** The goal of this task is to create an intensity image from the scanned direct plates from the 2\_m RCC telescope. For this we use the INTENAIP program from AIP context, implemented in MIDAS - [17]. The step wedge, superexposed on the plates, was used to determine the density to intensity conversion.

**Integrated photometry** This task is intended to derive accurate magnitudes for the selected objects.

Throughout the paper a Hubble constant  $H_0 = 75 \text{ km s}^{-1} \text{ Mpc}^{-1}$  is adopted.

A code for fitting the sky background from the AIP package for adaptive filtering is used, which constructs the background within the masked regions containing the objects and their environments. The instrumental flux is transformed into the apparent magnitude in the standard B photometric system.

The total uncertainty of the instrumental magnitudes includes the uncertainty of the background determination  $\sigma_{sky}$  and the photon noise  $\sigma_{obj}$ , which can be calculated from Poisson statistics. Therefore the instrumental magnitude uncertainty  $\sigma_{instr}$  is computed for the cumulative flux  $CF$  in the total mask area  $N_{pix}$  with the value of the local background dispersion  $SKY$ :

$$\sigma_{instr} = 2.5 \cdot \log \left( 1 + \frac{\sqrt{\sigma_{obj}^2 + \sigma_{sky}^2}}{\sigma_{obj}^2} \right), \quad (1)$$

where  $\sigma_{obj} = \sqrt{CF}$  and  $\sigma_{sky} = \sqrt{N_{pix} \cdot SKY}$ .

<sup>1</sup>MIDAS stands for Munich Image Data Analysis System, the data reduction package of the European Southern Observatory.

**Creation of surface brightness profiles** The next task creates  $B$  SBPs of each detected galaxy using the background-subtracted images, which were used for the calculation of the integrated magnitudes. To simplify our programs and to make them more robust with regard to the creation of SBPs for galaxies with different morphologies, our software generates SBPs in the same way as SDSS\_PHOTO does [12], i.e., by measuring magnitudes in circular apertures. As the analysis in [18] showed, good agreement exists between an elliptical fit and circular aperture photometry. We selected **1 arcsec** step sizes as the standard aperture steps for our work. The uncertainty of each SB level is calculated in the same way as the uncertainty of the integrated magnitude (with equation (1)). After a SBP is created, the effective (or half light) radius  $R_{\text{eff}}$ , the effective surface brightness  $\mu_{\text{eff}}$  (the mean SB inside  $R_{\text{eff}}$ ), the radius of the region containing 90% of the integrated flux ( $R_{90}$ ), and the concentration index  $C=R_{90}/R_{\text{eff}}$  are calculated. We also use some programs from the AIP package to calculate the following parameters: the PA of the major axis for each filter, the axis ratio  $b/a$  for each filter, effective surface brightnesses  $\mu_{\text{eff}}^{\text{AIP}}$ , and effective radii  $R_{\text{eff}}^{\text{AIP}}$ . All these additional parameters are calculated using the multilevel mask approach [15, 16], in which each region of intensities of the studied object is labelled with a different mask.

Apart from SBPs, our task calculates curves of growth to compare the final magnitude with the integrated magnitude calculated with the previous task.

**Fitting of SBPs** This task produces fits to the  $B$  SBPs created earlier. It is well known that the fitting functions reflect formation and evolution processes. Unfortunately, these processes are still not well understood and the most commonly used functions are derived empirically. The fitting functions for elliptical galaxies and spiral galaxy bulges include the King model [19] and de Vaucouleurs law [20]. Exponentials [24] (Freeman, 1970) and inner-truncated exponentials [25] are usually used for the disk components of spiral galaxies. For application of a generalized version of the de Vaucouleurs profile ( $n^{1/n}$ ) see, e.g. [21] and [22].

Our programs follow the scheme to fit both the disk and bulge components with exponential SB profiles to measure their central surface brightness  $\mu_0$  and scale length  $\alpha$ . The equation for such profiles in logarithmic scale usually looks as follows [26]:

$$\mu(R) = \mu_0 + 1.086 \cdot (R/\alpha)^n, \quad (2)$$

where  $R$  is the distance along the axis and  $n$  is taken to be 1 for both bulge and disk.

We also incorporated the de Vaucouleurs profile ( $n = 1/4$ ) in our fit task in order to have the possibility (1) to automatically classify our observed SBPs by comparing them to different model light profiles and finding the model profile with the smallest  $\chi^2$  difference, and (2) to get fit parameters for bright elliptical galaxies for future applications of our programs.

**Total magnitudes** Our method of galaxy detection and the subsequent deter-

mination of their integrated magnitudes is essentially isophotal photometry with a sufficiently deep isophotal detection limit ( $\mu_{\text{lim}}$ ). It is well-known that unlike for point sources, the fraction of light of a galaxy contained within the limiting isophote  $\mu_{\text{lim}}$  is a function of several parameters such as central surface brightness  $\mu_0$ , redshift, point-spread function (PSF), and cosmological dimming (see, e.g., [27]). In our case the minimal diameter of the selected galaxies of 5 arcsec is larger than the PSF, since the typical seeing in our images is  $\sim 1.5 \text{ arcsec}$  - for details see [10].

If one wishes to avoid all the systematic bias in the determination of magnitudes, it is possible to calculate total magnitudes  $m_{\text{tot}}$  integrated out to infinity (for equation (2) with  $n = 1$ ) using the equation

$$m_{\text{tot}}^d = \mu_0 - 5 \cdot \log_{10}(\alpha) - 2.5 \cdot \log_{10}(2\pi) \quad (3)$$

Here  $\alpha$  is in arcsec and index ‘d’ refers to a purely exponential disk. The mean difference  $\Delta B(\text{Apparent} - \text{Total})$  is  $0.04 \pm 0.04$ , which is well consistent with the mean value of the total error for this band,  $\sigma_{\text{tot}}(B) = 0.08$ . Our total magnitudes are derived only for the  $B$  filter, since we confine our SBP determinations to this filter. As was shown in [10], our apparent magnitudes are often a sufficiently good approximation of the total magnitudes.

