Virgo Cluster Membership Through Multivariate Normal Fitting And Hierarchical Clustering

Dmitrii Gudin[1]

[1] *Department of Mathematics, University of Maryland*

ABSTRACT

The problem of separating members of galaxy clusters from those of its surrounding intracluster medium (ICM) has remained one of the biggest challenges in the studies of the large scale structure of the Universe. Despite extensive theoretical developments, few conclusive observational results have been obtained. In this work, I develop an algorithm combining multivariate normal modelling with hierarchical clustering and apply it to the data set containing a large number of likely members of the Virgo galaxy cluster in order to split those members from unrelated galaxies. As the analysis is performed solely on the direct kinematic variables that are not quasi-conservative, both approaches produce inconclusive results, highlighting the difficulties in studying galaxy clusters. [a)]

1. INTRODUCTION

One of the challenges in extragalactic astronomy (i.e. the study of the Universe outside of the Milky Way) is the difficulty of studying the intracluster medium (ICM). According to the currently widely accepted Lambda-CDM model, the Universe is populated by a large number of galaxy clusters continuously accreting new galaxies, resulting in a top-down hierarchical structure. However, partially due to the technological limitations, most observational studies of nearby galaxy clusters have been concentrating on their densest central regions, while the accretion processes occur at the edges of the clusters. Observations of those edges until recently were made difficult by lack of wide-field CCD cameras and spectroscopic information (Ferrarese et al. 2012; Pak et al. 2014). As summarized in Kim et al. (2014), recent availability of large scale spectroscopic surveys such as the Sloan Digital Sky Survey (SDSS, York et al. 2000; Blanton et al. 2017) has revolutionized these studies, allowing us to pierce the veil of the ICM.

One of the primary targets of the ICM studies is the Virgo cluster, which is the nearest rich cluster to the Milky Way at a distance of 16.5 Mpc (Jerjen et al. 2004; Mei et al. 2007). Until recently a small catalog of 2096 galaxies (Binggeli et al. 1985) was predominantly used for the studies of the cluster – however, in the recent years much larger datasets have become available, such as the galaxy catalog assembled over the course of the 2MASS survey (Skrutskie et al. 2006).

With the number of observed galaxies in the region increasing, the problem of separating those belonging to the Virgo cluster from those populating the nearby ICM arises. The problem is confounded by the hierarchical structure of the Universe, making separation between the cluster region and the ICM somewhat arbitrary. Common metrics such as the critical radius r_{200} (see Section 2.1) are used for this purpose, but often require information not provided by digital large scale surveys.

Recently mathematical clustering methods have been increasingly used in order to detect hierarchies on various levels of the Universe structure, including stellar and galactic clusters. In particular, simple K-means (Lloyd 1982) clustering procedure was applied by Selim et al. (2020) to the problem of estimating the center of the Virgo cluster, in conjunction with the one-dimensional Gaussian fit. While successful in estimating the coordinates of the center and the number of Virgo cluster members in the employed dataset, the work did not shed light on the structure of the surrounding ICM, nor provide a quantitative assessment of the quality of the results.

In this work, I attempt to improve upon the work by Selim et al. (2020) with the following adjustments:

1. Rather than performing a one-dimensional Gaussian fit, I estimate the center coordinates by finding the MLE of the multinomial parameters.

2. I employ the increasingly used in astrophysics HDBSCAN (McInnes et al. 2017) hierarchical clustering method. Among its advantages are the minimized dependence on the clustering parameter choices, as well as the recreation of the hierarchical structure of the data set.

3. I compare the results of application of the two methods and quantitatively assess their performance.

[a)] The codes used in this work can be found on the Github page of the project.

4. I investigate the possibility of detection of the more complex structure by seeking to cluster the ICM data.

Below I provide an astrophysical background (Section 2.1), followed by the relevant introduction to the multivariate normal distribution (Section 2.2) and the HDBSCAN clustering method (Section 2.3). Then I describe the used data set (Section 3). Following is the description of the procedure and the results of the Virgo cluster center determination (Section 4.1), the clustering procedure (Section 4.2) and the comparison of the results of application of these two procedures (Section 4.3). The conclusion (Section 5) summarizes the results and outlines venues for future work.

2. BACKGROUND

2.1. *Astrophysics*

2.1.1. *Coordinates*

When surveying astrophysical objects, observers typically record their coordinates as points in a 6-dimensional space, corresponding to 3 spatial coordinates and 3 coordinates in the velocity space. The coordinates are as follows:

- Right Ascension (RA). It is the astronomical equivalent of longitude and measures the eastward angle along the celestial equator (see Figure 1). The celestial equator is the plane of the Sun's rotation around Earth at the March equinox (Seidelmann et al. 1992). It is measured in hours, from 0 to 24, linearly corresponding to 0° to 360° respectively. Automatically recorded by modern telescopes.

- Declination (DEC). The astronomical equivalent of latitude, it measures the northward angle with respect to the celestial equator (Figure 1). It is measured in degrees, from −90° to +90°. Just as RA, modern telescopes automatically record it.

- Parallax. Broadly, parallax is a displacement between the apparent position of the object viewed from two different locations (in practice, two positions of the Earth with respect to the Sun). A measure of distance to the object, while not always directly available (as distance can be inferred indirectly, without measuring said displacement), typically is listed in all astronomical survey data tables. Measured in regular distance units, such as parsec (1 parsec ≈ 3.26 light years).

- Radial velocity. Velocity of the object in direction away from Earth, most commonly measured in km/sec. Most commonly inferred from the redshift (the displacement of the observed spectral wavelengths due to the Doppler effect).

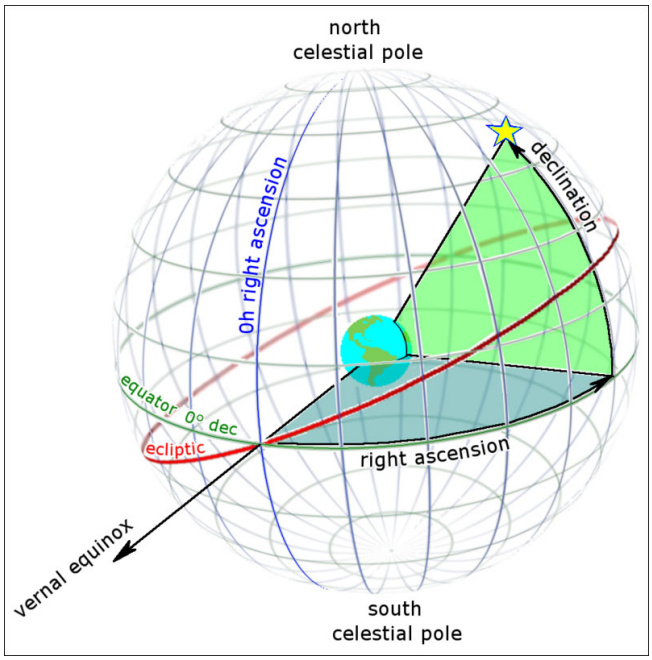

Figure 1. Right Ascention (RA) and Declination (DEC). Courtesy of Sky & Telescope.

- RA/DEC proper motions. Proper motions are the angular velocities of the object, i.e. the time derivatives of RA and DEC measurements. Commonly available for stars, but may be difficult to impossible to measure for galaxies due to their small values.

The data set used in this work (Huchra 2007) features accurate RA, DEC and radial velocity measurements for 27390 galaxies in a roughly rectangular region around the approximate center of the Virgo cluster. Parallax and proper motion measurements are unavailable, therefore the coordinates of the galaxies in the velocity space are unknown.

2.1.2. *Galaxy clusters and the ICM*

A detailed review of the physics of galaxy clusters and the ICM is provided in Walker et al. (2019). Here I provide a brief summary of the knowledge relevant to this work.

According to the Lambda-CDM model, the large scale structure of the Universe is mainly determined by the distribution of the dark matter (constituting approximately 85% of all matter in the Universe by mass). Small fluctuations result in overdensities of dark matter, and regular matter accumulates around those over-

densities, forming dense structures such as galaxy clusters. Each dark matter overdensity can be modeled as a dark matter halo with the boundary characterized by the "splashback radius" (Diemer & Kravtsov 2014), past which the dark matter density drops sharply (Mansfield et al. 2017; Diemer et al. 2017).

In contrast, the size of the galaxy cluster (characterized by the distribution of its regular matter) is considered to be much smaller than that of the dark matter halo. Typically it is defined by the overdensity radius, i.e. the radius at which the density of the matter becomes smaller than a given value. That value is taken to be a multiple of the critical density $\rho_{\rm crit}(z)$ which depends on the redshift z and equal the density at which the matter begins to collapse under the influence of its own gravity. The commonly used virial radius r_{178} (the radius of equilibrium, that is the radius within which the average kinetic energy of the matter is $-1/2$ of its potential energy) is usually approximated by r_{200}; here the index is the multiplier by the critical density $\rho_{\rm crit}(z)$.

This description illustrates the problem of determining the boundary of a galaxy cluster, making separation between the galaxies belonging to it and those belonging to the ICM difficult. A promising idea is to cluster the available galaxies in the distance-coordinate space, using the fact that the in-cluster and the ICM galaxies are expected to exhibit different kinematic behaviors.

2.2. *Statistics: multivariate normal distribution*

2.2.1. *Distribution and the maximum likelihood estimation*

A random variable with the multivariate normal distribution has the following p.d.f.:

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

where $\boldsymbol{x}$ is a k-dimensional vector, $\boldsymbol{\Sigma}$ is an $k \times k$ covariance matrix, and $\boldsymbol{\mu}$ is a k-dimensional mean vector equalling the element-wise expectation of $\boldsymbol{x}$. Given the set of data points $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}$, the log-likelihood function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be written as (up to a constant)

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}\left[\sum_{i=1}^{n}(\boldsymbol{x_i}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x_i}-\boldsymbol{\mu}) + n\log|\boldsymbol{\Sigma}|\right]. \quad (1)$$

The maximizer of this function can be found numerically (the method used in this work is described in Section 2.2.2), and the resulting values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ produce a multivariate normal fit for the data. Under the null hypothesis of the fitted model being the true distribution of $\boldsymbol{x}$, the likelihood of individual points being produced by the distribution can be characterized by their Mahalanobis distances:

$$d_i \equiv d(\boldsymbol{x_i}, \boldsymbol{\mu}) = \sqrt{(\boldsymbol{x_i}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x_i}-\boldsymbol{\mu})}. \quad (2)$$

For the random variable x, these distances have the χ^2 distribution with k degrees of freedom, therefore the probability of obtaining a given (d) or a smaller Mahalanobis distance is

$$P(d_i \geq d) = 1 - F_{\chi^2_k}(d), \quad (3)$$

where $F_{\chi^2_k}$ is the cumulative d.f. of the χ^2 distribution with k degrees of freedom. In the context of model fitting, this probability can be interpreted as the probability of the data point's membership in the target structure.

2.2.2. *Numerical optimization and the Limited-Memory BFGS*

Mathematical optimization is the process of selecting the best element among the available alternatives according to some criterion, in particular the process of finding the value of the argument maximizing the value of the function. *Numerical optimization* accomplishes this with a numerical algorithm, as a sequence of iterative steps. A good numerical optimization algorithm converges (i.e. closely approaches) to the true solution in a reasonable computational time and scales well with the size (for example, dimensionality) of the problem.

One of the simplest optimization algorithms is the Newton's method. For a function f of multiple variables $\boldsymbol{x} = (x_1, \ldots, x_k)'$, the algorithm searches for a local extremum as a sequence of steps; at the n-th step the assigned value of $\boldsymbol{x}$ is $\boldsymbol{x_n}$, and at the n-th step it is updated with

$$\boldsymbol{x_{n+1}} = \boldsymbol{x_n} - |J_f(\boldsymbol{x_n})|^{-1} f(\boldsymbol{x_n}),$$

where $J_f(\boldsymbol{x_n})$ is the Jacobian of $\boldsymbol{x_n}$. However, the Jacobian is not always available or computationally feasible to calculate, in which case it can be replaced with an approximation; methods employing this approach are called quasi-Newton methods. Typically, the exact function value in $\boldsymbol{x_{n+1}}$ is replaced with its second-order Taylor approximation as follows:

$$f(\boldsymbol{x_{n+1}}) \approx f(\boldsymbol{x_n}) + \nabla f(\boldsymbol{x_k})(\boldsymbol{x_{k+1}} - \boldsymbol{x_k}) + \\ + \frac{1}{2}(\boldsymbol{x_{k+1}} - \boldsymbol{x_k})' B (\boldsymbol{x_{k+1}} - \boldsymbol{x_k}),$$

where B is an approximation to the Hessian matrix of f. Upon convergence, setting the gradient of f to 0, we obtain the following update at the $(n+1)$-th step:

$$\boldsymbol{x_{n+1}} \approx \boldsymbol{x_n} - B^{-1}\nabla f(\boldsymbol{x_k}),$$

which, apparently, requires B to be non-singular. Typically the initial value of B is chosen to be $B_0 = I_k$ (the $k \times k$ identity matrix), and then the B estimate is updated at each step. In the popular BFGS method (Fletcher 1987), the $(n+1)$-th update is computed as follows:

$$\begin{cases} \boldsymbol{x_{n+1}} = \boldsymbol{x_n} - \alpha_n B_n^{-1} \nabla f(\boldsymbol{x_n}), \\ \Delta \boldsymbol{x_{n+1}} = \boldsymbol{x_{n+1}} - \boldsymbol{x_n}, \\ \boldsymbol{y_{n+1}} = \nabla f(\boldsymbol{x_{n+1}}) - \nabla f(\boldsymbol{x_n}), \\ B_{n+1} = B_n + \dfrac{\boldsymbol{y_{n+1}} \boldsymbol{y'_{n+1}}}{\boldsymbol{y'_n} \Delta \boldsymbol{x_{n+1}}} - \dfrac{B_n \Delta \boldsymbol{x_{n+1}} (B_n \Delta \boldsymbol{x_{n+1}})'}{\Delta \boldsymbol{x_{n+1}}' B_n \Delta \boldsymbol{x_{n+1}}}. \end{cases}$$

This method, however, can be computationally intensive. Rather than estimating a $k \times k$ Hessian approximation at each step, the Limited-Memory BFGS algorithm only stores the past $m \in \mathbb{N}$ updates of $\boldsymbol{x}$ and $\nabla \boldsymbol{x}$ and uses them to implicitly reproduce the operations done by the regular BFGS method with the B matrix. The details can be found in Liu & Nocedal (1989).

2.3. *Statistics: HDBSCAN clustering*

2.3.1. *Traditional methods and DBSCAN*

Many of the commonly used clustering methods, such as the K-means algorithm (MacQueen 1967; Lloyd 1982), use strong and sometimes unjustified assumption about the underlying distribution of the data. For example, the K-means algorithm assumes that the members of each cluster are normally distributed. In addition, there is the need to specify (priorly unknown) number of clusters k which strongly affects the result. Furthermore, the algorithm treats each cluster as having a simple structure and ignores the possibility of the common hierarchical structure. This is especially problematic when working with the large scale astrophysical data (such as galaxy clusters) which typically contains a multilevel hierarchy of structures.

DBSCAN (Ester et al. 1996) is a density-based clustering algorithm. It requires the user to set two parameters: the neighborhood parameter $\varepsilon > 0$ and the minimum number of points in a cluster $k \in \mathbb{N}$. It works as follows on a given set of data points $\{\boldsymbol{x_1}, \dots, \boldsymbol{x_n}\}$:

1. One of the data points $\boldsymbol{x_j}$ is selected.
2. The number of points within the ε-neighborhood of $\boldsymbol{x_j}$ is calculated; if it is $\geq k$, then a cluster C is formed of all the points $\{\boldsymbol{x_j}\}_{j \in C}$.
3. The procedure is applied to each of the points in C, updating C as needed.
4. Once no more points can be added to C, search for the next point determining a new cluster continues.
5. The procedure terminates once all points have been processed.

The output of the procedure is a set of clusters $\{C_1, C_2, \dots\}$ each containing a set of points from $\{\boldsymbol{x_1}, \dots, \boldsymbol{x_n}\}$, with empty intersections between any two clusters. Some of the points may remain unclustered.

2.3.2. *HDBSCAN*

While accounting for arbitrary structure of data, the regular DBSCAN algorithm still does not allow one to explore the possible hierarchical structure of the found clusters. The reason is the necessity to pre-set the ε parameter, while in the real data sets different clusters can have different scales. Allowing the value of ε to change dynamically would address this issue.

HDBSCAN (McInnes et al. 2017) is a hierarchical clustering algorithm that represents a modified version of the DBSCAN. In it, the dynamicity of the ε parameter is simulated by the altered distance metric: instead of the regular Euclidean distance as in the DBSCAN algorithm, the following distance metric (commonly referred to as the *mutual reachability distance*) between any two points $\boldsymbol{x}$ and $\boldsymbol{y}$ is used:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \max\left(d_E(\boldsymbol{x}, \boldsymbol{y}), d_c(\boldsymbol{x}), d_c(\boldsymbol{y})\right),$$

where d_E is the regular Euclidean distance, and d_c is the *core distance* which equals to the minimum value of ε containing k neighbors (the notation is borrowed from the previous section). d_c is a measure of local density of data, and the metric above results in denser regions discriminating stronger against outliers.

As the value of ε changes, groups of clusters merge into larger clusters, as schematically shown on Figure 2, forming a hierarchy of clusters. In the extreme cases $\varepsilon \to 0$ and $\varepsilon \to \infty$, each point forms its own single point cluster, and all points are united into a single cluster, respectively. *Persistence score* varying from 0 to 1 represents the stability of a given cluster: 0 represents a perfectly ephemeral cluster, while 1 represents a cluster that persists across all values of ε. Once the hierarchy of clusters is formed, the cluster tree is pruned such that the total sum of persistence of clusters is maximized.

As the value of ε needs not be specified, the only HDBSCAN parameter that needs to be varied is k, the minimum number of data points required to form a cluster core. This removes much of the subjectivity from the clustering process, compared to such methods as K-means or the regular DBSCAN.

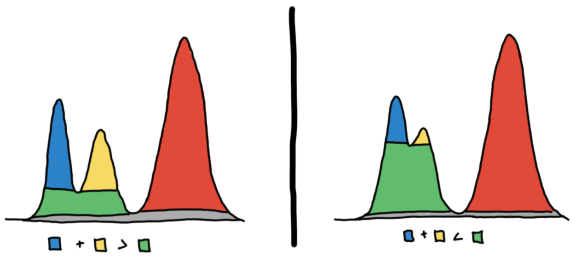

Figure 2. Visual representation of the hierarchy of HDBSCAN clusters from the official documentation.

3. DATA

3.1. *Description*

I use the data collected in the 2MASS survey (Skrutskie et al. 2006) and publicly distributed by the Center for Astrophysics (CFA) at Harvard University (Huchra 2007) that features various measurements and classifications of galaxies observed in a roughly rectangular region around the center of the Virgo cluster. A stub of data is shown in Table 1. The columns from the left to the right contain the following information:

1. Name: Numerical name of the galaxy (based on its RA/DEC coordinates).
2. RA: The Right Ascention of the galaxy: hours, minutes and seconds.
3. DEC: The Declination of the galaxy: degrees, months and seconds.
4. K: The absolute magnitude of the galaxy. The *absolute magnitude* is a measure of luminosity of the object which is on a logarithmic scale with respect to the energy flux from the object. Specifically:

$$F = F_{10} \times 100^{\frac{M-m}{5}},$$

where F is the energy flux from the object measured on Earth, F_{10} is the energy flux measured at the distance of 10 parsec from the object, M is the absolute magnitude, and m is the *apparent magnitude* (equal to the absolute magnitude at the distance 10 parsec). For reference, the Sun has the absolute magnitude of +4.85; lower absolute magnitude corresponds to higher luminosity.

5. V: The radial velocity of the galaxy (in km/s).
6. V_{err}: The error of the radial velocity measurement (in km/s).
7. Major axis: The angular size of the longer axis of the galaxy as viewed from Earth (in arcmin).

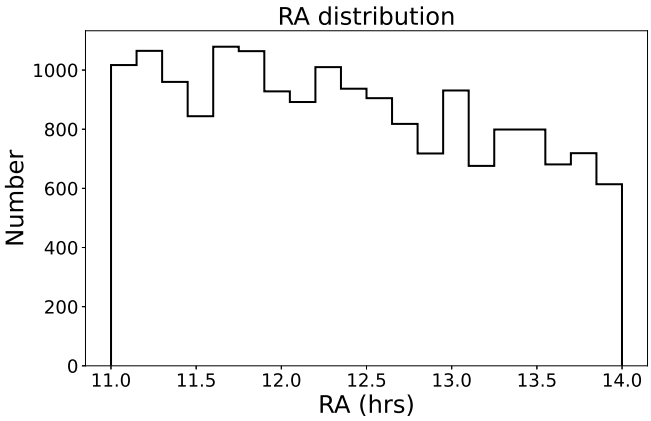

Figure 3. The Right Ascension distribution in the data set.

8. Minor axis: The angular size of the shorter axis of the galaxy (in arcmin).

3.2. *Investigation and processing*

Although the data shown in Table 1 contains such important characteristics of the surveyed galaxies as their magnitudes and axis measurements, no known relation between these characteristics and the membership in a galaxy cluster exists. Therefore I choose to only use three variables in all of our analyses: Right Ascension (RA), Declination (DEC) and radial velocity (V) (we do not make use of the radial velocity error $V_{\rm err}$ for brevity).

First, I convert all the RA and DEC values into hours and degrees respectively. The data contains the RA values in the range $[11, 14]$ hrs and the DEC values in the range $[-10, 35]$ degrees. Figures 3 and 4 show the distributions of RA and DEC in the data. We see that the RA distribution is fairly uniform, while the DEC distribution is roughly uniform in the range of $[-10, 10]$ degrees and experiences a dip in the $[10, 35]$ degree range. The latter is a result of the particular survey methodology and may be expected to introduce a bias in my analysis.

Upon examination, it can be seen that a large number of the radial velocity values V are missing (marked by the 0 entries in the table). Rather than imputing these values, I choose a more common approach in stellar and galactic astrophysics and remove all galaxies with missing radial velocity entries from consideration. This choice is justified by the fact that a given RA/DEC combination corresponds to a direction on the celestial sphere and can include galaxies at arbitrary distances from the Milky Way, hence one should not expect a strong correlation between the RA/DEC and the radial velocity values among the surveyed galaxies, and an attempt to perform ratio imputation or any other correlation-based imputation method is likely to lead to significant errors.

Table 1. Galactic cluster data

Name	RA			DEC			K	V	V_{err}	Major axis	Minor axis
	hh	mm	ss	ddd	mm	dd.d	mag	km/s	km/s	arcmin	arcmin
110000.95+2142354	11	0	0.95	21	42	35.4	12.66	0	0	0.46	0.28
110001.01+0106443	11	0	1.01	1	6	44.3	11.91	11706	21	0.34	0.23
110000.82+0921472	11	0	0.82	9	21	47.2	12.48	0	0	0.28	0.21
110001.01+1415375	11	0	1.01	14	15	37.5	13.36	52541	50	0.24	0.21
110001.46-0255362	11	0	1.46	−2	55	36.2	13.39	14160	39	0.19	0.19
110001.93+0146344	11	0	1.93	1	46	34.4	11.34	12163	38	0.65	0.49
110002.13-0257162	11	0	2.13	−2	57	16.2	13.29	11441	87	0.19	0.11
110002.33+2118504	11	0	2.33	21	18	50.4	13.04	0	0	0.38	0.21
110002.39+1450295	11	0	2.39	14	50	29.5	9.8	1436	4	1.86	1.6

NOTE—This table is a stub; the full data set is available at the project Github page.

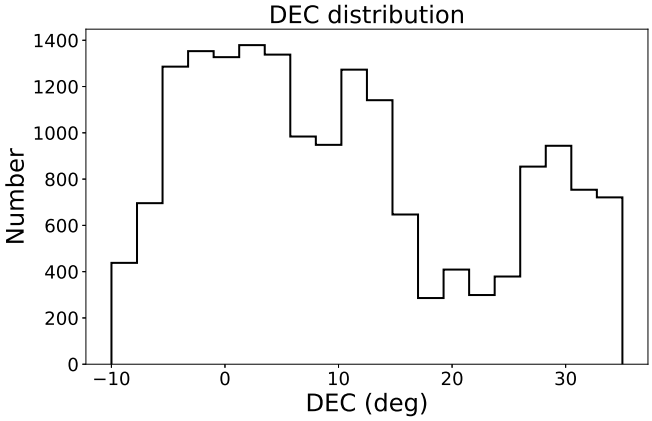

Figure 4. The Declination distribution in the data set.

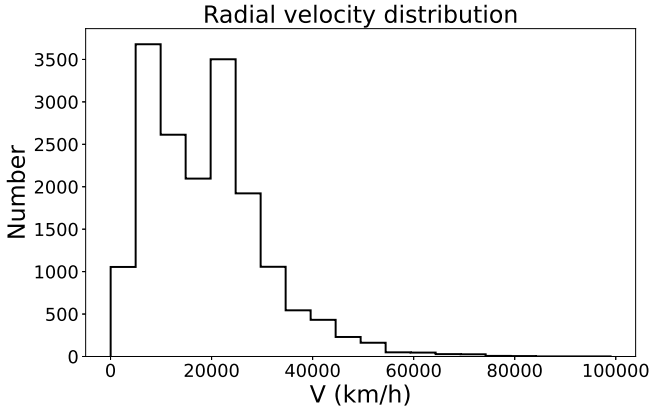

Figure 5. The radial velocity distribution in the data set.

After removing the galaxies with the missing V entries from the sample, the sample size is reduced from 27390 to 17456 galaxies; this is almost exactly the number (17466) used by Selim et al. (2020), although the source of the small discrepancy is unknown. This is the final sample I will be working with. The resulting distribution of the radial velocities is shown on Figure 5. Note first that all the radial velocities are positive; this is because only the absolute values are recorded in the data set. According to Hubble's law (Hubble 1929), on average, the more distant objects move faster away from any given location – and, given the large distance to even the closest major galaxies ($\sim$ 2.5 million light years, Ribas et al. 2005) and especially to the center of the Virgo cluster ($\sim$ 53.7 million light years, Mei et al. 2007), one should expect the vast majority of radial velocities to be positive. Hence the error introduced by using the absolute values instead of the true ones is likely to be small.

In the further analysis, considering the expected similarity in these values between the members of the Virgo cluster, I view the RA, DEC and radial velocity values as laying in a 3-dimensional Euclidean space. For the purposes of multivariate normal distribution fitting and HDBSCAN clustering, the values are normalized to have the mean of 0 and the standard deviation of 1 as follows:

$$\hat{x}_{k,i} = \frac{x_{k,i} - \overline{x_k}}{\sigma(x_k)}, \quad k \in \{1,2,3\}, \quad i \in \{1,\ldots,17466\}, \tag{4}$$

where x_k represents a column of values of the k-th dimension in the Euclidean space, $x_{k,i}$ represents its i-th entry, $\overline{x_k}$ is the mean of x_k, and $\sigma(x_k)$ is its standard deviation. $\hat{x}_{k,i}$ is the normalized entry.

4. PROCEDURE AND RESULTS

4.1. *Center determination*

In order to estimate the center of the Virgo cluster, I employ the assumption that the normalized coordinates

of the members of the cluster (RA, DEC and radial velocity) have the multivariate normal distribution. I note that this is a very strong assumption: while the observed mass distributions of large scale structures in the Universe are not very well known, a number of models (partially supported by the results of cosmological simulations) of dark matter density profiles exist, and none of the widely accepted models are multivariate normal. A summary of some of the most popular profiles is provided in Coe (2010), and the work by Mandelbaum et al. (2006) analysing the SDSS data in detail found these profiles to be consistent with the Navarro, Frenk & White (NFW) profile (Navarro et al. 1996).

As described in Section 2.2, the MLE estimation of the multivariate normal distribution parameters involves maximization of the log-likelihood function (Equation 1); I use the L_M BFGS optimization method for this purpose (see Section 2.2.2). The convergence criterion was set as $\max|\mathrm{proj}_f(\boldsymbol{x_n})| < 10^{-6}$, where $\boldsymbol{x_n}$ is the argument value at the n-th step of the algorithm, proj_f is a vector of elements of the projected gradient of f, and f is the log-likelihood function in question. The only constraint I set is that the Σ matrix is symmetric; while the L_M BFGS algorithm requires that it also be positive-definite, I do not enforce this constraint explicitly. The algorithm converged after 9 steps to the following parameters in the space of normalized RA, DEC and radial velocity values:

$$\boldsymbol{\mu} \approx (0,0,0)', \qquad \Sigma \approx \begin{pmatrix} 1.00 & 0.08 & -0.01 \\ 0.08 & 1.00 & -0.20 \\ -0.01 & -0.20 & 1.00 \end{pmatrix},$$

where $\boldsymbol{\mu}$ and Σ are the MLE estimates of the mean and the covariance matrix of the distribution respectively. The fact that the mean estimate is extremely close to 0 is expected from the data set selection: it had been pre-processed to be centered at the expected center of the Virgo cluster. The covariance matrix shows no significant correlation between RA and the radial velocity values, partially validating my choice to remove the data points with the missing radial velocity values made earlier (see Section 3). However, there is some correlation between the DEC and the radial velocity values, as well as between the DEC and the RA values; this could be attributed to the bias in the DEC distribution visible on Figure 4, although confirmation of this hypothesis requires further investigation.

Then, in accordance with Equations 2 and 3, the Mahalanobis distance and the cumulative probability are calculated for each of the data points; the latter are interpreted as the membership probabilities, i.e. the probabilities that the given points belong to the Virgo cluster and not the nearby ICM region. The distribution of these probabilities is shown on Figure 6 and appears to be skewed towards lower values, whilst for a true multivariate normal distribution they would be expected to be distributed uniformly.

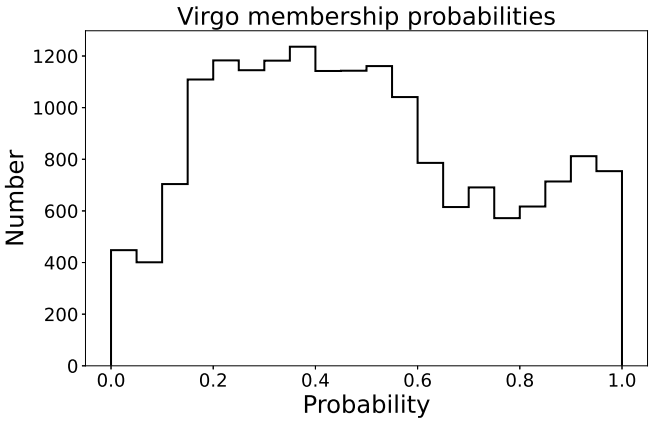

Figure 6. Probabilities of membership in the Virgo cluster under the multivariate normality assumption.

Under the assumption of multivariate normality, the true number of the Virgo cluster members can be estimated as

$$\hat{N} \approx \sum_{i=1}^{n} p_i,$$

where p_i is the Virgo cluster membership probability of the i-th datapoint. Numerical estimation of this value results in $\hat{N} \approx 8400$. Note that Selim et al. (2020) give a much more conservative estimate of $\hat{N} \approx 1300$ – however, they do not explain the details of the calculation and appear to have made this decision based on a somewhat arbitrary cut on the R-squared value of the univariate normal fit. It is worth noting that in my analysis I used all the data points available for the fit, while it is well known (and explicitly stated on the webpage of the data set, The Virgo Cluster) that many of the galaxies do not belong to the Virgo cluster, hence cannot have come from the same distribution. The data providers suggest employing a luminosity (magnitude) cut, although it has already been mostly done by removing the galaxies with no measured radial velocities (and redshifts) from consideration (the more luminous the galaxy, the more likely it is to have a reliable redshift measurement).

The estimated multivariate normal parameters correspond (after the de-normalization of the variables through inversion of Equation 4) to the center RA and DEC coordinates of ≈ 12.38 hrs and $\approx 10.18°$ respectively, and the center radial velocity of ≈ 18944 km/s. For comparison, the regularly cited work by Mahdavi & Geller (2001) lists the RA and DEC coordinates as 12.44

hrs and 12.72° respectively, which is fairly close to my values. However, Wu et al. (1998), old as the work may be, lists the radial velocity of the center of the Virgo cluster as 1137 km/h, drastically differing from my result – and the recent work by Kashibadze et al. (2020) still lists it as under 2000 km/h. A strong presence of the high radial velocity galaxies in CFA's sample, obvious from Figure 5, in conjunction with only the absolute values being listed, is the likely cause of this massive discrepancy.

Figure 7 shows the distribution of the average cluster membership probability p_i in the surveyed RA–DEC region. As expected, in general, the further away from the center the RA and DEC values lay, the lower the membership probability is – however, there is a finer structure observed in the central region, caused by the non-uniform distribution of radial velocities in that region. If one was interested in selecting the most likely candidates for the Virgo cluster membership, the most obvious way would be to enforce a cut on the p_i values, for instance, at the value of 0.95. The stricter the cut, the smaller the cluster region will be. In Section 4.3 I employ the (empirically selected) cut at 0.75, and cuts above ~ 0.85 appear to be too strict and result in strong mismatch between the results of the clustering procedure and the MLE fit (see the explanation below).

4.2. *Clustering*

In order to attempt to separate the Virgo cluster members from the members of the nearby ICM and potentially other galaxy cluster members, I employ the HDBSCAN clustering algorithm as described in Section 2.3. The single parameter I alter is the minimum number of data points required to form a cluster core, k. There appear to be three regions of k values producing different sets of clusters: "low", "medium" and "high". The clustering results for these choices are summarized in Table 2, where I list three general regions of k values in the first column, list of obtained clusters in the second column (the naming scheme is explained in Section 4.3), and the number of galaxy members in each in the third column.

From inspection of the Table, we can see that A_1 and B_1 clusters coincide, as do clusters A_2 and B_3, and B_2 and C_2. These coincidences are a result of the hierarchical structure of the HDBSCAN clusters. I explore these clusters in detail in the next Section.

4.3. *Comparison of methods and implications*

In order to assess the quality of the HDBSCAN clusters listed in Table 2, I introduce an empirical threshold on their members' Virgo cluster membership probabilities, p_i, of 0.75. Then I calculate the fraction of the members of each cluster that have the membership probabilities of above that threshold. The clusters then are sorted in decreasing order by that fraction; a good clustering outcome would be one in which the fraction of the first cluster after sorting is significantly higher than that of the second cluster, indicating that the first cluster corresponds to the Virgo cluster, and all the other clusters correspond to its surroundings.

The result is shown in Table 3. As can be seen, the clusters associated with the Virgo membership according to the multivariate normal fit are, at least, an order of magnitude more populated with the likely Virgo members than other clusters, indicating a good agreement between the multivariate normal fit and the HDBSCAN clustering. I note that this general agreement persists as long as the p_i threshold does not exceed a high value of ~ 0.85, after which the Virgo membership fractions in all clusters become extremely low. Such high thresholds might not be justified, given how loose the multivariate normal fit is and how many non-Virgo galaxy members it was fit over.

It is worth noting that the cluster associated with the likely Virgo membership is not always the largest cluster produced by the HDBSCAN: in case of $k \sim 100$ and $k \sim 1000$, the second largest cluster was associated with the Virgo cluster. This hints at one of the two possibilities: either the number of the Virgo cluster members is relatively small in the initial data set, causing HDBSCAN to overestimate the connectivity between the surrounding ICM members – or a larger cluster structure is present in the data set. Due to the scarcity of research involving this data set, neither of these two possibilities can be outruled at the moment. The latest generation of cosmological simulations of galaxy clusters including the Virgo cluster specifically, one example of which is the recent work by Sorce et al. (2021), is likely to shed light on the likelihood of these possibilities.

Lastly, Figures 8, 9 and 10 show the first cluster (in red) and the second cluster (in blue) members corresponding to the clusters in Table 2, for $k \sim 10$, 100 and 1000 respectively. Interestingly, the clusters associated with high Virgo membership fractions appear to occupy the high DEC value region and gravitate towards the high RA value region. At the same time, the fractions of associated Virgo members are close to 25% in all three cases: this consistency, in conjunction with the relatively small value, suggests that the employed clustering methodology does not allow for very reliable Virgo member detection. I am not aware of any other work applying mathematical clustering to this or a similar data set featuring Virgo cluster members; based on experience, I posit that it is quite likely that RA, DEC and

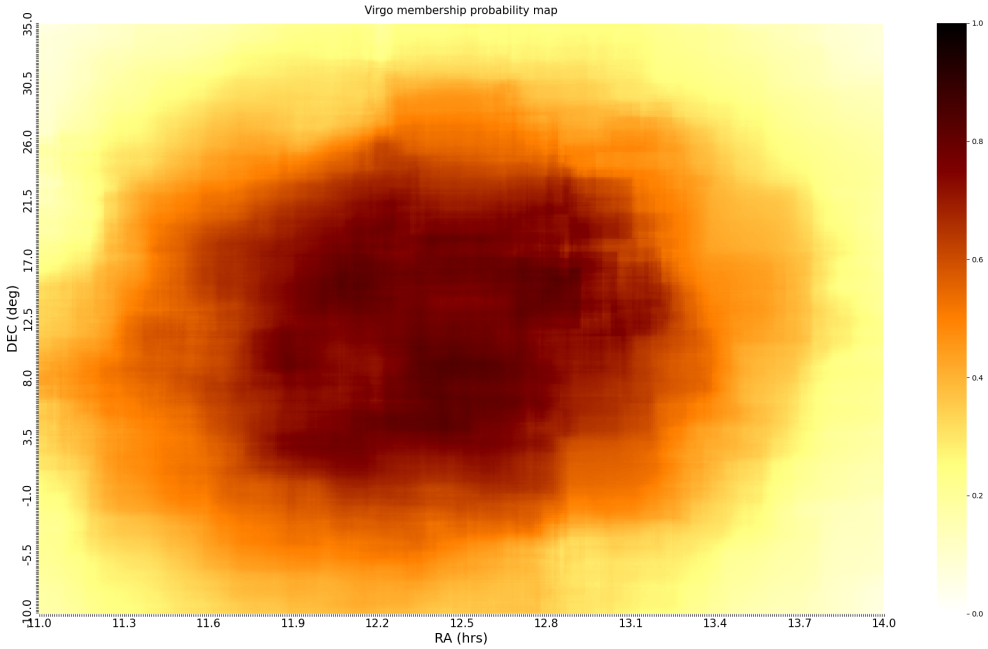

Figure 7. RA/DEC map of the average Virgo cluster membership probability (denoted by color).

Table 2. HDBSCAN clustering results

k	Cluster name	Number of galaxy members
~ 10	A_1	836
	A_2	363
	A_3	51
	A_4	37
	A_5	30
~ 100	B_1	836
	B_2	7065
	B_3	363
~ 1000	C_1	2497
	C_2	7065

Table 3. HDBSCAN clustering results: Virgo membership fractions

k	First cluster name	Virgo membership fraction	Second cluster name	Virgo membership fraction
10	A_1	26.2%	A_2	1.9%
100	B_2	26.2%	B_1	3.4%
1000	C_2	25.4%	C_1	3.4%

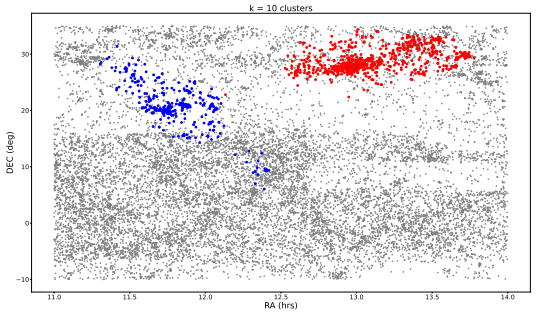

Figure 8. Two clusters with the highest Virgo membership fractions for $k = 10$ (see text). Red points: galaxies associated with the Virgo membership. Blue points: galaxies associated with the ICM or other structures.

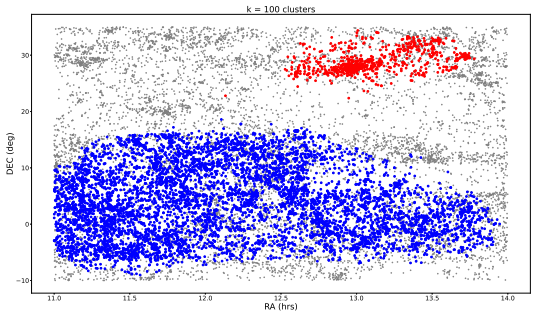

Figure 9. Two clusters with the highest Virgo membership fractions for $k = 100$ (see text). Red points: galaxies associated with the Virgo membership. Blue points: galaxies associated with the ICM or other structures. Grey points denote galaxies not belonging to either of these clusters.

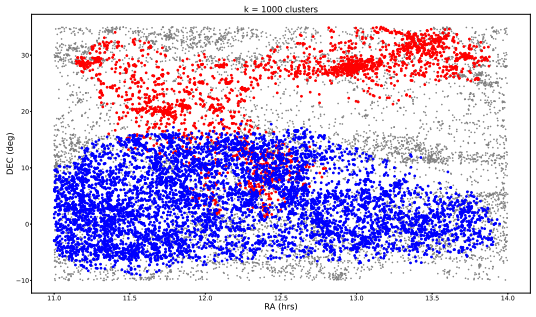

Figure 10. Two clusters with the highest Virgo membership fractions for $k = 1000$ (see text). Red points: galaxies associated with the Virgo membership. Blue points: galaxies associated with the ICM or other structures. Grey points denote galaxies not belonging to either of these clusters.

radial velocity alone are not sufficient to reliably separate members of different large scale structures from each other.

Mathematical clustering algorithms have recently been extensively applied to the Milky Way stellar datasets, with the purpose of linking stars with common origin to each other. Attempts to cluster stars over directly observed coordinates such as RA, DEC, radial velocity, parallax and proper motions have systematically failed; instead, researchers use these coordinates in order to reconstruct the stellar orbits around the Milky Way, and the clustering then is performed over the quasi-conserved parameters of those orbits. Some of the prominent results in this respect when applied to rare chemically peculiar stars on the outskirts of the Milky Way (which are as similar analysis-wise to galaxy clusters as one can find) were obtained by Roederer et al. (2018), Yuan et al. (2020) and Gudin et al. (2021). However, this procedure is not applicable to galaxies and galaxy clusters whose orbits around their approximate centers of gravity are practically incalculable. It is possible that mathematical clustering is not the best approach to study galaxy clusters and the processes in their surrounding ICMs.

5. CONCLUSION

Studies of galaxy clusters and their surrounding ICMs have been plagued by lack of substantial data and difficulties in estimating the kinematic parameters of distant galaxies. Despite the recent technological breakthroughs and rapidly increasing data set sizes, detailed studies of even the closest galaxy clusters to our Local Group have proven elusive. All studies to date have been fairly inconclusive and prone to a large number of biases and subjective choices.

In this work I attempted to combine multivariate normal modelling with the (widely recognized as the most promising in observational astrophysics) HDBSCAN clustering method in order to shed light on the galaxy membership in the Virgo galaxy cluster and its surrounding ICM. Ideally, I planned to discern a finer structure in the vicinity of the Virgo cluster than one discerned in previous works.

The results were inconclusive, and both the multivariate normal model and the HDBSCAN clusters featured heavy contamination of non-Virgo cluster members. Directly using kinematic parameters such as RA, DEC and radial velocity for the purposes of linking multiple objects together, as previously found in stellar studies, does appear to confirm strong links between different galaxies, and more indirect approaches are warranted.

While clustering methods of astrophysical objects rely on quasi-conserved values such as angular momenta or orbital energy typically unavailable for galaxies and galaxy clusters, distribution modelling methods only require theoretical knowledge of the galaxy cluster density distribution. Furthermore, density distribution of dark matter halos is expected to be generally the same regardless of the scale considered. Future observational studies of stellar clusters and galaxies, in conjunction with the rapidly increasing scales of cosmological numerical simulations, are likely to produce accurate knowledge of the most commonly occurring density distributions. Those distributions will then be usable for the plausible MLE estimates of their parameters, significantly improving upon my crude multivariate normal model.

REFERENCES

Binggeli, B., Sandage, A., & Tammann, G. A. 1985, AJ, 90, 1681, doi: 10.1086/113874

Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28, doi: 10.3847/1538-3881/aa7567

Coe, D. 2010, arXiv e-prints, arXiv:1005.0411. https://arxiv.org/abs/1005.0411

Diemer, B., & Kravtsov, A. V. 2014, The Astrophysical Journal, 789, 1, doi: 10.1088/0004-637x/789/1/1

Diemer, B., Mansfield, P., Kravtsov, A. V., & More, S. 2017, The Astrophysical Journal, 843, 140, doi: 10.3847/1538-4357/aa79ab

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96 (AAAI Press), 226–231

Ferrarese, L., Côté, P., Cuillandre, J.-C., et al. 2012, ApJS, 200, 4, doi: 10.1088/0067-0049/200/1/4

Fletcher, R. 1987, Practical Methods of Optimization, 2nd edn. (New York, NY, USA: John Wiley & Sons)

Gudin, D., Shank, D., Beers, T. C., et al. 2021, ApJ, 908, 79, doi: 10.3847/1538-4357/abd7ed

Hubble, E. 1929, Proceedings of the National Academy of Sciences, 15, 168, doi: 10.1073/pnas.15.3.168

Huchra, J. 2007, Center for Astrophysics, Harvard University

Jerjen, H., Binggeli, B., & Barazza, F. D. 2004, AJ, 127, 771, doi: 10.1086/381065

Kashibadze, O. G., Karachentsev, I. D., & Karachentseva, V. E. 2020, A&A, 635, A135, doi: 10.1051/0004-6361/201936172

Kim, S., Rey, S.-C., Jerjen, H., et al. 2014, The Astrophysical Journal Supplement Series, 215, 22, doi: 10.1088/0067-0049/215/2/22

Liu, D., & Nocedal, J. 1989, Mathematical Programming, 43, 503

Lloyd, S. 1982, IEEE Transactions on Information Theory, 28, 129, doi: 10.1109/TIT.1982.1056489

MacQueen, J. B. 1967, in Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, ed. L. M. L. Cam & J. Neyman, Vol. 1 (University of California Press), 281–297

Mahdavi, A., & Geller, M. J. 2001, The Astrophysical Journal, 554, L129, doi: 10.1086/321710

Mandelbaum, R., Seljak, U., Cool, R. J., et al. 2006, Monthly Notices of the Royal Astronomical Society, 372, 758, doi: 10.1111/j.1365-2966.2006.10906.x

Mansfield, P., Kravtsov, A. V., & Diemer, B. 2017, The Astrophysical Journal, 841, 34, doi: 10.3847/1538-4357/aa7047

McInnes, L., Healy, J., & Astels, S. 2017, The Journal of Open Source Software, 2, doi: 10.21105/joss.00205

Mei, S., Blakeslee, J. P., Côté, P., et al. 2007, ApJ, 655, 144, doi: 10.1086/509598

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563, doi: 10.1086/177173

Pak, M., Rey, S.-C., Lisker, T., et al. 2014, Monthly Notices of the Royal Astronomical Society, 445, 630, doi: 10.1093/mnras/stu1722

Ribas, I., Jordi, C., Vilardell, F., et al. 2005, The Astrophysical Journal, 635, L37, doi: 10.1086/499161

Roederer, I. U., Hattori, K., & Valluri, M. 2018, AJ, 156, 179, doi: 10.3847/1538-3881/aadd9c

Seidelmann, P. K., Britain., G., & Observatory., U. S. N. 1992, Explanatory supplement to the astronomical almanac / prepared by the Nautical Almanac Office, U.S. Naval Observatory ; with contributions from H.M. Nautical Alamanac Office, Royal Greenwich Observatory ... [et al.] ; edited by P. Kenneth Seidelmann, [rev. ed.]. edn. (University Science Books Mill Valley, Calif), xxviii, 752 p. :

Selim, I., Elkafrawy, P., & Dabour, W. 2020, International Journal of Astronomy and Astrophysics, 10, 1, doi: 10.4236/ijaa.2020.101001

Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163, doi: 10.1086/498708

Sorce, J. G., Dubois, Y., Blaizot, J., et al. 2021, Monthly Notices of the Royal Astronomical Society, 504, 2998, doi: 10.1093/mnras/stab1021

Walker, S., Simionescu, A., Nagai, D., et al. 2019, Space Sci. Rev., 215, 7, doi: 10.1007/s11214-018-0572-8

Wu, X.-P., Fang, L.-Z., & Xu, W. 1998, A&A, 338, 813. https://arxiv.org/abs/astro-ph/9808181

York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, AJ, 120, 1579, doi: 10.1086/301513

Yuan, Z., Myeong, G. C., Beers, T. C., et al. 2020, ApJ, 891, 39, doi: 10.3847/1538-4357/ab6ef7