Research Thesis: Declaration of Authorship

Print name: Michael Johnson

Title of thesis: Improving the Quality of Astronomical Survey Data

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;

2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;

3. Where I have consulted the published work of others, this is always clearly attributed;

4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

5. I have acknowledged all main sources of help;

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

7. Parts of this work have been published as: [please list references below]:

Signature:

Date:

References:

Johnson, Michael AC, et al. "Prospecting for periods with LSST–low-mass X-ray binaries as a test case." *Monthly Notices of the Royal Astronomical Society* 484.1 (2019): 19-30.

Strader, Jay, et al. "The Plane's The Thing: The Case for Wide-Fast-Deep Coverage of the Galactic Plane and Bulge." *arXiv preprint arXiv:1811.12433* (2018).

Johnson, Michael AC, et al. "Using the provenance from astronomical workflows to increase processing efficiency." *International Provenance and Annotation Workshop*. Springer, Cham, 2018.

UNIVERSITY OF Southampton

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given,

When referring to this thesis and any accompanying data, full bibliographic details must be given,
e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]

UNIVERSITY OF SOUTHAMPTON

Improving the Quality of Astronomical Survey Data

by

Michael A. C. Johnson

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

March 2020

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Michael A. C. Johnson

Astronomical survey telescopes are becoming increasing capable at generating large datasets. The quantities of data being produced necessitate the automation of the data processing which is commonly accomplished via astronomical workflows. The large scale of the data also means that small improvements in the quality of the data processing can have large implications for the value of the science gained. However, deciding on which workflow configuration is best is usually a qualitative process, achieved through trial and improvement which lacks a quantitative measure of the quality of the results produced by each workflow version. Consequently, the best workflow cannot be reliably chosen. Thorough analysis is typically applied to find specific outputs from astronomical workflows, such as the magnitude of an object. However, this targeted analysis focuses on specific components and does not utilise the wider workflow space or the provenance of the workflows. This thesis therefore outlines an approach to be applied to workflows to assess over different workflow versions and measure the quality of data that they produce. To test the approach, it was applied to three separate use cases. The first application used the approach to predict the completeness of period recovery of transient and variable astronomical sources with several candidate observing strategies from upcoming front line astronomical surveys. It was found that observing strategies which did not reduce the observations within the Galactic Plane increase the completeness by a factor of $\sim$3. The second was an investigation into the use of provenance to improve the timeliness of a differential photometry workflow. It was found that this method offered improvements of at least 96% in computational efficiency when analysing the outlined use cases. The third application was to improve the accuracy and completeness of a workflow designed to search for transients within a set of archival calibration data from an astronomical survey telescope. Workflow configurations were generated using the manual method in addition to via the approach. The best performing workflow found through the approach outperformed the workflow generated through the manual method and consequently found an additional $\sim$2,500 transient events. However, full evaluation of the approach could be a computationally expensive process, therefore the hill climbing algorithm was also investigated as a means to quickly find a verifiably good workflow configuration. The quality of the results produced by the workflow generated through this method were found to be within 0.2% of those produced by the highest quality workflow found.

Contents

List of Figures

List of Tables

Listings

Nomenclature

Air mass
The quantity of air between a celestial object and the observer

Asymptotic giant branch
A region on the Hertzsprung-Russell diagram populated by intermediate mass stars late in their evolution

Compact objects
The collective term for white dwarfs, neutron stars and black holes

Cosmic rays
High energy protons and atomic nuclei

Confusion limited image
An image where the limiting factor for object detection is from distinguishing them from other objects

Hertzsprung-Russell diagram
A scatter plot of luminosity vs temperature in order to separate the objects into their stellar classifications

Magnitude
A logarithmic measurement for brightness

Natal kick
Momentum imparted to compact objects after their formation via supernovae

Point spread function
The response of an imaging system to a point source

Provenance
Documentation describing the production of a piece of data or a thing

Reddening/Extinction
The scattering of electromagnetic radiation by gas and dust, typically strongest in the bluer wavelengths

Roche lobe
The region around a star in a binary system within which orbiting material is gravitationally bound to that star

Roche lobe overflow
A mechanism for mass transfer within a binary system when material lies outside of the Roche lobe

Acknowledgements

I was lucky enough to have three supervisors during my PhD and I would like to thank Dr. Adriane Chapman, Prof. Luc Moreua and Dr. Poshak Gandhi for their support and guidance throughout. I thank Luc for demonstrating the work ethic required to succeed in academia and for his patience while I got acclimatised to the world of computer science. I am grateful that Poshak was never too busy to support me whilst simultaneously being extremely busy. I thank Age for her ability to make connections between the most seemly disconnected of topics which was instrumental in making a cohesive interdisciplinary thesis.

I thank Prof. Phil Charles for his advice and support throughout my time a Southampton University, from being my personal tutor as an undergraduate to a co-author during my PhD. I also want to thank Prof. Christian Knigge for countless discussions on statistics and Dr. Adam Hill for answering all of my many data science questions.

I thank past and present members of Southampton University for their advice and friendship during my PhD, these include: Dr. Peter Boorman (for feeding me like he was my grandmother), Dr. Rory Brown, Dr. Jamie Court, Dr. Aru Beri, Dr. Steve Browett, Philip Grylls, James Leftly, Edward Parkinson, David Price and Miika Pursiainen.

I thank Ella Guise for putting up with all of the late nights spent writing and sleeping on the sofa far too many times so that I could work on the desk in the bedroom. I also truly thank her for all of the support which kept me (mostly) sane throughout my PhD, for laughing at my bad jokes and for making my University experience what it was.

I thank my mother and father for their continued support throughout my PhD. I am eternally grateful to my mother for proof reading all of my papers, reports and thesis versions.

This work was made possible with funding support from the Engineering and Physics Research Council.

Dedicated to my late Grandfather, Prof. A. Low Thomson for sparking my interest in Physics from an early age.

Chapter 1

Introduction

Data quality is composed of many different metrics which each describe different aspects of the data. Wand and Wang (1996) summarised the different dimensions of data quality and twenty six metrics were identified, among those included were accuracy, timeliness, and completeness. The relevance of each quality metric is dependant on the data itself. When datasets are large, even marginal improvements in the quality of the data can offer significant gain from the results. The scale of large datasets necessitate that the processing be automated, this is often realised through the use of workflows. These workflows are typically composed of many subcomponents, each with a large scope for customisation. Whilst finding the best workflow is important for the quality of the results, the numerousness of the potential workflow variants makes the determination of this workflow a cumbersome process.

Technological advances within the last few decades have enabled astronomers to collect far more data than previously possible. As a result, the lion's share of astronomical data collection is now due to survey telescopes, rather than the point-and-shoot method traditional to astronomy, some notable examples of this are: the Sloan Digital Sky Survey (SDSS, York et al. 2000); the Transiting Exoplanet Survey Satellite (TESS, Sullivan et al. 2015); and the Large Synoptic Survey Telescope (LSST, Tyson 2002). The datasets from these surveys are sometimes open access, specialised to a region within the electromagnetic spectrum, and describe a wide region of the night sky. Big multi-wavelength datasets can be a powerful tool in the hands of the astronomical community, however astronomical datasets can lead to astronomical problems. These datasets are notoriously noisy as they can be contaminated by sources such as cosmic rays, atmospheric refraction, and light pollution. The total dataset size can range from the terabyte scale for past survey telescopes such as SDSS and TESS to the petabyte scale for future telescopes, such as LSST.

The scale of the data production means that cleaning of these data must be an automated process, therefore each of these survey telescopes had a bespoke data management system

dedicated to the processing of astronomical data (SDSS-Ivezić et al. 2004, TESS-Jenkins et al. 2016, LSST-Juric et al. 2015). These data management systems contain many free parameters which can have a large impact on the quality of the results produced. As the scale of the data is so large, even small improvements in the data quality can offer large scientific gain. However, finding the best workflow configuration comes with its own complications such as the lack of ground truths in real data, the size of parameter space to investigate and the definition of good quality data.

Each of these data management systems also handles the documentation of the data processing - the provenance. This is essential for establishing reliability and reproducibility of the data products but introduces an initial overhead to the computational resources required to process the data. In an effort to offset this, the LSST data management system for example, will utilise pipeline provenance (Groth et al., 2010) to reduce the size and processing time required for generating provenance (Kantor, 2006). Whilst this method is certainly an improvement, it only has the means to reduce the cost of provenance recording and alternative methods need to be investigated in order to neutralise the cost completely.

Individual astronomers often lack the resources necessary to build such a data management system but still require automated processing to handle the large data volumes. As a result, the tendency is to utilise prebuilt software packages such as Source Extractor (SExtractor) (Bertin and Arnouts, 1996) for photometry or the High Order Transform of PSF ANd Template Subtraction code (HOTPANTS) (Becker, 2015) for difference imaging. Both of these examples contain over 50 categories of parameters for customisation. The astronomer must determine the best set of these parameters before processing each distinct dataset. Typically, this is achieved through a qualitative method of trial and improvement which is time consuming and lacks a quantitative measure of the overall quality of the data processing. The implications of the choice of SExtractor settings were investigated by Hetterscheidt et al. (2005), where they used two distinct versions of SExtractor settings to detect cluster-sized dark matter halos. They found a relation between the SExtractor settings chosen and both the number of objects detected within their images and the signal to noise ratio of the cluster detection. The restricted region of SExtractor parameter space investigated was likely to have been motivated by the computational cost of increasing it but without an analysis of the wider parameter space, they are unlikely to find the best choice of parameters for their use case. The problem of finding good SExtractor settings for each application was identified by Ryan Jr (2011) who presented a suite of IDL routines capable of interactively running SExtractor through a graphical user interface. While this solution may decrease the time required to find the ideal SExtractor settings, the interactive nature of the solution means that it lacks a quantitative assessment of the quality of data produced by each of the settings.

The quality of astronomical survey telescope data is not derived solely from the quality of the data processing; instrumentation, location, and observing strategy are also

fundamental within this regard. For example, the ongoing discussion surrounding the LSST observing strategy is outlined by Marshall et al. (2017) where candidate observing strategies are assessed via key quality metrics outlined by different scientific communities. One of the most controversial components of the current baseline observing strategy is the reduced cadence for all fields within the Galactic Plane, the motivation being it is thought that the high density of sources within the Galactic Plane will confusion limit the images so that no extra depth will be gained from additional exposures. However, the reduced cadence will also significantly limit the potential to gain temporal information within the Galactic Plane, where the majority of known objects reside.

The aim of the thesis was to define an approach to be applied to workflows in order to assess over different workflow versions and determine the quality of data that they produce. The approach consisted of a set of definitions and a formalism for determining the relative quality of possible workflow versions.The application of the approach therefore enabled the identification of the best workflow configuration. The approach was tested against three separate use cases which each utilise data from different astronomical surveys. The first workflow was an investigation into how to improve the quality of data to be taken by LSST in the future where the completeness of period recovery capable with different observing strategies was assessed. The second used data taken by the Faulkes telescope and focused on improving the efficiency of workflows in the present day by utilising the provenance of the workflow, the quality metric evaluated was timeliness. The third utilised past data from the Kepler spacecraft, attempting to find transients within a set of calibration data where the relevant quality metrics were accuracy and completeness. For this application, the number of possible workflows was large and consequently, so was the computation time required for evaluation of the approach. To solve this problem, the hill climbing algorithm was investigated as a method to quickly arrive at a verifiably good version of the workflow in a short time.

1.1 Research Statement

To improve the quality of results produced by workflows designed to analyse astronomical survey data through the use of an approach which formulates the relevant quality metrics and evaluates them over the workflow configurations.

1.2 Contributions

The focus of this thesis is to improve the quality of data produced by astronomical survey telescopes. A generalisable approach was developed to achieve this goal which was then applied to three separate astronomy use cases. Therefore, the major contributions within this thesis are:

1. The creation of a generalised approach for increasing the quality of data produced by astronomical workflows
2. The application and evaluation of the hill climbing algorithm as a method for quickly estimating the best workflow configurations with regards to data quality
3. Using this approach, the discovery and analysis of variable and transient objects within the Kepler Full Frame Images
4. Using the approach, an evaluation of candidate observing strategies for the Large Synoptic Survey Telescope with regards to variable, Galactic science
5. Investigating the use of provenance to analyse the final data products of astronomical workflows in order to increase their processing efficiency

1.3 Structure

The thesis is structured as follows, Chapter 2 is the literature review, which introduces all of the relevant background information as well as describing the related work. Chapter 3 describes the approach which was used throughout the thesis to improve the quality of astronomical survey data. Chapter 4 investigated the quality of Large Synoptic Survey Telescope data for Galactic science when utilising different candidate observing strategies. Chapter 5 describes a provenance enabled aperture photometry pipeline and investigated whether provenance can be used to increase its processing efficiency. Chapter 6 describes a workflow that was created to detect transients within the Kepler Full Frame Images and the results that it produced. This chapter also investigated the data quality of the workflow over a large region of parameter space and evaluated the use of the hill climbing algorithm as a method to quickly find good workflow settings. Chapter 7 discusses the applicability of the approach to each of the use cases and investigates changes in the data quality of each of the use cases when compared to their respective baselines. Finally, Chapter 8 outlines the future directions for this work.

Chapter 2

Literature Review

2.1 Astronomical Objects

2.1.1 Variable and Transient Objects

Variable and transient astronomical objects are defined as objects whose brightness varies with time. Variable objects do so consistently with a periodic or quasi-periodic nature, whilst transients are characterised by sporadic and often violent changes in brightness. Here, some of the more common examples of each are described. This list is by no means exhaustive but is included to provide the reader with some context to the objects that are discussed within this thesis.

2.1.1.1 Long Period Variable Stars

Long period variables (LPVs) are pulsating cool giant or supergiant variable stars characterised by periods ranging from a hundred to more than a thousand days. Most LPVs are thermally pulsating asymptotic giant branch stars with luminosities several times that of the Sun.

A sub category of LPV stars are the Mira variables. These objects are red giants in the very late stages of stellar evolution. They are a class of pulsating variable star characterised by periods longer than 100 days, very red colours and amplitudes greater than one magnitude. Within a few million years, these stars will expel their outer envelope and become white dwarfs.

2.1.1.2 Cataclysmic Variables

Cataclysmic variables (CVs) are a class of binary stars consisting of a white dwarf primary and a donor star. They exhibit dramatic increases in brightness followed by a return to their quiescent state. The outbursts observed from CVs result from the matter that is accreted from the companion star via Roche lobe overflow, forming an accretion disc around the white dwarf. Material from the inner edge of this disc accretes onto the white dwarf and a layer of hydrogen forms around its surface. Classical novae are observed when the temperature and pressure on the surface of the white dwarf are sufficient for hydrogen fusion and the hydrogen layer quickly fuses into helium. Dwarf novae are a subclass of CVs in which there is a small separation between the binary components. These exhibit more regular but smaller outbursts that are as a result of temporary increases in the accretion rate onto the white dwarf.

2.1.1.3 X-ray Binaries

Low mass X-ray binaries (LMXBs) consist of a neutron star or black hole accreting matter from a low mass (usually $< 1M_{\odot}$) orbiting companion donor star via Roche lobe overflow. The in-falling matter forms an accretion disc around the compact object, a diagram of the system is shown in Figure 2.1. This matter releases gravitational potential energy as it approaches the compact object, emitting up to a tenth of its rest mass as X-ray radiation (Tauris and van den Heuvel, 2006).

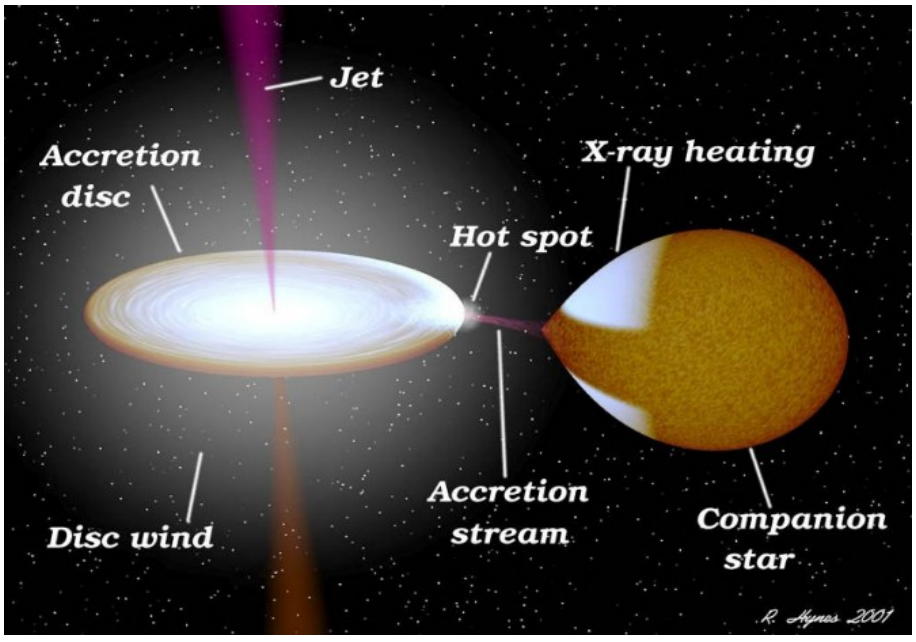


FIGURE 2.1: The anatomy of a low mass X-ray binary. Image credit: NASA/R. Hynes

LMXBs may either be in a state of quiescence or outburst, in quiescience the optical signature is dominated by the companion star whilst in outburst, it is dominated by the

reprocessed X-ray emission from the disc. As the companion star orbits, the gravitational attraction of the compact object elongates the star so that to an observer, the shape of the companion star's visible surface changes with the orbit. This produces the ellipsoidal modulation characteristic of LMXB light curves, as the two peaks per period correspond to points in the orbit where the observed surface area is greatest. The optical light curves of LMXBs have also been shown to contain fast optical variations superposed on the ellipsoidal modulation (Zurita et al., 2003), the origin of which remains debated. Figure 2.2 depicts a simulated LMXB light curve illustrating these phenomena.

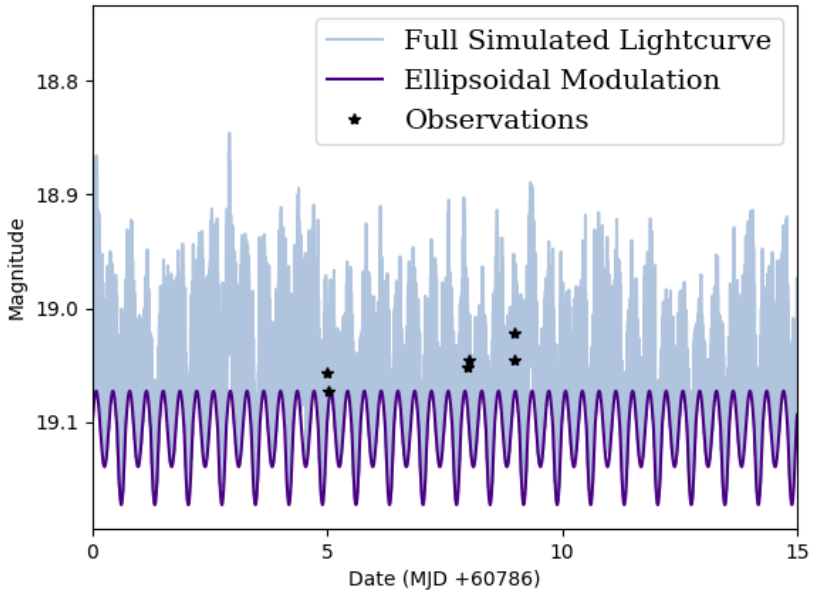


FIGURE 2.2: The optical light curve of a LMXB depicting the variations due to ellipsoidal modulation and the stochastic noise from the flaring shown in purple and blue, respectively.

Mineshige et al. (1994) observe $\frac{1}{f}$ (with f as frequency) X-ray fluctuations regardless of mass input. They attribute these red noise contributions to inconsistencies in the accretion flow - either occurring as gradual gas diffusion or as avalanches of infalling matter, triggered by instabilities in the accretion disk. These avalanches explain the transition of the LMXB from the quiescent state to outburst. The physical mechanism for which is outlined by Lasota (2001) in his disc instability model (DIM) whereby matter is accumulated on the inner radius of the accretion disc until a critical limit is reached. Once reached, the disc undergoes thermal runaway thereby triggering a shock wave which forms from the inner edge and propagates outwards - named an inside-out outburst. In practice, the shock wave never forms exactly at the inner edge so there is always an associated outside-in outburst. As the matter in this front propagates outwards from the disc, it takes angular momentum with it as the disc's angular momentum increases with increasing distance from the compact object. Inside-out outburst fronts may not reach the outer edge but they cause the outer regions to accumulate matter, enabling future

fronts to do so. As these fronts move outwards, the disc behind the front becomes very hot and begins to diffuse inwards, creating an initial rise in the emission and leads up to the outburst.

These characteristics have been observed in LMXBs GS 1354-64 and V404 Cyg by Koljonen et al. (2016) and Bernardini et al. (2016), respectively. Figure 2.3 depicts the optical light curve of the outburst of GS 1354-64, which has the characteristic slow rise in optical luminosity over the course of years, attributed to the gradual build up of matter in the accretion disc. This is followed by the sharp rise in optical luminosity in the weeks directly preceding the maximum as the accretion rate of the LMXB begins to increase.

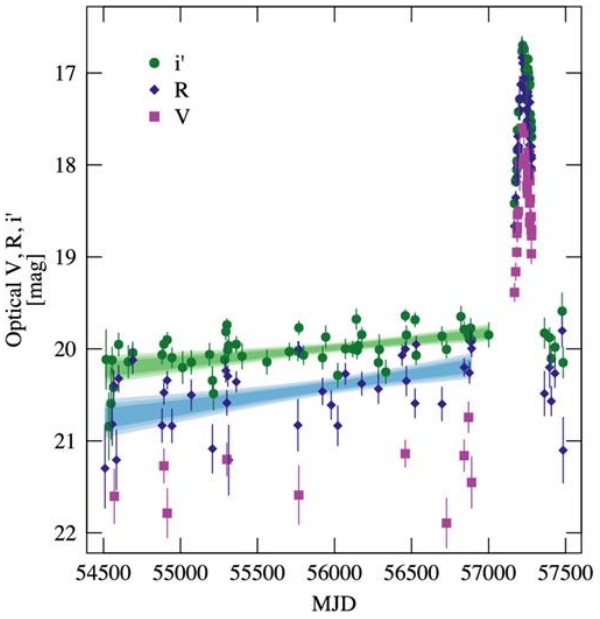


FIGURE 2.3: The optical light curve of LMXB GS 1354-64 on the rise to and during the outburst phase. Image credit: Koljonen et al. (2016)

In LMXBs the orbital period (P_{orb}) is the fundamental observable that can be combined with radial velocity information from spectra to determine the masses of the binary components. This can be achieved via the mass function shown in Equation 2.1, where: P_{orb} is the orbital period; i is the orbital inclination angle; K is peak orbital velocity of the companion star; G is the gravitational constant; M_1 and M_2 are the masses of the companion star and compact object, respectively. The mass of the companion star is typically estimated from observations of its spectral type and the equation is solved for M_2.

$$\frac{M_2^3 \, sin^3 i}{(M_1 + M_2)^2} = \frac{P_{orb} \, K^3}{2 \, \pi \, G} \tag{2.1}$$

Characterising the masses of LMXBs can gain insight into the processes that form these systems, for example Type Ibc and Type II supernovae. Properties such as the explosion energy, mass cut or the explosion mechanism can all have implications in determining the

final mass of the compact object. Additionally, measurements of X-ray binary motion and location in the Galaxy could help characterise the natal kicks that supernovae are expected to impart to the compact object (see: Van Paradijs and White 1995; Jonker and Nelemans 2004; Repetto et al. 2017; Gandhi et al. 2019). Knowing P_{orb} is also crucial for ultra-high precision astrometry due to the orbital wobble - when the flux-weighted centroid of emission wobbles at the P_{orb} of the binary system Casares (2014).

Fryer and Kalogera (2001) simulated the expected mass distributions of black holes via their formation processes. Their results predicted a continuous mass distribution between $3-5\ M_{\odot}$. However, the mass distribution of known sources exhibits a gap in mass between 2-5 $M_{\odot}$ (see Casares et al. 2017). Expanding the population of dynamically confirmed compact object masses will be able to determine whether or not there is a true gap in mass between neutron stars and black holes. Additionally, it may allow us to determine whether black hole masses tend to cluster around a particular value.

Approximately 200 LMXBs have been observed within the Milky Way (Liu et al., 2007) and ∼59 of these are thought to host a black hole, although only ∼20 have been dynamically confirmed (Corral-Santana et al., 2016). Many of the LMXBs in the Liu et al. (2007) catalogue are either steady or transient LMXBs that have only been seen in outburst as they are too dim to observe otherwise with current telescopes. The mass distribution of LMXB counterparts is shown in Figure 2.4 from which the clear separation in masses between neutron stars and black holes can be observed. While the neutron star masses are clearly confined and cluster around certain masses, no such conclusions can be drawn for black holes and higher statistics are required for this purpose. The spatial distribution of LMXBs in Galactic coordinates is shown in Figure 2.5 where their tendency to be distributed within the Galactic Plane is displayed as is expected as the Milky Way is where all known LMXBs were formed. Positions at a distance perpendicular to the Galactic plane can be explained by the natal kick imparted to LMXBs when their compact object forms via supernova (Janka, 2013).

2.2 Astronomical Data

Astronomical data may come in many forms such as images, spectra, photon lists, data cubes or just plain structured data. All of the aforementioned data types can be stored in the most common file format within the astronomical community - the Flexible Image Transport System (FITS) which are formatted as N-dimensional arrays or tables. The metadata for a FITS file is stored in the form of a human readable ASCII header of key-value pairs, allowing users to investigate a file of unknown origin.

Optical astronomical images are usually taken using a combination of optical equipment such as a mirror or lens and a camera such as a charged coupled device (CCD). Photons from the target object are focused through the optics onto the CCD where they trigger

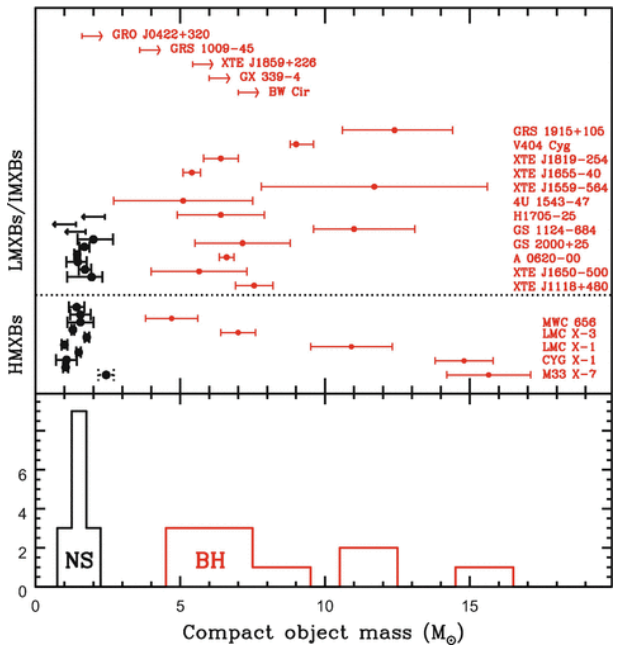


FIGURE 2.4: The mass distribution for LMXB compact objects. Image credit: (Casares et al., 2017)

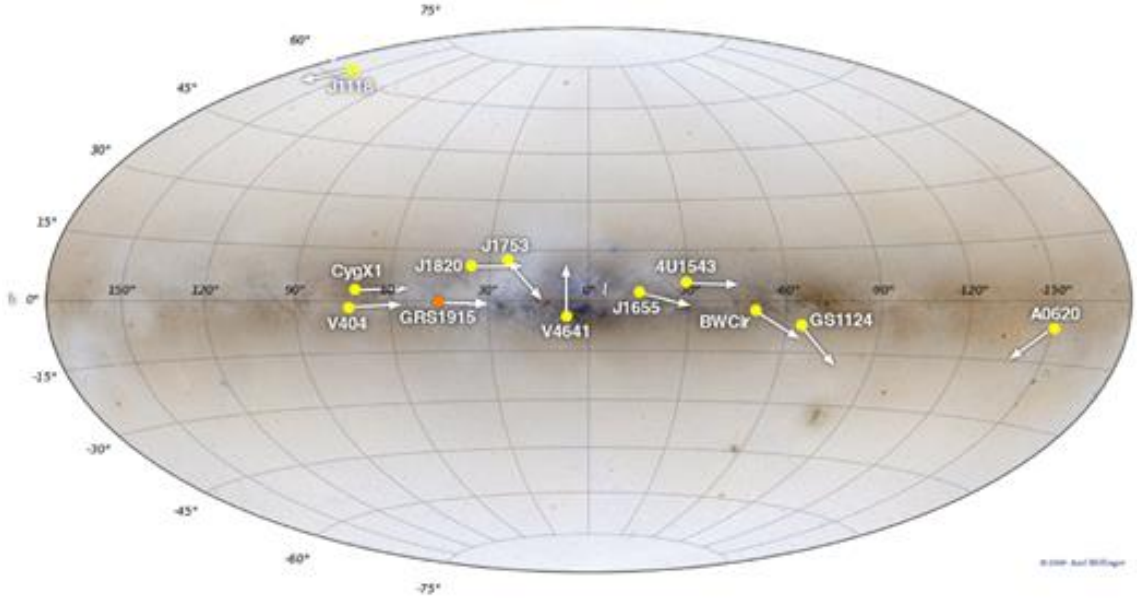


FIGURE 2.5: The spatial distribution of black hole LMXBs with measured proper motions shown in Galactic coordinates (Gandhi et al., 2019).

the release of electrons and the resultant current is measured as a signal. Before the images can be used for science, they need to be calibrated for systematic sources of noise. The three most commonly accounted for sources of systematic uncertainty are readout noise, thermal currents, and uneven illumination. The source of the readout noise is the CCD, which detects a current even when there is no signal. To account for this, a bias frame is taken where the exposure time is as small as possible and there is

no illumination, so that all signal is due to readout noise. This image is subsequently subtracted from the target image. Thermal noise is due to the release of current from thermally excited electrons. An image containing only thermal noise is made by taking an image without illumination for the same exposure time as the target image. This thermal frame is then subtracted from the image. Finally flat fields are taken to account for uneven illumination. These are usually taken of the sky at twilight so as to not be so bright as to saturate but sufficiently bright to not see individual stars, providing even illumination across the CCD. The bias frame is then subtracted from the flat field or, less commonly corrected with dark frames. Consequently, observed differences in brightness are due to either the CCD or the optics reacting differently to the same observed light source. The target image is divided by the flat field in order to finally leave a calibrated image. More information on correcting images for systematic sources of noise can be found in Widenhorn et al. (2007), and Seibert et al. (1998).

2.2.1 Photometry

The method for extracting information from the newly calibrated image depends on what information you want to find as well as the type of the target object. One of the most common techniques for extracting the photometry of an object is aperture photometry whereby an aperture (usually a circle) is centred on the object, entirely encompassing it and it collects all of the counts produced that object (Da Costa, 1992). In principle, this is a simple procedure however there are a number of complications, for example: the determination of the object centre, the choice of aperture, and the estimation of the background contribution. Other methods for photometry on astronomical images include differential photometry whereby aperture photometry is performed on two objects, one of known brightness and the other being the target object. The measured brightness of the known object can be compared to the true brightness in order to calculate a correction for the images sources of systematic uncertainty. This correction is then applied to the target object in order to find the "true" brightness of it.

2.2.1.1 Source Detection

Methods that locate stars in an image and their centres can leverage the typically Gaussian profile in the brightness of an observed star. One such method is implemented in DAOPHOT (Stetson, 1987) (a program for the automation of stellar photometry) with their FIND algorithm. The basis for this algorithm is to locate stars by going through the image, pixel by pixel and fitting a Gaussian profile to the brightness values of that pixel as well as the surrounding sub-array of pixels. If there is a star present, the Gaussian will have a good fit and a bright peak, which would not be the case if the pixel is in a region with no object or located at the tail-end of the Gaussian. If a candidate object is found, DAOPHOT then deblends the candidate objects in order to distinguish them from other

nearby objects. SExtractor (Bertin and Arnouts, 1996) is another astronomical photometry package capable of automatic object detection. SExtractor locates objects in much the same way but with a few additions such as additional scope for customisation. For example, the user has more control over the expected sizes and shapes of profiles of the stars, i.e. it can be Gaussian, Mexican hat or tophat and they can specify a pixel by pixel size of the expected objects.

Source detection may also be accomplished using empirical measurements of the point spread function (PSF) of the image. This method assumes that the PSF is constant in the image and that the pixels all have a linear response to incoming radiation. It may also only be done for point source objects as the shape of extended objects is not entirely governed by the PSF. However, it does have fewer assumptions on the shape of the PSF when compared to the aforementioned method as it is calculated from the image itself. This form of PSF fitting is included in DAOPHOT (Stetson, 1987) and once calculated, the PSF is used in place of the Gaussian model and fit to target stars. SExtractor is also capable of a PSF fitting but requires a model PSF as input Holwerda (2005).

2.2.1.2 Background Determination

When viewing a source through an aperture, not all the counts contained within the aperture will have originated from the source. This background flux can come from many places such as terrestrial night sky emission, contributions from other sources or scattered light inside the camera. To determine the background contribution, the usual method is to measure the signal in an annulus around the target object. An annulus is used to negate the effects of any potential gradient in the background and must be far enough from the source so as to not contain light emitted by it. Furthermore, the annulus must encompass a reasonable number of pixels (of order 100) which reduces the uncertainty in the background (Da Costa, 1992). With an idealised background the histogram of pixel intensity would be a Gaussian distribution and the appropriate value for the background would be equal to the mean. In real images, there is a positive skew on the mean owing to contamination from nearby astronomical objects which have very high pixel values. Therefore, the mode is used to represent the sky background. The mode does not need to be explicitly calculated and can be estimated with the formula in Equation 2.2 (Kendall and Stuart, 1977).

$$mode = 3\,median\ -\ 2\,mean \tag{2.2}$$

2.2.1.3 Image Registration

Before difference image analysis can be carried out over a set of images, they must first be spatially transformed so that they can be properly overlaid. Many sources of distortion

may be present in astronomical images such as spatial distortions due to differences in viewing angle or photometric distortions due to changes in the CCD. Image registration is the process of transforming the images to remove the differences between them that are due to the image acquisition only.

Cross-correlation is a statistical approach to image registration between a template image and target image. The basic function of this method is that it calculates a distortion function at each position of the image under examination which measures the similarity between the two images. The minimum distortion position is then taken to locate the corresponding positions on the two images (Sarvaiya et al., 2009). One of the disadvantages of this method is the low computational efficiency. This can be improved by using the Fast Fourier Transform (FFT) to compute the correlation as a product of Fourier transforms. Although this method is faster, it has the drawback of requiring a large memory capacity which scales with the log of the image area. This method was then further improved by using discrete Fourier transforms (DFTs) in place of FFTs. Guizar-Sicairos et al. (2008) found that the use of DFTs reduced both the computation time and required memory for this image registration whilst still achieving sub pixel accuracy on the result of the image registration.

2.2.1.4 Image Subtraction

Studying variability in astronomical images was previously achieved only by using catalogues of stellar brightnesses. While this method has proven useful, the scale of astronomical data production requires fast and accurate ways to study astronomical variability. Image subtraction is the method by which one image is matched to another using a convolutional kernel, in order to difference them and measure only the objects with brightness variations between images.

The first successful implementation of difference imaging was by Tomaney and Crotts (1996) whereby they measured the PSF for as many objects in the image as possible, given their strict conditions(high signal/noise, not saturated or corrupted, amplitude well above that of unresolved stars and isolated from nearest neighbours), totalling 220 objects. However, the spatial variation of PSF throughout their image dictated that in order to accurately match the entire 2048^2 pixel frame, 1500 subframes were required. In order to meet this requirement, they derived the spacial dependence of the PSF using a polynomial surface fit on the known PSFs, deriving one convolutional kernel for the entire image. They used this to compute the intermediate PSF models required for accurate image subtraction.

The High Order Transform of PSF ANd Template Subtraction (HOTPANTS) is an implementation of the Alard (2000) algorithm for image subtraction (Becker, 2015). The algorithm outlined in Alard (2000) instead of creating one kernel that is used to combine

both images, splits the image up into multiple regions and a kernel is constructed for each region. Each region is split into stamps with a size of 100 pixels and each stamp is composed of sub-stamps that are centred on astronomical objects, the purpose of which is to place constraints on the kernel with each sub-stamp. This method has a similar CPU cost to generating a single convolutional kernel but performs better when there are differential PSF variations or rotations present between the images.

2.2.2 Data Analysis

2.2.2.1 Period Determination

Astronomical objects such as X-ray binaries, Cepheid variables and transiting exoplanets all exhibit periodic variations in their brightnesses. The origin in many astronomical objects is orbital motion, however it can also be due to other physical processes as exemplified by the pulsations from cepheid variable stars. The length of the period is linked to other physical properties of the object.

If the period is due to orbital motion then the period will be related to the mass of the orbiting bodies as well as their separation. In the case of Cepheid variables, the period of pulsation is closely linked to the luminosity of the star, making period determination useful for measuring astronomical distances. These are examples of how accurate determination of the period of astronomical objects can provide insight into their other physical properties.

A Fourier series is a periodic function composed of harmonically related sinusoids. If the light curve of an astronomical object is assumed to be such a series, then a Fourier transform can be performed on it which decomposes the light curve into the weighted sum of the sinusoids. The Fourier transform of such a periodic function has the form shown in Equation 2.3, where $f_P(t)$ is the periodic function, P is the period, k is the wave vector, t is time and $F(k)$ denotes the Fourier transform.

$$F(k) = \frac{1}{P} \int_P f_P . e^{-i\,2\,\pi\,\frac{k}{P}\,t} dt \tag{2.3}$$

Equation 2.3 only applies to continuous data whereas real astronomical light curves are usually sampled at discrete points. In order to account for this the Fourier transform can have the form shown in Equation 2.4 (Grafakos, 2008).

$$F(k) = \frac{1}{P} \sum_{j=0}^{N-1} f_P . e^{-i\,2\,\pi\,\frac{k}{P}\,t_j} \tag{2.4}$$

There are further complications as Equation 2.4 only applies for evenly sampled data. A further extension is needed in order to analyse typical, unevenly sampled astronomical data. One solution is the Lomb-Scargle algorithm (Lomb 1976, Scargle 1982). This is a well known and widely used algorithm for characterising periods from unevenly sampled data via least squares fitting. As with the Fourier transform, the output of the Lomb-Scargle algorithm is a power spectrum showing the relative abundances of the fundamental frequencies present in the original data. One further extension to the Lomb-Scargle is presented by VanderPlas and Ivezic (2015) with the multi-band periodogram, which constructs the periodogram using information from all the different filters available. This technique has the potential to significantly improve the period determination of sparsely sampled, multi-band astronomical surveys, such as LSST.

The periodogram produced by the Lomb-Scargle algorithm will usually contain contaminating noise of two different types, white noise and red noise. White noise has a zero mean, constant variance and is uncorrelated in time whereas red noise has a zero mean, constant variance but is correlated in time and is typically weighted towards lower frequencies.

2.3 Large Scale Astronomical Survey Telescopes

2.3.1 The Large Synoptic Survey Telescope

In 2022, the Large Synoptic Survey Telescope (LSST) will begin a 10 year synoptic survey in six filters (*ugrizy*) from Cerro Pachón, Chile (Ivezic, 2014). The 8.4m primary mirror of LSST will enable it to observe down to $r \sim 27$ mag and the Paul-Baker three mirror optical design (shown in Figure 2.6) provides a 9.6 square degree field of view. The high optical sensitivity of LSST combined with large field of view means that its etendue is $319 m^2 deg^2$, where etendue is a measure of the throughput of a telescope defined as the product of the collecting area and field of view. Figure 2.7 shows the comparison with other large scale survey telescopes and demonstrates the expected improvements in astronomical data collection that is expected with the next generation of telescopes. The optical design of LSST enables it to scan the sky wide, fast and deep in order to simultaneously complete all of its scientific objectives using a single observing strategy.

The four primary scientific objectives for LSST are:

- The Nature of Dark Matter and Understanding Dark Energy
- Cataloguing the Solar System
- Exploring the Changing Sky
- Milky Way Structure and Formation

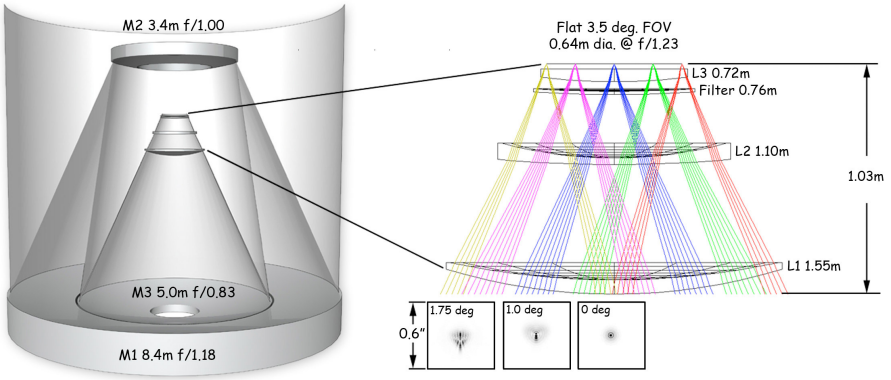


FIGURE 2.6: Paul-Baker optical design of LSST, science book

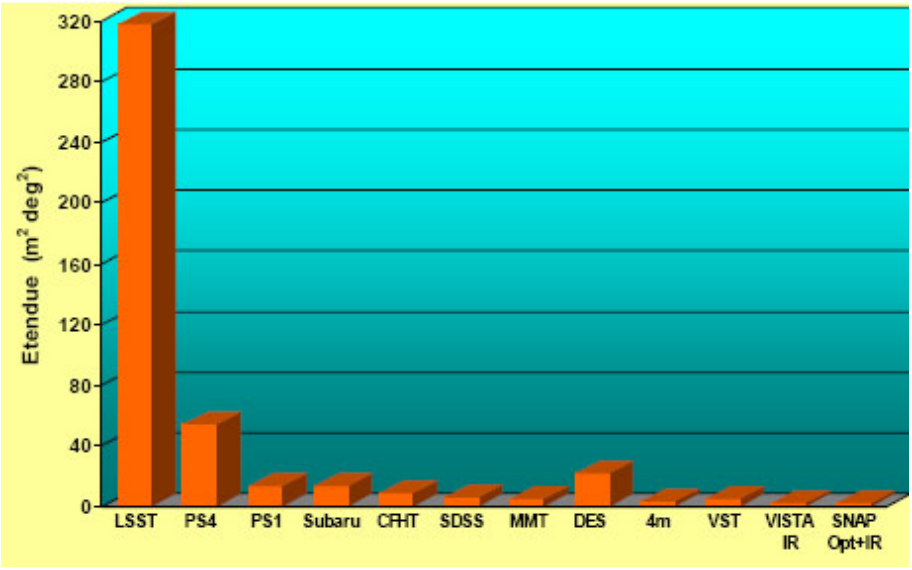


FIGURE 2.7: The etendue of different astronomical survey telescopes. Image credit: The Large Synoptic Survey Telescope Philip A. Pinto Steward Observatory University of Arizona for the LSST Collaboration Legacy Projects Workshop.

The current LSST baseline strategy will devote $\sim$ 90% of observing time to the main Wide-Fast-Deep survey mode. This survey region covers $\sim 20,000 deg^2$ of the sky and each field within it will be visited $\sim$ 200 times per filter where a visit consists of two 15 second exposures. The main survey region is limited in the north and south by an airmass of 1.4 (where airmass measures the quantity of air in-between celestial objects and the telescope). The remaining 10% of observing time will be spilt between several different mini observing strategies. Firstly, observations will be extended in the north and south, past the airmass limit but with a reduced cadence in order to observe a larger fraction of Near-Earth Asteroids in the Northern Ecliptic Spur and observe more of the Magellanic Clouds in the Southern Celestial Cap. Secondly, four or more Deep Drilling Fields will be chosen to receive $\sim$ 5 times more observations than fields in the main survey region, enabling observations to detect objects a magnitude fainter in the co-added images and to better characterise variable objects. Finally, LSST will also observe fields within the

Galactic Plane but with a reduced cadence - only 30 observations per filter per field. The reduced number of observations is because the high density of sources within the Galactic Plane is thought to make the images confusion limited, therefore there will not be extra depth with additional exposures. However, it also means less temporal characterisation of objects within the Galactic Plane - one of the most densely populated regions in the night sky. The current baseline observing strategy, baseline2018a, is shown in Figure 2.8 in equatorial coordinates with the colour denoting the number of observations each field will receive in all 6 filters over the full 10 year LSST lifetime.

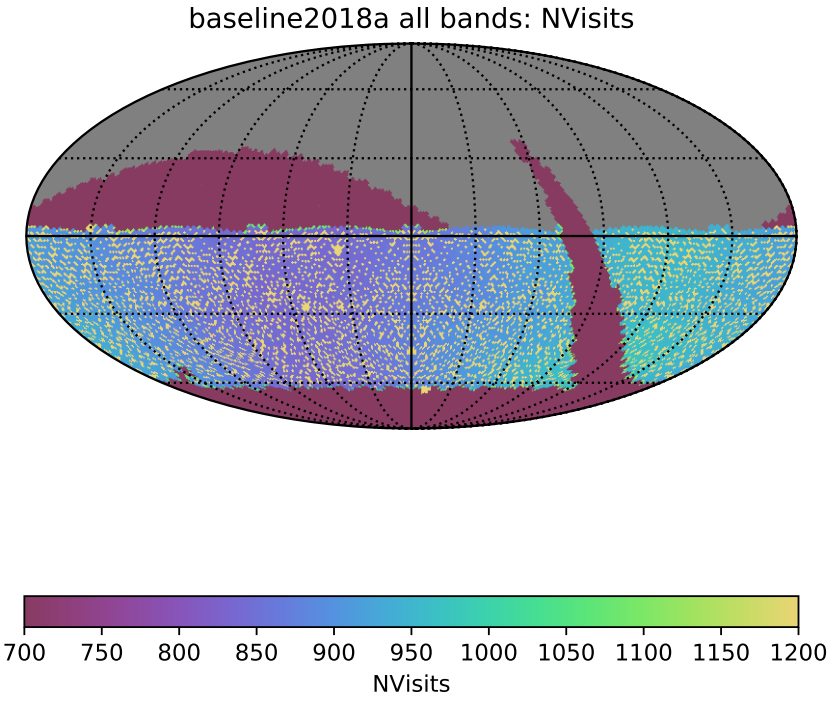


FIGURE 2.8: The current baseline LSST observing strategy - baseline2018a. The graph is shown in ecliptic coordinates and the colour corresponds to the number of observations each field will receive in all bands over the full 10 year LSST lifetime. Image credit: `http://astro-lsst-01.astro.washington.edu:8080/allMetricResults?runId=1`

Although LSST will be primarily tasked with completing the four main science goals, by virtue of its synoptic survey strategy, it will also be capable of completing many other secondary goals. Determining which observing strategy is best for the science cases of individual communities is the responsibility of the communities themselves. They are tasked to quantify the usefulness of LSST for their science case with a metric to have value between one and zero. These metrics are all compiled in the LSST observing strategy white paper (Marshall et al., 2017) and will be combined in order to provide the optimum observing strategy.

To aid the various scientific groups in determining the optimum observing strategy, a series of realistic simulators were made that emulate the operation and data collection of LSST. The simulators consist of: the Alert Simulator (Alert Sim), the Catalogue Simulator (CatSim), the Galaxy Simulator (GalSim), the Operations Simulator (OpSim)

and the Photon Simulator (PhoSim) where Alert Sim simulates the expected LSST alert stream; CatSim, GalSim and PhoSim simulate the creation of astronomical objects within realistic LSST images and OpSim generates a complete set of observational metadata for the ten-year simulated mission lifetime. Within the current version of the White Paper, there are three main observing strategies against which all metrics are evaluated - `minion_1016` (to be replaced by `baseline2018a` which is the same strategy but simulated using an updated OpSim version), `minion_1020` and `astro_sim_01_1004`. The first corresponds to the baseline observing strategy, the second observes all fields within the airmass limit with an even cadence and the final includes the Galactic Plane in the main Wide-Fast-Deep survey region.

One such example of a scientific community evaluating the usefulness of LSST observing strategies is presented by Strader et al. (2018), in which they challenge the reduced Galactic cadence of LSST with motivating examples of the potential scientific gain for Galactic objects with LSST. These use cases include enlarging the sample of known X-ray binaries, determining whether accreting white dwarfs are progenitors for Type Ia supernovae, and the study of Galactic supernovae. In Wells et al. (2017), they simulate pseudo-LSST light curves of eclipsing binaries and calculate the percentage of which are likely to have their periods recovered with LSST. They found that 71% of their sample had correctly recovered periods but noted that this decreased to 50% when considering periods longer than 10 days. The prospects for transiting planets studies with LSST has been assessed through a series of papers Lund et al. (2014), Jacklin et al. (2015), Jacklin et al. (2017) and Lund et al. (2018).

2.3.2 Gaia Space Observatory

The Gaia Space Observatory was tasked to find the positions, distances, space motions, brightnesses and astronomical parameters of stars within the Milky Way (Eyer et al., 2013). The mission aim was to construct a 3D catalogue of approximately 1 billion objects, $\sim$ 1% of the population of the Milky Way. Gaia utilises a broad photometric g band and is capable of observing objects up to magnitude $g = 20$. It observed each object an average of 70 times over the first five years alone.

Gaia measures many object parameters such as distances and space motions with the use of parallax. Parallax is the apparent displacement of an object when viewed from two different lines of sight and is measured as the angle of inclination between the two lines. Nearby objects have a larger parallax than farther away objects and it can therefore be used to measure distances. The relation between these two is shown in Equation 2.5, where d is the distance to the object in parsecs and p is the parallax angle in arc seconds.

$$d = \frac{1}{p} \tag{2.5}$$

Animals with two eyes use parallax in order to gain depth perception, the two viewing angles being the position of each eye. In the astronomical context, the motion of the Earth around the Sun is used to establish viewing angles, enabling the measurement of parallax and therefore distance to far away objects such as stars.

2.3.3 The Kepler Space Telescope

The Kepler Space Telescope (Borucki et al., 2010) was launched on the 7th of March 2009 and was designed to determine the frequency of Earth-sized planets in and near the habitable zone of Sun-like stars. Kepler's sole instrument was a photometer that continually monitored the same 115 $degree^2$ region of the Milky Way, shown in Figure 2.9. In order to keep the spacecraft's solar panels facing the Sun, Kepler would roll 4 times a year, resulting in observations of the same patch of the sky by four different CCDs. In order to detect transiting exoplanets, Kepler uses the transit photometry method (Koch et al., 2010) whereby the dimming of the star is observed as an orbiting planet passes between it and Earth.