**Astrometry and Photometry Results** The output data consist of a MIDAS table that contains all information about the detected galaxies: accurate positions ( $X, Y, \alpha(2000), \delta(2000)$ ), total fluxes and integrated magnitudes with their uncertainties, effective radii ( $R_{\text{eff}}$ ; in arcsec), effective surface brightnesses ( $\mu_{\text{eff}}$ ; in  $\text{mag arcsec}^{-2}$ ), radii of the regions containing 50% and 90% of the integrated flux ( $R_{50}, R_{90}$ ; in arcsec), concentration indices  $C = R_{90}/R_{50}$ , position angle (PA), axial ratio ( $b/a$ ). exponential fits for the central surface brightnesses ( $\mu_{E,0}$ ; in  $\text{mag arcsec}^{-2}$ ) and total magnitudes, integrated out to  $25 \text{ mag arcsec}^{-2}$  with their uncertainties. The distribution of  $\mu_0(B)$  for all studied galaxies fall in the range of  $16 < \mu_0(B) < 24 \text{ mag arcsec}^{-2}$ .

The data for all 1851 galaxies in the field – coordinates (2000), aperture and surface photometry, position angles, diameters, axis ratio and concentration are available at <http://www.astro.bas.bg/~petrov/papers/files/v16full.pdf> or upon request from [petrov@astro.bas.bg](mailto:petrov@astro.bas.bg) in FITS\_tbl format [17].

**Galaxy Number Counts.** The number counts of galaxies as a function of magnitude is one of the classical cosmological tests. The result for our data is plotted on the right panel of Fig. 1. Galaxy number counts are shown in 0.5 mag bins. The errors bars correspond to Poisson noise. The line in Fig. 1 shows a fit to the galaxy counts-magnitude relation expected in a homogeneous universe assuming Euclidian geometry for three-dimensional space. The observed galaxy counts are quite consistent with this line for  $17^m \leq B \leq 20^m$  and even fainter up to  $B=20.5 \text{ mag}$ . It means that we have complete data up to this magnitude. In our data we find big excess of bright galaxies ( $B < 17.0^m$ ).

#### 4 Cluster analysis algorithm

We used multivariate statistics to determine new or unknown objective laws or regularities in the raw data. These statistics are briefly reviewed in an astronomical context by [30], [31], [32] and are more thoroughly described by [33], [34] and [35]. Many monographs and WEB based courses presenting multivariate statistics are available, such as [36], [37], [38], [39]. A popular data mining technique involves the construction of decision trees, based on decision rules which define a partition of a dataset by splitting depending on key variables. Cluster analysis is an exploratory data analysis tool for solving classification problems. Its objective is to sort cases into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Each cluster thus describes the class to which its members belong. Cluster analysis partitions a large dataset into homogeneous subsets by grouping closely related cases into tight clusters. These core clusters, which map the main density, can then be used to re-construct the underlying model, for segmentation analysis and for exemplification.

**What is Clustering?** A large number of clustering definitions can be found in the literature, from simple to elaborate. The simplest definition is shared among all and includes one fundamental concept: the grouping together of similar data items into clusters. Clustering is an important problem that must often be solved as a part of more complicated tasks in pattern recognition, image analysis and other fields of science and engineering. Clustering is also needed for designing a codebook in vector quantization. The clustering problem contains two subproblems: (i) How many clusters there are in the data set and (ii) Find the location of the clusters.

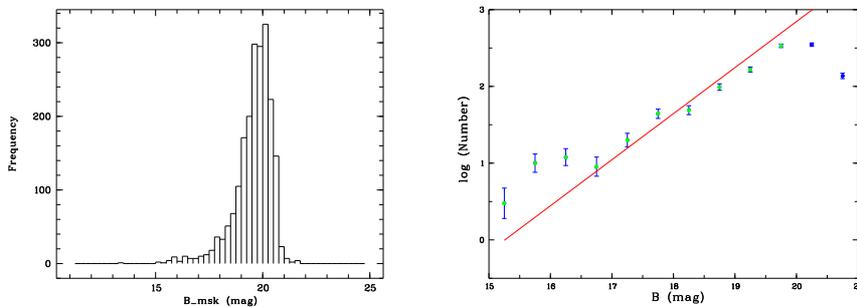


Figure 1. Left panel: Distribution of apparent magnitudes for all galaxies detected with our programs. Right panel: Number counts of all galaxies detected with our programs as a function of apparent magnitude. The errors bars on the galaxy counts are Poissonian. The line shows the count-magnitude relation expected for a homogeneous galaxy distribution in a universe with "Euclidean" geometry:  $N(B) = A_B \cdot 10^{0.6B}$ .

**Determining the number of clusters.** Determination of the number of clusters in data requires that we have both an algorithm that can seek for the correct number and a criterion that is capable of recognizing the correct number of clusters. The simplest algorithm is to use an existing algorithm for fixed number of clusters in a loop and select the best solution with some criterium. The rest is the usual trial-and-error approach.

**Algorithms in cluster analysis.** *Hierarchical clustering methods* do not require a priori knowledge of the number of groups. Two general methods of hierarchical clustering methods are available: *divisive* and *agglomerative*. *Divisive* methods start by putting all data vectors in a single cluster. New clusters are created by dividing existing ones. This approach involves two main design problems: which cluster to divide, and how the division is performed. The division can be made along a selected dimension of the vector space as in Median Cut algorithm or along the principal axis - [40], [41]. *Agglomerative clustering* generate the clustering by a sequence of merge operations. The clustering starts by initializing each data vector as its own cluster. Two clusters are merged at each step and the process is repeated until the desired number of clusters is obtained. Ward's method (1963) selects the cluster pair to be merged so that it increases the given objective function value least. Among the most used variations of the hierarchical clustering based on different distance measures are:

1. *Average linkage clustering.* The dissimilarity between clusters is calculated using average values. The average distance is calculated from the distance between each point in a cluster and all other points in another cluster. The two clusters with the lowest average distance are joined together to form the new-cluster.

2. *Centroid linkage clustering.* This variation uses the group centroid as the average. The centroid is dened as the center of a cloud of points.

3. *Complete linkage clustering (Maximum or Furthest-Neighbor Method).* The dissimilarity between 2 groups is equal to the greatest dissimilarity between a member of cluster  $i$  and a member of cluster  $j$ . This method tends to produce very tight clusters of similar cases.

4. *Single linkage clustering (Minimum or Nearest-Neighbor Method).* The dissimilarity between 2 clusters is the minimum dissimilarity between members of the two clusters. This method produces long chains which form loose, straggly clusters.

5. *Ward's Method.* Cluster membership is assigned by calculating the total sum of squared deviations from the mean of a cluster. The criterion for fusion is that it should produce the smallest possible increase in the error sum of squares.

Several *distance measures* are available and implemented within different statistical packages: SEUCLID squared euclidean distance. For any pair of cases the measure is  $\text{Distance}(x,y) = \sum (x - y)^2$ . Therefore SEUCLID is the sum of the squared differences between scores for two cases on all variables, i.e. the squared length of the hypotenuse. EUCLID - square root of SEUCLID. Note that for two variables this is the normal Pythagoras theorem. CHEBYCHEV -

this is a distance measure. In this the absolute maximum difference between variable scores is used.  $\text{Distance}(x,y) = \text{MAX } |x - y|$ . BLOCK -city block or Manhattan distance. So named because in most American cities it is not possible to go directly between 2 points. A route which follows the regular grid of roads is used.  $\text{Distance}(x,y) = \sum |x - y|$ .

*Iterative methods* are usually based on the *K-means algorithm* (Forgy 1965), which applies two optimization steps iteratively - [42], [37], [43]: (i) calculate optimal partition for the given codebook, and (ii) calculate new codebook as the cluster centroids. A popular method of classification is *k-means analysis*, which partitions a set of cases into  $k$  clusters so as to minimize the error or sum of squared distances of the cases about the cluster means. An alternative method with a more rigorous mathematical foundation is *k-means partitioning* (this we use here). It finds the combination of  $k$  groups that minimizes intragroup variance. However, it is necessary to specify  $k$  in advance. However, k-means analysis is usually only implemented with quantitative variables. K-means cluster analysis uses Euclidian or squared Euclidean distances. Initial cluster centers are chosen in a first pass of the data, then each additional iteration groups observations based on nearest Euclidian distance to the mean of the cluster. Thus cluster centers change at each pass. The process continues until cluster means do not shift more than a given cut-off value or the iteration limit is reached. (i) K-means cluster analysis assumes a large sample (ex.,  $\geq 200$ ). (ii) K-means cluster analysis usually generates different solutions, depending on the sequence of observations in the dataset.

Cluster k-Means is suitable for very large data sets. It will handle mixed data types that can contain missing values, contiguity constraints and allow for differential case weights and differential variable weights. The following criterion functions can be optimized: (i)Euclidean Sum of Squares, (ii)Euclidean Distance, (iii)City Block Distance, (iiii)Maximum Distance etc.

Euclidean Sum of Squares is the recommended criterion function because an exact relocation test has been implemented and hence k-means must converge if allowed sufficient iterations.

In the next steps we use mainly K-Means and Ward's methods results and the centroids of the clusters from all mentioned above methods are included in the final Table 6.

## 5 Clustering in the distribution of faint galaxies

THE HERCULES SUPERCLUSTER ([44] is made up of three Abell clusters A2151, A2147, and A2152, with a connection towards the A2197/A2199 supercluster [1]. The structures of the Hercules supercluster have been analysed with the non-hierarchical descendent taxonomical method, using  $\alpha$ ,  $\delta$  and  $vt$  as active coordinates, and morphological type, position angle and apparent magnitude as control parameters.

Here we present cluster analysis of faint galaxies in the direction of the Hercules void.

**Morphological types and concentration index.** Our photometric programs calculate a number of global morphological parameters for every galaxy. Some of these may be useful for morphological galaxy classifications (see earlier discussions by, e.g., [45], [46], [47], [11]). A particularly useful parameter is the concentration index ( $C$  hereafter), defined as the ratio of the radii containing 90% and 50% of a galaxy's light. For the classical de Vaucouleurs profile - [48],  $C$  is  $\sim 5.5$ , and for pure exponential disks,  $C \sim 2.3$ . These values are valid for the idealized seeing-free case, which we are approximating due to the limiting angular size of PSF that we impose in this work. On the Fig. 2 - dependence of effective surface brightness from Index of concentration is presented. While by K-means and Ward's methods the groups of E/S0 galaxies is well outlined ( $C \sim 5.5$ ) SB-Sd galaxies ( $C \sim 2.3$ ) even grouped well by K-means clustering are mixed amongst different surface brightness.

**Surface brightness of galaxies.** Figure 3 represents the results from cluster analysis of total Magnitudes and Large Diameters against Effective Surface Brightness distribution (a) and b) respectively) and Axis Ratios against Integrated magnitudes (c) distributions. Surface brightness of galaxies after correction for the absorption in the Galaxy (see [49] for details) is determined by equation (4):

$$B = m_p - 0.25 \operatorname{cosec}|bII| + 2.5Lg\pi \frac{ab}{4}. \quad (4)$$

Using relation equation (5)

$$SB = B + 0.22(a/b) + 0.73. \quad (5)$$

one get the surface brightness in Holmberg system [50]. As in [49] High Surface Brightness Galaxies (HSBG) are the galaxies with  $SB \leq 22.0$

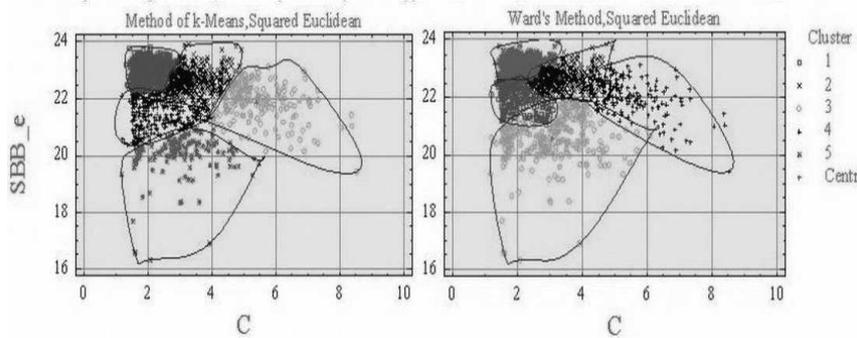


Figure 2. Cluster analysis of: Surface Brightness against Index of Concentration distribution

mag/sqr.arcsec. From the Eq.4 and Eq.5 faint galaxies (i.e. with bigger magnitudes) with large diameters could be a good candidates for HSBG group. This is very well demonstrated from the k-Means and Ward's clustering on Fig. 3\_a and b) and centroids of the clusters in the main Table. The two methods outliers a well populated group of galaxies with magnitudes ca. 17.5 and diamaters ca. 32 - 36 arcsec.

**Low Surface Brightness galaxies (LSBG).** Low-surface-brightness (LSB) galaxies are one of the main constituents of the realm of galaxies. They are usually defined as objects with a blue central surface brightness  $\mu_0(B)$  significantly fainter than the Freeman value of 21.65 mag arcsec<sup>-2</sup> [24]. However, the threshold value of  $\mu_0(B)$  to classify galaxies as LSB galaxies varies in the literature

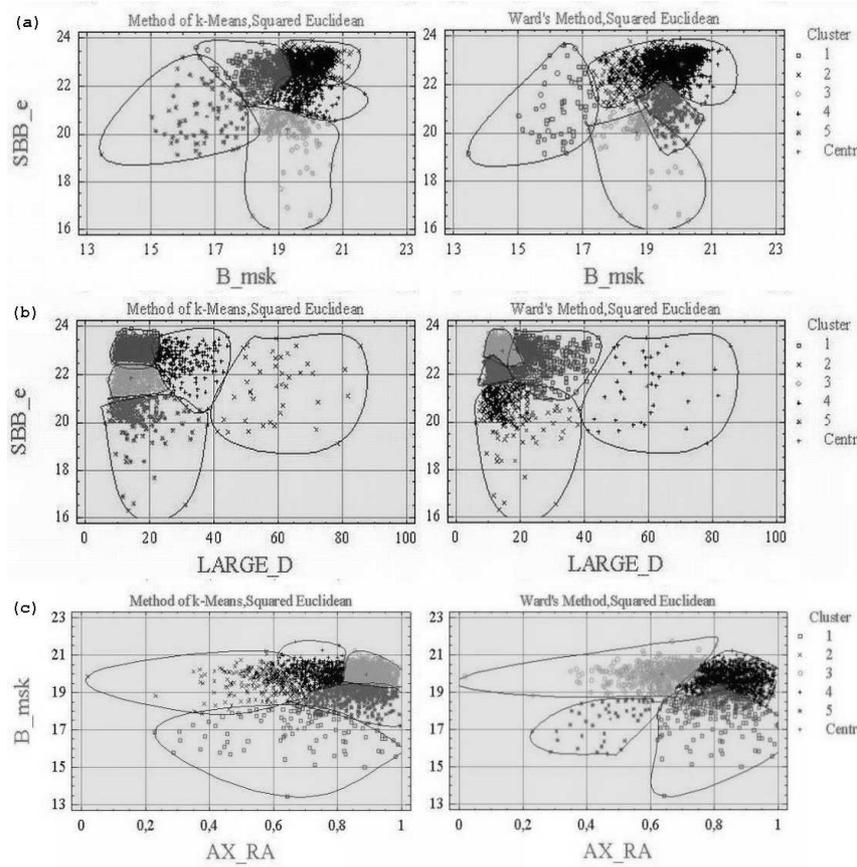


Figure 3. Cluster analysis of: a) Surface Brightnes against Magnitudes b) Surface Brightnes against Large Diameters and c) Magnitudes against Axis Ratio distribution

from  $\mu_0(B) \geq 23.0 \text{ mag arcsec}^{-2}$  [51] to  $\mu_0(B) \geq 22.0 \text{ mag arcsec}^{-2}$  [52]. There are many topics for which the knowledge of the properties of the LSB galaxy population is crucial. They include the following: a) the galaxy luminosity function, especially at its faint end, see e.g. [27, 53, 54], which in turn is related to the understanding of the primordial power spectrum of density fluctuations [55, e.g.]; b) the spatial distribution of lower-mass galaxies, which allows us to check the predictions of cold dark matter cosmology for large-scale-structure formation - see e.g. [56]; c) the physics of star formation at low gas surface densities - [57–59]; d) the role of interactions in galaxy evolution; and many others.

The detection of LSB galaxies is difficult due to their intrinsically low global luminosities and their characteristic low surface brightness. Despite more than 20 years of LSB galaxy studies, their census remains incomplete.

Past LSB surveys were usually either large area photographic surveys or deep CCD surveys with small area - for a nice review see [60].

An interesting aspect of this study is the identification of a substantial number of luminous distant galaxies with Low surface Brightness. As above bright galaxies (i.e. with lower magnitudes) with large diameters could be a good candidates for LSBG group. This again is very well demonstrated from the  $k$ -Means and Ward's clustering on Fig. 3.a and b) and centroids of the clusters in the main Table. The two methods outliers a group of ca.35 galaxies with magnitudes ca. 16.3 and diameters ca.60 arcsec!. These galaxies could fall into the category of the so-called giant LSB galaxies, or 'cousins' of Malin 1 [61], but the distances to galaxies are needed to say this definitely. In Fig. 3.b) the relations between the effective surface brightnesses  $\mu_{\text{eff}}$  and large diameters are plotted and clustered. Some LSBGs could be selected there - bright galaxies with  $B \sim 15.5-18$  mag or such ones with large diameters ( $> 40$  arcsec) and  $SB > 22$  mag/sqr.sec. Currently, only ca. 25 galaxies of this type are known. As [62] emphasize, giant LSB galaxies are quite enigmatic from the point of view that they normally formed their spheroidal component, but no conspicuous stellar disk ever formed around their bulge. Improved statistics for these objects will lead to a better understanding of the relationship between their fundamental parameters and the conditions/processes that led to their formation. The high detection rate of giant LSB galaxy candidates with our programs promises substantial progress for future systematic studies.

**Distribution of position angles of galaxies.** On the Fig. 4.b) histogram of the determined position angles are presented. Excluding  $\pm 90^\circ$  PA, where some misdetermination could take place, all direction of the orientation are almost equal populated - i.e. no preferable orientation of the galaxies in the direction of Hercules void is defined.

As one can see on Fig. 5.a), galaxies with large angular diameters tend to define a group with more compact position angles. This could be selection effect as for larger galaxies the accuracy of the position angles determination is better. Nevertheless, radial velocities for these "large" galaxies will clear the question.

**Edge-on galaxies** It is worth noting that our selection criteria produce a bias against edge-on galaxies. Since we separate objects with the total area within the limiting isophote greater than some minimal threshold, the lower limit for their major-axis diameter of 5 arcsec is valid only for round, face-on galaxies. The histogram of the distribution of Axis\_Ratio determined is shown on Fig. 4.c. For elongated galaxies, the lower limit for this parameter varies as

$$D_{\text{lim}} = \frac{5 \text{arcsec}}{\sqrt{b/a}} \quad (6)$$

Thus, for the most elongated galaxies ( $b/a \sim 0.1$ ) we will detect all objects with major-axis diameter larger than 50 arcsec. In the same time as one can see at Fig. 3.c) and Fig. 5.c), galaxies with different magnitudes diameters tends to be grouped acc. to axis ratio.

## 6 Conclusions

In the Table the results from cluster analysis of almost all the parameters are presented. The distance metric is squared euclidean and the different methods of clustering are tested as follow: centroid, furthest neighbor, group, *k-Means*, median, nearest neighbor and *Ward's* methods. The results from *k-Means* and *Ward's* method are in *italic* in the table.

1. Astrometric and photometric parameters for 1851 galaxies are determined
2. We found big excess of bright galaxies ( $B < 17.0^m$ ) co. to the galaxy counts-magnitude relation expected in a homogeneous universe assuming Euclidian geometry for three-dimensional space.
3. Using cluster analysis method for morphological classification of galaxies the groups of E/S0 galaxies is well outlined ( $C \sim 5.5$ ) SB-Sd galaxies ( $C \sim 2.3$ ) even grouped well by K-means clustering are mixed amongst different surface brightness.

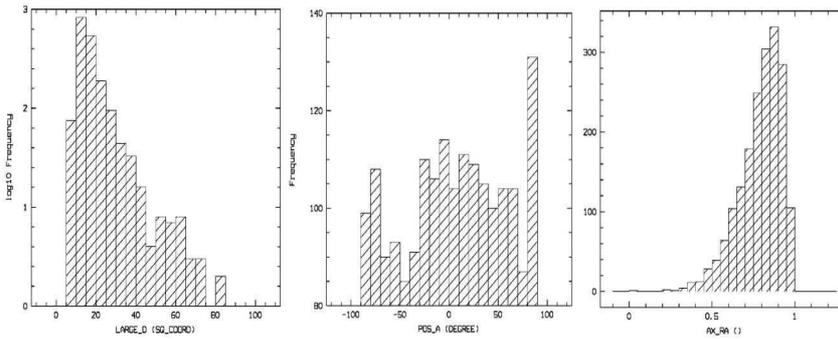


Figure 4. Histogram of the Large diameters (a), Position Angle (b) and Axis Ratio (c.)

4. A well populated group of galaxies with magnitudes ca. 17.5 and diameters ca. 32 - 36 arcsec - i.e. faint galaxies with large diameters forms HSBG group.
5. The method outliers a group of ca.35 galaxies probably falling into the category of the so-called giant LSB galaxies.
6. No preferable orientation of the galaxies in the direction of Hercules void is defined.
7. Galaxies with different magnitudes and diameters tends to be grouped acc. to axis ratio.

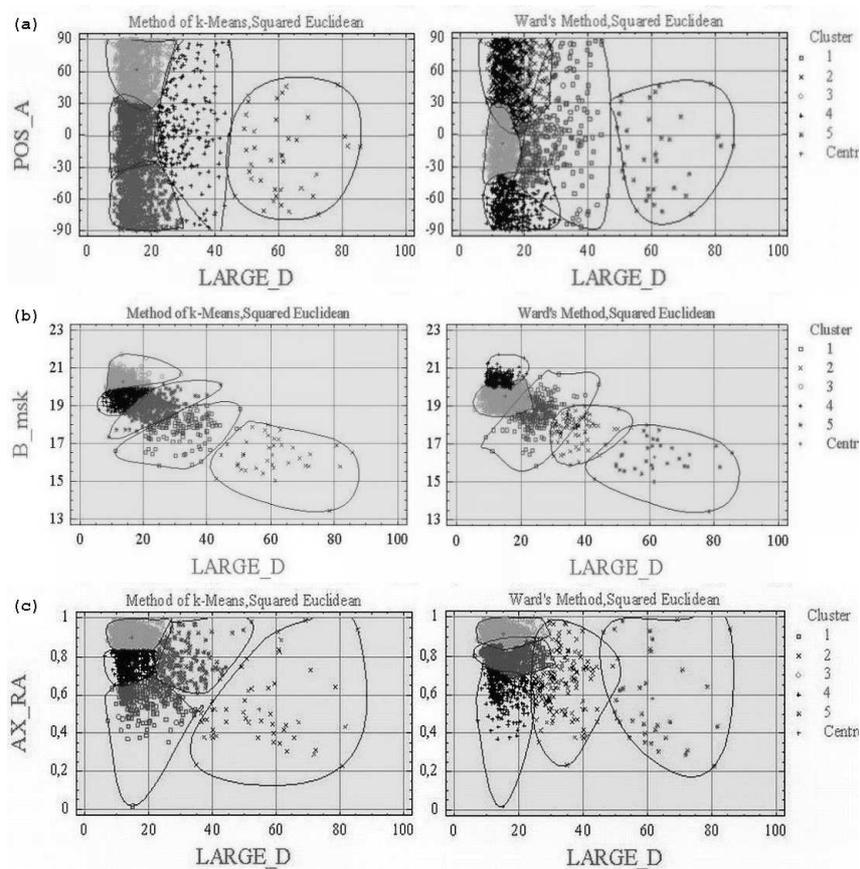


Figure 5. Cluster analysis of Axis Ratio, Magnitudes and Position Angle against the Large Diameters distributions

## References

- [1] G. Chincarini, H.J. Rood and L. Thompson (1981) *ApJ*, 249, L47
- [2] Gregory, S.A., Thompson, L.A. 1982, *Sci. Am.*, 246, 106.
- [3] J. Oort (1983) *Ann.Rev.A&Ap.*, 21, 373.
- [4] H.J. Rood (1988) *Ann.Rev.A&Ap.*, 26, 631.
- [5] H. Shapley (1934) *MNRAS*, 94, 53.
- [6] G. Petrov, B. Kovachev and A. Strigatchev (1993) *Astron.Ges. Abstr.Ser.*, 9, 46.
- [7] C.A.Christian, M. Adams, J. Barnes, H. Butcher, D. Hayes, J. Mould and M. Siegel (1985) *PASP*, 97, 363.
- [8] A.Y. Kniazev (1997) *Special Astrophys. Obs.*, Nizhnij Arkhyz, Ph.D. Thesis. (<http://precise.sao.ru>)
- [9] H. Lorenz, G.M. Richter, M. Capaccioli and G. Longo (1993) *A&A*, 277, 321.
- [10] A.Y. Kniazev, E.K. Grebel, S.A. Pustilnik, A.G. Pramskij, T.F. Kniazeva, F. Prada and D. Harbeck (2003) *AJ*, 127, 704.
- [11] K. Shimasaku, M. Fukugita, M. Doi and 14 couthors (2001) *AJ*, 122, 1238.
- [12] Stoughton, C., Lupton, R.H., Bernardi, M. et al. 2001, *aj*, 123, 485.
- [13] Salzer, J.J., Gronwall, C., Lipovetsky, V.A., Kniazev, A., Moody, J.W., et al., 2000, *aj*, 120, 80.
- [14] Kniazev, A.Y., Salzer, J., Lipovetsky, V.A., et al. 1998, in *IAU Symp. 179, New Horizons from Multi-Wavelength Sky Surveys*, ed. B.J.McLean, D.A. Golombek, J.J.E. Hayes, & H.E. Payne (Dordrecht: Kluwer), 302.
- [15] Vennik, J., Hopp, U., Kovachev, B., Kuhn, B., & Elsasser, H. 1996, *A&AS*, 117, 261.
- [16] Vennik, J., Hopp, U., & Popescu, C.2000, *A&AS*, 142, 399.
- [17] G. Petrov, A. Kniazev and J. Fried (2004) In.: *Balkan Astronomical Meeting, Proc. Of the Conference 14-18.06.2004 Rozhen, Bulgaria*, 108 (E.form at: [www.space.bas.bg](http://www.space.bas.bg)).
- [18] Bremnes, T., Binggeli, B., & Prugniel, P. 1999, *aaps*, 137, 337.
- [19] King, I.R. 1966, *aj*, 71, 64.
- [20] de Vaucouleurs, G. 1953, *mnras*, 113, 134.
- [21] Andredakis, Y.C., Peletier, R.F., & Balcells, M. 1995, *mnras*, 275, 874.
- [22] Andredakis, Y.C., & Sanders, R.H. 1994, *mnras*, 267, 283.
- [23] Beijersbergen, M., de Block, W., & van der Hulst, J. 1999, *aap*, 351, 903.
- [24] Freeman, K.C. 1970, *apj*, 160, 811.
- [25] Kormendy, J., 1977, *apj*, 217, 406.
- [26] Sersic, J.L. 1968, *Atlas de Galaxes Australes (Cordoba, Argentina: Observatorio Astronomico)*.
- [27] Dalcanton, J.J. 1998, *apj*, 495, 251.
- [28] Allen, R.J., & Shu, F.H. 1979, *apj*, 277, 67.
- [29] Impey, C.D., Bothun, G.D.,& Malin, D. 1988, *apj*, 330, 634.
- [30] J. Babu and E. Feigelson (1996) *Astrostatistics*, Chapman & Hill
- [31] E. Feigelson and J. Babu (1992) *Statistical chalenges in astronomy*, Penn State Univ
- [32] E. Feigelson and J. Babu (Eds.) (2003) *Statistical chalenges in astronomy*, Springer Verlag

- [33] F. Murtagh and A. Heck (1987) *Astr.Astroph.Suppl.Ser.*, 68, 113.
- [34] F. Murtagh and A. Heck (1987) Kluwer Academic Publishers, Dordrecht "Multivariate data analysis"
- [35] F. Murtagh (2002) Multivariate data analysis, WEB course
- [36] D.W. Stockburger (2000) Atomic Dog Publishing "Introductory Statistics: Concepts, Models, and Applications"
- [37] G. Fung (2001) A Comprehensive Overview of Basic Clustering Algorithms
- [38] Franti (2004) Algorithms in cluster analysis, WEB course
- [39] I. Karkkainen and P. Franti (2002) Advanced Concepts for Intelligent Vision Systems (ACIVS'2002), Gent, Belgium, 136-143, September 2002.
- [40] D. Wishart (1986) in Gaul, Shader (eds.) Classification as a tool of research, North Holland, Amsterdam, 453.
- [41] D. Wishart (1998) *Computing Science and Statistics*, 30, 257.
- [42] M. Xu and P. Franti (2004) *IEEE Int. Conf. on Image Processing (ICIP'04)*, Singapore, vol. 3, 3503.
- [43] D. Wishart (2001) GfKl, "k-Means clustering with outlier detection, mixed variables and missing values"
- [44] M. Moles, A. del Olmo and J. Perea (1985) *MNRAS*, 213, 365.
- [45] S.M. Kent (1985) *ApJS*, 59.
- [46] K. Kodaira, M. Watanabe and S. Okamura (1986) *ApJS*, 62, 703.
- [47] Fioc and Rocca-Volmerange (1999) *A&Ap*, 351, 869.
- [48] I. Strateva, Z. Ivezić, G. Knapp et al. (2001) *AJ*, 122, 1861.
- [49] M.A. Arakelian (1975) *Contrib.Bjurak.Obs.*, 47, 3
- [50] E. Holmberg (1954) *Medd. Lund. Astron. Obs.*, ser.II, No.136.
- [51] Impey, C.D., & Bothun, G.D. 1997, *araa*, 35, 267.
- [52] Impey, C.D., Burkholder, V., & Sprayberry, D. 2001, *aj*, 122, 2341.
- [53] Trentham, N. & Tully, R. B. 2002, *mnras*, 335, 712 .
- [54] Cross, N., & Driver, S. 2002, *mnras*, 329, 579.
- [55] Ostriker, J. 1993, *araa*, 31, 589.
- [56] Peebles, P.J.E. 2001, *apj*, 557, 495.
- [57] van Zee, L., Haynes, M. P., Salzer, J. J., & Broeils, A. H. 1997, *aj*, 113, 1618.
- [58] Ferguson, A.M.N., Wyse, R.F.G., Gallagher, J.S., & Hunter, D.A. 1998, *apjl*, 506, L19.
- [59] Noguchi, M. 2001, *apj*, 555, 289.
- [60] Dalcanton, J.J., Spergel, D.N., Gunn, J.E., Smith, M., & Schneider, D.P. 1997, *aj*, 114, 635.
- [61] G. Bothun, C. Impey, D. Malin and J. Mould (1987) *AJ*, 94, 23.
- [62] Bothun, G., Impey, C., & McGaugh, S. 1997, *pasp*, 109, 745.

Analysis Summary. Distance Metric: Squared Euclidean, Number of galaxies: 1851

Cl	Centroid			Furth.Neighbor			Group Average			k-Means			Median			Near.Neighbor			Ward's					
	LAR_D	AX_RA	SB_B	LAR_D	AX_RA	SB_B	LAR_D	AX_RA	SB_B	LAR_D	AX_RA	SB_B	LAR_D	AX_RA	SB_B	LAR_D	AX_RA	SB_B	LAR_D	AX_RA	SB_B			
1	1737	15.88	0.80	120	33.71	0.76	1792	16.48	0.86	249	18.37	0.58	1672	16.45	0.82	1846	17.48	0.79	754	15.74	0.80			
2	91	39.29	0.68	1672	15.38	0.80	13	61.27	0.86	50	54.75	0.53	144	50.63	0.53	1	85.68	0.94	159	32.27	0.66			
3	78	37	0.77	8	68.54	0.85	41	47.44	0.46	70	14.64	0.90	18	59.31	0.79	2	74.71	0.68	598	14.92	0.91			
4	19	62.45	0.40	35	35.82	0.45	1	14.92	0.02	603	14.56	0.76	16	64.69	0.38	1	14.92	0.02	305	14.81	0.62			
5	1	14.92	0.02	16	64.69	0.38	4	76.80	0.31	169	30.08	0.77	1	14.92	0.02	1	81.88	0.43	35	61.55	0.58			
1	14	16.61	22.66	566	18.83	22.59	61	16.58	21.24	391	18.70	22.61	430	18.33	22.24	1846	19.52	22.41	53	16.36	21.04			
2	1783	19.61	22.47	148	19.40	20.24	1600	19.70	22.67	947	20.04	22.98	452	19.59	21.36	2	20.15	16.45	493	18.84	22.65			
3	43	16.58	20.19	53	16.69	20.38	179	18.96	20.64	373	19.26	20.19	964	20.02	22.98	1	18.18	16.54	83	18.69	20.02			
4	10	19.63	17.64	1083	20.04	22.69	10	19.63	17.64	306	19.71	21.68	4	19.44	16.59	1	13.44	19.12	979	20.08	22.86			
5	1	13.44	19.12	1	13.44	19.12	1	13.44	19.12	70	16.74	20.72	1	13.44	19.12	1	19.28	16.91	243	19.61	21.14			
1	Memb	LAR_D	POS_A	Memb	LAR_D	POS_A	Memb	LAR_D	POS_A	Memb	LAR_D	POS_A	Memb	LAR_D	POS_A	Memb	LAR_D	POS_A	Memb	LAR_D	POS_A			
2	56	36.46	34.55	972	14.39	36.13	1166	14.66	27.12	37	60.92	-18.2	121	35.79	-9.52	28	57.09	-15.1	687	15.77	84.67			
3	33	58.40	-21.2	618	15.72	-55.0	488	16.68	-62.1	554	15.30	61.84	1163	15.05	-24.3	3	82.79	-7.27	439	13.06	-8.09			
4	4	81.72	6.55	28	60.05	-26.3	32	58.78	-20.0	212	30.16	6.98	21	62.82	-27.2	4	70.62	-54.7	434	15.47	-66.5			
5	22	36.87	-54.1	4	81.72	6.55	4	81.72	6.55	512	15.43	-61.7	4	81.72	6.55	1	78.49	48.0	36	61.33	-17.1			
1	Memb	LAR_D	SB_B	e	Memb	LAR_D	SB_B	e	Memb	LAR_D	SB_B	e	Memb	LAR_D	SB_B	e	Memb	LAR_D	SB_B	e	Memb	LAR_D	SB_B	e
2	27	43.25	20.28	373	13.15	21.10	15	51.63	20.40	40	59.57	21.50	46	50.10	21.20	1	85.68	23.20	281	15.09	20.61			
3	21	62.06	22.53	27	43.25	20.28	17	64.80	22.44	379	14.15	21.86	2	83.24	22.70	3	13.23	16.60	796	14.24	22.99			
4	10	15.90	17.64	25	64.33	22.15	4	17.68	16.58	231	29.17	22.65	10	15.90	17.64	1	31.04	16.54	40	59.57	21.50			
5	4	76.22	20.19	6	18.39	17.07	4	76.22	20.19	190	15.91	20.27	4	76.22	20.19	2	75.24	19.40	281	12.35	22.15			
1	Memb	C	SB_B	e	Memb	C	SB_B	e	Memb	C	SB_B	e	Memb	C	SB_B	e	Memb	C	SB_B	e	Memb	C	SB_B	e
2	1574	2.64	22.69	267	2.27	21.15	1371	2.45	22.83	894	2.06	23.00	1618	2.52	22.55	1845	2.79	22.41	275	2.03	22.07			
3	189	2.55	20.40	1408	2.62	22.82	310	2.84	20.75	401	3.44	22.65	64	2.46	19.59	1	1.51	17.70	424	3.28	22.68			
4	4	2.29	16.59	113	5.80	21.68	4	2.29	16.59	243	2.27	21.46	32	5.73	20.37	1	3.92	16.91	167	5.48	22.04			
5	4	8.34	20.71	36	4.27	20.09	1	8.53	19.42	129	3.08	19.97	4	8.34	20.71	1	8.53	19.42	720	2.06	23.11			
1	Memb	Lar_D	SB_B	Memb	Lar_D	SB_B	Memb	Lar_D	SB_B	Memb	Lar_D	SB_B	Memb	Lar_D	SB_B	Memb	Lar_D	SB_B	Memb	Lar_D	SB_B	Memb	Lar_D	SB_B
2	127	31.43	17.77	149	24.78	18.02	13	30.28	16.42	123	32.23	17.79	127	30.61	17.71	1845	17.44	19.53	230	24.75	18.67			
3	1686	15.61	19.72	147	31.40	18.80	1798	16.59	19.61	37	60.59	16.88	15.71	19.72	3	82.79	16.40	84	36.44	17.83				
4	2	47.38	19.48	35	60.00	16.40	8	74.84	16.18	661	14.51	19.48	3	82.79	16.40	1	43.07	15.15	528	13.33	20.34			
5	1	78.49	13.44	1	78.49	13.44	1	78.49	13.44	304	22.75	18.93	1	78.49	13.44	1	78.49	13.44	34	61.56	16.30			
1	Memb	AX_RA	SB_B	Memb	AX_RA	SB_B	Memb	AX_RA	SB_B	Memb	AX_RA	SB_B	Memb	AX_RA	SB_B	Memb	AX_RA	SB_B	Memb	AX_RA	SB_B	Memb	AX_RA	SB_B
2	71	0.83	17.00	19	0.82	15.77	43	0.79	16.56	99	0.68	17.03	1538	0.79	19.44	1847	0.79	19.52	291	0.81	18.21			
3	1728	0.80	19.68	563	0.83	18.79	1792	0.80	19.62	236	0.57	19.58	271	0.84	20.42	1	0.23	18.81	308	0.91	19.27			
4	2	0.17	17.69	329	0.60	19.73	2	0.13	19.33	626	0.89	19.95	20	0.86	15.86	1	0.02	19.85	432	0.64	19.95			
5	1	0.13	19.33	924	0.84	20.02	13	0.40	16.28	492	0.74	20.01	2	0.13	19.33	1	0.89	15.02	775	0.84	19.98			
5	1	0.64	13.44	16	0.40	16.48	1	0.64	13.44	398	0.86	18.80	20	0.44	16.51	1	0.64	13.44	45	0.47	17.58			

Figure 6. Results from Cluster analysis - Population and Coordinates of the centers of clusters determined