An Investigation Into The Local Universe Galaxy Count Using Computational Image Processing

Shona Curtis-Walcott

***Abstract*—A deep optical image of the intergalactic sky was analysed using algorithms in Python to extract the number count of galaxies in the local universe. The image was taken at the Kitt Peak Mayall 4m Telescope using a CCD mosaic camera with a Sloan r filter. A galaxy catalogue was built which contained the positions, sizes, and associates magnitudes of each object detected. A total of 682 galaxies were identified. A count-magnitude graph was plotted for magnitudes $8 < m < 16.4$, the gradient of which was found to be 0.242 ± 0.004 which differed from the expected value of 0.6 by 59.7%. The discrepancy is thought to be due to the insufficient scale considered in this investigation. The Cosmological Principle applies only to Cosmological scales, and therefore the small section of sky analysed in this investigation is too small to demonstrate the uniformity of the galactic distribution.**

I. Introduction

A key method of astronomical observation used to gain insight into the properties of the Universe is that of optical astronomy. Telescopes paired with charge-coupled devices allow measurements of electromagnetic radiation originating from a section of the sky to be visualised as 2D images. Via computational analysis of the images, information on astronomical objects can be gathered without the practical limitations faced when attempting scientific experiments in space. A comprehensive galactic survey can be built using the information gathered, and in turn aid research into topics such as galactic evolution. This report details the analysis of an image taken at the Kitt Peak Mayall 4m Telescope using a CCD mosaic camera, which covers $0.0469d^2$ of the sky. A Sloan SSDS r band filter with central wavelength 620 nm was used to capture the image in order to standardise the wavelength range. A computational algorithm was written to extract information about objects present within the image, and hence built a galactic survey from which the count-magnitude relationship of galaxies could be investigated.

II. Theory

A. Count-Magnitude Relationship

The first star catalogue was recorded in the secondary century BC by the Greek astronomer Hipparchus [1]. The visual brightness of the stars he recorded followed a logarithmic scale, like that of the human eye. Astronomers to this day measure star brightness in terms of apparent magnitude, defined in the following manner:

$$m_1 - m_2 = -2.5log_{10}\left(\frac{f_1}{f_2}\right) \tag{1}$$

where $m_1 - m_2$ is the apparent magnitude difference between two astronomical objects with flux f_1 and f_2 respectively. The minus sign indicates the brighter an object appears, the smaller the difference in apparent magnitude is between it and another fainter object. Hipparchus introduced a system for determining the apparent magnitude of a specific object by defining the magnitude of the AO star Vega to be zero. In doing so all star magnitudes can be measured relative to Vega, with (1) transformed to:

$$m_1 = -2.5log_{10}\left(\frac{f_1}{f_Vega}\right) \tag{2}$$

which can be further expressed as:

$$m = -2.5log_{10}(f) + c \tag{3}$$

where c is a 'zero point' constant. In the absence of absorption, the flux observed at a distance r from an object of luminosity L follows the following inverse square relation:

$$f = \frac{L}{4\pi r^2} \tag{4}$$

Substituting this into (3) we obtain,

$$m = -2.5log_{10}\left(\frac{L}{4\pi}\right) + 5log_{10}(r) \tag{5}$$

where we can consider L to be the luminosity of a galaxy. A few physical assumptions must be made in order to obtain the count-magnitude relationship, namely that the Universe is a homogeneous, isotropic, and Euclidean medium. The number of galaxies within a radius r of the Earth can then be expressed as the volume multiplied by the number density of galaxies, as in (6).

$$N(r) = \rho \int_0^r 4\pi r'^2 dr' = \frac{4}{3}\pi r^3 \rho \tag{6}$$

To proceed, equation (5) can be rearranged for r and then substituted in (6), leading to,

$$N(m) = c \times 10^{0.6m} \tag{7}$$

where c is a collective constant. Upon taking the log of both sides and rearranging, an expression for the number of luminous objects with magnitude below a limiting value m is obtained:

$$log_{10}(N(m)) = 0.6m + constant \tag{8}$$

It is worth drawing attention back to the assumptions underlying this count-magnitude relationship; assuming a uniform distribution of astronomical objects generates a lack of consideration towards the Galactic structure inside which the

Earth lies. As this investigation is based upon an image taken from an observatory on earth, things such as the increase in object density on the Galactic plane of the Milky Way axis deserve consideration. On a larger scale, the assumptions also ignore the time-evolution of the number of stars and galaxies contained, as well as not being consistent with the expansion of the Universe. It is predicted that turning a blind-eye to the dynamic nature and structure of the Universe will create a disagreement between results obtained in this investigation and that which is predicted in (8).

B. Charge-Coupled Device (CCD)

A CCD functions by assigning each pixel a p-doped metal-oxide-semiconductor (MOS) capacitor which is able to convert incoming photons to an electric charge using the photoelectric affect [2]. The number of electrons per pixel is then counted and stored using a shift register, and can be interpreted by software to create the final image. The use of CCD spurred advancements in the field of astronomical imaging due to their high sensitivity: they possess a quantum efficiency of around 90% meaning roughly 9 out of 10 incident photons are detected. Furthermore, being based on an electronic sensor rather than a chemical sensor establishes a strong linear relationship between the light intensity incident on each pixel and the total charge collected. A downfall of the CCD is its ability to become saturated above a certain charge count and hence lose this linear relationship. The Kitt Peak telescope contains a CCD detector with a 16-bit counter, meaning that a maximum of $2^16 = 65536$ photons are able to be stored per pixel. "CCD Blooming" occurs when the maximum count of a pixel is reached and all further counts leak over into nearby pixels, obscuring visual properties of the true image. Sensors in a CCD are designed to restrict the movement of charge between horizontal pixels, whilst allowing vertical shifting. Therefore, CCD Blooming is typically identified as a vertical streak across an image. A prime example of blooming can be seen in the CCD image under consideration in this report, shown in Fig.1, where the close proximity of the central brightest star to earth results in a blooming effect.

C. The Image: Noise

It is desirable to reduce noise present in astronomical images in order to capture as much faint detail possible. A typical method used to reduce noise in an image is to built it from a stack of multiple exposures [3]: the counts collected from bright astronomical objects will stay constant whereas the value of the noise will fluctuate due to its random nature. Averaging over multiple exposures therefore increases the Signal to Noise Ration (SNR) creating a better quality image overall. Unfortunately, using this method also causes a defect visible at the edges of an image. This is caused by a lower SNR in these regions due to fewer objects being present. This type of defect is also visible in Fig.1.

III. Experimental Procedure

A. Initial Masking and Background Determination

The CCD image was stored in the form of a FITS file which could be easily read using Python. The first task was to prepare the image for analysis by removing the obvious defects -namely CCD blooming and noise from the low SNR in the image edges- using a process called masking. A mask is a 2D Boolean array of identical dimensions as that of the FITS image under consideration which can be applied to the image to filter out unwanted details. The central bright star and CCD blooming were masked manually by setting the array inputs corresponding to the appropriate pixel values with digit 0. A simple histogram of pixel count over the whole unmasked image was plotted, exhibiting a Gaussian shape of mean count 3614 and standard deviation $= 2332$. This Gaussian displayed an outlier spike corresponding to the noise generate by the upper central bright star. A sigma-clipping operation -clipping the image pixel values to within 5σ of the mean- was hence used to allow for a more robust estimation of the background radiation.

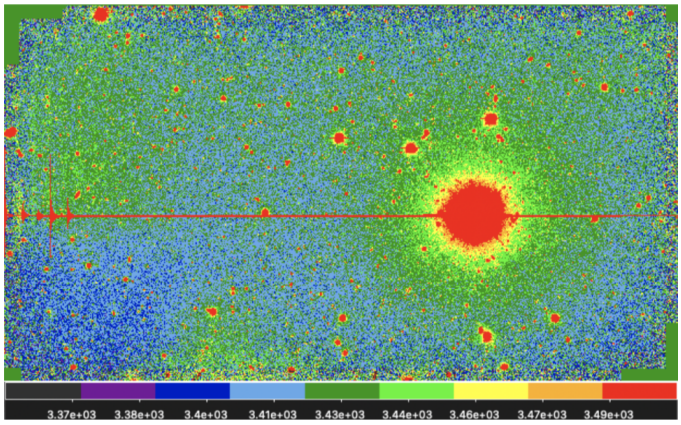

Fig. 1: The image taken at the Kitt Peak Mayall 4m Telescope using a CCD mosaic camera, covering $0.0469d^2$ of the sky. A Sloan SSDS r band filter with central wavelength $620nm$ was used to capture the image. The image is presented with the $zscale$ viewing option on the SAOImage DS9 software, and has been rotated to the right by $90°$. This scale allows fainter objects to be seen in the presence of much brighter objects.

An algorithm was written to determine the appropriate width at which the edges of the image should be masked. This was done by calculating the standard deviation of a portion of the image of area A as indicated in Fig.2. The area A was increased gradually by decrementing the values of L_0 and W_0 until the whole image was enclosed. At each stage, the standard deviation of pixel counts within A was calculated. A graph of standard deviation against the width of frame (where width of frame can be either W_0 or L_0; their values remained equal) is plotted in Fig.3. A steep rise in the count standard deviation was seen once a width of 100 had been surpassed. From a visual inspection of the image, defects were seen in each corner which stretched out beyond an edge width of $100mm$. Therefore, a final width of 144 pixels was chosen as the cut-off width. The updated image on which analysis could now be made is displayed in Fig.11 in the Appendix. For the following analysis, the upper limit of the mean background radiation was taken as:

$$n \times \sigma_p \tag{9}$$

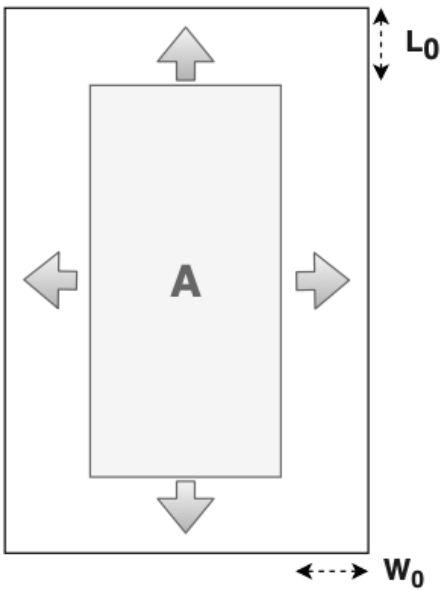

Fig. 2: Diagram to assist the explanation of how the width of image-edge to be masked was found. L_0 and W_0, which remained of equal value, were slowly decremented from 632 to 0, and after every decrement the standard deviation of pixel values inside area A was calculated.

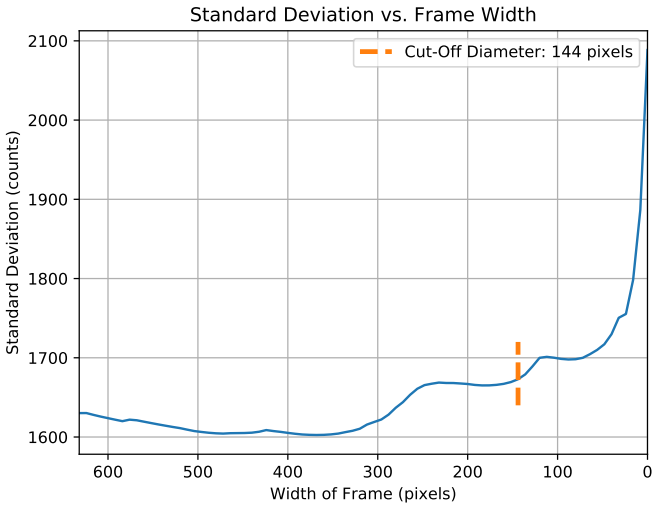

Fig. 3: Plot of Standard Deviation against Frame width. The width of frame refers to W_0, indicated in Fig.2, and the standard deviation of pixel value was calculated in the corresponding area A, again as illustrated in Fig.2.

where n is an integer varied between 1 and 5, and σ_p is the standard deviation of pixel value in the masked data. The parametrisation of n was explored at a later stage. The probability density distribution of pixel values in the masked data is displayed in Fig.4, with mean value 3421 and corresponding standard deviation $\sigma_p = 18$.

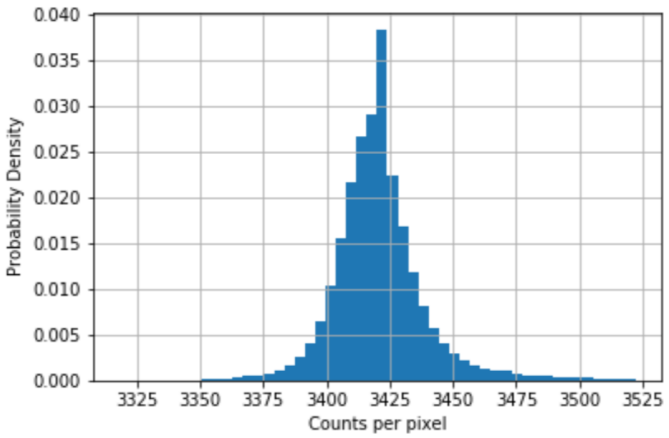

Fig. 4: The probability density distribution of pixel values in the masked FITS image clipped to 5σ. The plot conforms to a Gaussian profile where the average background count of the image corresponds to mean of the Guassian. The mean value is 3421 and the standard deviation $\sigma_p = 18$.

B. Object Detection and Classification

The central algorithm detected astronomical objects in descending order of the brightness of their central pixel. Given the unclipped and premasked dataset `image`, the built-in module Numpy was used to find the coordinate corresponding to the brightest pixel contained in the dataset by running the line `np.unravel_index(image.argmax(), imgage.shape)`. Once the brightest pixel had been located, the assumption was made that this pixel corresponded to the central coordinate of an astronomical object which could now be explored. The assumption was based on the spectral intensity profiles of galaxies, which follow a Moffat distribution [4] and hence have their brightest value in the centre. The approach taken to explore each detected object involved creating a circular aperture of variable radius around the central pixel. Both the size of object and the local background noise were then found by making measurements of the average pixel value around the perimeter of the aperture, and incrementing the radius until certain conditions were reached. Fig.5. illustrates how, given the coordinates of the central pixel, the surrounding pixels were then considered one increment of aperture radius at a time.

To determine the size of an object, a `while` loop was written in which, upon each iteration, the radius of the aperture around the central pixel was increased, a new set of points along the aperture perimeter was created, and the total sum of their pixel counts was calculated. An estimate for the background noise of pixels along the perimeter was found by taking the product of the number of pixels around the circle, with the upper limit of the expected background noise (9). The loop was broken once the difference between pixel count around the current aperture and the previous smaller aperture reached a value less than the expected background noise limit along the final aperture edge. A list of coordinates for the pixels contained within the object were then returned, along with the total sum of counts from the object and the final

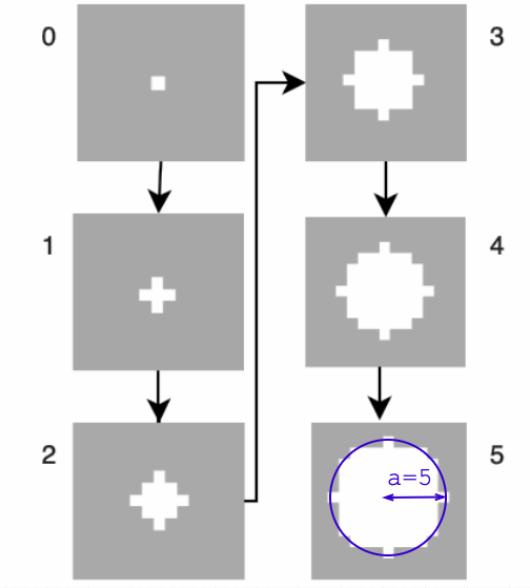

Fig. 5: Illustration of how, given a central pixel, the pixels along a circular aperture of incrementing radius a (the value of which is displayed beside each photo and illustrated in blue for $a = 5$) are masked. This aperture method was used in the algorithm for finding the size of each astronomical objects as well as their local background noise.

radius of aperture.

A similar procedure was followed for the calculation of the objects' local background noise. Upon each `while` loop iteration the radius of aperture was increased from a starting value of 6 in steps of A_p, the value of which was also parametrised later on. Once the mean count of pixels around the aperture reached a value less than the global background noise, the loop was broken out of and the final average count value was taken as the local background noise for the object under consideration.

Before applying these algorithms on the whole CCD Image, a fake image containing a star of known size and flux was created. The algorithms were applied to this image to ensure correct masking and flux calculations could be carried out. The fake image before and after masking is displayed in Fig.6.

To proceed, a small section of the true image was analysed and the algorithm was once again tested. The result is displayed in Fig.7.

Given the success of both validation tests, the algorithm was then applied to the whole image. The structure of the main algorithm is detailed by the Flowchart in Fig.12 of the Appendix. Information on the size, location, total pixel number, total count sum, and local background noise of each detected object were stored in a txt file referred to as the galaxy catalogue.

Once an object had been identified and its information had been calculated and stored, all pixels which had been identified as belonging to the object were masked to stop the algorithm from "re-detecting" the object as it searched for the next brightest pixel in the image. The algorithm was run until the next most bright pixel had a value within 4σ of the global

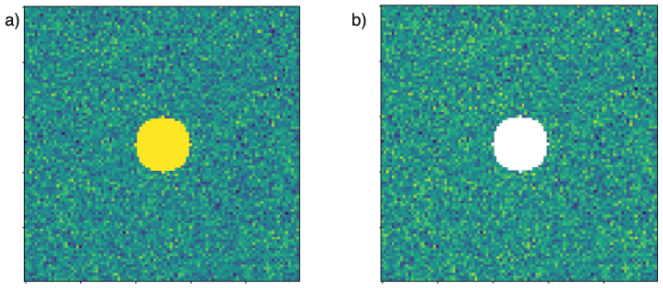

Fig. 6: As means of algorithm testing, a fake image of a star was created. The image contains a circle whose pixel intensity follows a Gaussian distribution. The object-finding algorithm was tested by checking that the full star could be successfully identified and masked. Image $a)$ displays the fake image, and $b)$ illustrates the image after the program was run: the full star has been successfully masked.

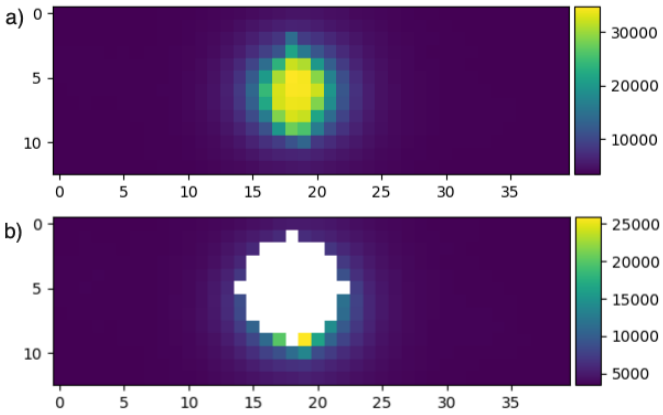

Fig. 7: A subsection of the true image was used to test the performance of the object-finding algorithm. Image $a)$ displays the chosen subsection prior to masking, and $b)$ displays the image once the program was run. The colourbar refers to the pixel count.

background noise.

In order to prepare the data for analysis, the count sum of pixels within each object was converted to a magnitude via the following relationship:

$$m = ZP_{inst} - 2.5\log_{10}(counts) \tag{10}$$

where $ZP_{inst} = 25.3 \pm 0.02$ is the instrumental zero point error.

IV. Results

Various catalogues were produced for a range of values of the parametrised variables: the aperture step for local noise determination A_p, and the number of sigmas n considered in determining the global background noise. It was decided that for the final galaxy count estimate, the parameter values $A_p = 6$ and $n = 2$ would be chosen. With these, a total object count of 1237 objects was obtained by the algorithm. The magnitude m of each object detected was calculated using (9) and was plotted against the logarithm of number of objects

with magnitude above m. It was assumed that the probability per unit solid angle of finding an object with magnitude below m was constant, and that the distribution of galaxies followed a Poisson distribution. As a result the error for each bin in the PDF distribution of objects against magnitudes was taken as $\sqrt{N}$. These errors were added in quadrature and scaled to obtain the cumulative errors on Fig.8.

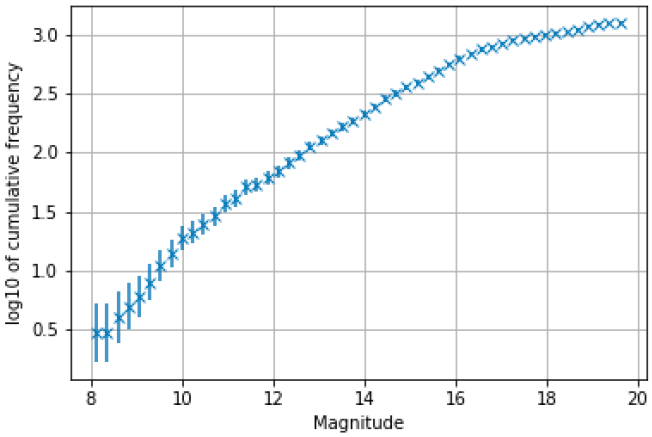

Fig. 8: A count-magnitude plot of the final results obtained using parameter values of $A_p = 6$ and $n = 2$, where the error bars have been calculated based on the assumption that the distribution of galaxies follows a Poisson distribution. The lower magnitude values, of unit a.u., correspond to brighter objects.

For magnitudes greater than 17, the slope of the data begins to approach a constant value. This is due to the detection limit of the survey; the objects with lowest detectable magnitudes have a magnitude similar to that of the background noise. It was decided that a more reliable count estimate would be obtained if analysis was done only on the linear region of the graph so as to ensure every object considered was definitely an astronomical object and not just background noise. The cut-off magnitude was calculated to be 16.4 using $4\sigma_{globalnoise} \times S$, where S is the average total pixel number of an object. The refined object count returned by the algorithm was 682, and the final count-magnitude relationship is displayed in Fig.9.

The section of Fig.9 corresponding to magnitudes below 10 deviates from the linear fit, suggesting that the particularly bright objects considered here are stars rather than galaxies. The gradient of the linear fit is 0.242 ± 0.004, which differs from the expected gradient of 0.6 found in Equation (8) by 59.7%.

A 2D spatial density plot of the results was created using a scatter graph of object position with aid of a Gaussian kernel density estimation technique to create a heat map of the resultant galaxy distribution [5]. This is displayed in Fig.10.

A spatial flux map of the image was also plotted, shown in Fig.13 of the appendix. The large variation in order of magnitude of $log_{10}(\Phi)$ where Φ is the flux of an object suggests that the original dataset was inhomogeneous. The assumption of an isotropic Universe remains unproven in this investigation as the distribution of galaxies within this small

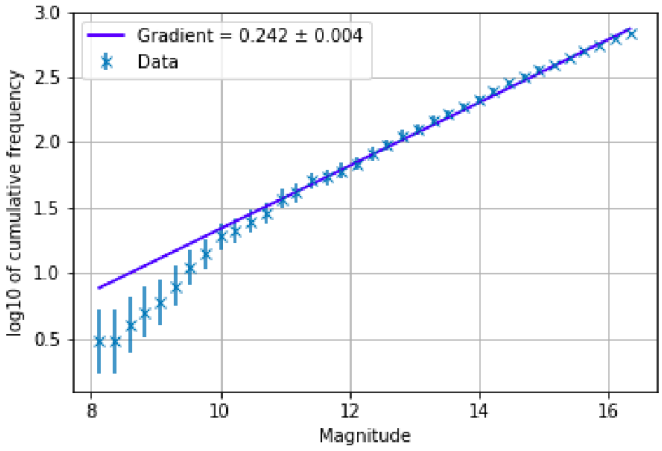

Fig. 9: A refined count-magnitude plot for the 682 astronomical objects detected. The line of best fit is plotted with gradient of 0.242 ± 0.004, with error bars calculated using propagated errors.

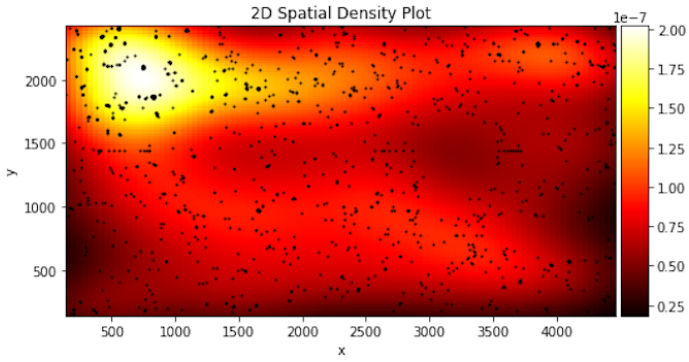

Fig. 10: The position of each objects' central pixel is plotted on a 2D scatter plot. A Gaussian KDE method is used to create the spatial density plot of the galactic distribution discovered using the object-finding algorithm explained in section III on the CCD Image under investigation.

section of the sky does not show a uniform distribution in either of the two spatial density plots.

An attempt was made to extend the algorithm to determine whether each detected object was a galaxy or a star by visualising the intensity profile across the centre of each object. The intensity distribution for a star should show a Gaussian or Moffat distribution [4], whilst the distribution for a galaxy would tend to have a Sersic profile [6]. The intensity distributions of a few objects were plotted however they showed no distinct shape; the results were not sufficiently clear. This will be due to the insufficient density of pixels in the given image. A further test would have been to calculate circularity of each detected object. This could be done by comparing each objects' bilateral symmetry to its four-fold symmetry. As elliptic objects were not considered in this algorithm, they will have been estimated as having a smaller flux.

In some cases multiple astronomical objects were located close together in the image, however this was not accounted

for in the algorithm. A different approach which could have been taken in the design of the object-identifying algorithm would have followed a computer vision method using pixel groupings. This would have been a much more effective method for the differentiation of nearby objects.

Upon first glace, this poor final slope result is in line with the concerns mentioned in Section I, regarding the ignorance of the underlying assumptions when deriving the count-magnitude relationship (8). The evolution of galaxies was predicted to have an effect on the results obtained in this study, however it can actually only be taken into account for magnitudes greater than 20 [7]. Insight into the reason behind the large error in the result was gained upon a comparison with the study by Yasuda N. et all. (2001) [8] which analysed images taken during the commissioning phase of the Sloan Digital Sky Survey. In this study over 900,000 galaxies were found, located across an image of $440deg^2$ of the sky. It was also found that the $log_{10}(N(m))$ vs m graph matches the predicted trend very closely for objects with a magnitude brighter than $m = 16$.

It is safe to conclude that the chance of obtaining a correct log relationship would only have been possible if this investigation was carried out with a significantly greater amount of data, as isotropy of the Universe can only be considered on the scale of millions of light-years [9].

V. Conclusions

A computational program to calculate the local Universe galaxy count was written in Python to analyse an image taken at the Kitt Peak Mayall 4m Telescope using a CCD mosaic camera with a Sloan r filter. Preliminary masking of the data was carried out to eradicate instrumental defects and artefacts that may have interfered with final information drawn from the image. A galaxy catalogue was built containing information on all galaxies successfully identified by the algorithm such as their size and magnitude. A count-magnitude graph was plotted and the gradient of the linear region was calculated. The result disagreed with the expected gradient by 59.7%, a disagreement so large that it could not be blamed on the effects of the expansion nor the inhomogenity of the Universe. It was concluded that an insufficient amount of data contained in the CCD image was to blame as the image covered only $0.0469deg^2$ of the sky, a significant deviation from the $440deg^2$ image used to calculate the expected result. A number of improvements to the object-detecting algorithm could have been made, such as the calculation of each objects' circularity and spectral profile. This would have allowed the algorithm to distinguish between stars and galaxies and therefore make sure it was only galaxies that were considered in the final data set. A computer vision approach could also have been implemented to better differentiate between objects in close proximity.

References

[1] T. Fujiwara, *"Magnitude Systems in Old Star Catalogues", Journal of Astronomical History and Heritage,*, 2004.

[2] M.-K. Sze, Simon Min; Lee, *MOS Capacitor and MOSFET". Semiconductor Devices: Physics and Technology. John Wiley Sons. ISBN 9780470537947*, 2012 Oct;25(4):239-46.

[3] Starzona, "Optimum exposures," Available at https://starizona.com/tutorial/optimum-exposures/ [Accessed: 19/12/2019].

[4] A. F. J. Moffat, "A theoretical investigation of focal stellar images in the photographic emulsion and application to photographic photometry," Available at https://ui.adsabs.harvard.edu/abs [Accessed: 4/1/2020].

[5] Scipy.org, "Scipy.stats.gaussian_kde," Available at https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.gaussian_kde.html [Accessed: 19/12/2019].

[6] J. L. S. (1963), *Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. Boletin de la Asociacion Argentina de Astronomia, Vol. 6, p.41*, 1963.

[7] S. D. M. P. F. P. B. T. e. a. Martin DC, Fanson J, *The Galaxy Evolution Explorer: A space ultraviolet survey mission. The Astrophysical Journal Letters. 619(1):L1.*, 2005.

[8] N. V. L. R. S. I. S. M. e. a. Yasuda N, Fukugita M, *Galaxy number counts from the Sloan Digital Sky Survey commissioning data. The Astronomical Journal. 122(3):1104.*, 2001.

[9] U. of Oregon., "Isotropy and homogeneity," Available at http://abyss.uoregon.edu/ js/cosmo/lectures/lec05.html [Accessed: 18/12/2019].

VI. Appendix

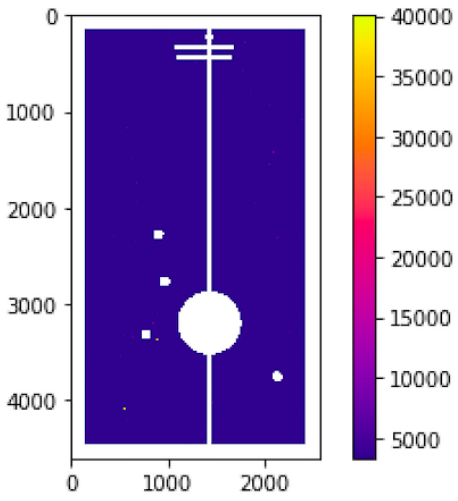

Fig. 11: The image after major defects have been masked. The brightest stars have been removed as well as a 144 pixel width around the image edge.

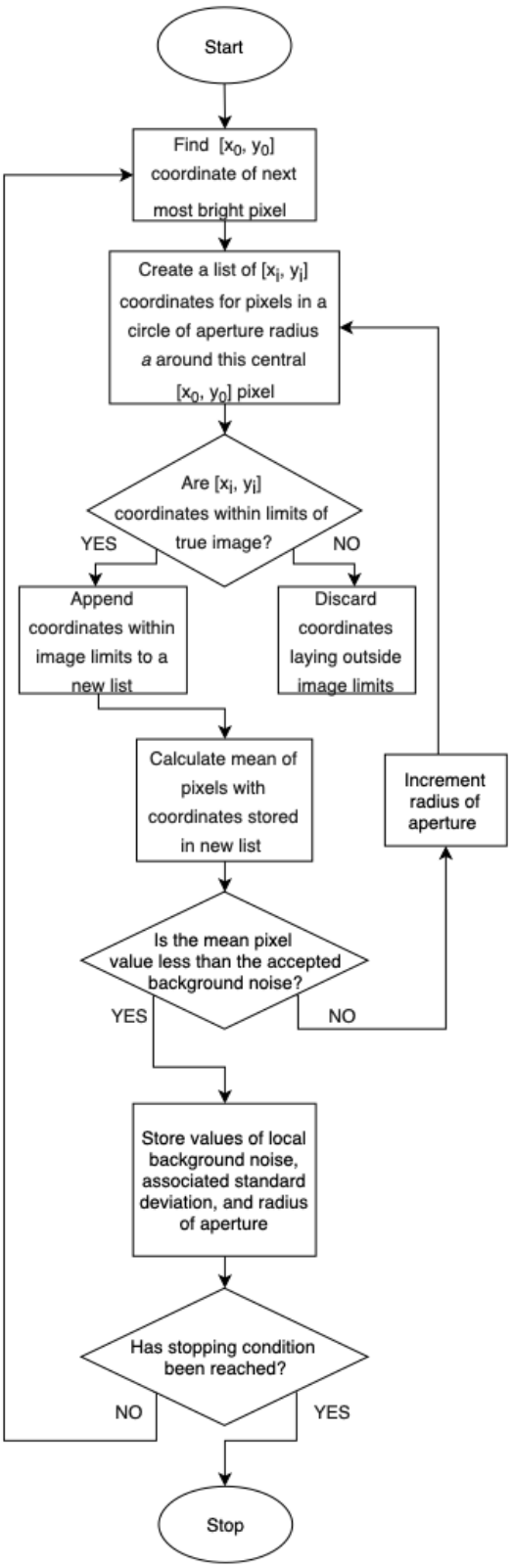

Fig. 12: Flowchart detailing the method used to find the local background noise of an object with central point $[x_0, y_0]$. Method consisted of creating a list of pixel coordinates along a circle centered on this central point, finding the average value of these pixels, comparing this with the accepted background noise (calculated by taking the mean of the whole dataset clipped to 5 sigma), and returning once the mean pixel value fell under that of the global background noise. The final mean pixel value corresponds to the local background noise around the astronomical object of centre $[x_0, y_0]$.

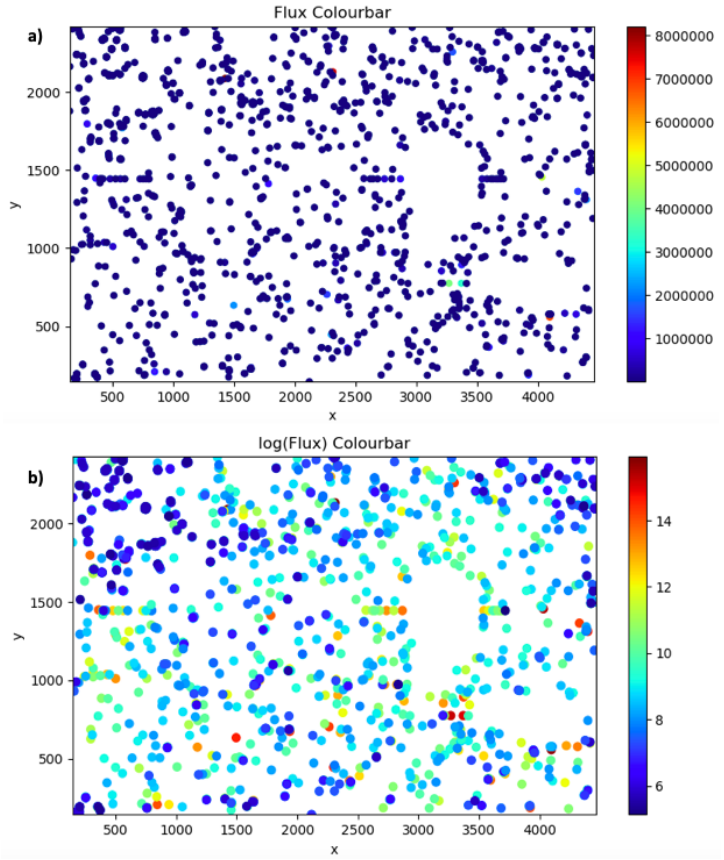

Fig. 13: A plot displaying the spatial flux density of detected astronomical objects for parameter values of $A_p = 6$ and $\sigma_p = 2$. The colourbar for image b) displays the logarithm of the flux value for each object, whilst image a) displays the raw Flux data.