The Kepler mission was originally planned for 3.5 years however higher than expected noise from the instruments and variability from Galactic stars meant that the mission goals were not met during the planned lifetime. The mission was therefore extended until 2016. However, the Kepler space telescope relied upon its four reaction wheels in order to maintain its precise pointing. On the 14th of July 2012 Kepler experienced failure in one of its reaction wheels. This happened again on the 11th of May 2013, and the Kepler Space Telescope lost the ability to point reliably at its targets, thereby ending the original Kepler mission (Cowen, 2013). Kepler continued to make observations again in 2014 with the new K2 mission which included the study of objects such as young open clusters, bright stars and supernovae where the decrease in photometric precision had less impact.

2.3.4 Instrumentation and Imaging

Kepler's only instrument, dubbed the photometer, was a one metre class Schmidt telescope with a 95-megapixel focal plane composed of 42 CCDs each of which were divided into two output channels (Koch et al., 2010). Schmidt optics were chosen to accommodate the large Kepler field of view while simultaneously providing excellent photometry. This did however come at the cost of the imaging quality of the telescope as the PSF of the Kepler telescope is dependant on the position of the pixel in each CCD. This necessitated the characterisation of the pixel response function (PRF) which is described by Bryson et al. (2010) as "a continuous representation that allows the pre-diction of a star's flux value on any pixel given the star's pixel position". The PRF was measured

during the commissioning phase of Kepler by observing bright unsaturated, uncrowded stars for 242 cadences of 14.7 minutes (Bryson et al., 2010).

Kepler took one exposure every 6.5 seconds which are combined onboard the spacecraft into either 1 or 30 minute coadded exposures. Due to the high cadence and number of pixels in the Kepler camera, only $\sim$ 6% of the data taken can be stored on the solid state recorder onboard the spacecraft. The imaging data chosen to be saved was in two different forms: targeted pixels and full frame images. The targeted pixels were in the form of postage stamp images centred on target objects of interest, the size of the stamps was dependant on the brightness of the object. The full frame images were 30 minute exposures of the entire Kepler field of view (FOV) taken approximately once per month. The purpose of these images was mainly for calibration of the other Kepler data.[1]

Although the full frame images were intended as calibration data, they do represent an evenly spaced data set of 53 images over a wield field of view with excellent photometery. Scientifically speaking, this dataset is relatively unexplored, however Montet et al. (2017) investigated the use of these images for identifying long term variability in Sun-like stars. They studied a sample of $\sim$ 4000 stars and found variability in roughly $\sim$ 10% of these. They also presented f3, a photometry package designed for aperture photometry within the Kepler Full Frame Images.

FIGURE 2.9: The Kepler field of view imposed on the Milky Way. Image credit:Carter Roberts.

[1] `https://keplerscience.arc.nasa.gov/data-products.html`

2.4 Handling Large Astronomical Datasets

2.4.1 Data Types & Format

As astronomers build survey telescopes that are capable of generating increasingly large datasets, tools must be created that are capable of storing, processing and analysing these complex datasets. Large relational databases are commonly used to store data products generated by survey telescopes such as Gaia (Brown et al. 2016,Brown et al. 2018), AllWISE (Cutri et al., 2014) and 2MASS (Cutri et al., 2003). Relational database management systems such as MySQL (MySQL, 2001), PostgreSQL (Momjian, 2001), SQLite (Owens, 2006) and Oracle Database (Loney, 2004) are example architectures for storing relational datasets and querying over it using the associated query language e.g. Structure Query Language (SQL). SIMBAD (Wenger et al., 2000) and Vizier (Ochsenbein et al., 2000) are examples of databases collated from the results produced by survey telescopes, research papers and other pre-existing astronomical catalogues. Consolidating the data into a single homogenous resource allows for powerful queries over the entire dataset.

Additionally, NoSQL databases are also being leveraged for the management of large astronomical data sets. An example being the China Near Earth Object Survey Telescope (CNEOST) Xin (2014) which collects over 10 TB of imaging data per year and has implemented a non-relational database to manage it. Another example is the New Vacuum Solar Telescope (Liu et al., 2014) which collects up to 120,000 FITS images per day. Originally, the data was stored in MySQL, a traditional relational database. Liu et al. (2016) investigated the relative performance of this database when compared to a NoSQL database they named "Fastbit". They found that Fastbit had a query repsonse time which was 15 times faster than the previous database system whilst also meeting all the previous requirements of the NVST.

The International Virtual Observatory Alliance[2] was set up in 2002 in an effort to standardise the dissemination of astronomical data. Their goal is to enable all astronomical datasets and other resources to act as a seamless whole. To achieve this, they have published data standards for the discovery, formatting and delivery of astronomical data in papers such as Hanisch (2006).

2.4.2 Astronomical Workflows

The analysis of complex datasets is increasingly achieved through scientific workflows composed of a series of computations, each of which may be composed of thousands of steps. A scientific workflow can be thought of as the broad term for any sets of processes

[2] `http://ivoa.net/`

that generate results from the raw data. Scientific workflows are currently the paradigm for analysing, visualising and managing large scientific datasets. This is in part due to their ease of use, wide scope for customisation and proficiency at performing repetitive tasks.

The Sloan Digital Sky Survey (SDSS) employed a variety of workflows in the SDSS data processing factory in order to produce the multitude of data products released by it; details of the pipelines can be found in Stoughton et al. (2002). The LSST has a dedicated data management team to develop pipelines that can process the large data quantities and release transient alerts in real time (Juric et al., 2015). With the scale of LSST data collection, this would not be possible without the automated processing enabled through scientific workflows. A prototype of the LSST data processing pipeline was also used as the basis for data management system of Subaru Telescope's Hyper Suprime-Cam (Bosch et al., 2017).

The challenges facing the creation of scientific workflows have been investigated by Gil et al. (2007) following a workshop on the subject. In their paper, they discuss the growing challenges of reproducibility in scientific workflows when they are distributed among both systems and scientists which fragments the documentation of the processing. In addition, they discuss the implications of the evolution of technology on the reproducibility of scientific workflows and whether re-execution will produce the same results when performed on different platforms.

2.5 Provenance

Provenance is necessary for the reproducibility of scientific workflows. The term provenance originated in the art world where it was used to determine the ownership and authorship of a work of art, in order to establish its value. With regards to astronomical data, it is the documentation of the data collection and processing. Traditionally, provenance was kept in the form of laboratory books in order to make scientific results reproducible. Most of the provenance about astronomical images is stored in the fits header, this includes information such as when it was taken, with which filters and how long the exposure was.

The World Wide Web Consortium (W3C) formalised PROV, a flexible specification for how to express provenance records. Their definition of provenance is "information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness"[3]. Entities refer to things whether they are physical, digital or conceptual. Examples of entities within astronomy would be telescopes, images or airmass. Activities are processes which

[3]https://www.w3.org/TR/prov-overview/

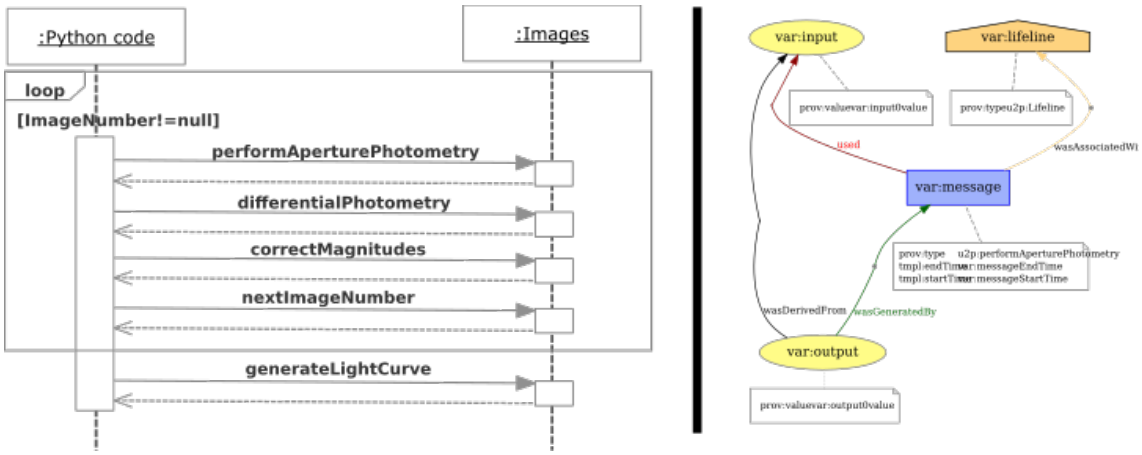


FIGURE 2.10: The left hand side is a UML sequence diagram depicting a simplified version of the differential photometry process. The right-hand side is a PROV template generated from *performAperturePhotometry*.

generate, consume or alter entities. For example, the process of taking an astronomical image is an activity which uses the telescope to produce the image. Additionally, the airmass would also be consumed by the process and impact the quality of the image. An agent is essentially an entity that is assigned responsibility for an activity taking place. The agent in the astronomical example is the operator of the telescope which may be a person or if it is a robotic telescope, the software that runs it. An example PROV diagram is shown in Figure 2.10 as well as the associated toy workflow depicted as a UML sequence diagram.

2.5.1 Template Provenance

Moreau et al. (2017) present a templating system to enable the creation of provenance compatible with the PROV standard of the W3C. This system deconstructs the provenance into two components: templates which are arranged by the designer to produce the topology of the provenance and contain variables as place holders for values; bindings containing a data structure and values which correspond to specific templates. The provenance is then created by using an expansion algorithm on these two components. One of the main advantages of this templating provenance system is that a single set of templates is required for an application, regardless of the number of executions. The runtime information of each execution is recorded in the bindings which can then be expanded into the full provenance on demand. As Moreau et al. (2017) found that the size of these bindings was typically $\sim$ 40% of the size of the full provenance, the templating system can therefore offer significant reductions in the storage cost of provenance recording. Complementary to template provenance, Sáenz-Adán et al. (2018) present UML2PROV - software designed to automate the creation of provenance templates. As

an input, UML2PROV takes a diagram of the workflow as described by a unified modelling language (UML) diagram and it produces a set of provenance templates which describe the workflow.

2.5.2 Provenance from Scientific Workflows

In order to address the challenges outlined by Gil et al. (2007) surrounding reproducibility and scientific workflows, scientific workflow management systems such as myGrid/-Taverna (Hull et al., 2006), Kepler (Altintas et al., 2004), VisTrails (Callahan et al., 2006) and Chimera (Foster et al., 2002) have been developed to manage, execute and visualise scientific workflows across hardware and software environments. Coordination of the workflow over one of these management systems facilitates the automated and standardised collection of provenance.

Many scientists have adopted the use of scripting languages rather than working within scientific workflow systems due to their relative proficiency in them. Fortunately for the modern astronomer, tools such as YesWorkflow (McPhillips et al., 2015) and NoWorkflow (Murta et al., 2014) have been developed to automate the generation of provenance from these scripts.

Provenance management systems have also been designed for the workflows of large scale survey telescopes, one notable example being LSST [4] (Becla et al., 2006). The motivations for LSST's provenance model extend beyond reproducibility, the pipeline centric model (Groth et al., 2010) that will be implemented also offers reductions in the computational storage required. This model stores information on what processes were performed, on which entities and with which versions. With this information, all intermediate data products can be discarded and regenerated at a later date, only the initial data and provenance are required to be stored which is comparably smaller.

2.6 The Five Characteristics of Good Quality Data

As astronomical survey telescopes become able to amass larger and larger datasets, small improvements on the quality of these datasets can offer large scientific gain. Before improving the quality of data, one must first define what is good quality data for their purpose. Data quality may refer to many different characteristics, Wand and Wang (1996) created a summary of all the cited data quality metrics and their occurences from the literature review by Wang et al. (1993), this is shown in Table 2.1.

Not only is there a great variety in categories for quality metrics, definitions for the same quality metric can also be inconsistent across the literature. For example, Kriebel

[4] `https://github.com/lsst-dm/provenance_proto/blob/master/Provenance.md`

Dimension	Cited	Dimension	Cited	Dimension	Cited
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Freedom from Bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of Detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2	-	-

TABLE 2.1: The types and number of citations of quality metric cited within literature. Credit: Wand and Wang (1996).

et al. (1979) define accuracy to be “The correctness of the output information", whereas Ballou and Pazer (1985) define it as “the recorded value is in conformity with the actual value". Ballou and Pazer (1985) as define the timeliness as “the recorded value is not out of date", whereas Larsen et al. (2009) define it to be the time taken for the data to be recorded, within their domain.

Due to the inconsistent nature of the quality definitions within the literature, five categories for quality metrics were chosen and motivated with relevant astronomical examples. In order to arrive at the five chosen categories, the metrics displayed in Table 2.1 were first refined by condensing quality metrics which had simlar meanings. Secondly, all metrics which did not have a definition within the literature that was relevant to astronomical datasets were removed. The final condition was that the quality metric needed to have a use case in the context of astronomical data within the literature. The chosen quality metrics:

- Accuracy and Precision
- Reliability
- Timeliness and Relevance
- Completeness
- Availability and Accessibility

2.6.1 Accuracy and Precision

Accuracy is defined as the degree to which the data resembles the “true" value of what the data represents. Although the definition of accuracy is not consistent throughout the literature, this definition conforms to the consensus for the definition discussed by. With regards to astronomical data, finding absolute truths in the properties of astronomical

objects is an impossibility. However, constraints can be placed on the measurements of these properties which define how far from the true value they are likely to deviate. Tightening these constraints can be considered an improvement in accuracy. Furthermore, knowing the degree of confidence in the results enables deduction of which physical mechanisms are feasible explanations for the observed properties.

Precision refers to the repeatability of the measurement of the data, the definition is consistent with that by Bailey and Pearson (1983) who define it as the variability of the output. Very precise data is characterised by measurements of the same quantity which have very consistent values. In astronomical data, an example for precise data would be consistent measurements of the brightness of a star when measuring it in the same way each time. The precision can be quantified via statistical measurements of variation, such as the standard deviation. The precision is therefore also very tightly related to the reliability of the measurement.

Data does not have to be accurate to be precise, nor precise to be accurate. Figure 2.11 illustrates this point with arrows being fired at a target. Arrows close to the target are accurate and those with little spatial variation are precise.

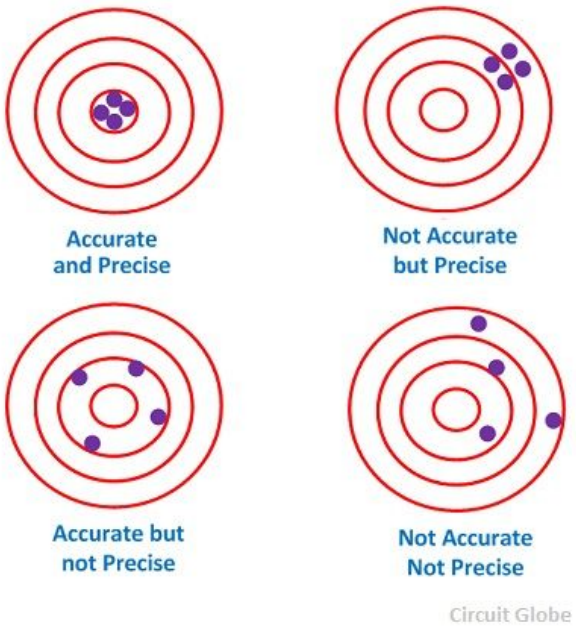


FIGURE 2.11: This figure depicts arrows being fired at a target. Arrows close to the centre have a high accuracy while arrows with little spatial variation have a high precision. Image credit: `https://circuitglobe.com/accuracy-and-precision.html`

2.6.2 Reliability

The defintion chosen for reliability was based on the work by Hansen (1983) which defines it to be a probability that something will not fail. An example of the assessment of this quality metric within astronomy was investigated by Zwaan et al. (2004) by calculating

the reliability of H_I Parkes All Sky Survey (HIPASS) catalogue which is an extra-Galactic blind H_I 21cm emission line survey within the southern hemisphere. This was achieved with the reobservation of subsamples of the catalogue with the Parkes telescope and determining the fraction of the previously found objects could be recovered and the likely fraction of spurious detections. They found the average reliability of the HIPASS catalogue was $\sim 95\%$. As the reliability is closely related to the signal to noise ratio, they also found that the reliability increased with increasing source brightness.

2.6.3 Timeliness and Relevance

This category refers to how the data reacts to or complements other external factors or datasets. The definitions used for each were from Halloran et al. (1978) where they define relevance as describing the usefulness of the dataset is to the world at large. Halloran et al. (1978) defined the timeliness as describing the usefulness of the data the date of data creation as well as whether the time taken to generate the data was satisfactory. Examples of these metrics occur when considering the construction of survey telescopes. With commissioning of LSST on the horizon, the data from an additional optical survey telescope would not be particularly relevant nor timely. Conversely, follow-up telescopes to compliment LSST observations would provide data that is both. This motivated the proposed use of telescopes such as the Southern Astrophysical Research Telescope (SOAR) to be used as follow up for the LSST transient alert stream (Elias and Briceño, 2016).

2.6.4 Completeness

Completeness refers to the quantity of the intended data space that is contained within the dataset. Incomplete datasets may contain unrepresentative populations of different regions of the total data landscape and bias the results. For this reason, this data quality category is important when performing population studies in astronomy. Zwaan et al. (2004) also investigated the completeness of the HIPASS telescope by inserting a number of simulated point sources to the images and attempting to recover them. They again found that the completeness was closely linked to the brightnesses of the objects being recovered and stated a 99% completeness for objects with a peak flux of 84 mJy and an integrated flux of $9.4 Jykms^{-1}$.

2.6.5 Availability and Accessibility

Availability and Accessibility describes the ease of access for those that need to use the data. The definition used in this thesis is similar to the one outlined by Kumar and Segev (1988) as the probability that both a read and write operation can be performed on the

data. Much astronomical data is available in the public domain and easily accessible, usually downloadable via a website. Data rights for large survey telescopes such as LSST depend on contributions to the project and will be unavailable to those who did not contribute. In addition, the data volume of LSST can make it inaccessible to users without the infrastructure in place required to process it. Therefore, LSST will do much of the basic photometry before disseminating the results to the public [5].

2.7 Summary

The large data collecting capabilities of astronomical survey telescopes make them a powerful tool in the hands of astronomers. This chapter outlined three notable examples in the Large Synoptic Survey Telescope, the Kepler Space Telescope, and the Gaia Space Telescope. The wide field of view and systematic monitoring of the sky characteristic of these telescopes makes them the ideal platform from which to discover and study variable and periodic astronomical objects such as long period variables, X-ray binaries, and cataclysmic variables.

The size of these datasets necessitates the automation of the data processing. This chapter outlined some of the most common computational techniques for processing astronomical data and introduced the astronomical workflow which typically utilise a number of these techniques to transform the raw data into the desired scientific result. The choice of techniques used during the execution of the workflow as well as which versions have implications for the results produced. Provenance was therefore introduced as a means of documenting the data processing to ensure that the results are reliable and reproducible.

Finally, when handling large datasets, small improvements in the data quality have the potential to offer large scientific gain. Therefore, five metrics for the quality of astronomical data were selected from the literature and the choice of each was motivated by relevant astronomical use cases. The following chapter outlines a model to be applied to astronomical workflows in order to improve the data quality of the results they produce with regards to these metrics.

[5] `https://www.lsst.org/about/dm/data-products`

Chapter 3

An Approach for Reasoning over Workflow Versions

Computational workflows are becoming a valuable tool to aid astronomers in processing the increasingly large datasets which they can create. This is in part due to the workflows being very customisable, quick to build and that they may make use of prebuilt processors. Using these pre-made processors such as SExtractor (Bertin and Arnouts, 1996), HOTPANTS (Becker, 2015) or DAOPHOT (Stetson, 1987) saves on development time but astronomers must be careful not to treat them all as black boxes as they usually have a large scope for customisation and the quality of the output is often heavily dependant on that of the input.

The determination of how to properly customise these tools within astronomy is typically a qualitative process, achieved through visual inspection and trial and improvement. Although experienced astronomers have developed an intuition for this method, it lacks any quantitative measure of the performance of their chosen workflow configuration as well as knowledge of whether it was the best choice. To address this issue, this chapter outlines an approach which can be applied to the workflows in order to reason over the relevant workflow versions to find the best configuration for the specified purpose. After defining the approach, the chapter then equates the problem of finding the best workflow version to the travelling salesman problem and investigates the applicability of the solutions of the latter to the former. To test the approach, it was applied to three separate astronomical use cases in Chapters 4, 5, and 6 in order to improve the quality of the data produced by each.

3.1 Problem Formulation

For ease of explanation, we will explain these concepts via the example of breakfast. A good quality breakfast can be interpreted in much the same way as good quality data, timeliness being an obvious example. Additionally, the accuracy of a breakfast may refer to how closely the recipe was followed or how close it resembles the intended breakfast. The completeness of a breakfast may refer to how much of the required nutrition the breakfast contained, for example calories, protein, $\frac{x}{5\,a\,day}$ etc.

The workflow for breakfast consists of two main processors: to choose the breakfast and to prepare it. The first processor has three parameters within the parameter space - cereal, scrambled eggs and a full English. This first choice also decides the choice of the second parameter space: add cereal, add milk; scramble the eggs; fry bacon, eggs, heat beans, toast bread. The output from each workflow is a breakfast and the decisions made around this workflow will decide the quality of the breakfast with regards to the different quality metrics. Figure 3.1 depicts this workflow and its components.

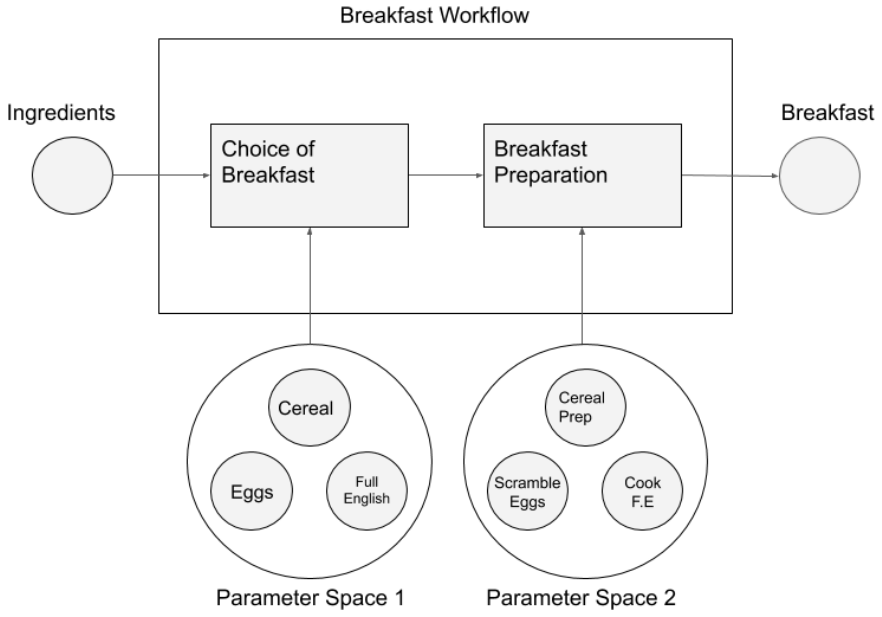


FIGURE 3.1: Schematic of the breakfast workflow.

The timeliness and relevance of the breakfast is important when there is a restricted amount of time allocated for breakfast preparation. The breakfast cereal performs very well within this metric as immediately after adding the milk, the breakfast can be considered complete. The full English has many components which each need preparing, decreasing the timeliness. Additionally, the relevance of the breakfast drastically decreases if the time required is longer than the allocated preparation time. However, the multiple components increase the quality of other metrics such as the completeness of the nutrition in the breakfast. The contradictory needs of these quality metrics result in

a mutual exclusivity between the timeliness and completeness of the breakfast. Therefore, the decisions made on which regions of parameter space are most suitable for the workflow need to be informed with the relative importance of the various quality metrics.

This same mutual exclusivity of optimum values for quality metrics is a common occurrence in scientific workflows also. This exemplifies the importance of defining the relevant data quality metrics for each application. A commonplace example in astronomy is completeness vs accuracy. It should be noted that the definition of accuracy used here is often referred to as the robustness or reliability within the astronomical community. When finding astronomical objects, the strictest requirements on the search will be the most accurate as it will detect the fewest false positives. On the other hand, it will likely reject many true events and the high accuracy will come at the cost of the completeness. The reverse is also true as a higher completeness will likely include a higher number of false positives. Within this example, it is necessary to find a balance between the two that will depend on the application. For example, when searching for a gravitational wave counterpart, one may be willing to detect 1000s of false positives and sift through them for the sake of a single object. However, for events that are not quite so rare such as supernovae, this kind of analysis would be prohibitively cumbersome and a lower completeness would be willingly sacrificed for a higher accuracy.

An additional complication is that the intra-processor interaction may also impact the data quality, depending on which metric is being assessed. For a monotonically increasing metric such as timeliness, the highest quality workflow will consist of the set of processors which take the least time. On the other hand, suppose we go back to the breakfast example and decide to make a breakfast sandwich (bear with me) and the quality metric is the best tasting sandwich. Our free parameters are bread, spread and filling. Although each stand-alone best tasting component may be white bread, nutella, and ham, the best overall tasting sandwich is probably not a nutella and ham on white. This can also be the case for astronomical data processing: the outputs produced by a processor may have the highest stand-alone quality, however its interaction with the rest of the workflow may not produce the highest quality final output.

3.2 Definitions

3.2.1 Processor

A processor is a subcomponent of a workflow which may consume parameters, act on one or more input datasets and produce one or more output datasets. In the context of astronomical workflows, a processor is usually a set of instructions written in a coding language intended to clean, extract or otherwise manipulate the input dataset. The output produced is in the form of a dataset which may correspond to the final dataset or

to an input to be consumed by another processor. This definition is consistent with the definition for processor from the workflow management system Taverna (Missier et al., 2010).

Definition 1. *Processor - A Processor, P, is a set of instructions which acts on Input Datasets, I, and consume Parameters, PM,in order to produce a set of outputs.*

In the context of our breakfast example, processor 1 is the decision of which breakfast to make and processor 2 is the corresponding preparation.

3.2.2 Parameter

Parameters are customisable inputs in processors that alter the outputs produced by a processor. Parameters are there to allow the same processors to be specialised for different datasets and different quality metrics. When a processor has multiple customisable parameters, each parameter may impact on how the other parameters impact the function of the workflow, therefore it is necessary to consider not just the applicability of each parameter to the situation but also the interaction of the parameters with each other.

Definition 2. *Parameters, PM, are key-attribute pairs that modify the function of a Processor, P.*

The parameters within the breakfast example are cereal, scrambled eggs and full English for processor 1 and the different methods of preparation for processor 2.

3.2.3 Parameter Space

The parameter space is the set of all potential values that a parameter may take in a given workflow. Each parameter may have its own parameter space but each parameter space must consist of at least two values. If a parameter has a single potential value within every conceivable evaluation of the workflow then it is just considered a parameter with no associated parameter space.

Definition 3. *Parameter Space, PS is the set containing all N potential attributes for a single Parameter key. $PS = PM_0...PM_N$*

The parameter space for processor 1 represents the space encompassing all of the aforementioned parameters for processor 1 and the same is true for the parameter space of processor 2.

3.2.4 Objects

Any given component of a single workflow may be considered to be an object, it denotes the broad class that inputs, outputs, processors and parameters belong to.

Definition 4. *Object O, may refer to either a Parameter, PM, Processor, P, Input, I or Output, O. It the encompassing term for any component contained within one Workflow.*

The objects within the breakfast example include the processors, parameters and parameter spaces related to both processor 1 and 2, i.e. the ingredients, decisions, preparation etc.

3.2.5 Workflow

The execution of a workflow consumes one or more inputs, acts on them through one or more processors which consume parameters and produce one or more outputs. Workflows are sequentially ordered such that the output of one processor becomes the input of the next. A simple model of a workflow made up of customisable processors components is shown in Figure 3.2. In this workflow the inputs are consumed by the first processor which can take either P1 or P2 as a parameter. The output produced by this processor is then consumed as an input to the second processor which also has two possible parameters (P1,P2) finally, this second processor produces the final output.

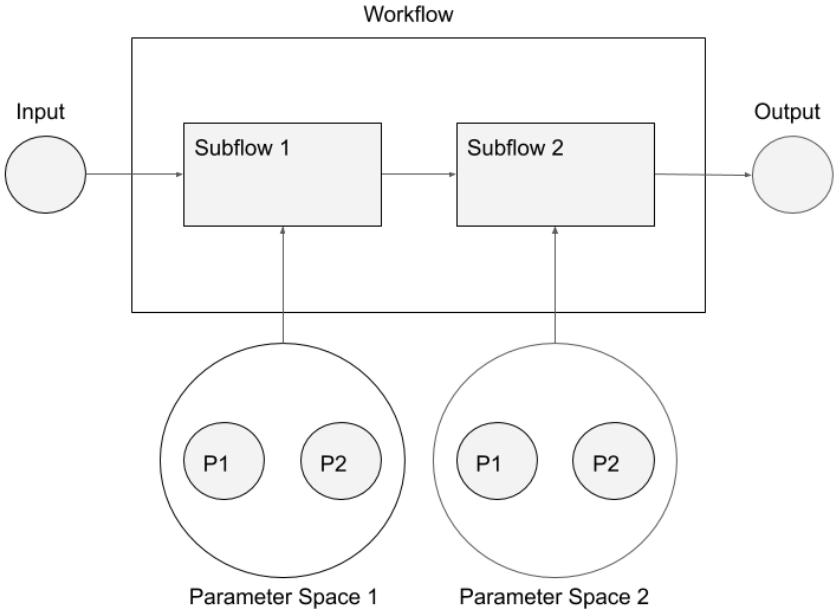


FIGURE 3.2: A diagram depicting an abstract workflow which consists of two processors, each with the option to use either parameter 1 or 2.

Definition 5. *Workflow - We base our definition for a Workflow on Ellkvist et al. (2009), as a set of one or more partially ordered Processors whose inputs include both static Parameters and the results of earlier computations. Edges exist in a workflow connecting Datasets, Parameters and Processors. The Processors are sequentially ordered such that the Output of Processor 1 becomes the Input of Processor 2 and so on.*

A workflow for our motivating example consists of one complete breakfast making a version where all decisions have been made and one parameter for each parameter space has been chosen - such as the decision to make and prepare cereal.

3.2.6 Edge

Edges in workflows represent the relationship between parameters, processors, inputs and outputs. The direction of the edge represents the direction of information flow.

Definition 6. *Edge - An Edge, E, is an ordered connection between a two distinct Objects, O.*

Edges are an abstract concept for the relation between workflow objects, the analogous concept in breakfast would be the transfer of the current state of breakfast from one processor to the next, i.e. removing the pancake mix from the bowl and into the frying pan.

3.2.7 Actual Quality Result

The actual quality result represents the quality of the final output produced by one single discrete workflow.

Definition 7. *Actual Quality Result - The Actual Quality Result, AR, is the quality of the final Output produced by the Workflow.*

Within our breakfast example, the quality of the produced breakfast within a specified quality metric is the actual quality result.

3.2.8 Expected Quality Result

The expected quality result is a baseline measure of quality for other workflows within a workflow space. The main use is in conjunction with the actual quality result in order to quantify the change in quality of output between workflows.

Definition 8. *Expected Quality Result - The Expected Quality Result, ER, is an example Output produced which represents the same object the Actual Quality Result. It can be either the ideal result expected with optimum Workflow settings or a baseline result produced by alternative solutions to the same problem.*

An example for the expected quality result within the breakfast workflow is the perfect breakfast. This breakfast takes zero time to make and is entirely complete in nutrition. Although this breakfast may not exist, its known perfect quality may still be used as a reference for measuring the other breakfast workflows.

3.2.9 Utility Function

The utility function is used to measure the difference in a quality metric between variations of workflows. It is a function of the expected and actual quality results, however the exact form depends on the relevant quality metrics and their relative importance.

Definition 9. *Utility Function - A Utility function, U_{qm}, is defined as a relative measure of Actual Quality Result AR and the Expected Result ER.*

Producing the utility function for the example would involve combining the quality of the breakfast produced by a single workflow and the perfect quality breakfast. To know the exact details for how these should be combined, a detailed analysis of the relative importance of the various quality metrics would be required.

3.2.10 Provenance

The provenance is the documentation of the processing. The record of the processing by the workflow as well as the relevant inputs/outputs/parameters are described by the provenance. This is used to relate the utility function to each workflow in a workflow space. The provenance is also required to measure the cost of each workflow.

Definition 10. *Provenance is documentation that describes the processing of a Workflow, W. It describes which processors were performed, on which objects and what versions of the processors were used.*

The provenance of the breakfast example would describe things such as the decision of which breakfast to make as well as the preparation process of the breakfast.

3.3 Problem Statements

The quality of the output produced by a workflow is dependent on factors such as: inputs, parameters, workflow configuration, the processors of the workflow and which quality metric the user decides defines good quality data. The number of free parameters which can impact the quality of the data makes the evaluation of every discrete workflow very computationally expensive. Figure 3.3 demonstrates the size of the workflow space when there are only two processors with two different parameters each. The total number of workflows in the space increases rapidly with an increase in either processors or parameters.

There is currently a lot of documentation surrounding the creation of breakfasts which each satisfy certain quality metrics. If someone wanted to improve the quality of their favourite breakfast they would first need to define:

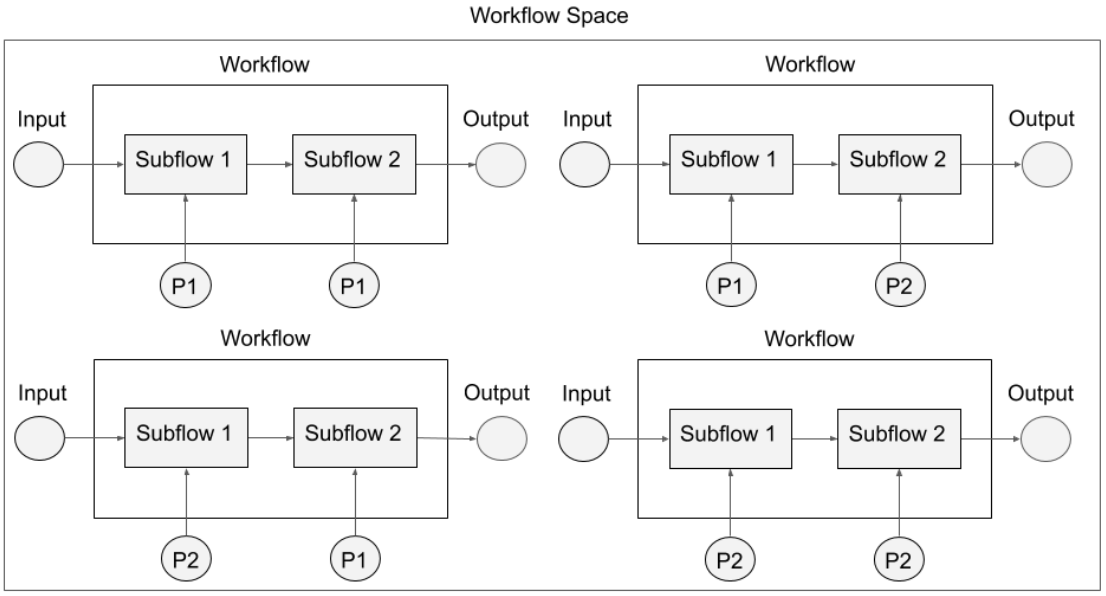


FIGURE 3.3: A diagram depicting the total workflow space for the example shown in Figure 3.2 where there are two processors each with two choices for parameters.

- The recipe
- The ingredients (I)
- The alterations to the recipe/ingredients that are likely to impact quality;
- The metric they use for good quality breakfast (AR)
- A baseline breakfast quality for comparison (ER)

In order to determine which breakfast has the best quality, this person could create a breakfast for each combination of ingredients, cooking times and kitchen equipment that they thought could have a positive impact on the quality of their breakfast (analogous to evaluating all workflows within the outlined workflow space). However this kind evaluation is too expensive to realistically carry out.

What the user needs to define is much the same when trying to determine which workflow creates the best quality data:

- The workflow
- The inputs, I
- The parameter spaces
- The relevant quality metrics

- The expected quality result ER

Once these quantities have been defined, the goal is to find the set of PMs which maximise the utility function. Figure 3.4 depicts the potential variations of the workflow and different paths that can be taken through the parameter space. Each PS must have exactly one PM selected. The order of PM selection is determined by the order of the processors within the workflow and depicted by the direction of the edges. The length of the edges represent the quality of the intermediate output produced by that processor-to-processor relation, the total length of all edges for one workflow represents the quality of the final output and in each case, shorter edges represent higher quality.

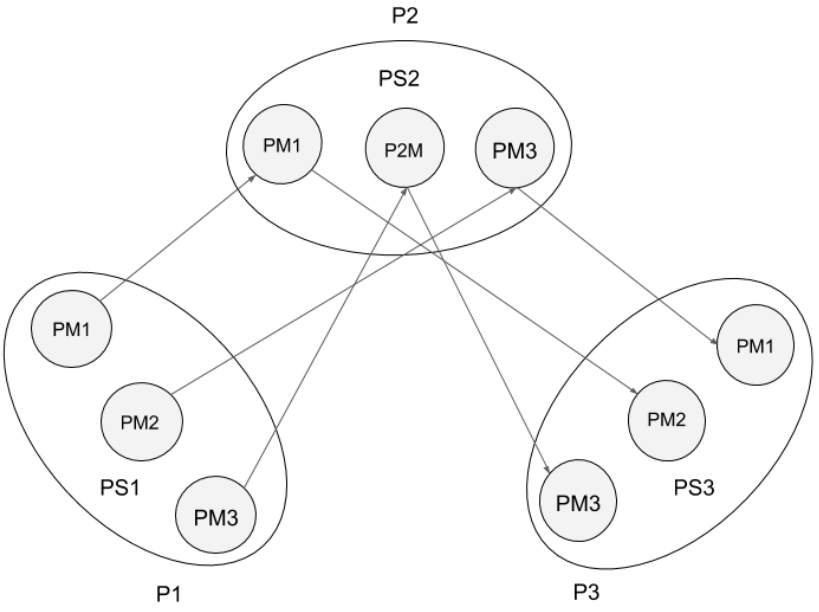


FIGURE 3.4: A diagram depicting three processors $(P1, P2, P3)$, three associated parameter spaces $(PS1, PS2, PS3)$ which each contain three parameters $PM1, PM2, PM3$. In a single workflow run, each parameter space must be visited exactly once with one value chosen each time. The quality of the output of the workflow is symbolised by the length of the line, shorter being of better quality. The goal is to find the choice of PM within each PS which maximises UF.

3.3.1 Theoretical Analysis of the Problem

The aim of this section is to prove that the instantiation of the approach is analogous to the travelling salesman problem which belongs to the set of NP-complete problems. If this analogy can be made, it then also proves that both problems must be NP-complete. NP (nondeterministic polynomial time) is a complexity class used to classify decision problems, it denotes the set of decision problems for which the problem instances have proofs in polynomial time. A problem is NP-hard if everything in NP can be transformed

into it in polynomial time meaning that it is at least as hard to solve as the hardest problems in NP. A problem is said to be NP-complete if it is both NP and NP-hard.

The original travelling salesman problem (TSP) asks the following question: "Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city and returns to the original city". This is a well known, NP-hard problem in combinatorial optimisation. The generalised travelling salesman problem (GTSP), presented in Laporte and Nobert (1983), has the addition of subsets of cities and at least one per subset must be visited by the salesman. The TSP can be thought of as the special case in the GTSP where all cities are confined to the same subset, therefore both are NP-hard.

The mathematical definition of the GTSP was taken from Pintea et al. (2007), it is as follows: $G = (V, E)$ is a n-node undirected graph with edges associated with non-negative costs.

$V_1, ..., V_p$ is a partition of V into p subsets named clusters, such that $V = V_1 \cup V_2 \cup ... \cup V_p$ and $V_l \cap V_k = \emptyset$ for all $l, k \in \{1, ..., p\}$. We denote the cost of an edge $e = i, j \in E$ by c_{ij}.

The solution to the GTSP is the minimum cost tour H which spans a subset of nodes such that H contains exactly one node from each cluster $V_i \in 1, ..., p$. Finding this solution involves two related decisions: choosing a node subset $S \subseteq V$, such that $\mid S \cap V_k \mid = 1$, for all $k = 1, ..., p$ and determining the minimum cost cycle in the subgraph of G induced by S.

The TSP may also be symmetric or asymmetric. It may be considered symmetric if $c(i, j) = c(j, i)$ for all $i, j \in V$, where c is the cost function to the edges of G. The problem is considered asymmetric otherwise.

The evaluation of the approach requires the ability to change parameters within the chosen parameter spaces, measure the quality of output that each discrete workflow produced and compare this to a baseline output. This maps to the GTSP as follows:

Proof. *Parameters* (PM) correspond to the nodes within the GTSP.
The Parameter Spaces corresponds to V the partitions of the nodes. Only one node per partition must be chosen, just as there can only be one Parameter (PM) chosen from each Parameter Space (Ps).
Actual Quality Result (AR) is the measure of quality of an individual workflow and therefore analogous to the total cost function of the edges between the nodes in graph G.
$AR = \sum_{(i,j) \in G} c(i, j)$.
Expected Quality Result (ER) as with AR, ER represents the total cost function of the edges between the chosen nodes but this time within the baseline graph G_{base}.

$ER = \sum_{(i,j) \in G_{base}} c(i,j)$

This baseline graph may represent either the initial starting parameters or the graph which minimises $\sum_{(i,j) \in G} c(i,j)$.

Utility Function (U_{qm}) is exactly the same for both the workflow and the GTSP, it is the comparison of AR by ER, which provides a relative measure of quality of the graph G with respect to the baseline. □

Since the GTSP is NP-complete and the approach can be thought of as an instance of the GTSP, the approach also belongs to the set of NP-complete problems. The equivalence of the approach to the GTSP allows the definition of the first problem:

Problem 1 - Assuming the knowledge of the length of all Edges between every possible combination of Parameters within each Parameter Space, determine the Workflow which minimises the Edge length, therefore maximising the Utility Function.

3.3.2 Implementation

When the number of discrete versions of the workflow is small, all can be evaluated and the best version can be easily determined. As the number of altered parameters increases, the number of discrete workflows rapidly increases and evaluation of all discrete workflows can become an impossibility. As this is common limitation of NP-hard problems, many algorithms have been devised to quickly arrive at an approximate optimum solution. Here, we discuss potential methods for reducing the computational intensity of full workflow space evaluation. This leads to the second and third problems:

Problem 2 - With no knowledge of the length of individual Edges, can the Workflow which maximises the Utility Function be estimated?

Problem 3 - Determine whether the estimated Utility Function (UF_e) is equivalent to the global Utility Function (UF_G) and if not, what confidence can we have in UF_e.

Evaluating all possible configurations of the workflow is the most concrete way to find the one that produces the best quality workflow but this may become unrealistic when there are many free parameters and each workflow takes a non-negligible amount of time to run. The approach currently includes one way to reduce this processing cost - the user identifies relevant parameters to alter. This initial assumption may introduce a bias to the final results but reduces the workflow space to only encompass parameters that are expected to have a reasonable effect on the data quality in the chosen metric. In many cases, the parameters may be numeric within a specified range (e.g. temperature on an oven) the user also defines the granularity with which this parameter will be tested.

Another possibility for reducing this cost is to measure the quality of the outputs produced by each processor individually, as opposed to that produced by the workflow as a whole. This can drastically reduce the total number of discrete workflows if there are multiple processors with free parameters, however as previously discussed, it is not necessarily the case that processor configurations that stand-alone produce the highest quality output will do the same when joined as a workflow.

The methods discussed here all rely upon assumptions and simplifications in order to reduce the total time taken to evaluate **all** configurations of the workflow. There are however additional ways to estimate the highest quality workflow within a workflow space. An example of one such way is discussed in the following section.

There are many algorithms for the GTSP in particular such as the work by Fischetti et al. (1997) who present a branch-and-cut algorithm for the exact solution to the symmetric GTSP. Noon and Bean (1991) present a Lagrangian approach for the symmetric GTSP. Renaud and Boctor (1998) present GI3 (Generalized Initialization, Insertion and Improvement), a composite heuristic for the solution of the GTSP. Finally, Pintea et al. (2007) present a solution to the GTSP designed to mimic the behaviour of ant colonies. The applicability of these solutions is reliant on the knowledge of city to city distances which are not always possible with our specialisation of the GTSP. However, solutions such as hill climbing (Jacobson et al., 2006), simulated annealing (Wang et al., 2015), and genetic algorithms (Silberholz and Golden, 2007) may still be applied to this regime.

The hill climbing algorithm was chosen to estimate the optima of our workflow. When applied, it made incremental changes to parameters and tested whether the quality of the output has improved. If it had, another incremental change in the same direction would be made until no positive change in quality was found. If the initial change did not make a quality improvement, then a change in the other direction was tested and the same procedure was followed as for the initial change. When the change in quality stagnated, another parameter was chosen for the same processor. This algorithm is repeated until a full cycle of changes to parameters had no positive effect on the quality. One potential caveat to the hill climbing method is the possibility to find local optima, not detect any quality improvement and end without necessarily finding the global optima. One way to mitigate this effect is at every incremental change made, also test the quality of a randomised set of parameters. If the quality of the output produced using these random parameters is higher than the current quality, then the algorithm will start hill climbing from this new maximum. This is shown in Algorithm 3.3.2 where X is a variable used to represent the direction of the hill climb.

Algorithm 1 Hill climbing algorithm in pseudo code. The variables here are as follows: PS denotes the parameter space; PS_n corresponds to the nth region in the parameter space; PM represents the parameters; PM_m is the mth parameter;UF is the utility function; UF_{PS_n,PM_m} is the utility function for the parameter space PS_n using the parameter PM_m; X is the coefficient which determines the direction of the hill climb.

```
for PS_n in {PS_1,...PS_p} do
  for PM_m in PS_n do
    for X in {−1, 1} do
      while UF_{PS_n,PM_m} < UF_{PS_n,PM_{m+x}} do
        PM_m ← PM_{m+x}
        if UF_{PS_n,PM_m} < UF_{PS_random} then
          PS_n ← PS_random
        end if
      end while
    end for
  end for
end for
return PS_n
```

3.4 Summary

This chapter outlined an approach to be applied to astronomical workflows in order to improve the quality of data that it produced. Each of the components of the approach were defined, most notably the utility function which represents a relative measure of data quality within the chosen quality metric(s). Instantiating the approach was subsequently proven to be a special case of the generalised travelling salesman problem where the distances between cities were not known. The applicability of the solutions for the GTSP to the approach instantiation were then investigated. The hill climbing algorithm was found to be satisfiable and its implementation to the approach was outlined.

The effectiveness of this approach was evaluated in Chapters 4, 5, and 6 by applying it to the three distinct use cases. When the parameter spaces were small, the complete evaluation of all workflow versions constructed using the parameter spaces could be evaluated easily. However, when the parameter space was large, the number of workflow configurations increased to a level where complete evaluation was cumbersome to compute, as was the case in Chapter 6. To circumvent this problem, the use of the hill climbing algorithm as a means to quickly find a good workflow version was investigated. To test the effectiveness of this method, the full workflow space was also evaluated so that the hill climbing results could be evaluated against all other workflow versions.

The following chapter utilises the approach to assess the completeness of several candidate LSST observing strategies. The total parameter space for this example was large, however brute force evaluation was required so that the results were representative of the entire range of expected physical properties of observed objects and observing conditions, therefore the hill climbing algorithm was not utilised.

Chapter 4

Prospecting for Periods with the Large Synoptic Survey Telescope

This first use case for the approach aims to assess the completeness of the period recovery of Galactic LMXBs with several candidate observing strategies. Quiescent LMXBs are typically far too faint for monitoring with small telescopes. This is because they are located throughout the Galaxy with typical distances of order several kilo parsec, or more. Furthermore, high and patchy extinction from gas and dust, especially in the plane of the Milky Way, renders them weak and red. Therefore, even though studies predict the existence of order $\sim$1300 Galactic black hole transients (e.g. Corral-Santana et al., 2016), many are too faint to be detected, and the majority remain uncharacterised.

The high sensitivity and broadband wavelength coverage of LSST will allow it to probe through Galactic gas and dust in the Milky Way, particularly in the redder filters. Therefore LSST has the potential to expand the known population of LMXB counterparts (down to $r \sim 27$ mag) of which we have only seen a fraction during the short history of X-ray astronomy.

Other than LSST, the two wide area, broad-band optical survey telescopes that are most suited to observations of LMXBs are the Zwicky Transient Facility (Bellm, 2014) and Pan-STARRS (Kaiser et al., 2002). Although both of these surveys regularly observe large regions of the Galactic Plane, their 5 σ single visit depth is several magnitudes brighter than that of LSST and will therefore observe a much smaller fraction of the Galactic LMXB population than LSST.

An alternative route to LMXB discovery was presented by Casares and Torres (2018), where they present the H α Width Kilo-degree survey (HAWKs) and demonstrated the photometric discovery of LMXBs down to $r \sim$22. Again the optical sensitivity of this survey does not rival that of LSST, however the observations from HAWKs will

potentially complement that of LSST for LMXBs by classifying binary systems that are without X-ray follow-up.

At present, very few LMXBs have been observed with sufficient cadence in order to recover P_{orb}. One example of the limitations of such a small sample was investigated by Arur and Maccarone (2017) where they found that the current distribution of LMXB periods could equally be described by two potential period distributions. Furthermore, they deduced that a sample size of $\sim$ 275 LMXB periods would be required to break this degeneracy at the $3\,\sigma$ level.

In order to characterise the variability of transient Galactic objects, such as LMXBs, long-term monitoring with good sampling is required. However, the current LSST baseline observing strategy includes a reduced cadence for all fields within the Galactic Plane. The choice of observing strategy is yet to be finalised but the case for full Galactic Plane coverage must be backed by a quantitative analysis of its expected impact. As the strategy will be decided before the telescope begins operations, the impact of the reduced observations on Galactic transient science must be studied through simulation. Therefore, the contribution of this chapter was to use the approach to evaluate candidate observing strategies for the Large Synoptic Survey Telescope with regards to variable, Galactic science.

To do so, a workflow was designed to simulate LMXB characterisation in realistic LSST observations in order to investigate the potential for LSST to measure the orbital periods (P_{orb}) of LMXBs in quiescence. The parameter space was composed of: several candidate LSST observing strategies; the magnitude and period ranges for simulated objects; a selection of potential LSST fields. The first parameter space was used to reason over the completeness of each potential observing strategy, the second was used to ensure that the results were representative of the expected LMXB population and the third was chosen so that the results were representative of the full LSST field. The quality metric being assessed was the completeness of LMXB period recovery with each observing strategy. LMXBs are simply a test case of the more generic class of periodic variables which LSST should be able to characterise, so the results can be interpreted more broadly while keeping peculiarities specific to LMXBs in mind, such as stochastic flaring, described later.

The work in this chapter was published in the Monthly Notices of the Royal Astronomy Society as part of the ongoing discussion to the improvement of the LSST observing strategy (Johnson et al., 2018a).

4.1 The Requirements of the Workflow

In order to examine the ability of LSST to measure the variability properties of LMXBs, first the LMXB lightcurves shall be simulated, and then they shall be combined with simulations of several potential LSST observing cadences to find the expected sampling of the lightcurves. Finally,the multi-band Lomb-Scargle algorithm (VanderPlas and Ivezić, 2015) shall be used to recover P_{orb}.

4.1.1 LMXB Lightcurve Simulations

Quiescent LMXB lightcurves were simulated to represent the range of optical counterparts that LSST is expected to observe. P_{orb} and apparent magnitude were varied to encompass a broad area of parameter space outlined by the properties of LMXBs with known counterparts together with the observational constraints of LSST.

In quiescence, the optical flux of LMXBs is dominated by the companion star, with additional contributions due to the disc and stochastic flaring. The spectral profile was assumed to be of a K-type star, a typical late-type companion in many known LMXBs (see e.g. Casares and Jonker 2014). The spectral profile of a typical K-type star[1] was convolved with the LSST's filter transmission coefficients (Marshall et al., 2017) in order to calculate the expected magnitudes in the LSST filters. For an object with $r = 0$, the full set of LSST magnitudes would be as follows: $u = 4.14$, $g = 3.24$, $r = 0.0$, $i = 0.33$, $z = 1.05$, $y = 2.36$. Note these these magnitudes also include atmospheric transmission effects, which are important at both extremes of the optical spectral regime. In order to account for the additional optical contribution of the disc, which is essentially a flat power-law, a further, constant contribution of 35% was added to each filter.

To reflect the ellipsoidal modulation expected in an LMXB light curve, a peak to peak brightness variation of 0.1 mag was assumed. This was split 2:1 between the primary and secondary peaks, chosen so as to be consistent with the sample of quiescent sources published by Zurita et al. (2003). The lightcurves were constructed using alternating portions of two sinusoids with an amplitude ratio of 2:1. The limits of the simulated P_{orb} range were defined to be from 0.0063 days (9 minutes) to 50 days in twenty logarithmically spaced intervals. The minimum value includes the ultra compact LMXBs such as 4U 1820–30 (with $P_{orb} = 11$ minutes; Stella et al. 1987) and the maximum includes systems such as GRS 1915+105 (33.5 days; Greiner et al. 2001). The magnitude range used represented the expected, quiescent LMXB magnitude before reddening was applied. This was defined to be from 13 to 22 in the r band. The lower limit to the de-reddened magnitude range corresponds to a typical LMXB with $M_{\rm V} = 5$ ($M_r = 4.6$) at a distance of 0.48 kpc and the higher limit of 22 corresponds to the same object at a distance of 30

[1] `http://classic.sdss.org/dr5/algorithms/spectemplates/` (ID 11)

kpc. This range encompasses both the closest candidate LMXB GS 1354–64 at a possible distance ~ 1 kpc (Gandhi et al., 2018) and the farthest edge of the Milky Way from the Sun.

The flaring present in the optical emission of LMXBs was simulated according to the flaring power spectra reported by Zurita et al. (2003) together with the lightcurve generation algorithm outlined by Timmer and König (1995). The input parameters of the flare algorithm were $\beta = -1$ representing the slope of the power spectrum and the standard deviation of the flare amplitude was 0.04 mag. Only flares positive in flux (i.e. brightening the source above the ellipsoidal modulation) were simulated. The simulated amplitude slightly exceeded those exhibited in four of the five systems presented in Zurita et al. (2003), so it is conservative in terms of P_{orb} recovery. The simulated flaring was sampled at 30 second increments as the smallest temporal increment detectable by LSST. The absolute value of the simulated flaring was then taken to reflect the flaring being an additive component to the ellipsoidal modulation. To represent the uncertainties due to shot, instrumental and background noise, the signal to noise ratio prescription suggested by the LSST operations simulation framework[2] was used. Figure 4.1 depicts a representative segment from a simulated lightcurve displaying both the final lightcurve as well as the separated contribution from the ellipsoidal modulation alone.

Galactic reddening is an important factor to consider, however the clumpy nature of interstellar dust means that this reddening is uncertain, especially within the plane of the Milky Way. As the entire sky could not be simulated within a reasonable time, the reddening to five different LSST fields was used during the simulations. Three of the selected fields were chosen so as to encompass a wide range of potential Galactic reddenings. A further two were also chosen as they are located such that one field was observed using a different LSST mini-survey and the other resided in the main WFD region. Therefore, they were comprised of fields with OpSim field ID's: 1304, 1322, 630, 1929, 3311. Field 1304 includes the globular cluster NGC 6522 and covers a substantial part of Baade's Window which contains relatively low columns of interstellar dust. Field 1322, corresponds to an LSST field aimed at the Galactic Centre which shows very strong interstellar extinction. Fields 630 and 1929 correspond to two fields which contain famous LMXBs GX 339-4 and Scorpious X-1, respectively. Additionally, field 1929 resides in the main WFD survey region. Finally, field 3311 was included as it is a field which resided on the opposite side of the Galactic longitudinal axis to the other chosen fields and well as being located such that it will be observed by the south celestial pole mini-survey. The three Galactic Plane fields were chosen so as to gain meaningful statistics on LSST's P_{orb} recovery in this region and the other two demonstrated how the P_{orb} recovery changes with changing cadence in each observing strategy. The position of these fields in the Galactic Plane is shown in Figure 4.2. The E(B-V) to each target field was found using the dust maps from Schlafly and Finkbeiner (2011) and this was converted to the

[2] `https://smtn-002.lsst.io/`

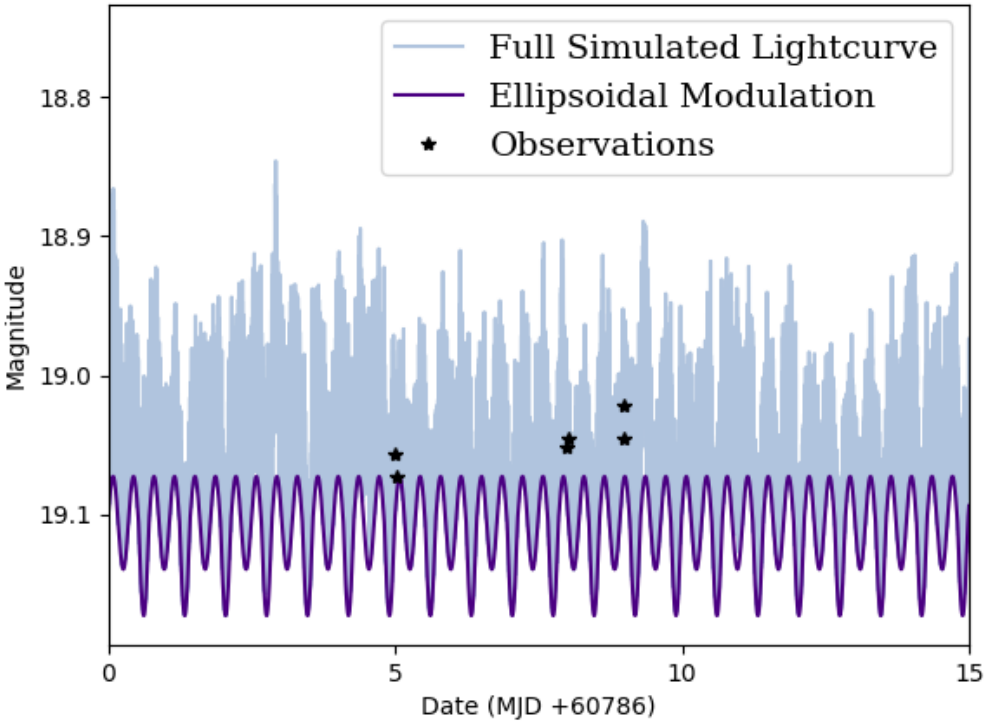


FIGURE 4.1: Segment of a mock LMXB lightcurve using r band observations of LSST field 1304 with the `astro_lsst_1004_01` observing strategy. The continuous solid purple lightcurve represents the underlying ellipsoidal modulation; light blue includes the additional flaring and noise. Stars symbolise observations made by LSST in the r filter.

expected reddening for each LSST filter using the values for R_V found also in Schlafly and Finkbeiner (2011). For each LSST field, the extinction that corresponded to that field was added to the original magnitude range. Observations which had a final magnitude that was either saturating during a single visit or fainter than LSST's 5-σ sensitivity limit as described in Marshall et al. (2017) were not used when determining P_{orb}. If there were no usable observations in a simulated lightcurve then the period was automatically assumed to not to have been recovered.

4.1.2 Observing Strategy

OpSim (Delgado et al., 2014) generated mock multi-filter observations `Minion_1016`, `Minion_1020`, `astro_lsst_01_1004` and `baseline2018a` were downloaded from the LSST simulations page[3]. Figure 4.3 displays all observations made by the new baseline strategy, `baseline2018a` (simulated using OpSim 4), of each LSST field over the full ten-year

[3] `http://astro-lsst-01.astro.washington.edu:8081/`, `http://astro-lsst-01.astro.washington.edu:8080`

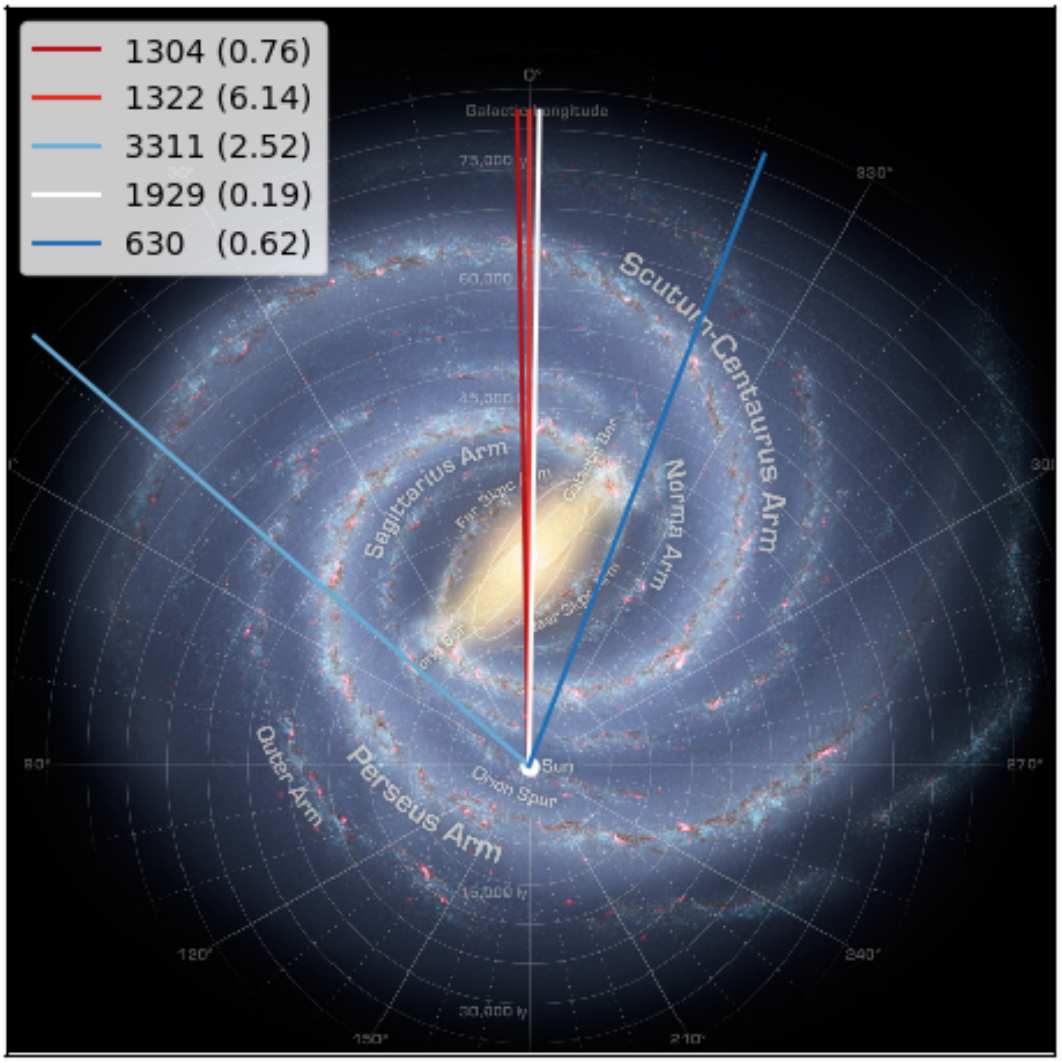


FIGURE 4.2: Figure depicting the positions of the five chosen LSST fields in the Galactic Plane. The key denotes their LSST field ID and Galactic reddening in r magnitudes. (Milky Way image: NASA/JPL-Caltech, ESO, J. Hurt.)

survey, in all filters. In the map, the regions with distinct cadences from the main WFD can be clearly seen in the north, south and Galactic Plane. As with `Minion_1016` (the previous baseline strategy), `baseline2018a` will observe all Galactic Plane fields, in all LSST filters, at a reduced cadence. One key difference between the old and new baseline strategies is that in `Minion_1016`, all Galactic Plane observations occur within the first ten months of operation, whereas these observations are spread out over the ten year survey for `baseline2018a`. `astro_lsst_01_1004` is identical to the baseline strategy `Minion_1016` except that it observes the Galactic Plane with the same cadence as the main survey region. `Minion_1020` utilises a Pan-STARRS-like cadence, with uniform coverage for all observable fields. Maps showing the total number of observations per field for each observing strategy have been included in Appendix A.1.

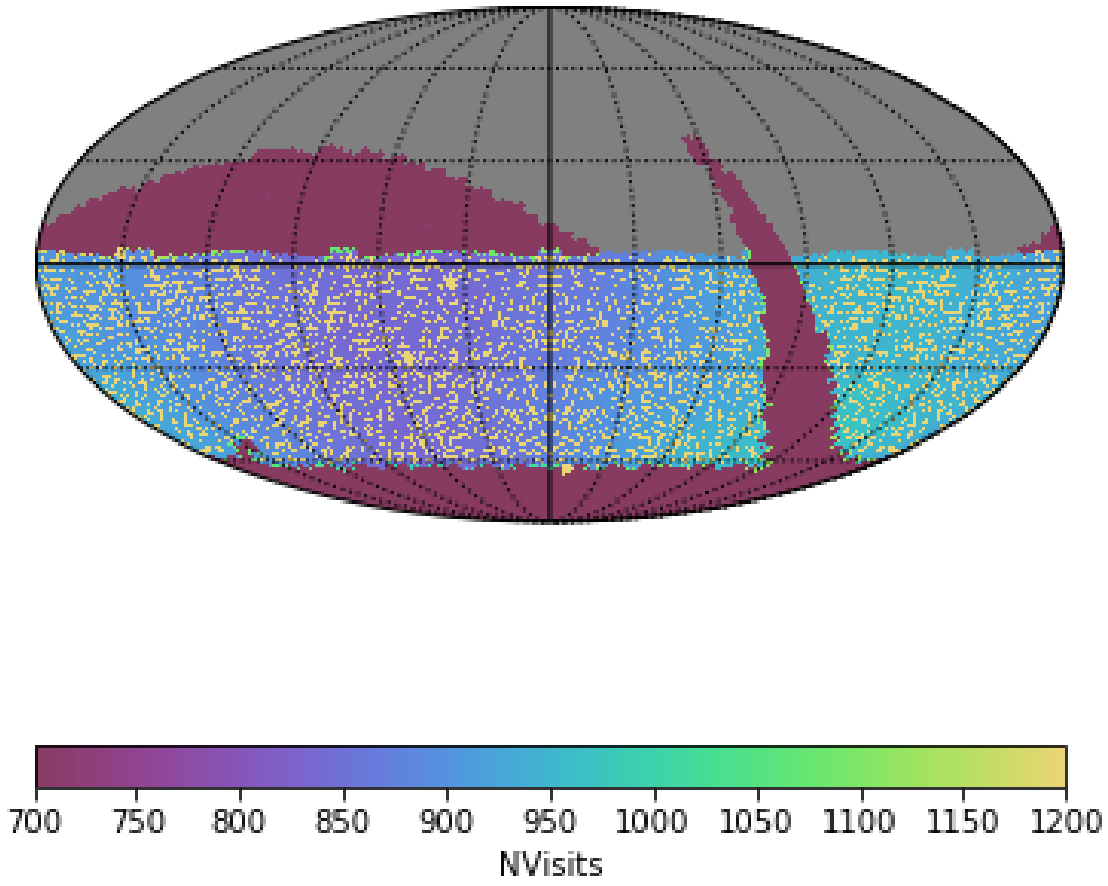


FIGURE 4.3: Total number of observations in all bands made using the `baseline2018a` observing strategy, shown in celestial coordinates where zero RA corresponds to the black line in the plane of the y-axis and North=up, East=left. Image credit: http://astro-lsst-01.astro.washington.edu:8080.

Simulated lightcurves were constructed using the observations that each observing strategy is predicted to make of each target field. The reddening used for each lightcurve corresponded to the line of sight reddening for the field whose observations were used.

LMXBs enter outburst with recurrence timescales of years to decades, during which the companion star is outshone by the disc and hence the characteristic ellipsoidal modulation cannot be observed. To reflect this in the observations, a randomly selected segment, comprising a consecutive 25% of the total observing time from the 10 year survey, was removed in all filters for `Minion_1020`, `baseline2018a` and `astro_lsst_01_1004`. However, this was not implemented for `Minion_1016` as all observations within the Galactic Plane occur within the first year.

4.1.3 Multiband Lomb-Scargle Period Measurement

To take full advantage of the randomly sampled, multi-filter data, the multi-band periodogram outlined by VanderPlas and Ivezić (2015) was used to determine P_{orb}. This

approach computes the periodogram for each LSST filter separately and regularises them on a common base model to produce a composite.

The strongest peak in the periodogram was taken to correspond to the orbital period measured for that system and its significance was determined as follows: the dates of all observations and the ellipsoidal modulation magnitudes were shuffled; in order to preserve the red noise inherent in the stochastic flaring, the flaring magnitudes (in their original order) were then added to the ellipsoidal modulation magnitude; the Lomb-Scargle periodogram was recomputed over this new modified dataset and the power of the maximum peak in this uncorrelated dataset was compared to that of the original simulated data. This process was repeated 10,000 times and the significance level was then determined as $\sigma = \frac{x}{N}$ where x represents the number of times that the peak power of the period in the original data was greater than that of the uncorrelated ensemble and N is the total number of shuffles. This formula therefore has a maximum of 1, corresponding to a 100% recovery rate. If the period was determined incorrectly, defined as $\pm$ 5% difference between the measured and input periods, the significance was set to zero. This period cut was chosen so as to provide a conservative estimate for P_{orb} recovery herein, and it should be noted that if the period were recovered incorrectly due to aliasing then the correct period may be able to be recovered with further dedicated observations. This measurement of the significance was used as a measure for the completeness of the period recovery within each region of parameter space. The decision to keep the flaring magnitudes ordered was motivated as the correlations in the flaring may artificially boost the power of the peaks in the Lomb-Scargle periodogram.

4.2 The Quality of Results Produced by the Workflow

The goal of the workflow was period recovery therefore completeness was used as the quality metric. Completeness of the data in this use case refers to the relative number of correctly recovered periods from the simulated objects for each observing strategy used. The measurement of the significance of the Lomb-Scargle periodogram measurement outlined in section 4.1.3 was used to represent the completeness of period recovery for each discrete region in parameter space. To find the completeness for each observing strategy, the significance of period recovery was averaged over the corresponding magnitude-period parameter space.

4.2.1 Instantiation of the Approach

The schematic of the workflow used for this investigation is shown in Figure 4.4. The mapping of this use case to the components of the approach are as follows:

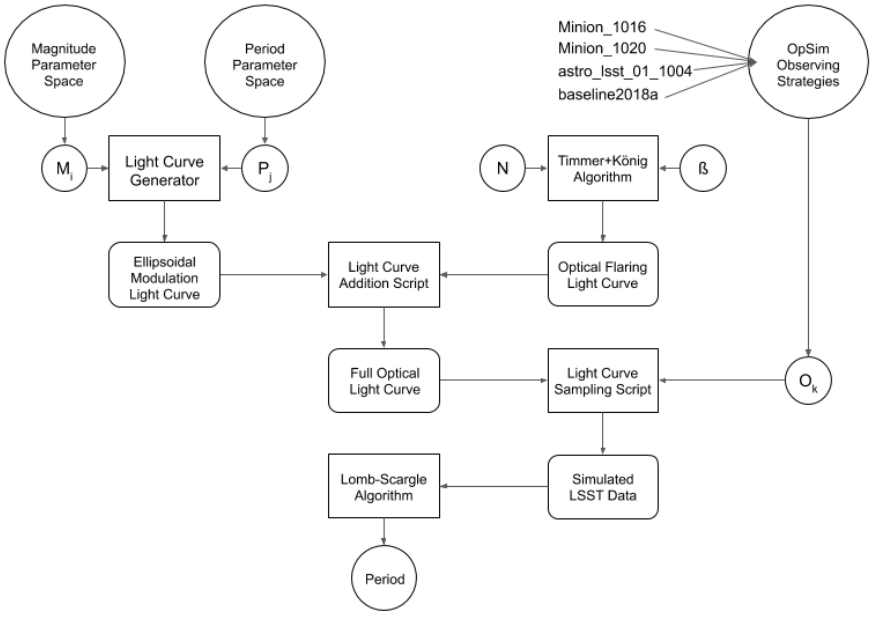


FIGURE 4.4: Schematic of the workflows main components for a single run.

- Workflow - Simulate realistic LMXB light curves, sample with potential LSST observing strategies and recover the periods.
- Processors - Light curve generator, Timmer + König algorithm, light curve addition script, light curve sampling script, multi-band Lomb-Scargle algorithm.
- Parameter Space(s) - A selection of potential LSST observing strategies produced using OpSim which have differing cadences of the Galactic Plane. Magnitude and period ranges representative of the LMXB population. A selection of LSST fields, representative of the full sample.
- Expected Result - The total number of simulated LMXBs.
- Actual Result - The number of LMXB periods that were correctly recovered from the simulations.
- Utility Function - The fraction of the simulated LMXBs with correctly recovered periods, simply:
  $U_{qm} = AR/ER$.
- Data Quality Metric - Completeness

The application of the approach in this chapter differs from that in Chapters 5 and 6 as although it aimed to compare the data quality when using each observing strategy, the decision for which strategy to choose will be decided by the LSST team. While the approach can be used here to determine the observing strategy with the highest

completeness for period recover of LMXBs, there are many other objects of interest and quality metrics that are important to achieve LSST's science goals. Therefore,the results aim to quantitatively discuss the quality of each observing strategy in this respect in order to aid the decision of the LSST team.

4.3 Investigating Versions of the Workflow

The goal for approach evaluation was to find the completeness of individual candidate LSST observing strategies. This parameter space consisted of `Minion_1016`, `Minion_1020`, `astro_lsst_01_1004` and `baseline2018a`. These strategies needed to be assessed over a range of potential LSST fields so that the results were representative of Galactic Plane observations LSST is expected to make. Therefore, a parameter space was added which included LSST Galactic Plane fields: 1302, 1322, 3311, 1929 and 630. Furthermore, to make the results representative of the LMXB population, a magnitude range of 13-22 mags and a period range of 0.0063-50 days were added to the parameter space. Workflow configurations were constructed for each distinct region in parameter space so that their completeness could be evaluated. It should be noted that whilst the LSST fields, magnitude ranges and period ranges are parameter spaces, they were employed to test the effectiveness of the different observing strategies for a wide range of potential systems and pointings. They were not the focus of the investigation which was purely about the observing strategy. Therefore, although there were 8,000 unique workflows to investigate, they can be thought of as four categories, one for each observing strategy that encompass all variations of the light curves.

All computation for evaluating the approach was performed on the IRIDIS Compute Cluster nodes at the University of Southampton. The jobs were run on the cluster's nodes which have dual 2.6 GHz Intel Sandybridge processors, 16 CPUs and 64 GB of memory, per node. The total time of computation for the brute force evaluation of all 8,000 combinations of observing strategies within all regions of magnitude-period parameter space was $\sim$16,000 CPU hours.

4.3.1 Evaluating the Completeness

As previously discussed, the completeness for a single region in parameter space (consisting of a single magnitude, period, LSST field and observing strategy) was calculated as the significance of the Lomb-Scargle periodogram generated for simulated LMXBs within that region. However, the goal of the approach evaluation was only to compare the completeness of the candidate observing strategies. Therefore, the completeness was averaged over all magnitude, period and LSST field parameter spaces for each observing strategy to result in their overall completeness.

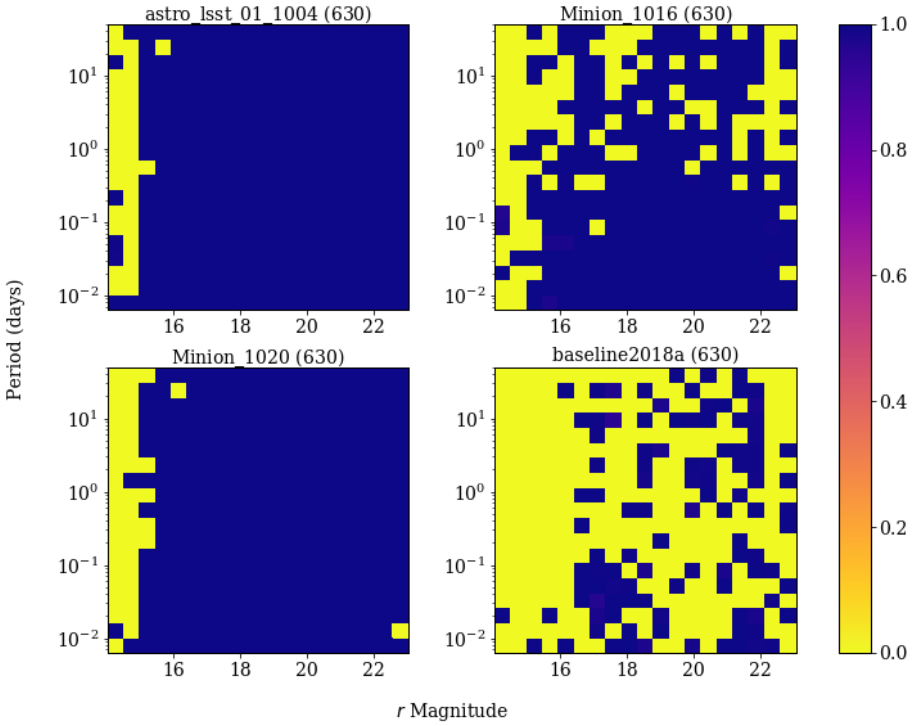


FIGURE 4.5: Colour maps displaying the period determination of LMXBs possible in LSST field 630 with observing strategies `astro_lsst_01_1004`, `Minion_1020`, `baseline2018a` and `Minion_1016`. Y axis denotes the orbital period in days, X axis the reddened r mag before adding contributions from ellipsoidal modulation, flaring and noise. The colour denotes the completeness of the period recovery. If the measured period differed from the actual period by more than 5%, then the completeness was set to zero. The graph shows a bimodality in the completenesses of period recovery as recovered periods that had low completeness were often incorrect and manually set to zero.

4.3.2 Evaluation of the Approach

Figure 4.5 shows the P_{orb} determination possible with the `astro_lsst_01_1004`, `Minion_1016`, `Minion_1020` and `baseline2018a` observing strategies over the P_{orb}-mag parameter space. The simulated observations for this graph were all within the LSST field 630; similar figures covering the other LSST fields have been included in Appendix A.2. The magnitude on the x-axis of this figure refers to the mean base r magnitude, as it would be observed after including contributions from reddening for field 630, but without adding any of the introduced stochastic variations. In other words, it corresponded to the mean flux relevant for orbital period determination. The colour denotes the completeness of the period recovery and if the period was returned incorrectly, the completeness was set to zero. The summaries for the P_{orb} recovery over the full parameter space are shown in Table 4.1, describing both the prospects per field and prospects averaged over all Galactic

TABLE 4.1: The fraction of the simulated parameter space for which P_{orb} was correctly recovered for each observing strategy, both for the individual LSST fields and the total, combined over all three Galactic Plane fields. The initials denote which cadence was used for that field; South Celestial Pole (SCP), Galactic Plane (GP) or Wide-Fast-Deep (WFD). The reddening is listed in magnitudes. The reddening and coordinates refer to the centre of the field.

		Field (Cadence)				
	Average	3311 (SCP)	1322 (GP)	1304 (GP)	630 (GP)	1929 (WFD)
Reddening:						
E(B–V)		2.52	6.14	0.76	0.62	0.19
Galactic Coordinates:						
l (°)		49.11	-0.66	0.30	-21.58	-1.50
b (°)		0.80	-0.90	-3.49	-5.24	24.81
Observing Cadences:						
`Minion_1016`	0.46	0.37	0.06	0.63	0.69	0.83
`baseline2018a`	0.23	0.16	0.02	0.4	0.28	0.74
`Minion_1020`	0.70	0.63	0.30	0.92	0.89	0.83
`astro_lsst_01_1004`	0.69	0.36	0.27	0.90	0.90	0.81

Plane fields tested.

The P_{orb} recovery was worst for the `baseline2018a` observing strategy as it only correctly recovered 0.23 of the simulated parameter space. This was to be expected as although it had a similar number of observations per field as `Minion_1016`, a 25% segment of the observations corresponding to potential outburst durations was removed from the full survey lifetime. Therefore, it offered the fewest *usable* observations per Galactic Plane field of any strategy. The next worst performing strategy was `Minion_1016` which correctly recovered 0.46 of the parameter space, the low fraction was again due to the relatively small number of Galactic Plane observations per field. The two strategies that performed best were `Minion_1020` and `astro_lsst_01_1004` which correctly recovered P_{orb} for 0.70 and 0.69 of the simulated magnitude-P_{orb} parameter space, respectively, averaged over the Galactic Plane fields. The vast majority of the incorrectly recovered periods for both strategies had *no* observations in LSST's visible magnitude range, within that region of parameter space. In these regions, there is no potential for good recovery of P_{orb}, regardless of the number of observations.

In order to evaluate the relation between Galactic reddening and period determination, the P_{orb} recovery and reddening were plotted against both magnitude and P_{orb}. In order to construct the magnitude-reddening graph, the completeness of the P_{orb} recovery was first averaged over all twenty periods for each region of parameter space with a distinct magnitude, field and strategy. This average P_{orb} recovery completeness per magnitude was then plotted against Galactic extinction. Each completeness-extinction graph therefore had three points, one which corresponded to each Galactic Plane field.

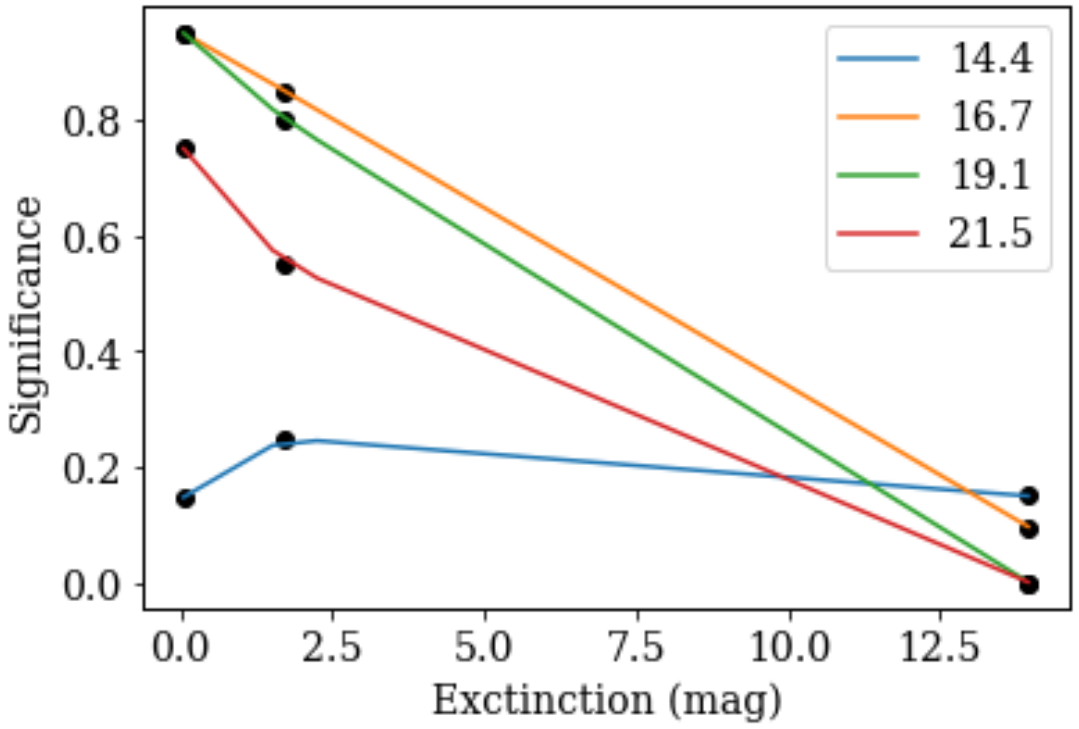


FIGURE 4.6: Figure displaying the P_{orb} recovery significance (the proxy for the completeness) interpolation for the observing strategy `Minion_1016` with pre-reddened r magnitudes shown in the key. Each point represents the P_{orb} recovery for a Galactic LSST field, with the significance of recovery on the Y axis and the field's extinction on the X axis. The line represents the corresponding extinction and P_{orb} recovery completeness for the twenty chosen, linearly spaced extinction values that are being interpolated.

These graphs were repeated for all observing strategies. These graphs were then linearly interpolated in order to find the P_{orb} recovery completeness at twenty linearly spaced reddening values, ranging from 0 to 13.9 magnitudes. An example is shown in Figure 4.6, each point represents the actual completeness of P_{orb} recovery with an LSST field and the line represents the interpolated completeness. Using the interpolation, the relation between r magnitude, r band reddening and P_{orb} recovery was then plotted. This process was then repeated, except the average was taken over the magnitudes in order to produce a graph showing the relation between P_{orb}, r band reddening and P_{orb} recovery, this figure is included in Appendix A.3.

4.4 Extrapolation to the Underlying Milky Way Population

In order to calculate the fraction of the underlying LMXB population that LSST is expected to observe with each observing strategy, the above simulations were combined with P_{orb} and magnitude distributions for systems in the Milky Way.

Firstly, to find the expected magnitude distribution, the reddening to each sight-line in the Milky Way was calculated by using the dust map of the Galaxy from Schlegel et al. (1998). This reddening was then converted to a mag ($A_{\rm r}$) assuming an $R_{\rm V}$ of 3.1 and the LSST reddening factors from Schlafly and Finkbeiner (2011). An absolute r mag of 4.6 ($M_{\rm V} = 5$) was then used to represent the LMXB quiescent counterpart main sequence K-type star. A main sequence K-type star was chosen for the companion over a sub-giant because they are typically fainter and will therefore correspond to a more conservative prediction for period determination. Finally, the distribution of systems was assumed to follow the Galactic distribution of LMXBs in the disc and bulge as outlined by Equations 4 and 5 in Grimm et al. (2002), combined with the constants from Table 4. The mass ratio used for the disc:bulge was 2:1 and a Milky Way radius of 15 kpc was also assumed. This choice of mass ratio was justified by using the bulge mass estimate from Picaud and Robin (2004) and generating a disc mass using Equation 3 and the parameters from Table 2 of McMillan (2011), giving an approximate ratio of 2:1. The contribution from the spheroid component, as described by Equation 6 in Grimm et al. (2002) was not included as we were not able to reproduce the mass ratio for it. It is also likely to be a relatively minor contribution to the total mass of black hole binaries (BHBs).

The Milky Way was then modelled as a disc with radius 15 kpc (from the Galactic Centre) and height 0.4 kpc, chosen to match the scale height of LMXBs stated in Grimm et al. (2002). This disc was then divided into segments using the Galactic coordinate system, l and b were each segmented in degree intervals and r was segmented each 0.1 kpc. The expected probability that an LMXB resided in each section was assigned and these probabilities were integrated over the entire Galaxy and then normalised. The expected reddening and magnitude was then also calculated at each Galactic segment in order to determine what region of the simulated parameter space it corresponded to and therefore, what the completeness of P_{orb} recovery in that segment is expected to be. If the region had a magnitude or reddening that was not simulated in the parameter space, then that segment was assigned a P_{orb} recovery completeness of zero.

The P_{orb} distribution of known systems was characterised by fitting a Gaussian function to the logarithm of the known BHB orbital periods from Corral-Santana et al. (2016). In log space, the distribution had a mean and standard deviation of -0.12 days and 0.47 days, respectively. Figure 4.7 displays the expected BHB P_{orb} distribution calculated using P_{orb} of known BHBs from Corral-Santana et al. (2016).

The final expected fraction of LMXBs for which LSST could determine periods was then calculated by multiplying the P_{orb} completeness, the expected period distribution probability and magnitude distribution probability at all points in parameter space. The equation for this process is shown in Appendix A.4. This was then normalised to a conservative estimate of the LMXB population of 1040 objects to determine the total number of BHB periods that LSST observations could be expected to recover, as shown in Table 4.5. The population estimate used was a combination of the total population

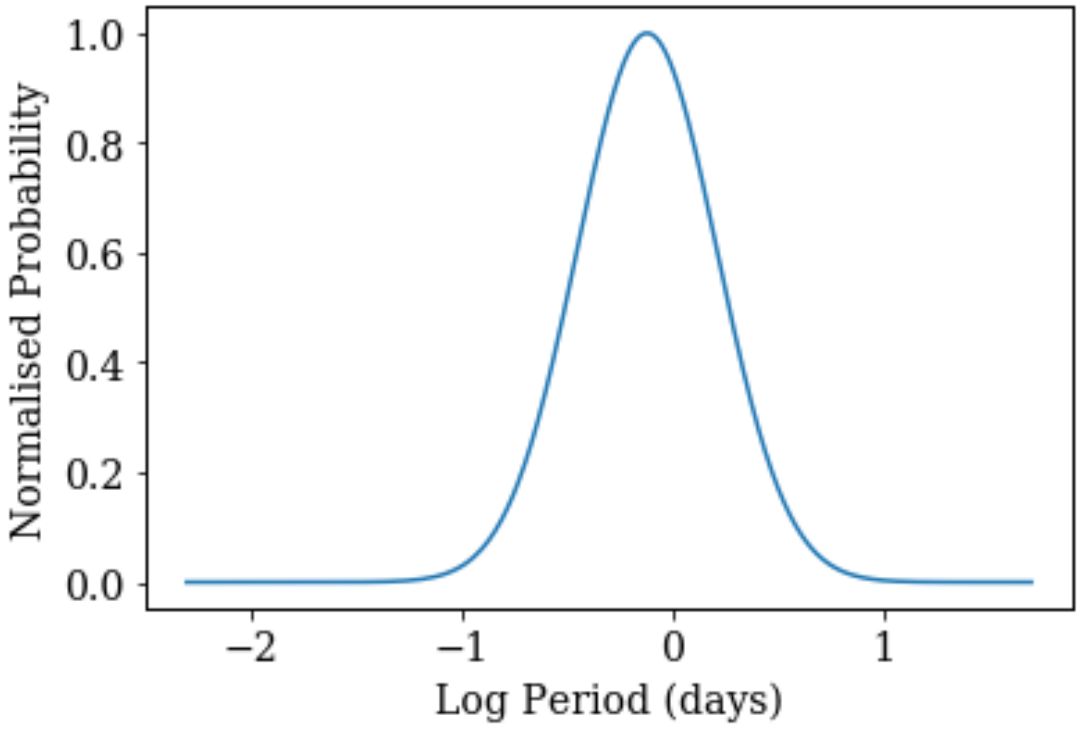


FIGURE 4.7: P_{orb} distribution of BHBs, generated by fitting the logarithm of the BHB periods from the BlackCat catalogue (Corral-Santana et al., 2016). The probability is normalised to one at peak.

estimate of BHBs from Corral-Santana et al. (2016) (1300), combined with the fact that 80% of known LMXBs reside within the LSST defined Galactic Plane.

4.5 Discussion

The approach was applied to a workflow in order to investigate the prospects for periodic signal extraction from LSST light curves. The parameter space being evaluated over was the candidate LSST observing strategies. This could be expanded to include additional or custom built strategies which have the potential to produce higher quality results for signal extraction. However, purpose built strategies will likely reduce the quality of results for other LSST science goals. Niche strategies are unlikely to be assessed in the detail that the chosen four strategies were in, for example, the LSST white paper (Marshall et al., 2017). These statements both motivate the choice for the current set of observing strategies and deter the addition of strategies to this parameter space. However, a potential improvement is to replace observing strategies `Minion_1016`, `Minion_1020`, and `astro_lsst_01_1004` with their counterparts generated by OpSim4.

The test case was determination of orbital periods of quiescent LMXBs, but the results can be used more generally for assessing various proposed observatory cadencing

strategies, especially those relevant for the Galactic Plane. These additions could be realised by making alterations to the magnitude and period parameter space, in addition to adding the other nuanced optical behaviour that each object exhibits. Expanding the simulations in this way could help broaden the reach of the investigation so that it has a higher impact on the final decision of LSST observing strategy.

P_{orb} recovery with LSST was shown to be affected by the total number of the observations in the observing strategy; the observing strategies with the highest numbers of observations had the best P_{orb} recovery and those with the fewest had the worst. Observing strategies that did not have a reduced cadence in the Galactic Plane (`astro_lsst_01_1004` and `Minion_1020`) resulted in excellent results for P_{orb} recovery over the simulated mag-P_{orb} parameter space, correctly recovering periods for nearly all of the parameter space which contained observations with magnitudes within the observing range of LSST (within the saturation mag and 5 σ limit). Conversely, the baseline strategies (`Minion_1016` and `baseline2018a`) with a reduced cadence in the Galactic Plane, correctly recovered far fewer periods. Furthermore, the current baseline strategy (`baseline2018a`) correctly recovered on average a factor of ~ 3 fewer periods than either `astro_lsst_01_1004` or `Minion_1020`.

The P_{orb} recovery was shown not to vary much between `astro_lsst_01_1004` and `Minion_1020`, even though the former has an additional 100 observations. The most common reason that the period was recovered incorrectly for these two strategies was that there were no usable observations for that region in parameter space (i.e. all observations were so heavily obscured that all observations had magnitudes that were above LSST's $5\,\sigma$ limit). In fact, there were very few regions with an incorrectly recovered period where this was not the case, meaning that the difference in total observations between these two strategies had little effect. This suggests that the number of observations required for good period determination of LMXBs is higher than the number in either `Minion_1016` or `baseline2018a` however also lower than in `astro_lsst_01_1004` and potentially lower than in `Minion_1020` also. However, as both strategies that had good P_{orb} recovery increase the total number of fields in the Wide-Fast-Deep survey region, without increasing its priority, the median co-added depth achieved is then reduced by 0.04 and 0.15 mags for `astro_lsst_01_1004` and `Minion_1020`, respectively when compared to `Minion_1016` (Marshall et al., 2017). This is a slight reduction in depth whose impact on other scientific programs would need to be assessed.

One factor that may artificially boost the LMXB recovery for the observing strategy `Minion_1016` is that no observations were removed to account for the potential time that the LMXBs would be in outburst. Removing a randomly selected 2.5 years from this strategy, as was done for the others, is not sensible as the baseline of the observations for this strategy was only 1 year. This will however mean that a fraction of the LMXB population will not be observable through ellipsoidal variability for the entirety of this baseline strategy lifetime, although this is likely to be a negligible effect.

In order to calculate a conservative estimate for the recovered P_{orb} of LMXBs, observations were only considered if they had measured magnitudes within LSST's observable range. However, LSST will perform forced photometry at the location of known objects even if they lie below the $5\,\sigma$ limit during intermediate data releases (Jurić, 2018). This could be relevant for known LMXBs in quiescence that have only been bright enough to be observed in outburst. Therefore, LSST may also be able to determine periods for objects that are outside of this limiting magnitude. There may also be fringe cases where r is $\sim$ 24 mags and its optical variability raises it occasionally above the $5\,\sigma$ limit, thereby producing more usable observations than considered here. However, the combined impact of both of these scenarios is not likely to be significant.

The overlap and dithering of LSST fields also has the potential to impact the period recovery of LMXBs possible with LSST. These effects could mean that some LMXBs are visible in several LSST fields. Equally, they could also mean that the systems may fall within chip gaps in some images and not others. The impact of these effects will be investigated in the future, however it is not expected to be substantial. Examples of dithering investigations are presented in Chapter 9 (Cosmology) of Marshall et al. (2017).

The average and standard deviation value of the flaring had a sizeable effect on the overall period determination. The choice of 0.04 mags for the standard deviation of the flaring was justified as it was representative of the majority of the sample outlined in Zurita et al. (2003). However, one of the sample included a system with standard deviation $>$ 0.1 mags. When implementing the simulations with this value, the completeness of P_{orb} recovery decreased significantly. This region of parameter space will be explored in the future.

It should also be noted that the predictions made by using the dust maps are only estimates as the maps used (Schlegel et al., 1998) represent the integrated reddening along each line of sight, therefore information on the radial change of extinction in the Galaxy is lost. Another limitation to these dust maps is their angular resolution of 6". It should also be noted that the reddening used per field was used assuming a single pointing, corresponding to the centre of the field, whereas there are potentially many different reddening values per field.

By combining the LMXB period recovery fraction of LSST with a fairly conservative estimate for the LMXB population of 1040 systems, a lower limit on the number of systems for which LSST can be expected to determine periods was found, as shown in Table 4.5. LSST will likely correctly determine P_{orb} for $\sim$200 systems and $\sim$180 systems, while implementing the baseline strategies simulated by OpSim 3 and 4, respectively (`Minion_1016` and `baseline2018a`), whereas for observing strategies that do not have a reduced cadence in the Galactic Plane (`astro_lsst_01_1004` and `Minion_1020`), LSST will likely correctly determine P_{orb} for $\sim$300 LMXBs. This sample is sufficient to satisfy

Observing Strategy	Total Number of Observations [1]	Period Recovery Fraction	Period Recovery (No. of systems)	Description
`Minion_1016`	180	0.23	239	OpSim 3 baseline
`baseline2018a`	134	0.18	187	OpSim 4 baseline
`Minion_1020`	548	0.32	333	Pan-STARRS-like
`astro_lsst_01_1004`	613	0.32	333	WFD in Galactic Plane

TABLE 4.2: Fraction of Galactic LMXBs with Measurable Periods as a function of LSST Observing Strategy. The fraction was combined with a total population estimate of 1040 to calculate the total number of systems expected with correctly recovered periods for each observing strategy. Total number of observations represents observations made, averaged over the three Galactic Plane fields (minus a 25% segment for `Minion_1016`, `astro_lsst_01_1004` and `basline2018a`).

the example science case mentioned in the Introduction, as Arur and Maccarone (2017) deduced that a sample of ∼275 LMXB periods would be required in order to distinguish the two different LMXB P_{orb} distributions at the 3σ level.

Although LSST will have the *capability* to determine the periods for many LMXBs, identification of these sources will require further evidence, and there are numerous other Galactic entities that exhibit similar behaviour to the ellipsoidal modulation of LMXBs. However, there are several potential routes for discerning potential LMXBs. The characteristic X-ray emission seen during the outburst phases of LMXBs can be observed via follow-up with current X-ray telescopes such as Chandra (Weisskopf et al., 2002). The high sensitivity future instruments such as Lynx (Team, 2018) may also be able to observe the X-ray emission of many of the LMXBs in quiescence. However, X-ray follow-up will not be feasible for all LMXB candidates detected by LSST as X-ray telescopes have relatively small fields of view. All-sky X-ray surveys such as eROSITA (Cappelluti et al., 2010) will observe the entire Galaxy, however they are limited by their sensitivity and will not be capable of observing the entire Galactic, quiescent LMXB population.

Spectroscopic follow up can be used for source characterisation and radial velocity determination in order to make mass measurements of the LMXB population. The current generation of spectroscopic telescopes may struggle to observe some of the fainter systems simulated, however this will be feasible with the next generation of instruments available in 2032, after LSST's 10 year lifetime. As the number of LMXBs with dynamically confirmed compact object masses is currently fewer than 20, LSST has the potential to help in improving this by at least a factor of ∼10 and potentially a factor of 15+. The implications of this result also extend to many other classes of stellar phenomena involving binary systems which are likely to benefit in exactly the same way as outlined here.

4.6 Conclusions

In conclusion, it was found that the period recovery of LSST for LMXBs improved by a factor of ∼3 when utilising observing strategies which did not reduce the number of observations in the Galactic Plane. However,these strategies also reduced the number of observations in all other fields by ∼5%. Therefore the adoption of these strategies over the baseline will likely depend upon whether the original science drivers for LSST could be achieved with 5% fewer observations.

The LSST observing strategy is just one of the components currently being scrutinised in order to improve LSST operations. Another major component is the data management system which handles the data processing as well as the documentation of the provenance. Recording this provenance introduces an initial overhead to the computation. The following chapter investigates methods for reducing this computational cost of provenance recording.

Chapter 5

Using the Provenance from Astronomical Workflows to Improve their Timeliness

Over the last few decades, the ability of the astronomer to collect and process data has increased dataset sizes from giga to tera to now peta-byte scale. This is in part due to the creation of large scale survey telescopes such as the Sloan Digital Sky Survey (SDSS) (York et al., 2000), the Palomar Transient Factory (Law et al., 2009) and, in future, the Large Synoptic Survey Telescope (LSST) (Tyson, 2002). As astronomy is increasingly becoming a data-driven science, many frameworks and tools have been designed to automate the generation of the accompanying provenance. Producing this detailed record of the provenance requires additional storage and introduces an initial runtime overhead to the execution time. However, it can also allow for a significant reduction in resources when analysing the final data products.

With the advent of new survey telescopes, such as LSST, which have extremely large datasets, it is becoming ever more crucial for the astronomer to make the most efficient use of the computational resources available. The contribution of this chapter was a quantitative demonstration of the use of provenance on the final data products of astronomical workflows as a means to improve their processing efficiency.

In order to achieve this, firstly, PROV-TEMPLATES were used to generate the provenance of an astronomical image processing pipeline which was designed to measure the brightness variation of black hole binary systems. Secondly, within the context of this workflow, two use cases were identified for which provenance is vital for the astronomy community. Use Case 1 was to investigate the origin of an observed variation in a target astronomical objects brightness and in Use Case 2, a star was found to be incorrectly measured and it was investigated whether this star was used in the calibration process.

These use cases were then evaluated with and without the use of the generated provenance and the relative resources required by each method were quantified. Finally, the total impact of provenance capture and usage was measured by comparing the computational resources required for implementation and use case evaluation with and without the use of provenance. This chapter investigated the use of provenance as a means to improve processing efficiency - contrary to the initial overhead that is required to produce it. It also determined whether the increase was at a sufficient level to completely offset the initial overhead introduced.

The work in this chapter was published as a contribution to the International Provenance and Annotation Workshop (Johnson et al., 2018b).

5.1 Astronomy Application

The motivation for this piece of work is to investigate the potential for provenance to increase the efficiency of processing astronomical data, therefore we outline an astronomical dataset and image processing pipeline in this section. The astronomical images used throughout were all taken of the low mass X-ray binary (LMXB), GS 1354-64 which consists of a star in orbit around a black hole. The pipeline identifies the objects in the images, measures the brightness of all objects and calibrates them to account for changing viewing conditions in order to find the variations in flux that GS 1354-64 exhibits over time. These optical variations can be used to determine properties of the system such as its orbital period, which can then be used, inconjunction with spectral information, to infer the masses of the binary components. Currently, this is the only method to robustly measure the mass of stellar mass black holes and increasing the sample of known black hole masses enables us to better understand their properties. Survey telescopes are the ideal equipment in order to discover new systems as they are designed to systematically observe large swathes of the sky. As we are looking to discover new LMXBs, we do not know their position, although we may know areas of the sky where they are more likely to be. This means that large quantities of data must be analysed in order to find the objects of interest and it is essential to utilise any advantage in computational efficiency available motivating our use of provenance for this investigation.

5.1.1 The Image Processing Pipeline

The image processing pipeline had two main functions: differential photometry and pattern recognition. The measured brightness of the object in an image is dependent on conditions such as atmospherics conditions, proximity to the Moon and other sources of light pollution. The images must therefore be calibrated via differential photometry, whereby stars of known and constant brightness within the same image are used to correct

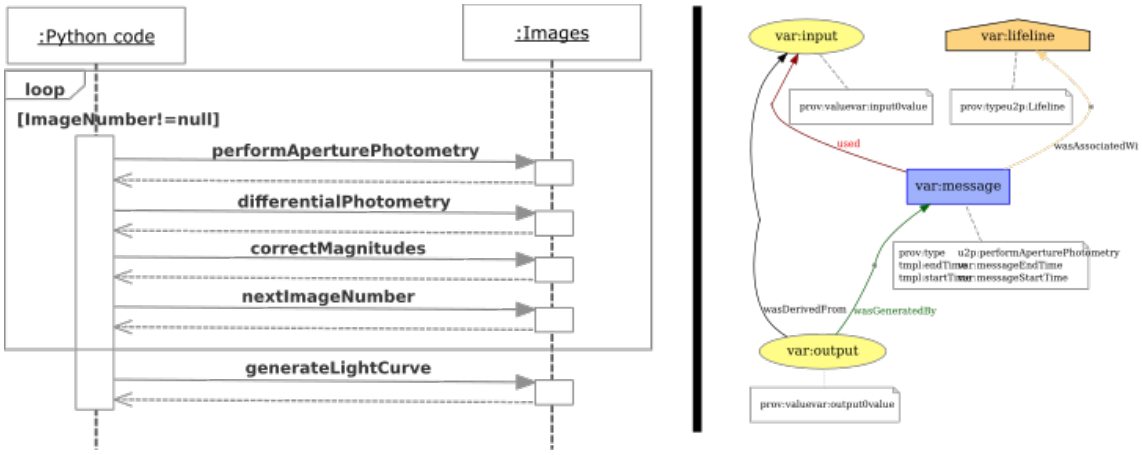


FIGURE 5.1: The left hand side is a UML sequence diagram depicting a simplified version of the differential photometry process. The right-hand side is a PROV template generated using UML2PROV (Sáenz-Adán et al., 2018).

for the measured brightness for differences in observing conditions. Pattern recognition was required in order to determine which source in the image corresponded to which astronomical object. The use cases are both concerned with differential photometry, therefore the explanation of the workflow will focus on this aspect.

The left hand side of Figure 5.1 is a UML sequence diagram depicting a simplified version of the differential photometry in the image processing pipeline. The two lifelines of the UML diagram represent the script itself and the astronomical images. The first message, *performAperturePhotometry*, measures the brightness of all objects within the image. Then, *differentialPhotometry* compares the measured brightness of known objects (standard stars) to their true brightness in order to calculate the brightness correction needed for that particular image. The pipeline determined which stars should be used as standard stars for each image individually. Multiple bright standard stars were used in order to obtain a more consistent calibration as individual stars and faint stars are more effected by things such as noise or systematic uncertainties. Once a sample of candidate stars had been selected, they were cross-referenced with the SIMBAD astronomical database (Wenger et al., 2000) to determine whether they were non-variable stars. If they were found to be so, then their true brightness was retrieved and compared to the measured value and the brightness correction for that image could be calculated. This process was repeated for each standard star in the image and the final correction was the averaged value. The brightness of the target object (in this case GS 1354-64) was then adjusted using this correction. This process was then repeated for all images. Finally, the corrected brightness of the object across all images was plotted against time to give the lightcurve which demonstrated the object's temporal optical variation.

5.1.2 Use Cases

In order to assess the usefulness of provenance for the astronomical community, the following use cases have been identified.

USE CASE 1. *Variation Investigation* - An Astronomer, Alice, detects a change in luminosity in a star between two images taken on two different nights. Alice *determines whether the change was intrinsic to the object or a result of the image processing pipeline.*

Solving this use case requires a record of the version of the pipeline that was used for the image processing. The change in brightness could also be the result of the choice of standard stars used to correct the measurement.

If the image processing is found to be consistent between the observations, then the change in observed brightness can be deduced to originate from the object. However if there are inconsistencies then the images must be reprocessed to determine the true origin of the variation.

Without accompanying provenance, the processing would have to be repeated, ensuring the pipeline was identical in order to dispel any doubt in the origin of the variation.

The aim of evaluation of Use Case 1 asserts absolute certainty that the origin of the optical variation was not due to the image processing pipeline. There is a standard assumption that the origin of the variation is from the object. Therefore, it is likely that Use Case 1 would only be evaluated when the astronomer, Alice, detects an unexpected result. An unexpected result from astronomical images is not uncommon, however, quantifying how often this will occur is difficult to determine as this kind of data is typically poorly documented within the astronomical community. Consequently, estimated probabilities of 1%, 10% and 30% were all investigated in order to assess the impact of evaluating Use Case 1 on the total computational resources required.

USE CASE 2. *Calibration Propagation* - A star that was previously thought to be standard has been shown to demonstrate variability. Alice *determines which objects used this star for calibration and recalculates the photometry for them.*

Standard stars are objects of known and constant luminosity that astronomers use to calibrate images. If a standard star that was used for calibration had a miscalculated brightness then the calibration could be incorrect. An incorrect calibration means that the measured brightness of the target object is wrong, invalidating the results.

Without the use of provenance, there are two possible solutions for this calibration propagation: firstly, with no knowledge of the standard stars used for calibration, all images which contain the previously standard star would have to be re-processed, ensuring that

this star is not selected; secondly, the workflow could be re-run up until the standard stars are selected from each image, and with this information, only the images which use the previously standard star in the calibration would be repeated.

Conversely, when evaluating this use case with provenance, the provenance can be queried to return the list of standard stars used in the calibration process for each image. From this, only the images which contain the newly variable star have to be re-processed.

The invalidation of the use of a standard star could also be due to an incorrectly measured brightness as well as incorrectly determining the object to be variable. Determining how often Use Case 2 is likely to be evaluated is not trivial by any means as an object may be incorrectly measured or identified if: the object saturated the image; a cosmic ray interfered with the image; there were unaccounted for artefacts or systematics; the standard object exhibited sporadic variation or it transitioned into a variable object. Taking into account all of these scenarios, an estimated 1% probability that Use Case 2 would need to be evaluated was assumed. It should be noted that this number could be calculable if provenance use was more ubiquitous within the astronomy community.

5.2 Provenance in Astronomy Simulations

Although the aim of this investigation was to demonstrate the use of provenance to reduce the overall processing cost, the initial overhead introduced by provenance capture must also be addressed. The PROV-TEMPLATE Moreau et al. (2017) approach was used to generate PROV-compatible provenance which described the workflow. The full pipeline was modelled as a UML Sequence Diagram and later, UML2PROV (Sáenz-Adán et al., 2018) was used to generate *templates* that described the design of the provenance to be generated for each function. During the execution of the workflow, *bindings* were generated every time a function was called which contained the variable-value pairs (such as inputs or outputs) that were specific to that call of the function and had corresponding variables on the template for that function. On the right-hand side of Figure 5.1 a template generated from *performAperturePhotometry* can be seen. After completion of the workflow, these *bindings* were then expanded with their corresponding *templates* using the ProvToolbox[1] to yield the individual provenance files. These were then merged to produce the full provenance that described the system.

The image processing pipeline analysed a series of 10 images of LMXB GS 1354-64 taken by the Faulkes Telescope. All of the computation was repeated twenty times and the results in Figure 5.3 A) represent the average and standard deviation of these execution times. It should be noted that the only relevant time increase for workflow execution time is the addition of *bindings* as the merging and expansion can both be done post-pipeline. The size of the products of the workflow with and without provenance were

[1] `https://lucmoreau.github.io/ProvToolbox/`

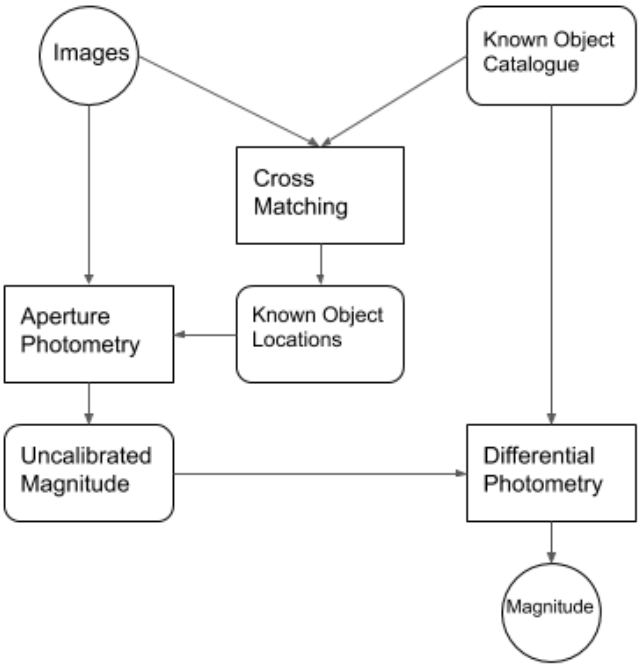


FIGURE 5.2: A diagram depicting the workflow outlined within this chapter. The parameter space here is the decision of whether or not to record provenance. This would be recorded at each edge in the workflow.

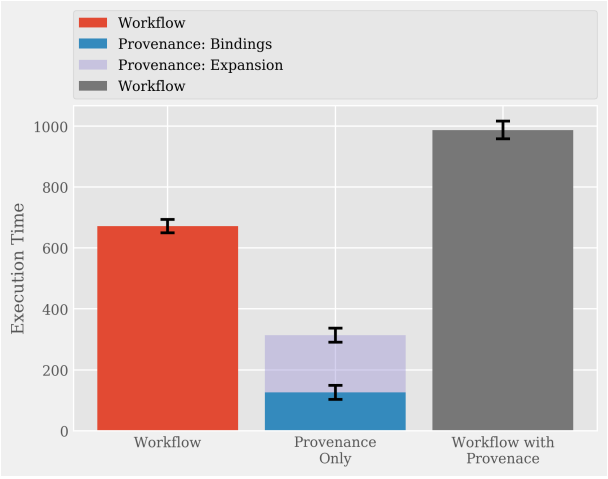
also assessed and are shown in Table 5.1. The size of the inputs are also included to demonstrate that whilst the provenance files are large when compared to the outputs, they are still inconsequential in the scale of the full workflow.

All simulations were run on a Dell Latitude E7470 laptop with the following specifications: 8GB of system memory; an Intel®Core™ i5-6200U CPU @ 2.30GHz. The machine was running Ubuntu 16.04, kernel: 4.4.0-112-generic.

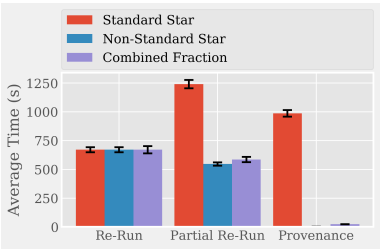
5.3 Instantiation of the Approach

Figure 5.2 shows the workflow and a diagram of the associated provenance. The mapping of this use case to the components of the approach are as follows:

- Workflow - Measure the optical variation of a LMXB over a series of images.
- Processors - Cross matching, aperture photometry, differential photometry, light curve generation.
- Parameter Space - Whether or not provenance is recorded by the workflow.
- Expected Result - The time of workflow execution ($T_w(ER)$) and use case evaluation ($T_e(ER)$) without recording provenance.



((A)) Timing: Workflow Execution, with and without Provenance



((B)) Timing: Analysis of Use Case 2

FIGURE 5.3: A) Average processing times for workflow execution in seconds, with and without provenance generation. B) Computational resources required to evaluate Use Case 2, when implementing different solutions. Execution times vary depending on whether the newly variable star was used as a standard star in the calibration on not, so both times are shown. The combined fraction convolves these processing times with the probability that any star in the image was used as a standard star. Both sets of results are the average found over twenty simulations and the error bars represent their standard deviation.

- Actual Result - The time of workflow execution ($T_w(AR)$) and use case evaluation ($T_e(AR)$) when recording provenance.
- Utility Function - The execution/evaluation time for the workflow with provenance, normalised to that without provenance and combined with estimates for the probability (P_uc)) of the use case evaluation taking place. Therefore:
  $U_{qm} = \frac{T_w(AR)+(T_e(AR)\,P_uc)}{T_w(ER)+(T_e(ER)\,P_uc)}$
- Data Quality Metric - Timeliness and Relevance

The parameter space only has two components - to record provenance or do not. Whilst intermediate granularities of provenance recording may have produced workflows with an overall higher quality, this was intentionally not investigated to avoid biasing the results towards recording provenance specialised to the outlined use cases. This also meant that the work was more generalisable to other astronomical workflows and that it could be expanded to include additional use cases not considered here.

5.4 Evaluation

5.4.1 Use Case 1 - Variation Investigation

The astronomical pipeline may not always perform a consistent analysis from image to image. It may have different parameters during the calibration such as which stars were used as standard stars. It may also use different library versions of the pipeline and the path that each data product made through the pipeline may not always be the same. Use Case 1 investigated an observed change in brightness from one image to another and tried to determine whether this variation was inherent to the object itself or whether its origin was due to inconsistencies in the image processing pipeline.

In order to evaluate Use Case 1 without provenance, the workflow must be re-run over the series of images where the variation was observed, with the pipeline versions and calibration settings made to be the same throughout. To evaluate Use Case 1 with the use of provenance, SPARQL queries were written to determine which versions of the pipeline and which standard stars were used for each image. The queries were < 10 lines long and had a negligible run time (< 1 second).

It was found that the same standard stars were used throughout the series of images and the versions of the pipeline used were the same throughout as well. Therefore, the observed variation could be deduced to not be due to the image processing and the data did not need to be reprocessed. This information resulted in a $\sim$99% increase in computational efficiency over evaluating the use case without provenance. Table 5.2 shows the processing time necessary for evaluating each use case, as well as the length of the code required to do so.

Table 5.1: The size of inputs consumed by and outputs produced by the image processing pipeline with and without provenance generation.

Method	Total Input Size	Total Output Size
Workflow Only	21MB	20kB
Workflow with Provenance	21MB	546kB

Table 5.2: Computational resources required to evaluate Use Case 1, including the average run time and an order of magnitude of the lines of code needed to evaluate the use case with and without the use of provenance.

Method	Use Case Analysis Computation Time (s)	SD (s)	Lines of Code (Approximate)
Workflow Only	671	22	500
With Provenance	1	0	10

5.4.2 Use Case 2 - Calibration Propagation

Use Case 2 was to determine whether a star that was recently determined to be variable was used in the image processing as a standard star and therefore invalidated the calibration for that image. Three ways of evaluating Use Case 2 were investigated: firstly, the workflow was completely re-executed, ensuring that the variable star was not used in the calibration process; secondly the workflow was executed up until the selection of standard stars, and this information was recorded and the images which contain the variable object were re-computed and finally, the provenance of the workflow was queried to determine which images should be re-processed.

For the first method of evaluation, the time to evaluate Use Case 2 was the same as the original execution time as there was no information on which images did or did not use the variable object for calibration so all must be repeated. For the second scenario, the evaluation time is reduced when the variable star was found not to be used as a standard star as the workflow had to only be partially re-run. However, if it were found to be used as a standard star then the workflow must also be completely re-run with this star not being used in the calibration in addition to the partial run to find the standard stars used. The third method of evaluation queried the provenance in order to determine whether the newly variable star was used as a standard star. In summary, the first evaluation method assumed no knowledge of the workflow and was always completely re-ran. The second method determined information on the standard stars used by partially re-running the workflow then decided whether it should all be re-run. The final method leveraged provenance information in order to determine whether the workflow should be re-run.

As the computational efficiency of two of the methods rely on whether the newly variable star was used as a standard star, the probability that any star in the image was used in the calibration as a standard star was calculated. This probability was convolved with the computation time required by each method of use case evaluation for whether the star was used as a standard star and if it was or not. This probability, P, was defined as $P = n/A$ where n is the number of standard stars per image and A is the average number of objects per image. For this example, 10 standard stars were used and the total number of objects in the image was $\sim$450, therefore, assuming all objects were treated equally, there was a $\sim$2% chance that any star in the image was used as a standard star. By combining this probability with the two timings, the average cost of use case evaluation was calculated if any given star in the image was found to be variable.

Figure 5.3 B) shows the results for evaluating Use Case 2 with the three possible solutions. The time represents the average execution time after repeating the simulation twenty times. The columns in Figure 5.3 B) represent time taken when the object found to be variable was used as a standard star, when it was not used as a standard star and both

these results combined with the probability that any star in the image was used as a standard star (the combined fraction).

It was found that the computational processing cost of Use Case 2 evaluation if the star is found to be standard decreased by 21% with provenance when compared to partially re-running the workflow. However, it was also found that the processing time increased in this respect with the use of provenance by 47% when compared to simply re-running the workflow. This is due to three reasons: firstly, the initial overhead of provenance production; secondly the relatively small cost of querying the provenance and finally, the workflow must be completely re-run in either case as the fact that star was used as a standard star invalidated the initial results. It was also found that the cost of evaluation was greatly reduced if the star was not used as a standard star because the only computational cost was for querying the provenance which was negligible when compared to re-running the workflow. This increased the efficiency by ∼99% when compared to either evaluation without the use of provenance. Finally, these efficiencies were combined with the probability that the star would be used as a standard star, it was found that with the use of provenance, the computational efficiency of evaluating Use Case 2 increases by a factor of 97% and 96% when compared to evaluating it by re-running and partially re-running the workflow, respectively.

5.5 Discussion

In this chapter, the approach was applied to an astronomical workflow which measured the optical variation of LMXBs with the goal of improving the timeliness. Timeliness is a particularly important quality metric when investigating changes in brightness as a swift identification of the change allows follow up observations to be triggered before the event ends. For example, an early detection of an LMXB outburst enables the study of the rise to outburst which is critical for understanding their physical mechanism. Few LMXB outbursts have been observed prior to the peak but some examples include Koljonen et al. (2016) and Bernardini et al. (2016). The parameter space consisted of two components: the decision of whether or not to record provenance. As the parameter space was small, evaluation of all possibilities was not an issue and no optimisation was attempted, as with in Chapter 6. It was found that recording the provenance of an image processing pipeline increased the initial processing cost by ∼45%. However, we have also demonstrated that the use of provenance resulted in an increase in computational efficiency of 99% and 96% when evaluating Use Cases 1 and 2, respectively. It was speculated that evaluation of Use Case 1 would occur from 1% to 30% of the time and Use Case 2 would likely need to be evaluated ∼1% of the time. By combining the processing cost of provenance production, use case evaluation and the probability that the use cases would need to be evaluated, it was computed that the total net change in processing efficiency of the workflow by introducing provenance generation as a decrease in computational processing efficiency

of 13-44%, depending on how often Use Case 1 needed to be evaluated. The full results are shown in Table 5.3.

These results demonstrate that there is the potential for substantial improvements in timeliness but the initial overhead is also fairly high. This is in part due to the only option for recording the provenance within the parameter space being a full and detailed set. This was included so that the provenance could be used to solve many more use cases than those outlined here, however there are additional options for provenance recording that could be included without sacrificing its potential to describe a multitude of additional use cases. For example, the full templates were generated in this example but the bindings alone could be generated on execution and then queried over which would negate the time taken for their expansion. It should also be noted that the expansion does not need to take place during the execution of the workflow and can be done post processing. Another addition to the parameter space could be the inclusion of recording provenance at targeted intervals instead of for the entire workflow. This could drastically reduce the total size of the provenance and by extension the processing cost but would necessitate an investigation into where the important capture points are in the workflow.

It was also found that when including provenance, the total size of artefacts produced by the workflow increased by a factor of $\sim$6. Whilst these results do represent a large increase in data products, it should be noted that they are completely un-optimised for storage savings. Also the provenance is fairly fine-grained and has the potential to evaluate many other use cases not investigated here. This means that there is the possibility for a significant reduction in both the size of the final provenance and its intermediate products. Furthermore, the combined data products from provenance production and the workflow still represented $< 1\%$ of the total data products consumed by the pipeline as the size of the input images dwarfed that of the data products.

These results pertain to the image processing pipeline used during this paper and it is likely to change from pipeline to pipeline. Other pipelines which are designed to achieve the same goals will likely be similar in operation and correlate with the results found in this investigation. One interesting investigation would be the comparison between results

TABLE 5.3: Total computational processing cost of running the workflow with and without provenance. Including processing cost of use case analysis combined with the probability that the use case must be evaluated. Use Case 1 results are combined with the probability the use case would need to be evaluated 1%, 10% and 30% of the time.

	Workflow Run Time (s)	Use Case 1 Run Time (s) (1%,10%,30%)	Use Case 2 Run Time (s)	Total Run Time (s) (1%,10%,30%)
Workflow Only	671	7, 67, 201	6	684, 744, 878
Workflow with Provenance	987	<1	<1	988

obtained with the use of PROV standard provenance vs the custom built provenance solutions developed by astronomers as part of their scripts.

One limitation of the approach was determining the probability that the use cases would need to be evaluated as estimated probabilities could only be postulated. The more often these use cases need to be evaluated, the more provenance positively impacts the computational efficiency of the workflow. The results therefore only serve as an estimation of the impact of provenance recording on the computational efficiency of astronomical workflows.

5.6 Conclusions

In conclusion, it has been demonstrated that significant improvements to the processing efficiency of astronomical workflows are possible through the use of provenance. However, the initial cost of provenance recording in this example was too high to produce a positive impact to the workflow's timeliness. Having said this, there are additional methods to reduce the size of this initial overhead which will be investigated in the future such as: changing the type of provenance recorded; identifying more use cases and increasing the size of the parameter space to include different granularities of provenance.

Although one of the limiting factors of this approach was the small parameter space, it did enable brute force evaluation as the total number of possible workflows was small. The next chapter presents a workflow in which this is not the case and the number of workflow configurations is ∼40,000. Therefore, both brute force evaluation and the hill climbing algorithm were investigated as methods for evaluation.

Chapter 6

Finding Transients with Kepler

This chapter applies the approach in order to improve the accuracy and completeness of a transient detection workflow analysing the Kepler Full Frame Images (FFIs). The FFIs are a set of calibration data for use with the Kepler Spacecraft. These were taken each month by Kepler and consisted of 6.25 second exposures, summed up to 30 minutes. Kepler had 42 charged coupled devices (CCDs) which captured images of the sky, each CCD had two channels, therefore each FFI FITS file consisted of 84 individual images. There were 53 FFIs deemed as being good quality meaning a total dataset of 4452 individual images. The Kepler Spacecraft rolls during its observations meaning that different CCDs were used to capture images of the same field, depending upon when the observations were taken. Each CCD comes with its own set of systemic uncertainties and spatial variations in the PSF as this is primarily due to the position of the CCD with regards to the optics. As a result, data taken with different CCDs should be treated as independent datasets.

Although the FFIs were collected to be calibration data, they still offered excellent photometric precision, a wide field of view and several years of evenly sampled monitoring. It is very likely that they therefore contain many previously undiscovered transient objects. In order to better utilise these data and extract these transients, a workflow was required as well as an assessment of the quality of the results it produced. The workflow was designed to search through the FFIs and discover these objects through a combination of difference imaging, aperture photometry and cross matching. Furthermore, an experiment was set up to find the best workflow configuration for each of the components. The quality metrics analysed were accuracy and completeness and determined by the evaluation of the entire parameter space. As this was very resource intensive a method was devised to arrive at an estimated set of best settings very quickly which utilised the hill climbing algorithm. The purpose of this chapter is therefore twofold, to quantitatively find the SExtractor parameters which produce the best quality results (within our chosen parameter space) and to quickly arrive at an estimate of the best parameters with confidence that they will produce good quality results. Furthermore,

the contributions of this chapter are a comprehensive assessment of the quality of all discrete workflow configurations constructed from the parameter spaces and the application and assessment of the hill climbing algorithm to improving the quality of astronomical workflows.

6.1 The Requirements of the Workflow

In order to detect and refine the transient objects within the Kepler FFIs, the workflow shall first perform difference image analysis to produce images containing only objects which vary in brightness. Next, the workflow shall perform source detection and aperture photometry to detect and measure the variable objects within the images. The workflow shall then apply restrictions to these detected objects to produce subsamples of objects which displayed the highest variability. Finally, the workflow shall match the subsamples to known astronomical databases to remove known and variable objects, leaving behind only newly discovered transients.

6.1.1 Difference Image Analysis

Difference image analysis was used to remove all non variable objects present in the Kepler FFIs. The images were first separated into groups which only contained images taken of the same field with the same CCD. The following process was then carried out for each group individually:

1. Align all the images
2. Median combine the aligned images
3. Subtract the median image from the aligned images

The images were aligned using the cross correlation shifts method of the astropy image registration package[1]. The template images were made by median combining the FFIs, this removed the majority of the spurious increases in magnitude that are characteristic of transient sources so that they would be present in single images but not the template image. The images were both aligned and median combined in 100 individual, evenly spaced segments. HOTPANTS (Becker, 2015) was then used to subtract the median images from the original Kepler images, the default HOTPANTS settings were used. One of the difficulties with difference imaging comes from the PSF which can vary greatly dependant on the location with the image - particularly within the Kepler fields. HOTPANTS was therefore chosen for the image subtraction as the algorithm implemented within it

[1] `https://image-registration.readthedocs.io/en/latest/`

creates multiple convolutional kernels for the image. It also matched each 100x100 pixel region of the image individually. Both of these processes help in decreasing the effect of a spatially varying PSF.

Figure 6.1 shows the an example of the image subtraction processor for Kepler image kplr2013038133130 extensions 44, 63 and 79. The original image for each is on the left, the difference image on the right and the histogram of the pixel value in the difference image is in the middle. As Kepler's optical instrument is optimised for photometry and not imaging, it can have a fairly large PSF. This explains the origin of the artefacts within the difference images. Although present, the panels showing the original and difference images in Figure 6.1 are not displayed on the same scale and the amplitude of the artefacts was fairly small. This is demonstrated by the histogram of pixel values within the difference image, shown in Figure 6.1, as the mean value of the difference images is zero and there is a fairly small spread in the distribution of the pixel values.

6.1.2 Source Detection and Aperture Photometry

SExtractor (Bertin and Arnouts, 1996) was used on the difference images in order to identify the variable objects present. SExtractor has of order fifty different categories for customisation which can impact the results of the aperture photometry. Many of these are simple to decide as they are characteristics of the telescope such as: the magnitude of the zero point, the gain and the pixel scale. These were set to 25 mags(Aigrain et al., 2015), $110e^-$ [2]and 4 pixels [3], respectively. Other settings are inconsequential to the transient detection such as naming conventions and aperture dimensions for photometry. These settings (excluding the naming settings) were left as default. The three most relevant settings with regards to transient detection were identified as: the detection threshold (in σ), the minimum number of pixels to be above this threshold and the filter applied to the detection. The filter was used to smooth the image based on the shape of the light profile - equivalent to convolving the image with the relevant function, the typical function is to remove noise for the image and it helps to detect faint, extended objects. The options for this setting were the default kernel, Gaussian, tophat or Mexican hat. Each filter shape had different options for pixel by pixel size and it is recommended to use a size similar to that of the PSF. Typically, once an astronomer has identified these settings, they would either be selected based on prior experience or estimated through visual inspection of each of their results. Each case contains no quantitative measure of how suitable the chosen settings were for their application. Therefore we investigated the selection of these settings, discussed in Section 6.3.

[2] `https://keplergo.arc.nasa.gov/CalibrationSN.shtml`
[3] `https://keplerscience.arc.nasa.gov/the-kepler-space-telescope.html`

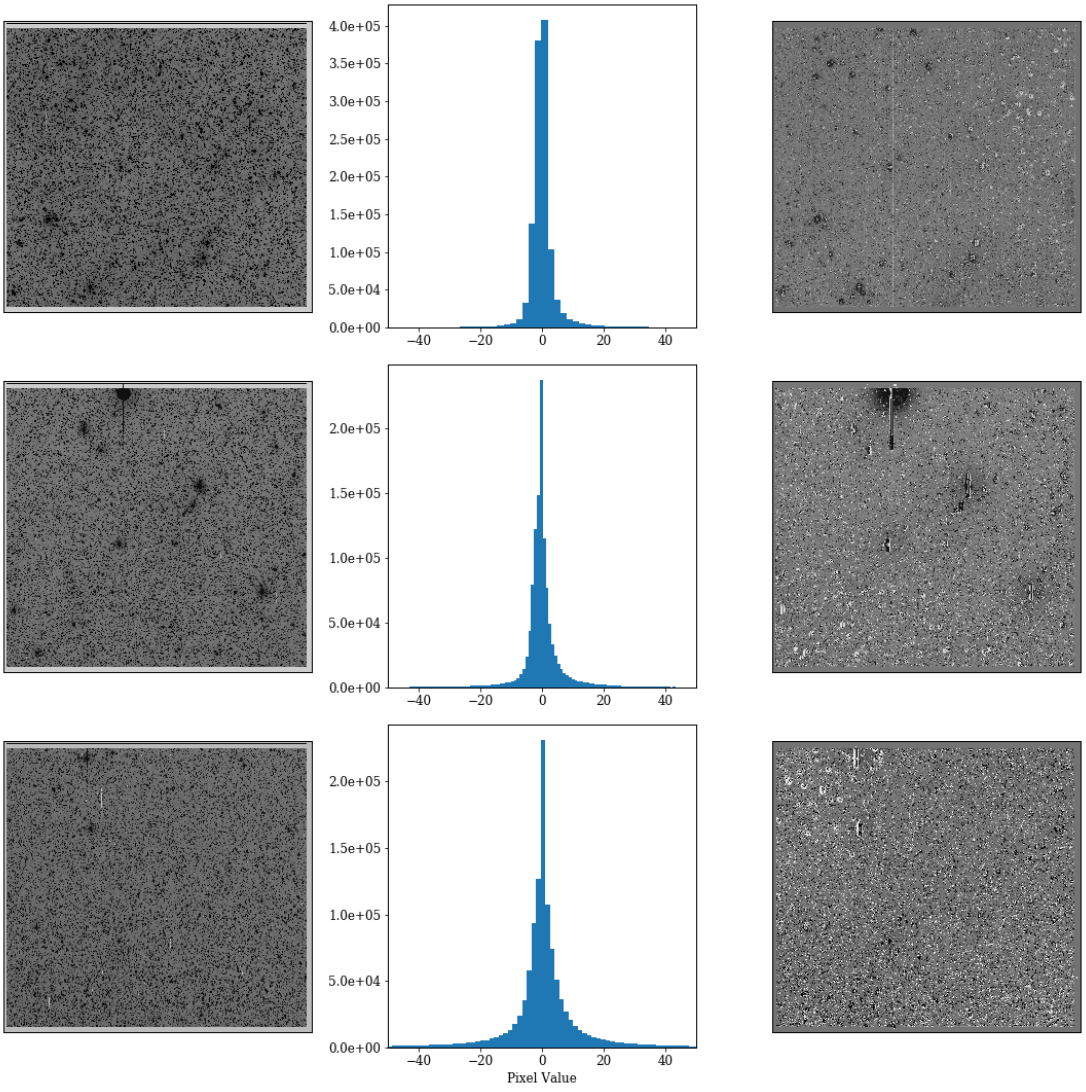


FIGURE 6.1: An example of the difference imaging processor when applied to the Kepler image kplr2013038133130 on channel 44 (top), 63 (middle), and 79 (bottom). The original Kepler image is shown on the left, the difference image on the right and the histogram of the difference images is in the middle. It should be noted that the scales for the original and difference images are not the same. They were changed to emphasis the defects within the difference image. The histogram demonstrates that the amplitude of the artefacts in the difference image is generally small.

6.1.3 Refining the Objects

Once the appropriate settings were selected, the source detection was run on all images and difference images. The objects found within the difference were then reduced into three subsamples by restricting them to only include objects which displayed either: an increase in magnitude of one; an increase in magnitude of three; or a signal to noise ratio (SN ratio) above 10. Each subsample produced had its own benefits. The SN ratio subsample contained a statically significant sample of variable objects but tended to favour the selection of bright objects with reasonable increases in brightness rather than the increase itself. Whereas, the magnitude restricted subsample was composed entirely of objects with a specified increase in brightness which has relevance outside of the Kepler dataset.

The magnitude increase between the median and original images was found by cross matching objects between the original Kepler image and the corresponding difference image. An object was considered a match only if there was less than a 2.5" difference in the object location between both images, representing sub-pixel accuracy. The difference image was produced by subtracting the median image from the original image, therefore:

$$C_D = C_I - C_M \tag{6.1}$$

where C_D, C_I and C_M represent the counts of an object in the difference image, original image and median image respectively. By extension:

$$C_M = C_I - C_D \tag{6.2}$$

The equation for the difference in magnitude between the median and original image can therefore be written as:

$$\delta M = 2.5.log_{10}\frac{C_I}{C_I - C_D} \tag{6.3}$$

This method was chosen to find the increase in magnitude over attempting to find a counterpart in the median image as transients did not always have one. The median counts would then have to be estimated by finding the counts within an aperture around where the object was expected to be and would result in a less reliable measurement.

The signal to noise ratio was calculated according Equation 6.4 where SNR is the signal to noise ratio, N is the noise, n is the total number of images combined to make the median image, S_D, S_I and S_M are the signals in the difference, original and median images, respectively.

$$SNR = \frac{S_D}{N} = \frac{S_D}{\sqrt{S_I + (\frac{S_M}{n})}} \tag{6.4}$$

The signals were equal to the counts divided by the gain of $110\,e^-$. The noise in the median image is reduced by a factor of $\frac{1}{\sqrt{n}}$ due to the dampening of random fluctuations

when median combining multiple images. Using Equation 6.2, we can transform Equation 6.4 to Equation 6.5.

$$SNR = \frac{S_D}{N} = \frac{S_D}{\sqrt{S_I + \frac{S_I - S_D}{n}}} \tag{6.5}$$

The signal from the original and median images include contributions from the sky background, thermal currents and read noise which are included in this calculation of the noise. Their contributions contain Poisson noise due to the random nature of the motion of discrete packets of photons. No additional noise should be accounted for the difference image as its formation was not a random process, it was a combination of the other two images whose noise has already been accounted for. Additionally, as it is a subtraction of the other two images, the extra sources of signal, such as the background should already be subtracted and the flux measured within the difference image can be attributed to the desired source alone.

6.1.4 Matching to Astronomical Databases

The next process in the workflow was to match the potential transients with catalogues of known astronomical objects and remove all variable objects from the sample. This step was also used to confirm that the workflow was producing the expected results and verify that it was functioning as intended. The catalogues chosen were Gaia and Simbad. Objects were considered a match if they had a position within 5" of the catalogue distance within Simbad and 2.5" within Gaia. The 5" search was chosen as it was representative of the average expected Kepler PSF. This was reduced for Gaia searches due to the extra Galactic crowding expected from this catalogue.

The final step to remove variable objects from our detected sample was to search through the Kepler observations and determine whether the object has been viewed in multiple Kepler images. If it had, light curves were constructed and searched for possible periodicities. If the object had only one observation and no discernable period, it was likely to be a transient event.

6.2 The Quality of Results Produced by the Workflow

Before the quality of the transient detection workflows results can be calculated, good quality must first be defined within this context. The goal of the workflow is detection, therefore the completeness is important to recover as much of the potential sample as possible. However, if the workflow had very relaxed requirements for an object to be considered real, it will have a very high completeness but also be contaminated with a

high number of false positives. The completeness must therefore also be balanced with the accuracy which in this example needs to be proportional to the false positive rate. The relation between these two metrics and the overall quality is shown in Equation 6.6, where Q is the quality, C is the completeness metric, A is the accuracy metric and W is a coefficient, with a value between 0 and 1, where 1 gives a high importance to the completeness and 0 to the accuracy. C and A were normalised to also have a value between 0 and 1, such that 0 indicated the highest quality and 1 the lowest.

To investigate the quality, Kepler image kplr2013038133130 was chosen at random to be representative of the full sample. Within this image, channels 44, 63 and 79 were selected to incorporate a range of the expected PSFs. The PSF for a channel is dependant on the channels location as the primary source of point source spread is from the optics. Therefore, the channels were selected such that each resides in one of the three concentric rings of CCDs within the Kepler field of view, shown in Figure 6.2.

$$Q = WC + (1 - W)A \tag{6.6}$$

6.2.1 Instantiation of the Approach

The schematic of the workflow in this investigation is shown in Figure 6.3. The mapping of the components of the approach to this workflow is as follows:

- Motivation - To find transient objects within the Kepler FFIs.
- Workflow Overview - Difference image the FFIs, perform aperture photometry to find variable objects, cross match these with astronomical databases to help determine if it was a transient.
- Inputs - HOTPANTS settings (H_s), SExtractor settings, Simbad Database
- Processors - Image aligning (Astropy), Difference imaging (HOTPANTS), transient detection (SExtractor), transient filtering, matching to databases, removing variable objects (lightcurve inspection)
- Parameter Space - Settings within SExtractor.
- Actual Result - The number of transients correctly recovered ($Tr_{aug} - Tr_{orig}$) and the total number of objects found in the image (N_x).
- Expected Result - The total number of transients inserted in the image (Tr_{tot}) and the total number of objects recovered with the least strict settings (N_{max}).
- Utility Function - The fraction of inserted transients recovered combined with the normalised accuracy with a weight, W, with a value between 0 and 1 that describes

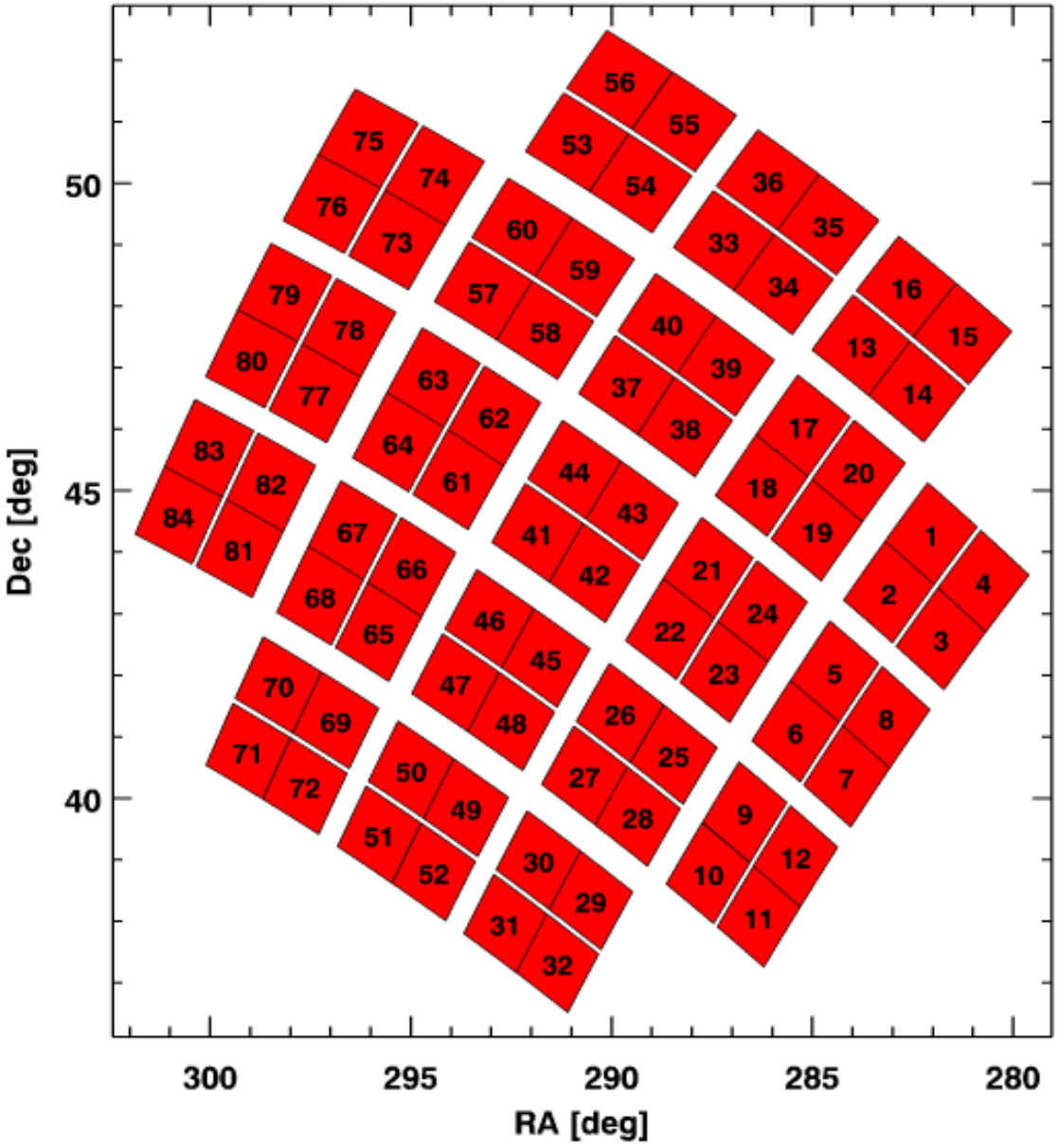


FIGURE 6.2: The field of view of the Kepler Spacecraft with each CCD channel numbered.

the relative importance of the two data qualities. Therefore:
$U_{qm} = (W \frac{Tr_{aug} - Tr_{orig}}{Tr_{tot}}) + (1 - W)(1 - \frac{N_x}{N_{max}})$

- Data Quality Metric - Completeness and Comprehensiveness & Accuracy and Precision

6.3 Investigating Versions of the Workflow

As previously discussed, the chosen settings for SExtractor can have a large impact on the quality of the source detection but determining the best settings to use can be difficult,

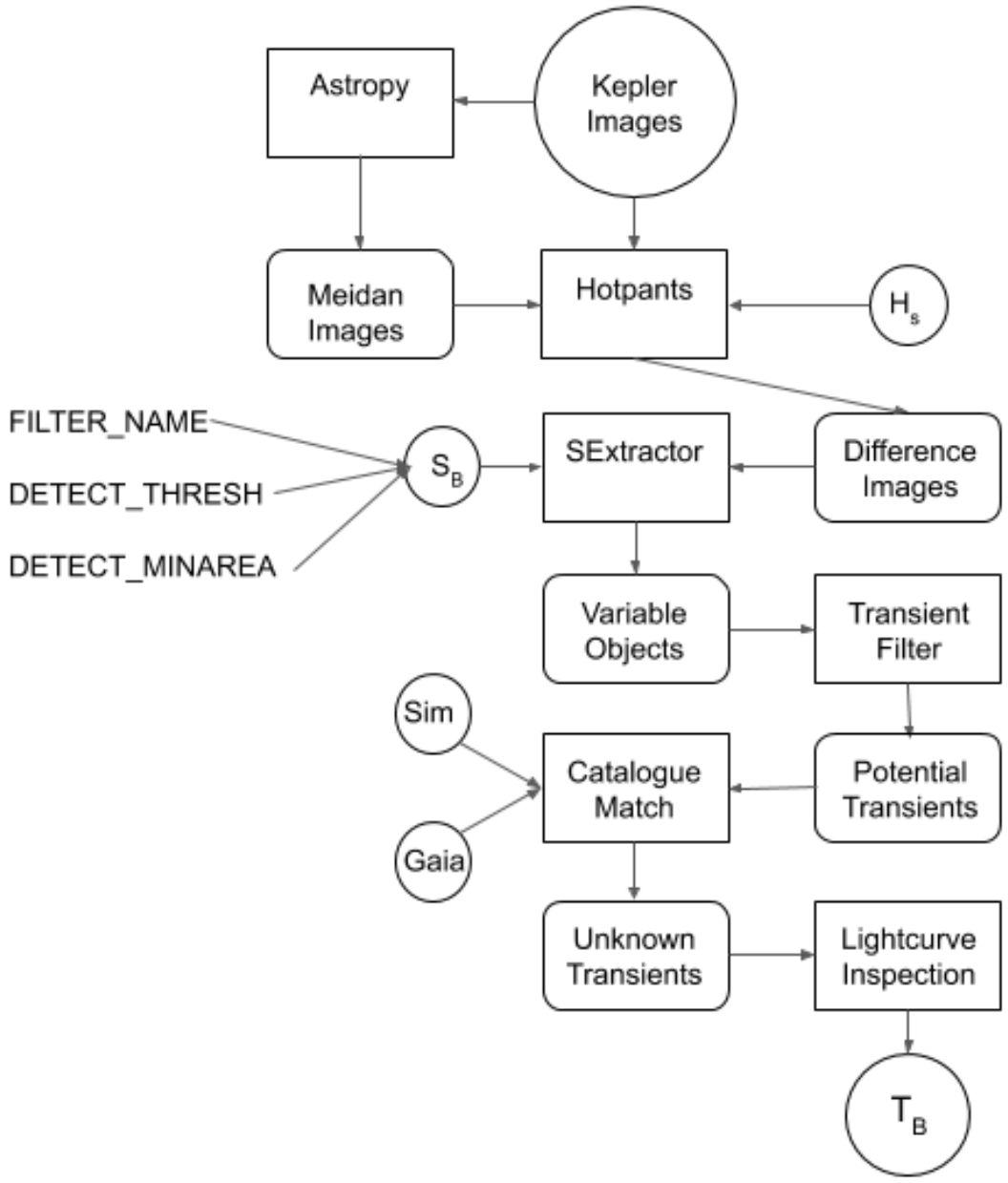


FIGURE 6.3: Schematic of the workflows main components for a single images.

especially for a workflow such as this where the goal is to discover a wide variety of unknown objects. The relevant SExtractor parameters for transient detection are named DETECT_THRESH, DETECT_MINAREA and FILTER_NAME within the context of the SEXtractor settings. The outlined parameter space for DETECT_THRESH was 1,2,3,4,5,6,7,8,9,10 and 1,2,3,4,5,6,7,8,9 for DETECT_MINAREA. These ranges were decided upon after discussion with an astronomer as to what the total relevant parameter space was for transient detection in this dataset. The FILTER_NAME parameter space was composed of every possible SExtractor filter which included different size kernels for Gaussian, Mexican hat and tophat profiles as well as the default convolutional kernel.

Further details of these filters can be found in Appendix B. This section is composed of two main parts: the first is dedicated to quantitatively finding the SExtractor parameters which produce the best quality results (within our chosen parameter space); the second is dedicated to quickly arrive at an estimate of the best parameters, followed by an evaluation of the quality of the results produced.

6.3.1 Evaluating the Completeness

Completeness within astronomical images is not easily measured as there are no ground truths - there is no way to verify whether each potential transient object was real nor know how many there were in the image. One common method to circumvent this problem is to add simulated objects to the image and recover those objects. This was applied to our images using the IRAF MKobjects routine (Tody, 1986). An alternative method is to entirely fabricate images but purely simulated images lose all of the systematic defects characteristic to it.

MKobjects uses a PSF model to generate artificial objects that are characteristic of those found in the image. As the PSF varies over the Kepler images, the PSF was separately calculated for 16 different, evenly spaced segments within the original Kepler images (4x4 squares). The decision for this number of segments was motivated by the conflicting arguments for small segments which provide good spatial PSF resolution and segments large enough to contain a reasonable sample of high quality targets to form the PSF (bright point sources but not saturated). Ultimately, the compromise was reached by finding the most possible sections which still produced a PSF that resembled the objects in the image. IRAF's PSFmodel function was used to generate the PSF within each section and these PSF models were then consumed by MKobjects to generate the simulated stars. 992 simulated objects were added to each difference image. 62 objects were simulated for each section, using the corresponding PSF and inserted with a random spatial distribution. To ensure that simulated objects from one section were not detected in neighbouring sections, a 10 pixel boundary between each segment edge was placed in which there were no objects added. In order to investigate the relation between the magnitude of the objects and the quality of transient detection, 7 separate images were made for each difference image, each contained 992 object randomly distributed in magnitude within the selected magnitude range for that image. These ranges began at -7,-6,-5,-4,-3,-2 and -1, each were one magnitude in width. To determine the magnitude that these ranges corresponded to as measured by SExtractor, it was evaluated over each augmented image to find the corresponding apparent magnitude. Figure 6.4 shows histograms of the observed magnitude distribution within the images. The median values of each magnitude range are as follows: 11.7, 12.7, 13.7, 14.7, 15.7, 16.7, and 17.6 for magnitude ranges -7,-6,-5,-4,-3,-2 and -1, respectively.

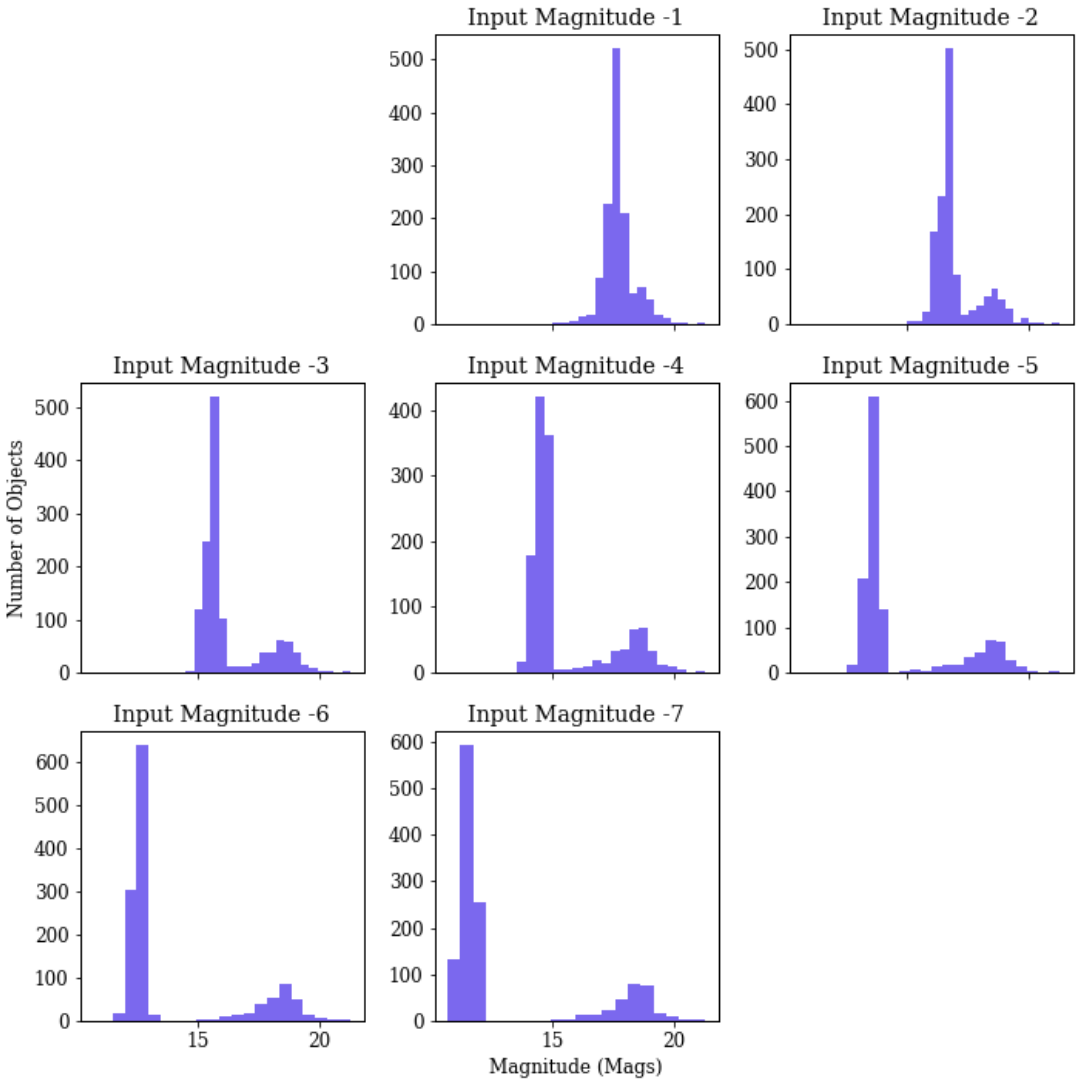


FIGURE 6.4: Histogram showing the distribution of magnitudes of objects within the Kepler image kplr2013038133130[44] after difference imaging and insertion of simulated objects from each of the seven magnitude ranges.

To determine the completeness, the transient detection workflow was evaluated over these augmented images and the objects found were cross matched with the locations of the simulated objects to determine which were recovered. This process was repeated for each possible combination of SExtractor parameters. Objects were considered a match if they were within half a pixel of the simulated position (2"). In order to remove the sample of objects that were present in the image before augmentation, the same procedure was performed on the original difference image, with the same SExtractor settings as the augmented image and the total number of objects which matched by coincidence was subtracted from the total recovered from the augmented image. Finally, the maximum possible value for completeness was 62 for each region, any beyond this would not be counted.

The final value for the completeness quality metric can then be calculated as in Equation 6.7, where Tr_{aug} and Tr_{tot} were the number of objects found in the augmented image and simulated stars list, respectively.

$$C = 1 - \frac{Tr_{aug}}{Tr_{tot}} \tag{6.7}$$

6.3.2 Evaluating the Accuracy

A good metric for the accuracy of the workflow is the false positive rate - the ratio of real results to spurious ones. However, determining this value is hindered by the same problems as with the completeness - the lack of ground truths. Furthermore, determining the accuracy can not be accomplished by inserting simulated objects as it is still unknown which objects inherent in the image are true or false. Simulating entire images again has the caveat of removing all of the images defects and systematic uncertainties. Therefore, a proxy was chosen to represent the accuracy which is proportional to false positive rate.

The first step in generating the proxy was to evaluate SExtractor over the images using the settings within the parameter spaces which would accept the highest number of objects as transients and record the total number found (N_{max}), these were: DETECT_THRESH = 1, DETECT_MINAREA = 1 and FILTER = mexhat_1.5_5x5.conv. Then for each x separate version of the workflow, the total number of objects recovered (N_x) when using that specific combination of SExtractor parameters was also calculated. The estimate for accuracy was then calculated using Equation 6.8. This proxy now represents the ratio between the total objects recovered with the chosen settings and the total objects it was possible to recover with the most relaxed settings. As with the completeness, A was a value between 0 and 1 where 0 was the highest accuracy and 1 was the lowest. Maximum accuracy with this metric corresponds to the detection pipeline finding zero objects, which as with the case for the completeness metric, is clearly not the optimum workflow. This emphasises the importance of balancing these two mutually exclusive quality metrics to find the overall best quality workflow.

$$A = \frac{N_x}{N_{max}} \tag{6.8}$$

6.3.3 Evaluation of the Approach

This approach was evaluated using two different methods to achieve the two goals of the investigation. To find the region in parameter space which produced the highest quality workflow, all possible configurations of parameters were evaluated via brute force and each quality metric was calculated for every discrete workflow configuration. To quickly arrive at the approximate solution, hill climbing was used. The results produced by the

hill climbing algorithm were then compared with the brute force results to deduce how good the quality of the estimated solution was.

All computation was performed on a machine running Ubuntu 16.04 with the following specifications: 12GB of system memory; an Intel®Xeon™ W3520 CPU @ 2.67GHz The brute force evaluation of the 36,960 workflow versions within the outlined space took ∼50 hours.

6.3.4 Brute Force

The Evaluation of all discrete workflows involved the execution of 1890 (10 DETECT_THRESH, 9 DETECT_MINAREA, 21 FILTER) discrete workflows for 7 magnitude ranges across 3 different Kepler images, totalling 39,690 individual workflows. Therefore, this was a computationally expensive task with high storage requirements. The computation was all carried out using a Spark cluster with 6 cores, 500 Mb of executor memory per core and 1 Gb of driver memory per core. Spark dataframes were chosen for the analysis for their efficient handling of multiprocessing and fast in-memory data processing. In order to transfer the data processing from python scripts to Spark dataframes, one change was made to the cross matching of objects. Instead of finding a maximum radial separation in arcsec between the two objects, an inner join was performed between the columns containing the x and y pixel positions of the detected objects and four columns which were ±0.5 pixels of both the known x and y position for each object. This effectively searched a box with sides of length 1 pixel, centred around the found object instead of a radial search. This small difference made a negligible impact on the search quality whilst providing significant increases in computational efficiency. The data was stored on a MySQL database making use of the python packages PySpark and mysqlconnector to interoperate the processing and storage.

The spatial dependence of the accuracy was not investigated as the metric relied almost entire on the distribution of the original sources within the image which is not a known quantity. However, the completeness metric had been constructed to be somewhat consistent between the 16 different sections of the images that were tested. Figure 6.5 displays the relation between the spatial distribution and the completeness, averaged over all SExtractor settings, magnitudes and images. The top row is the completeness metric when measured in that section of the image and the bottom row shows the corresponding FFI.

As previously discussed, the two quality metrics were thought to have conflicting requirements from the settings. This behaviour was observed across all of our simulated images and this effect is demonstrated in Figures 6.6 and 6.7. These graphs show the quality of the completeness and accuracy vs the detection threshold (6.6) and the detection minimum area (6.7) averaged over all other SExtractor settings, image positions

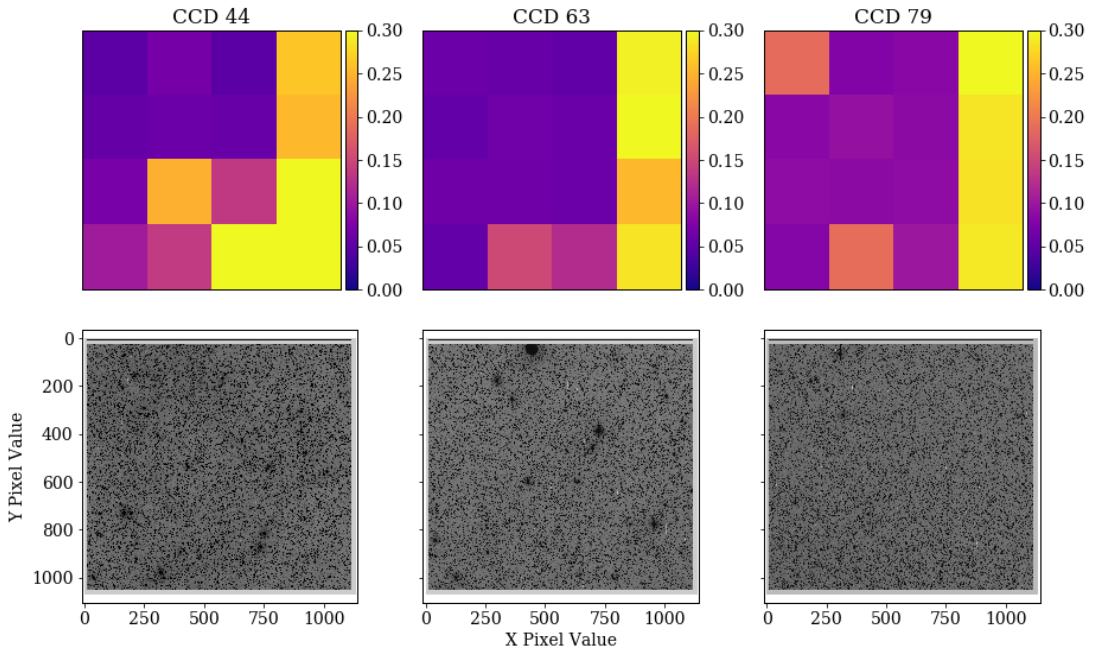


FIGURE 6.5: A figure displaying the relationship between spatial distribution and completeness of transient recovery within the Kepler FFIs. Top depicts the completeness for each of the 16 sections tested, 0 representing maximum completeness and 1 minimum. The bottom row depicts the corresponding Kepler FFI.

and magnitudes for Kepler channels 44, 63 and 79 (from left to right in each Figure). The quality measure for each metric is maximised at zero and minimised at 1. These graphs therefore demonstrate that the higher restrictions the workflow imposed for an object, the higher the accuracy and the lower the completeness and vice versa.

Although these graphs have some distinguishing trends, they do not show the effect of the interactions between different SExtractor settings. To display these, 3-D surface plots were made, one for each filter type (Gaussian, Mexican hat, tophat) with all sizes of that filter plotted. The x and y axes were the detection threshold and detection minimum area, respectively. The z axis was the quality metric - either completeness or accuracy. The quality metrics for each SExtractor setting combination were then averaged over the spatial distribution, image, and magnitude to produce each graph

Figures 6.8, 6.9 and 6.10 show the relation between detection threshold, minimum area and accuracy of the transient detection workflow for the Gaussian, Mexican hat and top hat filters, respectively. The x, y and z axes correspond to the detection threshold, minimum area and accuracy, respectively. Each figure contains all different sizes of the filter which were tested, each represented by an individual slice. Each slice was plotted on its own figure in Appendix B.2. Each of the figures demonstrated that the higher the detection threshold and minimum area, the better the accuracy. The accuracy is also

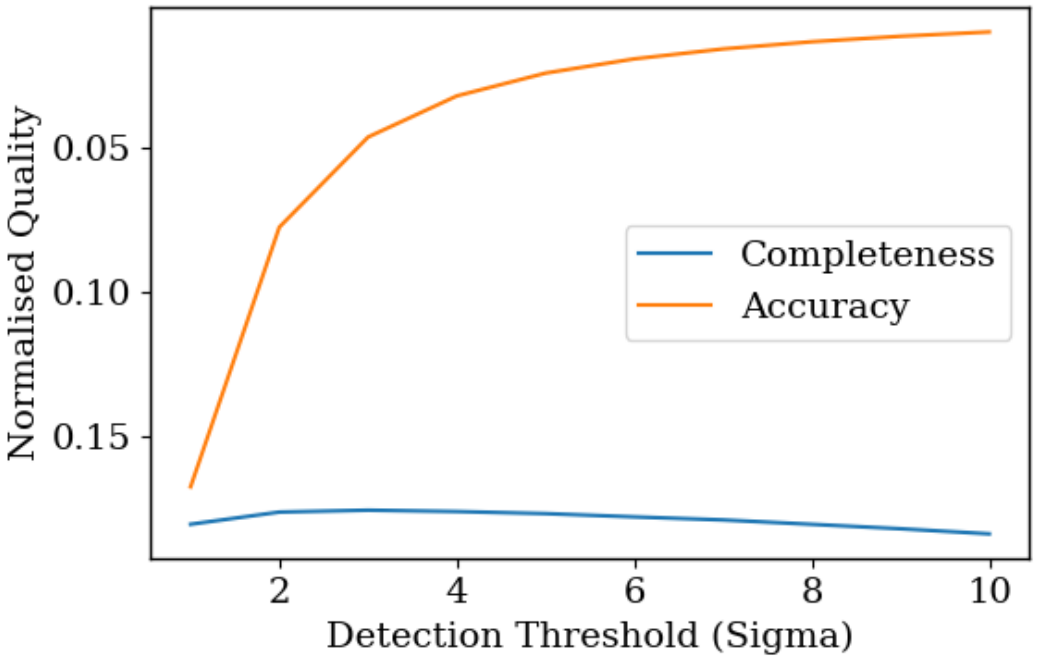


FIGURE 6.6: Schematic of the workflows main components for a single images.

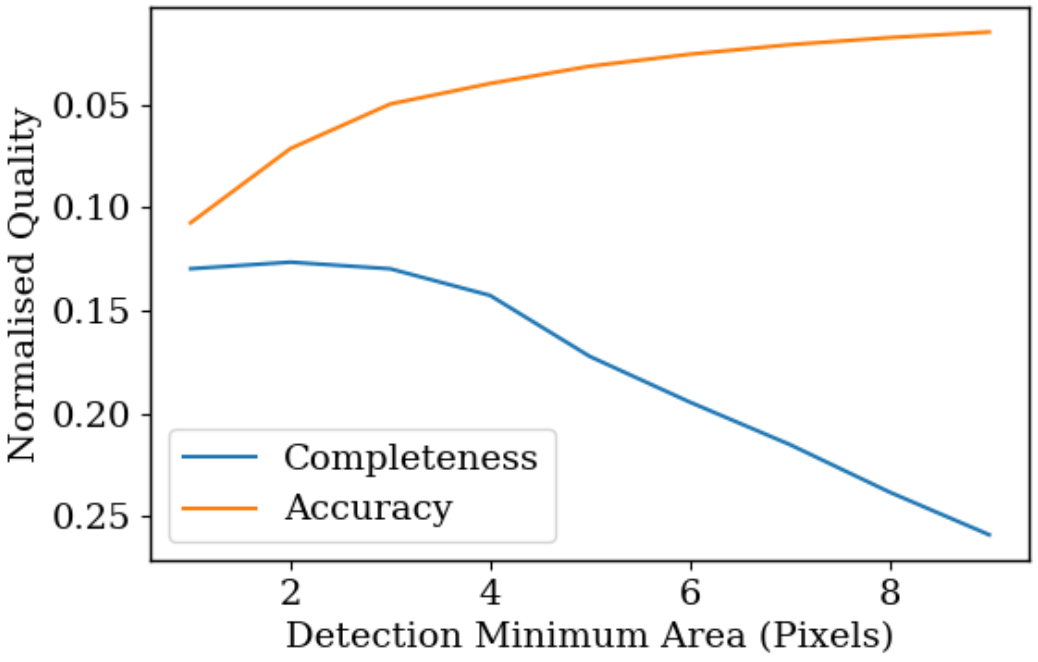


FIGURE 6.7: Schematic of the workflows main components for a single images.

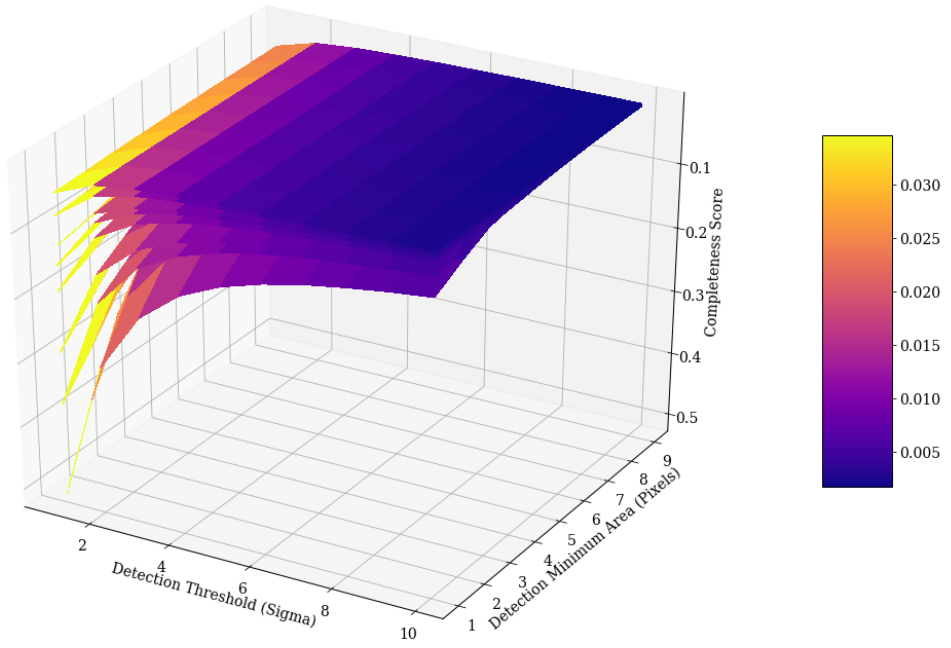


FIGURE 6.8: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for all sizes of the Gaussian filter. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

observed to be more dependant on minimum area than detection threshold but both are important for achieving the highest accuracy.

Figure 6.8 shows the 3-D accuracy plot for all sizes of the gauss filter, each represented by a single slice. This figure demonstrates the expected results, the lower the restrictions on what a transient is, to more objects get detected which in turn lowers the accuracy. Although both the detection threshold and minimum area contribute to this effect, it is far more dependant on the minimum area than the threshold.

Figures 6.11, 6.12 and 6.13 show depict the same information as in Figures 6.8, 6.9 and 6.10 except for completeness, instead of accuracy. These figures again represent the Gaussian, Mexican hat and top hat filters, respectively.

Within Figure 6.11 the overall trend is that the completeness is best with smaller values for both detection threshold and minimum area. However, the best value for completeness is not at the minimum value for each setting as it improves in quality from a detection threshold of one to three. One explanation for this is that as the requirement to be a transient is so low, it is detecting the outer regions of the objects as transients and not the center such that this new position is no longer included within the cross matching tolerance and it is not accepted to be one of our input objects.

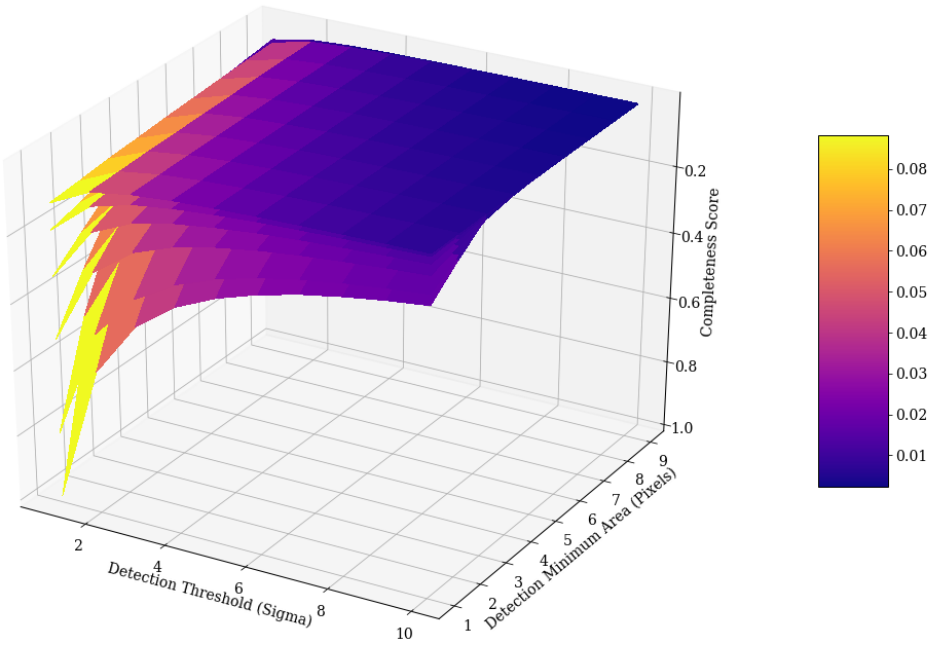


FIGURE 6.9: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for all sizes of the Mexican hat filter. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

The results within Figure 6.12 are fairly poor. The completeness has little relation with the detection threshold but closely linked to the detection minimum area. The best results with regards to completeness are found with a small detection minimum area and as this value increases, the completeness sharply decreases. It is clear from Figure 6.12 that the Mexican hat filter is the wrong choice for the Kepler FFIs.

In Figure 6.13 the completeness with the tophat filter shows a negative correlation between increasing detection threshold and minimum area. The worst completeness is found with the highest values of both, whereas the lowest is found with a combination of low values for the two but not the minimum value, this may be due to the same effect as with the Gaussian filter. Figure 6.13 also shows that high minimum area or threshold alone does not decrease the completeness, it the combination of high values for both.

To demonstrate the effect of both detection threshold and minimum area on the overall quality of the workflow, the values for quality of completeness and accuracy were combined in Equation 6.6 with weights ranging from 0.1-0.9, in steps of 0.1. This utility function was then averaged over filter, spatial distribution, magnitude and CCD channel to demonstrate the relationships between the utility function, the detection threshold and the detection minimum area. The results are shown in Figure 6.14. When a small weight is used, a high detection threshold and minimum area produce the best results as

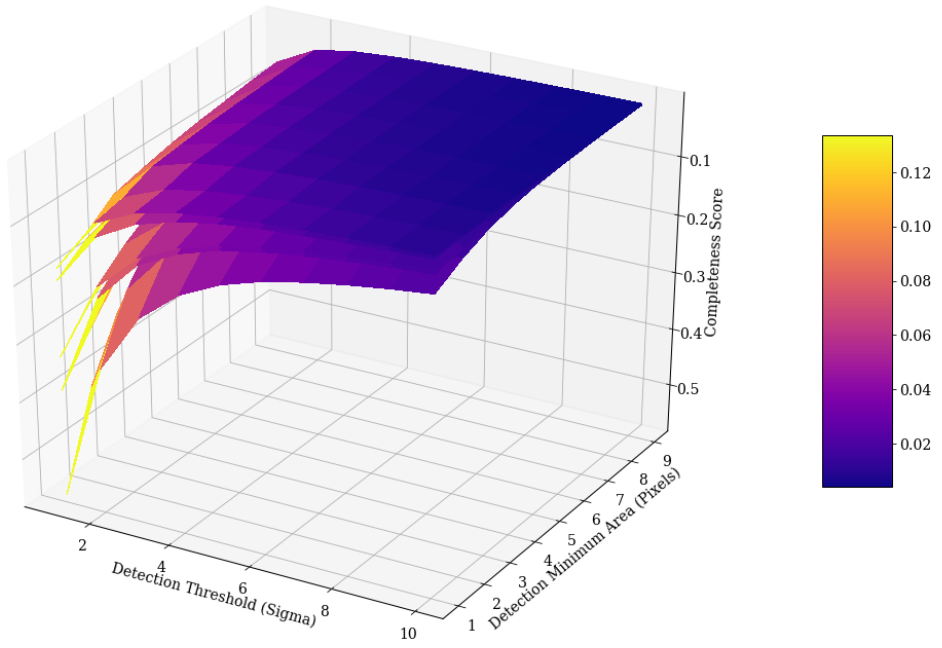


FIGURE 6.10: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for all sizes of the top hat filter. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

Weight	Detection Threshold	Detection Minimum Area	Filter	Average Accuracy	Average Completeness	Utility Function
0.1	10.0	8.0	gauss_4.0_7x7.conv	0.007	0.143	0.021
0.2	10.0	7.0	gauss_4.0_5x5.conv	0.011	0.120	0.033
0.3	10.0	7.0	default.conv	0.019	0.096	0.042
0.4	10.0	6.0	tophat_2.0_3x3.conv	0.023	0.090	0.046
0.5	8.0	7.0	default.conv	0.025	0.087	0.056
0.6	7.0	7.0	default.conv	0.029	0.084	0.062
0.7	7.0	7.0	default.conv	0.029	0.084	0.068
0.8	7.0	7.0	default.conv	0.029	0.084	0.073
0.9	9.0	4.0	gauss_1.5_3x3.conv	0.046	0.080	0.077

TABLE 6.1: The best workflow configuration found through the brute force simulations for each of the different quality weights.

this increases the accuracy. Whereas, with a large weight, these settings produce a poor quality workflow. When a high importance is placed on the completeness, a relatively small but not the minimum values for these SExtractor settings produce the best results, as with in the 3-D plots. The SExtractor settings which produced the best quality workflow for each weight is shown in Table 6.1.

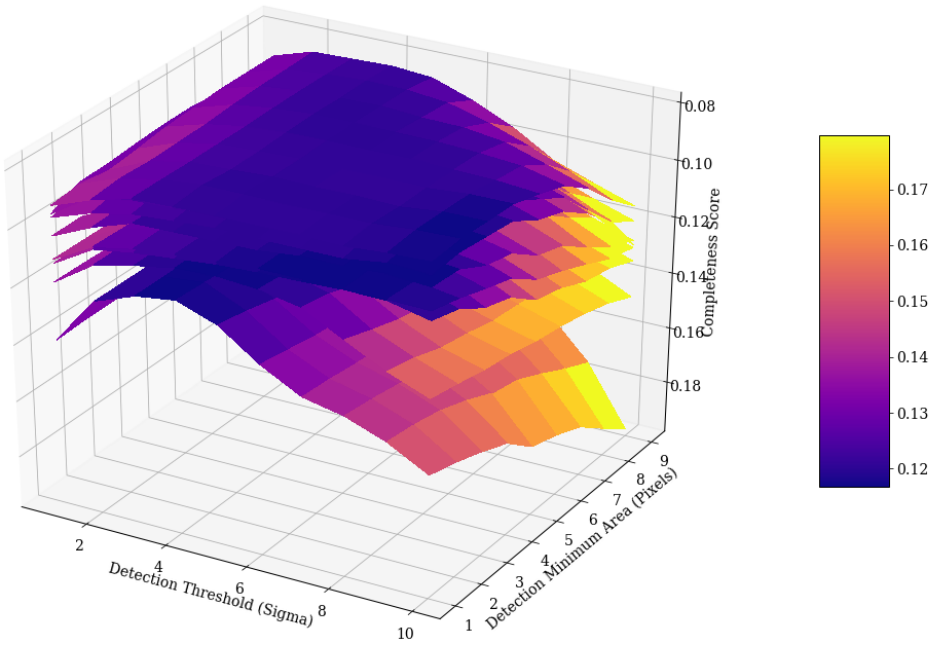


FIGURE 6.11: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for all sizes of the Gaussian filter. The completeness quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

6.3.5 Hill Climbing

The hill climbing algorithm consisted of first choosing a set of starting parameters, chosen to be the minimum and maximum values for each parameter as well as a random selection from each parameter space. Once the initial parameters were chosen, the workflow with those parameters would be evaluated and then the DETECT_THRESH parameter was moved one space up in parameter space then this new workflow was evaluated. If the quality of the second workflow was greater than the first, changes would be continually made upwards in parameter space until the quality was not found to increase. If the quality was found to be worse in the second iteration of the workflow, a negative step in parameter space was made and the same procedure was followed by in the negative direction. The steps would cap out at the minimum or maximum values. Once no further improvement could be found, the DETECT_MINAREA was altered in the same way and finally FILTER_NAME. This process was repeated twice to assess whether changes in the latter two parameters impacted the decision for the former. To avoid finding local optima in the workflow quality, jumps to a workflow configuration situated randomly in parameter space were made at each climb either up or down the parameter space hill. If the quality of the output produced by the workflow with these random settings was greater than that of the current hill, then the algorithm would switch to this region

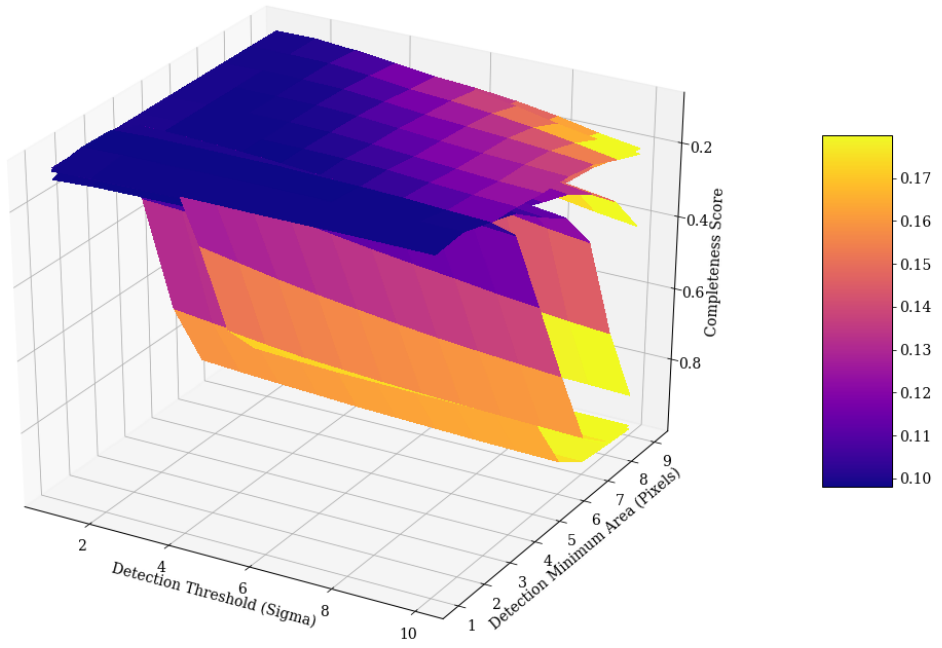


FIGURE 6.12: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for all sizes of the Mexican hat filter. The completeness quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

in parameter space. Once no further improvement in parameter space was found, the algorithm stopped and the best estimated settings had been reached. This was repeated for each image with different magnitude ranges of simulated objects.

The weight, W, was varied from 0.1 to 0.9 in steps of 0.1 providing the estimate for the best quality workflow when accuracy and completeness were valued in different ratios. Each hill climb evaluation took an average of 90 seconds to run, and this was repeated for each starting range, Kepler channel and magnitude range of the simulated stars.

The hill climbing algorithm was evaluated over each image which had a discrete combination of magnitude and Kepler channel within the parameter space. This was repeated nine times, changing the weighting of the quality metrics each time. At the end of the hill climbing algorithm, it produced a selection of settings which it determined was the best for evaluating images in the tested region of parameter space with the specified weight. An example run of the algorithm is shown in Figure 6.15 which shows the changes in the quality of the data processing as the algorithm proceeds and the final settings to be chosen. The algorithm continues to test settings until no improvement is found which explains why the algorithm does not end as soon as the best settings have been found.

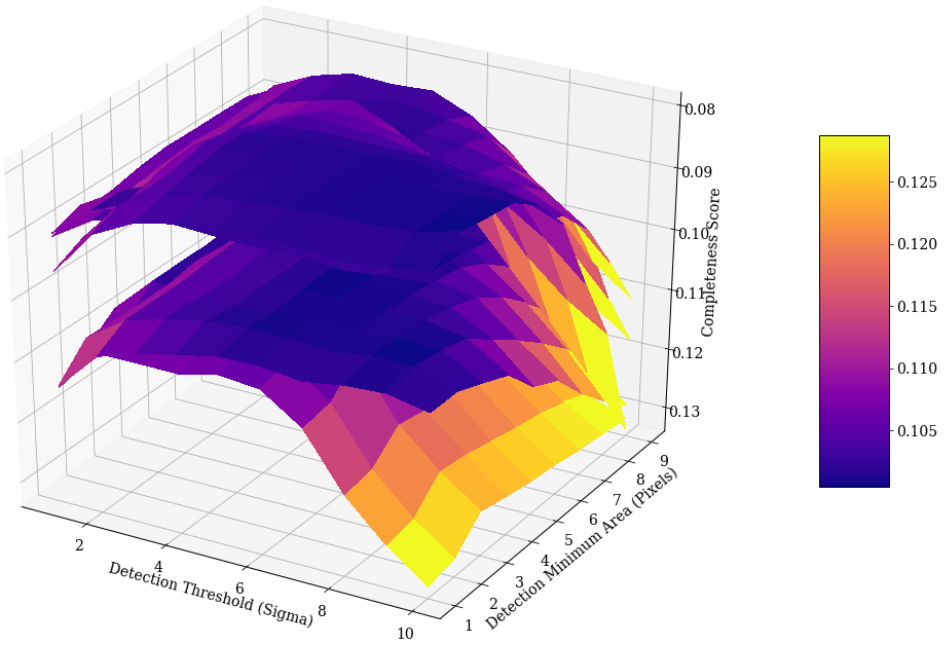


FIGURE 6.13: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for all sizes of the top hat filter. The completeness quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

To determine which of the starting parameter sets performed the best, the score of the selection of parameters returned by each for each region in image parameter space was subtracted from the best possible score produced by settings in that space as determined by the brute force simulations. The mean and standard deviations of these differences in quality are displayed in Table 6.2. The best results were found to be when using the minimum starting parameters by a reasonable margin, making them the choice of starting parameters for the hill climbing algorithm. Figure 6.16 shows the quality of the output produced by the workflow when evaluated on each Kepler channel, simulated star magnitude range and relative quality weighting for the: average and median of all combination of SExtractor settings, shown in green and red, respectively; the settings produced by the hill climbing algorithm, shown in blue; and the best possible settings found with the brute force simulations, show in orange. For all images within the parameter space, the settings produced by the hill climbing algorithm perform better than the median and average results. They also closely follow the quality of best possible results from the brute force simulations. The mean difference between these hill climbing qualities and the brute force qualities was -0.0013. The same difference between the hill climbing qualities and the mean and median qualities were 0.058 and 0.020, respectively.

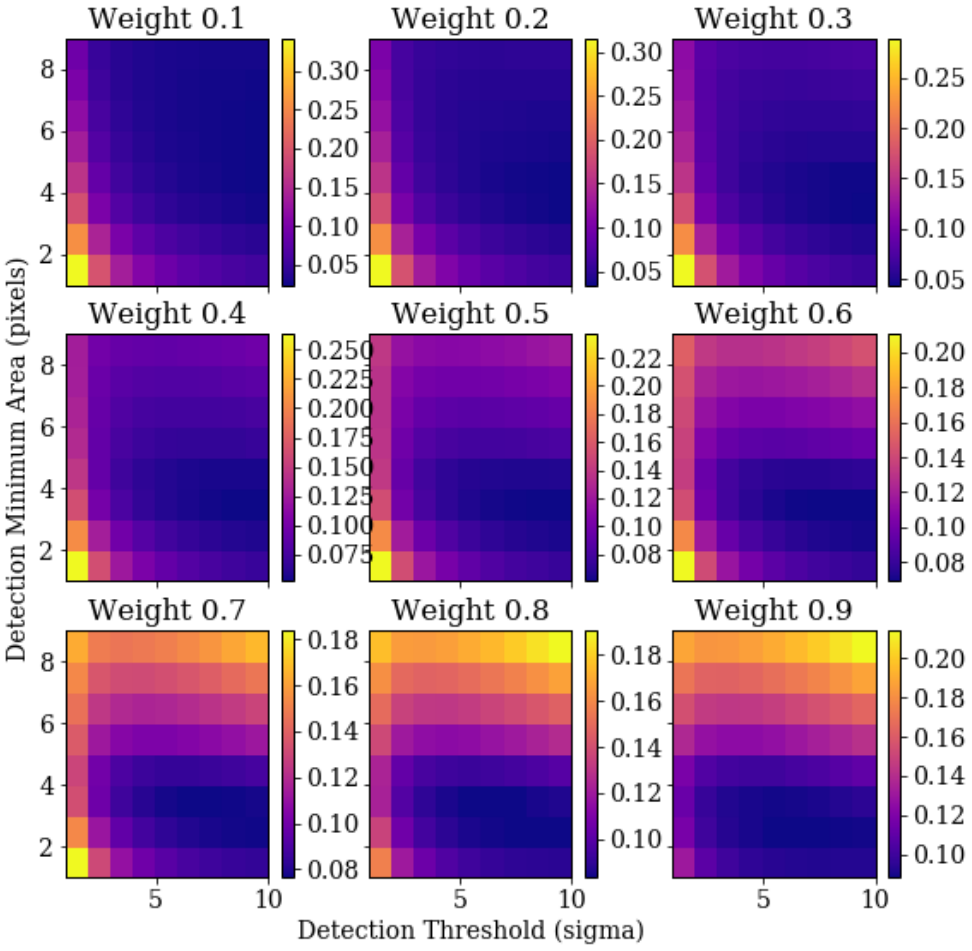


FIGURE 6.14: Colour maps showing the relation between the utility function, detection threshold and detection minimum area when using values for the weight from 0.1 to 0.9. These results were averaged over the spatial distribution, magnitude range, SExtractor filter and CCD channels.

Starting Parameter Set	Mean Difference	Standard Deviation
Minimum	-0.0013	0.0025
Maximum	-0.046	0.076
Random	-0.088	0.0529

TABLE 6.2: The mean and standard deviation of the difference between the quality scores produced by best settings found by the hill climb algorithm when using different starting parameters and the best possible settings as found by the brute force algorithm.

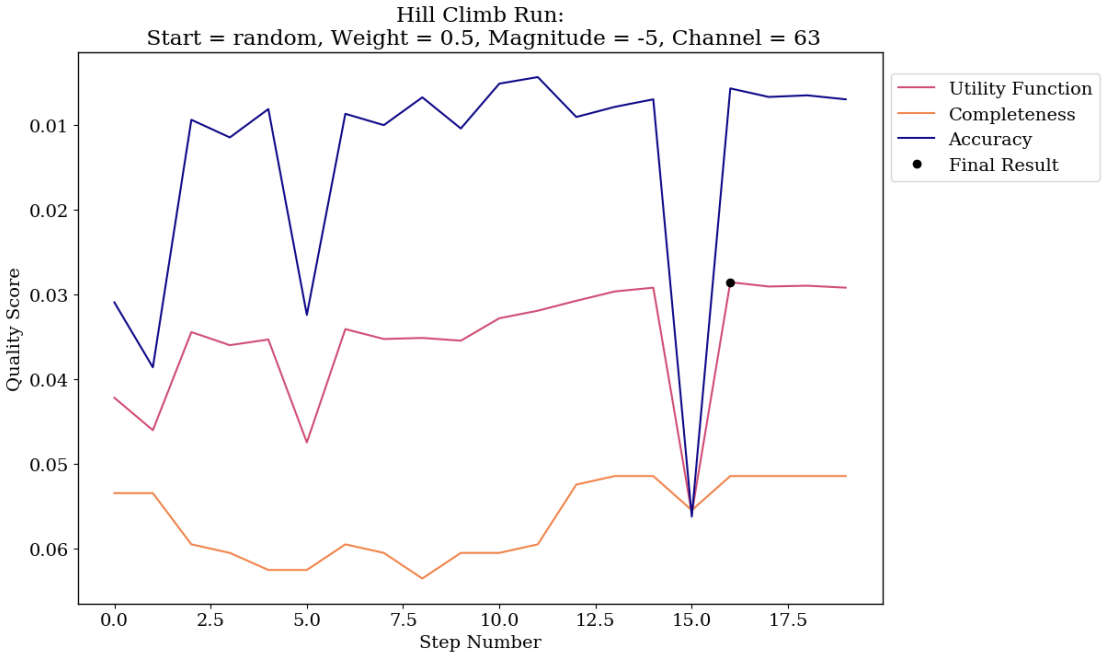


FIGURE 6.15: An example of the path that the hill climbing algorithm too through parameter space. The quality of the results are plotted after each evaluation of the workflow, the completeness is shown in green, the accuracy in orange and the utility function in blue. The quality of the final results is represented by the red dot. These results are for the region in parameter space where: the starting values were random, the weight was 0.5, the magnitude was -5 and the Kepler channel was 63. The y axis denotes the quality measure and the x axis represents each step taken within this hill climb.

6.4 Finding Transients Using the Workflow

A weight of 0.5 was selected as the ideal quality for transient detection as both components are important. Therefore, the SExtractor settings used were: DETECT_THRESH = 7, DETECT_MINAREA = 7 and FILTER_TYPE = default.conv. To test the accuracy of the magnitudes produced by these settings, SExtractor was evaluated over Kepler image kplr2010078174524[25] (chosen randomly) and the objects detected were cross matched with the Gaia catalogue. Gaia was chosen as it uses a similar broad G filter to the one used in Kepler. The SExtractor magnitude was then plotted against the Gaia phot_g_mean_magnitude and is shown in Figure 6.17, median separation between the two was found to be 0.49 mags.

The total number of objects found in all difference images was 2,091,872. A histogram was constructed displaying the distribution of magnitudes of this total sample, shown in Figure 6.18. The mean magnitude of all detected objects was 17.9.

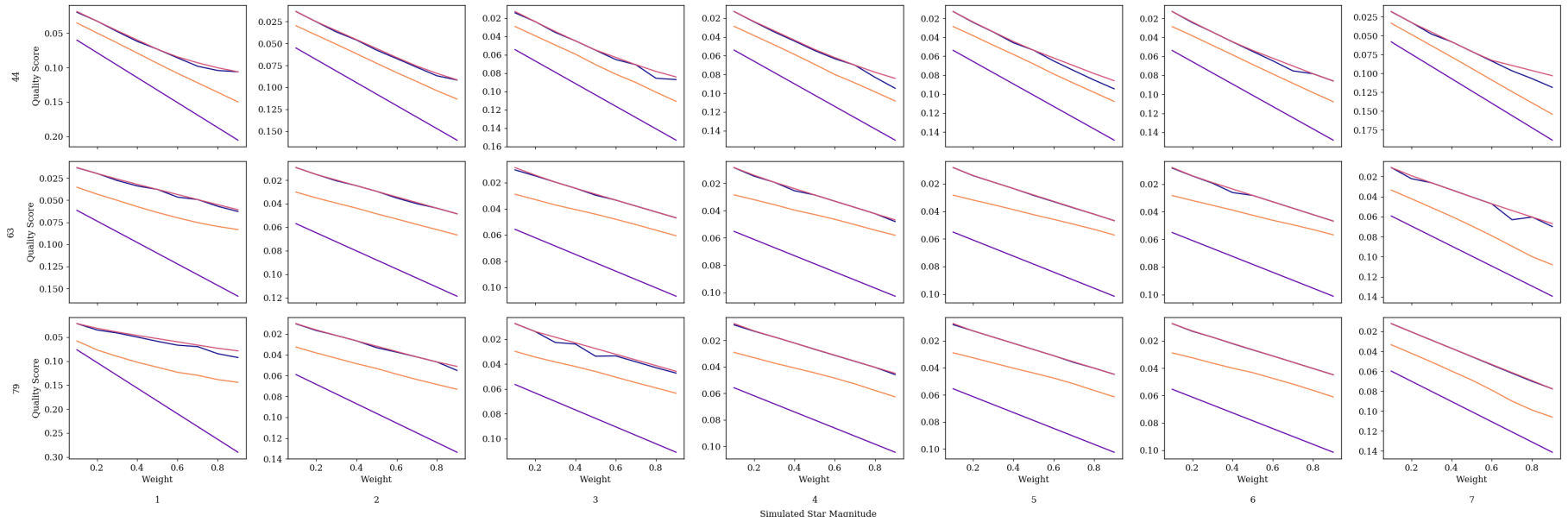


FIGURE 6.16: These graphs depict the quality score of the results produced vs weight when SExtractor was evaluated over Kepler images 44, 63, and 79 when simulated transients were inserted into in one magnitude ranges, begining at -1 and ending at -7. The green and red lines indicate the median and average qualities produced using all combinations of SExtractor settings tested. The blue line represents the quality of the output produced by the settings suggested by the hill climbing algorithm when whose starting values were the minimum of each parameter space. The blue line represents the best possible quality for that region of parameter space as found by the brute force simulations.

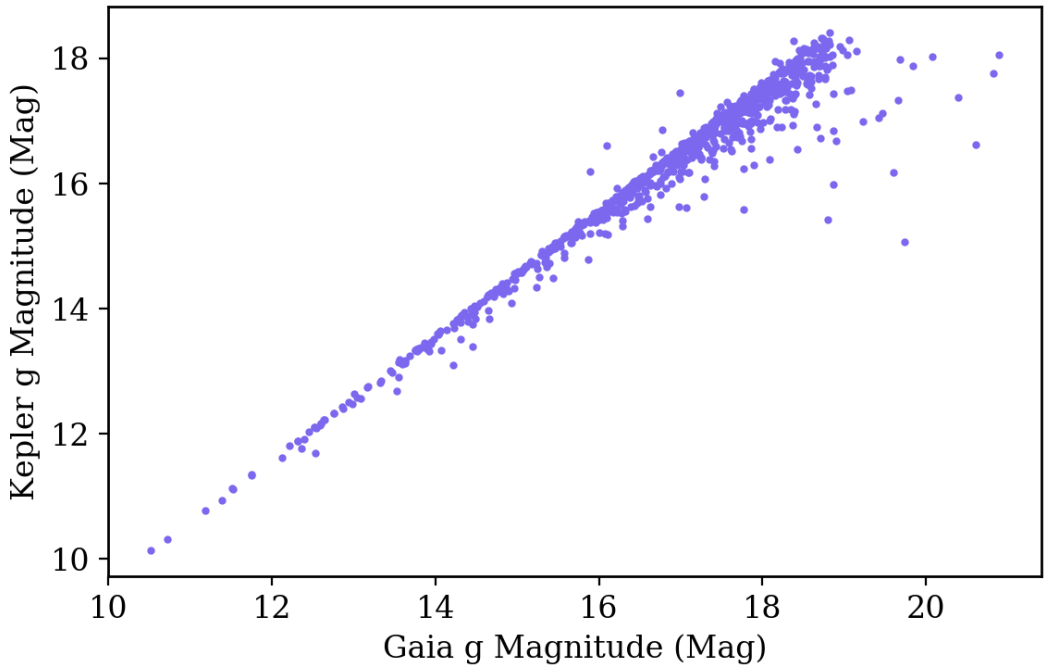


FIGURE 6.17: The Gaia g band magnitude of sources found within the Kepler FFIs vs their respective SExtractor magnitude when measured using the best settings found with the brute force simulations for a weight of 0.5.

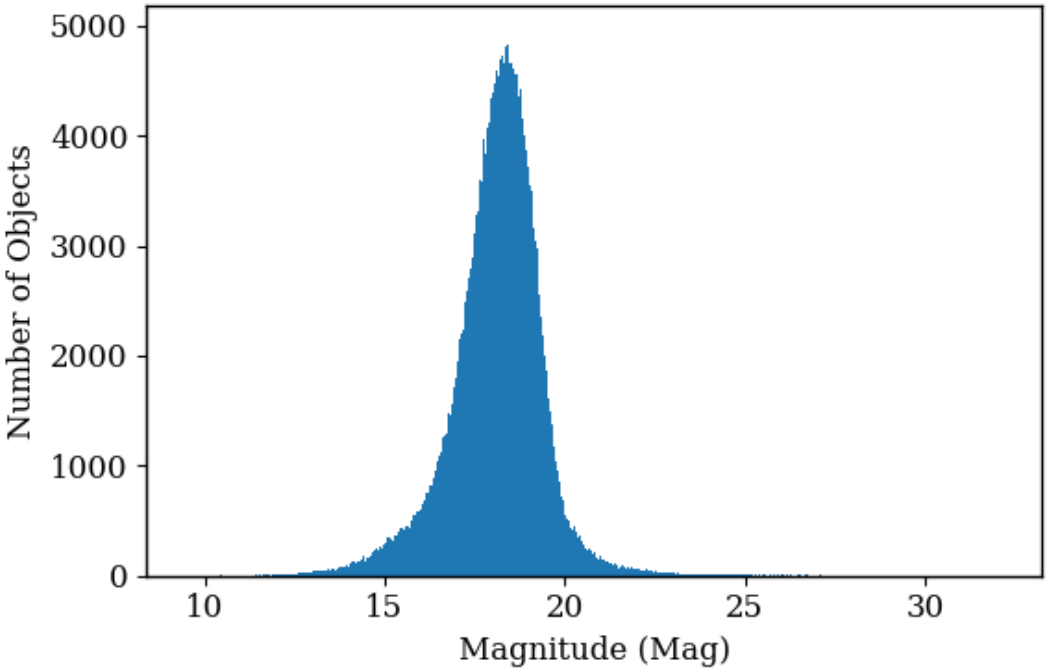


FIGURE 6.18: A histogram of magnitudes for all variable objects found within all channels over every Kepler image, totalling 4,452 images and 2,091,872 variable sources. The mean magnitude over all sources was 17.9 mags.

RA	DEC	IMAGE_NAME	Num_Obs	FLUX_BEST	MAG_BEST	X_IMAGE	Y_IMAGE	magDiff	MAG_ORIG	S/N
280.426667	47.745559	kplr2011208112727_ffi-cal.fits[75]	1	2310.065	16.591	296.464	70.278	-12.712	16.591	4.583
282.8531	47.956328	kplr2009260000800_ffi-cal.fits[76]	1	2889.191	16.348	670.972	861.332	-7.278	16.347	5.121
284.7372	49.055956	kplr2009260000800_ffi-cal.fits[79]	1	2045.174	16.723	481.96	775.935	-7.188	16.722	4.309
280.557104	43.236477	kplr2010019225502_ffi-cal.fits[32]	1	1992.572	16.752	684.766	272.396	-6.747	16.749	4.251
286.02756	37.831268	kplr2009351005245_ffi-cal.fits[75]	1	3926.344	16.015	517.701	230.401	-6.573	16.012	5.967
286.910688	40.388775	kplr2012121122500_ffi-cal.fits[9]	1	13505.36	14.674	198.201	80.03	-6.163	14.67	11.06
284.789151	38.976109	kplr2012310200152_ffi-cal.fits[76]	1	5919.693	15.569	351.907	140.052	-5.113	15.559	7.3
280.26128	43.235157	kplr2013038133130_ffi-cal.fits[32]	1	2337.293	16.578	559.99	122.247	-4.984	16.567	4.584
287.826932	47.209204	kplr2012004204112_ffi-cal.fits[47]	1	2255.791	16.617	1001.959	635.14	-4.982	16.606	4.504
282.508667	42.697489	kplr2011053174401_ffi-cal.fits[31]	1	3542.318	16.127	344.52	963.938	-4.782	16.113	5.637

TABLE 6.3: All transients found within the Kepler FFIs which displayed an increase in magnitude of at least three mags. MAG_BEST refers to the magnitude in the difference image whereas MAG_BEST_ORIG is the magnitude found in the original Kepler image. The largest shown magnitude differences are not expected to be true, the origin of these large rises is the object not being detected in the median image at all. Having said this, as Kepler images are confusion limited at around 20th magnitude, constraints can be put on the minimum magnitude difference of these events.

MAIN_ID	OTYPE	RA	DEC	IMAGE_NAME	Num_Obs	MAG_BEST	X_IMAGE	Y_IMAGE	magDiff	MAG_ORIG	S/N	Parallax
2MASS J19293151+3742406	Mira	292.381	37.711	kplr2011271191331_ffi-cal.fits[3]	40	13.553	838.085	439.081	-1.238	13.134	15.129	-0.263
IRAS 18554+4753	LPV*	284.205	47.953	kplr2009351005245_ffi-cal.fits[69]	52	13.41	27.925	798.66	-1.529	13.105	17.078	-0.213
KIC 12055999	Star	288.148	50.575	kplr2012341215621_ffi-cal.fits[31]	2	15.728	328.145	830.047	-1.302	15.338	5.635	1.423
USNO-B1.0 1360-00297059	CataclyV*	284.662	46.035	kplr2011303191211_ffi-cal.fits[65]	1	15.256	360.362	825.101	-1.444	14.923	7.193	1.486
V* V1119 Cyg	Mira	291.436	51.159	kplr2010111125026_ffi-cal.fits[53]	51	13.13	46.633	728.84	-1.561	12.836	19.521	0.039
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2009115080620_ffi-cal.fits[46]	14	13.847	933.806	121.833	-1.092	13.353	12.733	1.897
V* V344 Lyr	DwarfNova	281.163	43.375	kplr2012151105138_ffi-cal.fits[4]	6	15.415	849.79	661.178	-1.125	14.939	6.24	0.94

TABLE 6.4: Objects found with the Kepler FFIs which displayed an increase in brightness greater than one magnitude and had a counterpart within the Simbad astronomical database. The parallax measurements were found within the Gaia DR2 database. MAG_BEST refers to the magnitude in the difference image whereas MAG_ORIG is the magnitude found in the original Kepler image found through the same method.

6.4.1 Increase in Magnitude Sub-Sample

The first magnitude restriction was a three magnitude increase in brightness which produced a subsample of 85 transients. None of these objects had counterparts in the Simbad or Gaia database which is somewhat expected if the objects are true transients as it is unlikely that they have been observed by other means. All of the subsample were also only observed in one image. This result is also likely as the difference in magnitude was calculated compared to the median image, additional detections would increase the brightness of the object in this image and reduce the effective magnitude increase. The ten brightest objects are shown in Table 6.3 which contains information on their positions, observation dates and all relevant magnitudes. A table with all of the objects is included in Appendix C.1.

A one magnitude restriction was also applied to the sample producing a subsample of 857 transient events. There were 28 events which had a counterpart in Simbad and these corresponded to 11 separate astronomical objects, 7 of which had counterparts in Simbad. These objects and their Simbad counterparts are displayed in Table 6.4 along with information on their magnitude increases and dates of observation. This table only shows one event per object, the full table of events in shown in Appendix C.2

The distribution of these objects with regards to the Kepler field of view is shown in Figure 6.19, where each black dot represents a single Kepler observation. Counter-intuitively, the vast majority of the transients are observed outwards from the Galactic

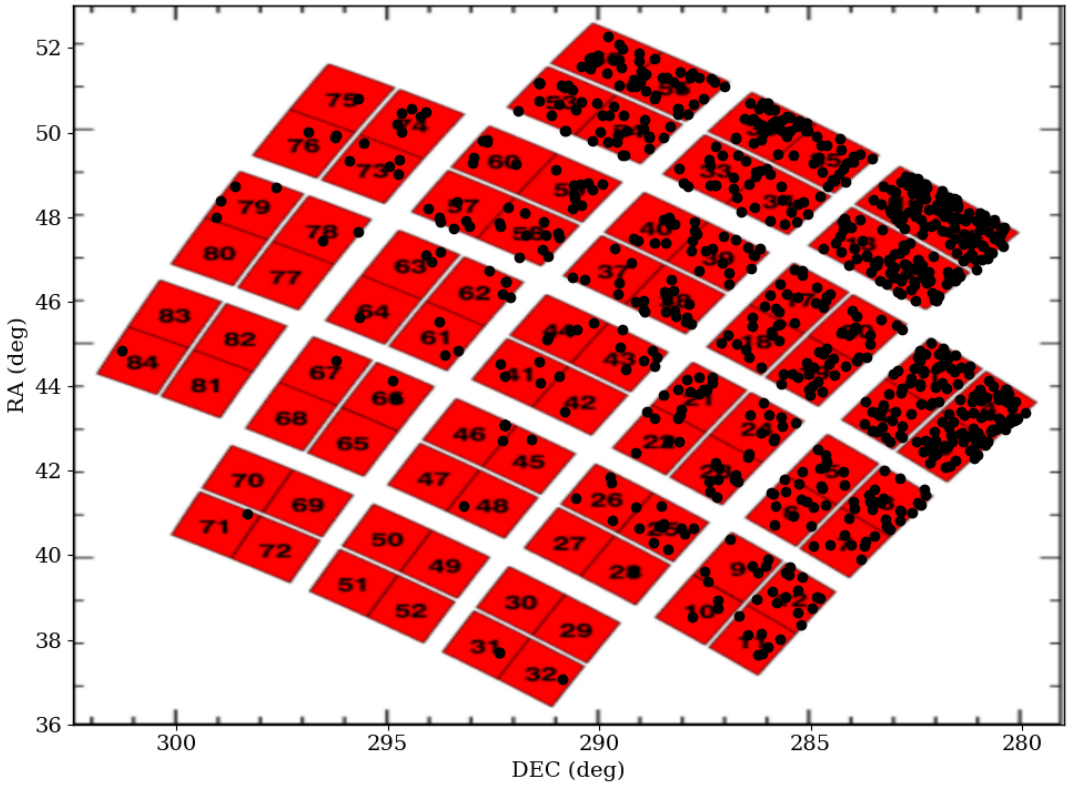


FIGURE 6.19: Schematic of the workflows main components for a single images.

Plane even though more sources reside within it. This can be explained as being a result of the high crowding in the Galactic Plane, the light pollution caused by nearby sources may obstruct the view to many transient sources, effectively dampening their optical variation. There is also a higher likelihood that relatively nearby objects are in the line of sight closer to the Galactic Plane, these objects will typically be very bright and obstruct the view to farther away, dimmer objects. Therefore, although the density of objects that can be seen in the Galactic Plane is much higher, the total volume of observable objects is higher outside of the Plane. The volume is particularly important for transient objects as their high optical amplitude variations enable them to be observed at large distances.

To discover more about the characteristics of our discovered population, a Hertzsprung-Russel diagram was made utilising Gaia data. Firstly, the Gaia DR2 database was

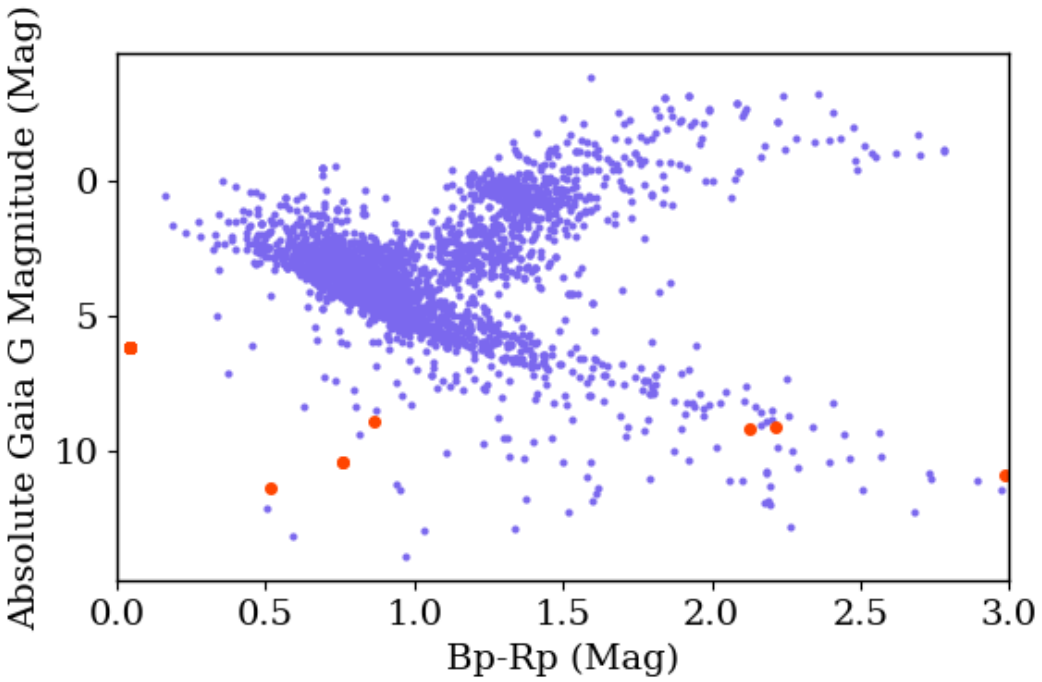


FIGURE 6.20: Hertzprung-Russel diagram of Kepler sources found to have a one magnitude increase in brightness. The G band magnitude, parallax and bp-rp values were found by matching to the Gaia DR2 database. Fifty thousand variable Kepler objects were matched, shown in purple. The one magnitude and above variable sources were also matched, shown in orange. Only sources which had a parallax error < 0.5 times the size of the parallax were used.

queried with 10,000 sources which were detected with the workflow, followed by the subsample which had at least one magnitude increase in brightness. The photometric G band mean magnitude, parallax and Bp-Rp were recorded from each of the 58 events which had a match in the Gaia catalogue. These values are also included in Table 6.4. Sources with a parallax error < 0.5 of the parallax value were not used in the creation of the diagram. Equation 2.5 was then used to determine the absolute G band magnitude. Finally, this was plotted against Bp-Rp, shown in Figure 6.20. This diagram shows that the majority of the sources reside either in the main sequence or the white dwarf branch of the HR-diagram.

6.4.2 Signal to Noise Sub-Sample

A signal to noise restriction of 10 was also placed on the full sample of detected variable events. In total, difference images contained objects which had a signal to noise ratio above 10. The signal to noise sub-sample tended to favour bright objects over the increase in brightness itself. For this reason, 53 events within the sample had a counterpart within the Simbad astronomical database and each of these was also included in the Gaia DR2

MAIN_ID	OTYPE	RA	DEC	IMAGE_NAME	Number_of_Observations	MAG_BEST	magDiff	MAG_BEST_ORIG	S/N	parallax
2MASS J19293151+3742406	Mira	292.381	37.711	kplr2011271191331_ffi-cal.fits[3]	40	13.553	-1.238	13.134	15.129	-0.263
2MASS J19382873+3913505	Mira	294.62	39.231	kplr2011024134926_ffi-cal.fits[80]	46	13.81	-0.675	12.974	10.993	0.018
IRAS 18554+4753	LPV*	284.205	47.953	kplr2009351005245_ffi-cal.fits[69]	52	13.41	-1.529	13.105	17.078	-0.213
IRAS 19545+4512	Mira	299.045	45.35	kplr2010078174524_ffi-cal.fits[54]	52	13.043	-0.689	12.224	15.777	0.004
Mis V0148	Mira	299.234	44.59	kplr2010049182302_ffi-cal.fits[53]	31	13.094	-0.793	12.38	16.208	0.17
V* FX Cyg	Mira	296.04	39.79	kplr2010356020128_ffi-cal.fits[35]	51	13.288	-0.44	12.095	11.811	0.13
V* V1119 Cyg	Mira	291.436	51.159	kplr2010111125026_ffi-cal.fits[53]	51	13.13	-1.561	12.836	19.521	0.039
V* V1155 Cyg	Mira	297.109	42.558	kplr2012242195726_ffi-cal.fits[28]	18	12.838	-0.352	11.445	13.225	0.087
V* V1253 Cyg	LPV*	289.737	44.957	kplr2012179140901_ffi-cal.fits[43]	52	13.702	-0.577	12.739	10.888	0.021
V* V1292 Cyg	Mira	297.002	41.877	kplr2010234192745_ffi-cal.fits[9]	52	13.526	-0.528	12.49	11.395	-0.174
V* V1503 Cyg	Mira	289.972	43.762	kplr2011271191331_ffi-cal.fits[43]	52	13.255	-0.321	11.776	10.482	-0.113
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2009115080620_ffi-cal.fits[46]	14	13.847	-1.092	13.353	12.733	1.897
V* V1670 Cyg	Mira	292.931	41.427	kplr2011177110110_ffi-cal.fits[48]	51	13.447	-0.592	12.506	12.363	0.033
V* V355 Lyr	RRLyr	283.358	43.155	kplr2010140101631_ffi-cal.fits[2]	33	13.894	-0.645	13.022	10.402	0.191
V* V390 Cyg	Mira	292.98	48.459	kplr2012179140901_ffi-cal.fits[57]	28	12.742	-0.204	10.827	10.827	0.167

TABLE 6.5: This table displays the sub-sample of transient events discovered by the workflow in the difference images which all have a signal to noise ratio above ten, only one event per object has been included in this table. MAG_BEST and FLUX_BEST refer to the magnitude and flux measured in the difference image, whereas MAG_ORIG refers to the magnitude measured in the target Kepler image, found through the same method.

database. The full subsample of these objects is shown in Appendix C.3. Table 6.5 is a more compact representation of these data which displays one event per object with a counterpart.

6.4.3 Light Curves from Kepler Full Frame Images

This chapter investigated the quality of different workflow configurations to improve the quality of the data processing and consequently increase the scientific throughput of the dataset. The application of the workflow to the Kepler FFIs also achieves the same goal. As these images were primarily for calibration, they were previously unexplored in terms of transients. Therefore, the study of the transients within them presents an additional opportunity to increase the scientific gain from the Kepler dataset. As a result, light curves were constructed for objects which displayed large increases in brightness, had known counterparts, and were detected in multiple Kepler images. To achieve this, additional images containing any of the one magnitude subsample were searched for, again within a 2.5" radius around their original position. The number of observations each object had in the full set of FFIs is shown in Table 6.4. Light curves were plotted for all objects which were detected in at least two Kepler FFIs, totalling six objects each of which are discussed below.

6.4.3.1 Long Period Variable Stars

2MASS J19293151+3742406 is classified by Simbad as a Mira variable. It was observed forty times within the Kepler FFIs and was detected three times with a brightness one magnitude greater than its brightness in the median image. The lightcurve of this object is shown in Figure 6.21 where the characteristic periodicity and large amplitude for the lightcurve of a Mira variable can clearly be seen. The observations that were one magnitude above the median image are circled. Note that each magnitude increase was calculated treating observations from each CCD as a separate data set, therefore it is

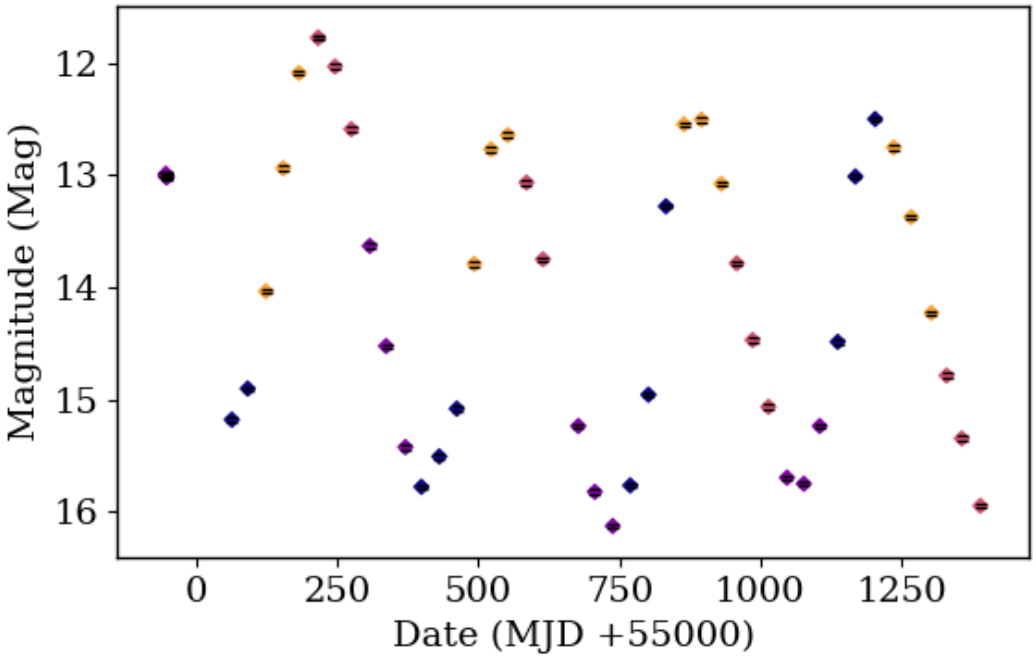


FIGURE 6.21: Light curve of 2MASS J19293151+3742406 constructed from observations within the Kepler FFIs, the separate colours denote the CCD that was used to take the image.

not simply the highest recorded magnitudes that are designated as the highest increases in magnitude.

IRAS 18554+4753 is classified by Simbad as a long period variable star. It was observed in all 52 of the Kepler FFIs with one observations being one magnitude above the median value. The lightcurve is shown in Figure 6.22. The lightcurve demonstrates the long periodicity for which the system gets its name.

V* V1119 is again classified as a Mira variable by Simbad. It was also observed in all 52 FFIs and it was recorded to be one magnitude above the median image in six of them. The lightcurve for this object is shown in Figure 6.23 which again shows the characteristic high amplitude and long period variations characteristic of Mira variables.

6.4.4 Cataclysmic Variables

KIC 12055999 is classified as a star within the Simbad catalogue however it was also identified as a potential CV candidate by ASAS-SN [4], being observed at a magnitude of V=13.7. It was only twice observed within the FFIs, one of these observations was above the one magnitude threshold. The lightcurve is shown in Figure 6.24. Within the

[4] http://www.astronomerstelegram.org/?read=7809

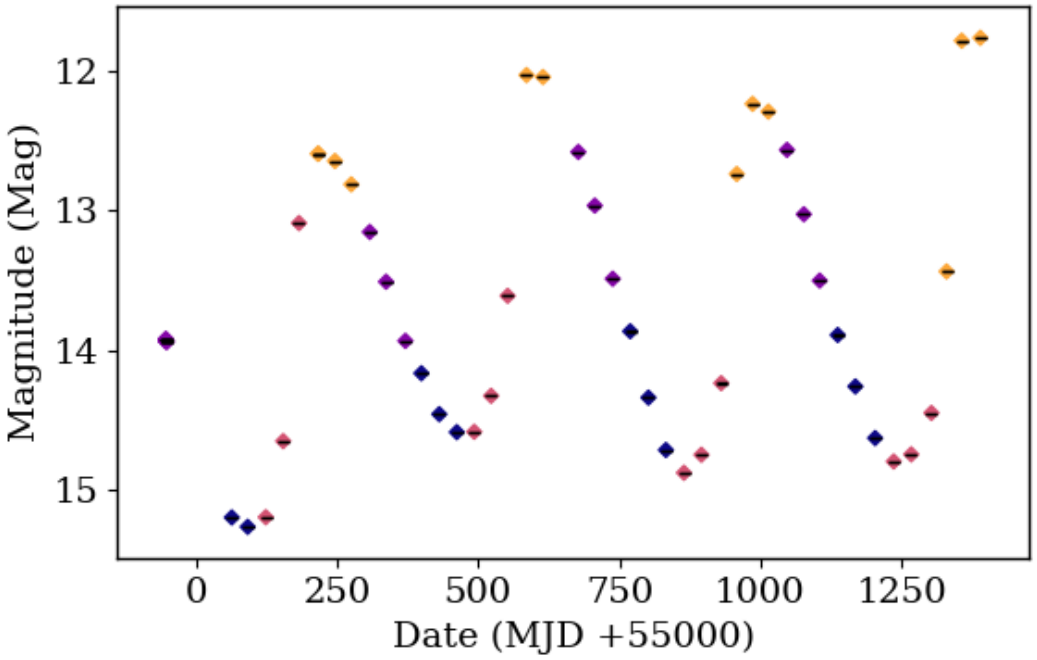


FIGURE 6.22: Light curve of IRAS 18554+4753 constructed from observations within the Kepler FFIs, the separate colours denote the CCD that was used to take the image.

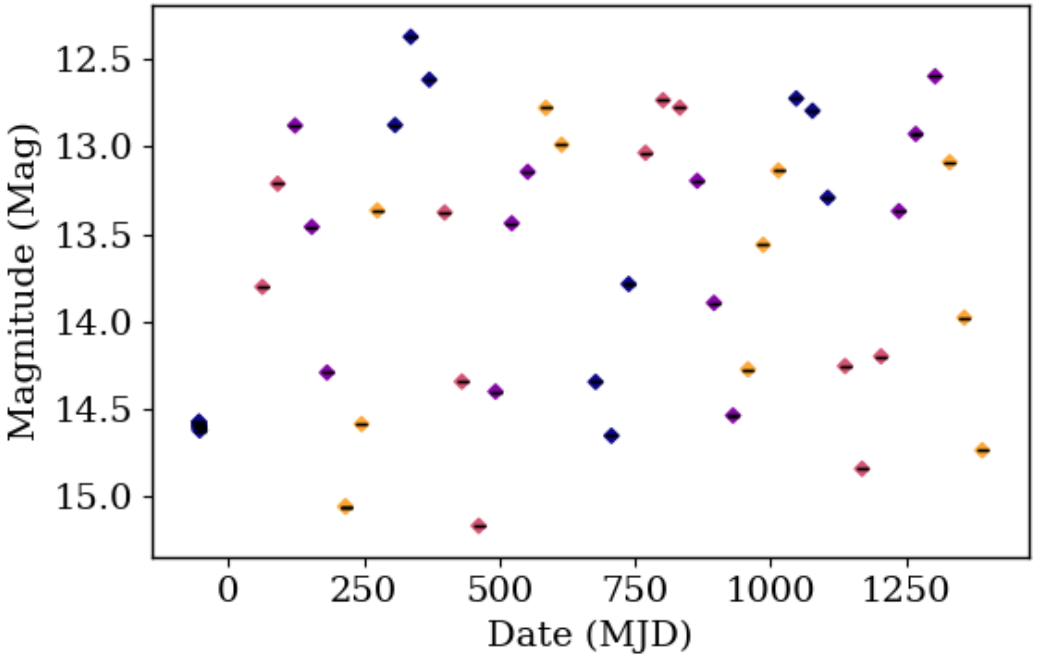


FIGURE 6.23: Light curve of V* V1119 constructed from observations within the Kepler FFIs, the separate colours denote the CCD that was used to take the image.

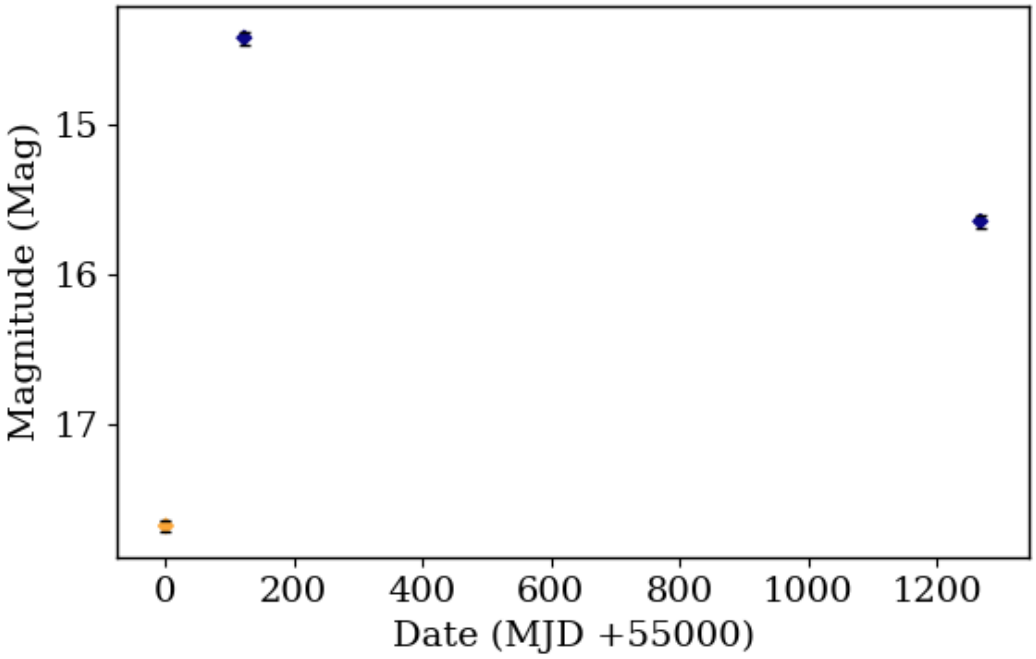


FIGURE 6.24: Light curve of KIC 12055999 constructed from observations within the Kepler FFIs, the separate colours denote the CCD that was used to take the image.

Kepler Input Catalogue, this object is stated as having a g band magnitude of 19.386 [5] which explains why it was not observed in the vast majority of the FFIs as the limiting magnitude is $\sim$ 19 mags. This object was therefore observed to increase in brightness by five magnitudes. The outburst observed by ASAS-SN was during 2015 and the increases in brightness observed within the FFIs were during 2012 and 2013. This means that these observations document previously unknown outbursts in this system.

V* V1504 Cyg is classified by Simbad to be a Dwarf Nova. It was observed in the Kepler images seventeen times, with ten of these observations having a greater than one magnitude increase in brightness when compared to the difference image. The lightcurve for this object is shown in Figure 6.25 where several periods of outburst can be seen.

V* V344 Lyr was also classifed as a Dwarf Nova in the Simbad database. It was observed in 8 Kepler FFIs and one of these was observed as having a one magnitude increase in brightness over the median image. The lightcurve for this object is shown in Figure 6.26 and as with the previous Dwarf Nova lightcurve, multiple rises to outburst can be observed.

[5] `http://vizier.u-strasbg.fr/viz-bin/VizieR-S?KIC%2012055999`

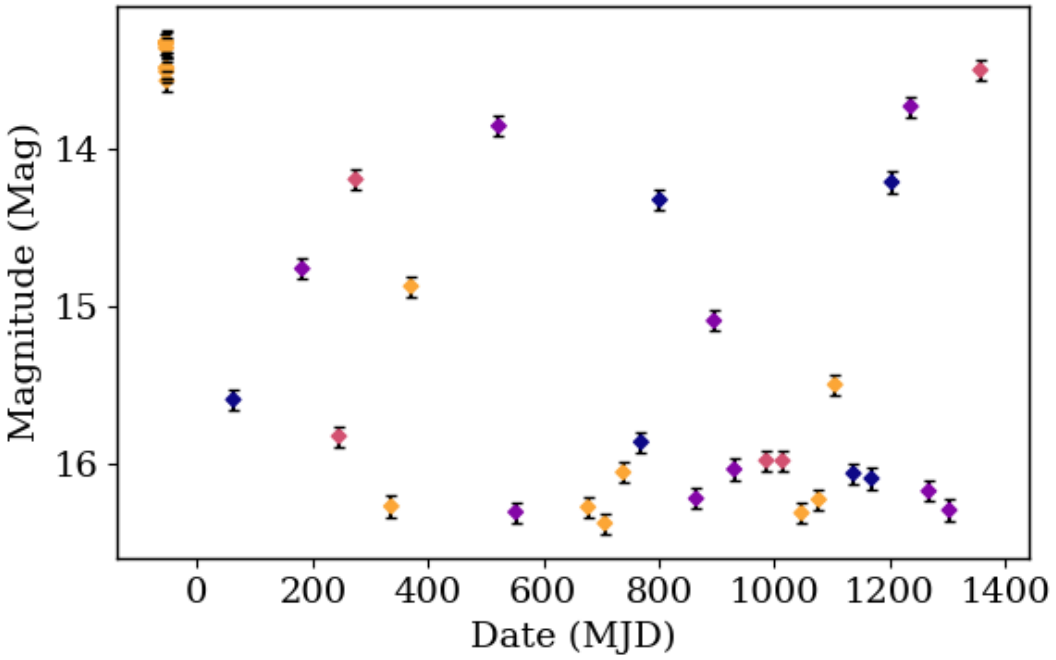


FIGURE 6.25: Light curve of V* V1504 Cyg constructed from observations within the Kepler FFIs, the separate colours denote the CCD that was used to take the image.

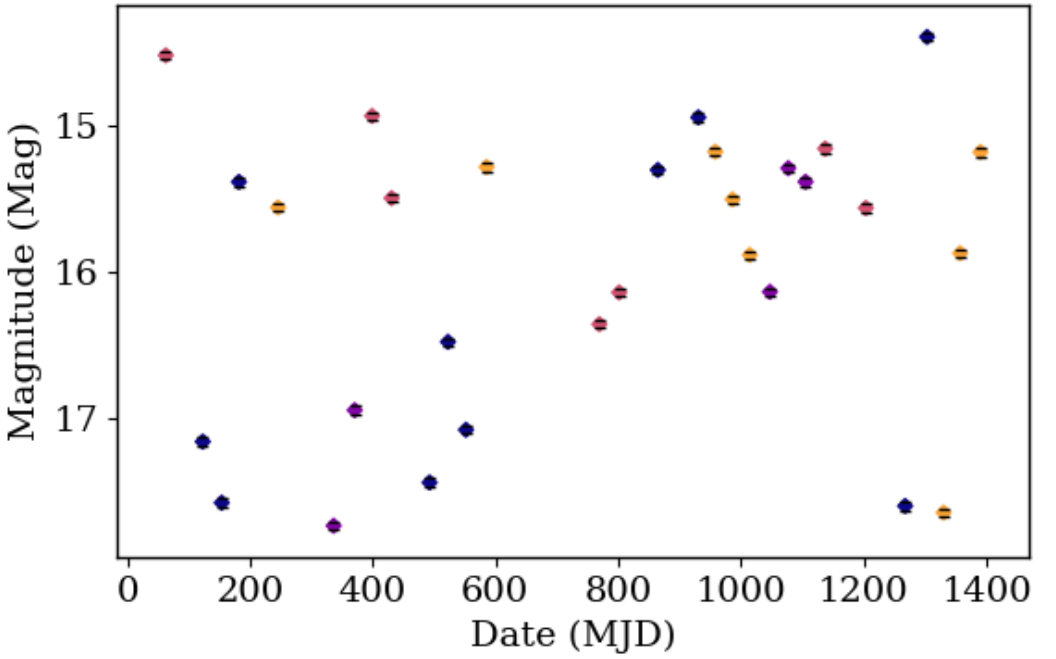


FIGURE 6.26: Light curve of V* V344 Lyr constructed from observations within the Kepler FFIs, the separate colours denote the CCD that was used to take the image.

6.5 Discussion

This chapter applied the approach to a transient detection workflow in order to improve the accuracy and completeness of the transient detection over the Kepler FFIs. The approach reasoned over the the different parameters to be consumed by the SExtractor processor. The first method for approach evaluation was brute force whereby 39,690 distinct versions of the workflow were evaluated and the utility function was calculated for each. Three dimensional surface plots were made which demonstrated the effect of quality that each combination of parameters displayed. The quality of the accuracy of the of the workflow was found to decrease with both detection threshold and minimum area for all filters tested. Each parameter was found to have a similar impact on the accuracy but with a slight bias towards the detection minimum area. This suggests that a one pixel increase in minimum area has more implications for detection than a one sigma increase in detection threshold. As both parameters constrained the conditions required for an object to be detected, it is expected that a higher number of false positives will also be detected. The inverse effect was observed when assessing the completeness as tighter constraints on what is to be detected inevitably resulted in fewer of the simulated transients being recovered. Although both detection threshold and minimum area effected the completeness in this way, which had more of an impact on the completeness depended on the filter - for Gaussian filters it was even, for Mexican hat filters the detection threshold was more important and for top hat filters the detection threshold was more important. The shape of the filters therefore impacts the interaction between the three parameters and the quality. An example of each filter shape is shown in Figure 6.27. The top hat transformation smooths pixels into a profile as shown in Figure 6.27. The flat line of this profile compared to the peaks observed for the other two filters decreases the maximum value that a pixel may contain. This effect explains why the completeness is at its worst for high detection thresholds as the smoothing has reduced the maximum amplitude of the population of the simulated transients. The shape of the Mexican hat filter causes the opposite effect, maintaining the high peaks so that the detection minimum area is most impactful on the completeness. Overall, of shapes of the Mexican hat filter by far performed the worst out of all other other filter shapes with regards to completeness. The results were particularly bad for small kernel sizes of Mexican hat filters and large values for detection minimum area. It can therefore be concluded that the smoothing of the kernel reduced the size of the artificial transients such that they did not meet the requirements of the detection threshold within the minimum area.

This enabled the determination of the best workflow configuration but had a processing time of $\sim$50 hours on a standard computer. The high computational cost of this method then motivated the investigation of the hill climbing algorithm as a means to quickly arrive at a workflow configuration which performed reasonably well. This method took an average of 100 seconds to run and had an average quality within 0.2% of the best possible settings as found by the brute force evaluation. To compare these results to

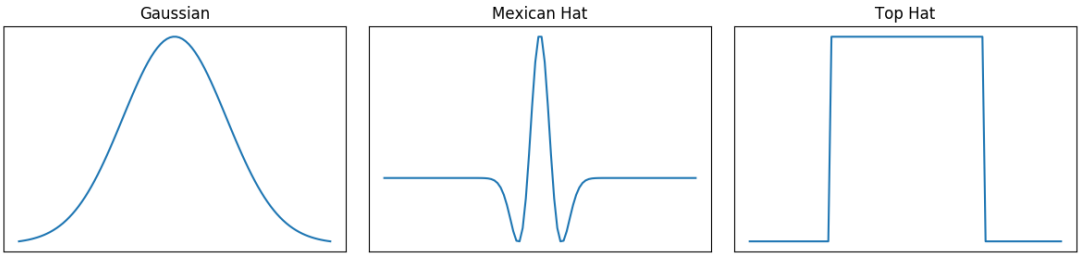


FIGURE 6.27: The shapes of the three different convolutional kernels used to filter the images using SExtractor.

the traditional method of finding workflow configurations, an additional experiment was performed where the workflow settings were found through manual trial and improvement on the images. This consisted of inserting fake transients into the images as before and evaluating the images using the workflow. The parameter space was cycled through and changes were made to the parameters according to which set appeared to produce the best results when looking at the images. Example indications of the performance of the workflow settings used were the number of inserted transients returned and the number of false transients returned. Whether an object was deemed to be false was by visual inspection of the PSF of the object, those with similar PSFs to other objects in the image were considered real whereas uniquely shaped objects were regarded as false. This process took ∼4 hours and the best returned settings were returned to be: DETECT_THRESH = 10, DETECT_MINAREA = 6, and FILTER_NAME = tophat_2.0_3x3.conv. A comparison of the times taken and quality of results produced by the three separate methods is shown in Figure 6.28. From these results we can conclude that not only is the hill climbing method much faster than traditional means, it also provides better results. On average, the results outperformed the traditional means by ∼1%. Although this seems like a minor improvement, the scale of the dataset translates this into sizeable scientific gains. For example, ∼250,000 transient objects were detected within the FFIs, a 1% improvement in completeness here equated to an additional 2,500 potential transients.

With regards to the actual objects found through the workflow, instant scientific gain can be derived from the objects with known counterparts. Most notably, three of the CVs detected within the FFIs have observations which predate their discovery. In addition, outburst phases were observed within two of these objects which were previously unknown. The entirety of the three magnitude subsample of objects unfortunately had no known counterpart, nor more than one observation within the Kepler FFIs. Although little science can be achieved with a single data point, the information that these objects exhibited dramatic increases in brightness may prove useful if they are detected in the future.

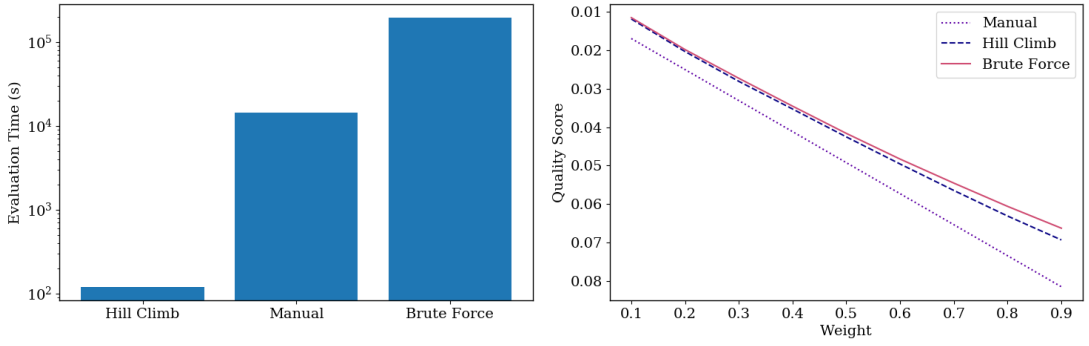


FIGURE 6.28: The left hand figure depicts the time taken for the best SExtractor settings to be found by a single run of the hill climb algorithm, the manual determination by an astronomer and the full brute force evaluation. The right hand figure shows the quality of the workflow produced by each method for a variety of weights, averaged over all magnitude ranges, Kepler channels and spatial distributions.

6.6 Conclusions

In conclusion, the best workflow configuration for transient detection was found via three different methods - manual investigation, brute force evaluation and by using the hill climbing algorithm. The traditional manual method for finding workflow versions was out performed by both the brute force simulations and hill climbing algorithm. Although the best workflow configuration was found via the brute force evaluation, it was very computationally expensive, taking ∼50 hours of computation. On the otherhand, the hill climbing algorithm took ∼100 seconds to evaluate and produced a workflow version with a quality within 0.2% of that of the brute force algorithm. The results of this investigation as well as that in Chapters 4 and 5 as discussed in the following chapter.

Chapter 7

Discussion and Conclusions

The research statement was to improve the quality of the results produced by workflows which analyse astronomical data through the use of the approach outlined in Chapter 3. Three separate use cases were investigated in this regard in Chapters 4, 5 and 6. In each case, the relevant parameter spaces, quality metrics and baseline qualities were identified. Workflow configurations were constructed utilising each discrete version of parameter space and the quality of each was measured. Finally, these measures of quality were combined with the baseline quality to find the workflow configuration which provided the highest quality. This chapter describes the applicability of the approach to each of the use cases as well as the benefits, drawbacks and potential improvements. A summary of the approach instantiation to each chapter is shown in Table 7.1, which includes information on their quality metrics, parameter spaces and workflow functions.

Brute force evaluation was utilised in each of these use cases and was found to be an effective method for assessing the quality of the different workflow configurations. The biggest limitation of this method was the computational cost of evaluation. For example, Chapter 4 contained 8000 discrete configurations and the brute force evaluation required ∼16,000 hours of computation. Similarly, Chapter 6 contained 36,960 workflow versions and the brute force evaluation required ∼50 hours of computation. From these figures, it is clear that not only the number of discrete workflows affect the computation time but

Chapter	**4**	**5**	**6**
Workflow function	Period recovery	Differential photometry	Transient detection
Quality metric	Completeness	Timeliness	Accuracy & completeness
Parameter space	Observing strategies	Provenance recording	SExtractor settings

Table 7.1: A table showing a summary of the approach application to the use cases within this thesis.

also the time required for a single workflow execution. The time required per execution can be influenced by a number of factors. An example of one such factor is shown in Chapter 4 where a single workflow version required 10,000 instances of the Lomb-Scargle algorithm in order to find the completeness. On the other hand, the number of potential workflows is decided exclusively by both the number and size of the parameter spaces. When the evaluation time per workflow and/or the number of workflow versions are large, the usefulness of the brute force method becomes overshadowed by the time required for evaluation. It is in these regimes where the hill climbing algorithm is required. Large parameter spaces are commonplace in astronomical workflows as fundamental processes such as aperture photometry, difference imaging and source detection all often contain many free parameters. The brute force evaluation of the example of this in Chapter 6 was only possible by restricting the parameter space to the three most impactful categories but in total, 100+ were available for customisation.

7.1 Prospecting for Periods with the Large Synoptic Survey Telescope

Chapter 4 utilised the approach to determine the effect of different LSST observing strategies on the period recover of LMXBs. The parameter space consisted of the candidate observing strategies, the quality metric used was completeness and the utility function was the number of simulated LMXBs which had their period correctly recovered divided by the total number of simulated systems.

By evaluating the approach, it was found that the current baseline observing strategy will likely recover periods for 18% of the Galactic population, this improved to 32% with observing strategies that do not reduce observations within the Galactic Plane. Although these strategies are promising for Galactic Plane science, they do result in a 5% reduction in observations for all fields outside of the Galactic Plane. The case of improved observations for LMXBs alone is unlikely to impact the decision on which strategy is chosen. To increase the impact of this work, the parameter space could be expanded to include other Galactic periodic objects such as long period variables or cataclysmic variables. Having said this, such an option is also being researched by the LSST community, as discussed in Chapter 2.

7.2 Using the Provenance from Astronomical Workflows

In Chapter 5, the approach was applied to an aperture photometry workflow to determine whether provenance could be used to increase the processing efficiency. The parameter space was the choice of whether or not to record the provenance. The quality metric

was timeliness and the utility function was constructed as a combination of the time cost from recording the provenance and the time saved when evaluating the identified use cases.

It was found that when evaluating the two outlined use cases, the processing efficiency increased by 99% and 96%. However, recording the provenance decreased the processing efficiency of the workflow itself by 45%. This resulted in a net decrease in processing efficiency that was estimated to be between 13% and 44%. Thereby demonstrating that although there is potential for provenance to be used in this way, the cost outweighed the benefits. As discussed in Chapter 5, a few potential methods to reduce the cost of provenance would include increasing the number of use cases identified, transferring the recording of provenance from provn files to bindings, and being more selective with the points at which provenance is captured. With regards to the approach, several alterations could also be investigated. Firstly, the parameter space was binary as it only included the option to record or not record the entirety of the provenance. The motivation for this decision was to not specialise the provenance so that it could only solve the outlined use cases and it could be applicable to any possible scenario. Knowing which key components of the provenance to record could drastically reduce the cost of recording the provenance while still enabling efficient evaluation of the use cases. Finding which components are needed for the vast majority of use cases could be found via an extensive investigation into potential use cases for provenance within the astronomy workflow. A second method would be to add different granularities of provenance recording to the parameter space. With only two identified use cases this may tend to specialise the results. This method therefore would also rely on the identification of additional use cases.

7.3 Finding Transients with Kepler

Chapter 6 applied the approach to a transient detection workflow which was evaluated over the Kepler Full Frames Images. The parameter space consisted of SExtractor settings which impacted the source detection of the workflow. The relevant quality metrics were accuracy and completeness and the utility function was constructed as a weighted combination of the two. The quality metrics were scaled to be maximum at zero and minimum at one, the baseline quality for the utility function was therefore defined as that produced by the perfect workflow. Two separate methods were used to evaluate the approach - brute force and hill climbing. To compare the results of these methods to real world results, a workflow configuration was also chosen by visual inspection of impact of different settings on the quality of transient detection. This qualitative method was effectively similar to the hill climbing algorithm, where incremental changes were made to the settings and the workflows were evaluated over the Kepler images which contained simulated objects.

The first application of the approach involved the brute force evaluation of all discrete workflows configurations constructed within the parameter space over all simulated magnitude ranges in each chosen Kepler channel. These 39,690 workflow evaluations took ∼50 hours to complete on a standard computer. These simulations enabled the determination of the best workflow configuration for every combination of magnitude range, Kepler channel, spatial distribution and quality weighting. The computationally expensive nature of this evaluation was both the biggest draw back and limiting factor. The number of free parameters in this example is fairly low, so the computing time was able to be kept at a reasonable level whilst maintaining a reasonable granularity within the parameter spaces. However, as the number of parameter spaces increases, the computation time increases rapidly. One potential solution is to analyse a very coarsely sampled parameter space and use a machine learning algorithm to infer the relationships between each parameter and extrapolate to a more finely sampled grid. This method comes with its own caveats as a reasonable region of the parameter space still needs to be evaluated and it is assumed that the relations within this coarse grid are representative of the entire parameter space. For these reasons, the hill climbing method was chosen instead. This algorithm on average found settings that had a quality within 0.2% of the best found via brute force and took ∼100 seconds to complete per run. The results from the traditional method for manually selecting the best workflow versions took around four hours to generate and found a workflow version with a data quality within 1% of the best settings from the brute force simulations. Therefore, not only did the hill climbing algorithm find the results much faster, it also produced a higher quality workflow.

The brute force simulations found that the quality of the average workflow configuration was only ∼5% different from the best quality workflow found. This may seem like a small improvement, however the workflows were finding approximately two million objects within the full set of Kepler images, meaning that a 5% increase in completeness will correspond to roughly an additional one hundred thousand potential transient objects. The same can be said with the accuracy, where a 5% increase in accuracy will likely remove thousands of false positives from the sample. These results exemplify the importance of small improvements in data quality when analysing large datasets.

Chapter 8

Future Work

8.1 The LSST Observing Strategy

Chapter 4 was able to successfully characterise the LMXB period determination of LSST when using different candidate observing strategies. However, the discussion surrounding the LSST observing strategy is still ongoing and the impact of LMXB science alone is unlikely to be influential. Therefore, the parameter space of this work will be expanded to include other Galactic, periodic objects such as binary stars or cataclysmic variables (CVS). These types of object have similar optical variability to that of LMXBs and their addition would not require the alteration of the majority of the current workflow. The changes that would need to be made are solely within the light curve generation procedure. The magnitude and period parameter space would be expanded to accommodate the populations of these objects. Alterations would also need to be made to include the additional peculiarities in their optical signals, the example for LMXBs was the additions from stochastic flaring and the power law disc component.

With regards to updating the observing strategy parameter space, the simulations will be repeated for the same observing strategies made with the most up to date version of the operations simulator - OpSim4. There is currently no plan to include additional types of observing strategy as niche versions are unlikely to receive the same analysis from the LSST community and therefore unlikely to be selected. However, exceptions will be made for promising new strategies which gain traction with the community.

One potential use for for the orbital period of LMXBs is in determination of the binary mass function, however in order to calculate the mass of the binary components, additional information on the velocity of the binary orbit is required. This is usually obtained by looking at the Doppler shift in the distinct emission lines observed in the spectra of LMXBs, however it is unlikely that it will be possible to gather sufficient spectral information on the full LMXB population discovered by LSST.

Antokhina and Cherepashchuk (1997) calculate the theoretical light curves of X-ray binaries in a quiescent state for a wide range of parameter values. They discovered dependences on the parameters of the light curves, such as amplitude of the minima and maxima, can be related to the parameters of the system such as mass or orbital inclination. Therefore their techniques will be applied to simulated LMXB data in order to recover parameters of the binary system. Light curves will be simulated which correspond to LMXBs with well defined parameters and optical behaviour, these will subsequently be sampled by potential LSST observing strategies. Table 3 in Antokhina and Cherepashchuk (1997), which lists the minima and maxima that correspond to simulated LMXB light curves with different parameters such as ellipticity or temperature of the donor star, will be used to relate the observed light curve to these system parameters. These results will be compared to the parameters quoted in the literature in order to determine the usefulness of this technique in this context and what constraints it could potentially put on the parameters of the observed LMXBs. If successful, this technique could expand the impact of LSST Galactic Plane coverage for LMXBs to also include direct constraints and the binary component masses.

8.2 Improving the Timeliness of Astronomical Workflows Through Provenance

One of the main drawbacks of the work in Chapter 5 was that the cost of collecting and storing the provenance was too high for the provenance analysis to offset. A number of ways to reduce this cost will be tested in the future. The first is to record the provenance within a graph database - not as files. At present, each provenance record is stored in its own provn file. By switching to a graph database the time for opening and closing each of these files will be eliminated. This method also promises significant reduction in the amount of required storage. Secondly, not generating the full provn files and instead only recording the bindings and querying over them directly. This method would eliminate the time required for transforming the json files to the provn format and the merging of the files. Finally, reducing the granularity of the provenance by selecting capture points throughout it. This would require an investigation into the location of these capture points which would involve the identification of a multitude of potential use cases and which regions of the provenance would be required to solve them.

The impact of reducing the computational overhead of recording provenance scales with the size of the dataset. LSST will generate one of the largest astronomical datasets ever and their data management pipeline already utilises pipeline provenance in order to reduce the computational resources required to capture and store it (Juric et al., 2015). It will therefore be investigated whether the aforementioned methods could also be applied to the LSST data management pipeline in order to further reduce the computational overhead of recording provenance for the LSST pipelines.

8.3 Finding Transients in the Archival Datasets

An in depth investigation will be performed into the objects that were found within the Kepler FFIs within Chapter 6. So far, there has been the identification of highly variable and transient objects within the FFIs and a preliminary investigation into their significance. In future, highly variable objects with multiple Kepler observations will be searched for periodicities using the Lomb-Scargle algorithm (Lomb 1976, Scargle 1982). Additional, object dependant investigation will also be undertaken, for example the examination of the outburst phases in the cataclysmic variables.

All objects which displayed a greater than three magnitude increase in brightness were found in only one FFI and were not present in the Gaia DR2, or Simbad databases. The first extension to this work is the extension to include the Vizier database. This was currently not implemented as Vizier is an amalgamation of many separate catalogues and there are inconsistencies in the conventions with which they store data. These inconsistencies require additional cleaning which will be handled in the future. The next place to search for additional observations of these events is within other astronomical survey telescope datasets such as that produced by the All Sky Automated Survey for Supernovae (ASASSN) (Pojmanski, 1997), the Transiting Exoplanet Survey Satellite (TESS) (Ricker et al., 2014) and the Zwicky Transient Facility (Bellm, 2014) which all regularly observe the same region as the Kepler field.

The evaluation for the completeness of the workflow will be used to investigate the populations of object classes within the Kepler FFIs. The completeness metric provides a defined measure for the fraction of objects that are present within the FFIs, which are not detected. This will be used in combination with the total number of objects detected of each type in order to deduce additional information on the object's population within the Kepler FFIs. Furthermore, this population estimate can then be extrapolated for the entirety of the Milky Way using an expected spatial distribution for the object type - similar to the analysis in Chapter 4 for the extrapolation of LMXB period recovery with LSST within the Milky Way.

The work described in Chapter 6 on improving the transient detection workflow with the approach will be expanded in two separate ways: by specialising the purpose of the workflow to individual types of objects; by applying the transient detection workflow to additional datasets.

Chapter 6 generated simulated objects according to the average PSF of an object within each image. When these images were processed by the workflow, the quality of the results it produced then represented the expected results for the average object in the workflow. The simulated images will be made using preselected types of objects only. This is likely to have the largest impact on objects such as galaxies which have a light profile that is likely to differ from the PSF of the image. This would then mean that the filter that best

described them could be of a different shape. The filter shape was also demonstrated to impact the effect on quality that both the detection threshold and minimum area had, so the different filter choice will also likely change the choice of the other two parameters.

The Kepler FFIs were ideal candidates for the workflow as they covered a wide area of the sky, were taken in regular intervals and most importantly, they were previously unexplored with regards to transient science. These qualities are shared by many other survey telescopes, the Transiting Exoplanet Survey Satellite (TESS) (Ricker et al., 2014) being a good example. The TESS dataset covers an area 400 times the size Kepler's and is an ideal candidate for the workflow. The larger size of the TESS dataset when compared to the Kepler FFIs motivated the choice to develop the approach using Kepler as a relatively small subsample of Kepler data could be used to represent the full set of FFIs. Now that the approach is more developed, it can be applied to the workflow again to find the ideal parameters for the TESS dataset.

An additional improvement will also be made to the brute force method whereby a very coarse region of the parameter space is simulated and used to train a machine learning algorithm. The quality of the workflows which use the intermediary regions of the parameter space can then be determined. Preliminary investigations have been made using Scikit-learn's implementation of the random forest algorithm (Pedregosa et al., 2011). The algorithm was able to return the quality of the test workflow configurations with an accuracy of 99.8%. The high accuracy is likely to be due to the smooth and regular patterns that can be observed in the quality vs parameters plots shown in Appendix B. This technique has the potential determine workflow quality for a much finer mesh of parameter space that through brute force evaluation alone.

Appendix A

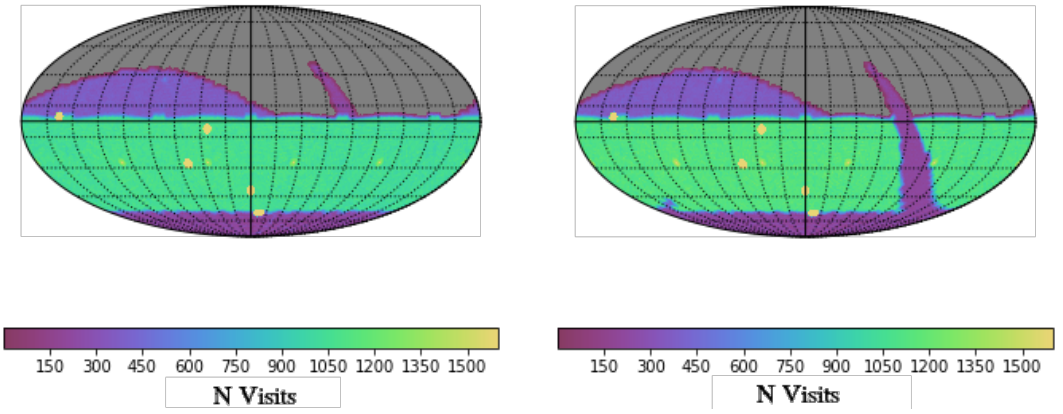
Appendix: Prospecting for Periods

A.1 Observing Strategies

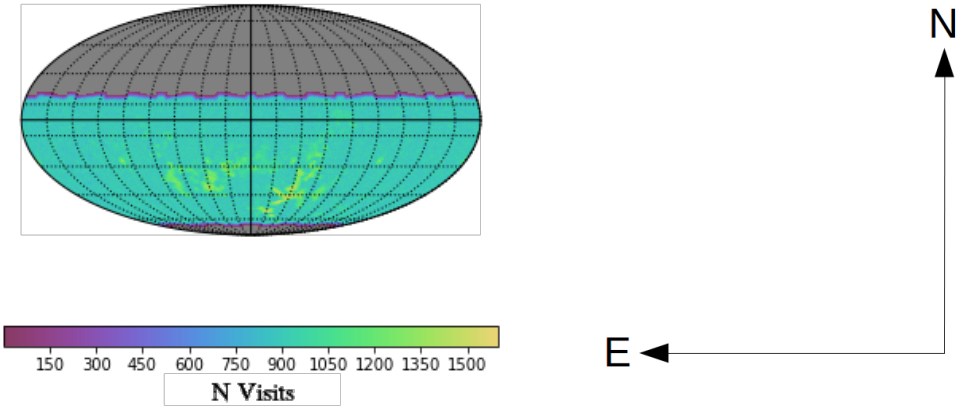
Figure A.1 depicts the total number of observations per field made during candidate observing strategies `Minion_1016`, `Minion_1020` and `astro_lsst_01_1004` in all bands over the full 10 year survey as simulated with the OpSim (Delgado et al., 2014).

A.2 Orbital Period Determination in Simulated LSST Fields

Figure A.2 depicts the P_{orb} recovery over the P_{orb}-mag parameter space for the four simulated LSST fields which did not have their corresponding diagram included in the main text. The left panel of Figure A.2 (b), shows the P_{orb} determination of each strategy with LSST field 1929, which is in the main WFD survey region. This figure demonstrates that when observed with this cadence, the recovery of P_{orb} is very good under all strategies, as there is not the reduced Galactic cadence present. In the right panel of Figure A.2 (b), the P_{orb} determination for LSST field 3311 is displayed. This is located such that it will be observed by the South Celestial Pole cadence and the observations are reduced in all strategies except `Minion_1020` due to airmass restrictions. For `Minion_1020`, P_{orb} recovery is reduced only by the relatively high reddening in this field.



((A))

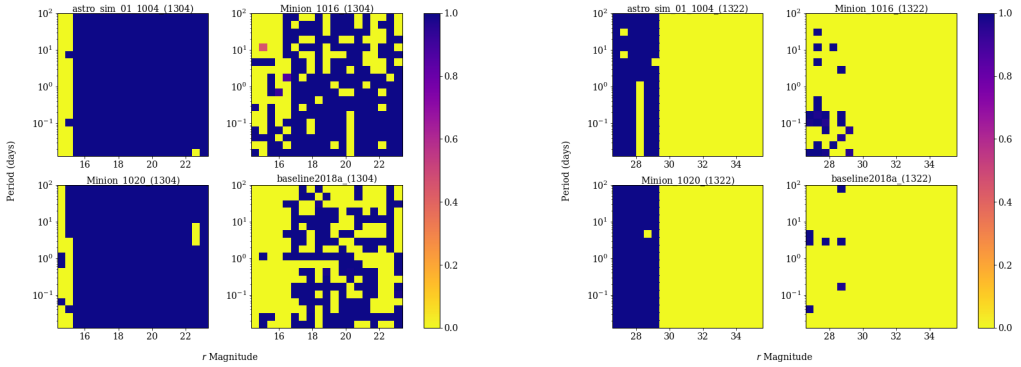


((B))

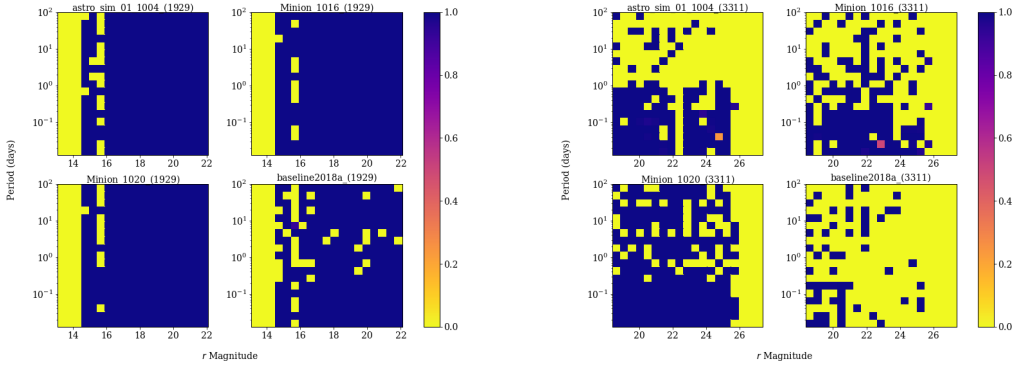
FIGURE A.1: Total number of observations in all bands made using the `astro_lsst_01_1004` (left, a), `Minion_1016` (right, a), and `Minion_1020` (left, b) observing strategies, shown in celestial coordinates where zero RA corresponds to the black line in the plane of the y-axis and North=up, East=left. All graphs were made using the LSST Metrics Analysis Framework.

A.3 Reddening-Orbital Period and Reddening-Mag Relationships

Figures A.4 and A.5 depict the relationship between reddening-magnitude-P_{orb} and reddening-period-P_{orb} recovery, respectively. In both graphs we observe a negative correlation between reddening and P_{orb} recovery, which is as to be expected as in most cases, the higher the reddening, the fewer usable observations that are available. One notable exception to this rule is shown in Figure A.4, where at the very low reddenings and low magnitudes, the objects are too bright, reducing the number of observations in LSST's visible range. For these particular magnitudes, P_{orb} recovery is shown to increase with reddening. The relative lack of bi-modality in P_{orb} recovery completeness in Figures A.4 and A.5 when compared to that in graphs that represent signal fields (Figure A.2)



((A))



((B))

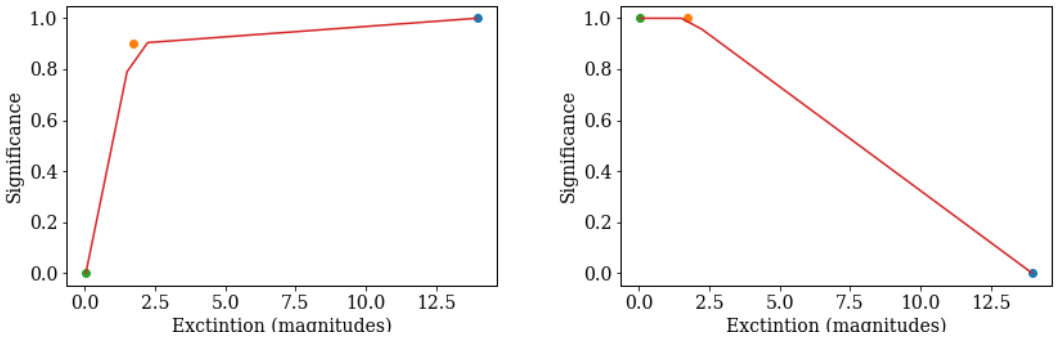
FIGURE A.2: Colour maps displaying the relationship between magnitude, reddening and period determination of LMXBs possible with observing strategies `astro_lsst_01_1004`, `Minion_1020`, `baseline2018a` and `Minion_1016`. X axis denotes the r band magnitude after reddening had been applied individually for each field and before adding contributions from ellipsoidal modulation, flaring, noise. The X axis denotes the period in days. The colour denotes the completeness of the period recovery. Simulations using observations of LSST field 1304 (left,a), 1322 (right,a), 1929 (left,b), and 3311 (right,b) are displayed in this figure.

is a result of how much P_{orb} recovery is expected to change based on which field and consequently, which Galactic extinction was used.

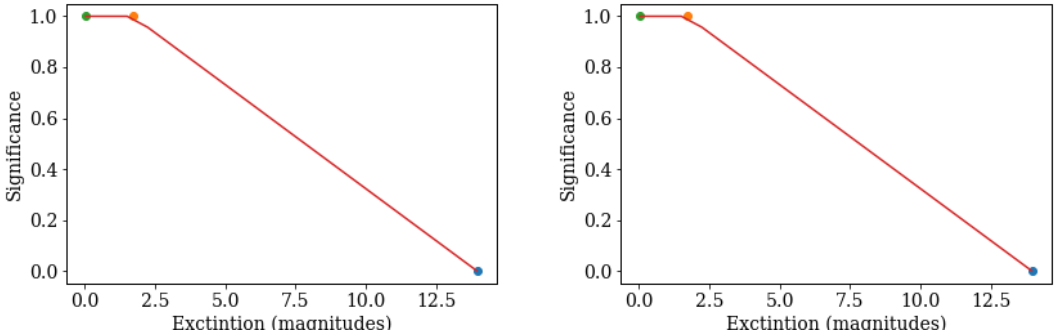
A.4 Galactic Period Recovery Integration

Equation A.1 details the procedure for summing up the total P_{orb} recovery over the Milky Way.

P_{tot} is total fraction of the LMXB population that will likely get accurately recovered periods. r represents distance from the Galactic Centre. $P_p(P_{orb})$ is the probability of an LMXB having P_{orb}, p; $P_m(M_{obs}, r, \theta, \phi)$ is the probability of an LMXB having an observed, post reddening magnitude , m (when calculating $P_m(M_{obs}, r, \theta, \phi)$, r, θ and



((A))



((B))

FIGURE A.3: Figures displaying the P_{orb} recovery completeness interpolation for the observing strategy `astro_sim_01_1004` with pre-reddened magnitudes 14.4 (left,a), 17.2 (right,a), 20.1 (left,b), and 23.4 (right,b). Each point represents the P_{orb} recovery for a Galactic LSST field, with the completeness of recovery on the Y axis and that fields reddening on the X axis. The red line represents the corresponding extinction and P_{orb} recovery completeness for the twenty chosen, linearly spaced extinction values that are being interpolated.

ϕ were transposed to l, b and the radial distance from the Sun using a distance of 7.9 kpc from the Sun to the Galactic Centre). $S_{P_{orb},m}(P_{orb}, M_{obs}, r, \theta, \phi)$ is the P_{orb} recovery completeness with P_{orb}, p, and magnitude, m. M_{obs} is the magnitude of the LMXB before reddening corrections. Finally, θ and ϕ represent angles, in the Galactic Plane and perpendicular to the Galactic Plane, respectively.

Equation A.1 is held if $12 \leq M_{obs} \leq 22$ and $0 \leq A_r \leq 13.9$. Otherwise, $P_{tot} = 0$.

(A.1)

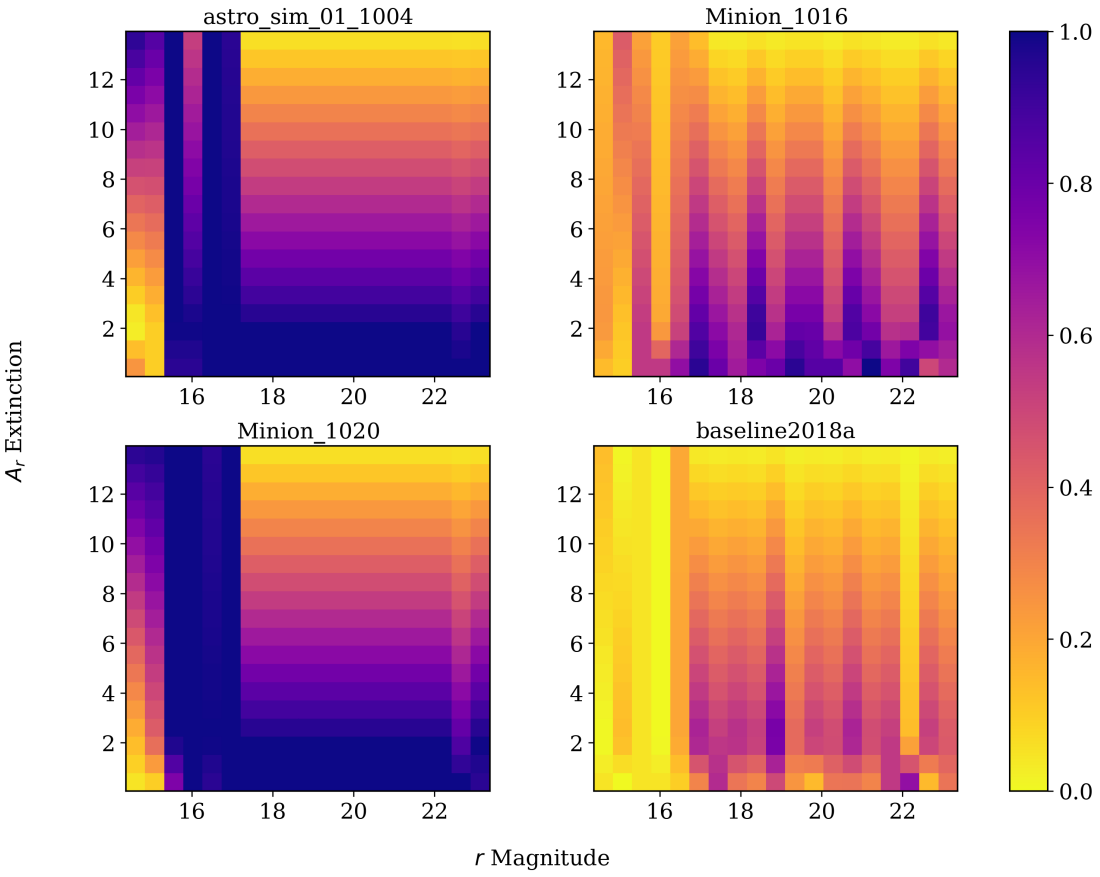


FIGURE A.4: Colour maps displaying the relationship between magnitude, reddening and period determination of LMXBs possible with observing strategies `astro_sim_01_1004`, `Minion_1020`, `baseline2018a` and `Minion_1016`. Y axis denotes the r band reddening in magnitudes, X axis r mag before adding contributions from ellipsoidal modulation, flaring, noise and reddening. The colour denotes the completenss of the period recovery.

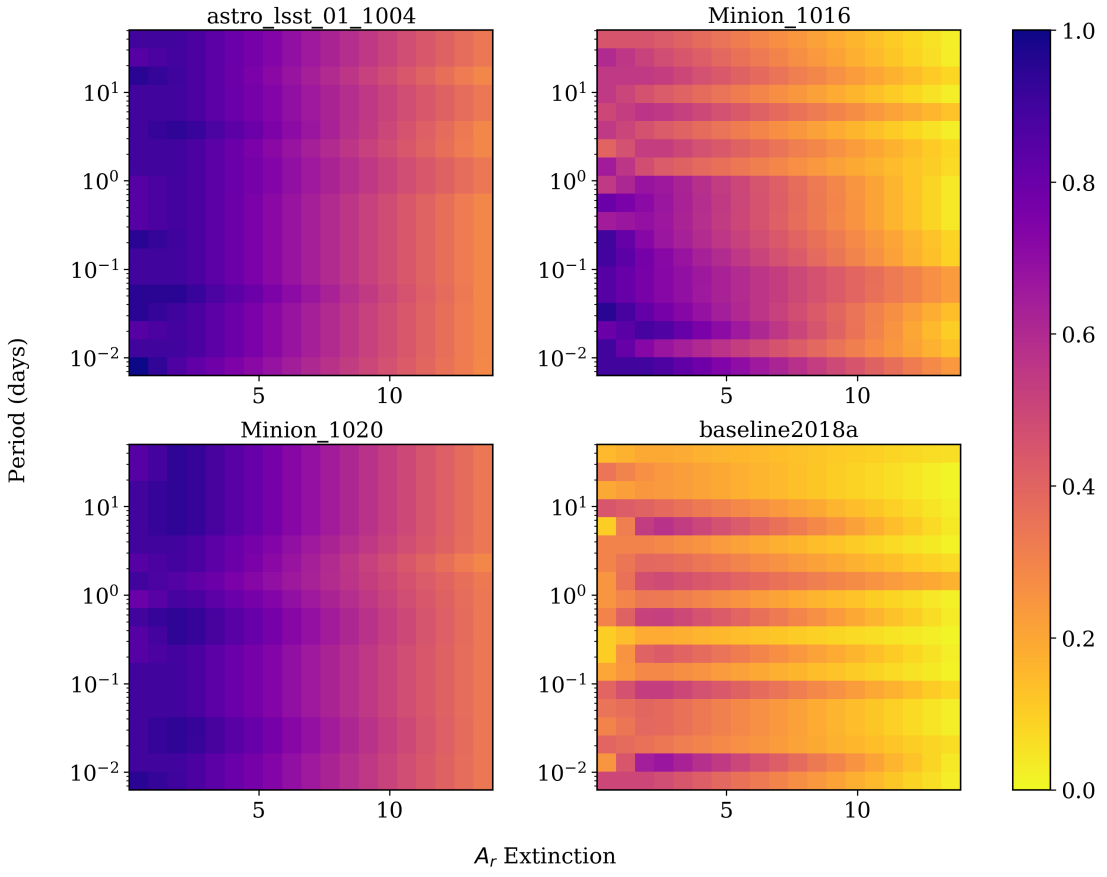


FIGURE A.5: Colour maps displaying the relationship between period, extinction and period determination of LMXBs possible with observing strategies `astro_lsst_01_1004`, `Minion_1020`, `baseline2018a` and `Minion_1016`. Y axis denotes the period in days, X axis r band reddening in magnitudes. The colour denotes the completeness of the period recovery.

Appendix B

Appendix: Finding Transients Surplus

B.1 SExtractor Settings

The full list of SExtractor filters used in the simulations was:

- default.conv
- gauss_1.5_3x3.conv
- gauss_2.0_3x3.conv
- gauss_2.0_5x5.conv
- gauss_2.5_5x5.conv
- gauss_3.0_5x5.conv
- gauss_3.0_7x7.conv
- gauss_4.0_7x7.conv
- gauss_5.0_9x9.conv
- mexhat_1.5_5x5.conv
- mexhat_2.0_7x7.conv
- mexhat_2.5_7x7.conv
- mexhat_3.0_9x9.conv
- mexhat_4.0_9x9.conv

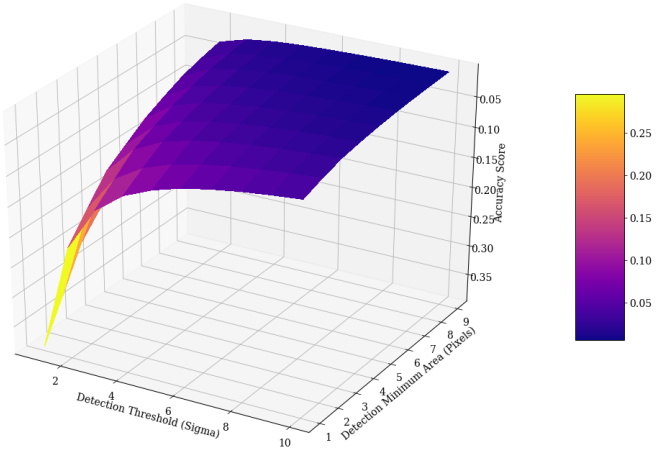


FIGURE B.1: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'default.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

- mexhat_5.0_11x11.conv
- tophat_1.5_3x3.conv
- tophat_2.0_3x3.conv
- tophat_2.5_3x3.conv
- tophat_3.0_3x3.conv
- tophat_4.0_5x5.conv
- tophat_5.0_5x5.conv

B.2 Completeness and Accuracy for Each Filter

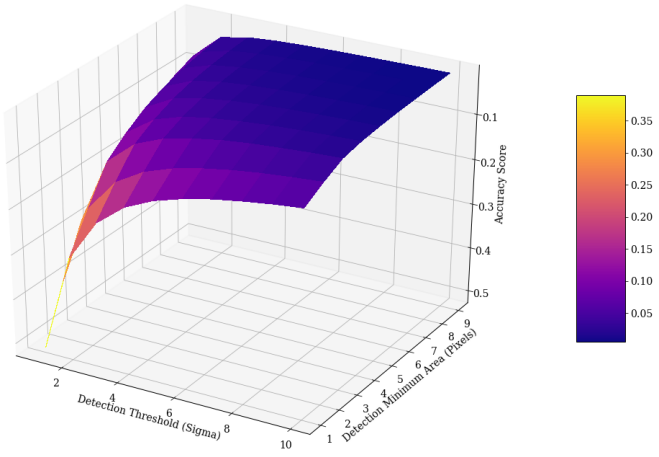


FIGURE B.2: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_1.5_3x3.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

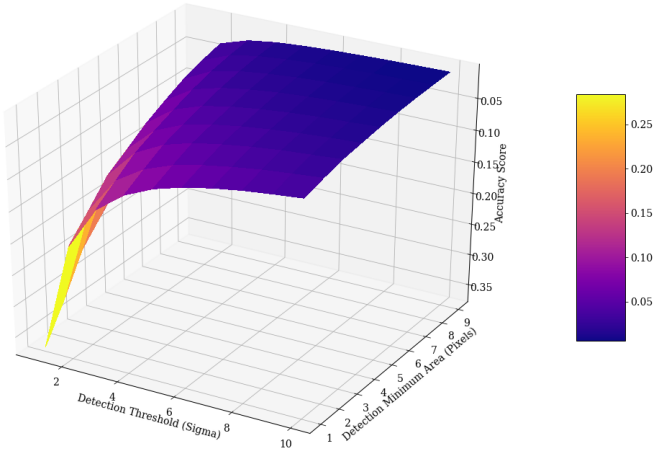


FIGURE B.3: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_2.0_3x3.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

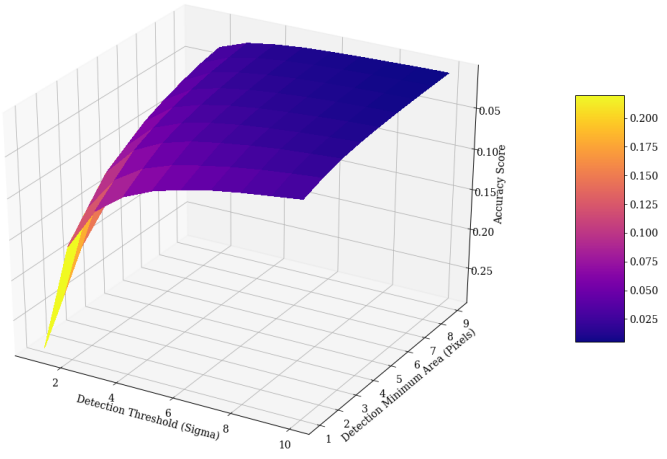


FIGURE B.4: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_2.0_5x5.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

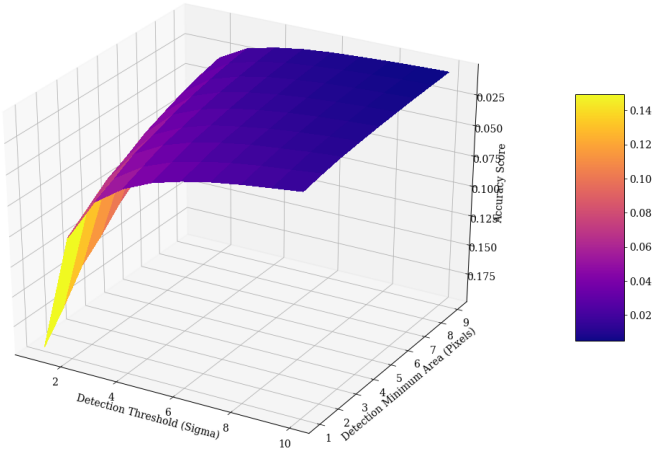


FIGURE B.5: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_2.5_5x5.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

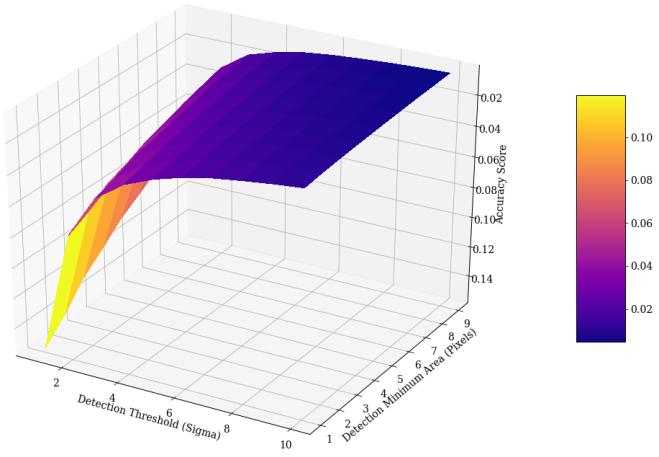


FIGURE B.6: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_3.0_5x5.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

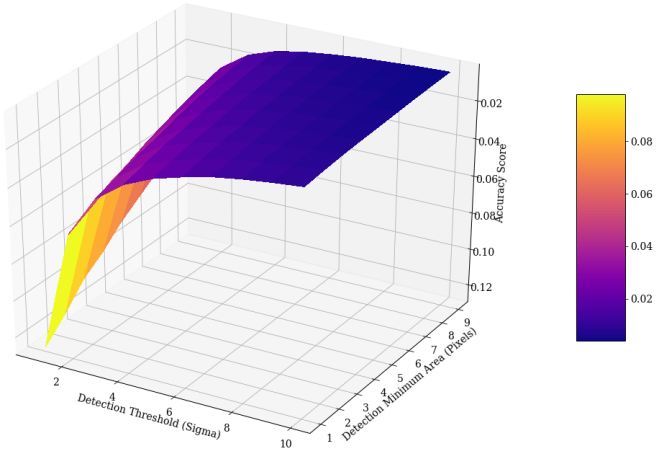


FIGURE B.7: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_3.0_7x7.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

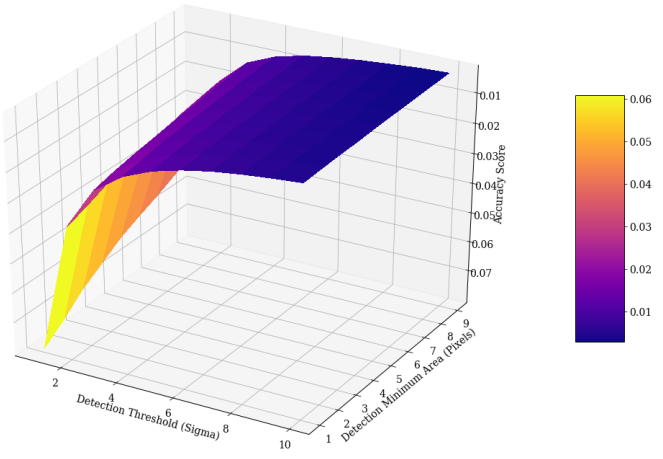


FIGURE B.8: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_4.0_7x7.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

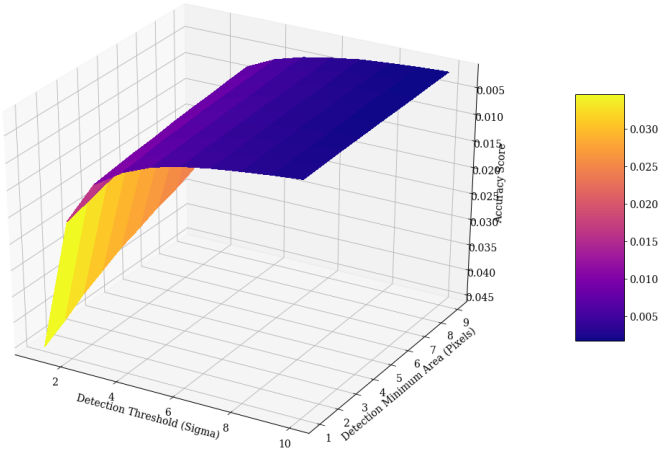


FIGURE B.9: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_5.0_9x9.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

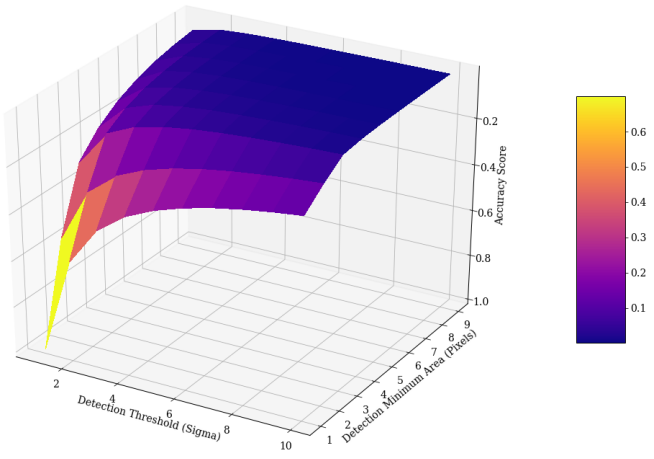


FIGURE B.10: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'mexhat_1.5_5x5.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

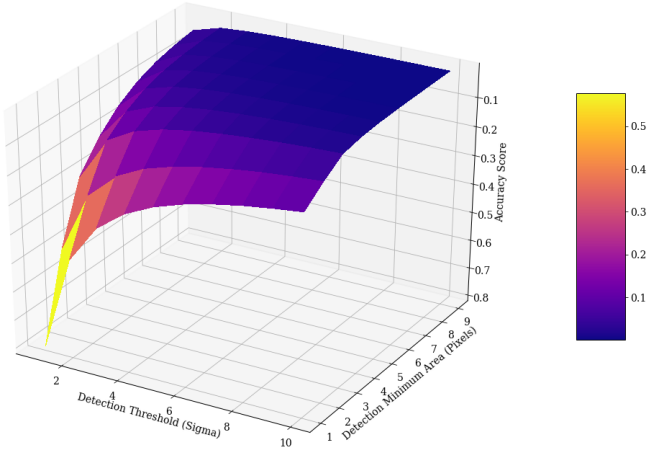


FIGURE B.11: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'mexhat_2.0_7x7.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

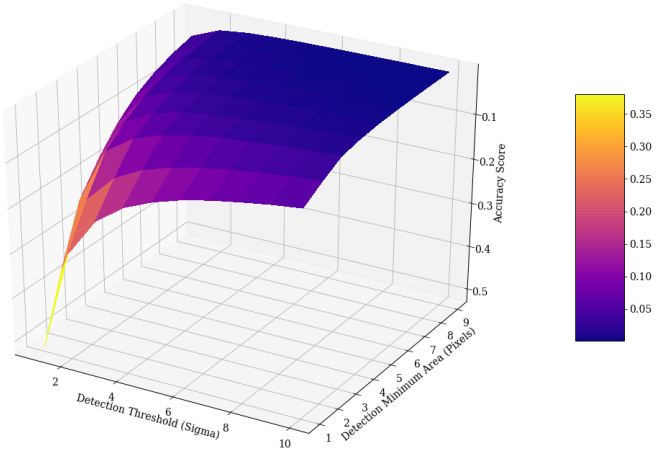


FIGURE B.12: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'mexhat_2.5_7x7.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

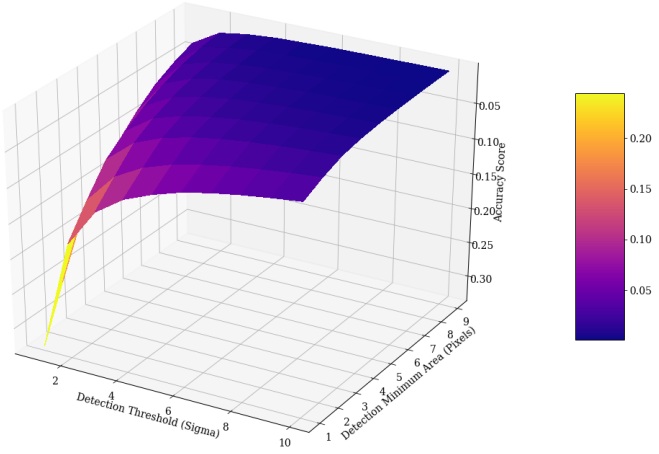


FIGURE B.13: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'mexhat_3.0_9x9.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

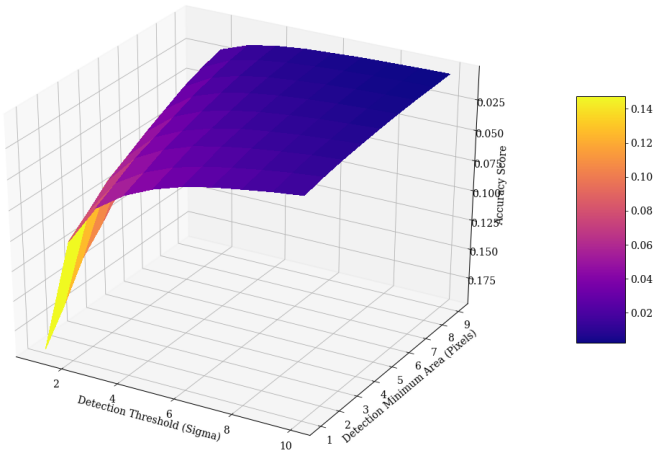


FIGURE B.14: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'mexhat_4.0_9x9.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

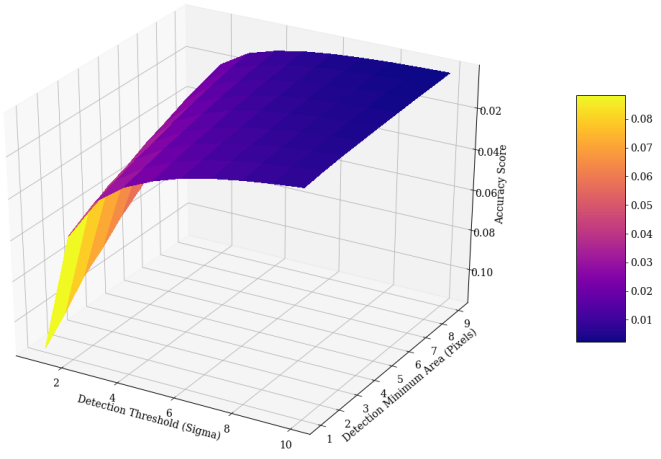


FIGURE B.15: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'mexhat_5.0_11x11.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

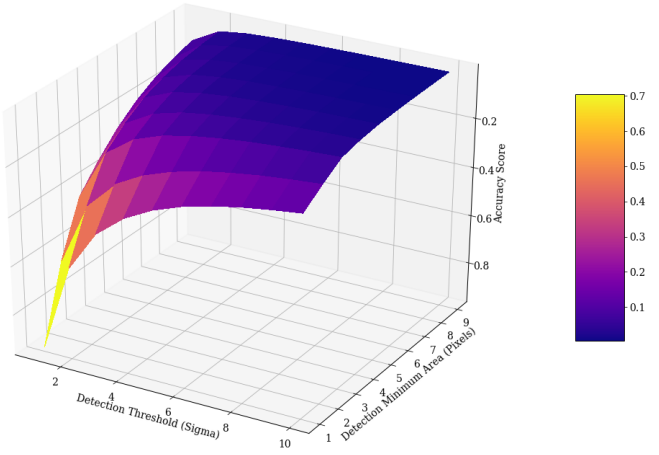


FIGURE B.16: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'tophat_1.5_3x3.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

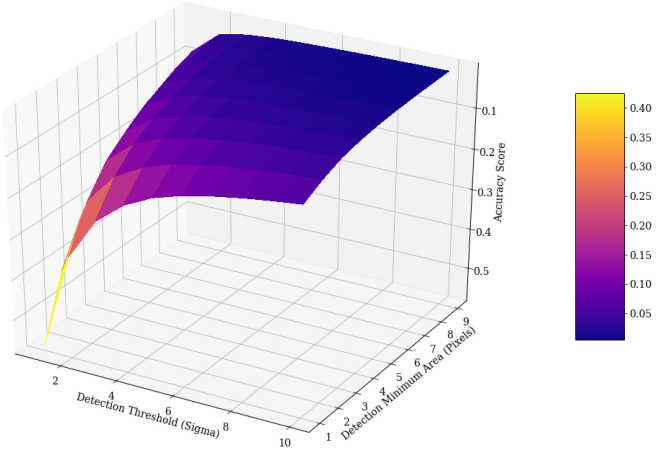


FIGURE B.17: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'tophat_2.0_3x3.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

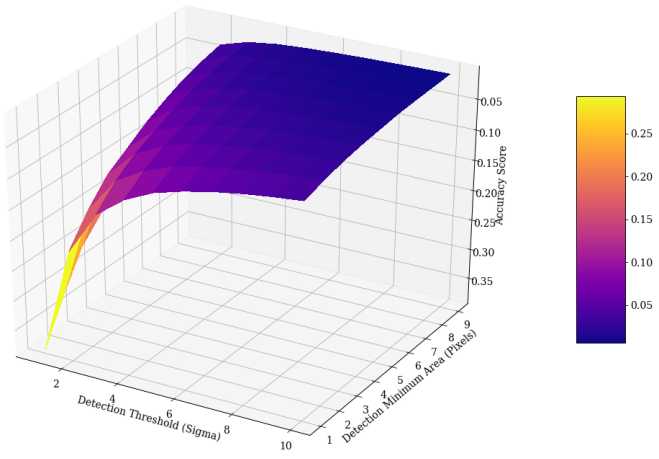


FIGURE B.18: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'tophat_2.5_3x3.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

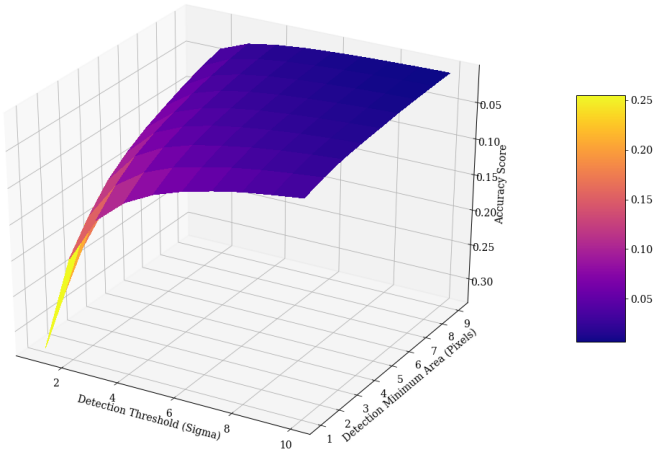


FIGURE B.19: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'tophat_3.0_3x3.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

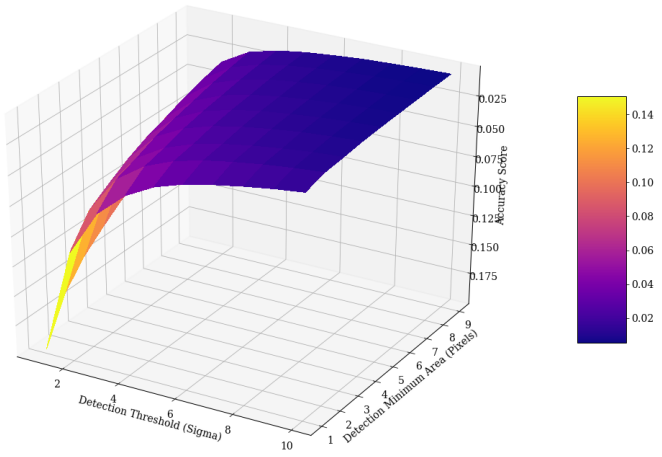


FIGURE B.20: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'tophat_4.0_5x5.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

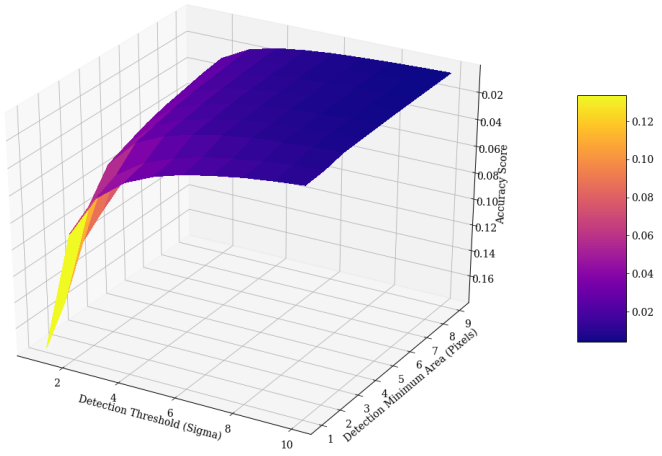


FIGURE B.21: The accuracy of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'tophat_5.0_5x5.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

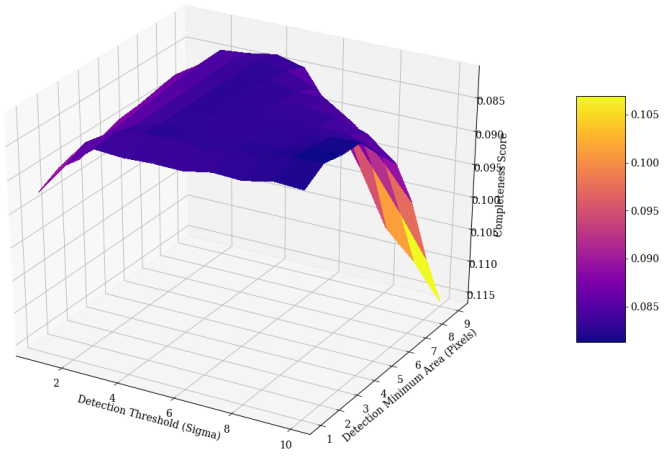


FIGURE B.22: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'default.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

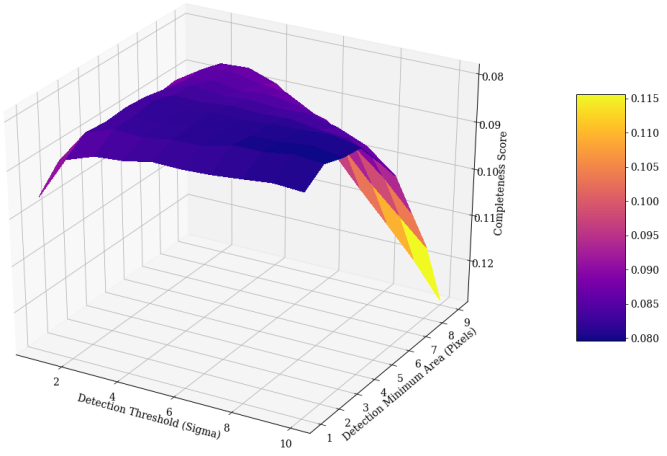


FIGURE B.23: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_1.5_3x3.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

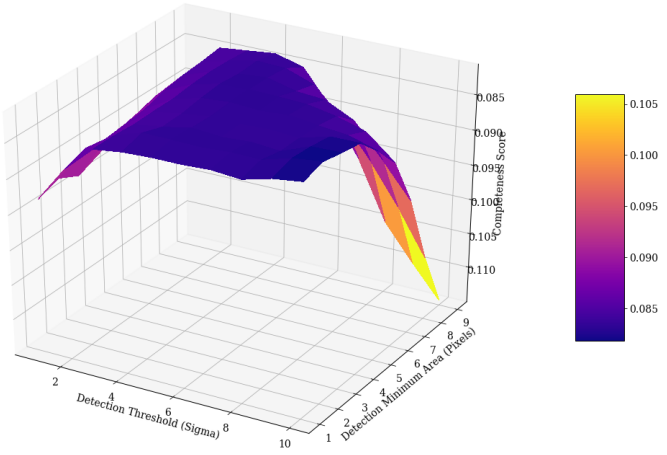


FIGURE B.24: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_2.0_3x3.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

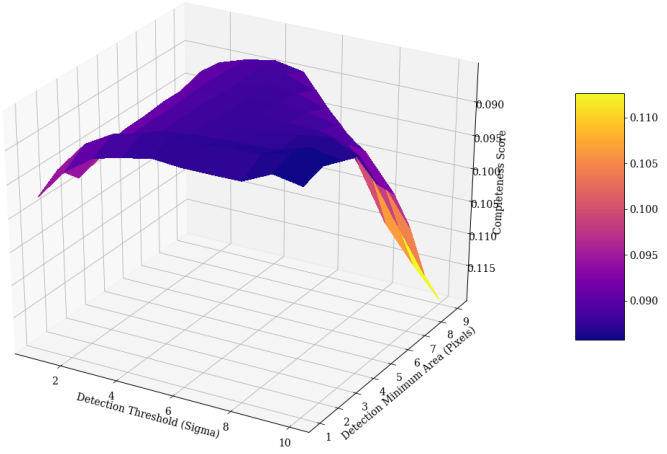


FIGURE B.25: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_2.0_5x5.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

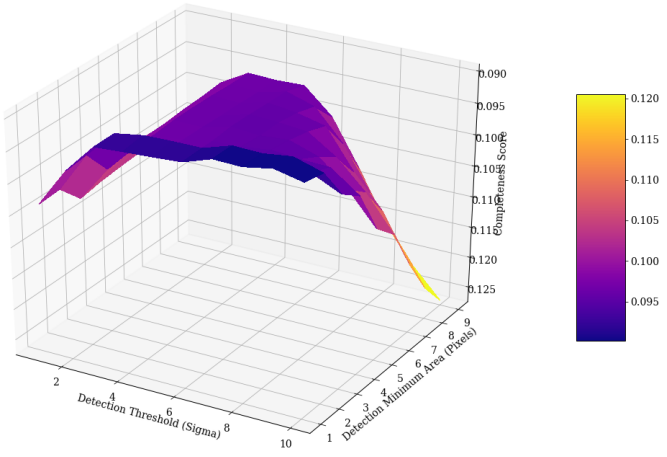


FIGURE B.26: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_2.5_5x5.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

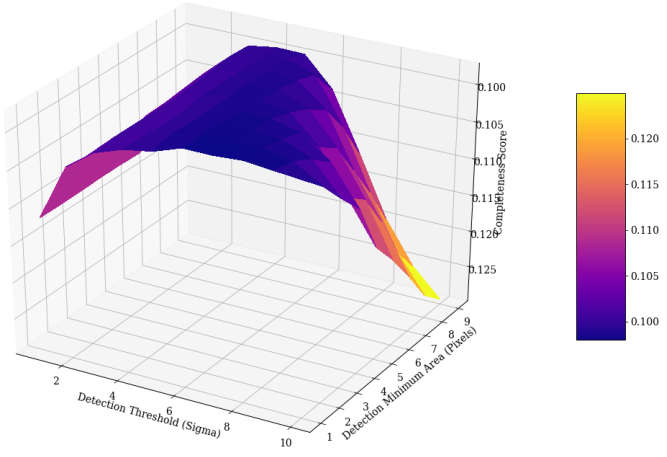


FIGURE B.27: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_3.0_5x5.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

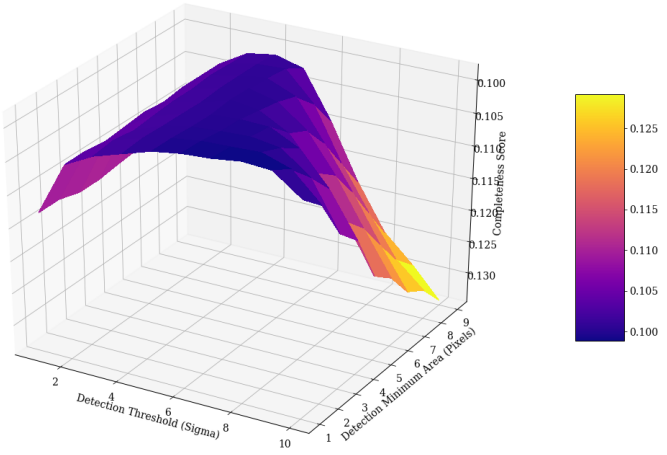


FIGURE B.28: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_3.0_7x7.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

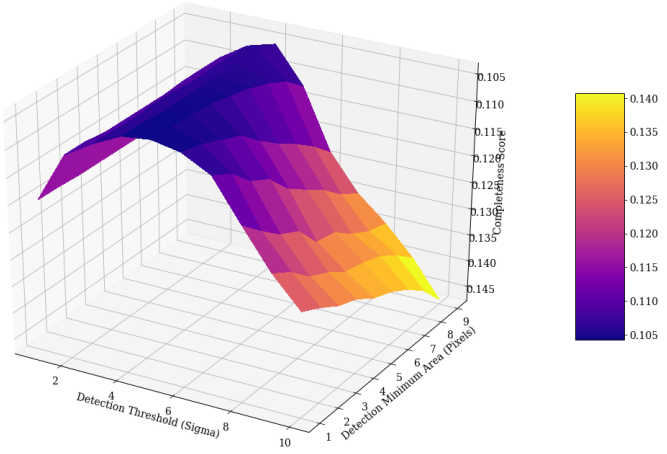


FIGURE B.29: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_4.0_7x7.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

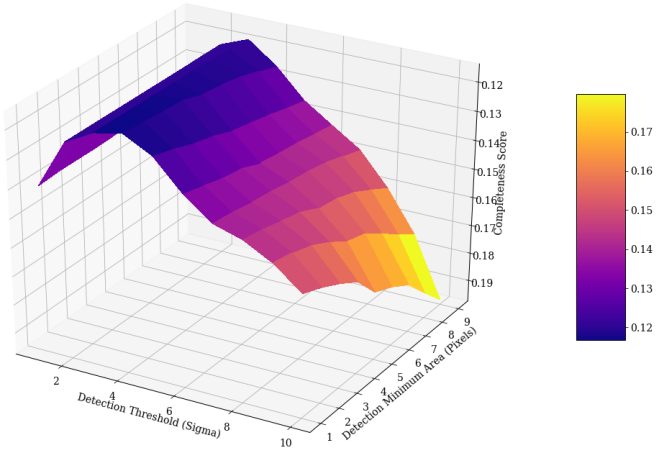


FIGURE B.30: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'gauss_5.0_9x9.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

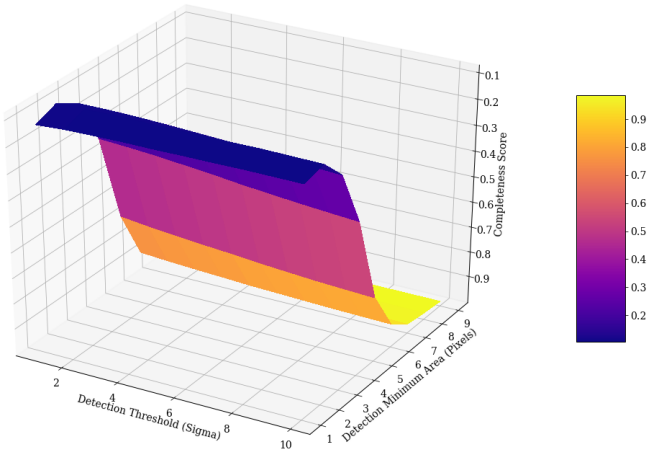


FIGURE B.31: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'mexhat_1.5_5x5.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

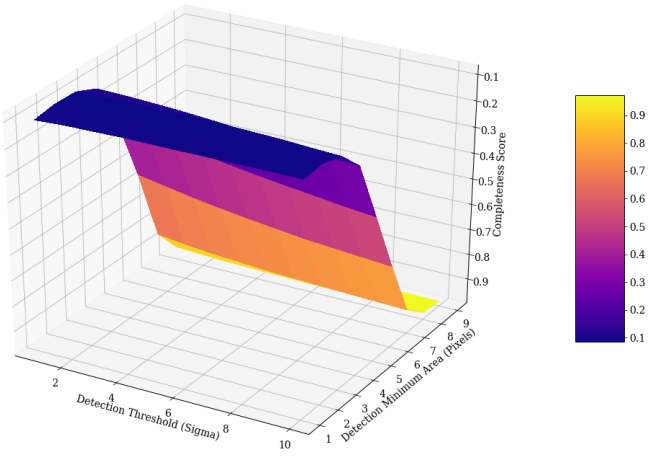


FIGURE B.32: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'mexhat_2.0_7x7.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

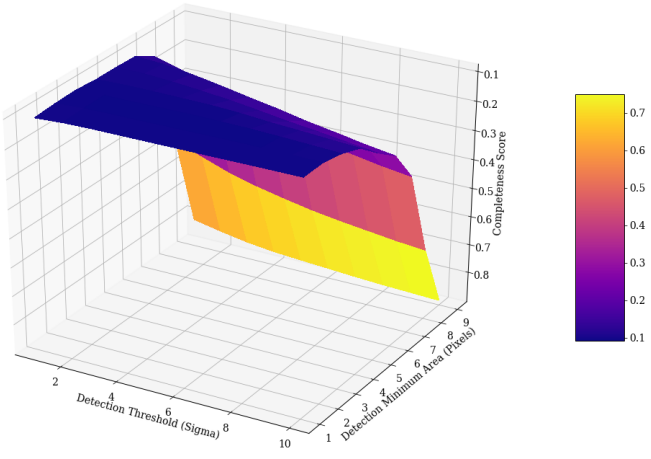


FIGURE B.33: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'mexhat_2.5_7x7.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

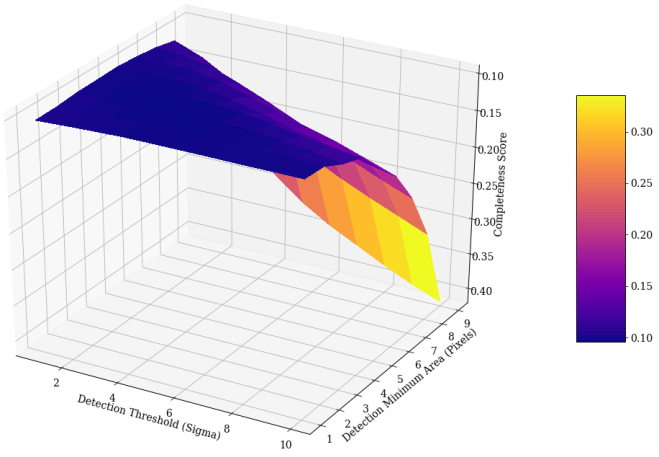


FIGURE B.34: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'mexhat_3.0_9x9.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

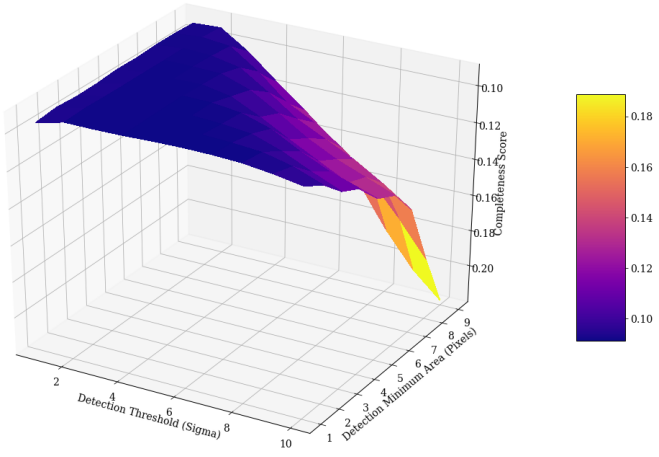


FIGURE B.35: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'mexhat_4.0_9x9.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

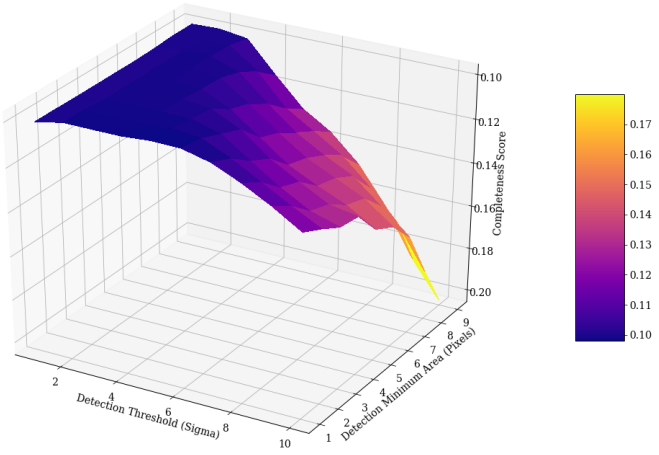


FIGURE B.36: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'mexhat_5.0_11x11.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

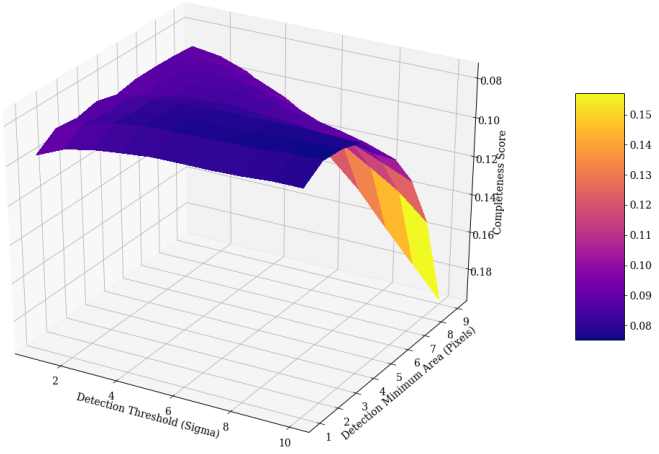


FIGURE B.37: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'tophat_1.5_3x3.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

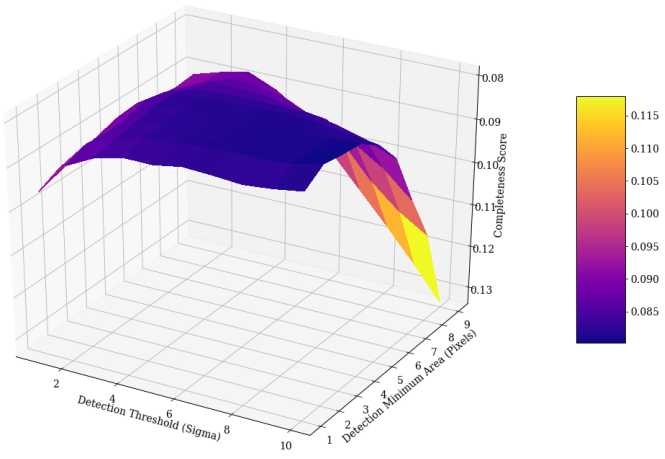


FIGURE B.38: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'tophat_2.0_3x3.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

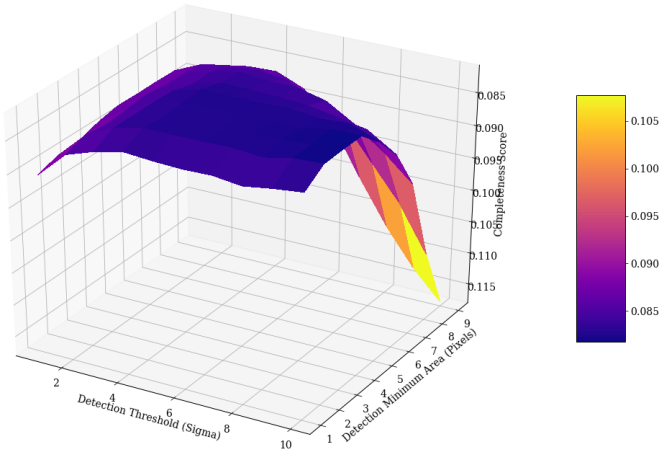


FIGURE B.39: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'tophat_2.5_3x3.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

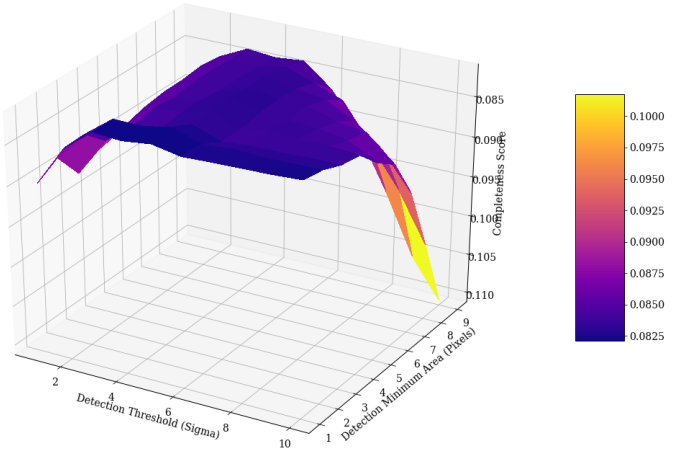


FIGURE B.40: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'tophat_3.0_3x3.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

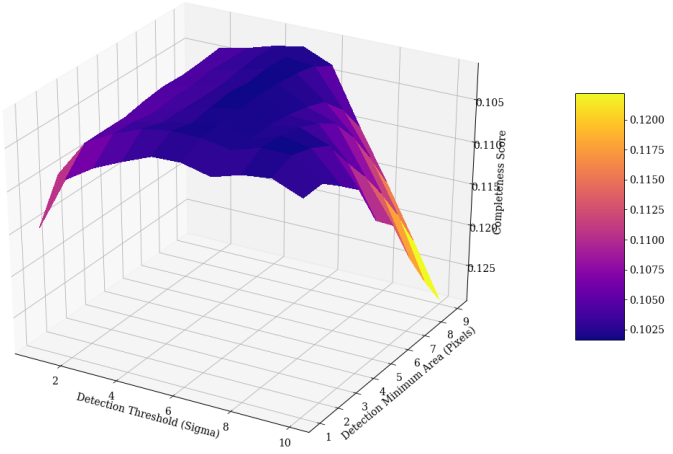


FIGURE B.41: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'tophat_4.0_5x5.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

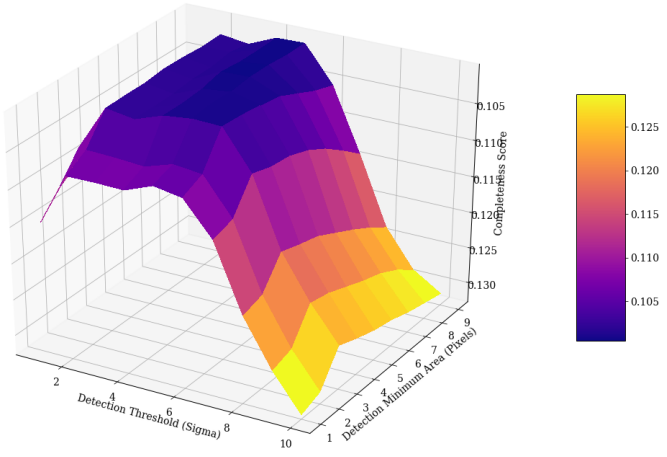


FIGURE B.42: The completeness of the detection workflow when using different combinations of detection threshold and minimum area for the filter 'tophat_5.0_5x5.conv'. The accuracy quality metric was averaged over the spatial distribution, magnitude range and all three CCD channels used.

Appendix C

Appendix: Kepler Object Tables

C.1 Three Magnitude Subsample

C.2 One Magnitude Subsample

C.3 Signal to Noise Subsample

RA	DEC	IMAGE_NAME	Observations	FLUX_BEST	MAG_BEST	X_IMAGE	Y_IMAGE	magDiff	MAG_BEST_ORIG	S/N
283.449	46.912	kplr2009114174833_ffi-cal.fits[13]	1	1563.474	17.015	977.231	335.001	-4.066	16.989	3.722
285.11	50.034	kplr2009114174833_ffi-cal.fits[36]	1	2366.871	16.565	1036.981	203.49	-4.073	16.539	4.58
288.391	46.044	kplr2009114174833_ffi-cal.fits[38]	1	5791.752	15.593	683.066	458.652	-3.34	15.542	7.074
283.143	43.818	kplr2009114204835_ffi-cal.fits[2]	1	1997.534	16.749	872.868	244.026	-3.846	16.717	4.195
291.334	47.224	kplr2009114204835_ffi-cal.fits[58]	1	3272.575	16.213	211.005	340.045	-3.428	16.166	5.328
284.838	42.514	kplr2009114204835_ffi-cal.fits[5]	1	1907.738	16.799	392.007	72.979	-3.019	16.729	4.024
297.649	48.704	kplr2009115053616_ffi-cal.fits[79]	1	5967.598	15.56	41.88	591.88	-3.13	15.498	7.141
281.467	42.863	kplr2009115080620_ffi-cal.fits[3]	1	1981.667	16.757	890.552	521.852	-3.473	16.712	4.15
283.846	47.95	kplr2009115131122_ffi-cal.fits[13]	1	1581.959	17.002	206.894	925.194	-4.599	16.986	3.763
281.852	42.547	kplr2009115131122_ffi-cal.fits[3]	1	2053.565	16.719	506.641	539.568	-3.514	16.675	4.229
282.192	42.66	kplr2009115173611_ffi-cal.fits[3]	1	3443.877	16.157	443.936	780.318	-3.212	16.099	5.437
281.912	46.137	kplr2009116035924_ffi-cal.fits[14]	1	2716.361	16.415	62.005	386.222	-8.695	16.415	4.969
282.987	46.563	kplr2009116035924_ffi-cal.fits[14]	1	1910.534	16.797	826.698	264.587	-3.087	16.732	4.035
282.09	47.067	kplr2009260000800_ffi-cal.fits[74]	1	2340.553	16.577	685.131	969.03	-3.727	16.541	4.532
282.853	47.956	kplr2009260000800_ffi-cal.fits[76]	1	2880.242	16.351	670.958	861.313	-5.911	16.347	5.105
284.737	49.056	kplr2009260000800_ffi-cal.fits[79]	1	2042.471	16.725	481.946	775.912	-6.441	16.722	4.303
282.364	46.862	kplr2009322233047_ffi-cal.fits[70]	1	1773.411	16.878	700.304	722.31	-3.778	16.844	3.948
284.001	41.263	kplr2009322233047_ffi-cal.fits[80]	1	3168.766	16.248	948.387	961.877	-3.872	16.217	5.285
282.277	48.768	kplr2009351005245_ffi-cal.fits[72]	1	2080.474	16.705	479.996	69.157	-4.169	16.681	4.298
286.028	37.831	kplr2009351005245_ffi-cal.fits[75]	1	3921.754	16.016	517.714	230.414	-6.135	16.012	5.96
280.164	43.631	kplr2009351005245_ffi-cal.fits[84]	1	1970.285	16.764	245.835	311.711	-3.085	16.698	4.098
284.246	45.277	kplr2010019225502_ffi-cal.fits[28]	1	1902.258	16.802	747.011	710.905	-3.186	16.742	4.038
280.557	43.237	kplr2010019225502_ffi-cal.fits[32]	1	1991.641	16.752	684.767	272.437	-6.519	16.749	4.249
285.594	39.601	kplr2010019225502_ffi-cal.fits[72]	1	2901.56	16.343	234.292	927.921	-3.48	16.298	5.023
280.317	43.812	kplr2010020005046_ffi-cal.fits[32]	1	1381.485	17.149	184.83	486.866	-3.189	17.09	3.442
281.101	42.812	kplr2010049182302_ffi-cal.fits[31]	1	1472.71	17.08	1014.999	303.879	-3.081	17.014	3.542
290.347	46.525	kplr2010078174524_ffi-cal.fits[21]	1	2991.176	16.31	283.999	141.994	-3.155	16.249	5.059
284.81	49.408	kplr2010111125026_ffi-cal.fits[35]	1	4732.54	15.812	705.773	549.001	-3.55	15.77	6.424
288.255	46.25	kplr2010111125026_ffi-cal.fits[38]	1	2054.668	16.718	719.349	660.895	-3.29	16.664	4.209
282.859	46.193	kplr2010140101631_ffi-cal.fits[14]	1	1787.397	16.869	555.753	52.574	-3.319	16.817	3.928
289.771	50.397	kplr2010140101631_ffi-cal.fits[54]	1	2127.889	16.68	1004.02	669.565	-3.091	16.615	4.259
282.429	44.064	kplr2010174164113_ffi-cal.fits[1]	1	2136.207	16.676	888.903	469.89	-3.381	16.627	4.301
287.802	38.545	kplr2010203012215_ffi-cal.fits[14]	1	3689.878	16.082	304.791	529.6	-3.661	16.044	5.684
290.815	43.395	kplr2010203012215_ffi-cal.fits[43]	1	3751.609	16.064	219.632	620.69	-3.572	16.023	5.722
283.347	41.235	kplr2010234192745_ffi-cal.fits[36]	1	2376.854	16.56	699.983	593.608	-3.391	16.511	4.537
284.701	45.873	kplr2010234192745_ffi-cal.fits[57]	1	1632.154	16.968	497.939	894.014	-3.828	16.936	3.791
287.943	46.448	kplr2010234192745_ffi-cal.fits[62]	1	2738.449	16.406	666.002	920.837	-6.767	16.404	4.984
284.283	47.191	kplr2010234192745_ffi-cal.fits[73]	1	3092.372	16.274	412.014	217.548	-3.262	16.219	5.159
283.517	48.412	kplr2010234192745_ffi-cal.fits[76]	1	1653.425	16.954	100.677	786.978	-4.26	16.932	3.836
284.753	41.626	kplr2010265195356_ffi-cal.fits[33]	1	2411.984	16.544	998.965	600.225	-4.607	16.528	4.646
283.043	48.171	kplr2010296192119_ffi-cal.fits[72]	1	3096.144	16.273	459.948	778.747	-3.219	16.215	5.157
282.765	46.649	kplr2010326181728_ffi-cal.fits[70]	1	2219.893	16.634	772.982	416.318	-3.688	16.597	4.411
284.18	47.873	kplr2010356020128_ffi-cal.fits[69]	1	3245.061	16.222	84.014	750.364	-3.958	16.193	5.355
280.449	47.311	kplr2010356020128_ffi-cal.fits[71]	1	4151.407	15.954	51.565	377.907	-3.324	15.902	5.987
280.51	43.195	kplr2011024134926_ffi-cal.fits[32]	1	2285.962	16.602	693.748	224.036	-3.124	16.539	4.419
288.315	47.834	kplr2011053174401_ffi-cal.fits[24]	1	1938.77	16.781	659.166	341.798	-3.395	16.732	4.098
282.509	42.697	kplr2011053174401_ffi-cal.fits[31]	1	3533.898	16.129	344.527	963.932	-4.591	16.113	5.623
281.894	48.563	kplr2011145152723_ffi-cal.fits[16]	1	2089.83	16.7	779.993	73.929	-3.0	16.629	4.209
282.647	44.088	kplr2011145152723_ffi-cal.fits[1]	1	1872.953	16.819	960.375	345.468	-4.083	16.793	4.074
283.69	47.081	kplr2011208112727_ffi-cal.fits[73]	1	2167.62	16.66	761.134	363.283	-4.197	16.637	4.389
280.175	43.18	kplr2011240181752_ffi-cal.fits[56]	1	1893.198	16.807	564.786	51.293	-3.187	16.748	4.029
287.762	50.681	kplr2011240181752_ffi-cal.fits[83]	1	3024.814	16.298	200.503	630.946	-3.016	16.228	5.066
285.517	39.54	kplr2011271191331_ffi-cal.fits[16]	1	2820.82	16.374	251.994	855.92	-4.138	16.35	5.003
281.243	42.818	kplr2011271191331_ffi-cal.fits[55]	1	3057.423	16.287	955.38	383.896	-3.51	16.243	5.159
284.856	46.345	kplr2011271191331_ffi-cal.fits[57]	1	2911.947	16.34	215.508	557.785	-3.847	16.308	5.065
289.293	49.652	kplr2011271191331_ffi-cal.fits[82]	1	1448.479	17.098	402.0	264.932	-10.821	17.098	3.629
288.209	47.833	kplr2011303191211_ffi-cal.fits[48]	1	1549.625	17.024	709.014	305.781	-3.002	16.954	3.624
287.827	47.209	kplr2012004204112_ffi-cal.fits[47]	1	2245.84	16.622	1001.952	635.16	-4.594	16.606	4.483
282.159	47.789	kplr2012004204112_ffi-cal.fits[72]	1	1613.083	16.981	1094.007	712.722	-3.15	16.92	3.715
285.314	39.156	kplr2012004204112_ffi-cal.fits[76]	1	3021.807	16.299	441.706	533.645	-3.984	16.271	5.169
281.681	48.389	kplr2012060123308_ffi-cal.fits[12]	1	1426.346	17.114	977.998	108.763	-3.608	17.075	3.531
292.752	49.812	kplr2012060123308_ffi-cal.fits[20]	1	2915.664	16.338	50.004	281.993	-3.138	16.276	4.993
281.877	44.557	kplr2012060123308_ffi-cal.fits[29]	1	4237.023	15.932	311.149	467.074	-3.407	15.884	6.06
281.872	44.888	kplr2012060123308_ffi-cal.fits[29]	1	2127.698	16.68	77.009	280.858	-3.443	16.634	4.298
286.911	40.389	kplr2012121122500_ffi-cal.fits[9]	1	13498.66	14.674	198.199	80.055	-6.017	14.67	11.054
286.532	49.334	kplr2012151105138_ffi-cal.fits[33]	1	4027.266	15.988	744.986	958.871	-3.287	15.934	5.891
280.397	43.921	kplr2012151105138_ffi-cal.fits[4]	1	2411.027	16.544	148.979	592.224	-3.226	16.487	4.551
286.464	48.791	kplr2012179140901_ffi-cal.fits[33]	1	3040.686	16.293	1066.939	584.724	-3.038	16.224	5.083
292.606	46.732	kplr2012211123923_ffi-cal.fits[46]	1	3181.333	16.244	63.999	614.419	-4.121	16.219	5.312
280.778	47.661	kplr2012211123923_ffi-cal.fits[75]	1	4252.868	15.928	409.928	267.25	-3.082	15.863	6.02
282.388	47.949	kplr2012242195726_ffi-cal.fits[76]	1	1853.191	16.83	895.068	690.002	-4.465	16.812	4.068
284.76	49.687	kplr2012277203051_ffi-cal.fits[79]	1	2154.875	16.666	838.821	329.902	-4.446	16.648	4.386
284.188	48.839	kplr2012310200152_ffi-cal.fits[51]	1	2152.764	16.668	103.959	728.281	-3.546	16.625	4.332
286.564	50.338	kplr2012310200152_ffi-cal.fits[52]	1	2112.741	16.688	192.154	479.722	-3.16	16.627	4.253
287.261	42.182	kplr2012310200152_ffi-cal.fits[63]	1	2559.199	16.48	637.912	641.087	-3.426	16.432	4.712
284.789	38.976	kplr2012310200152_ffi-cal.fits[76]	1	5917.287	15.57	351.917	140.044	-5.065	15.559	7.297
281.785	43.938	kplr2012310200152_ffi-cal.fits[81]	1	3516.285	16.135	709.047	861.905	-3.601	16.095	5.543
283.483	40.46	kplr2012341215621_ffi-cal.fits[79]	1	2617.89	16.455	918.486	236.798	-3.949	16.426	4.809
282.893	41.811	kplr2012341215621_ffi-cal.fits[80]	1	4067.363	15.977	95.95	673.824	-3.499	15.933	5.949
281.144	42.286	kplr2012341215621_ffi-cal.fits[83]	1	6691.445	15.436	629.349	28.036	-3.381	15.387	7.611
280.261	43.235	kplr2013038133130_ffi-cal.fits[32]	1	2329.518	16.582	559.989	122.253	-4.679	16.567	4.568
294.097	47.099	kplr2013038133130_ffi-cal.fits[39]	1	3183.856	16.243	250.876	593.983	-4.442	16.224	5.331
281.548	47.15	kplr2013065115251_ffi-cal.fits[11]	1	1905.201	16.8	470.99	921.995	-3.067	16.734	4.027
291.026	49.118	kplr2013065115251_ffi-cal.fits[19]	1	3078.284	16.279	981.346	275.369	-3.164	16.219	5.134
286.289	38.949	kplr2013098115308_ffi-cal.fits[72]	1	4782.51	15.801	997.994	973.936	-3.695	15.764	6.475

TABLE C.1: All objects within the Kepler Full Frame Images that were found to exhibit a three magnitude or greater increase in brightness between an image and the corresponding median image. None of this sample were found within any other Kepler images nor the Simbad or Gaia databases.

MAIN_ID	OTYPE	RA	DEC	IMAGE_NAME	Number_of_Observations	FLUX_BEST	MAG_BEST	X_IMAGE	Y_IMAGE	magDiff	MAG_BEST_ORIG	S/N	parallax	parallaxError	phot_g_mean_flux	bp_rp
2MASS J19293151+3742406	Mira	292.381	37.711	kplr2011271191331_ffi-cal.fits[3]	40	37926.73	13.553	838.085	439.081	-1.238	13.134	15.129	-0.263	0.123	17777.813	5.777
2MASS J19293151+3742406	Mira	292.381	37.711	kplr2012242195726_ffi-cal.fits[3]	40	51622.38	13.218	838.032	439.229	-1.497	12.903	18.558	-0.263	0.123	17777.813	5.777
2MASS J19293151+3742406	Mira	292.381	37.711	kplr2012277203051_ffi-cal.fits[3]	40	89740.15	12.618	838.082	439.118	-2.001	12.43	26.045	-0.263	0.123	17777.813	5.777
IRAS 18554+4753	LPV*	284.205	47.953	kplr2009351005245_ffi-cal.fits[69]	52	43268.85	13.41	27.925	798.66	-1.529	13.105	17.078	-0.213	0.109	50027.034	5.752
KIC 12055999	Star	288.148	50.575	kplr2012341215621_ffi-cal.fits[31]	2	5116.245	15.728	328.145	830.047	-1.302	15.338	5.635	1.423	0.341	250.874	0.758
N/A	Object Not Found	284.647	41.188	kplr2009115053616_ffi-cal.fits[6]	1	2314.183	16.589	952.188	898.863	-1.038	16.062	3.547	0.286	0.275	438.334	1.505
N/A	Object Not Found	282.269	46.442	kplr2009115080620_ffi-cal.fits[14]	1	3649.695	16.094	409.359	459.208	-1.312	15.709	4.769	1.132	0.203	538.154	2.213
N/A	Object Not Found	281.204	47.118	kplr2009115131122_ffi-cal.fits[15]	1	1790.891	16.867	289.282	814.907	-2.971	16.795	3.892	-0.765	1.262	86.718	0.982
N/A	Object Not Found	281.981	44.765	kplr2010174164113_ffi-cal.fits[1]	1	1197.028	17.305	212.345	297.4	-2.221	17.154	3.063	0.729	0.448	266.062	0.866
N/A	Object Not Found	292.208	46.464	kplr2011177110110_ffi-cal.fits[62]	3	3011.133	16.303	139.034	272.727	-1.307	15.916	4.328	0.294	0.51	162.391	1.275
N/A	Object Not Found	292.208	46.464	kplr2011334181008_ffi-cal.fits[22]	3	3280.754	16.21	134.581	275.843	-1.248	15.796	4.459	0.294	0.51	162.391	1.275
N/A	Object Not Found	285.191	46.633	kplr2012004204112_ffi-cal.fits[65]	1	3070.549	16.282	128.0	235.262	-1.743	16.038	4.687	0.176	0.204	449.648	0.989
N/A	Object Not Found	281.277	47.992	kplr2012032101442_ffi-cal.fits[11]	1	2091.945	16.699	830.021	230.047	-1.227	16.275	3.544	0.689	0.449	184.063	2.127
N/A	Object Not Found	292.208	46.464	kplr2012211123923_ffi-cal.fits[46]	3	2952.152	16.325	140.142	276.712	-1.106	15.838	4.085	0.294	0.51	162.391	1.275
N/A	Object Not Found	283.917	44.599	kplr2012277203051_ffi-cal.fits[60]	1	2667.026	16.435	1107.267	172.403	-1.033	15.905	3.802	0.0	0.0	64.358	0.0
N/A	Object Not Found	289.365	44.393	kplr2013038133130_ffi-cal.fits[42]	1	3834.07	16.041	727.662	325.493	-1.561	15.746	5.109	-0.989	1.041	99.133	0.408
USNO-B1.0 1360-00297059	CataclyV*	284.662	46.035	kplr2011303191211_ffi-cal.fits[65]	1	7895.583	15.256	360.362	825.101	-1.444	14.923	7.193	1.486	0.855	117.496	0.515
V* V1119 Cyg	Mira	291.436	51.159	kplr2010111125026_ffi-cal.fits[53]	51	55979.6	13.13	46.633	728.84	-1.561	12.836	19.521	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2010140101631_ffi-cal.fits[53]	51	97939.29	12.523	46.605	728.849	-2.042	12.343	27.311	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2010174164113_ffi-cal.fits[53]	50	75195.51	12.81	46.517	728.939	-1.807	12.582	23.37	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2012121122500_ffi-cal.fits[53]	51	68378.33	12.913	46.628	728.84	-1.694	12.656	21.981	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2012151105138_ffi-cal.fits[53]	50	63194.7	12.998	46.588	728.855	-1.744	12.755	21.266	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2012179140901_ffi-cal.fits[53]	51	34454.98	13.657	46.524	728.929	-1.215	13.228	14.344	0.039	0.055	53177.273	3.688
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2009115080620_ffi-cal.fits[46]	14	28915.06	13.847	933.806	121.833	-1.092	13.353	12.733	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2009115131122_ffi-cal.fits[46]	14	27873.98	13.887	933.805	121.832	-1.071	13.38	12.429	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2009115173611_ffi-cal.fits[46]	14	28571.27	13.86	933.806	121.832	-1.087	13.363	12.64	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2009116035924_ffi-cal.fits[46]	14	25883.04	13.968	933.806	121.83	-1.014	13.425	11.773	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2010078174524_ffi-cal.fits[62]	14	17408.26	14.398	932.33	116.54	-1.158	13.94	10.055	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2010326181728_ffi-cal.fits[38]	14	24754.64	14.016	928.236	115.407	-1.327	13.637	12.46	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2011240181752_ffi-cal.fits[22]	14	15169.22	14.548	928.063	120.025	-1.033	14.018	9.068	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2012277203051_ffi-cal.fits[22]	14	17395.65	14.399	928.06	120.013	-1.174	13.949	10.091	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2012310200152_ffi-cal.fits[38]	14	28411.51	13.866	928.226	115.413	-1.514	13.557	13.806	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2013065115251_ffi-cal.fits[62]	14	36705.66	13.588	932.34	116.521	-1.803	13.359	16.321	1.897	0.048	23115.667	0.042
V* V344 Lyr	DwarfNova	281.163	43.375	kplr2012151105138_ffi-cal.fits[4]	6	6822.164	15.415	849.79	661.178	-1.125	14.939	6.24	0.94	0.076	5702.46	-0.063

TABLE C.2: The subsample of variable events found which have the restrictions that: they displayed at least a one magnitude increase in brightness between the target and median images; the objects had a counterpart in the Simbad astronomical database. MAG_BEST and FLUX_BEST refer to the magnitude and flux measured in the difference image, whereas MAG_BEST_ORIG refers to the magnitude measured in the target Kepler image.

MAIN_ID	OTYPE	RA	DEC	IMAGE_NAME	Number_of_Observations	FLUX_BEST	MAG_BEST	X_IMAGE	Y_IMAGE	magDiff	MAG_BEST_ORIG	S/N	parallax	parallaxError	phot_g_mean_flux	bp_rp
2MASS J19293151+3742406	Mira	292.381	37.711	kplr2011271191331_ffi-cal.fits[3]	40	37926.73	13.553	838.085	439.081	-1.238	13.134	15.129	-0.263	0.123	17777.813	5.777
2MASS J19293151+3742406	Mira	292.381	37.711	kplr2012242195726_ffi-cal.fits[3]	40	51622.38	13.218	838.032	439.229	-1.497	12.903	18.558	-0.263	0.123	17777.813	5.777
2MASS J19293151+3742406	Mira	292.381	37.711	kplr2012277203051_ffi-cal.fits[3]	40	89740.15	12.618	838.082	439.118	-2.001	12.43	26.045	-0.263	0.123	17777.813	5.777
2MASS J19382873+3913505	Mira	294.62	39.231	kplr2011024134926_ffi-cal.fits[80]	46	29914.74	13.81	877.024	803.766	-0.675	12.974	10.993	0.018	0.14	37304.545	5.622
IRAS 18554+4753	LPV*	284.205	47.953	kplr2009351005245_ffi-cal.fits[69]	52	43268.85	13.41	27.925	798.66	-1.529	13.105	17.078	-0.213	0.109	50027.034	5.752
IRAS 18554+4753	LPV*	284.205	47.953	kplr2010111125026_ffi-cal.fits[13]	52	26469.42	13.943	33.214	794.03	-0.716	13.153	10.574	-0.213	0.109	50027.034	5.752
IRAS 18554+4753	LPV*	284.205	47.953	kplr2011116104002_ffi-cal.fits[13]	52	42100.35	13.439	33.337	794.107	-0.648	12.571	12.847	-0.213	0.109	50027.034	5.752
IRAS 18554+4753	LPV*	284.205	47.953	kplr2011145152723_ffi-cal.fits[13]	52	37337.67	13.57	33.214	794.05	-0.921	12.963	13.709	-0.213	0.109	50027.034	5.752
IRAS 18554+4753	LPV*	284.205	47.953	kplr2012121122500_ffi-cal.fits[13]	52	43733.82	13.398	33.325	794.104	-0.678	12.565	13.316	-0.213	0.109	50027.034	5.752
IRAS 18554+4753	LPV*	284.205	47.953	kplr2012151105138_ffi-cal.fits[13]	52	34128.14	13.667	33.204	794.073	-0.874	13.024	12.876	-0.213	0.109	50027.034	5.752
IRAS 19545+4512	Mira	299.045	45.35	kplr2010078174524_ffi-cal.fits[54]	52	60628.25	13.043	552.774	939.221	-0.689	12.224	15.777	0.004	0.083	73954.318	5.444
IRAS 19545+4512	Mira	299.046	45.35	kplr2011024134926_ffi-cal.fits[54]	51	62104.21	13.017	552.97	939.56	-0.512	11.955	14.231	0.004	0.083	73954.318	5.444
IRAS 19545+4512	Mira	299.045	45.35	kplr2011303191211_ffi-cal.fits[2]	52	65437.53	12.96	552.819	945.465	-0.35	11.561	12.461	0.004	0.083	73954.318	5.444
Mis V0148	Mira	299.234	44.59	kplr2010049182302_ffi-cal.fits[53]	31	57846.21	13.094	1005.362	734.516	-0.793	12.38	16.208	0.17	0.119	42717.605	5.011
Mis V0148	Mira	299.234	44.59	kplr2010326181728_ffi-cal.fits[1]	33	47545.71	13.307	1005.595	740.984	-0.594	12.369	13.205	0.17	0.119	42717.605	5.011
Mis V0148	Mira	299.234	44.59	kplr2012242195726_ffi-cal.fits[29]	32	52028.72	13.209	1000.626	739.675	-0.832	12.531	15.634	0.17	0.119	42717.605	5.011
N/A	Object Not Found	298.473	45.769	kplr2011271191331_ffi-cal.fits[30]	50	30896.52	13.775	54.079	795.322	-0.734	13.003	11.523	0.034	0.107	53379.275	5.648
N/A	Object Not Found	299.137	43.627	kplr2012060123308_ffi-cal.fits[53]	21	33147.73	13.699	257.358	282.223	-0.468	12.56	10.028	0.073	0.093	42896.1	5.483
N/A	Object Not Found	298.473	45.769	kplr2012211123923_ffi-cal.fits[30]	50	47273.95	13.313	54.036	795.206	-0.891	12.684	15.258	0.034	0.107	53379.275	5.648
N/A	Object Not Found	298.473	45.769	kplr2012242195726_ffi-cal.fits[30]	50	43497.98	13.404	54.095	794.966	-0.615	12.494	12.8	0.034	0.107	53379.275	5.648
N/A	Object Not Found	298.474	45.769	kplr2012310200152_ffi-cal.fits[2]	50	59005.44	13.073	49.007	797.408	-0.442	11.884	13.069	0.034	0.107	53379.275	5.648
V* FX Cyg	Mira	296.04	39.79	kplr2010356020128_ffi-cal.fits[35]	51	48404.14	13.288	232.174	748.123	-0.44	12.095	11.811	0.13	0.156	221865.57	6.725
V* V1119 Cyg	Mira	291.436	51.159	kplr2010111125026_ffi-cal.fits[53]	51	55979.6	13.13	46.633	728.84	-1.561	12.836	19.521	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2010140101631_ffi-cal.fits[53]	51	97939.29	12.523	46.605	728.849	-2.042	12.343	27.311	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2010174164113_ffi-cal.fits[53]	50	75195.51	12.81	46.517	728.939	-1.807	12.582	23.37	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2011053174401_ffi-cal.fits[1]	51	32114.22	13.733	46.778	735.558	-0.715	12.942	11.639	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2011240181752_ffi-cal.fits[81]	51	31865.85	13.742	42.069	728.993	-0.516	12.687	10.23	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2011271191331_ffi-cal.fits[81]	49	33021.05	13.703	41.926	729.107	-0.568	12.728	10.817	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2012121122500_ffi-cal.fits[53]	51	68378.33	12.913	46.628	728.84	-1.694	12.656	21.981	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2012151105138_ffi-cal.fits[53]	50	63194.7	12.998	46.588	728.855	-1.744	12.755	21.266	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2012179140901_ffi-cal.fits[53]	51	34454.98	13.657	46.524	728.929	-1.215	13.228	14.344	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2013011160902_ffi-cal.fits[29]	50	49646.75	13.26	41.763	735.21	-0.81	12.562	15.132	0.039	0.055	53177.273	3.688
V* V1119 Cyg	Mira	291.436	51.159	kplr2013038133130_ffi-cal.fits[1]	50	27296.17	13.91	46.726	735.674	-0.651	13.045	10.361	0.039	0.055	53177.273	3.688
V* V1155 Cyg	Mira	297.109	42.558	kplr2012242195726_ffi-cal.fits[28]	18	73266.34	12.838	25.369	932.23	-0.352	11.445	13.225	0.087	0.143	104531.199	5.686
V* V1253 Cyg	LPV*	289.737	44.957	kplr2012179140901_ffi-cal.fits[43]	52	33066.68	13.702	839.609	410.683	-0.577	12.739	10.888	0.021	0.048	95164.398	4.145
V* V1253 Cyg	LPV*	289.738	44.957	kplr2012341215621_ffi-cal.fits[41]	52	72292.12	12.852	842.28	406.738	-0.458	11.694	14.674	0.021	0.048	95164.398	4.145
V* V1292 Cyg	Mira	297.002	41.877	kplr2010234192745_ffi-cal.fits[9]	52	38861.06	13.526	910.934	91.445	-0.528	12.49	11.395	-0.174	0.105	44856.205	5.12
V* V1292 Cyg	Mira	297.002	41.877	kplr2012088132324_ffi-cal.fits[73]	52	28640.87	13.858	906.498	86.015	-0.692	13.04	10.857	-0.174	0.105	44856.205	5.12
V* V1503 Cyg	Mira	289.972	43.762	kplr2011271191331_ffi-cal.fits[43]	52	49899.73	13.255	800.586	340.575	-0.321	11.776	10.482	-0.113	0.105	100801.406	5.443
V* V1503 Cyg	Mira	289.972	43.762	kplr2012242195726_ffi-cal.fits[43]	49	113937.7	12.358	800.64	340.874	-0.315	10.862	15.712	-0.113	0.105	100801.406	5.443
V* V1503 Cyg	Mira	289.972	43.762	kplr2012277203051_ffi-cal.fits[43]	50	121919.6	12.285	800.746	340.626	-0.264	10.62	15.021	-0.113	0.105	100801.406	5.443
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2009115080620_ffi-cal.fits[46]	14	28915.06	13.847	933.806	121.833	-1.092	13.353	12.733	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2009115131122_ffi-cal.fits[46]	14	27873.98	13.887	933.805	121.832	-1.071	13.38	12.429	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2009115173611_ffi-cal.fits[46]	14	28571.27	13.86	933.806	121.832	-1.087	13.363	12.64	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2009116035924_ffi-cal.fits[46]	14	25883.04	13.968	933.806	121.83	-1.014	13.425	11.773	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2010078174524_ffi-cal.fits[62]	14	17408.26	14.398	932.33	116.54	-1.158	13.94	10.055	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2010326181728_ffi-cal.fits[38]	14	24754.64	14.016	928.236	115.407	-1.327	13.637	12.46	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2012277203051_ffi-cal.fits[22]	14	17395.65	14.399	928.06	120.013	-1.174	13.949	10.091	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2012310200152_ffi-cal.fits[38]	14	28411.51	13.866	928.226	115.413	-1.514	13.557	13.806	1.897	0.048	23115.667	0.042
V* V1504 Cyg	DwarfNova	292.235	43.094	kplr2013065115251_ffi-cal.fits[62]	14	36705.66	13.588	932.34	116.521	-1.803	13.359	16.321	1.897	0.048	23115.667	0.042
V* V1670 Cyg	Mira	292.931	41.427	kplr2011177110110_ffi-cal.fits[48]	51	41792.39	13.447	897.45	516.218	-0.592	12.506	12.363	0.033	0.071	57602.128	4.449
V* V355 Lyr	RRLyr	283.358	43.155	kplr2010140101631_ffi-cal.fits[2]	33	27690.26	13.894	313.45	503.624	-0.645	13.022	10.402	0.191	0.023	31639.805	0.752
V* V390 Cyg	Mira	292.98	48.459	kplr2012179140901_ffi-cal.fits[57]	28	80022.41	12.742	574.117	760.217	-0.204	10.827	10.827	0.167	0.099	90311.73	5.15

TABLE C.3: This table displays the sub-sample of transient events discovered by the workflow in the difference images which all have a signal to noise ratio above ten. None of these objects were found in any other Kepler FFI and non had counterparts within the Simbad database or the Gaia DR2 catalogue. MAG_BEST and FLUX_BEST refer to the magnitude and flux measured in the difference image, whereas MAG_BEST_ORIG refers to the magnitude measured in the target Kepler image.

Bibliography

Suzanne Aigrain, Simon T Hodgkin, Michael J Irwin, Jim R Lewis, and Stephen J Roberts. Precise time series photometry for the kepler 2.0 mission. *Monthly notices of the royal astronomical society*, 447(3):2880–2893, 2015.

C Alard. Image subtraction using a space-varying kernel. *Astronomy and Astrophysics Supplement Series*, 144(2):363–370, 2000.

Ilkay Altintas, Chad Berkley, Efrat Jaeger, Matthew Jones, Bertram Ludascher, and Steve Mock. Kepler: an extensible system for design and execution of scientific workflows. In *Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004.*, pages 423–424. IEEE, 2004.

EA Antokhina and AM Cherepashchuk. The ellipticity effect in x-ray novae light curves. *Astronomy Reports*, 41(3):364–377, 1997.

K Arur and TJ Maccarone. Selection effects on the orbital period distribution of low-mass black hole x-ray binaries. *Monthly Notices of the Royal Astronomical Society*, 474(1):69–76, 2017.

James E Bailey and Sammy W Pearson. Development of a tool for measuring and analyzing computer user satisfaction. *Management science*, 29(5):530–545, 1983.

Donald P Ballou and Harold L Pazer. Modeling data and process quality in multi-input, multi-output information systems. *Management science*, 31(2):150–162, 1985.

Andrew Becker. Hotpants: High order transform of psf and template subtraction. *Astrophysics Source Code Library*, 2015.

Jacek Becla, Andrew Hanushevsky, Sergei Nikolaev, Ghaleb Abdulla, Alex Szalay, Maria Nieto-Santisteban, Ani Thakar, and Jim Gray. Designing a multi-petabyte database for lsst. In *Observatory Operations: Strategies, Processes, and Systems*, volume 6270, page 62700R. International Society for Optics and Photonics, 2006.

Eric C Bellm. The zwicky transient facility. *arXiv preprint arXiv:1410.8185*, 2014.

F Bernardini, DM Russell, AW Shaw, F Lewis, PA Charles, KII Koljonen, JP Lasota, and J Casares. Events leading up to the 2015 june outburst of v404 cyg. *The Astrophysical Journal Letters*, 818(1):L5, 2016.

Emmanuel Bertin and Stephane Arnouts. Sextractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series*, 117(2):393–404, 1996.

William J Borucki, David Koch, Gibor Basri, Natalie Batalha, Timothy Brown, Douglas Caldwell, John Caldwell, Jørgen Christensen-Dalsgaard, William D Cochran, Edna DeVore, et al. Kepler planet-detection mission: introduction and first results. *Science*, 327(5968):977–980, 2010.

James Bosch, Robert Armstrong, Steven Bickerton, Hisanori Furusawa, Hiroyuki Ikeda, Michitaro Koike, Robert Lupton, Sogo Mineo, Paul Price, Tadafumi Takata, et al. The hyper suprime-cam software pipeline. *Publications of the Astronomical Society of Japan*, 70(SP1):S5, 2017.

AGA Brown, A Vallenari, T Prusti, JHJ De Bruijne, C Babusiaux, CAL Bailer-Jones, M Biermann, Dafydd Wyn Evans, L Eyer, Femke Jansen, et al. Gaia data release 2-summary of the contents and survey properties. *Astronomy & astrophysics*, 616:A1, 2018.

Anthony GA Brown, A Vallenari, T Prusti, JHJ De Bruijne, F Mignard, R Drimmel, C Babusiaux, CAL Bailer-Jones, U Bastian, M Biermann, et al. Gaia data release 1-summary of the astrometric, photometric, and survey properties. *Astronomy & Astrophysics*, 595:A2, 2016.

Stephen T Bryson, Peter Tenenbaum, Jon M Jenkins, Hema Chandrasekaran, Todd Klaus, Douglas A Caldwell, Ronald L Gilliland, Michael R Haas, Jessie L Dotson, David G Koch, et al. The kepler pixel response function. *The Astrophysical Journal Letters*, 713(2):L97, 2010.

Steven P Callahan, Juliana Freire, Emanuele Santos, Carlos E Scheidegger, Cláudio T Silva, and Huy T Vo. Vistrails: visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 745–747. ACM, 2006.

N Cappelluti, P Predehl, H Böhringer, H Brunner, M Brusa, V Burwitz, Evgeniy Churazov, K Dennerl, M Freyberg, A Finoguenov, et al. erosita on srg: a x-ray all-sky survey mission. *arXiv preprint arXiv:1004.5219*, 2010.

J Casares. J. casares and pg jonker, space sci. rev. 183, 223 (2014). *Space Sci. Rev.*, 183: 223, 2014.

J Casares and PG Jonker. Mass measurements of stellar and intermediate-mass black holes. *Space Science Reviews*, 183(1-4):223–252, 2014.

Jorge Casares, Peter G Jonker, and Garik Israelian. X-ray binaries. *arXiv preprint arXiv:1701.07450*, 2017.

Jorge Casares and Manuel AP Torres. A feasibility study on the photometric detection of quiescent black hole x-ray binaries. *Monthly Notices of the Royal Astronomical Society*, 481(4):4372–4380, 2018.

Jesus M Corral-Santana, Jorge Casares, Teo Munoz-Darias, Franz E Bauer, Ignacio G Martinez-Pais, and David M Russell. Blackcat: A catalogue of stellar-mass black holes in x-ray transients. *Astronomy & Astrophysics*, 587:A61, 2016.

Ron Cowen. The wheels come off kepler. *Nature News*, 497(7450):417, 2013.

RM Cutri, MF Skrutskie, S Van Dyk, CA Beichman, JM Carpenter, T Chester, L Cambresy, T Evans, J Fowler, J Gizis, et al. Vizier online data catalog: 2mass all-sky catalog of point sources (cutri+ 2003). *VizieR Online Data Catalog*, 2246, 2003.

RM e Cutri et al. Vizier online data catalog: Allwise data release (cutri+ 2013). *VizieR Online Data Catalog*, 2328, 2014.

GS Da Costa. Basic photometry techniques. In *Astronomical CCD Observing and Reduction Techniques*, volume 23, page 90, 1992.

Francisco Delgado, Abhijit Saha, Srinivasan Chandrasekharan, Kem Cook, Catherine Petry, and Stephen Ridgway. The lsst operations simulator. In *SPIE Astronomical Telescopes+ Instrumentation*, pages 915015–915015. International Society for Optics and Photonics, 2014.

Jonathan H Elias and Cesar Briceño. Soar telescope operation in the lsst era: real-time follow-up on large scales. In *Observatory Operations: Strategies, Processes, and Systems VI*, volume 9910, page 99100W. International Society for Optics and Photonics, 2016.

Tommy Ellkvist, Lena Strömbäck, Lauro Didier Lins, and Juliana Freire. A first study on strategies for generating workflow snippets. In *Proceedings of the First International Workshop on Keyword Search on Structured Data*, pages 15–20. ACM, 2009.

Laurent Eyer, Burkhard Holl, Dimitri Pourbaix, Nami Mowlavi, Christos Siopis, F Barblan, DW Evans, and Pierre North. The gaia mission. *arXiv preprint arXiv:1303.0303*, 2013.

Matteo Fischetti, Juan José Salazar González, and Paolo Toth. A branch-and-cut algorithm for the symmetric generalized traveling salesman problem. *Operations Research*, 45(3):378–394, 1997.

Ian Foster, Jens Vockler, Michael Wilde, and Yong Zhao. Chimera: A virtual data system for representing, querying, and automating data derivation. In *Scientific and Statistical Database Management, 2002. Proceedings. 14th International Conference on*, pages 37–46. IEEE, 2002.

Chris L Fryer and Vassiliki Kalogera. Theoretical black hole mass distributions. *The Astrophysical Journal*, 554(1):548, 2001.

Poshak Gandhi, Anjali Rao, