University of Kent

Long-Term Time-Series Photometry in Cygnus

Author:
Niall Miller

Supervisor:
Dr Dirk Froebrich

November 14, 2023

Contents

Abstract

In this thesis we present the data reduction, database creation and photometric correction of a large, high cadence data-set provided by an amateur astronomer. We also discuss the potential science projects available from this data. The database and catalogue produced via the methods detailed in this report feature 19,858 stars within a 2.8° x 2.8° square centred on 317.0° $+46.5^\circ$ (J2000). Each star has $\approx$64,000 measurements collected between 2003-09 to 2009-09. The final, photometrically corrected magnitudes have an approximate error of $\pm\,0.025$ mag for bright stars and $\pm\,0.040$ mag for dimmer stars. The shortest cadence for the data is 1 minute (much of the data is spaced 1 minute apart with larger gaps due to nightly/seasonal observations). The data is 99% complete up to a magnitude of GAIA R $\approx$ 15 mag and begins to saturate at a magnitude of GAIA R $\approx$ 7 mag.

We have shown how the database was created with the intention of making light curves easy to retrieve. During this we also explored how certain features of the data-set, such as the seeing and resolution, were instrumental in the design features of the database, particularly when designing the software to match the same objects across multiple images.

We have also cross-matched the objects in this database with other publicly available databases such as GAIA, 2MASS and WISE in order to gain further information on the population present in this data-set. We may use the additional information provided by GAIA's astrometric data (parallax and proper motion) and stellar colour (provided by GAIA, 2MASS and WISE) to further investigate this stellar population.

We found that the flat frames used for calibration did not produce data of sufficient quality for accurate photometric measurements. A substantial amount of structure is present in some of the flat fields, as it is not possible to know if the structure is truly due to the optical path of the telescope or due to incorrect flat fielding methods. Hence, a photometric correction was performed. The correction procedure removed any systematic photometric offset caused by inhomogeneous flat frames. This was achieved with modelling the photometric offset in a given image that is present in non-variable stars. The model is a function of magnitude, colour and CCD position, and is subtracted from all stars in the image.

We have outlined some of the potential future science projects that can be performed with this data, and show that the database presented in this report is very good for conducting research in the field of time-based astronomy. A preliminary investigation of periodic variable stars was performed. It was

found that multiple different types of periodic variables are present in our catalogue such as W Uma binaries and Delta Cepheids. We also investigated the possibility of detecting exoplanet transits and found that it is possible to obtain a photometric accuracy high enough to detect exoplanet transits at the sacrifice of temporal resolution. We found that, if we bin our data to reduce the temporal resolution to 42 minutes, we have a 90% probability of detecting a hot Jupiter.

I would like to thank Dr Dirk Froebrich, for providing excellent supervision and guidance, not just for this project but for my future career as a scientist.
I would also like to thank Dr Timothy Kinnear for his continued help and support.
A much needed thank you to my family and friends for support and encouragement.

1 Introduction

The goal of this project is to prepare a large data-set provided by an amateur astronomer. This was done to allow for investigation into long-term time-series relative photometry of field stars in the region of Cygnus. In this project, we present the data reduction pipeline from raw science and calibration 'FITS' files to an indexed queryable database. We also show how the data was calibrated with the goal of performing time-series relative photometry.

Time-series relative photometry is the study of how the photometric properties of a stellar object may change as a function of time. Variable objects can be categorised into two subcategories, extrinsic and intrinsic. An intrinsic variable is a variable object whose source of variability is from the object itself, such as Delta Cepheid variables or FU Orionis-type stars. Extrinsic variables are variable objects whose perceived variability is due to the geometry between the observer, the object and some other object which perturbs the observation of the original object. An example of an extrinsic variable is a planetary transit, where the planet is the other object blocking light from the star, thus making the star appear variable (Charbonneau et al., 2000). The General Catalogue of Variable Stars (GCVS) (Samus' et al., 2017) lists seven main groups of variable stars (grouped according to their reason for varying); eruptive, pulsating, rotating, cataclysmic, eclipsing binary system, intense variable X-ray sources and uncategorised/miscellaneous stars. The GCVS version 5.1, which is the most up-to-date at the time of writing, contains 870,571 variable objects.

Time-series photometry and the study of variable stars is an essential and rapidly growing branch of astronomy. Most stars that are not on the main sequence vary to some capacity (e.g. Gautschy & Saio (1995) & Gautschy & Saio (1996)). Studying a star's variability can give key insight into a star's formation, environment and evolution. The detection of extra-solar planets is also possible with time-series photometry (Seager & Mallén-Ornelas, 2003). Most extra-solar planets are discovered via planetary transits which is achieved with the use of time-series photometry. The many features of stars in binary systems can be determined using their orbital period with their neighbour which is measured via time-series photometry (e.g. Manfroid et al. (1987) & Pustylnik (1998)). Standard candles are used in astronomy as a way of measuring distances. Often determining a distance in astronomy is difficult as a star's magnitude is often extincted due to interstellar matter, the effect of which is often little-known, hence making distances harder to determine. Delta Cepheid stars are variable stars which exhibit a period-luminosity relationship, that is, the period of a Delta Cepheid's

variability is proportional to is absolute magnitude. Thus, if we use time-series photometry to measure a Delta Cepheid's period, we can determine its absolute magnitude. Comparison between the star's absolute magnitude and apparent magnitude will yield information about any present extinction and distance to the star. The above examples are not comprehensive and as previously stated, there are many types of variable objects.

The data presented here consists of multiple years of continuous, very high cadence observations of a region of field stars looking through the galactic plane. The area we have used features $\approx 50,000$ field stars, allowing for study of multiple different types of objects such as contact binaries and Delta Cepheids. As this data is of a very high temporal resolution ($\approx$ 1 min) and over a long period of time ($\approx$ 6 years), the period and other features of variability of objects can be determined with high accuracy. We also show how the data can be binned, exchanging temporal resolution for increased photometric accuracy (see Sect.6.3.3).

The properties of an object's variability can change significantly depending on the source of variability. For example, a W Ursae Majoris (W UMa) variable is a relatively common low mass contact binary of typically F, G or K type stars. W UMas generally have an orbital period of less than a day with an observed amplitude (Δm) of a few tenths of a mag. A common area of study for a W UMa type variable is the mass transfer between the two contact binaries as well as interacting magnetic fields. This is known as the Applegate mechanism (Applegate, 1992) and it can be studied by measuring the rate of change of the W UMa's period. Typically this can be on the order of $\frac{\Delta P}{P} \approx 10^{-5}$. Given these properties, one may design an observational study of W UMas to be sets of high cadence data with large multiple month/year gaps in-between. Thus one might only produce single filtered data of multiple regions of field stars in order to increase sample size.

However, if a survey was aimed toward studying FU Orionis-type stars, the features of the survey would change significantly. FU Orionis type stars are pre-main-sequence stars commonly found in star-forming regions, which have displayed an extreme change in their magnitude and spectral type, known as FUor events (Herbig, 1966). Typically, this change happens over the course of approximately 1 year and is expected to last on the order of decades, however, no FU Orionis stars have been observed returning to their original state. Hence, if one was to design an observation study for FU Orionis stars it is likely that a higher cadence would be exchanged for a wider field of data in multiple different star-forming regions to increase the probability of detection. This is necessary as FUor events are rare with fewer than 13 such events being discovered in the past decade (Hillenbrand et al., 2018). Due to

such a high variation in the properties of variable stars, there have been many surveys and missions aimed at studying them.

We will discuss some of the projects that have been undertaken with the aim of studying time-series photometry. By no means is this a comprehensive list; there have been many projects and surveys aimed at studying time-domain astrophysics.

1.1 Previous Time-series Projects

1.1.1 UKIDSS and VVV

The UKIRT Infrared Deep Sky Survey (UKIDSS) (Lawrence et al., 2007) is a large-scale near-infrared (NIR) survey conducted on the United Kingdom Infrared Telescope (UKIRT). UKIDSS, which began operations in May 2005, serves as a successor to 2MASS (Section 4.3) and surveyed 7500 $\deg^2$ of the Northern sky. UKIDSS consists of five individual surveys, of which, the UKIDSS Galactic Plane Survey (GPS) is one (Lucas et al., 2008). GPS surveyed 1868 $\deg^2$ of the Galactic plane with Galactic latitudes $|b| > 5°$ in the J, H and K filters. The GPS provides two epochs of K-band photometry and has an aim of investigating phases of stellar evolution via the detection of high amplitude NIR variability (Contreras Peña et al., 2014). GPS will investigate eruptive young stellar objects (YSOs) also known as FUor and EXor events (Contreras Peña et al., 2014; Lucas et al., 2017; Montmerle, 1990). As the GPS is a large area but only two epochs, it is suitable for the study of such high amplitude variability with long quiescence periods. The main caveat of the GPS is the fact that it only features two epochs of data. Hence differentiating between FUor events and other sources of variability, such as Miras and Novae, proves difficult. Another caveat of the GPS is the relatively low dynamic range. The conservative saturation limit for GPS is $m_k < 12.0$, $m_H < 12.75$ and $m_J < 13.25$ and a 90 % completeness of $m_K = 18.0$, $m_H = 18.75$ $m_J = 19.5$ (Lucas et al., 2008). If we consider that an average FUor event has an increase in brightness of ≈ 6 mag this significantly reduces the range of FUor events that can be fully studied.

GPS, and the UKIDSS more generally, was not initially proposed for time-series studies, which were considered after the project had begun. In Contreras Peña et al. (2014) it is discussed that GPS was largely used as a precursor for the VISTA Variables in the Vía Láctea (VVV) survey (Saito et al., 2012). The VVV survey, which started in 2010, is a multiple epoch survey with observations carried out on the 4-meter VISTA telescope in the $Z, Y, J, H \,\&\, K_s$ filters. VVV has a target area of 562 $\deg^2$

centred on the galactic bulge and adjacent plane region. The survey has $\approx$ 80 epochs of data in the K_S band. The VVV survey saturates at $K_S \approx 12$ and is 90 % complete at $K_S \approx 16.8$ however it is noted the magnitude limit is strongly dependent on the crowding of the field. Due to its multiple epochs in and around the galactic centre, VVV is useful for a wide array of studies, particularly in time-domain astrophysics. In Ferreira Lopes et al. (2020) it is discussed that the 4^{th} VVV data release (VVVDR4) has NIR light curves for 288,378,793 sources. After analysis, it was determined that 44,998,752 of them are variable star candidates. However, the caveats of VVV lie within it having a highly complete sample in a densely populated area. It is reported in Ferreira Lopes et al. (2020) that 1 in 10 variable stars suffer photometric contamination from none variable stars.

Furthermore, as VVV is only NIR, classifying the sources based on colour is difficult. It is stated in Contreras Peña et al. (2017) that AGB stars are the largest source of contamination when studying YSOs and comparison of the star's variability is often used to identify contaminants. Much like in UKIDSS, differentiating between YSOs and AGB stars is difficult. Both YSOs and AGB stars have IR excess caused by surrounding material being heated by the star. In the case of YSOs, this material is the circumstellar disk while for AGB stars this is the cooling atmosphere. AGB stars radiate in IR due to the thick circumstellar envelopes.

1.1.2 HOYS

The Hunting for Outbursting Young Stars (HOYS) project aims to produce long-term, multi-filter, high cadence monitoring of large samples of YSOs (Froebrich et al., 2018b). YSOs were originally discovered due to their large amplitude optical variability (Joy, 1945).

It was found that YSOs vary due to changes in accretion from the disk to the star, varying amounts of obscuration from an inhomogeneous disk as well as quasi-periodic variations due to hot spots from accretion on the surface of the rotating star and cold spots similar to those found on the sun. As such, studying these periods can help gain insight into the angular momentum of YSOs and subsequently their interactions with their circumstellar disks, giving insight into the evolution of the protostellar stages and beyond.

The HOYS project obtains its data via many observatories. Much of the data is from willing amateur astronomers whom are provided with a list of targets, any images of which they can submit to the HOYS project. The targets are chosen as they are star forming regions that are also generally easier to observe and more popular with amateur astronomers, such as the Pelican Nebula and the

Elephants Trunk Nebula (see Froebrich et al. (2018b) and Froebrich et al. (2018a)).

The amateur images are calibrated in such a way as to obtain magnitude measurements appropriate for science use (see Froebrich et al. (2018b)). The Beacon Observatory accounts for 27 % of all images in the HOYS project and hosts a 17" astrograph with Johnson-Cousins B, V, R, I & H_α filters. The Beacon Observatory is based at the University of Kent and aims for nightly observations, weather permitting.

The main complication of HOYS arises from the use of data from other telescopes, particularly amateur data. As the data obtained from different telescopes has a different optical path and potentially a different filter, a star's observed magnitude can likely vary as a function of the observer. Hence, some software calibration is required to decrease the inconsistencies across observers (see Froebrich et al. (2018b)). As stated in a discussion of a comparable methodology in Sect. 5, this software calibration is imperfect and still induces some errors. Secondly, the images from different observatories have different resolutions. This can be problematic in crowded fields where lower resolution images may merge multiple stars into one. Hence, it is difficult to identify the same star across multiple images taken from different sources. An issue similar to this is discussed in Sect. 9 except in that case, they are all the same source. However, given the higher variation in the resolution of the data provided, implementing the same methods discussed in Sect. 9 is significantly more difficult.

Another problem for the HOYS project is the inconsistency in how fields are observed. As all of the observatories present are ground-based, they are subject to the movement of the earth. The targets on the target list are in or close to the celestial equator, as a result, there are times of the year where targets are in-observable due to the geometry of the earth, sun and target. This leads to inconsistencies in the amount of data for each target. Consequently, comparisons between the population of YSOs are difficult and often statistically limited by smaller sample size. However, this does have the benefit of reducing any aliasing in the data as a function of time. If an area is routinely observed at a set cadence, this will mean that when a search for a variable star's period is performed, that cadence, and multiples of it, will become apparent, such aliasing is removed with randomly spaced observing cadences.

1.1.3 PTF and ZTF

The Palomar Transient Factory (PTF) was a time-domain survey designed to search for variable stars, supernovae and comets (Law et al., 2009). The survey ran from 2009 to 2012 and was performed on the Samuel Oschin Telescope at the Palomar Observatory. The nature of PTF was to perform an R-band 5 d cadence search for transient objects such as novae and cataclysmic variables. PTF was fitted with a Mould-R filter which is very similar to the SDSS r' filter. PTF had a data reduction pipeline which performs near real-time data reduction to allow for the identification of transient objects within minutes of observation. The PTF provided followup multi-band observations with a large fraction of the 60-inch telescopes at the Palomar site. Later, the PTF changed to the Intermediate Palomar Transient Factory (iPTF) (Cao et al., 2016) with an upgrade to the data reduction pipeline and the inclusion of some g-band observations. The iPTF operated until 2017 where it was superseded by the Zwicky Transient Facility (ZTF) (Bellm et al., 2019).

ZTF, while being substantially different to the Large Synoptic Survey Telescope (LSST), is discussed as acting as a precursor to the LSST (Bellm et al., 2019). ZTF observes with a cadence of as high as 38.3 seconds which allows for the study of asteroid light curves. ZTF improves on PTF and iPTF with a new set of filters, ZTF-g, ZTF-r, ZTF-i. All of PTF, iPTF and ZTF have a 95 % completion limit of $m_r \approx 20$. PTF and iPTF are saturated at $m_r \approx 15$ and ZTF saturates at $m_r \approx 12.5 - 13.2$. PTF, iPTF and ZTF have a median image quality of $FWHM \approx 2.1$".

Similar to UKIDDS, the relatively shallow depth of these surveys means that some transient events may only have measurable magnitudes within the completeness of the telescope for half of their activity. As the Samuel Oschin Telescope is a ground-based telescope situated in southern California its observing time and survey area is largely determined by the motion of the earth. ZTF's lack of a range of filters means that investigation into a lot of the variability in colour is severely limited, while follow up observations can perform multi-band observations, these will not have the same cadence of ZTF. This problem was greater for PTF and iPTF which observed predominantly in a single filter.

As PTF, iPTF and ZTF rely on fast data reduction, they are limited by computational power. In order to achieve data reduction at the speed required for fast observational follow-ups, high-performance computing is necessary. All of the data processing required for fast follow up for PTF and iPTF was performed at the National Energy Research Scientific Computing Center (NERSC) [1] which is located at the Lawrence Berkely National Laboratory, ≈ 500 miles separated from the Palomar Ob-

[1] atnf.csiro.au

servatory. ZTF uses the Infrared Processing and Analysis Center (IPAC) [2] located at Caltech which is $\approx$ 120 miles separated from the Palomar Observatory. Having such a dependency on high-performance computing so physically separated from the observatory of these surveys introduces a great potential point of failure, which would lead to wasted observing time. It is stated in Bellm et al. (2019) that the High Performance Wireless Research and Education Network (HPWREN) is used to transfer images with typical transfer times of $<$ 25 seconds. While this is sufficient to keep up with the observing cadence of up to 38.3 seconds, it relies on a second external system operating perfectly in addition to the observatory.

1.1.4 COROT, TESS and KEPLER

The original stated aim for the data-set discussed in this report is to detect exoplanets via the transit method. We discuss the possibilities of detecting transiting exoplanets in Sect. 6.3.3. Due to the nature of an exoplanet transit, a high signal-to-noise is required to reliably detect transits and multiple observed transits are often required to reliably determine features of the system. In Sect. 6.3.3 we show how the detection of a Jupiter mass planet orbiting around a solar mass star would require a photometric precision of 0.005 as the change in flux is $\frac{\Delta F}{F} = 0.01$. The sources of photometric uncertainty can be internal or external. The internal sources of photometric uncertainty, such as dark current, can mostly be corrected for with proper calibration techniques. External sources of photometric uncertainty can not always be removed from ground-based observations. For instance, differences in the seeing in each image can cause differing amounts of light to be captured by the `Source EXtractor` software. The PSF of the star may vary as a function of atmospheric conditions (seeing). This can cause contamination from neighbouring stars, particularly in crowded areas, where the PSF of one star grows sufficiently high that it can overlap with the photometry of other sources.

While these errors are unavoidable for ground-based astronomy, selectively searching for transits with larger changes in flux as well as the binning of data can increase the probability of accurately detecting a planetary transit (as seen in Sect. 6.3.3). Removal of many external sources of error, such as those caused by the atmosphere, space based telescopes, and the high photometric accuracy they provide, are of interest when operating an exoplanet transit-oriented project.

CoRoT was a space telescope aimed at investigating stellar seismology and extra-solar planetary transits (Auvergne et al., 2009). CoRoT entered a polar orbit of the earth in December 2006 at an

[2] www.ipac.caltech.edu

altitude of $\approx 900\,\text{km}$. CoRoT features two channels for its optical path, one for astroseismology (AS) and one for planetary transits (PF). The PF channel features two CCDs each with 400 windows of 10 x 10 pixels, each CCD has a pixel size of 2.32 arcseconds and the telescope has a field-of-view of $2.7^\circ \times 3.05^\circ$. It is reported in Auvergne et al. (2009) that this setup achieves a photometric precision of 0.4 %. The smallest transit depth detectable by CoRoT is $\approx 6 \times 10^{-4}$ for a star of magnitude $m_R = 12$.

The Kepler space telescope is a retired space Schmidt-type telescope launched by NASA in March 2009 with the aim of detecting Earth-sized planetary transits. Kepler hosts an array of 42 CCDs covering the telescopes $100\,\text{deg}^2$ field-of-view (Borucki et al., 2003). Kepler saturates at 9^{th} visual magnitude and is complete up to 15^{th} visual magnitude. Kepler has a photometric accuracy of 20 ppm. As of September 15^{th}, 2020, Kepler has 5,481 exoplanet candidates with 4,276 of them as confirmed exoplanets [3].

The Transiting Exoplanet Survey Satellite (TESS) is a space telescope operated by NASA that was launched in April 2018 (Ricker et al., 2015). TESS effectively succeeds Kepler in that it continues the search for exoplanet transits. TESS hosts 4×4 megapixel CCDs and a $2,300\,\text{deg}^2$ field-of-view. TESS, like Kepler, saturates at 9^{th} mag and is complete up to 15^{th} mag with a target photometric precision of 50 ppm.

The main caveat of space-based telescopes is the lack of human intervention once they are operational. This can lead to irreparable issues and, due to finite fuel resources, generally creates an unavoidable time of operations for all space-based telescopes. As a result of this, space-based telescopes are rarely fully operational for the same amount of time as ground-based telescopes. This means that the variability of stars on longer time scales, such as FU Ors, is more difficult to investigate via space based observatories. Further to this, the limited time of operations for these telescopes means the prediction for many transits can have a large uncertainty. The lower operations times means less repeated observations of these transits and thus higher uncertainty. To meet the requirements for photometric precision, these telescopes are space-based. This means it is virtually impossible to independently confirm their findings.

In March 2009 CoRoT suffered a loss of communications with one of its Digital Processing Unit (DPU) and hence loss of communications for one of the CCDs from each of the AS and PF channels effectively halving the field-of-view for CoRoT. Due to the nature of this telescope, repairs are not an

[3] exoplanets.nasa.gov

option and hence CoRoT had permanently lost access to one of the two DPUs. In November of 2012, CoRoT suffered a computational failure which rendered the retrieval of observational data impossible. Thus, in June of 2013 CoRoT was decommissioned and its orbit lowered to allow it to burn up in the atmosphere. In July 2012 one of the four reaction wheels used for fine pointing of the Kepler telescope failed, with a second reaction wheel failure in May of 2013. These failures severely hindered the Kepler space telescope as a minimum of three wheels are required for the telescope to accurately point. This ultimately lead to the end of planetary transit hunting for the Kepler space telescope and in 2014 Kepler was re-purposed for more general transient detection. This phase of operation (K2) relied on solar radiation pressure to accurately position the telescope and was in operation until all RCS fuel was used in November of 2018.

1.2 Calibration and Data Reduction

Calibration and general data reduction steps are a requirement for all photometric observations aiming to produce reliable results. When comparing data of consistent quality with the same source, these procedures are trivial. Often, extra considerations are required to ensure the quality of the data is of the highest achievable standard. These extra steps are usually a byproduct of the design of the project. For example, with the CoRoT satellite, due to its polar orbit, it was necessary to take into account the 'South Atlantic Anomaly' (SAA). The SAA is an area in the south Atlantic where the inner part of the Van Allen belt is as low as 200 km in altitude. This leads to satellites transiting the SAA being bombarded with protons with energies in the interval of 10 keV to 300 MeV and imparting on average 5 keV (Auvergne et al., 2009). These impacts can leave semi-permanent hot pixels which can take time to return to usual. This means that an up-to-date profile on how the CCD functions on the pixel scale would be necessary for CoRoT to accurately produce photometric measurements.

If one is performing relative photometry where the photometric measurements are from different sources, extra calibration is required. In Evitts et al. (2020) it is shown how the HOYS project deals with accepting data from multiple amateur astronomers. This method focuses on removing any differences between the observational setup of all of the participating observatories, such as filters. This is achieved by modelling these differences as polynomials and using this model to subtract any differences. This process produces data with a suitably low photometric uncertainty that would be otherwise unusable. The HOYS project amateur data has a typical uncertainty of 0.2 mag for fainter objects and 0.05 mag for brighter stars (Froebrich et al., 2018b), making it suitable for the types of

variability it investigates.

The data-set we are discussing in this thesis was produced by an amateur astronomer, taken with the intention of finding exoplanet transits. As such the astronomer designed his observations to be of very high cadence data with a focus on a high-density area of field stars. This was done to increase the probability of finding an exoplanet transit. A substantial part of this report discusses how data were properly calibrated and organised into a practical, accessible format. Here we will discuss the less than adequate calibration data (flat field measurements) that were provided, and the impact this calibration data had when used to perform data reduction on the science data. We will discuss the steps we have taken to reduce the impact of this (Discussed in Sect. 5).

2 Data

2.1 The Cygnus Project and Imaging data

Amateur astronomer Mr Waterman approached Dr Dirk Froebrich regarding the Hunting Outbursting Young Stars 'HOYS' project (Froebrich et al., 2018b) [4]. The HOYS project is a citizen science project aimed at working with amateur astronomers to perform long-term photometric observations of young stellar clusters. The amateur astronomers can submit images of an object in the HOYS target list. Although Mr. Waterman had intended to provide data to the HOYS project, his data chiefly constituted high cadence observations of a region in Cygnus which were not a HOYS target. Nevertheless, it was decided that the data would be worth investigating and thus will be fully calibrated and analysed.

Mr Waterman operates a fully automatic observatory located North East of Luton, England (51.934 179°N, -0.302 138°E). From this observatory he runs the 'Cygnus Project'[5]. The Cygnus project started as 'The Planet Project' where Mr Waterman was exclusively looking for exoplanets. Hence this data was taken with the intention of being high cadence and featuring a high number of field stars (to maximise the likelihood of detecting a transit). During the analysis of his data at the start of the project, Mr Waterman decided to expand the search to all variable stars. The field observed has the galactic coordinates of 88.175° +0.761°, thus it can be considered to be within the galactic plane. A mosaic of nine images centred on 317.0° +46.5° (J2000) each with a field-of-view of 2.8° x 2.8° was used for this project. Figure 1 shows a DSS2 colour image with a 17° x 17° FOV in order to provide context to where the images are situated. Figure 1 also shows a rough outline for each of the targets with the labels used by Mr Waterman. While we were provided with all images for all nine targets this report only focuses on image region 'a', the centre most region, and the region with substantially more images available. Table 1 shows the amount of images belonging to each region.

The Cygnus project started in August 2003 and the data we received runs up to September 2009. A full distribution of the data, as well as a display of when each of the master calibration frames were taken, can be seen in Fig. 2. Notable is that Flat and Dark frames where collected regularly throughout the six-year period, while the Bias frames were collected over a span of only three months. From here we will be referring to the collection of non-indexed data provided to us as the 'data-set'.

[4] astro.kent.ac.uk/ df/hoyscaps

[5] stanwaterman.co.uk

Table 1: Shows the amount of images taken for each region

Region	Image count
a	64,293
b	413
c	584
d	366
e	716
f	240
g	421
h	643
i	750

Some seasonal variation can be seen in the distribution of science images. Cygnus is at it's highest transit altitude, and thus longest visible, in October.

The data-set was divided into 'nights' of observations with each night containing on average 260 science frames. Figure 3 shows a distribution of how many images are taken per night. It can be seen that the majority of nights have between 0-400 images with nights with more than 7 hours of exposure time becoming increasingly rare.

Figure 4 shows an example science frame taken on 2003-10-16 at 18:08 UT. Here, we can see the vignetting for each image, this is predominantly caused by the large CCD size. The vignetting is apparent in the corners of the image. Here, the images appear to feature fewer stars, however, this is just due to the CCD being exposed to less light. As each image contains $\approx$50,000 field stars with no obvious structure, Fig. 4 appears noisy.

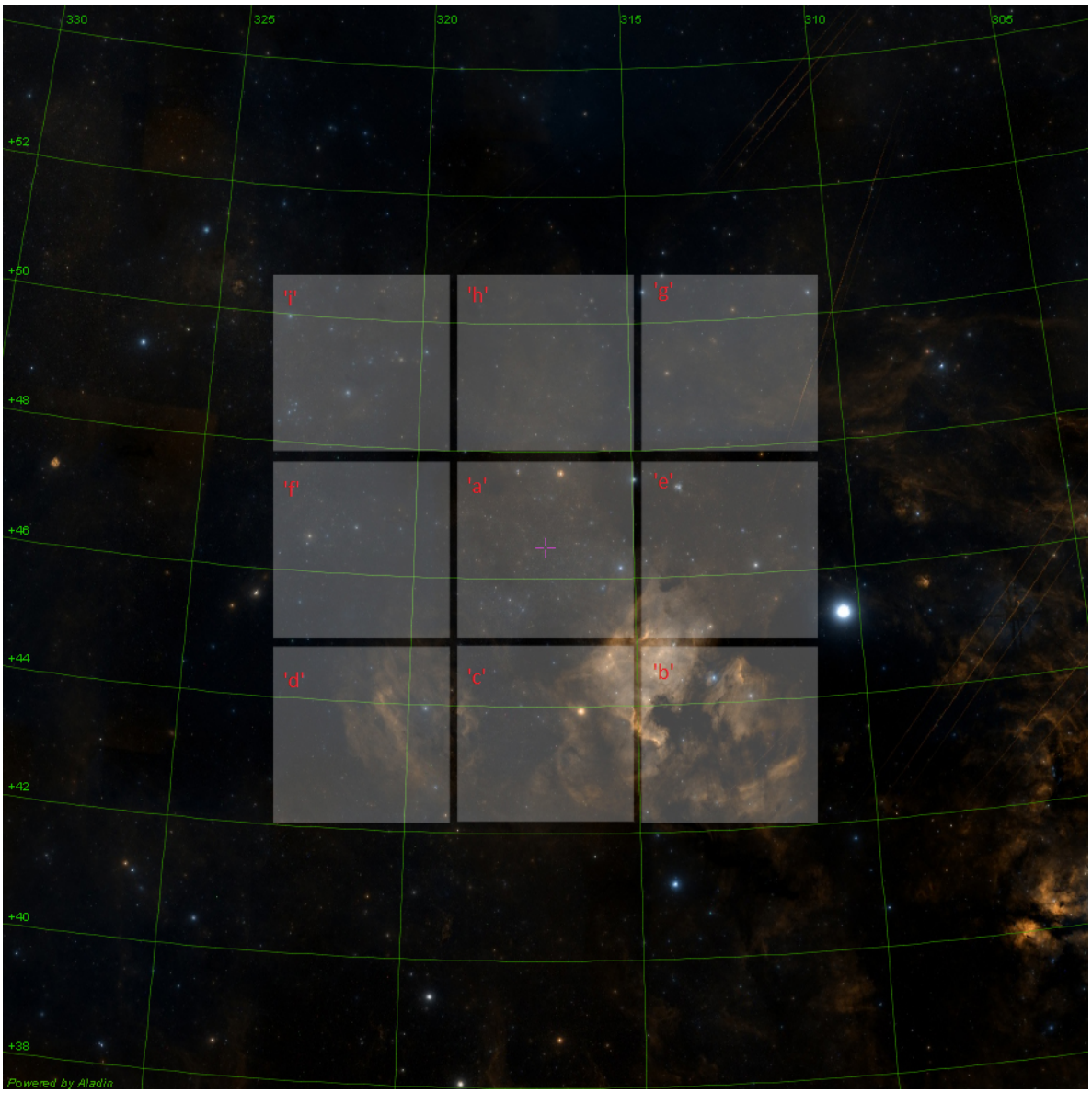

Figure 1: DSS2 colour image centred on 317.0° +46.5° (J2000) with a FOV of 17° x 17°. The overlaid squares represent a rough outline of the regions provided by Mr Waterman with their corresponding labels. It can be seen that the North American and Pelican nebula is south of the field. Deneb can be seen at 310.35° +45.28° (just outside of region 'b'). The Cocoon nebula can also be seen at 328.35° +47.26°.

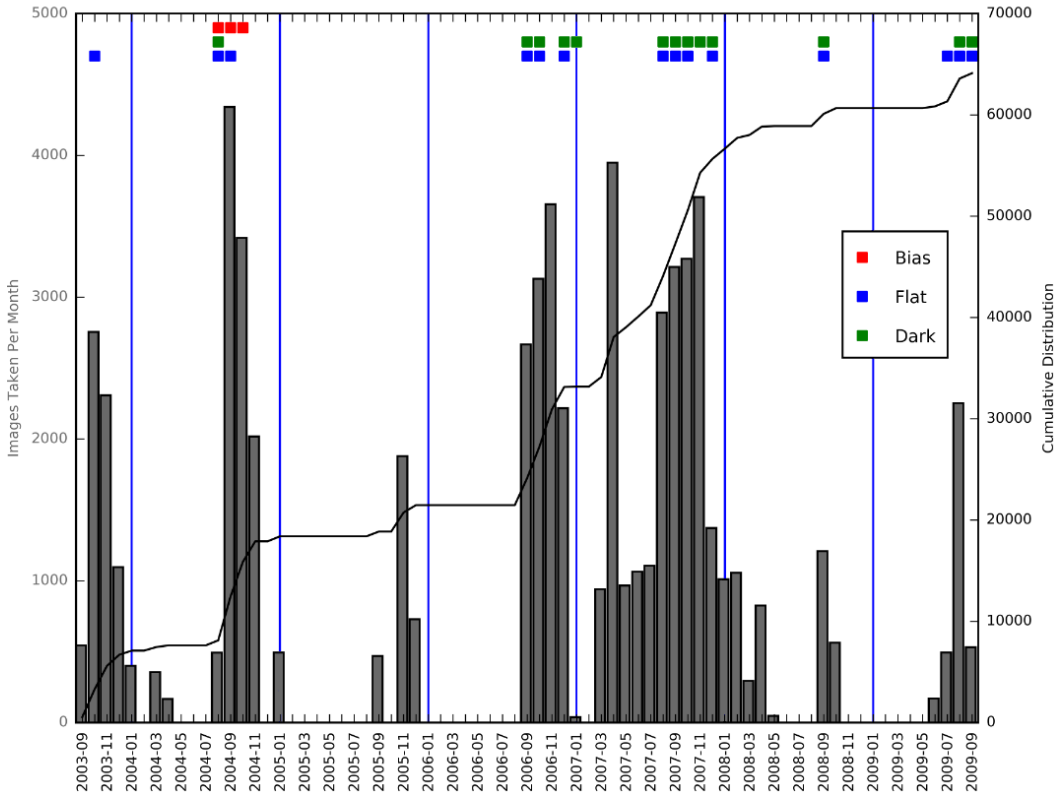

Figure 2: Showing the distribution of images for the set we are using taken per month over the whole dataset. The grey histogram and cumulative function show the number of images taken each month. When each calibration frame was taken can also be seen towards the top of the plot. The vertical blue line shows the start of each year.

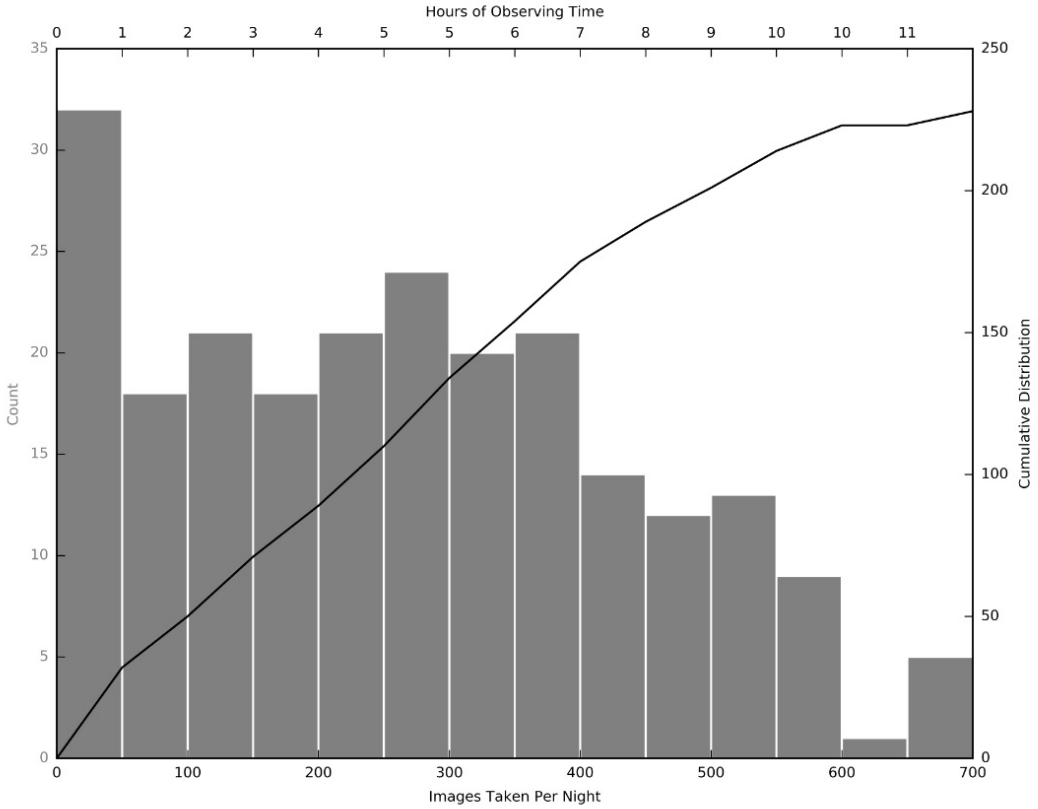

Figure 3: Showing the distribution of science images taken per night. The hours of exposure time include the ≈ 30 second CCD readout time and are calculated as 1 image per 60 seconds.

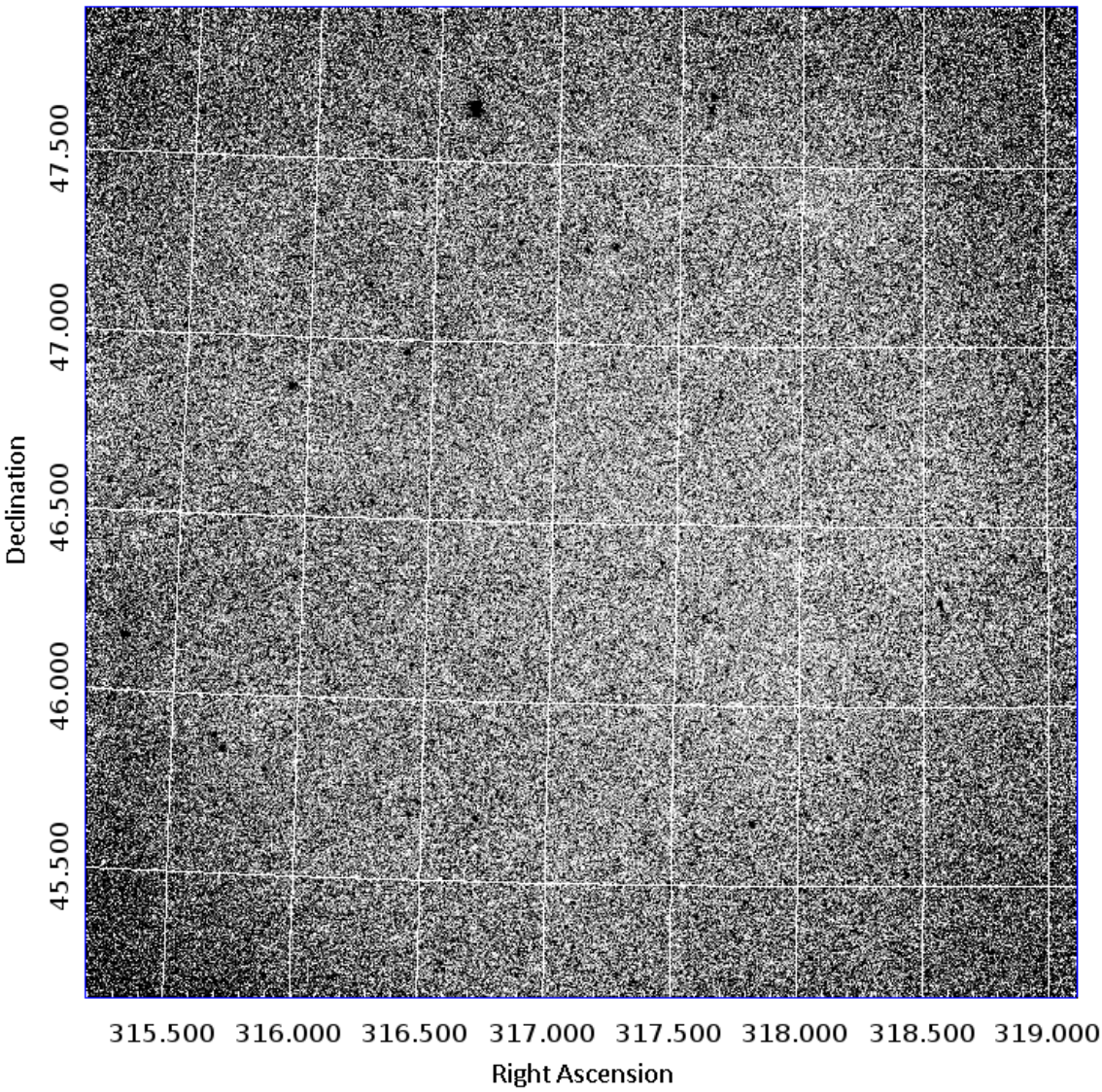

Figure 4: Example image from the data-set scaled with the inverted `IRAF` ZScale (Tody, 1986) algorithm taken on 2003-10-16 at 18:08 UT with a 30 second exposure on a 5" refractor hosting an Astrometrik Red filter. The coordinate system of (J2000) is overlaid.

2.1.1 Equipment

The data was taken on a 5" achromatic refractor with a focal length of ≈ 750 mm. The telescope hosts an Apogee Kaf 16801E CCD held at $-20°C$ with a size of 4096x4096 giving a resolution 2.48 arcseconds per pixel. The angular resolution of the telescope with the red filter used is calculated with Eq. 1.

$$\alpha = 1.22 \cdot \frac{\lambda}{D} \tag{1}$$

A Johnson-Cousins 'R' filter was used for the entirety of this data-set. The Jonson-Cousins 'R' filter has a peak wavelength of $\lambda = 634.9\,\text{nm}$. The telescope used has a diameter of $D = 0.127\,\text{m}$. We can use Eq. 1 to obtain the diffraction limit of the telescope at this wavelength. This gives a diffraction limit of 1.258".

This data-set was taken with the intention of searching for periodic variability such as exoplanet transits. Thus observations were made with the goal of creating a high cadence data-set. The images were mostly taken at a 30-second exposure although a 20 second exposure time was used from 2006-09-08 to 2006-11-20. The CCD has an average readout time of ≈ 30 seconds this gives a sampling rate of ≈ 1 minute and thus the temporal resolution of ≈ 2 minutes due to Nyquist sampling.

2.2 Image reduction and Initial Calibration

2.2.1 Files

The data was provided in the form of 64,293 FITS files each with a size of ≈ 31.7 Mb with a collective size of 8.6 TB. The original FITS headers contained information regarding the time and date of observation as well as exposure time and CCD temperature at the time of exposure. Given that each FITS file contains measurements for $\approx 50{,}000$ stars and there are 64,239 FITS files, a star well within the completeness limit of this data-set will have been measured 64,239 times. We can consider each individual measurement of a star to be a data-point, this gives up to 3,214,650,000 data-points. During this report, we will refer to a single measurement of a star as a 'data-point'.

2.2.2 Calibration Frames

The data was provided with 11 master bias frames, 15 master dark frames and 19 master flat frames (see Fig. 2 for distribution). Each science frame was bias and dark subtracted as well as flat-field corrected. Each science frame was calibrated using the calibration frame taken closest in time.

Each master bias frame is comprised of a stack of individual bias frames taken at either $-20°$C or $-25°$C. The number of individual bias frames used to create a master bias ranges from 1 to 30 frames. The bias frames were taken as dark frames with the telescope shutter closed at 1 second exposure.

The master dark frames are comprised of a stack of individual dark frames that have been bias subtracted. The bias frame taken closest in time to the dark frame is used for subtraction. Each dark frame used to create the master frame was a 30 second exposure taken at either $-20°$C or $-25°$C. Each master dark is comprised of between 30 to 100 individual frames.

Like the dark frames, each master flat frame was comprised of between 30 to 100 individual flat frames. The flat frames were taken with a 30 second exposure at either $-20°$C or $-25°$C. The master flat frames are comprised of a stack of individual flat frames where each flat frame has been bias and dark subtracted. The bias and dark frames used for subtraction on the flat frames were chosen as the bias and dark frames taken closest in time to the flat frame.

2.2.3 Image reduction

The science frames were bias and dark subtracted as well as flat-field corrected. The `Source EXtractor` program (Bertin & Arnouts, 1996) is given the approximate coordinates of a single pixel as well as the pixel size in degrees. The `Source EXtractor` program performs aperture photometry on each image to build the catalogue of measured instrumental magnitudes. The `Source EXtractor` program uses this information to assign an approximate coordinate for each star it finds. A star is measured if it covers four or more pixels on the CCD and its magnitude is two sigma above that of the background noise. Further, the `Source EXtractor` program will signal the quality of a data-point using 'Extraction Flags'. Data-points flagged as poor quality are likely caused by a star merging (or close to merging) with another star, a star with at least one pixel saturated, or a star close to the edge of the CCD. Any images with less than 50 stars found by the `Source EXtractor` were not calibrated and removed from the data-set.

`SCAMP` software (Bertin, 2006) is used to accurately determine the coordinate system used in each image with information obtained from the FITS header. The `SCAMP` software compares the coordinates assigned to each star by the `Source EXtractor` program to coordinates from a known catalogue (in this case the USNO catalogue was used[6]). The `SCAMP` software uses χ^2 optimisation with a 3^{rd} order polynomial to accurately fit a coordinate frame with each image.

2.2.4 Initial Calibration

A photometric calibration program was run on each image. The photometric calibration consists of comparing the photometry of each image to a reference image. The reference image is a stacked, deep image consisting of images taken under photometric conditions (no visible clouds). The program calculates the offset each measurement of a star's magnitude (m^i) has against the star's magnitude from the reference file (m_r). Then the least-squares regression from the `SciPy Curve Fit`[7] python package is used to minimise $|m_r - f(m_i)|$ where $f(m_i)$ is the calibrated instrumental magnitude Eq. 3 (Froebrich et al., 2018b) shows how the calibrated instrumental magnitude is formed.

$$f(m_i) = A \cdot log(10^{B\cdot(m_i - C)} + 1) + \mathcal{P}_4(m_i) \tag{2}$$

[6] www.usno.navy.mil/USNO

[7] docs.scipy.org

Where A,B and C are free parameters for the least-squares and $\mathcal{P}_4(m^i)$ is a fourth-order polynomial whose coefficients are also free parameters. The equation is weighted such that the brighter stars hold a higher weight for the correction. This is done to ensure that data-points with a higher signal-to-noise affect the free parameters more.

$$\sigma_i = 1 + m_i - min(m_i) \tag{3}$$

Here $min(m_i)$ is the brightest star included in the image being corrected. σ_i is the weight factor given to the `SciPy Curve Fit` package. The package applies a weight to each star i, that is inversely proportional to σ_i. Once the stars have been calibrated against the reference file giving us $f(m_i)$ the errors associated with $f(m_i)$ are calculated. The associated error for $f(m_i)$, '$E(m_i)$' was calculated by taking each star within range of ± 0.1 mag of m_i and calculating the difference between each star and their calibrated magnitudes '$m_i - f(m_i)$'. Then taking the RMS of the scatter of the differences as the associated error, $E(m_i)$.

The calibrated data-set is stored as a series of text files. Each text file held all the data extracted from its corresponding FITS files. The calibrated data-set totalled to 63,930 files with a size of 283 GB.

2.2.5 Flat Frame Issues

The flat frames used to calibrate this data-set were mostly taken as sky flats with sidereal tracking. Mr Waterman explained two conditions under which flat fields and science images were taken that were cause for issue. Firstly, observations where water was present in the optics of the telescope. Secondly, flat fields where Mr Waterman used the perspex as a 'flat field generator'. In neither case, a log was made and so we do not know which science and flat frames are affected.

Typically sky flat frames are not taken with sidereal tracking, as this prevents the flat-field from being consistently exposed to the same point in the sky. If sidereal tracking is used when exposing for flat frames, the same part in the sky is exposed for the whole exposure. This can lead to brighter stars to become apparent on the frame. Regions on the flat-field with affected stars would render the same region on a science frame using that flat-field inaccurate for any photometric measurements. Figure 5 shows an example of the stars visible due sidereal to tracking while taking flat fields. The brightest star present holds a pixel value of ≈ 1.92 with the surrounding pixels holding a value of ≈ 1.04. Figure 5 shows five stars, as it is a 900×750 pixel segment of a 4090×4090 pixel image we can extrapolate to expect ≈ 124 stars for this image. We know from Sect. 2.2.1 that each star covers

Figure 5: Showing a 900×750 pixel segment from a master sky flat taken on 2004-09-01. This image highlights the issue caused by sidereal tracking while taking flat fields. The pixel value surrounding the brightest star featured is ≈ 1.04 whereas the central pixel value of the star is ≈ 1.92 .

on average ≈ 6.16 pixels, thus each flat taken with sidereal tracking will have ≈ 763 pixels affected by stars present on the flat frame.

Given the low number of flats used, each master flat frame was manually inspected and any found to have a large amount of structure to them were recorded. Figure 6 shows on the left the two of the flat with the most apparent structure that were used in this data-set and on the right a science frame after being flat fielded. These two flat fields were used to reduce 2.4% of the data in this data-set.

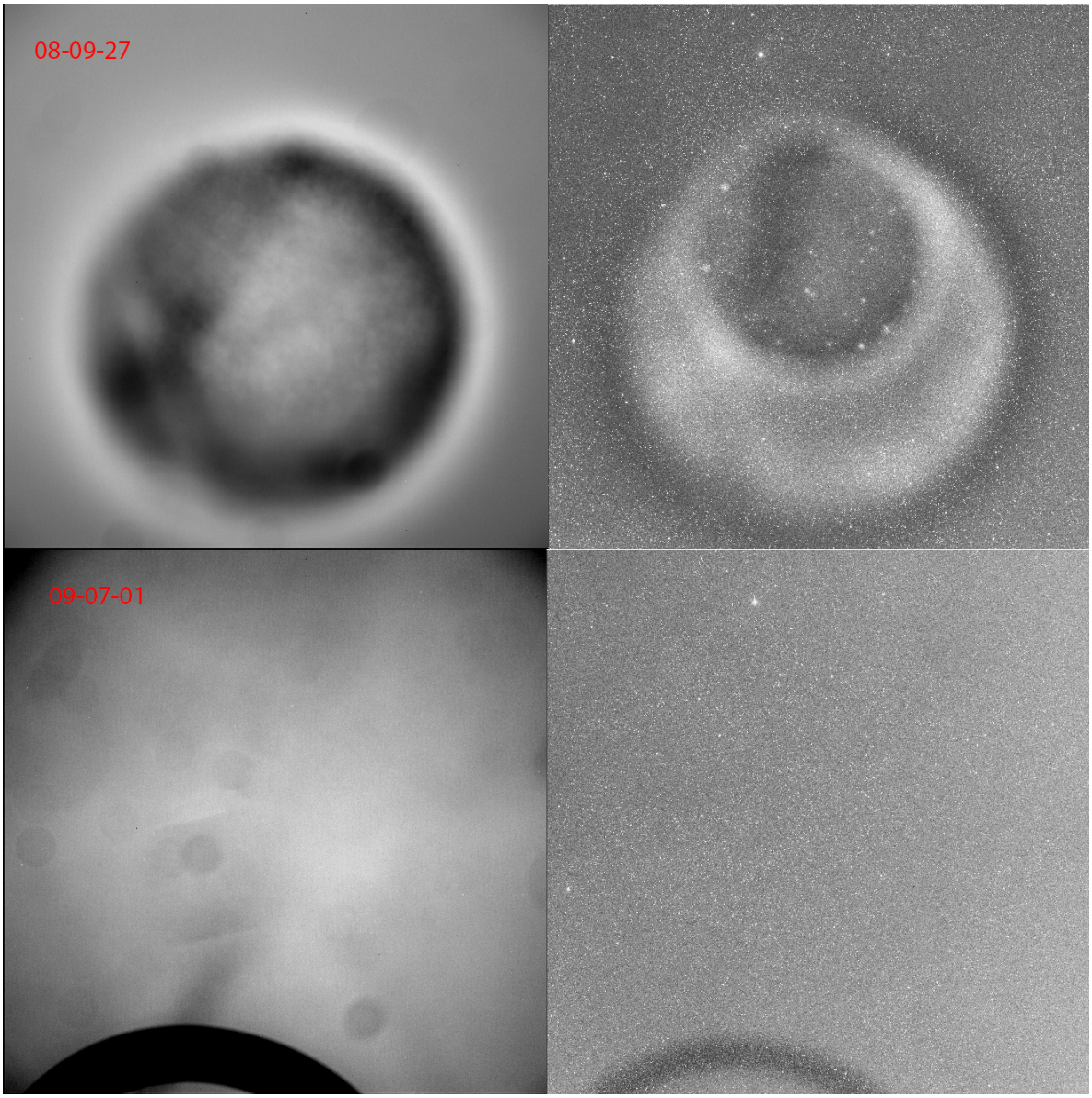

Figure 6: Flat `08-09-27`: The source of inhomogeneity in this flat is unknown, it is comprised of 50 flat frames taken at a 30 second exposure. This flat is used on 1.9% of the data. Flat `09-07-01`: The structure at the bottom of the flat causes a significant difference in the photometry of the calibrated science frame. This master flat is comprised of 50 flat frames taken at a 1 second exposure. This flat is used on 0.47% of the data. The images on the right show a science frame that has been calibrated using the corresponding flat frame on the left.

Due to vignetting and the poor quality flat fielding shown in Figs 5 & 6 it is relatively common for measured magnitude to vary as a function of the position on the CCD. The astronomer who provided this data has reported intermittently and briefly changing telescopes. While the different telescopes are reported to be of a similar 750 mm focal length and aperture size it is possible flat-fields taken on one telescope will be used for science frames taken on a separate telescope. There is no way of quantifying which flat fields are of poor quality and/or belonging to a different telescope until a photometric catalogue of each star has been made. While each flat frame can be manually investigated, without also manually investigating each science image it is unknown what structures are unique to just the flat frame and not also present on the science frames. For instance, a flat frame with a large amount of dust might appear to be of poor quality but it could also be true that the corresponding science frames also have these features and thus the apparent poor quality flat frames would be appropriate.

In Sect. 5 we discuss a method of correcting for the photometric offset caused by the inhomogeneous properties of the master flat frames. This correction is performed after all of the data has been converted into a database of indexed stars. This is to allow for each measurement of a star's magnitude to be compared with its average magnitude. This method also benefits from only calibrating against non-variable stars, which can only be identified after the data-points have been grouped into a database.

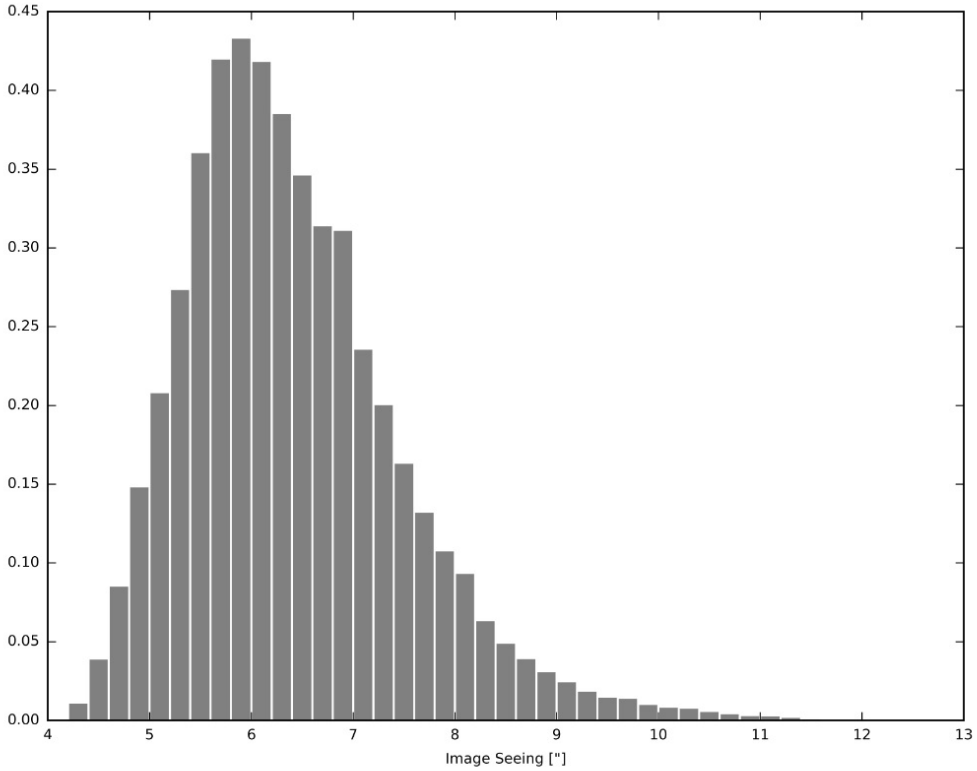

Figure 7: A histogram showing the distribution of seeing over every image in the data-set.

2.2.6 Image Seeing

The seeing for each image is calculated by finding the point spread function (PSF) for each star found within an image by the `Source EXtractor`. The FWHM of the PSF is calculated for every detected object in an image and the median of all of those is taken as the image seeing. The median seeing for all images is 6.25 arcseconds. The best image in the data-set has a seeing of 4.09 arcseconds. Figure 7 shows the distribution of seeing for all of the images. It can be seen in Fig. 7 that most images have a seeing within a distribution between 4.5 and 9 arcseconds. Given a resolution of 2.48 arcseconds per pixel we can say each star has a PSF that covers on average 6.16 pixels.

Table 2: Shows the header for each text file generated after initial image reduction and calibration

NUMBER	Running object number
MAG_AUTO	Kron-like elliptical aperature magnitude
MAGERR_AUTO	RMS error for MAG_AUTO
BACKGROUND	Background count at the centroid position
X_IMAGE	Object position on the CCD along the X position
Y_IMAGE	Object position on the CCD along the Y position
ALPHA_J2000	Right ascention of barycenter (J2000)
DELTA_J2000	Declination of barycenter (J2000)
FWHM_WORLD	FWHM of PSF assuming a gaussian core
FLAGS	Quality flag given by Source EXtractor
MAG_CALI	Magnitude after initial calibration
MAGERR_CALI	RMS error for MAG_CALI
FLAGS_CALI	Flags given from calibration process

3 Database Set Up and Population

Given the size of this data-set, care needed to be taken when planning how all of the data provided would be stored and accessed. Thus a relational database was required to allow for scientific analysis of the data. After the image reduction and initial calibration was performed, information extracted was stored in the form of one text file per image. The text files hold information extracted from each star in a given image. Table 2 shows the information held in each text file. Any other information regarding the conditions during observation were left as the meta data in the FITS header.

3.1 Data Structure Formatting

A problem with storing information in the form of individual text files is that none of the data-points are linked. While a single measurement of a star in any given image is accessible, all measurements of any given star are not linked together. Thus rendering any investigation into time-based variations very slow and tedious. A linked database must be constructed allowing for each star to be identified and each data-point corresponding to each star to be grouped. Rather than one large table, the linked database will contain different linked tables. This will be achieved with the use of comma-separated variable (CSV) files. Each text file will be collated into a data structure of CSV files, where each CSV file represents a table in a linked data structure. This collection of CSV files forms a '`NoSQL`' database. A `NoSQL` database is a database that provides a mechanism for the storage and retrieval of data without the use of Structured Query Language '`SQL`'. This is achieved by treating each CSV file as a table within our `NoSQL` database and using various python packages such as `Pandas`[8] (Wes McKinney, 2010) and `NumPy`[9] (Oliphant, 06) to perform any tasks on our database. For this report, the collection of these six tables which are used to organise and index the data-set provided to use will be referred to as the 'database'.

A total of six tables were decided to be appropriate for this database. An example of each table is found within the appendix (Sect. 7). Figure 8 shows the structure of each table within the database. It can be seen that each table has at least one identification number. Each identification number is shared with one other table, with the exception of `Data ID` as it is only appropriate for that to be in `STAR_DB`. The arrows, along with the 'ID' numbers they point to, represent how the '`NoSQL`' database is linked. A search for an object using its coordinates within the `ID_DB` could be performed. After that the `Object_ID` would be known so one could query `STAR_DB` with `Object_ID`. All possible information stored in this database regarding that object would now be available as `STAR_DB` links to each table.

For each column of data inside a data table, the smallest appropriate data type was chosen. All of the identification numbers and flags were stored as integers as per the 'int32' data type. Some 'int32' object could be stored as 'int16' or 'int8' as this would only provide a small storage decrease at the cost of reducing potential future additions to the database it was decided to forgo using integers less than 32-bit lengths. It is possible the other regions shown in Fig. 1 or any stars removed due to a completeness cut may be re-added later.

[8] `pandas.pydata.org`
[9] `numpy.org`

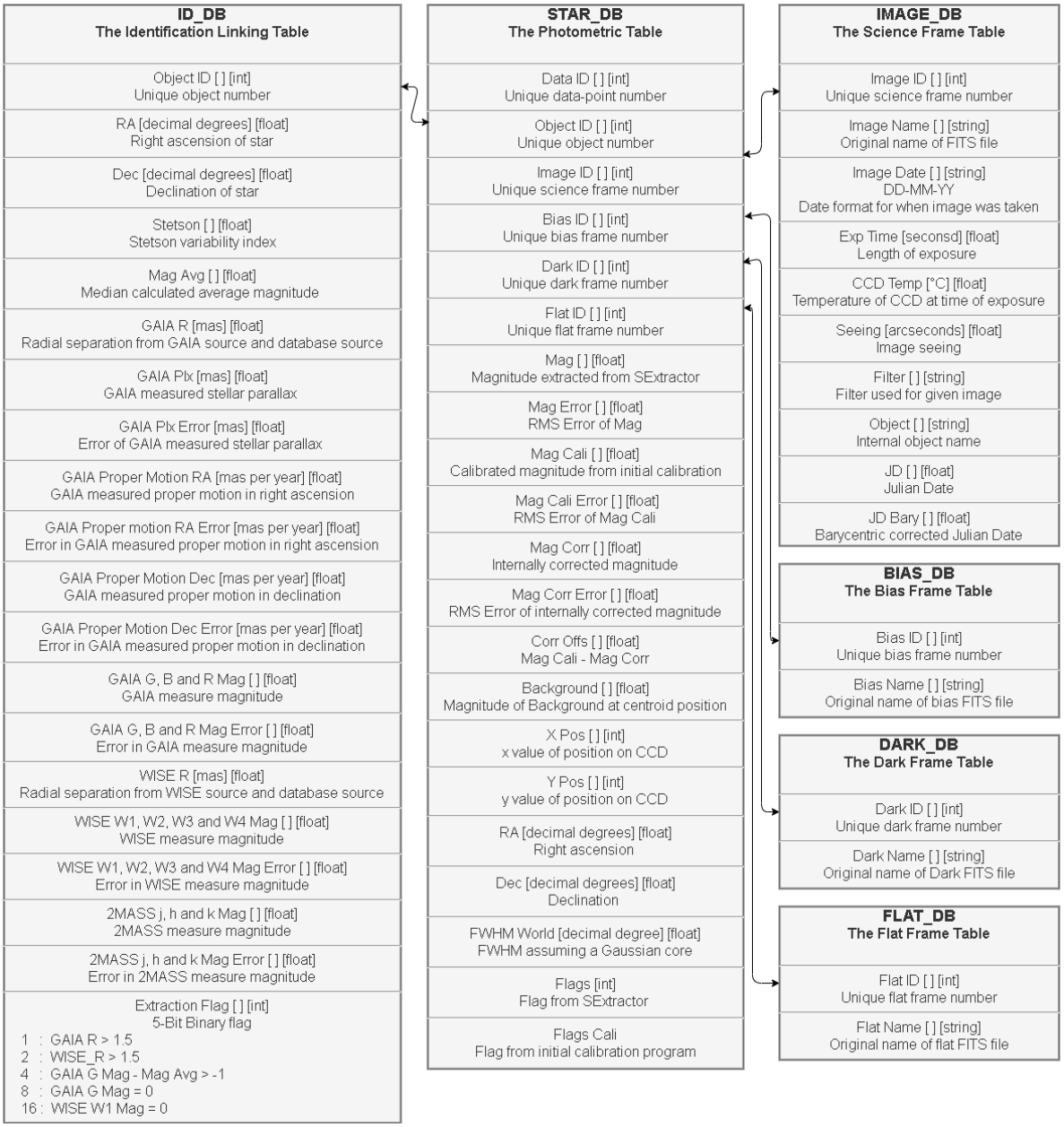

Figure 8: Showing the layout of each table within the database and how they are linked together. The arrows indicate how each table is linked together by is corresponding 'ID'.

3.1.1 Image Table

The image table '`IMAGE_DB`' holds all the relevant information regarding the conditions that an image was taken under as well as the time it was taken. This includes exposure time '`Exp_Time`', CCD temperature at the time of exposure '`CCD_Temp`', time in Julian date '`JD`' and the images identifier '`Image_ID`'.

This table will also hold the barycentric corrected Julian date '`JD_Bary`'. The barycentric corrected Julian date is the time of exposure adjusted to a single constant location. As the earth orbits the sun the light travel distance from an object to the observer is not constant and can be up to 2 AU different dependent of time of year. This gives a difference in light travel time of up to 998 seconds. To correct for this we can calculate the light travel time between the earth and the barycenter of the solar system at the time of measurement. We can then add the time to the recorded time of measurement if the earth is closer to the object then the sun or subtract the time if the earth is further away from the object than the sun. As the temporal resolution of our data-set is ≈ 120 seconds, this is a necessary correction. Three calibration tables are used for the master bias, dark and flat frames. Each table holds the names and identifiers of each of the master calibration frames. The rightmost set of tables in Fig. 8 all represent the image tables.

3.1.2 Photometric Table

The photometric table '`STAR_DB`' holds any astrometric and photometric information obtainable for each individual data-point. This table will hold all information that varies per image and is not shared with every star in a given image. The photometric table also holds every data-point's object identifier '`Object_ID`' (that being the object/star this data-point belongs to) and image identifier (that being the image the data-point is from) allowing for all data-points belonging to the same object or image to be grouped.

The middle table in Fig. 8 represents `STAR_DB`, the photometric table. It can be seen that `STAR_DB` holds three separate magnitude measurements, '`Mag`', '`Mag_Cali`' and '`Mag_Corr`'. `Mag` is the raw magnitude extracted by the `Source EXtractor` program, it also has the associated errors '`Mag_Error`' and extraction flags '`Flags`'[10]. `Mag_Cali` is the calibrated magnitude described in Sect. 2.2.4 it, like `Mag`, also has its associated errors and flags. `Mag_Corr` is the corrected magnitude described in Sect. 5

[10] sextractor.readthedocs.io/en/latest/Flagging.html

this corrected magnitude has an associated error 'Mag_Corr_Error' and a measurement of Mag_Cali − Mag_Corr as 'Corr_Offs'. Corr_Offs is used as an indicator for how much of a correction was necessary during the photometric correction (see Sect. 5).

3.1.3 Identification Linking Table - The Catalogue

A main linking table 'ID_DB' is used for querying individual stars. This table holds the coordinates and unique object id for every star. This table forms what will be referred to as the catalogue. This table will also hold any information about each star that does not change from image to image such as average magnitude and colour. Section 4 describes how our database was cross-matched with other catalogues in order to obtain information such as colour. All of the catalogues cross-matched in Sect. 4 are single measurements per star, so to avoid repetition within the photometric table, all data from external catalogues was added to the identification linking table.

The leftmost table in Fig. 8 shows the identification linking table. It can be seen that ID_DB holds information from GAIA, 2MASS and WISE, Sect. 4 describes the process of matching our data-set with other catalogues. ID_DB also holds the radial separation each match our data-set has with the matched catalogues, 'GAIA_R' and 'WISE_R'.

3.1.4 Small Database

The un-indexed STAR_DB (i.e not featuring a populated Object_ID) CSV is 72.4 GB. In order to be able to construct and calibrate the database at a reasonable rate, it was decided to create a second data-set with a reduced temporal resolution. This was achieved by only using one image from each night, a 'best file'. The best file was determined by the number of stars in each text file, not within 10 pixels of the edge of the CCD. The second, smaller database will be used to prototype and then construct a data reduction and calibration pipeline. The complete data reduction pipeline will then be applied to the full database. The CSV's created from the 'best file' data-set are analogues to the CSV's that will be made from the full data-set. After the 'best file' from each night was found, it was copied, along with its corresponding FITS file, to a separate location. The data-set holds images from 213 nights therefore the small database holds 213 images. This shrinks the un-indexed STAR_DB to 979 MB. For the purpose of this report, unless stated otherwise, all processes described are performed on the small database. The small database was entirely due to time constraints, all of the processes described are being performed on the large database a the time of writing. The catalogue ID_DB that

was generated with the small database is also applicable for the large database, hence the iterative matching method described in Sect. 3.2.1 was not repeated for the large database. Instead, the large database was matched to the preexisting catalogue.

3.2 Star Indexing

In order to investigate magnitude as a function of time for each individual star, the same stars need to be identified across all of the images. Without this, each measurement of a star in an image can only be treated as a single isolated data-point. Given that each star in this database will be given a celestial coordinate by the `Source EXtractor` and `SCAMP` software we may use the coordinates of each data-point as a method of grouping data-points belonging to the same star.

Each image has an uncertainty in the measured coordinates. The random, statistical parts of uncertainty in a coordinate are caused by fitting a function to an intensity distribution. The spread, and thus uncertainty, of the function, is caused by astronomical seeing and the size of the Airy disk from the instruments diffraction limit. We know that on average `Seeing` ≈ 6.95 arcseconds and the Diffraction limit$= 1.26$ arcseconds. From this, we can see that the dominant source of random statistical noise is the atmospheric seeing. We can approximate the seeing and diffraction limit as normally distributed.

There is also some systematic uncertainty associated with the astrometry. The systematic uncertainty arises partly from the pixel size of 2.48 arcseconds. There is also a systematic uncertainty arising from the projection of a coordinate frame onto the images. Due to the optics of the instrument, there is some warping of the coordinate system towards the edges of the images.

As explained in Sect. 2.2.3 the `SCAMP` software uses a third-order polynomial with χ^2 optimisation to fit the coordinates of pre-catalogued stars to an image. The fit of the astrometric solution may not have enough terms to properly account for any warping, this effect can skew the distribution of data-points surrounding a coordinate to not be centred around a point and instead be elongated. The amount of elongation would be a function of the radial distance from the centre of the CCD. It is also likely to only make a noticeable effect on the extreme edges of the CCD.

The compound effect of all of the uncertainties causes, in general, all the coordinates of the data-points associated with a star to form a distribution around the true coordinates of that star. The distribution will be dominated by the image seeing, thus a good approximation would be a Gaussian distribution. To identify a unique star across all of the images we must create a search parameter. We will use this search parameter to find all data-points associated with the same star. Equation 4

compares one set of coordinates to a second set of coordinates and calculates the angular separation between the two coordinates.

$$AngularSeparation = \arccos(\sin(\delta_1)\sin(\delta_2) + \cos(\delta_1)\cos(\delta_2)\cos(|\alpha_1 - \alpha_2|)) \quad (4)$$

Where δ_1 and α_1 is the declination and right ascension of the first set of coordinates and δ_2 and α_2 is the declination and right ascension of the second set of coordinates.

We can compare each data-point within `STAR_DB` to every other data-point within `STAR_DB`. For each comparison, we can obtain a list of angular separations between a single data-point and every other data-point in the photometric table. Any data-point found to have an angular separation within a set search area (determined in Sect. 3.2.2) is considered to be a data-point obtained from the same star. Every matching data-point is assigned the same object identification number `Object_ID` and subsequently removed from any future comparisons.

If no matching data-points are found for a data-point within its search area, that data-point is removed from the photometric table. Figure 9 shows the schematic representation of how a data-point would be matched to a star. For this report, this process will be referred to as the 'star matching' process.

After each star had been matched and assigned an identifier the identification linking table was created. The identification linking table known as `ID_DB` holds each star's `Object_ID`, as well as that star's coordinates in right ascension and declination. The coordinates for each star in `ID_DB` are calculated by taking the median of all of the coordinates with the same `Object_ID` within `STAR_DB`. For example, a star with five associated data-points might have the coordinates:

- 317.456 69° +45.642 79°
- 317.456 78° +45.642 85°
- 317.456 79° +45.642 94°
- 317.456 81° +45.642 96°
- 317.456 85° +45.643 04°

Thus its coordinates listed in `ID_DB` would be 317.456 79° +45.642 94° as the median of those coordinates.

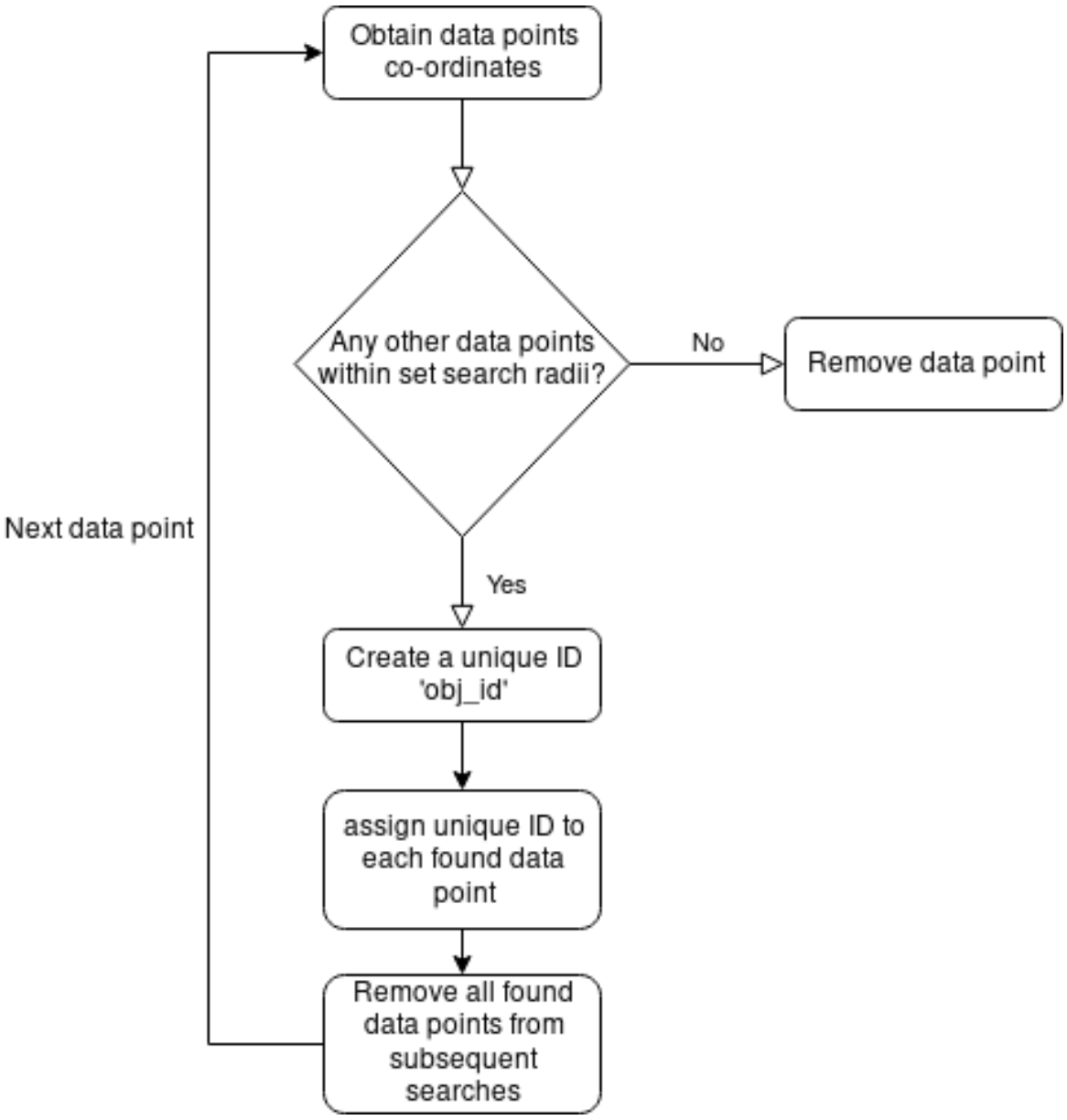

Figure 9: Flow chart outlining the process of 'star matching'. The above flow chart shows the process of finding each star's associated data-points. This is done by taking a data-point at random from the set of data-points found in `STAR_DB` and applying Eq. 4 to their associated coordinates in order to find any data-point whose angular separation falls within a set search area

3.2.1 Iterative Matching

At the start of Sect. 3.2 the uncertainty in the measurements of each data-points coordinates is discussed. In Sect. 3.2, we also discuss how a random data-point's coordinates are used as the basis for the star matching search.

A caveat of picking a random coordinate for the process of 'star matching' is that it assumes the coordinates of the distribution of corresponding data-points are centred around the randomly selected data-point. This is not always going to be the case. It is more likely that the coordinates of any randomly selected data-point are not at the centre of its associated distribution. Without ensuring that each search starts in the middle of a distribution of data-points substantial problems regarding the matching process may occur.

A single star with a distribution of data-points may be falsely detected as two stars. This might occur if the first detection fails to detect a sufficient amount of data-points due to being close to the edge of the distribution. Then, when all of the matched data-points have been removed, some data-points will remain, this allows the program to erroneously recognise the remaining data-points as a second star. Figure 10 shows a diagrammatic representation of how this erroneous second match may occur.

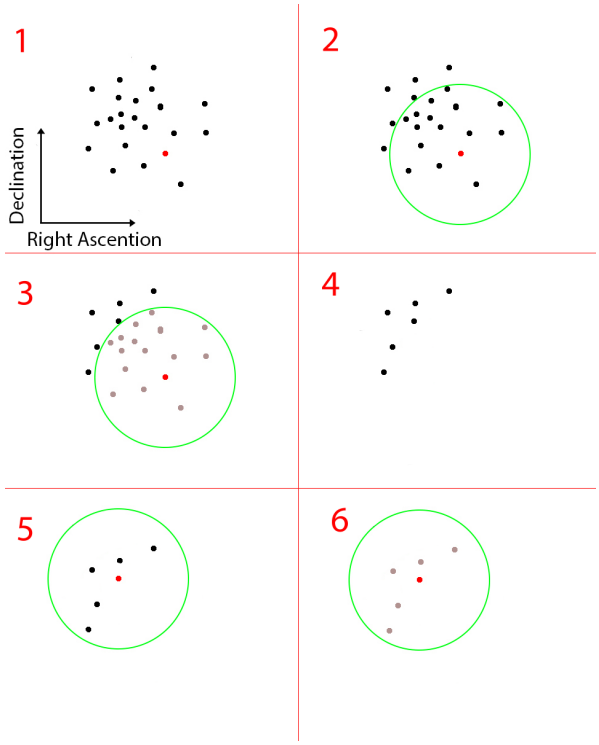

Figure 10: If we consider each square to be a plot of a distribution of data-points in Ra and Dec. Panel 1: Shows a distribution of data-points where the point marked red will be used as the randomly selected point for the 'star matching' process. Panel 2: Shows the search area used, any data-points within the green circle are considered to belong to the same star. Panel 3: Shows the data-points found to be associated with the initial data-point. Panel 4: Now the matched data-points have been removed from any subsequent searches, however, some data-points associated with the same star remain. Panel 5: Now a second random data-point has been selected and has found matches within its search area. Panel 6: An erroneous match has been found from the data-points not removed from the original search.

Another issue may occur if two stars have associated data-points within the search radii of each other. This could cause two stars to be falsely classified as one star if the original data-point used for searching is close to both stars and not in the centre of one of them.

We can ensure a search is at the centre of a distribution of data-points by iteratively running the 'star matching' program where each iteration after the first uses the coordinates found within `ID_DB`, rather than using a random coordinate found in `STAR_DB`. Providing the instrumental uncertainty

does not dominate the distribution of each data-points coordinates, the distribution of data-points associated with a star can be approximated by Gaussian centred on the star's true coordinates. It is possible to accurately calculate the seeing of each image by median averaging the FWHM of each star's PSF. This process relies on the predominant source of the PSF to be atmospheric seeing. As the distribution of data-points belonging to a star can be approximated as Gaussian, each calculation of the median matched data-points should be closer to the true median of all data-points associated with a star. A diagrammatic representation of this process can be found in Fig. 11. Figure 11 shows how the coordinates used for 'star matching' converges onto the centre of distribution using this method.

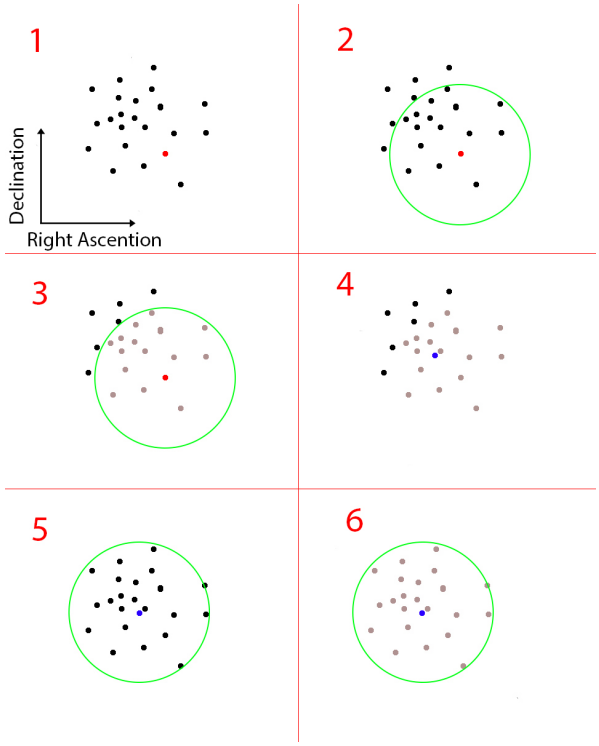

Figure 11: If we consider each square to be a plot of a distribution of data-points in Ra and Dec. Panel 1: Initially we select a random point in this distribution (coloured red). Panel 2: Now we search for any data-points within a set radius (shown by the green circle). Panel 3: We now assign each data-point within this radius the same identifier and consider it to belong to the same star. Panel 4: We take the average of each of the data-points (coloured blue) and consider that as the new coordinate to search with. Panel 5: We now repeat the second step, searching for each star within a given area. Panel 6: We now assign each data-point within the search radii to a single identifier.

With this, we can conclude the coordinates of a star found within ID_DB will converge on to the average of all data-points for that star. Figure 12 shows a schematic representation, similar to Fig. 9, of how the iterative process may be applied to our database.

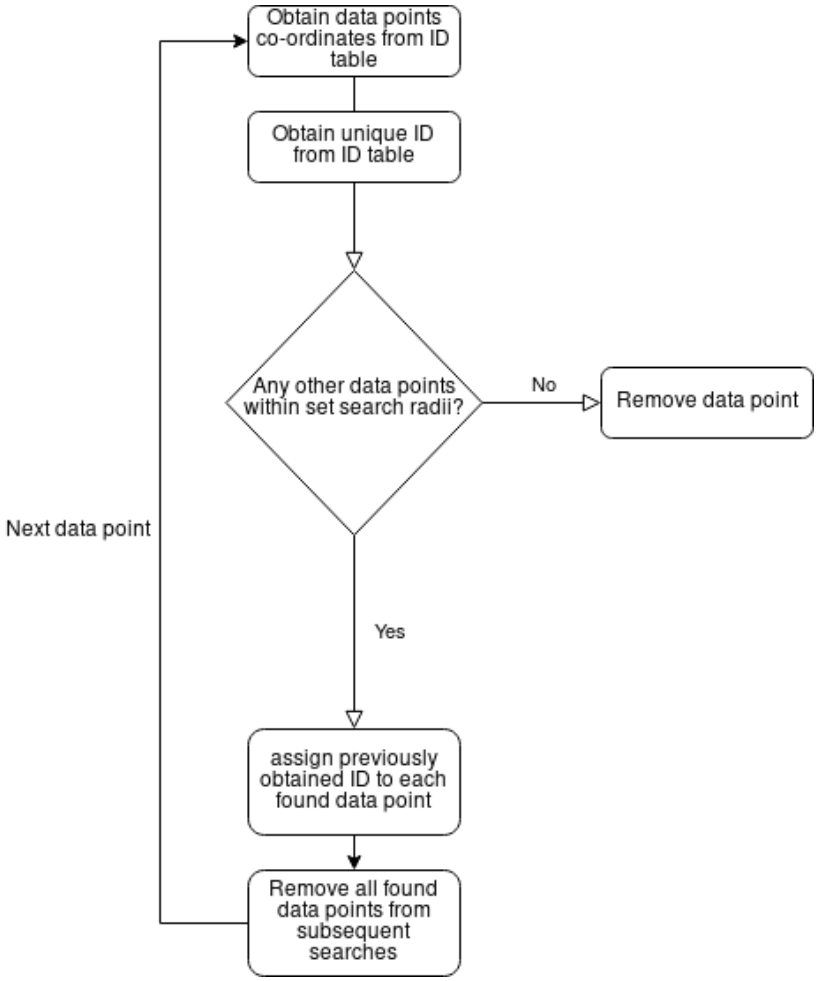

Figure 12: Flow chart outlining the process of subsequent iterations 'star matching'. This shows that the 'star matching' process following the initial matching is started by first obtaining a coordinate from ID_DB and then applying Eq. 4 to find any data-points in STAR_DB that fall within a set radii of angular separation

3.2.2 Search Radii Justification

The basis for the matching process is to define an area around a searched coordinate, all data-points with coordinates within that area are considered to be associated with the original search coordinates and all data-points outside of that area are not considered to be associated. The size of the area with which we search for matching data-points is critical for finding a balance between not missing any potential matches but also not matching any false positives. A larger search area may reduce the amount of missed data-points but it also increases the potential for erroneously matching data-points not associated with the star in question. Any data-points erroneously matched with a star will contaminate any photometric investigations of that star. To ensure that no false matches are made and no potential matches are missed, the astrometric uncertainty must be considered. As each coordinate given to a data-point by the `SCAMP` software has an error in associated measurement we must consider the range of coordinates that a data-point can have. From Sect. 3.2 we know that the dominant source of astrometric uncertainty is the random component of the uncertainty. Thus we can use the image seeing as the basis for our justification of the search area. If we compare the average separation each data-point has from its parent star with the distribution of image seeing we can find a search radius that should increase the chance of capturing all data-points without risking false matches.

To investigate the average angular separation each data-point has with its parent star we must first perform the matching procedure using a less precise method of determining the search area. As this method will be used to gauge the range of angular separation between data-points and their host star it is not critical to use a properly defined search area. It was suspected that a search radius of 3.5 arcseconds would be a good heuristic choice given an average seeing of 6.25 arcseconds.

The 'star matching' process was iterated five times as per the iteration process discussed in Sect. 3.2.1. After each iteration `ID_DB` was inspected to check for changes. It was determined that three iterations were sufficient as no changes were made to `ID_DB` after the third iteration.

Figure 13 was made comparing the seeing of each image to the angular separations each data-point has to its parent star (i.e the angular separation each coordinate found in `STAR_DB` has from its parent coordinate in `ID_DB`). The distribution in black on the left represents the distribution of angular separations each of the data-points have with their corresponding parent star. We can use the `SciPy Curve Fit` python package to fit a Gaussian to the distribution.

From the fitted Gaussian distribution we can obtain the standard deviation of the fit, 'σ_f'. We can compare the standard deviation of the fitted Gaussian (seen as the red vertical line) to the standard

error of the distribution, 'σ_d' (seen as the blue vertical line). If the data has been successfully fitted with a Gaussian, the difference between the standard error of the distribution and the standard deviation is small. From this, we can verify that the uncertainty of the coordinates of the data-points are dominated by random uncertainty. This further confirms that it is appropriate to base the area with which we search for star matches on our image seeing.

We can compare the standard deviation of the distribution of separations to the distribution of seeing. From this, we can see to what extent we can increase our search radii without it being larger than the seeing. Figure 13 shows the comparison between each data-points angular separation and image seeing. It can be seen from Fig. 13 that the distribution of data-points is unanimously smaller than seeing of the images. The rightmost blue line represents $5 \times \sigma_d$. The grey line represents the lowest image seeing in this data-set. The green line represents the average image seeing in this data-set.

Figure 13 shows that $5\sigma_d$ is less than the minimum seeing. It was decided that 5σ was an appropriate search size as should account for $< 99\%$ of the data-points. It was found in reality that $5\sigma_d$ accounts for 98.8%. This is because the errors in our coordinates are not purely random and feature some systemics, hence our distribution is not a true Gaussian.

While up to $10\sigma_d$ is less than the minimum seeing a smaller search area still decreases the probability of erroneously matching to the wrong star. If the accuracy of the astrometry is sufficiently poor it could cause some data-points to be incorrectly matched with a different star, a larger search area could increase the probability of incorrectly matching a data-point. It was decided that sacrificing some potentially correct data-points in order to decrease the possibility of a mismatched data-point was appropriate for a data-set this large.

The matching radii used to construct the linked catalogue is $5\sigma_d = 5 \times 0.437$ arcseconds=2.185 arcseconds. The new matching radii was then used on the data-set over three iterations. Figure 14 shows the results of using 2.185 arcseconds as the matching radii

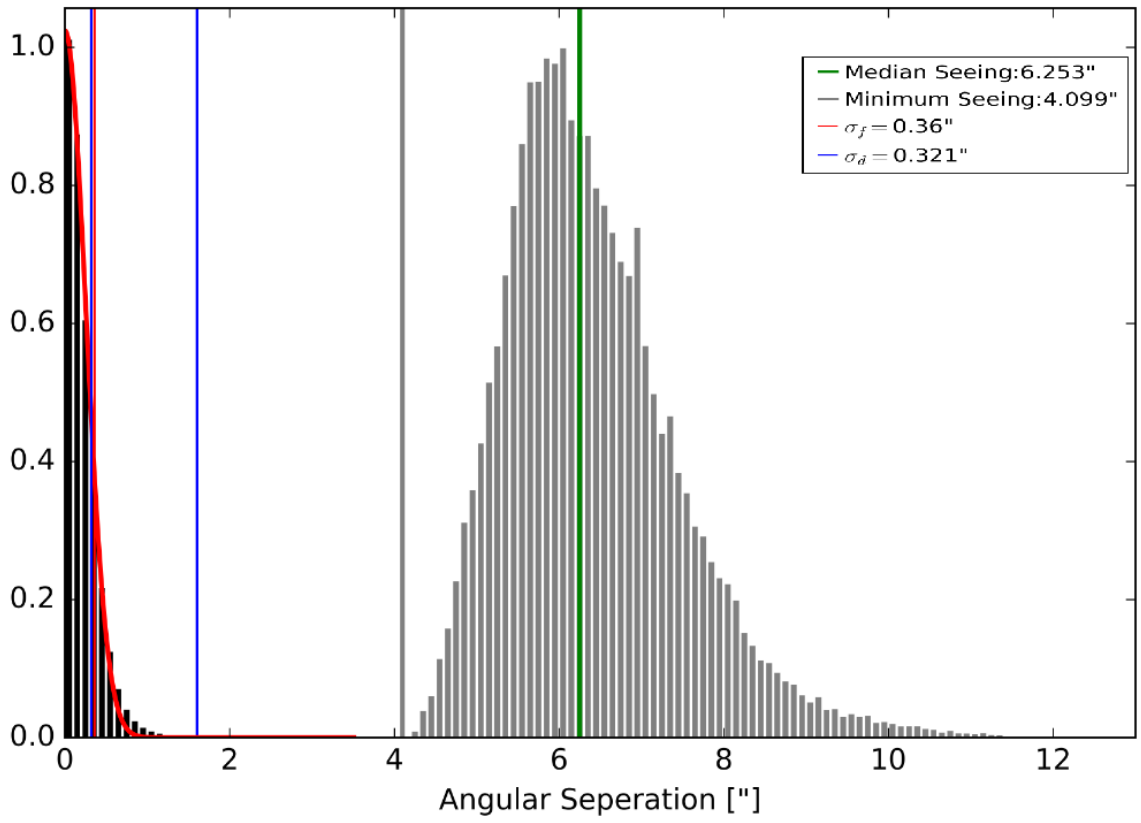

Figure 13: Showing the Seeing in arcseconds for each image in grey. The smallest image seeing can be seen as the grey line. The distribution of angular separation between each data-point and its parent can be seen as the black bars with a fitted Gaussian in red. The first blue line represents σ_d and the rightmost blue line representing $5\sigma_d$. It can be seen that the left-most blue line, σ_d, is very similar to the fitted Gaussian's standard deviation σ_f represented by the red line. The y-axis has been normalised so both distributions are between 0 and 1

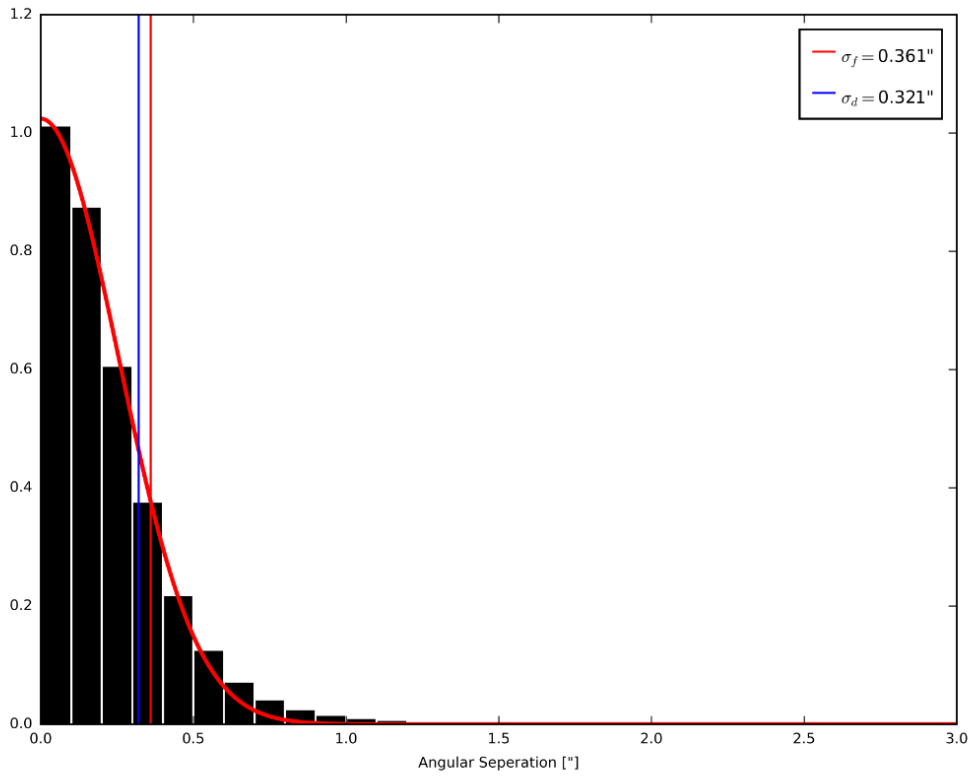

Figure 14: Histogram showing the distribution of angular serrations each data-point has with its parent coordinates. The red line is the fitted Gaussian. It can be seen that the standard deviation of the fitted Gaussian, the red line, is very similar to that of the data, the blue line.

3.2.3 Matching Verification

In order to verify the matching process a plot of the coordinates of each data-point was made. We can plot all data-points relating to the same star and over-plot the 'master' coordinates found in `ID_DB`. Figure 15 shows the data-points associated with two stars of coordinates 316.525 97° +45.311° (J2000) and 316.521 75° +45.316 16° (J2000). It can be seen that a two data-points fall outside of the search radii for star 316.525 97° +45.311°(J2000). Any data-points associated with a star on the initial iterations but not later iterations are not removed unless associated with another star.

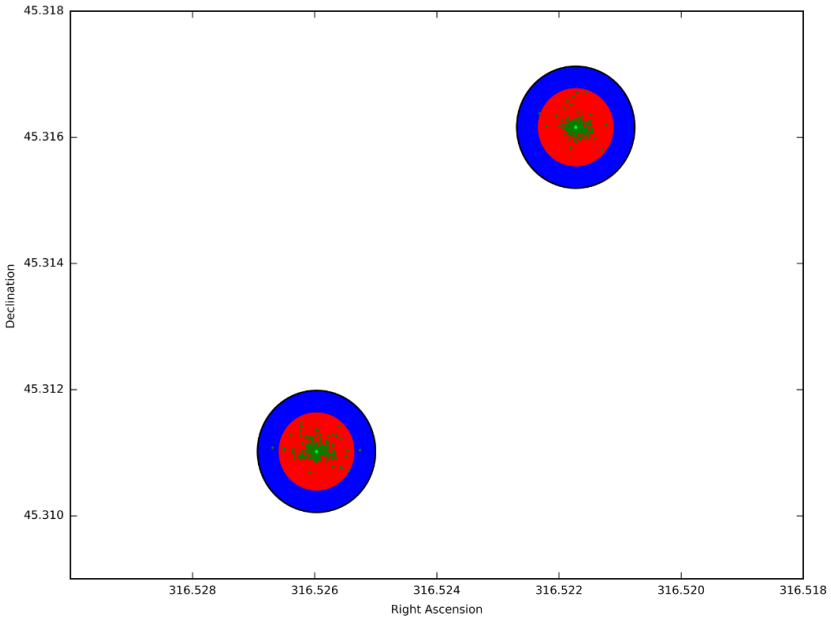

Figure 15: A Ra-Dec plot of each two stars with their corresponding data-points. The bottom left star has coordinates 315.526° +45.311° (J2000) it has 196 data-points and an average instrumental magnitude of 18.44 mag The top right star has coordinates 316.522° +45.316° (J2000) it has 209 data-points and an average instrumental magnitude of 17.56 mag Each dark green point represents an individual data-point with each light green point being the coordinate given to the star in `ID_DB`. This plot shows the distribution of data-points around a star. The blue circle represents the average seeing and has a diameter of 7 arcseconds whereas the red circle represents the search area used to identify each data-point with a diameter of 4.37 arcseconds.

As each star has now been identified and its corresponding data-points linked to it we can investigate the average magnitude of each star. From this, we can find the completeness of our data-set. Figure 16 shows the distribution of magnitudes. The magnitude values shown here are instrumental magnitudes. The magnitudes are not corrected to apparent magnitude as the main focus of this data-set is to perform relative photometry. The true apparent magnitude can be obtained in Sect. 4 whereby we will obtain the magnitude from GAIA, 2MASS and WISE. Figure 16 shows an 99% completion limit of 18.45 instrumental magnitude. The completion limit is found by using the `SciPy Curve Fit` python package to fit a straight line to the middle two quarters of the distribution. This was done to avoid fitting to the lower sample size found at either end of the distribution. The blue line represents the point at which the data-set falls below 99% of the fitted red line, the completion limit. In this distribution, the brightest star in our data-set was 8.4 instrumental magnitude with the dimmest being 20.6 instrumental magnitude. The data-set at this stage has 64,439 stars.

We can now remove any star with an average instrumental magnitude dimmer than 18.45 and any star with less than 25 measurements. As the smaller database holds 213 images, any star within the completeness limit should have ≈ 213 measurements. As such, any star within our completeness limit that is present in less than 25 images is likely to be problematic and will be removed. Although most stars that have only a magnitude within the completeness limit for less than 25 images are likely to be problematic, this cut will also remove any short term outbursting stars. Figure. 16 also allows us to investigate the linearity of the detector. This has allowed us to conclude that for the level of photometric accuracies present here, the detector is effectively linear. Performing these cuts leaves the data-set with 19,858 stars. The brightest star in the data-set is 8.51 instrumental magnitude. Some brighter stars are removed as it is likely they saturated the CCD, any measurements with any saturated pixels were removed in Sect. 2.2.3. It is likely that stars at the saturation limit will occasionally not saturate the CCD, allowing them to be measured, however, if this happened less than 25 times they are removed.

Figure 17 shows the distribution of the number of data-points associated with each star in the small database. It can be seen that the majority of stars have over 150 data-points. The small number of stars whose magnitude is greater than 12 mag are likely to be above the CCD saturation limit and hence some of its data-points are rejected by the `Source EXtractor` software.

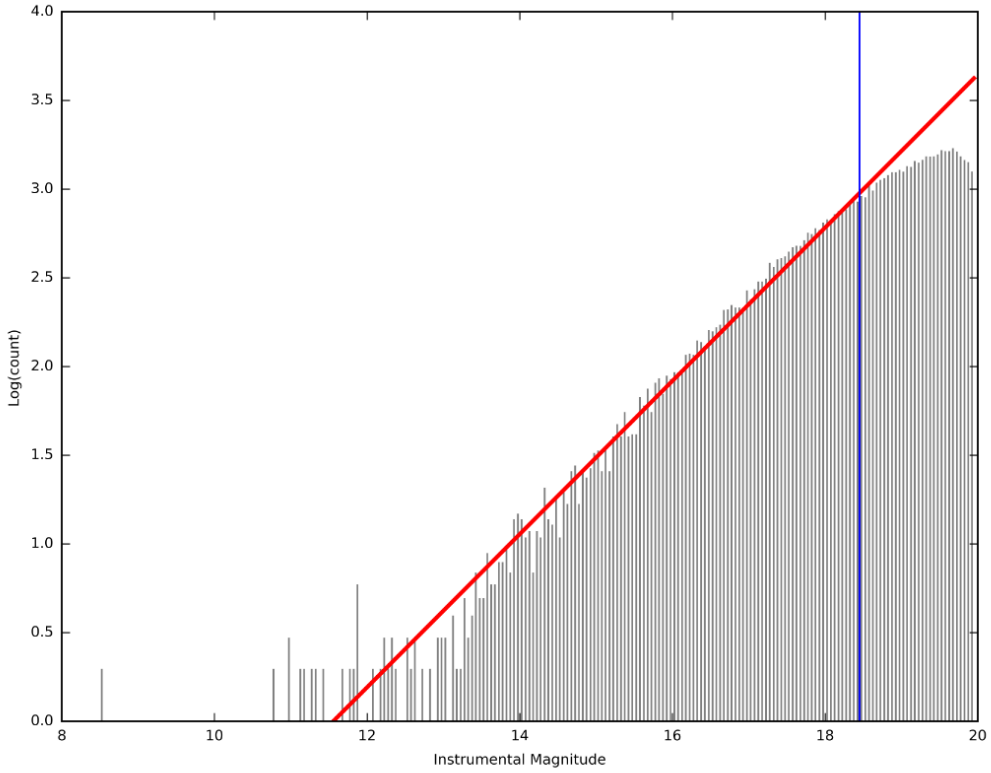

Figure 16: Histogram showing the distribution of stars using their average magnitude. A fitted line is used to show completeness limit. This data-set has a 99% completeness limit of 18.45 instrumental magnitude.

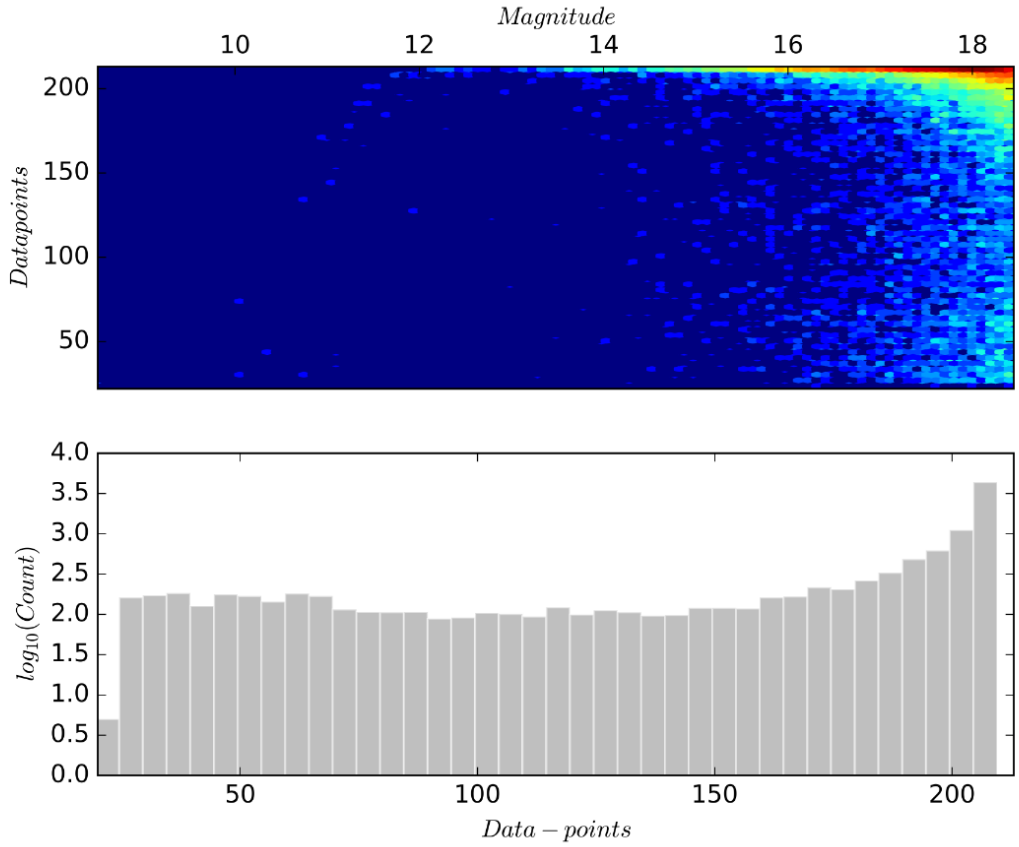

Figure 17: The top panel shows a heatmap of the magnitude of a star against the amount of data-points associated with a star. The bottom panel shows a $\log_{10}$ space histogram of the distribution of data-points associated with each each star in the database.

As each data-point has now been linked to a star and the time of observation is known for each we can construct a light curve. Figure 18 show a light curve of two stars where each star was chosen at random. The first star has an identification number of 141, it has 178 data-points and is located on 317.916° +45.1650° (J2000). Star 141 has an average instrumental magnitude of 14.471 and a GAIA B-R colour of 1.393. The second star has an identification number of 152, it has 128 data-points associated with it and it is located at 315.566° +45.125° (J2000). Star 152 has an average instrumental magnitude of 16.556 and a GAIA B-R colour of 2.155. Each light curve was made using the barycentric

corrected modified Julian date along with the calibrated magnitude calculated in Sect. 2.2.4 along with their corresponding errors. Each data-point was also plotted with a colour corresponding to the master flat frame that was used to calibrate it.

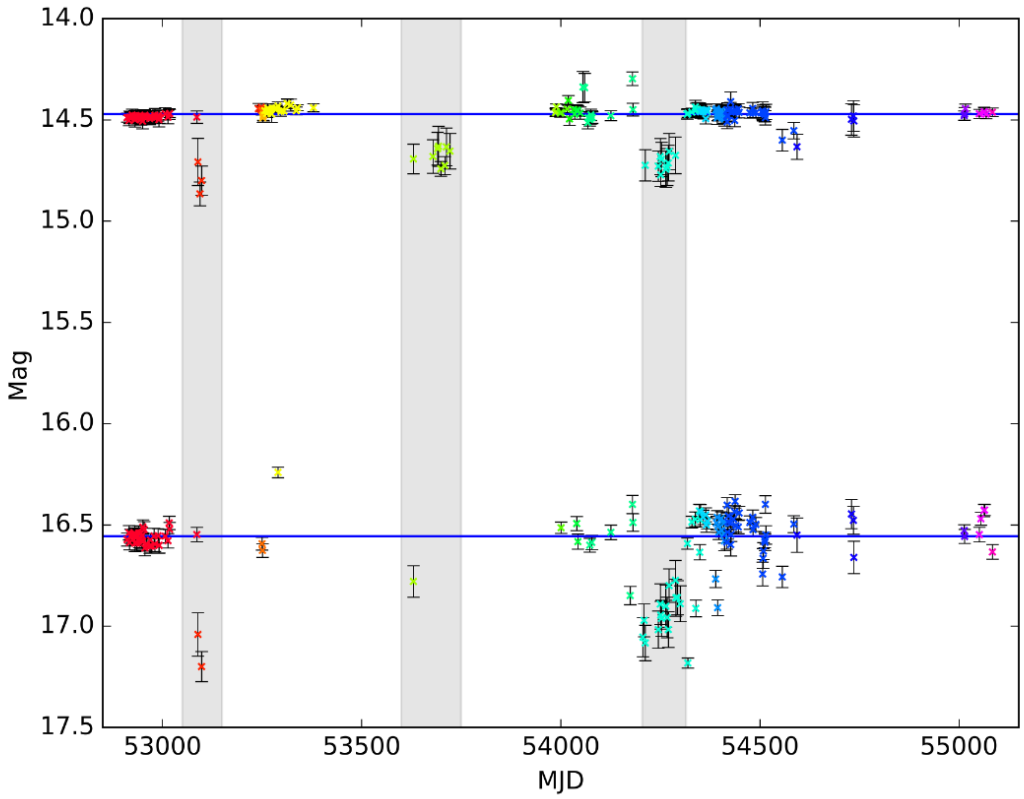

Figure 18: Showing a light curve of two stars using the calibrated magnitude along with the barycentric modified Julian date. The 14.5 instrumental magnitude star is positioned at 317.916 09° +45.165 035° (J2000) and has a GAIA B-R colour of 1.393. The 16.6 magnitude star is positioned at 315.566 26° +45.124 86° (J2000) and has a GAIA B-R colour of 2.1553. Point of the light curve that feature significant systematic magnitude shifts have been highlighted in grey.

Figure 18 show some systematic offset in both light curves. It can be seen that the offset happens at various points in curve although there are three main sections identified as being most effected this

is between MJD=53049.5 - 53149.5, 53599.5 - 53749.5 and 54204.5 - 54314.5. This offset is largely due to the flat fields used for initial reduction and calibration. Sect. 5 discusses how this systematic offset is corrected.

We now have a 99% complete catalogue of stars from our data-set, '`ID_DB`'. The star matching pipeline can now be performed on the full data-set of 64293 images. Performing this gives a `STAR_DB` containing 632123248 data-points with a total size of 72 Gigabytes.

4 Catalogue Cross-Matching

In order to obtain magnitude measurements in multiple filters as well as astrometric measurements, we can match our current data-set with that of GAIA (Gaia Collaboration et al., 2018a), 2MASS (Skrutskie et al., 2006) and WISE (Wright et al., 2010). Matching with an existing catalogue that is present on the 'Vizier' online catalogue library[11] is achievable using the `Astroquery` python package (Ginsburg et al., 2019).

We cross-match with GAIA in order to gain astrometric information about the stars in our catalogue. This will allow us to determine the stars absolute magnitude and thus allowing us to create Hertzsprung-Russel diagrams. We cross-match with 2MASS and WISE in order to obtain photometric information about the stars in our catalogue. As our data-set is single-filtered this will allow for further insight into the population of stars in our catalogue via photometric classifications.

4.1 GAIA

The GAIA space observatory was launched in 2013 and is operated by the European Space Agency. GAIA hosts two filters, GAIA's blue 'G_B' $\lambda_{mean} = 0.51\mu$m and GAIA's red 'G_R' $\lambda_{mean} = 0.8\mu$m. GAIA also takes 'filter-less' measurements referred to as green measurement. Due to the optics of the GAIA spacecraft, it is treated as a filter. GAIA's green 'G' has $\lambda_{mean} = 0.67\mu$m.

In GAIA's second data release 'DR2' can resolve stars with a separation up to ≈ 0.1 arcseconds and has an effective angular resolution of 0.4 arcseconds. The GAIA DR2 covers the whole sky and is complete between G=12 and G=17 [12]. The second data release from GAIA includes positional, parallax and proper motion measurements for ≈ 1.3 billion stars. Our catalogue was cross-matched with GAIA to obtain astrometric and colour measurements for each star in the GAIA data release 2 catalogue that was also in our catalogue.

The proper motion and parallax measurements from GAIA will allow for the investigation into the astrometric features of the stars in our catalogue not provided by our original data-set. The GAIA colour measurements 'G_{BP} and G_{RP}' will be used for internal photometric calibration and classification of the stars in this database. Given that GAIA has a higher angular resolution than our data-set it was common for a comparison between our catalogue and GAIAs to yield multiple matches. If multiple

[11] vizier.u-strasbg.fr

[12] cosmos.esa.int/web/gaia/dr2

matches were found the brightest match was used as this will dominate the photometry in our image. When matching to GAIA the GAIA/IPHAS 'GIPHAS' catalogue (Scaringi et al., 2018) was used. The GIPHAS catalogue was used in an attempt to also obtain measurements from the IPHAS survey (Corradi et al., 2008), however, none of the coordinates provided by our data-set yielded any results for IPHAS. It is likely that this is a software error and future crossmatching with IPHAS may be possible.

4.2 WISE

The Wide-field Infrared Survey Explorer (WISE) is a space telescope that was launched in December 2009 and saw first light on January 2010. The WISE space telescope hosts four filters in the infrared spectrum, 'W1' with W1λ_{mean} = 3.35μm, 'W2' with W2λ_{mean} = 4.6μm, 'W3' with W3λ_{mean} = 11.6μm and 'W4' with W4λ_{mean} = 22.1μm. WISE has an angular resolution of $\approx$6 arcseconds. The WISE survey covers the whole sky and is 95% complete at; W1 $<$ 17.1, W2 $<$ 15.7, W3 $<$ 11.5 and W4 $<$ 7.7. The WISE catalogue was matched with the intention of using WISE colour as a means of classifying any young stellar objects present in our catalogue (Koenig & Leisawitz, 2014).

4.3 2MASS

The Two Micron All-Sky Survey (2MASS) was a survey of the whole sky, taking place between 1997 and 2001 at the Fred Lawrence Whipple Observatory and the Astronomical Observatory Cerro Tololo. 2MASS uses three filters; 'J' with λ_{mean} = 1.2μm which is 99% complete at 16.1 mag, 'H' with λ_{mean} = 1.7μm which is 99% complete at 15.5 mag, and 'Ks' with λ_{mean} = 2.2μm which is 99% complete at 15.1 mag. '2MASS' has a spatial resolution of 4 arcseconds. The 2MASS catalogue was obtained at the same time as the 'WISE' catalogue as part of the 'ALLWISE' catalogue (Cutri & et al., 2014). In addition to the WISE catalogue, the 2MASS catalogue was also matched to perform stellar classifications based on colour.

4.4 Obtaining Cross-match

The python package `Astroquery` (Ginsburg et al., 2019) was used to match the catalogue of stars from our data-set with other catalogues. `Astroquery` allows the user to input a coordinate for querying. The coordinate is then queried with `Astroquery` using a circle with a diameter of 4.37 arcseconds around it (as described in Sect. 3.2.2). Each coordinate in our data-set was matched with the GIPHAS and ALLWISE catalogues. Given that these catalogues have a higher angular resolution than our catalogue multiple stars may fall within our search area.

More than one star could be found in a query if the angular separation of the stars is less than the angular resolution of our catalogue or one of the stars returned is below the completeness of our catalogue. In either of these cases, the photometry of the brightest star found in the match would dominate the photometry of our catalogue; thus, if multiple stars are found for a single search the brightest star is chosen. The brightest stars in the 'G' and 'W1' filters were used from GIPHAS and ALLWISE, respectively.

Each match found with `Astroquery` returned a measurement of the angular separation between the search coordinates and the coordinates of any stars found. This was also added to the database to be used as a quality check. A match with a large angular separation might be indicative of a false or otherwise problematic match.

All of the data extracted from the cross-match was appended to our catalogue `ID_DB`. From Fig. 8 (within Sect. 3) we can see that `ID_DB` also features a 5-bit binary flag; the flag is used to indicate any potential issues with information obtained from an external catalogue. Each binary bit in the 5-bit flag represents a different condition. The flag is stored as a single number ranging from 0 to 31. The first 2 conditions '+1' and '+2' describe the magnitude of `GAIA_R` and `WISE_R`. The third condition '+4' represents a significant magnitude offset between `GAIA_G` and the average magnitude obtained from our data-set. If a star is not found within either the GAIA or ALLWISE data-set it is given a 0 magnitude, the last two conditions in the 5-bit flag '+8' and '+16' represent a no match found for GAIA or ALLWISE. The nature of an N-bit binary flag is that each combination of conditions can be represented by a unique number whose constituent conditions can be easily identified. A flag of 7 would be formed of $1 + 2 + 4$, this represents a radial offset of `GAIA_R` and `WISE_R` larger than 1.5 arcseconds and an average magnitude less than `GAIA_G`+1.

Of the 19,898 stars, 18,596 stars found a match with the GAIA catalogue and 17,556 stars found a match with the ALLWISE catalogue. Any stars that can not find a match were given a magnitude

of 0 mag and the corresponding flag (discussed in Sect. 3.1.3) was raised. Figure 19 shows a right ascension vs declination plot of all data-points queried with the `Astroquery` software. Points shown in green represent a query that yielded a result for both GAIA and ALLWISE. Points shown in red represent a query that did not match with either the GAIA or ALLWISE catalogues. Points shown in blue represent a match with GAIA but not ALLWISE, and points shown in yellow represent a match with ALLWISE but not GAIA.

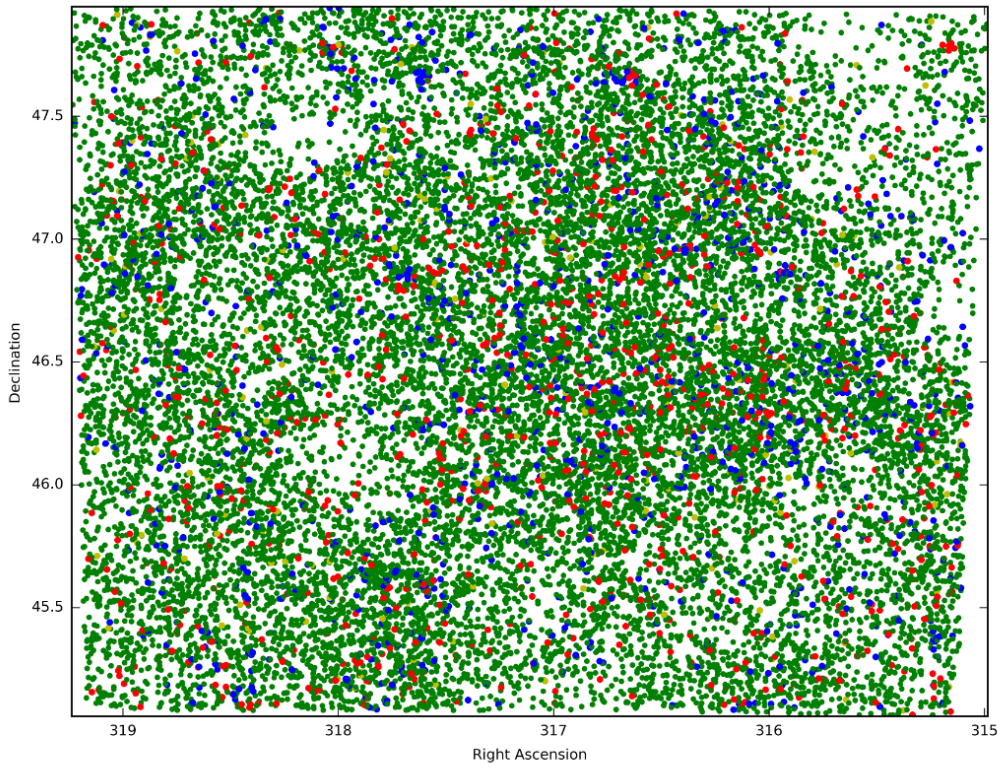

Figure 19: RA - Dec plot of all stars queried with the `Astroquery` software. Points shown in blue represent a match with GAIA but not ALLWISE and points shown in yellow represent a match with ALLWISE but no GAIA. The green points represent a successful match with both GAIA and ALLWISE and the red points represent an unsuccessful match with either.

Compared to our data-set, the GAIA, 2MASS, and WISE catalogues all have a higher angular resolution and a deeper completeness limit. Given this, there should be significantly less failed queries. The coordinates of each unsuccessful query were over-plotted onto an example image from our data-set. Figure 20 shows a small segment of an image from our data-set where each query that failed to return a match has been highlighted with a red circle of a diameter equal to that used in the query (4.37 arcseconds). It can be seen from Fig. 20 that all of the four unsuccessful queries are in between two relatively bright stars.

Images with poor seeing are likely to merge some apparent binaries. In addition, it is possible for these merged stars to be detected as a single star by the `Source EXtractor` software.

The coordinates given to these erroneously detected stars will not be the true coordinates of any constituent star. If enough of these false stars are found by the extraction software they will have been added to the database. It is possible that these false stars could have existed in the database undetected until this point. Thus we can also use the lack of a GAIA or ALLWISE match as an indicator for a potentially none real star. It could also indicate a potential cataclysmic variable that was not undergoing an outburst at the time these surveys collected their data.

Table 3 shows the data obtained via the `Astroquery` python package.

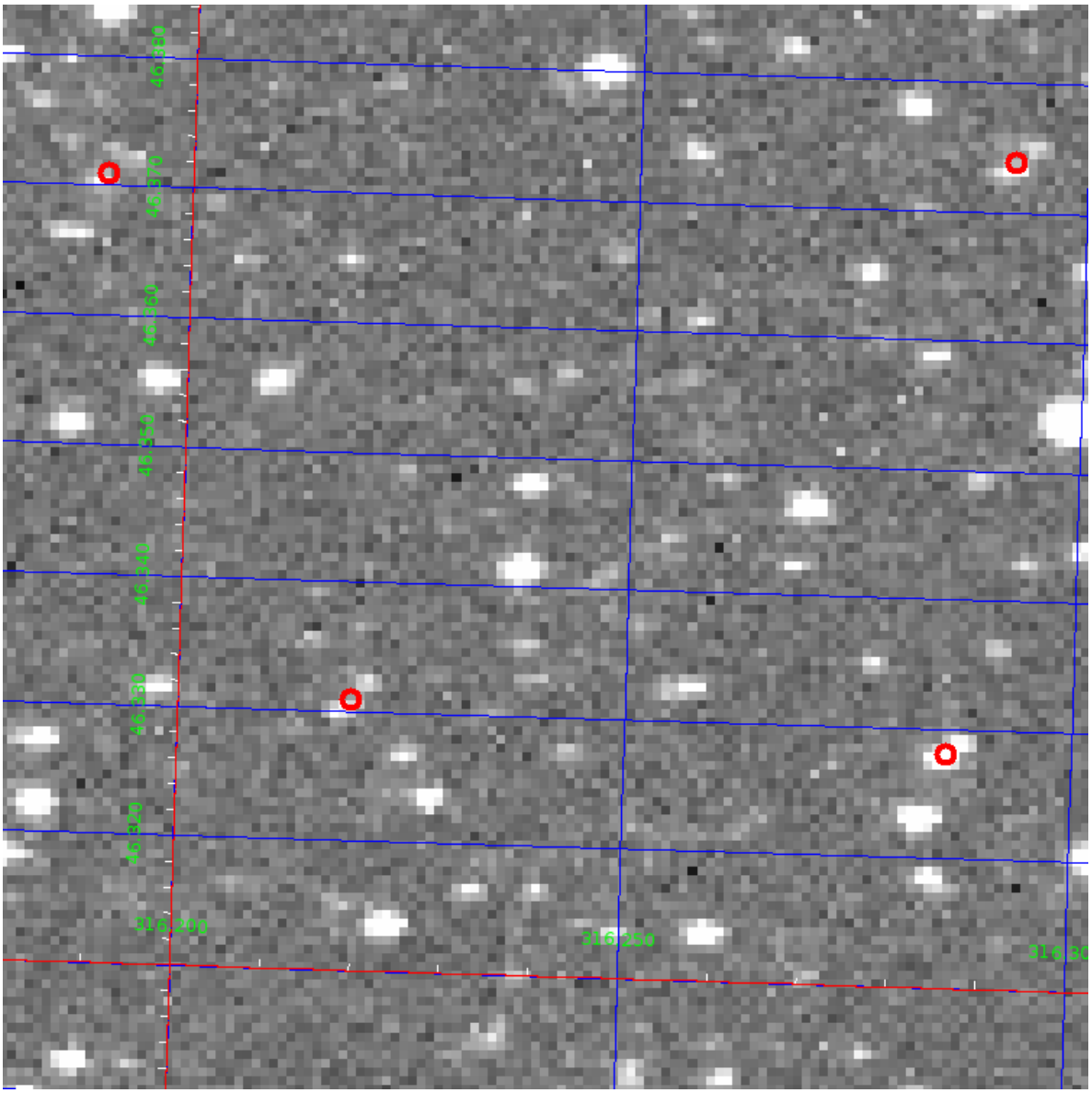

Figure 20: A small segment of an image from our data-set. Each unsuccessful query of the `Astroquery` software is over-plotted with a red circle with a diameter of 4.37 arcseconds.

Table 3: A table representing all of the data extracted from the `Astroquery` python package

0	GAIA_r	radial seperation between search coordinates and GAIA coordinates
1	GAIA_plx	GAIA measured parallax
2	GAIA_e_plx	error of GAIA measured parallax
3	GAIA_pmRA	GAIA measured proper motion in RA
4	GAIA_e_pmRA	error of GAIA measured proper motion in RA
5	GAIA_pmDE	GAIA measured proper motion in DE
6	GAIA_e_pmDE	error of GAIA measured proper motion in DE
7	GAIA_Gmag	GAIA measured magnitude in 'G' filter (very broad-band filter)
8	GAIA_e_Gmag	error of GAIA measured magnitude in 'G' filter
9	GAIA_Bmag	GAIA measured magnitude in 'B' filter
10	GAIA_e_Bmag	error of GAIA measured magnitude in 'B' filter
11	GAIA_Rmag	GAIA measured magnitude in 'R' filter
12	GAIA_e_Rmag	error of GAIA measured magnitude in 'R' filter
13	WISE_r	radial seperation between search coordinates and WISE coordinates
14	WISE_w1mag	WISE measured magnitude in 'W1' filter
15	WISE_e_w1mag	error of WISE measured magnitude in 'W1' filter
16	WISE_w2mag	WISE measured magnitude in 'W2' filter
17	WISE_e_w2mag	error of WISE measured magnitude in 'W2' filter
18	WISE_w3mag	WISE measured magnitude in 'W3' filter
19	WISE_e_w3mag	error of WISE measured magnitude in 'W3' filter
20	WISE_w4mag	WISE measured magnitude in 'W4' filter
21	WISE_e_w4mag	error of WISE measured magnitude in 'W4' filter
22	2MASS_jmag	2mass measured magnitude in 'j' filter
23	2MASS_e_jmag	error of 2mass measured magnitude in 'j' filter
24	2MASS_hmag	2mass measured magnitude in 'h' filter
25	2MASS_e_hmag	error of 2mass measured magnitude in 'h' filter
26	2MASS_kmag	2mass measured magnitude in 'k' filter
27	2MASS_e_kmag	error of 2mass measured magnitude in 'k' filter

4.5 Verifying Cross-Match

The goal for performing a cross-match between our catalogue and other catalogues is to gain insight into the population of our catalogue not provided by our original data-set. We can use the acquired magnitude measurements made with multiple filters to aid in the classification of the stars in our catalogue. We can also use the astrometric measurements from GAIA as a way of distinguishing any physically associated clusters of stars from our population of otherwise individual field stars. We can determine clusters of stars by investigating for co-moving groups of stars in physical proximity of each other using GAIA's astrometric measurements.

4.5.1 GAIA Parallax

One of the reasons for matching our data-set with the GAIA catalogue is to make use of GAIA's parallax measurements. Figure 21 shows the distribution of parallax measurements obtained from GAIA (i.e with a cross-match flag of less than 8). Figure 21 shows 171 negative parallax values. It is explained in Luri et al. (2018) that negative parallax arises from parallax measurements with large uncertainty. It is explained in Luri et al. (2018) that the large uncertainties which lead to negative parallaxes are caused by 'noisy observations' of stars with a high proper motion which leads to GAIA's orbit around the sun to be incorrectly accounted for.

It is discussed in Schönrich et al. (2019) that a colour and magnitude dependent uncertainty is present in GAIA's parallax (see Fig. 9 of Schönrich et al. (2019)). These offsets have not been considered in the provided error of the parallax 'GAIA_e_plx'. Furthermore, dimmer stars will suffer for lower signal-to-noise, as will redder, more dust obscured stars. From this Schönrich et al. recommend the following quality cutoffs for using GAIA parallax:

- $0.5 < G_{BP} - G_{RP} < 1.4$
- $G_{BP}, G, G_{RP} > 0$

Luri et al. (2018) discusses the caveats of converting GAIA's parallax 'ρ' to distance by using $distance = 1/\rho$. Luri et al show that treating $distance = 1/\rho$ is sufficient for an approximate indication of distance for distances of $0.5\text{mas} < \rho < 2\text{mas}$ and where $\sigma_\rho << \rho$. Performing these quality cuts reduces the number of usable parallax measurements from 17171 to 4695. As the above-listed cuts significantly reduce the number of parallax measurements, parallax measurements outside of our cuts

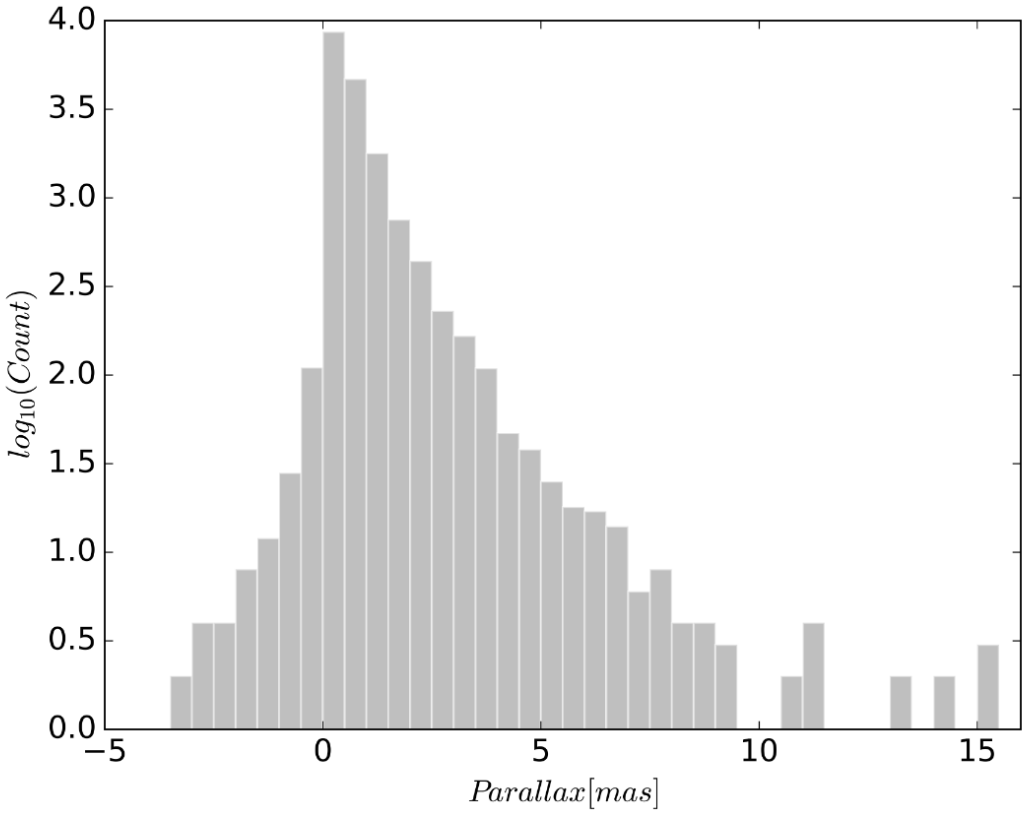

Figure 21: A $\log_{10}$ space histogram showing the distribution of available parallax measurements from GAIA that have been added to our catalogue. It can be seen that there are 171 erroneous negative parallax measurements provided.

are not removed from the database and are instead used as an indicator of quality. We will also apply the suggested zero point correction of -0.0523mas from Leung & Bovy (2019). Figure 22 shows a distribution of stars as a function of distance at a set volume. Figure 22 was calculated from the 4695 quality-filtered parallax measurements. We can see a steadily decreasing population, which is due to the magnitude limit in our data.

We can pair the parallax measurements of these stars with the coordinates in our catalogue to build a 3-dimensional representation of our catalogue. The left of Fig. 23 shows to plots of either right ascension (top) or declination (bottom) vs distance. The right graph of Fig. 23 shows a right ascension - declination plot where distance has been plotted as colour. It can be seen that there is not much structure in the 3D representation of our catalogue.

There is a lack of depth between $\approx 315.0^\circ$ $+47.0^\circ$ (J2000) and $\approx 316.0^\circ$ $+48.0^\circ$ (J2000) where there is are proportionally less stars at a distance larger than 800pc. The lack of depth in this region is likely due to extinction at $\approx$ 800pc. Investigating this region, we can see that there is a dark cloud of extinction known as 'LDN 954'. It is stated in Tomita et al. (1979) that 'LDN 954' is $\approx$ 300pc away with a radii of $\approx$ 0.61pc. Figure 24 shows how the region in question shows substantially fewer stars in the DSS2 colour image.

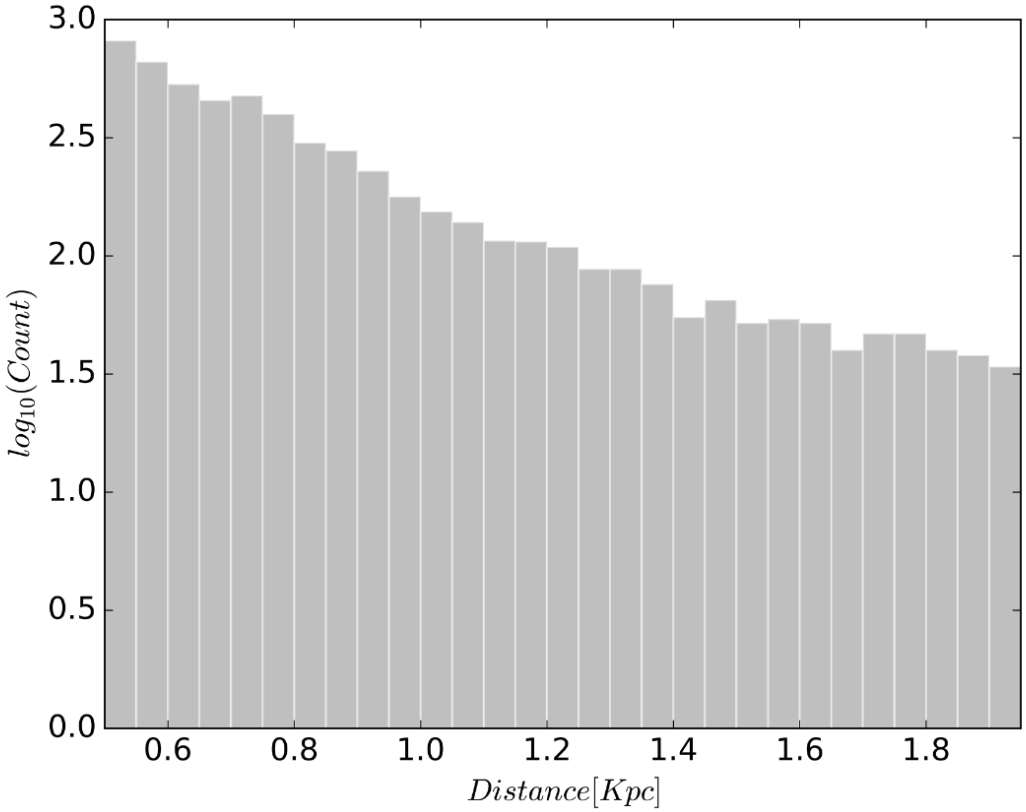

Figure 22: A $\log_{10}$ space histogram showing the distribution of distance measurements from the 4695 quality cut parallax measurements in our catalogue. The graph shows a steadily decreasing population in our catalogue as a function of distance which highlights the extinction present and the magnitude limit of our data. The volume considered in each bin has been held constant as distance increases.

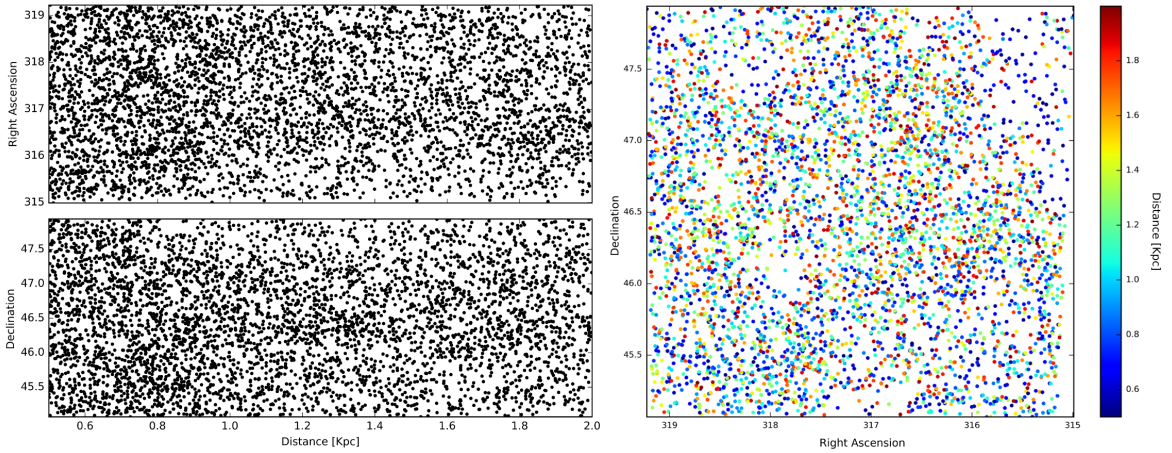

Figure 23: Left: An RA vs Distance plot on the top and a Dec vs Distance plot on the bottom. Right: An RA vs Dec plot with Distance plotted as colour. Both plots highlight some areas of lower population count. Such areas are likely caused by high extinction.

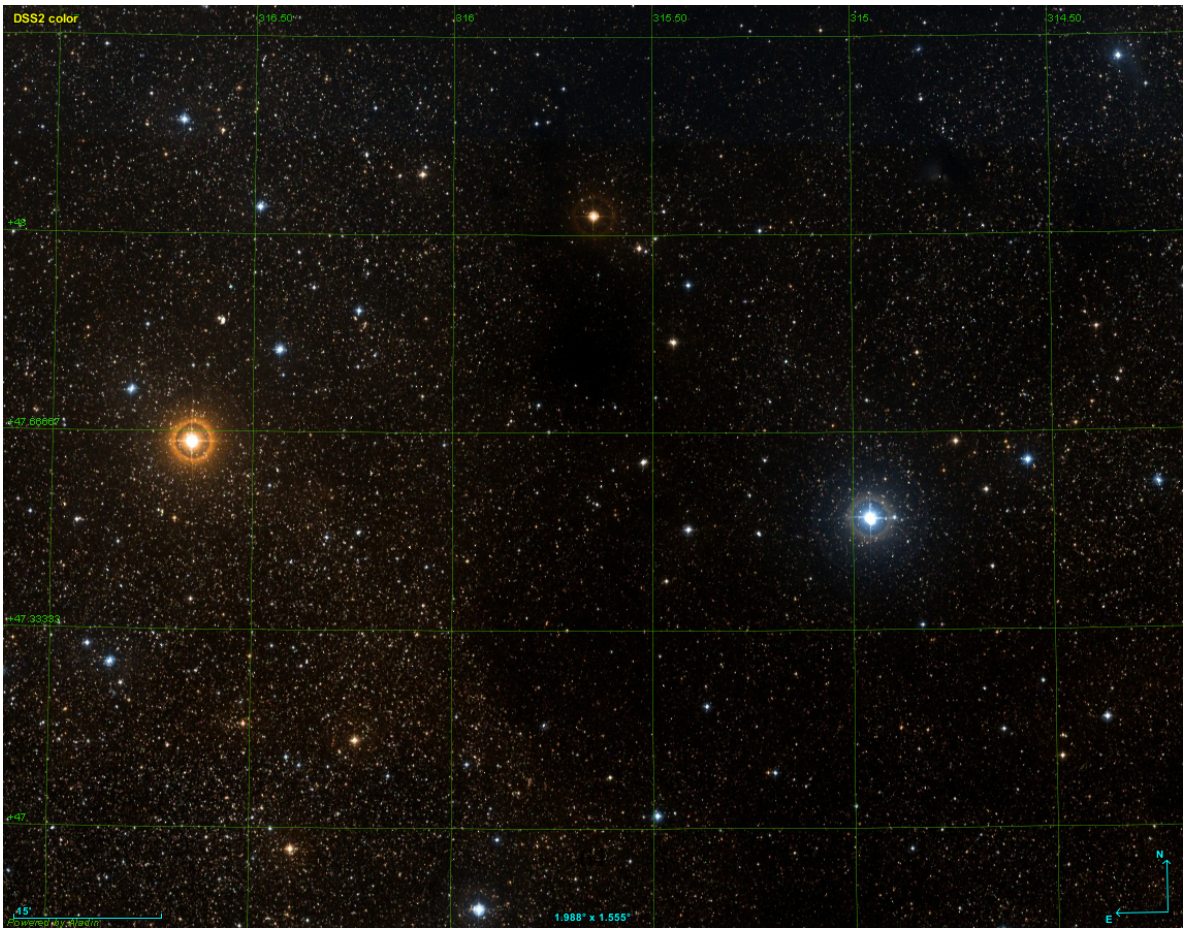

Figure 24: DSS2 colour image highlighting the lack of stars inside LDN 954. This shows how +47.0° - +48.0° (J2000) and 315.0° - 316.0° (J2000) fall within the dark cloud.

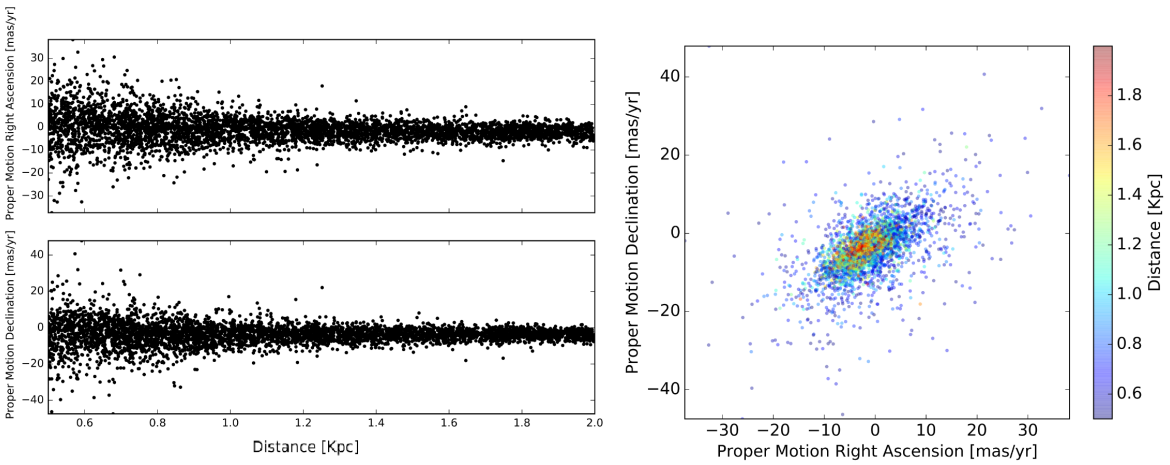

Figure 25: Left: a proper motion plot vs distance. Declination is plotted on the top and right ascension plotted on the bottom. Right: Proper motion in right ascension vs proper motion in declination with the third colour axis of distance. Both graphs highlight the expected inverse relationship between proper motion and distance

4.5.2 GAIA Proper Motion

We can use GAIA's proper motion measurements to search for co-moving clusters in our catalogue. The proper motion measurements provided by GAIA are calculated as the change in a star's angular position over time, measured in mas per year. Proper motion is the observed change in the apparent position of an object in the sky. Thus a star with constant space velocity would have a proper motion inversely proportional to its distance from the observer.

Figure 25 shows two plots both of which represent proper motion vs parallax. These plots show the expected inverse relationship between proper motion and radial velocity. It can be seen that the range of proper motions decreases as a function of distance. This helps verify the validity of the data acquired by the catalogue cross-match. Only parallax measurements that meet the parallax criteria stated above are used here.

Figure 26 shows a plot of all of the available proper motion measurements in our catalogue. The right panel is a heatmap where the colour scale is in $\log_{10}$ space. It can be seen from the slightly negative proper motion in both right ascension and declination that the majority of stars are following Keplerian orbit in the galaxy, so they are travelling with the rotation of the milky way.

Some of the stars with a high proper motion relative to their distance are possible halo stars (Brown

et al., 2005). Halo stars do not travel in the plane of the galaxy and do not orbit with the net rotation of the milky way. Halo stars orbit the milky way separately to the milky way disk and as such have a relatively high proper motion.

As mentioned in Sect. 3.2 the magnitudes used in our database are instrumental apparent magnitudes. The primary use for our database is to perform relative photometry, thus the absolute value of our recorded magnitude is irrelevant. To properly remove the instrumental offset from our database we must account for magnitude and colour dependent instrumental offsets. As this method would be beyond the scope of this report and would currently not be of any use, such a precise method of determining instrument offset has not been performed.

We can instead perform a less precise comparison. We compare the median magnitude ('Mag Avg' in Fig. 8) for each star in our catalogue with a corresponding GAIA measurement to that of GAIA's magnitude, 'Mag Avg$-G_{RP}$'. We can then group the differences to all magnitudes above 14 instrumental magnitude and all magnitudes below 14 instrumental magnitudes. If the difference between the two groups is substantial enough that it would have been relevant even for a less precise method then a separate method would have been taken. It was found that the instrumental offset was:

- Where Mag Avg$>$ 14$-$Mag Avg$-G_{RP} = 4.90$
- Where Mag Avg$<$ 14$-$Mag Avg$-G_{RP} = 4.96$

Thus for the purpose of this report, we will treat the instrumental offset as ≈ 4.93 mag We can use Eq. 13 (Gaia Collaboration et al., 2018b) to generate absolute magnitudes using the magnitude measurements obtained from our data-set with GAIA's parallax measurements:

$$M = MagAvg + 5 + 5log_{10}(\rho/1000) \tag{5}$$

Where 'M' is the absolute magnitude based on the average magnitude for each star 'Mag Avg'. 'Mag Avg' is the apparent magnitude of a star determined as the average of all measurements of that star in our data-set. The instrumental offset discussed above has been accounted for with 'Mag Avg'. 'ρ' is the parallax measured by GAIA. Figure 27 also shows a clear main-sequence line, and it is apparent that our catalogue consists predominately of main-sequence stars and giant stars.

Figure 27 also shows some extinction in the upper main sequence. As it was possible to create a Hertzsprung–Russell diagram using magnitudes taken from our data-set along with colour and astro-

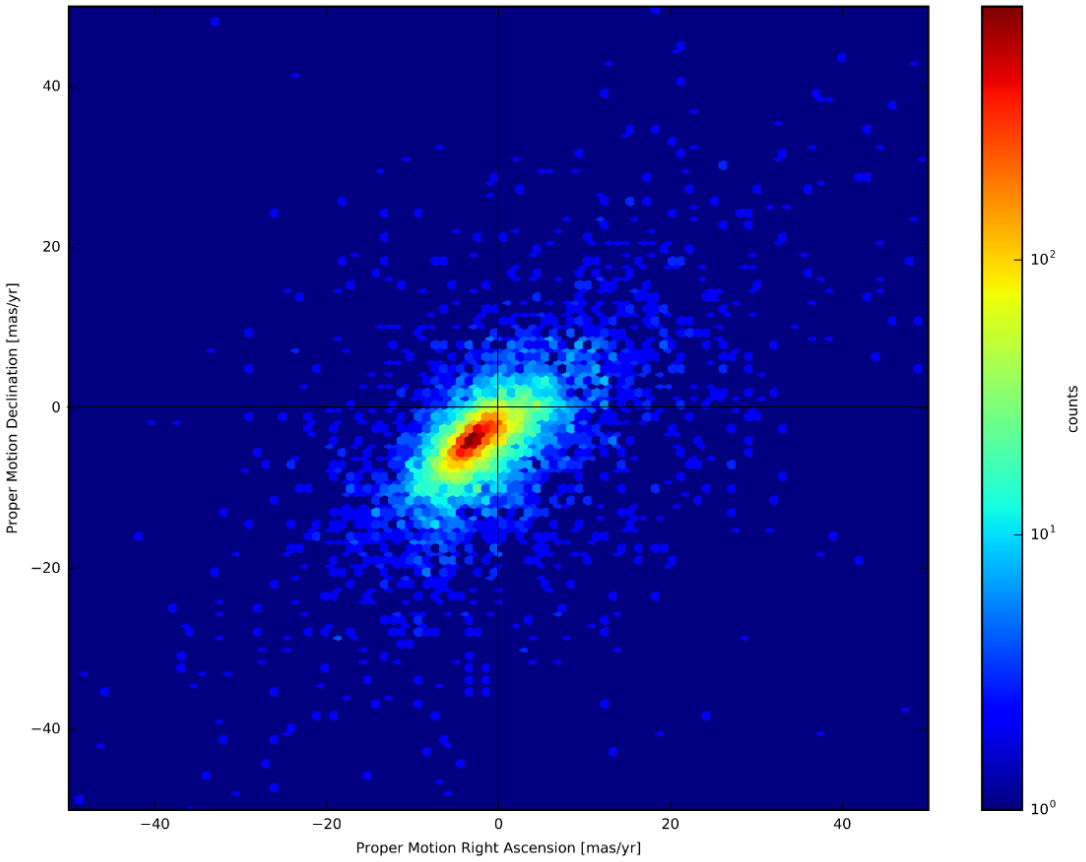

Figure 26: A heatmap showing the distribution of proper motion measurements. It can be seen that the majority of stars shown are travelling with the rotation of the galaxy. A line through 0 on each axis has been added to clarify the rotation.

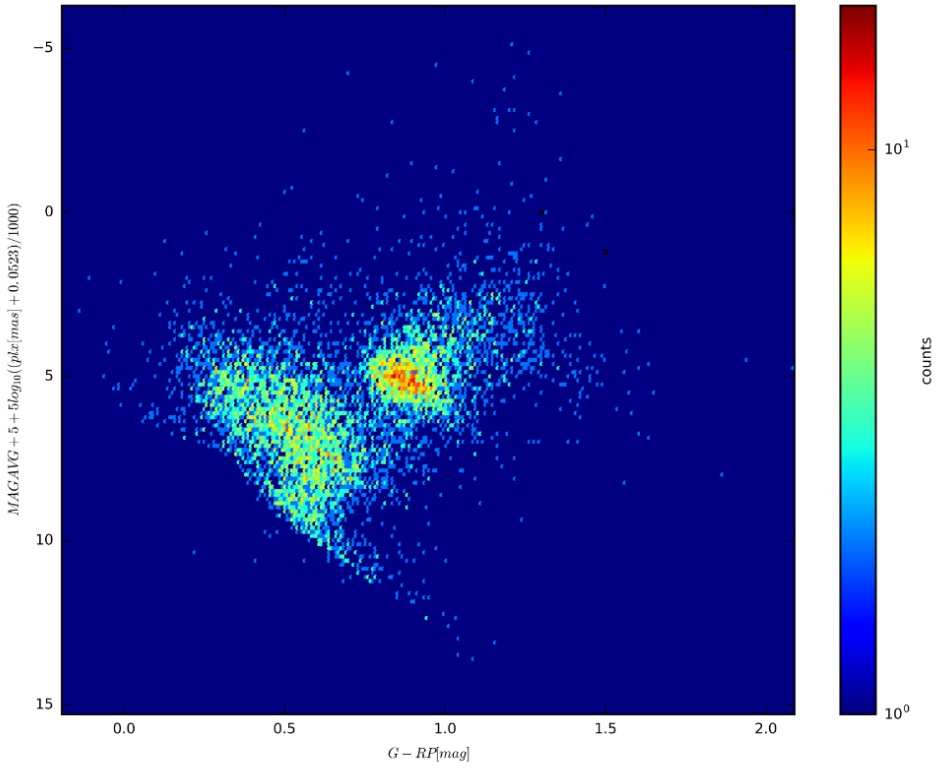

Figure 27: Hertzsprung–Russell diagram of all of the stars in our catalogue with a GAIA measurement. Absolute magnitude has been calculated using the average magnitude of each star from our database and its corresponding parallax from GAIA

metric measurements taken from our cross-matched data, we can conclude that the cross-match was successful.

5 Internal Photometric Calibration

5.1 Systematic Magnitude Shift

At the end of Sect. 3.2.3 we can see that Fig. 18 highlights a systematic shift in the magnitude of the star that is present at certain times within our data. To accurately investigate real variations in the brightness of stars in our catalogue, we must remove the systematic offsets without destroying any real variations in the magnitude of a star. If the non-real changes in magnitude are left in then any measurement of a stars variability would be untrustworthy as the source of the variability would be affected by the non-real photometric shifts.

To remove the problematic photometric shifts we must identify the source of them. In Sect. 2.2.5 we discussed the inhomogeneities in the flat fields provided. As discussed in Sect. 2.2.5, Mr Waterman explained to us that he often used an 'evenly illuminated' perspex sheet as a way of generating flat fields. As the perspex sheet was likely not manufactured with the intention of being used for flat fields, it is likely that the flat fields generated via the perspex sheet are not homogeneously illuminated as the perspex is not guaranteed to have a homogeneous density. The result of using potentially inhomogeneous light for a flat field is that any science frames reduced using a flat field with this issue will gain a CCD position-dependent magnitude shift. Mr Waterman also informed us of a problem where water would condense onto the optics of the telescope. Water present in the optical path of the science of flat field frames would also cause a magnitude offset dependent on CCD position for both the flat and science frames. There could also be a CCD position-dependent colour offset for any science frames taken with water condensation in the optical path. Mr Waterman also used a separate telescope for a short time. It is possible that some calibration frames taken on one telescope have been used to calibrate science frames taken on a separate telescope. No logs were taken to indicate when a different telescope was used. We were not previously aware of these problems at the start of this project and they only came to light as the project was worked on, hence the photometric correction being performed at this stage.

Using calibration frames from a telescope separate to the one used for science frames will cause the flat fields used to incorrectly account for any inhomogeneities present due to the optics of the telescope.

Of particular problem would be the difference in the vignetting of the images. Mr Waterman informed us that the second telescope he had used had 'much less vignetting'. A miss-match of flat fields and science frames from this telescope to another will cause the vignetting to be incorrectly

Table 4: A table showing the instrumental magnitude, colour ($G_{BP}-G_{RP}$) and average CCD position in 'X' and 'Y' for each of the stars featured in Fig. 28

Mag	Colour (B-R)	X	Y
11.984	-0.15	3353.428	744.774
12.980	0.0071	1389.192	906.063
14.006	1.4774	712.864	102.508
14.957	0.6943	2442.158	220.739
16.000	0.2392	2573.578	121.652
17.047	1.7524	2336.673	98.313
18.004s	1.0194	2794.307	96.516

calibrated and thus a photometric offset will arise in the form of vignetting.

As the camera has remained constant for this data set and hence pixel to pixel variation have been appropriately corrected for. Thus, we are concerned with fixing the large scale structure present due to the issues described above.

Figure 28 shows the light curves of seven stars each approximately a magnitude apart. This figure uses '`Mag_Cali`' and '`Mag_Cali_Error`' as seen in Fig. 8. Figure 28 highlights the photometric offset (As seen in Fig. 18). The figure also shows which flats were used for each data-point by the use of the colour bar. It can be seen that while the flats associated with noticeable changes in the photometry are the same, they are also sometimes used without causing a noticeable error in the photometry. This is indicative of some of the flats being appropriate for some of the science frames but not others (as discussed in Sect. 2.2.5). Hence, simply removing the problematic flats will not fix the issue of the photometric offset. Table 4 has been provided to show relevant information pertaining to each of the stars featured in Fig. 28.

5.2 Photometric correction methodology

The goal of performing this photometric correction is to remove any erroneous photometric offset present in the images of our data-set. We can do this by comparing the magnitudes of several none variable stars we expect to have in an image with the magnitudes they actually have in that image. Onwards we will refer the difference between a stars average magnitude and the magnitude in a given

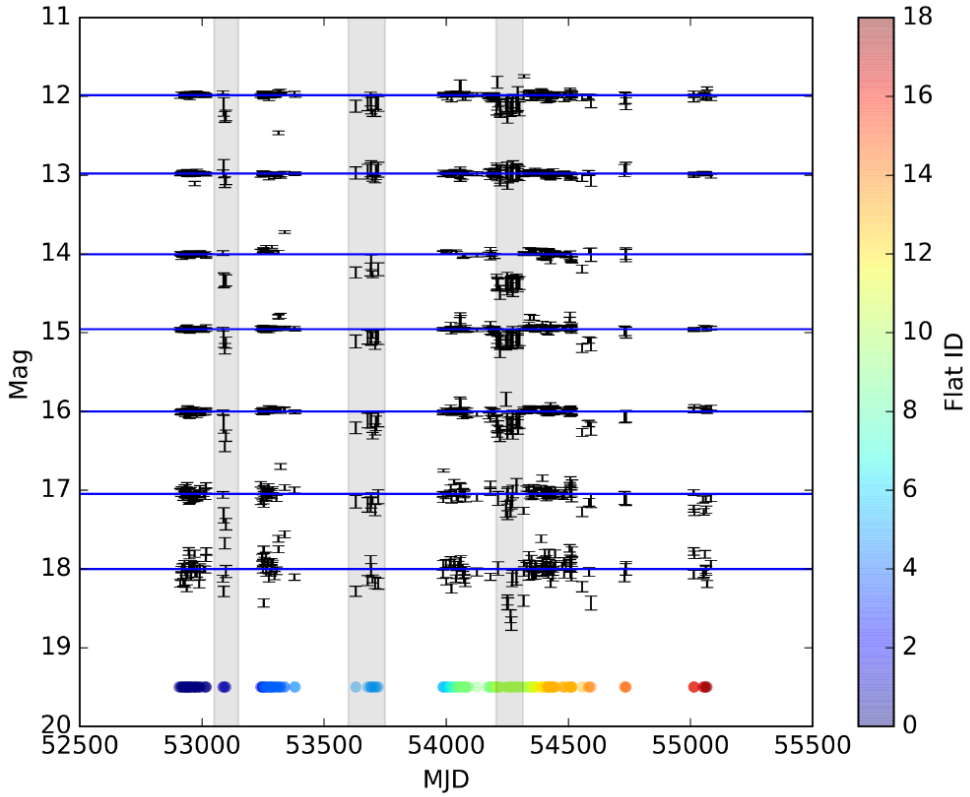

Figure 28: A light curve featuring seven stars spaced approximately 1 magnitude apart. The flat field calibration image used for each of the data-points can be seen by the colour bar. Areas of significant erroneous photometric offset are highlighted with grey bars. Table 4 has been provided to show relevant information pertaining to each of the stars shown.

image as the 'photometric offset'. Equation 6 shows how the photometric offset is defined. Where the magnitude that is being corrected (M_{Image}) is the `Mag_Cali` that was recorded in a given image. In Sect. 2.2.4 we describe how `Mag_Cali` is derived.

$$Offset = M_{Image} - M_{Average} \tag{6}$$

The correction is intended to remove any photometric offset from M_{Image} such that $M_{Corrected}$ =M_{Image}−Offset. It is possible to use a different magnitude for $M_{Average}$ to correct for the photometric offset that will simultaneously correct our magnitude into real apparent magnitudes and not just instrumental magnitudes. However, this will involve a further matching of all of our stars to magnitude measurements specifically in the filter that was used in our data-set (Cousins R). This will have the added issue of increasing the importance of the colour term. Any observations of the stars from the matched catalogue not taken under perfect photometric conditions will induce a colour term. If we were to use an external catalogue for the photometric correction we will need to ensure that the potential differences in colour are accounted for, otherwise, we risk inducing a further photometric offset. Given that the science goals of this database are to perform relative photometry for each star individually, calibrating our stars this way would not provide any information that will be currently useful. The actual magnitude of the star is not of high importance such that we require more than what is provided by our cross-match with GAIA, 2MASS and WISE. Hence, we will be using the internally calculated average magnitude to calculate the offset as described in Eq. 6. It would be possible to re-perform this calibration with a matched catalogue at a later date with minimal adjustments to the program.

The photometric offset can be modelled with the use of a multi-variable N^{th} order polynomial. We can formulate this polynomial so that it models the photometric offset as a function of magnitude, colour and CCD position '$\mathcal{P}_N$(X,Y,m,col)' (Eq. 9) (Evitts et al., 2020). If we have a sufficient amount of stars covering a range of magnitudes, colours and CCD positions, we can use least-squares regression to generate the coefficients of the multi-variable polynomial. For this purpose, we used the `Curve Fit` python package. Once the least-squares regression program has provided a fit to the photometric offset for all of the none variable stars, we can apply the fit to every star in the image, inclusive of variable stars. Figure 29 shows an outline of the steps taken in the correction procedure.

Most images have $\approx 50,000$ stars, while testing this program on the smaller database (discussed in Sect. 3.1.4) it was found that enough calibration stars were always present. However, when performing the calibration on the entirety on the large database, there was an issue. Some poor quality images

can have fewer stars than the number of free parameters given to the least-squares regression. Hence, a check was performed before the correction to ensure that the number of stars was larger than the number of free parameters. Equation 9 shows that there are 23 free parameters for the least-squares regression to solve.

5.3 Identifying Calibration Stars

To determine the photometric offset we compare the magnitude of a star in a given image against the average magnitude of that star. We calculated the average magnitude each star has with the sections of photometric offset removed. This allowed us to identify where the majority of the photometric offset occurs by eye using Fig. 28. Any difference between the average magnitude and the magnitude measured in the image is considered to be a photometric offset generated internally and is not considered to be a real magnitude difference. As it is common for a star to have is magnitude vary as a function of time naturally, we must ensure that any stars we use for calibration do not vary naturally. This method of only using non-variable stars will allow us to preserve the real variations in a star's magnitude while removing non-real magnitude variations. It also improves the accuracy of our calibration as we will gain a more reliable measurement for the photometric offset.

A caveat of this method is that it does not consider any colour variations a star holds. Our data is only single filtered, and we only have colour data provided by the cross-matching with other catalogues (Sect. 4). Hence, we do not have the same high cadence information for the colour of a star as we do its magnitude. Thus, we can only select stars based on how much they vary in magnitude, and we have no consideration of variability in the colour of a star. To do this we will make use of the 'Stetson Variability Index' as a method of quantifying the real variability of stars.

5.3.1 Stetson Index

The Stetson Index (Stetson, 1996) 'I' is a method of measuring the variability of a star that is based on how much a star's magnitude varies while taking into account the error associated with the magnitude of that star.

$$\delta = \sqrt{\frac{N}{N-1}} \frac{m - \overline{m}}{\sigma_m} \tag{7}$$

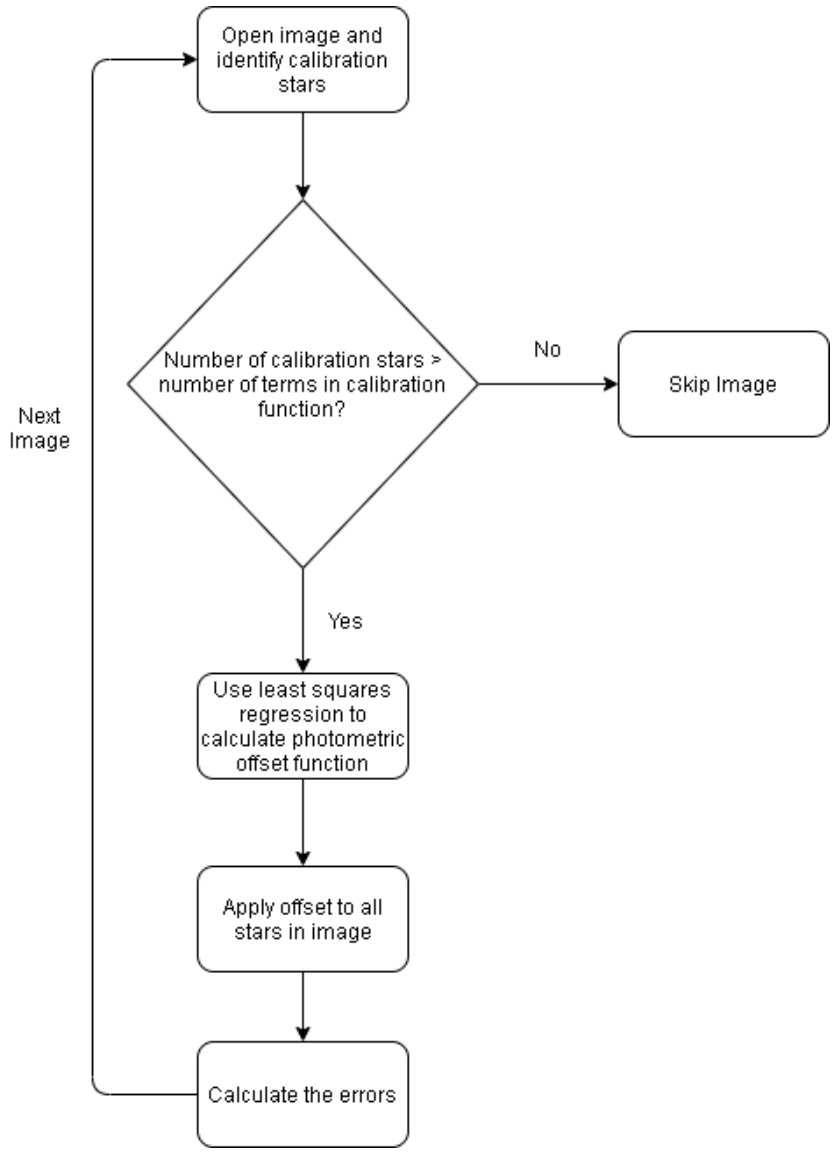

Figure 29: Flow chart outlining the basic process of the internal photometric calibration. An image is opened and checked to see how many stars in the image are present in a predetermined list of non-variable stars. If the number of calibration stars is less than the number of terms in the function used for calibration then the image is ignored. Otherwise, least-squares regression is used to calculate the coefficients for a pre-determined polynomial. The polynomial models the photometric offset as a function of magnitude, colour and CCD position. After the coefficients have been determined, the polynomial is applied to all-stars in the image.

Where N is the number of magnitude measurements for the star (in our case this is known as the number of data-points), m is the magnitude and $\overline{m}$ is the average magnitude.

$$I = \frac{\sqrt{\delta^2 - 1}}{N + 1} \tag{8}$$

We can calculate the Stetson Index of every star in our catalogue. Figure 30 shows a plot of Stetson index vs instrumental magnitude. It is seen that there is some magnitude dependency on the variability index a star receives. This is likely due to a lower signal-to-noise for dimmer stars thus appearing more variable. While it is hoped that the Stetson Index will capture this, this is not perfect, as our errors at this stage are likely not fully accurate. In Evitts et al. (2020) it is discussed that a Stetson Index of I$<$ 0.1 is appropriate for defining a star as non-variable. Hence, we have 6580 non-variable stars to use for calibration.

All of the stars determined to be suitable for calibration are added to a list that can be later queried via their `Object_ID` in the correction.

5.4 Generating Correction

For each image, the calibration stars used to generate the correction are found by querying the list of calibration stars previously determined. `Object_ID` along with `Image_ID` is used for finding all the data-points for a given image corresponding to the calibration stars. If the amount of calibration stars in a given image is less than the number of terms in the polynomial used for calibration, the image is skipped and subsequently removed from the database as it is likely to be of poor quality and is not correctable.

Once all the calibration stars have been identified in a given image their average magnitudes, magnitude in that image, position on the CCD in that image and their GAIA colour (B-R) is recorded.

To accurately model the photometric offset as a function of magnitude, colour and CCD position, we must construct an equation which appropriately models all of the features without being needlessly computationally intensive. The computation times for the calibration of the full data-set are likely to be on the order of months. We must ensure that there are no wasteful terms in the equation that do not provide improvement to our fit of the photometric offset.

The equation was started as a 3^{rd} order polynomial with all cross-terms in magnitude, colour and CCD position $\mathcal{P}_3$(X,Y,m,col). From here we added and removed terms in an iterative process. After every change to the equation, the correction process was performed and the fit was investigated where

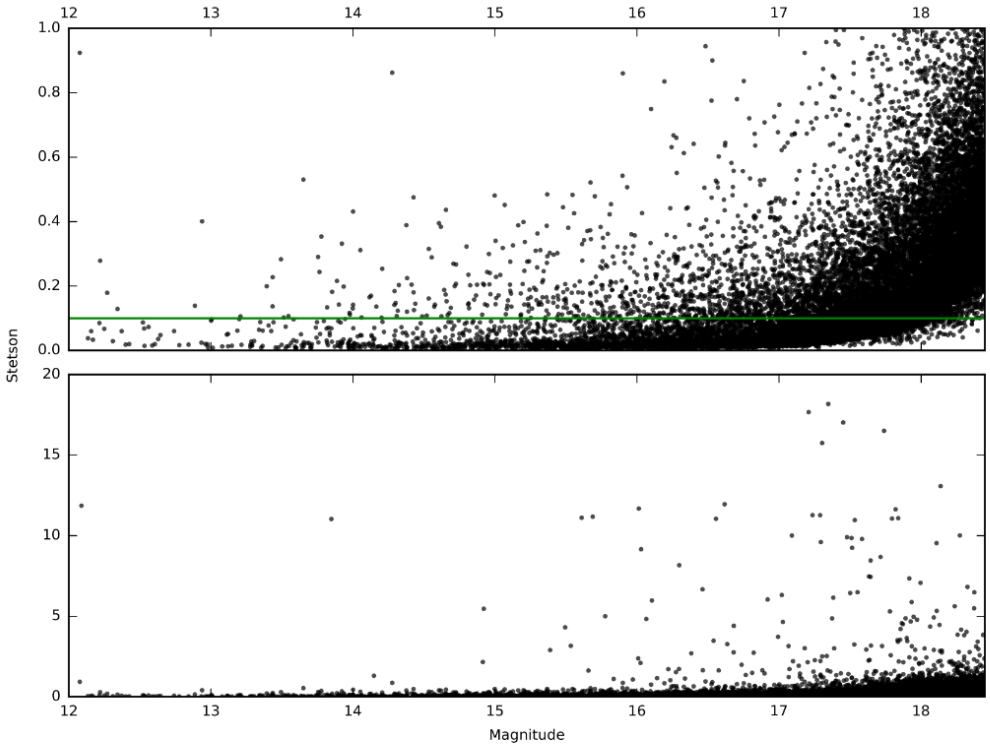

Figure 30: Magnitude vs Stetson Index. Shows the range of variability present in our catalogue. Top: Shows Stetson Index between 0-1. The green line is at a Stetson Index value of I=0.1. Any stars with a Stetson Index of less than 0.1 are considered to be non-variable (Evitts et al., 2020). Bottom: Shows Stetson Index between 0-20.

the final offset was compared to the previous iteration. After multiple iterations, it was found that the CCD position was the most important group of terms in the equation. Equation 9 is a 4^{th} order polynomial in magnitude, colour and CCD position. It has cross-terms in the X and Y CCD position up to 3^{rd} order and one cross term in colour-magnitude.

We can not expect to fully remove all the fine structure present in some flat fields as some of it would be virtually impossible to model via a polynomial. Hence, we would still recommend using nearby comparison stars for more accurate photometry after the photometric correction procedure.

Equation 9 shows the final form of the equation used for the photometric calibration.

$$\begin{aligned}\mathcal{P}_{22}(X,Y,m,col) = \\ & P_0 + P_1X + P_2X^2 + P_3X^3 + P_4X^4 \\ & + P_5Y + P_6Y^2 + P_7Y^3 + P_8Y^4 \\ & + P_9XY + P_{10}X^2Y + P_{11}XY^2 + P_{12}X^2Y^2 + P_{13}X^3Y + P_{14}XY^3 \\ & + P_{15}m + P_{16}m^2 + P_{17}m^3 + P_{18}m^4 \\ & + P_{19}\,col + P_{20}col^2 + P_{21}col^3 + P_{22}col^4 \\ & + P_{22}\,m\,col\end{aligned} \tag{9}$$

Where 'P_N' are the free parameters that the least-squares regression can vary in order to minimise the difference between our fit and the photometric offset. 'X' and 'Y' are the X and Y positions on the CCD. 'm' is the stars instrumental magnitude `Mag_Cali` and 'col' is the (B-R) colour of the star as measured from GAIA (see Sect. 4).

Figure 31 shows each of the four variables being used for the fit vs the magnitude offset. The fit is shown as the red line. The image used as an example appears to have a systematic photometric offset as a function of the y-axis on the CCD. It can be seen that the fit on the y-axis follows the trend of the photometric offset. To simplify, the correction subtracts the red line from the magnitude of all of the stars, thus 'flattening' the photometric offset.

As expected and seen in Fig. 16 we have substantially more faint stars than we have bright stars. So in order to more fairly weight the brighter stars, we will apply a magnitude dependent weighting on each of the stars being used for the fit Eq. 10 shows how the weight factor for preferably using brighter stars is calculated.

$$Weight = (m - min(m) - 2.0)^2 \tag{10}$$

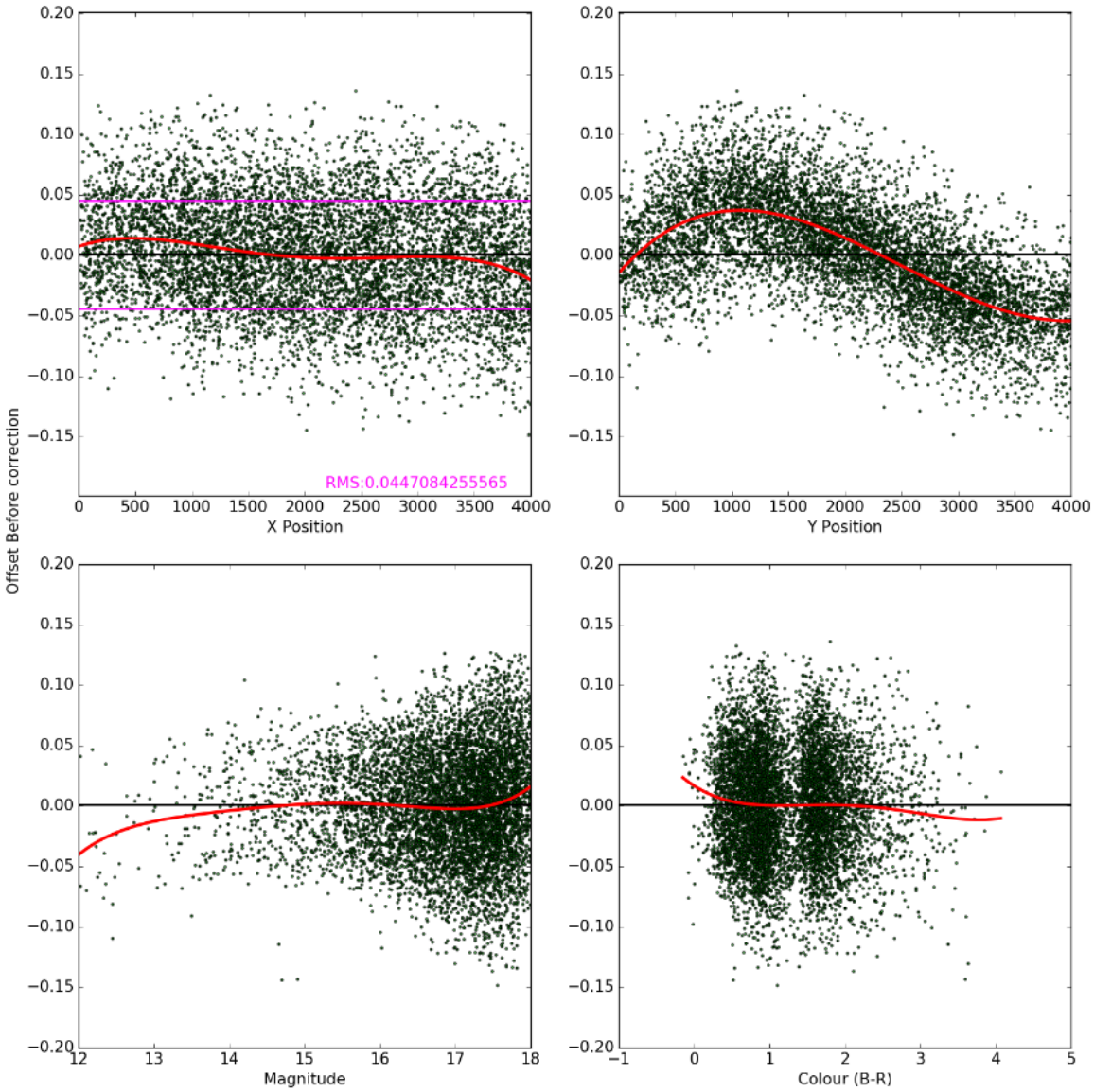

Figure 31: Showing the offset vs each of the variables in the polynomial. Each of the red lines have been plotted with only the terms belonging to the variable used. (Example: For the fit of the offset vs the X CCD position only terms P_0 to P_4 in Eq. 9 were used.) The magenta line shows the RMS of the offset with the RMS written in magenta.

Where m is the calibration stars magnitude and min(m) is the magnitude of the brightest calibration star in the image. The '-2.0' was added to control the scale at which the weight acts and was determined by trial and error. Once the weight factor has been calculated and a fit of how the magnitude of a star varies as a function of magnitude, colour and CCD position we can combine them to calculate a weighted offset Eq. 11 shows how the weighting factor calculated in Eq. 10 is used with Eq. 9 to generate the final fit of the photometric offset for each image.

$$Fit = \frac{\mathcal{P}_{22}(X, Y, m, col)}{Weight} \tag{11}$$

The least-squares regression program varies $\mathcal{P}_N$ to try to match the photometric offset as close as possible. Equation 11 was given to a least-squares regression program in order to produce the fit. The least-squares regression program was allowed to perform 1,000,000 iterations before returning its results.

5.4.1 Sigma Clipping

To account for any extraneous points with a higher than average offset we will use sigma clipping. The sigma clipping process iteratively performs the least-squares regression to fit the photometric offset. After each iteration, any stars outside $3 \times$ Standard Deviation of the distribution of photometric offsets are removed. This process is performed either until none of the calibration stars are outside $3 \times$ Standard Deviation or the clipping process has been performed five times. Figure 32 shows a flowchart which highlights the process of sigma clipping.

This process is performed to ensure the fit provided by the least-squares regression is not significantly influenced by stars with an extraordinary offset in the image. It could be that one or more of the calibration stars happens to land on a star present in the flat field (as seen in Fig. 5). If a calibration star has its photometry disproportionately affected by such an occurrence, it should be disregarded from the fit as to not erroneously affect how the least-squares regression program generates its terms in the polynomial.

Figures 35 and 36 show the before and after for the photometric correction process. They also highlight the sigma clipping process, as it is seen that before the correction there are several stars above 3σ that are not used in the formation of the fit (as seen by the red line) they are removed after the correction.

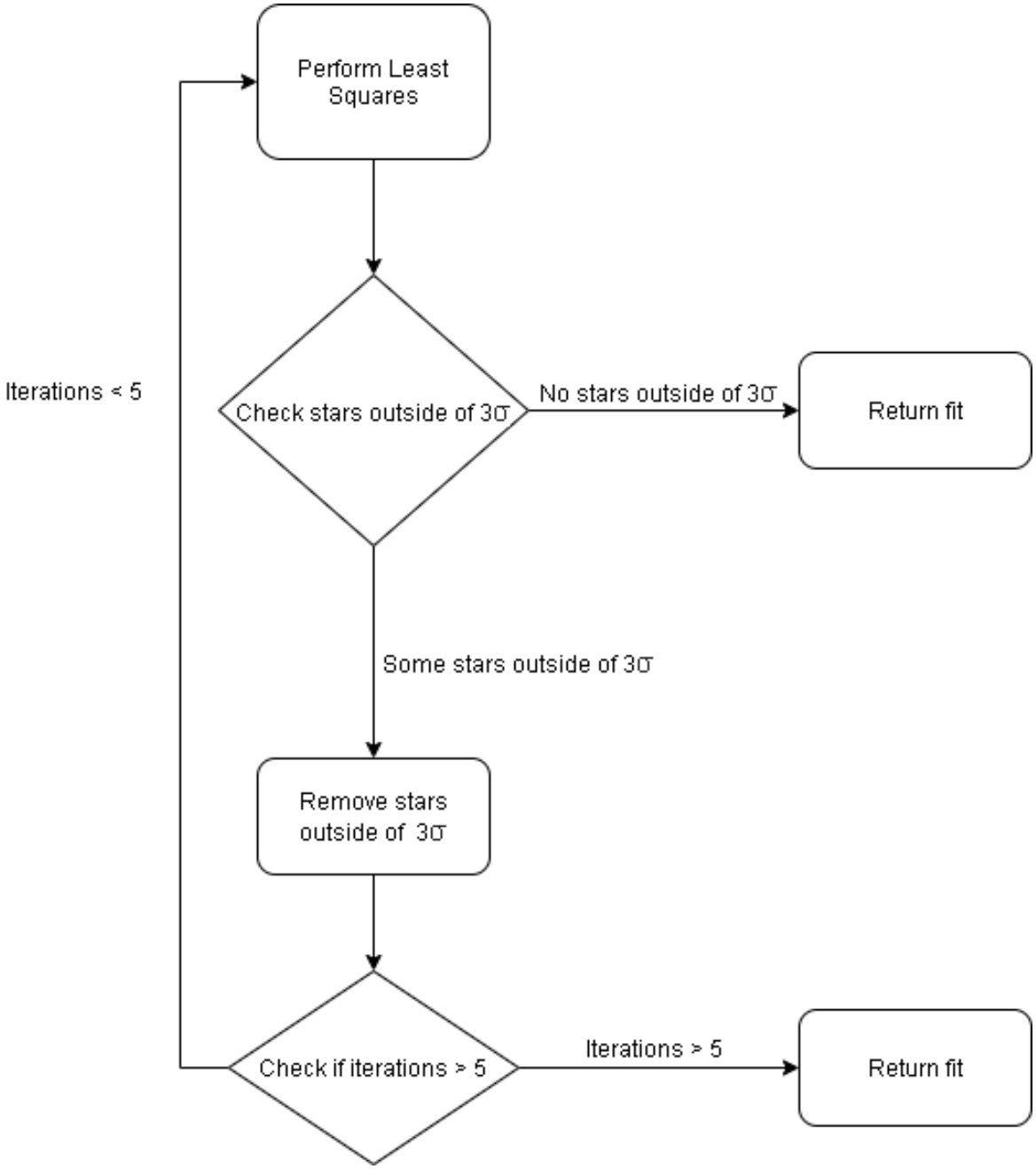

Figure 32: Flow chart describing how the sigma clipping process works. After each iteration of using least-squares regression to fit the offset, any stars outside 3σ of the distribution offsets is removed. The process is repeated until no stars are found outside of three sigma (or after five iterations).

5.5 Errors

The errors that will be attributed to the new corrected magnitudes are calculated as the root mean square (RMS) of all the offsets after the photometric correction within ±0.3 mag of each star.

Firstly a list of magnitudes is generated starting at the lowest magnitude calibration star in the current image and ending at the highest magnitude calibration star in the image. The magnitudes in the array are in steps of 0.1 mag For each magnitude in the array, all calibration stars within ±0.3 mag are found. Then the RMS of the offset after correction for the calibrations stars. A plot of this can be seen in Fig. 33 as the red line.

However, it can be seen that the population of stars decreases as a function of magnitude where we have less bright stars. This can cause the calculations of RMS to occasionally erroneously increase. Any increase in the RMS at brighter magnitudes is not likely to reflect a larger error and are more likely to be due to a lower amount of stars. As we have fewer stars at a brighter magnitude, it would only take a small number of unusually large offsets to substantially skew the RMS calculations with this method. Figure 33 shows a calculation of the RMS at ≈ 13 mag that is not reflective of the true RMS. It is likely that the two points that have been circled in red are the cause for this, however as they are still within 3σ, RMS they have not been removed.

Hence in order to correct for this, we will fit an equation to the RMS using a similar method to that of the photometric offset fit. The equation required must be able to model an increase in the RMS as a function of magnitude while also being a bounded function, as an RMS of ∞ or 0 is nonsensical. A sigmoid function is appropriate for this. Hence a least-squares regression program was provided with Eq. 12 where E_n are the free parameters that the least-squares regression program can vary and 'm' is the magnitude.

$$Fit = \frac{E_1}{1 + E_2 e^{E_3 m}} + E_4 \tag{12}$$

The blue line seen in Fig. 33 is the result of the fitted sigmoid.

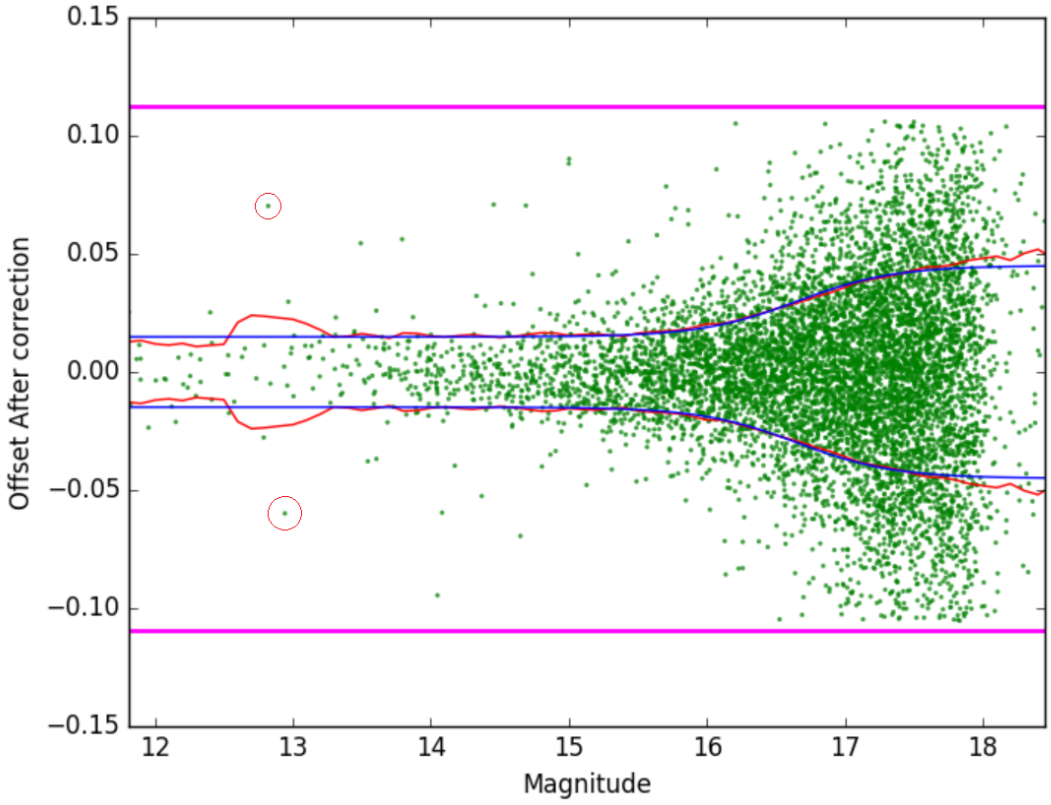

Figure 33: A plot of magnitude vs offset. The magenta line shows 3× RMS. The red line shows the RMS calculated for all points ±0.3 mag in steps of 0.1 mag The blue line shows a sigmoid fitted to the red line by the same way the offset was fitted.

After the correction is applied the errors assigned to the corrected magnitude `Mag_Corr_Error` are generated by taking the RMS as calculated by the sigmoid at each stars corrected magnitude `Mag_Corr`.

The distribution of the error associated with the corrected magnitudes in our data-set can be seen in Fig. 34. Here we can see a 2D histogram showing the error as a function of magnitude. We can see that there is a strong correlation between the magnitude and error of the magnitude. In Fig. 33 we can see how that correlation takes place, where fainter stars with a lower signal-to-noise have a higher error associated with their magnitudes.

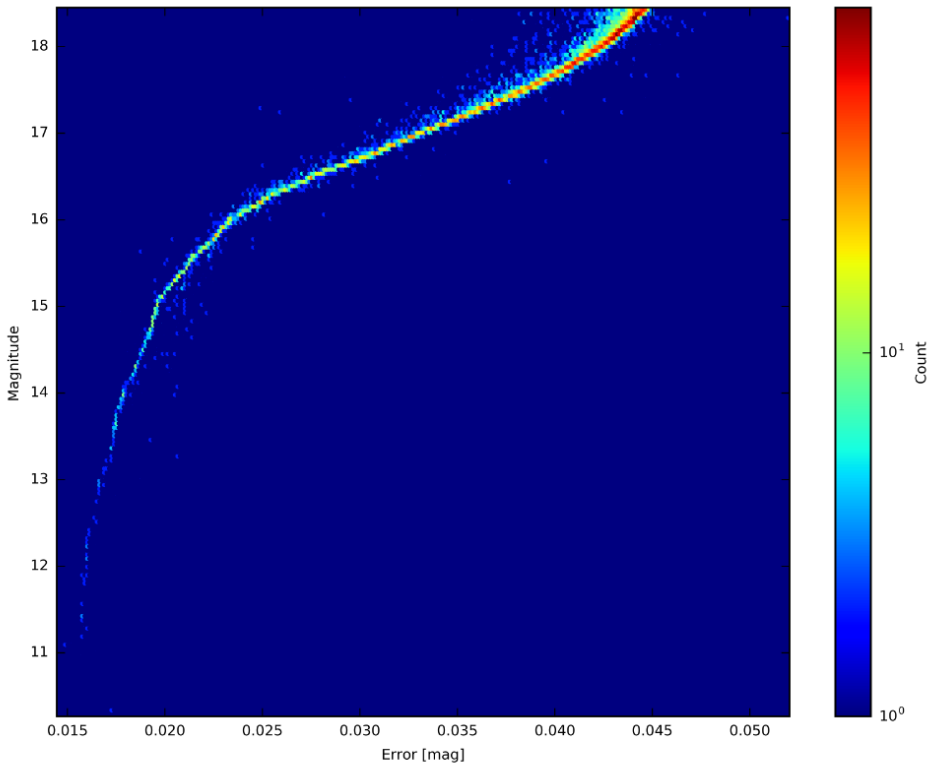

Figure 34: Showing the distribution of errors associated with the photometrically corrected magnitudes as a function of the corrected magnitudes. It can be seen that there is a strong correlation between a stars corrected magnitude and the error for that magnitude.

5.6 Applying correction

After the fit has been generated it can be stored as a series of parameters, these are each of the 'P_N' values seen in Eq. 9. Now we have an equation which models the photometric offset as a function of magnitude, colour and position on the CCD for a given image. We can now apply this equation to the data-points of every star in that image and not just for the calibration stars data-points. This process will remove all the photometric offsets that can be modelled by the use of a multi-variable N^{th} order polynomial. This process will not remove any small scale structure present in the images, nor will it

remove any of the stars apparent in the sky flats (described in Sect. 2.2.5).

Figures 35 and 36 show the before and after offsets as a function of magnitude, colour and CCD position. Figures 35 and 36 show how the fitted red line is subtracted from the offset thus removing any offset as a function of magnitude, colour or CCD position. Figure 35 also highlights the split between main sequence and AGB stars that are present in this dataset.

Figure 37 shows the same seven stars seen in Fig. 28 after the photometric correction process. This figure uses `Mag_Corr` and `Mag_Corr_Error` as seen in Fig. 8. It can be seen that the photometric correction significantly reduced the offset observed in Fig. 28. It can also be seen that the errors associated with these magnitudes have also decreased. It should be noted that some points are still outliers. Any data points with sufficiently large deviations can fail to be captured by the correction discussed above.

Figure 38 shows the light curve for the 14 instrumental magnitude star seen in Fig. 28 and 37 for both before (top) and after (bottom) the photometric correction. The scale in both has been kept the same to allow for easier comparison. It can be seen that the photometric offset has been largely removed in the light curve after the photometric correction. The errors are also smaller in the corrected light curve.

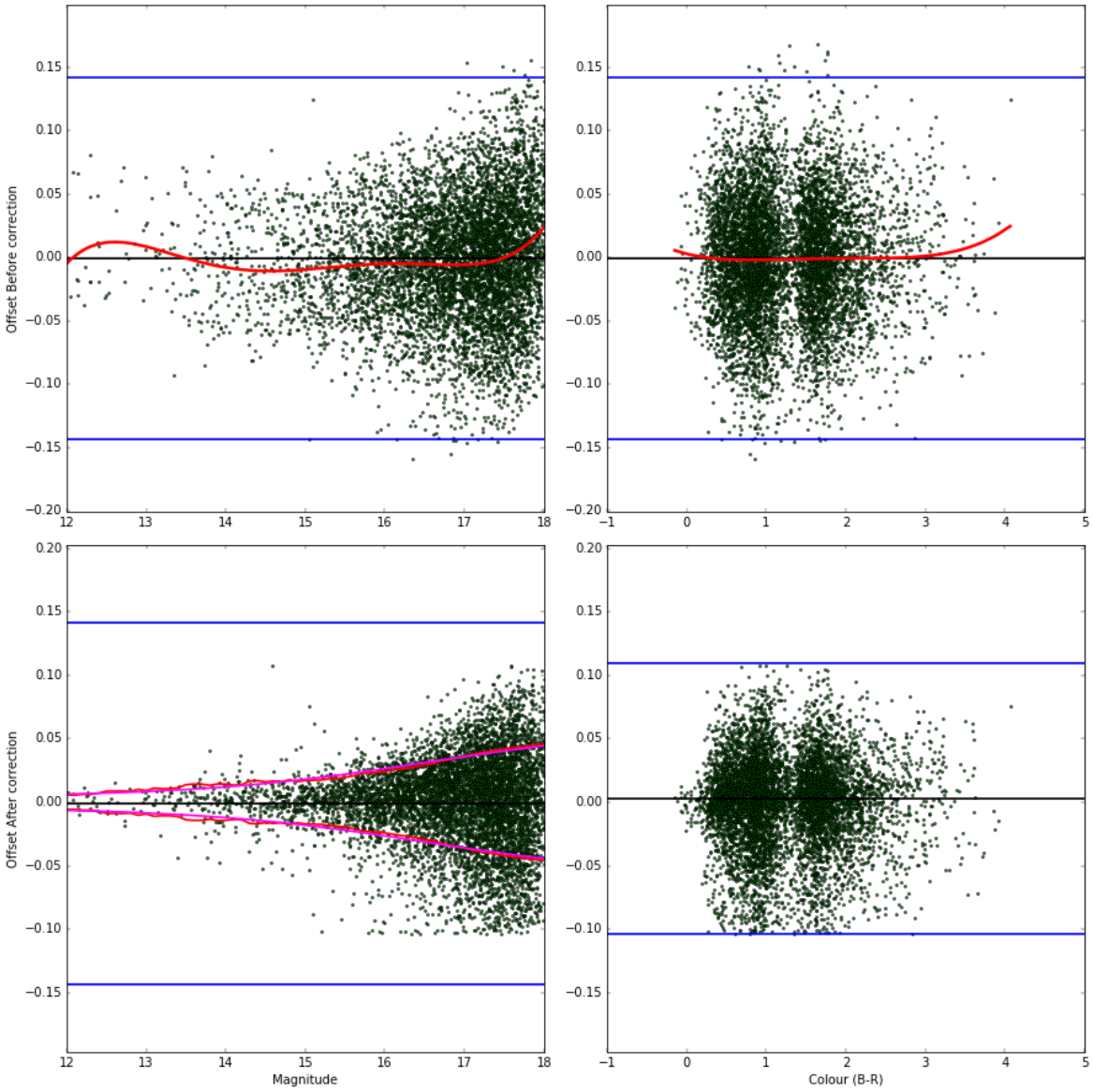

Figure 35: A scatter plot of offset versus either magnitude or colour. The blue lines represent $\pm 3\sigma$ of the distribution of offsets. Top: Shows the offset before the correction where the red line is the fit determined as a function of either magnitude or colour (as explained in Fig. 31). Bottom: Shows the offset after the correction procedure. The red line represents the RMS of the offset and the magenta line represents a sigmoid fitted to the RMS of the offsets.

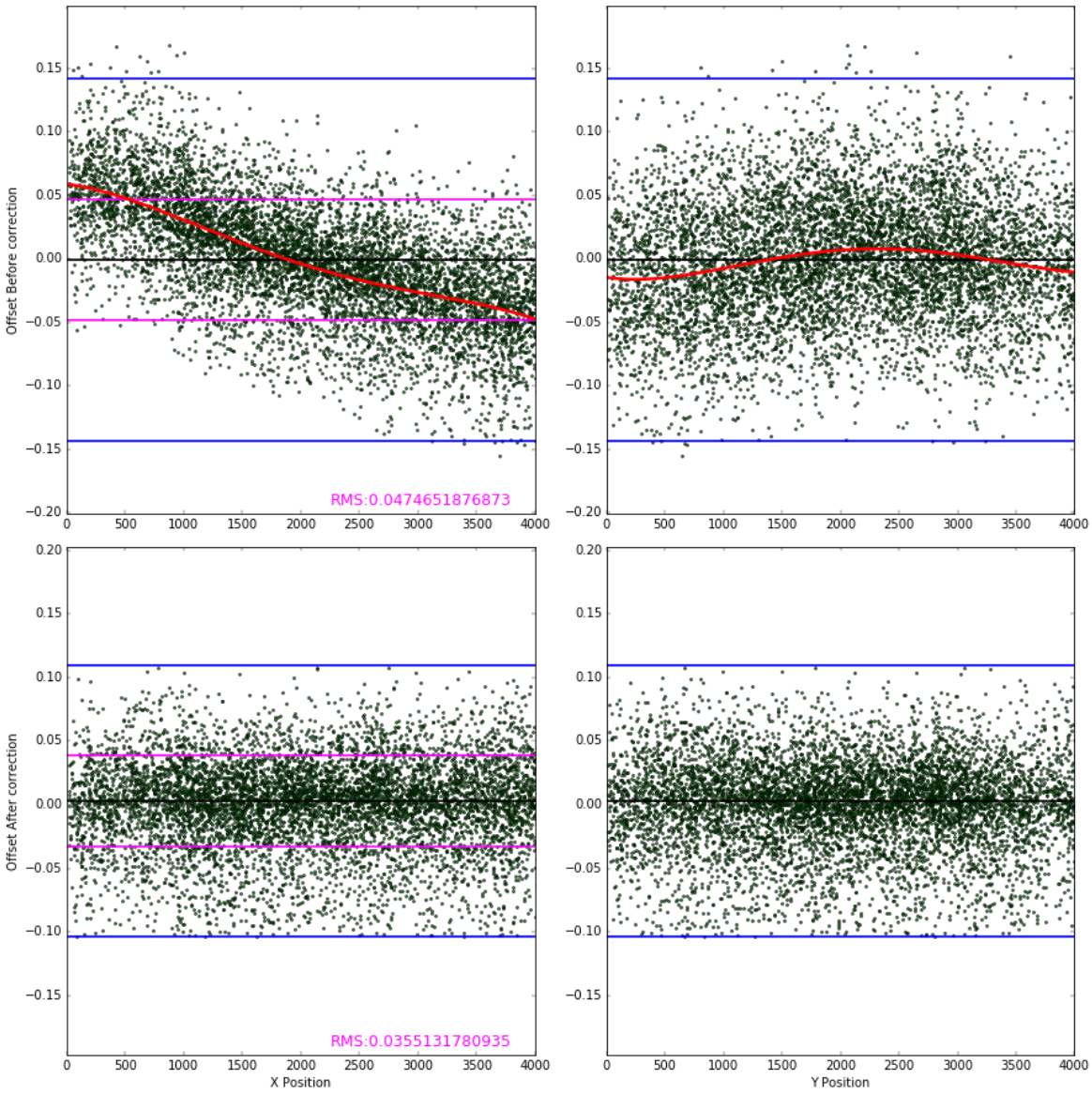

Figure 36: A scatter plot of offset versus position on the CCD in 'X' and 'Y'. The blue lines represent $\pm 3\sigma$ of the distribution of offsets. Top: Shows the offset before the correction where the red line is the fit determined as a function of either magnitude or colour (as explained in Fig. 31). Bottom: Shows the offset after the correction procedure.

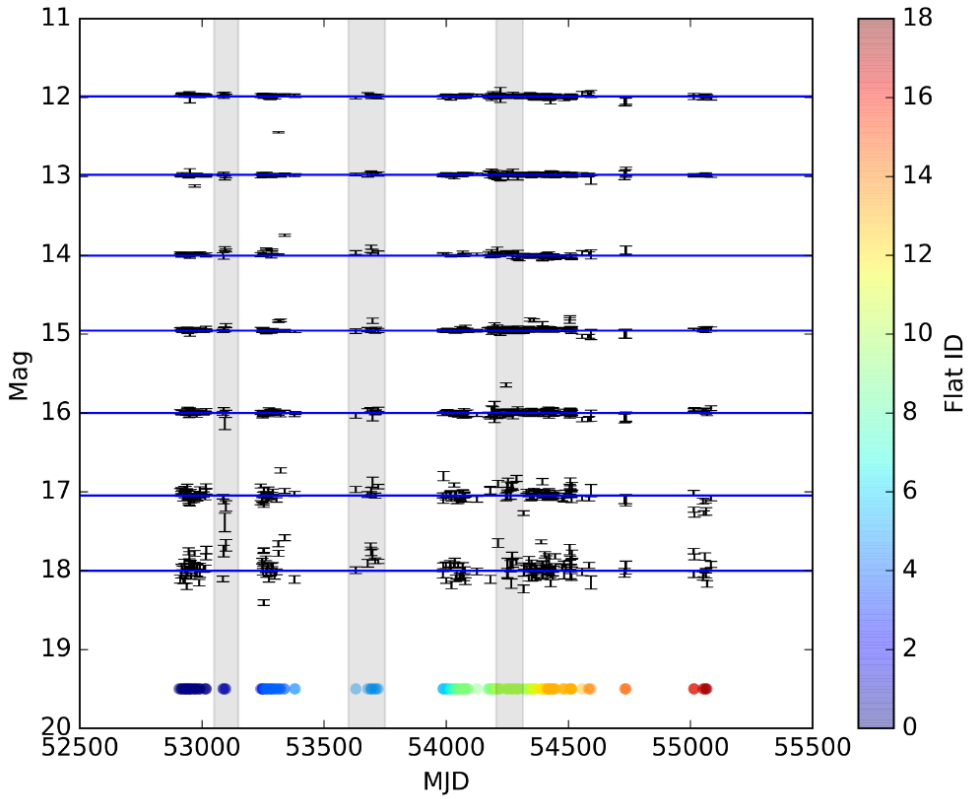

Figure 37: A light curve featuring the same seven stars seen in Fig. 28. Table 4 has been provided to show relevant information pertaining to each of the stars shown.

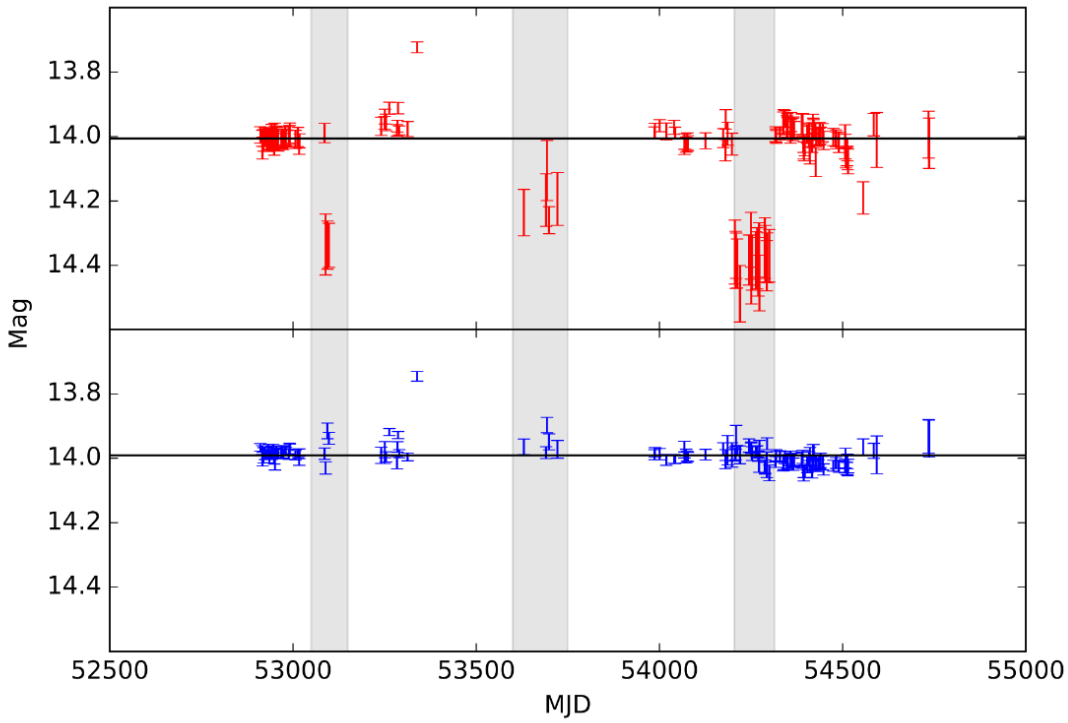

Figure 38: A light curve showing the before and after the photometric correction for the 14 magnitude star featured in Fig. 28 and 37. Top: Showing the light curve prior to the correction, the black line represents the median magnitude for all the points excluding any points highlighted by the grey bars. Bottom: Showing the light curve after the correction, the black line represents the median magnitude for all the points including the points highlighted by the grey bars.

6 Potential Science Projects

After the completion of the photometric correction, we can start to investigate some features of the stars in our catalogue. Below we discuss a preliminary investigation into the classification of stars in our catalogue via a Hertzsprung–Russell diagram as-well-as using GAIA, 2MASS and WISE colours to further classify the stars.

We also discuss investigations into periodic variability and show some of the tools that can be used to identify extrinsic and intrinsic periodic variable stars. Mr Waterman's original goal for the 'The Cygnus Project' was to detect exoplanets, hence, we discuss how it may be possible to detect exoplanets with the signal-to-noise of our data and how we may improve reduce the temporal resolution to increase the signal-to-noise.

6.1 Stellar Classifications

6.1.1 Hertzsprung–Russell diagram

We can use Eq. 13 (Gaia Collaboration et al., 2018b) to generate absolute magnitudes from GAIA's 'G' filter measurements and GAIA's parallax measurement's.

$$\mathrm{M}_G = \mathrm{G} + 5 + 5log_{10}(p/1000) \tag{13}$$

Where M_G is the absolute magnitude calculated, 'G' is the apparent magnitude as measured with GAIA's G filter `Gaia_Gmag` and 'p' is the parallax in mas measured by GAIA. Figure 39 shows a heat map Hertzsprung–Russell diagram of all the stars in our catalogue with a GAIA measurement.

Figure 39 was generated by comparing the absolute magnitude 'M_G' obtained via Eq. 13 to GAIA colour G-G_{RP}. All parallax measurements greater than 0 were included in the calculation of the absolute magnitude. The zero point correction provided by (Leung & Bovy, 2019) was applied to all parallax. Stars whose absolute magnitudes were calculated with parallax measurements that met the quality cuts described above were over-plotted with a green marker. The red arrow shows the reddening vector taken from (Kuhn et al., 2020) for a pre-main sequence star with an unreddened colour of G-$\mathrm{G}_{RP} = 1.3$ with $\mathrm{A}_V = 2$.

Figure 39 shows that our catalogue is populated by main sequence 'F' and 'G' type stars and giant 'K' type stars, also known as FGK stars. It can be seen that almost the entirety of the stars with more

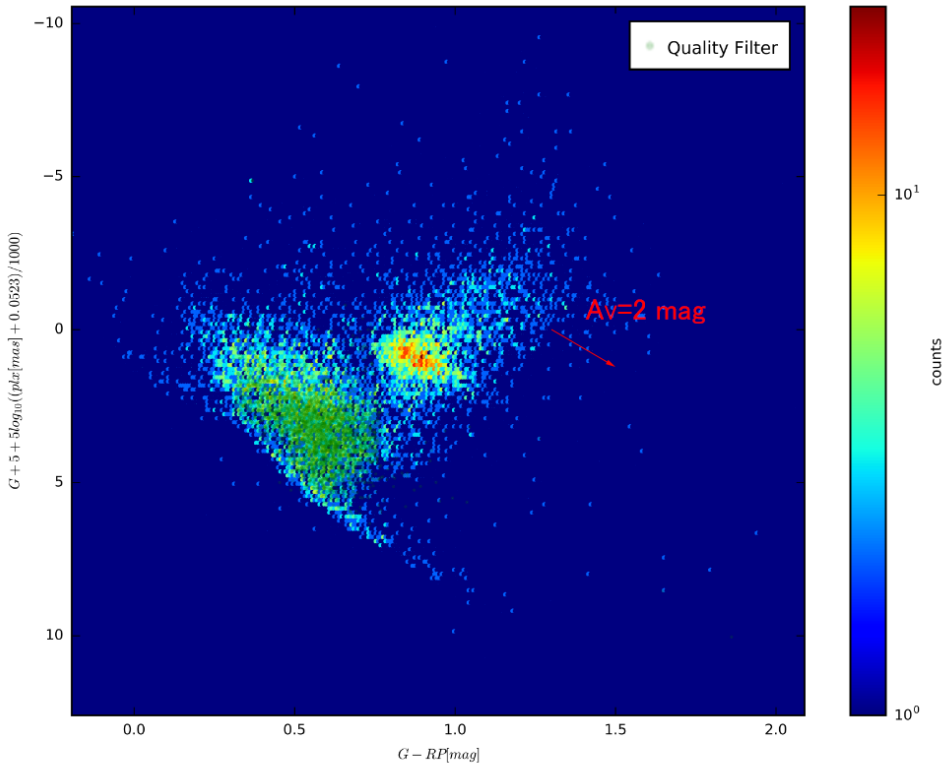

Figure 39: Hertzsprung–Russell diagram of all the stars in our catalogue with a GAIA measurement. The green points show absolute magnitudes calcualted with a GAIA parallax that meets the quality cut described in Sect. 4. The extinction vector is taken from (Kuhn et al., 2020) for a pre-main sequence star with an unreddened colour of G-G_{RP} = 1.3 with A_V = 2.

trustworthy parallax (described in Sect. 4) are on the main sequence. Virtually none of the quality cut parallax measurements encompass the giant stars, however, giant stars are likely to be further away with a high apparent magnitude and redder colour.

6.1.2 Colour classification

In Koenig & Leisawitz (2014) a method of identifying young stellar objects (YSOs) is provided. The method also shows how we can distinguish between YSOs and asymptotic giant branch stars (AGBs). The area discussed in Koenig & Leisawitz (2014) is towards the outer galaxy and is focused on high galactic latitude fields $|b| > 30°$. Hence Koenig & Leisawitz (2014) also include a filter to identify and remove active galactic nuclei (AGNs) and star-forming galaxies (SFGs). As the depth of our catalogue does not exceed the size of the milky way in the direction we are looking, this classification was disregarded as the area of our field is as such a low galactic latitude that we should not expect any extragalactic sources in our catalogue.

Figure 40 shows the Hertzsprung–Russell diagram shown in Fig. 39 where stellar classifications for YSO and AGB stars are shown in white and magenta respectively. A total of 8622 AGB stars and 13 YSOs have been identified by this method. The small amount of YSOs suggests that there is not much physical association between the field stars in our catalogue and the neighbouring star-forming region NGC7000 (The North American Nebula). It can be seen that the majority of the AGB stars have been classified into their correct position on the red giant branch of the Hertzsprung–Russell diagram, however, it can be seen that a substantial number of main-sequence stars have been miss-classified as AGB stars.

Koenig & Leisawitz (2014) provides a method for distinguishing between YSOs and AGBs (which is often necessary due to both having an IR excess) and hence, it is not fully appropriate to classify AGB stars in our catalogue with this method as we are also considering main sequence stars. It is also possible that this method of classifying stars is not fully appropriate given the difference in galactic positions we are looking in. It is also mentioned in Koenig & Leisawitz (2014) that they specifically investigated star-forming regions. Our catalogue consists of field stars and there does not appear to be any star-forming region within our catalogue.

The distinction between the main sequence stars and AGB stars was made by comparing colour and absolute magnitude. We used Fig 40 to generate Eq. 14 which allowed us to determine an approximate value for how many stars are likely main sequence or AGB. From this we were able to calculate that

of the 19,858 stars we have, ≈13,000 are considered to be main sequence while the rest (≈ 7000) are AGB stars.

$$\text{MS: Absolute Magnitude} > 15(\text{G-RP}) - 10.5 \tag{14}$$

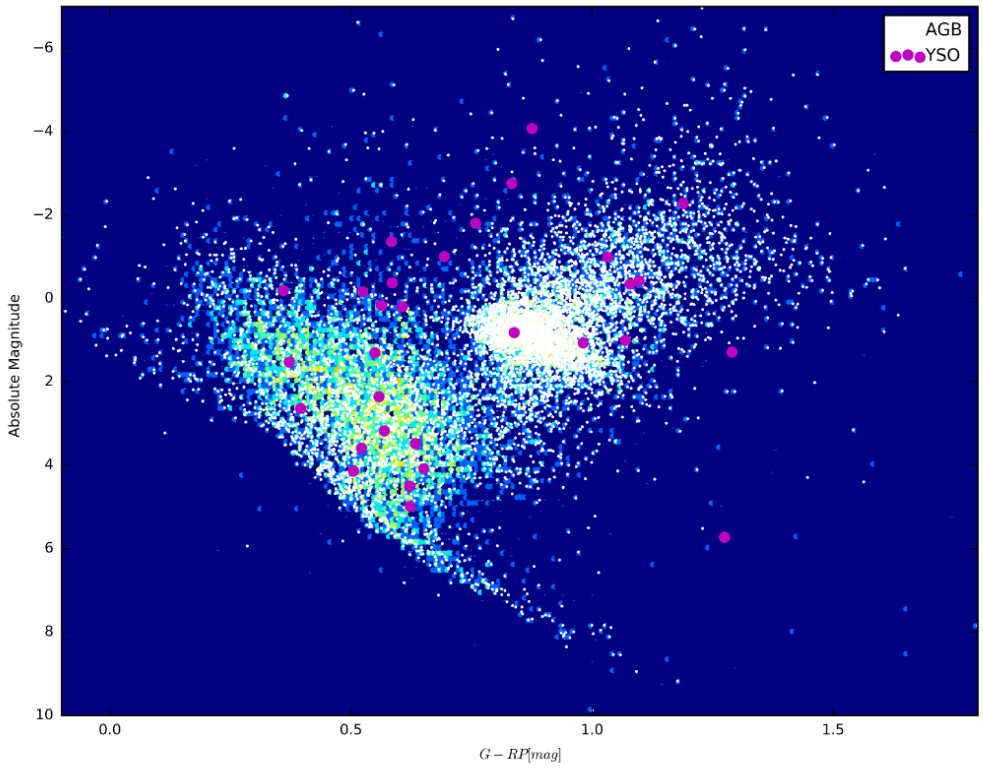

Figure 40: The same Hertzsprung–Russell diagram from 39. The YSOs are highlighted in magenta and the AGB stars are highlighted in white. The AGB points have been made smaller than the YSO points due to them being substantially more numerous.

6.2 Microlensing

Due to the large number of sources present, it is worth investigating whether we can identify microlensing events. We can use the value quoted in Wozniak (2000) to approximate how many microlensing events we could have. In Wozniak (2000) it is noted that difference image analysis was used with red giant stars to gain a value of $\tau_{OGLE} = 2.55^{+0.57}_{-0.46} \times 10^{-6}$ Where 'τ_{OGLE}' is the probability of seeing a microlensing event. When we compare this to our data we see that it is unlikely that we will find an event in all of our 19858 sources. In Wood (2007) multiple values for the probability of seeing a microlensing event are shown. It is noted that these values change significantly depending on the survey source and the method of identification. Hence, without undertaking further significant study of this data, it appears we can not reliably predict the frequency of microlensing events.

6.3 Periodic Variables

In order to search for periodic variables we can use the `Lomb-Scargle Periodogram` (Lomb, 1976) (Scargle, 1982). The Lomb-Scargle periodogram is a statistical tool that can detect periodic variations in unevenly spaced data. We can use the `astropy` (Astropy Collaboration et al., 2013) Lomb-Scargle package to search for periodic variable stars in our catalogue. We ran the Lomb-Scargle program over some of the stars in our catalogue, which are discussed below. For each star, the Lomb-Scargle program produced an output of 'Power' (which has a range of 0-1) as a function of frequency (seen as the top left plot in Fig. 41). We set a threshold where any star with peak > 0.4 in power will be considered a periodic variable. If a star was found to be variable the largest peak in power given by the Lomb-Scargle program was considered to be the period at which the star varied. Once the period of the variable star is found the light curve is folded in accordance with the period.

Table 5 shows a list of some variable stars found in our catalogue. These stars were found by manual investigation of their folded light curves after the Lomb-Scargle process was run. All of the 'types' of the stars were given by GAIA (Gaia Collaboration, 2018) with the exception of star V* V1898 Cyg whose classification was provided by 2MASS and TYC 3588-196-1 which has no official classification. The period given for each star was calculated using the Lomb-Scargle method described above. The absolute magnitude was calculated using Eq. 13.

Table 5: A table showing information about three of the periodic variable stars found within our catalogue.

Name	**Type**	**Object_ID**	**Period**	**Amplitude**	**Coords [J2000]**	**MAG_AVG**	**Absolue Magnitude 'M$_G$'**	**B-R**	**Parallax [mas]**	**Stetson Index**
V* V356 Cyg	Delta Ceph	22164	5.05721	0.526	316.944 +46.737	16.2354	-3.289	1.782	0.141	1.023
Gaia DR2 2162640763705905920	Delta Ceph	1133	3.56902	0.478	316.089 +45.239	17.938	-5.561	1.878	0.066	0.862
V* V2578	Delta Ceph	28710	1.76632	0.182	317.786 +47.168	14.477	-2.486	1.263	0.343	0.084
NSVS 5840174	Eclipsing Binary	18028	1.06857	0.606	315.547 +46.447	17.611	-1.046	1.065	0.178	1.085
V* V1898 Cyg	Algol Type	16282	1.51311	0.208	315.974 +46.331	12.221	-2.329	0.005	1.023	0.278
TYC 3588-196-1	W UMa Type	22053	1.12305	0.106	315.501 +46.697	15.771	-0.175	0.361	0.544	0.067

6.3.1 Delta Cepheid Variables

Delta Cepheids are a type of population I variable stars. Delta Cepheids have a mass of $4-20\mathrm{M}_{\odot}$ and a radius of $10-1000\mathrm{R}_{\odot}$. Spectroscopically, Delta Cepheids range from F-type to G-Type stars which evolved from B-type main sequence stars (Turner, 1996).

Delta Cepheid variables have a distinctly shaped triangular light curve. The three Delta Cepheid variables shown in table 5 were identified by manual inspection of the shape of each light curve, followed by checking if the period and amplitude is indicative of a Delta Cepheid. We expect the period of a Delta Cepheid to be on the order of a couple of days and the peak to peak amplitude of the light curve to be on the order of three quarters of a magnitude. Figure 41 shows the Lomb-Scargle, light curve and folded light curve for one of the Delta Cepheids found in our catalogue (V* V356 Cyg). The Lomb-Scargle program returned a period of 5.05683. However through manual inspection of the phase folded light curve we determined the period to be closer to 5.05721 d, as this period produces a phase folded plot with the least scatter. Figure 46 shows the phase folded light curve of this star when using the Lomb-Scargle given period. For this and every subsequent phase plot we have also plotted a running median and standard deviation which can be seen by the blue line and green filled area respectively.

Delta Cepheids also have a well defined Period-Luminosity relationship. Equation 15 (Benedict et al., 2002) shows how we can determine the absolute magnitude 'M_V' of a Delta Cepheid if we know its period 'P' in days.

$$M_V = (-2.43 \pm 0.12) log_{10}(P-1) - (4.05 \pm 0.2) \tag{15}$$

From Eq. 15 we can calculate that the absolute magnitude 'M_V' for star V* V356 Cyg as -5.528 ± 0.273, the M_V for star Gaia DR2 2162640763705905920 as -5.05 ± 0.151 and the M_V for star V* V2578 as -3.77 ± 0.214. While our calculated M_V for star Gaia DR2 2162640763705905920 is within error for the 'M_G' calculated with Eq. 13, the other two stars are not. The M_G for star V* V356 Cyg is -3.29 and the M_G for star V* V2578 as -2.49. The likely source of this error is the parallax used to calculate M_G. We discuss the issues and errors associated with the GAIA parallax in Sec. 4.5.1.

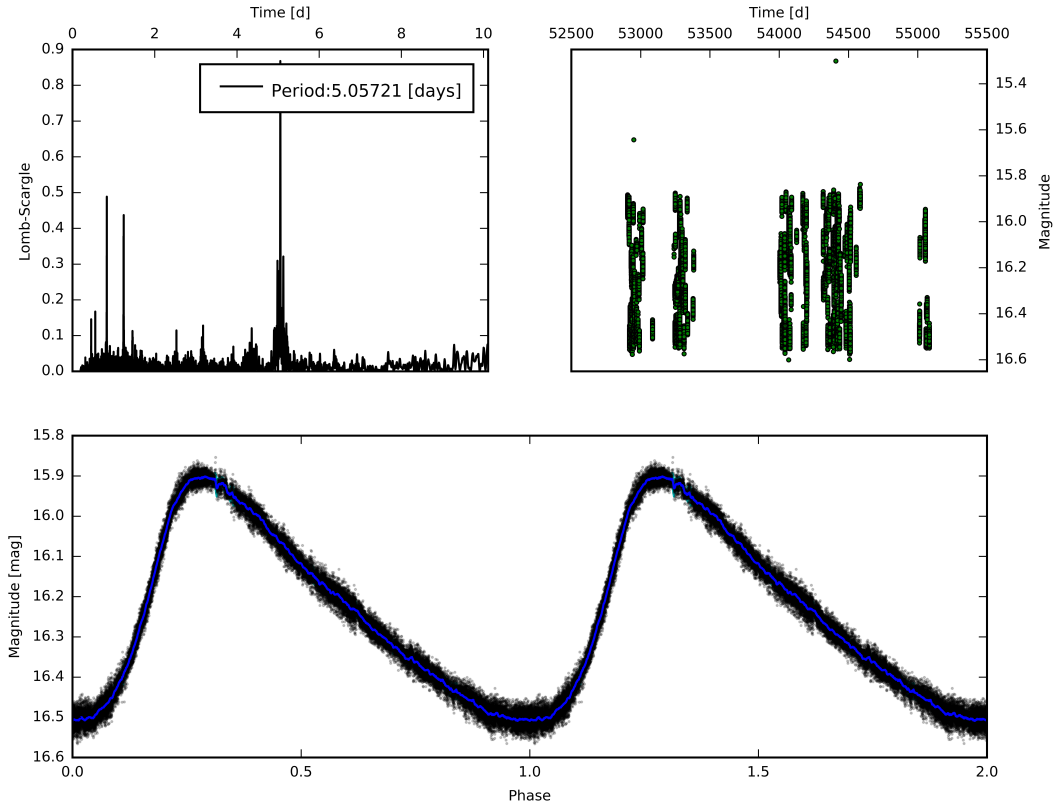

Figure 41: Shows the light curve of a Delta Cepheid variable found in our catalogue (V* V356 Cyg). Top Left: The output of Lomb-Scargle showing frequency vs power plot. Here the period is shown to be at 5.05721 d (whereas the Lomb-Scargle reported period is 5.05683 d). Top Right: Showing the unfolded light curve of this star. Bottom: Shows the folded light curve of the star.

6.3.2 Eclipsing Binaries

Several Eclipsing Binary stars were found by manual inspection of folded light curves. Figure 42 shows the light curve, folded light curve and Lomb-Scargle of an eclipsing binary (NSVS 5840174). For this eclipsing binary, we found a period of 1.06857 d, which is double that given by the Lomb-Scargle (0.53428 d). In Hoffman et al. (2009) it is stated that this star has a period of 2.29623 d. However, when this star is folded with that period we do not obtain a light curve indicative of a correct period (see Fig. 47 and Fig. 48).

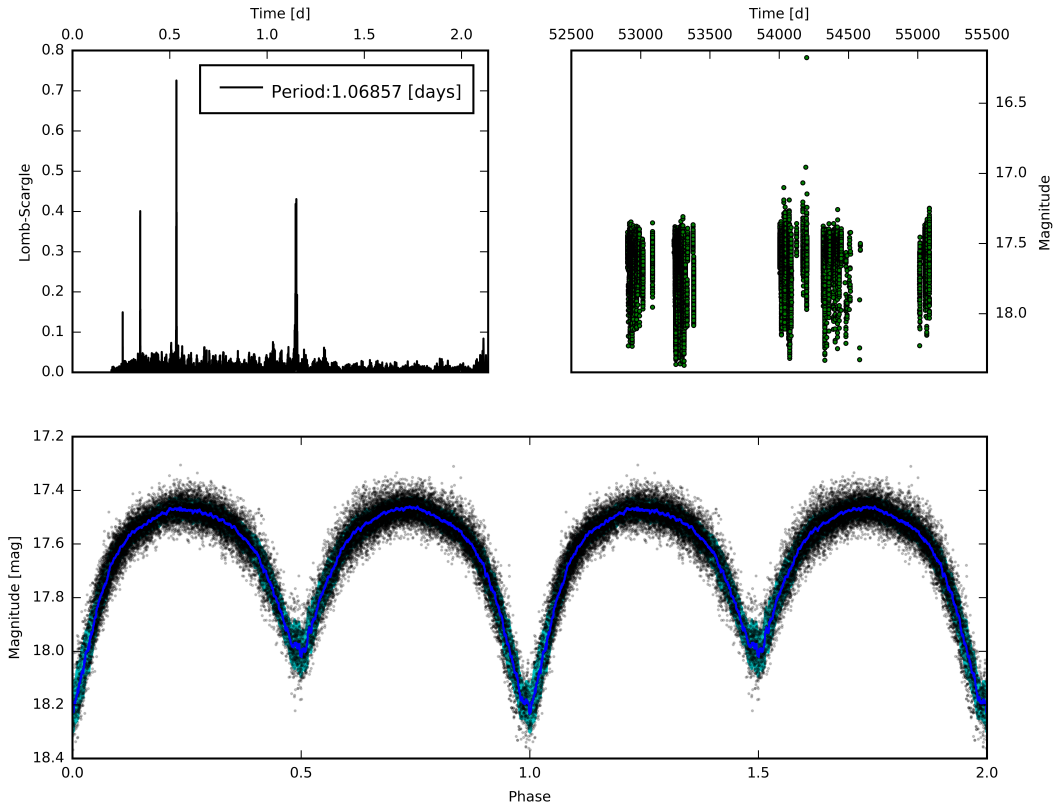

Figure 42: Shows the light curve of an eclipsing binary found in our catalogue (NSVS 5840174). Top Left: The output of Lomb-Scargle showing frequency vs power plot. Here it is shown that the period is is at 1.06857 d (whereas the Lomb-Scargle reported period is 0.53428 d). Top Right: Showing the unfolded light curve of this star. Bottom: Shows the folded light curve of the star.

The Algol type eclipsing binary is a semi-detached system of stars where the less massive star transfers mass to the more massive star. The database we have has a temporal resolution of ≈ 1 minutes. This will allow study into the long term variations in the stars period caused by the change in angular momentum of the more massive star as it gains mass. Figure 43 shows the light curve, folded light curve and Lomb-Scargle of an Algol variable star (V* V1898 Cyg). For this star, the Lomb-Scargle gave us a period of 1.51317 d. However, through manual inspection of the light curve (similar to V* V356 Cyg), we found the actual period to be 1.51311 d. Figure 49 shows the phase folded light curve of this star when using the Lomb-Scargle given period.

W UMas are a type of eclipsing binary variable star. W UMa stars are close binary stars of spectral type F, G or K (FGK stars). W UMa stars share an envelope of material where at least one of the stars overfills its Roche lobe and transfers mass onto the other, hence these stars are known as contact binaries. Figure 44 shows the light curve, folded light curve and Lomb-Scargle of a W UMa variable star (TYC 3588-196-1). We suspect this W UMa star is a binary system where the two stars are of very similar mass. This is because the shape and size of the transits for each star are similar thus indicating they are of similar size and luminosity. The periodogram indicates a period of 0.56153 d. However, the true period is double that (1.12306 d) of the tallest spike, this is because this binary system is likely near-equal mass and so the transits appear similar. Figure 50 shows the phase folded light curve of this star when using the Lomb-Scargle given period.

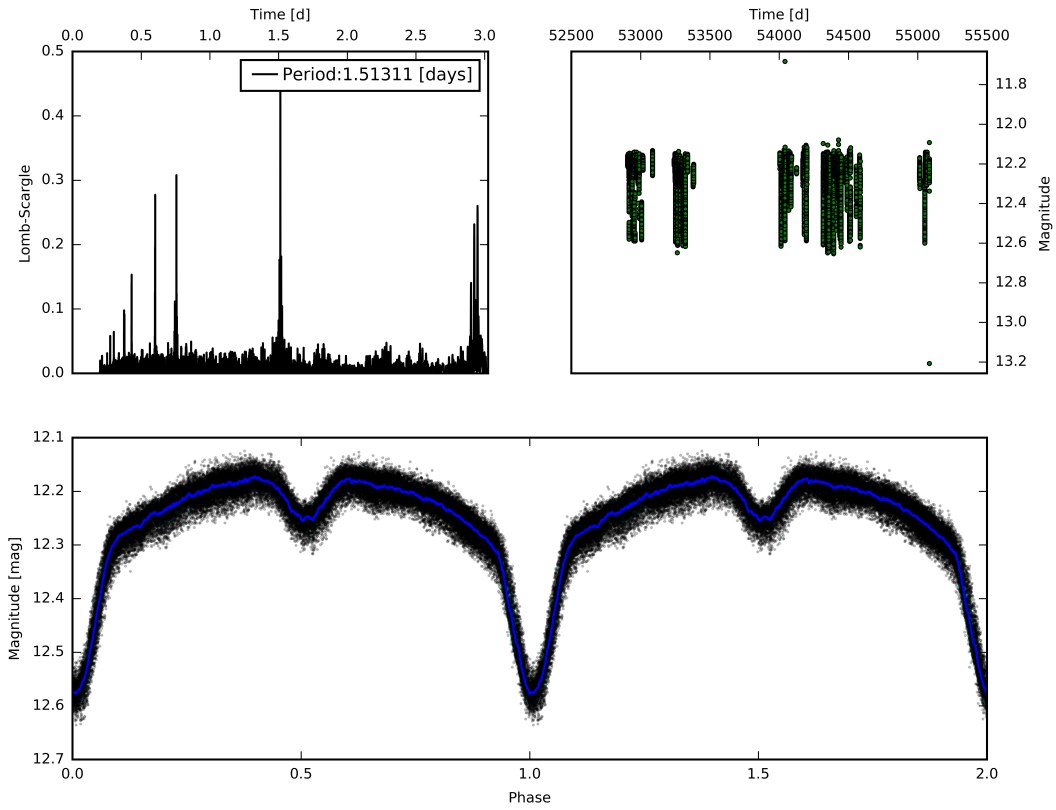

Figure 43: Shows the light curve of an Algol variable found in our catalogue (V* V1898 Cyg). Top Left: The output of Lomb-Scargle showing frequency vs power plot. Here it is shown that the period is is at 1.51311 d (whereas the Lomb-Scargle reported period is 1.51317 d). Top Right: Showing the unfolded light curve of this star. Bottom: Shows the folded light curve of the star.

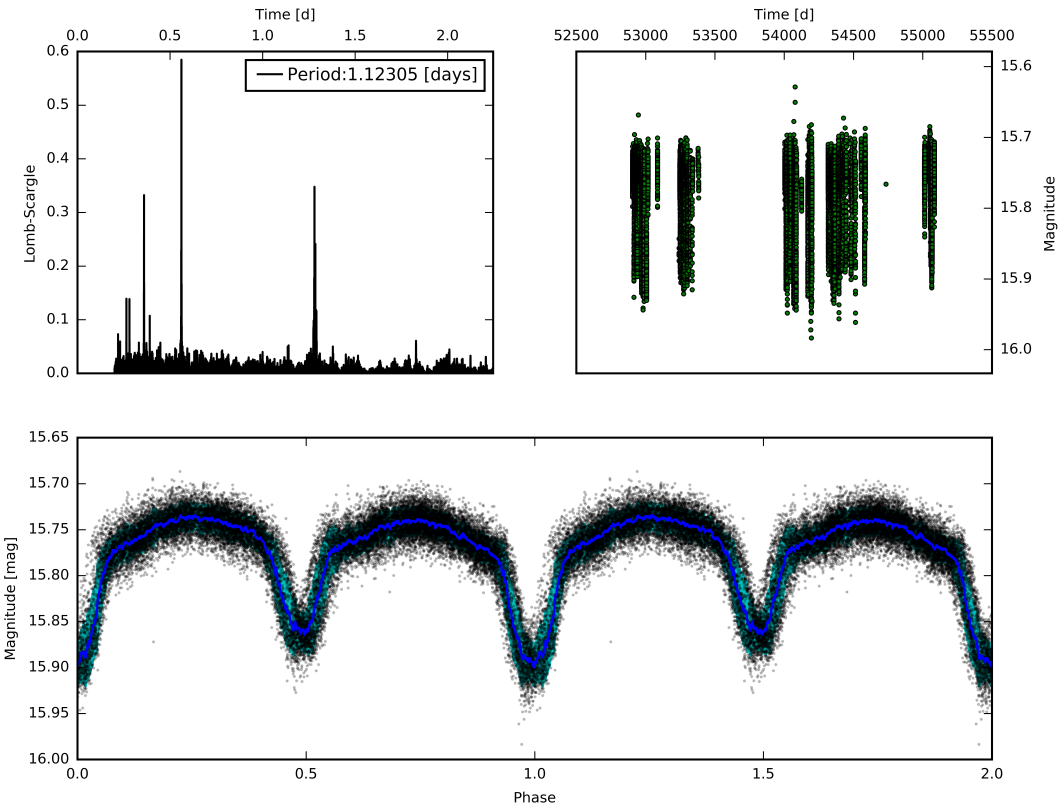

Figure 44: Shows the light curve of a W UMa variable found in our catalogue (TYC 3588-196-1). Top Left: The output of Lomb-Scargle showing frequency vs power plot. Here it is shown that the clearest spike is at 0.56153 d. However, the true period is double that (1.12306 d) of the tallest spike. Top Right: Showing the unfolded light curve of this star. Bottom: Shows the folded light curve of the star.

6.3.3 Planetary Transits

A planetary transit is where a planet orbiting a star crosses the line of sight between the observer and the star. This is detected as a decrease in the observed brightness of the star. The amount that the stars brightness decreases (the depth of the transit) can be determined as a function of the radius of the star and the radius of the planet, Eq. 16 describes the depth of the transit. Where '$\frac{\Delta F}{F}$' is the ratio of an observed change in flux to that of stellar flux, R_{Planet} and R_{Star} are the radii of the planet and star respectively.

$$\frac{\Delta F}{F} = \frac{R^2_{Planet}}{R^2_{Star}} \tag{16}$$

To reliably detect a planetary transit we must ensure that the difference in magnitude from the transit $\frac{\Delta F}{F}$ is sufficiently larger than the error associated with the magnitude `Mag_Error`.

The rate at which a planetary transit occurs will govern how likely we are to detect it. The probability of detecting a planetary transit is determined by Eq. 17. Where R_{Star} and R_{Planet} is the radius of the planet and the star respectively. 'a' is the semi-major axis of the planets orbit with respect to the star.

$$P_{Transit} = \frac{R_{Star} + R_{Planet}}{a} \tag{17}$$

Thus, to increase the probability of being able to detect a planetary transit we can preferably search for a planet with a large radius and a small semi-major axis. Such planets are known as 'hot Jupiters'. A 'hot Jupiter' is a gas giant planet with an orbital period of fewer than 10 d, and so they usually have a semi-major axis of less than 0.1 AU. In Marcy et al. (2005) it is stated that the occurrence of hot Jupiter's within 0.1 AU is $1.2 \pm 0.2\%$ around FGK stars.

From Eq. 17 we can calculate that there is a $\approx$ 5.1% chance of finding a hot Jupiter where R_{Star} =R_{Sun} and R_{Planet} =$\mathrm{R}_{jupiter}$ and a= 0.1 AU. We know that the catalogue after the completeness cut has 19,858 stars within it, of those 19,858 stars, $\approx$13,000 of them are main sequence FGK stars. As we know that the occurrence of hot Jupiter's within 0.1 AU is $1.2 \pm 0.2\%$ around FGK stars, this means we should expect $\approx$156 of those stars to host a hot Jupiter. However, this assumes that we are able to reliably do so with the signal-to-noise of our data.

From Eq. 16 we can calculate that a hot Jupiter orbiting a star where R_{Star} =R_{Sun} and R_{Planet} =$\mathrm{R}_{jupiter}$ will give us $\frac{\Delta F}{F} = 0.01$. In order to accurately measure a transit under Nyquist sampling, we will re-

quire the error to be half of $\frac{\Delta F}{F}$. The transit described above will need the error of the magnitude `Mag_Corr_Error` to be < 0.005mag. We can improve the signal-to-noise by combining 'N_{Image}' measurements taken after each other, creating a new standard error. The new standard error is smaller than the standard deviation by a factor of $1/\sqrt{N_{Image}}$. This will reduce the temporal resolution of the measurements in order to decrease the uncertainty of the photometric measurements.

We can investigate for how many 'N_{Image}' measurements do we need to combine for a star with a given average '`Mag_Corr_Error`' in order to improve the signal-to-noise to the point where we could reliably detect a hot Jupiter transit. Figure 45 shows how the magnitude error that would be usable increases as a function of $\sqrt{N_{Image}}$. For example, we could use a star with an average error of $\approx \pm 0.02$ mag if we bin the data into groups of 15, thus reducing the temporal resolution to 15 minutes but effectively decreasing the average error to $\approx \pm 0.005$ mag. The blue line represents how many stars we have in our database with an error less than $\sigma_{mag,N_{Image}} = 0.005 \times \sqrt{N_{Image}}$ for each 'N_{Image}' on the x-axis. We can see from the graph that we have 1653 stars with an error below $\sigma_{mag,N_{Image}} = 0.005 \times \sqrt{43} = 3.279$. Hence, if we reduce the temporal resolution of our data-set to 43 minutes, we can gain a large enough sample of stars from our catalogue to expect to be able to detect a hot Jupiter. This is done under the assumption that the transit duration of the hot Jupiter would be over double that of the temporal resolution to avoid Nyquist sampling. If one were to undertake an automated search for exoplanets using a method such as Box-least-squares, (Kovács et al., 2002) then special care would need to be taken to not remove real transit data via this binning method.

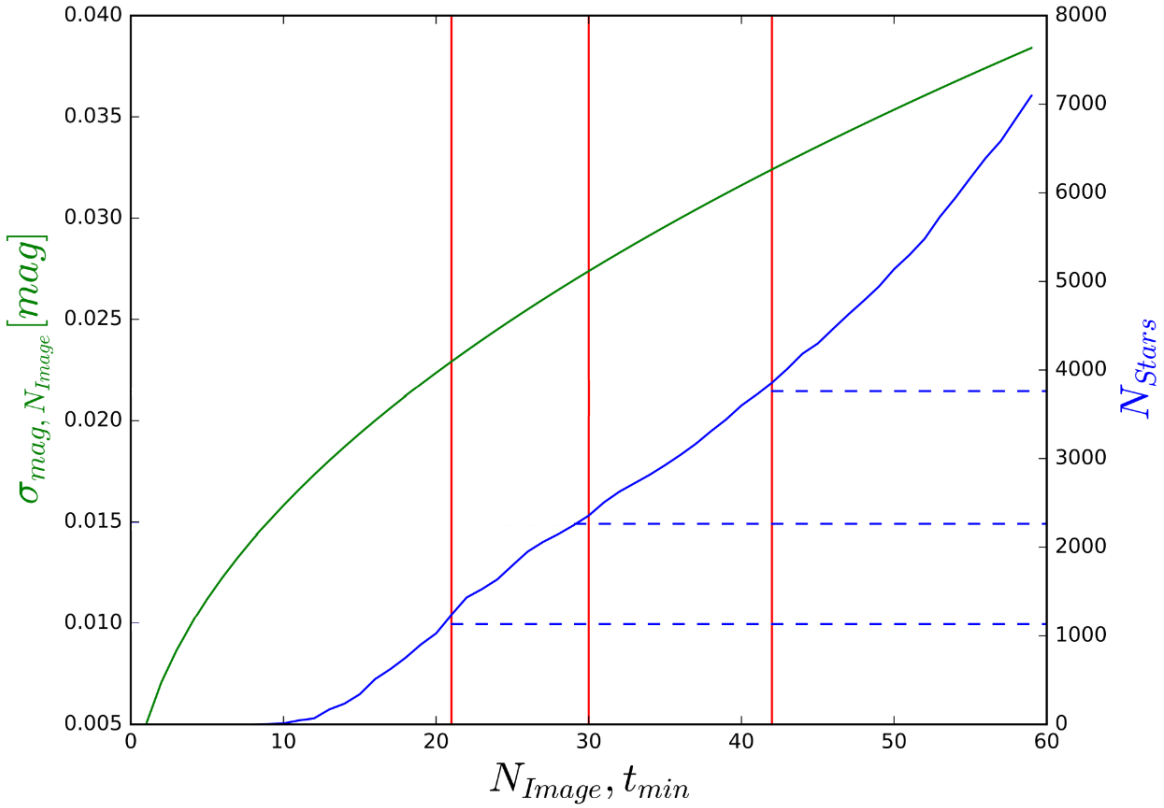

Figure 45: Where N_{Image} represents the number of data-points being combined (i.e the number of measurements being binned together). Which in our case, as we have an original cadence of 1 minute, $N_{Image} = t_{min}$ such that if we combine 5 measurements our cadence will be 5 minutes. $\sigma_{mag,N_{Image}}$ represents the maximum error allowed when N_{Image} are being binned together to give an uncertainty low enough to reliably detect a hot Jupiter transit. Therefore the green line represents a plot of $\sigma_{mag,N_{Image}} = 0.005 \times \sqrt{N_{Image}}$. N_{Stars} represents the number of stars where `Mag_Corr_Error` $= \sigma_{mag,N_{Image}}$. The blue line represents how many stars we have with an error low enough to detect a hot Jupiter transit when combining N_{Image} data-points together. The horizontal blue dashed lines are at $N_{Stars} = 1132, 2264$ and 3762 which corresponds to the number of stars required for us to have a 50%, 75% and 90% probability of detecting a hot Jupiter transit, respectively. The red vertical line shows at what value of 'N_{Image}' do we have $1132, 2264$ and 3762 stars with an error below $0.005 \times \sqrt{N_{Image}}$, these are at 21, 30 and 42 N_{Image}.

7 Summary

The data-set and subsequent database discussed in this thesis will provide useful insight into the study of many facets of time-domain astrophysics. We expect the data presented here to be useful for investigation into multiple types of periodic and aperiodic variables such as Binaries and Cepheid's (a cursory investigation is seen in Sect. 6.3). It will also possible to search for exoplanet transits when combining data-points to increase photometric accuracy. The database presented here is of a 1 minute cadence over 6 years and is 99% complete up to an instrumental r-band magnitude of 18.45 mag (GAIA R $\approx$ 15 mag). The high temporal resolution for this database is amongst its best features. This allows for much more accurate measurement of a stars variability whilst also allowing for the flexibility in exchanging temporal resolution for increased photometric accuracy.

In this thesis we have shown the data reduction and initial calibration pipeline (Sect. 2). We have shown how the data was initially reduced with the calibration frames and how each data-point was extracted from the FITS files. We have discussed the relatively low amounts of calibration frames (11 sets of bias frames, 15 sets of dark frames and 19 sets of flat frames with 64,293 science frames). Notably, we identified the issue with the flat frames that were used for data reduction and how identifying specifically which flat frames are of poor quality is not possible due to the nature of flat frames. The steps that were taken to initially calibrate the data after the data reduction are also shown. This calibration sought to increase the consistency of the data by comparing the magnitude of every star in a given image to their magnitudes in a master image that was taken under photometric conditions (no obvious colour or magnitude terms brought on by the atmosphere).

Due to the size of the data-set presented in this thesis, a database was necessary to allow for the querying of stars (Sect. 3). We have shown the construction and formatting of the database and the catalogue that subsequently formed from its construction (Referred to as `ID_DB`). We have discussed the contents and justification of each table inside the database. The range of available queries is highlighted via each of the 6 tables identifier (`Object_ID`, `Data_ID`, `Image_ID`, `Bias_ID`, `Dark_ID` and `Flat_ID`). An investigation into the distribution of the seeing across the data-set was used as a justification for how each star was identified across all of the images. Here we found that the distribution of coordinates that an individual star may have can be represented with a Gaussian distribution. Hence, we know that the astrometric uncertainty is dominated by random processes such as atmospheric seeing. From this, we decided to use $5\sigma_d = 2.185$ arcseconds as the matching radius for our star matching process

(where σ_d is the standard deviation of the distribution of coordinates for a given star).

The catalogue generated from this database was cross-matched with other publicly available catalogues (Sect. 4). We cross-matched with GAIA, 2MASS and WISE. GAIA's second data release was used and, from this, we were able to gain astrometric data to allow for the study of proper motions as well as parallax measurements and thus absolute magnitudes. The extra filter measurements from 2MASS and WISE will be used to supplement the otherwise single filter catalogue. While the data we obtained from GAIA, 2MASS and WISE is only a single measurement from each we can use the colours to further classify the stellar population in our database.

A substantial portion of the work presented in this thesis was spent designing and performing a post-data reduction calibration (Sect. 5). While there was an initial calibration process (see Sect. 2.2.4), this process assumed the calibration files where of adequate quality. After initially creating the database and producing the first light curves it was identified that there was occasionally significant photometric shifts for the stars (see Fig. 18). Further investigation showed that this offset was due to large inhomogeneities present in the flat fields used to reduce the data. These inhomogeneities are present for two main reasons: partly due to some of the flat frames being taken as 'sky flats' with sidereal tracking and hence some stars became present (see Fig. 5); partly as some flats were taken with an evenly illuminated perspex sheet (see Fig. 6). It is likely that the perspex sheet used was not manufactured with the intention of being used for flat field calibration, hence, the perspex sheet did not provide homogeneous illumination for flat field measurements.

In this photometric correction we compare the magnitude of all non-variable stars (see Sect. 5.3.1) in a given image to the average magnitude those stars have across all images. We then designed a correction program which fits a polynomial to the difference between the magnitudes from the image and average magnitude as a function of magnitude, colour and CCD position (see Eq. 9). Here we found that the largest offsets were CCD dependent, further confirming that the photometric offset was due to poor quality flat fields. This method is appropriate as it will remove the overarching structure present whilst not destroying any real variability. The model does not remove the smaller structure, however, some of the small structure can be further mitigated by using nearby stars as reference. A comparison between Fig. 28 and Fig. 37 (which shows the before and after correction for the same seven stars) displays the effectiveness of this correction procedure.

After performing the correction, we were able to explore some of the scientific projects this database can provide. We have shown how we can utilise the colours provided by the cross-matched catalogues to

generate a Hertzsprung–Russell diagram (see Fig. 39). We can also use the measurements in different filters to classify the population of objects in our catalogue. We also investigated some periodic variables in our catalogue (see Sec. 6.3). Here, we found a number of periodic variable stars via the Lomb-Scargle method (Lomb, 1976) (Scargle, 1982). Finally, we addressed the original goal that the astronomer who took this data had: exoplanet transits. We selectively looked at hot Jupiters in order to give a best-case scenario. From this, we calculated that our catalogue should hold $\approx$ 156 hot Jupiters. However, when accounting for the observed change in flux '$\frac{\Delta F}{F}$' we found that the photometric accuracy of our database is not high enough, even after the correction procedure. To rectify this we investigate how many data points we need to combine to get a high enough photometric accuracy to reliably detect a planetary transit. From Fig. 45 we found that combining 42 data points (and thus reducing the temporal resolution from 1 minute to 42 minutes) gives us a 90% probability of detecting a hot Jupiter.

References

Applegate J. H., 1992, ApJ, 385, 621

Astropy Collaboration et al., 2013, A&A, 558, A33

Auvergne M., et al., 2009, A&A, 506, 411

Bellm E. C., et al., 2019, PASP, 131, 018002

Benedict G. F., et al., 2002, AJ, 124, 1695

Bertin E., 2006, Automatic Astrometric and Photometric Calibration with SCAMP. p. 112

Bertin E., Arnouts S., 1996, A&AS, 117, 393

Borucki W. J., et al., 2003, in Fridlund M., Henning T., Lacoste H., eds, ESA Special Publication Vol. 539, Earths: DARWIN/TPF and the Search for Extrasolar Terrestrial Planets. pp 69–81

Brown W. R., Geller M. J., Kenyon S. J., Kurtz M. J., 2005, ApJ, 622, L33

Cao Y., Nugent P. E., Kasliwal M. M., 2016, PASP, 128, 114502

Charbonneau D., Brown T. M., Latham D. W., Mayor M., 2000, ApJ, 529, L45

Contreras Peña C., et al., 2014, MNRAS, 439, 1829

Contreras Peña C., et al., 2017, MNRAS, 465, 3011

Corradi R. L. M., et al., 2008, A&A, 480, 409

Cutri R. M., et al. 2014, VizieR Online Data Catalog, p. II/328

Evitts J. J., et al., 2020, MNRAS, 493, 184

Ferreira Lopes C. E., et al., 2020, MNRAS, 496, 1730

Froebrich D., et al., 2018a, Research Notes of the American Astronomical Society, 2, 61

Froebrich D., et al., 2018b, MNRAS, 478, 5091

Gaia Collaboration 2018, VizieR Online Data Catalog, p. I/345

Gaia Collaboration et al., 2018a, A&A, 616, A1

Gaia Collaboration et al., 2018b, A&A, 616, A10

Gautschy A., Saio H., 1995, ARA&A, 33, 75

Gautschy A., Saio H., 1996, ARA&A, 34, 551

Ginsburg A., et al., 2019, AJ, 157, 98

Herbig G. H., 1966, Vistas in Astronomy, 8, 109

Hillenbrand L. A., et al., 2018, ApJ, 869, 146

Hoffman D. I., Harrison T. E., McNamara B. J., 2009, AJ, 138, 466

Joy A. H., 1945, ApJ, 102, 168

Koenig X. P., Leisawitz D. T., 2014, ApJ, 791, 131

Kovács G., Zucker S., Mazeh T., 2002, A&A, 391, 369

Kuhn M. A., Hillenbrand L. A., Carpenter J. M., Avelar Menendez A. R., 2020, arXiv e-prints, p. arXiv:2006.08622

Law N. M., et al., 2009, PASP, 121, 1395

Lawrence A., et al., 2007, MNRAS, 379, 1599

Leung H. W., Bovy J., 2019, MNRAS, 489, 2079

Lomb N. R., 1976, Ap&SS, 39, 447

Lucas P. W., et al., 2008, MNRAS, 391, 136

Lucas P. W., et al., 2017, MNRAS, 472, 2990

Luri X., et al., 2018, A&A, 616, A9

Manfroid J., Heck A., Lunel M., Bergeat J., 1987, A&A, 176, 180

Marcy G., Butler R. P., Fischer D., Vogt S., Wright J. T., Tinney C. G., Jones H. R. A., 2005, Progress of Theoretical Physics Supplement, 158, 24

Montmerle T., 1990, in Klare G., ed., Accretion and Winds. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 209–233

Oliphant T., 2006–, NumPy: A guide to NumPy, USA: Trelgol Publishing, http://www.numpy.org/

Pustylnik I., 1998, Astronomical and Astrophysical Transactions, 15, 357

Ricker G. R., et al., 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003

Saito R. K., et al., 2012, A&A, 537, A107

Samus' N. N., Kazarovets E. V., Durlevich O. V., Kireeva N. N., Pastukhova E. N., 2017, Astronomy Reports, 61, 80

Scargle J. D., 1982, ApJ, 263, 835

Scaringi S., et al., 2018, MNRAS, 481, 3357

Schönrich R., McMillan P., Eyer L., 2019, MNRAS, 487, 3568

Seager S., Mallén-Ornelas G., 2003, ApJ, 585, 1038

Skrutskie M. F., et al., 2006, AJ, 131, 1163

Stetson P. B., 1996, PASP, 108, 851

Tody D., 1986, The IRAF Data Reduction and Analysis System. p. 733, doi:10.1117/12.968154

Tomita Y., Saito T., Ohtani H., 1979, PASJ, 31, 407

Turner D. G., 1996, J. R. Astron. Soc. Canada, 90, 82

Wes McKinney 2010, in Stéfan van der Walt Jarrod Millman eds, Proceedings of the 9th Python in Science Conference. pp 56 – 61, doi:10.25080/Majora-92bf1922-00a

Wood A., 2007, Monthly Notices of the Royal Astronomical Society, 380, 901

Wozniak P. R., 2000, Acta Astron., 50, 421

Wright E. L., et al., 2010, AJ, 140, 1868

Appendix

7.0.1 Photometric Table

The photometric table is the table that holds most of the data that will be used for research and is linked with every other table.

data_id	Object ID	image_id	bias_id	dark_id	flat_id	mag	magerr	background	x_image	y_image	ra	dec	fwhm_world	mag_cali	mag_cali_err	flags	flags_cali
15	0	0	0	0	0	18.1705	0.0657	30.24354	3844.563	12.8122	318.9130132	45.1597399	0.00449694	18.0215	0.0481	1.0	1.0
36409	0	2	0	0	0	18.3186	0.0664	25.87227	192.5772	4019.2031	318.9127603	45.1596878	0.00317579	18.0791	0.0559	0.0	0.0
889	0	4	0	0	0	18.2129	0.0787	44.06126	3900.1089	46.5386	318.9128568	45.1597575	0.00278629	18.0521	0.0508	0.0	0.0
555	0	5	0	0	0	18.0308	0.0697	41.93408	3886.2964	22.2969	318.9128755	45.1595722	0.00507708	17.8729	0.0502	0.0	0.0
117	0	7	0	0	0	18.3569	0.0674	34.51875	3867.5017	26.0992	318.912908	45.1597781	0.00307249	18.1495	0.0476	0.0	0.0
76	0	8	0	0	0	18.3775	0.0868	34.70331	3857.8066	22.9418	318.912978	45.1597156	0.0028584	18.1246	0.0473	0.0	0.0
230	0	10	0	0	0	18.3188	0.0628	38.132	3872.4788	47.0157	318.9131082	45.1597148	0.00353917	18.1491	0.0495	0.0	0.0
833	0	11	0	0	0	18.2108	0.0612	27.98271	3862.6523	40.1558	318.9131781	45.1597959	0.00439764	18.0582	0.0521	0.0	0.0
31261	0	13	0	0	0	18.081	0.0905	39.46214	220.7748	4022.5356	318.9126571	45.1594849	0.00270943	17.8398	0.0458	0.0	0.0
36433	0	15	0	0	0	18.2457	0.0653	35.6738	213.4803	4015.6758	318.9129606	45.1595397	0.00302489	18.0493	0.0468	0.0	0.0
36349	0	17	0	0	0	18.2227	0.0786	58.46521	201.068	4020.7153	318.9128676	45.1596312	0.00242166	17.9596	0.054	0.0	0.0
860	0	19	0	0	0	18.2551	0.0666	61.94218	3875.5315	52.1868	318.91281	45.1596543	0.00266559	18.0292	0.0449	0.0	0.0
36232	0	20	0	0	0	18.3124	0.0683	48.44613	191.0656	4034.5537	318.9127692	45.1597853	0.00301282	18.0763	0.0498	0.0	0.0

7.0.2 Identifier Linking Table

The Identifier table '`ID_DB.csv`' holds the right ascension, declination and unique identifier for each star that is has been found in the process outlined in Sect. 3.2.

data_id	ra	dec
0	318.91288	45.1597169
1	317.889531	45.158791
2	317.9886951	45.16054725
3	318.0119337	45.1602028
4	316.3128244	45.13502325
5	316.5782057	45.14165045
6	317.18275855	45.1521247
7	317.4307205	45.1554157
8	318.4148643	45.1630002

7.0.3 Image Table

The image table stores all of the relevant information about the FITS file. 'JD_BARY' is the Barycentric Julian Date, this is necessary as the investigations of this data are time sensitive.

image_id	image_name	image_dir	exp_time	CCD_TEMP	SEEING	FILTER	OBJECT	JD	JD_BARY
0	a-02756a.fit	2003-09-23	30.0	-20.1339285714	0.00195876299404	Red	fits_2003_4_area_a	2452906.34883	2452906.35176
1	a-03145a.fit	2003-09-28	30.0	-19.7470238095	0.00175197201315	Red	fits_2003_4_area_a	2452911.52473	2452911.52758
2	a-03348a.fit	2003-10-03	30.0	-20.5208333333	0.00171251699794	Red	fits_2003_4_area_a	2452916.4313	2452916.43403
3	a-03568a.fit	2003-10-04	30.0	-20.5357142857	0.00189896544907	Red	fits_2003_4_area_a	2452917.58155	2452917.58426
4	a-03687a.fit	2003-10-07	30.0	-20.1488095238	0.00136741402093	Red	fits_2003_4_area_a	2452920.34899	2452920.35163
5	a-03985a.fit	2003-10-11	30.0	-20.4761904762	0.00176759704482	Red	fits_2003_4_area_a	2452924.30808	2452924.31061
6	a-04414a.fit	2003-10-16	30.0	-20.1488095238	0.00183320348151	Red	fits_2003_4_area_a	2452929.44891	2452929.45128
7	a-04547a.fit	2003-10-17	30.0	-20.2827380952	0.00143815448973	Red	fits_2003_4_area_a	2452930.27832	2452930.28066
8	a-04870a.fit	2003-10-18	30.0	-20.3273809524	0.00175197201315	Red	fits_2003_4_area_a	2452931.33346	2452931.33576
9	a-05131a.fit	2003-10-20	30.0	-19.6130952381	0.00197823997587	Red	fits_2003_4_area_a	2452933.40772	2452933.40995
10	a-05380a.fit	2003-10-23	30.0	-20.0744047619	0.00171499699354	Red	fits_2003_4_area_a	2452936.31722	2452936.31934
11	a-05614a.fit	2003-10-25	30.0	-18.869047619	0.00192889594473	Red	fits_2003_4_area_a	2452938.28652	2452938.28856
12	a-05716a.fit	2003-10-26	30.0	-20.6547619048	0.00175499694888	Red	fits_2003_4_area_a	2452939.31609	2452939.31809

7.0.4 Bias, Dark and Flat tables

The three tables used for store information about image calibration are mostly used for debugging. Should the situation arise where some sets of data had a systematic offset the the calibration files used would be investigated. For the entirety of the data only 11, 15 and 19 Bias, Dark and Flat files where taken.

flat_name	flat_id
2003-10-16_m20_Red_sky_masterflat.fits	0
2004-08-30_m20_Red_sky_masterflat.fits	1
2004-09-01_m20_Red_sky_masterflat.fits	2
2004-09-03_m25_Red_sky_masterflat.fits	3
2004-09-12_m25_Red_sky_masterflat.fits	4
2006-09-12_m25_Red_sky_masterflat.fits	5
2006-09-20_m25_Red_sky_masterflat.fits	6
2006-10-24_m25_Red_sky_masterflat.fits	7
2006-10-27_m25_Red_sky_masterflat.fits	8
2006-12-01_m25_Red_sky_masterflat.fits	9
2007-08-26_m25_Red_sky_masterflat.fits	10
2007-09-18_m25_Red_sky_masterflat.fits	11
2007-10-14_m25_Red_sky_masterflat.fits	12
2007-12-07_m25_Red_sky_masterflat.fits	13
2008-09-27_m25_Red_sky_masterflat.fits	14
2009-07-01_m25_Red_sky_masterflat.fits	15
2009-07-03_m25_Red_sky_masterflat.fits	16
2009-08-19_m25_Red_sky_masterflat.fits	17
2009-09-09_m25_Red_sk y_masterflat.fits	18

bias_name	bias_id
2004-08-29_m20_masterbias.fits	0
2004-08-30_m20_masterbias.fits	1
2004-09-01_m25_masterbias.fits	2
2004-09-02_m25_masterbias.fits	3
2004-09-03_m25_masterbias.fits	4
2004-09-04_m25_masterbias.fits	5
2004-09-05_m25_masterbias.fits	6
2004-09-08_m25_masterbias.fits	7
2004-09-12_m25_masterbias.fits	8
2007-10-14_m25_masterbias.fits	9
2004-09-09_m20_masterbias.fits	10

dark_name	dark_id
2004-08-31_m20_masterdark.fits	0
2006-09-10_m25_masterdark.fits	1
2006-09-20_m25_masterdark.fits	2
2006-10-27_m25_masterdark.fits	3
2006-12-01_m25_masterdark.fits	4
2007-01-14_m25_masterdark.fits	5
2007-08-15_m25_masterdark.fits	6
2007-08-27_m25_masterdark.fits	7
2007-09-18_m25_masterdark.fits	8
2007-10-14_m25_masterdark.fits	9
2007-11-21_m25_masterdark.fits	10
2007-12-07_m25_masterdark.fits	11
2008-09-25_m25_masterdark.fits	12
2009-08-19_m25_masterdark.fits	13
2009-09-09_m25_masterdark.fits	14

7.0.5 Additional Light Curves

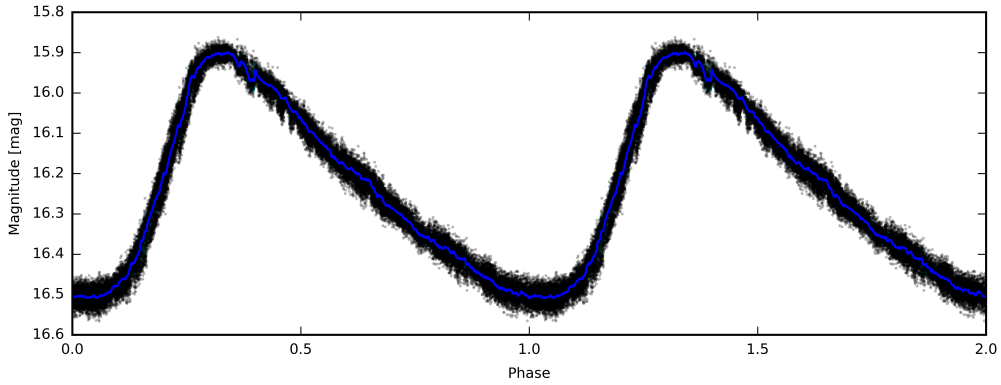

Figure 46: Shows the phase folded light curve of star V* V356 Cyg folded with a period of 5.05683 days, as given by the Lomb-Scargle.

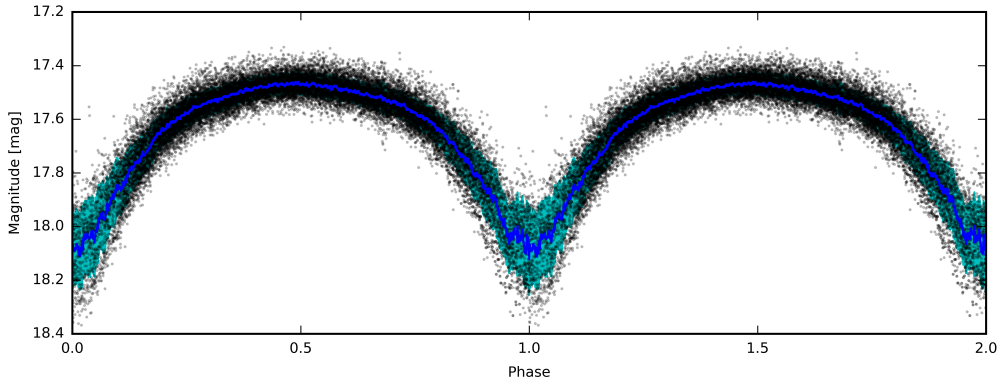

Figure 47: Shows the phase folded light curve of star NSVS 5840174 folded with a period of 1.06857 days, as given by the Lomb-Scargle.

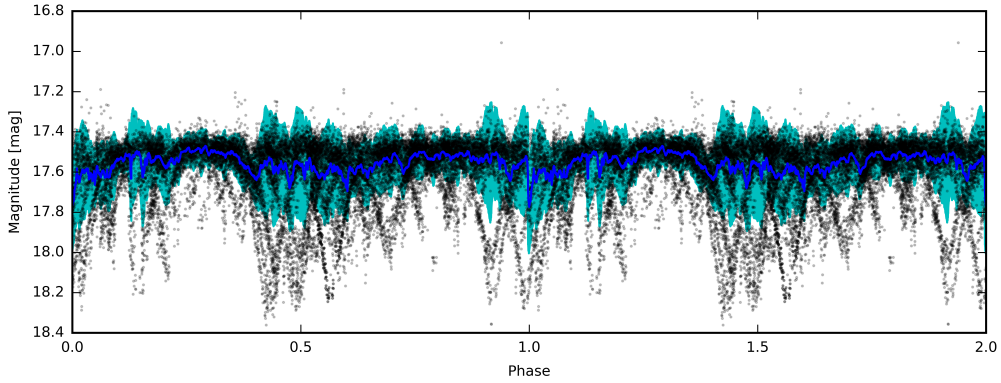

Figure 48: Shows the phase folded light curve of star NSVS 5840174 folded with a period of 2.29623 days, as stated in Hoffman et al. (2009).

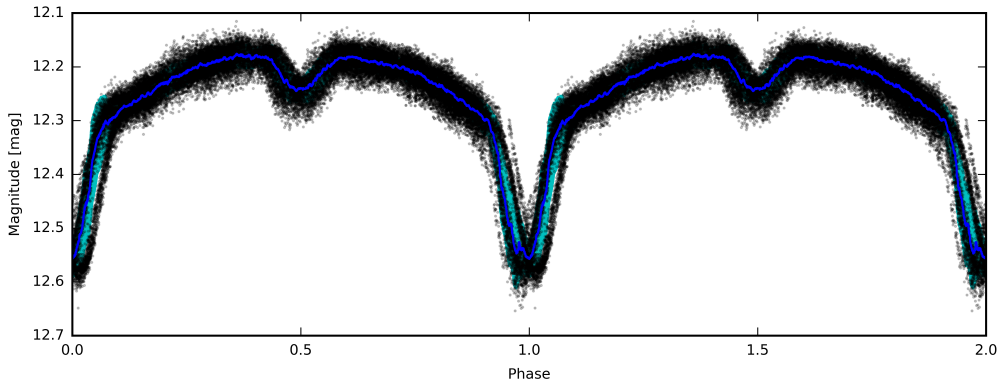

Figure 49: Shows the phase folded light curve of star V* V1898 Cyg folded with a period of 1.51317 days, as given by the Lomb-Scargle.

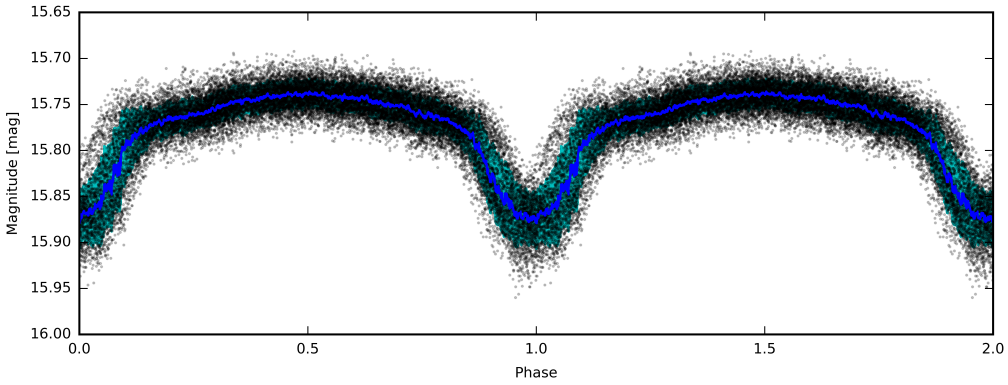

Figure 50: Shows the phase folded light curve of star TYC 3588-196-1 folded with a period of 0.56153 days, as given by the Lomb-Scargle.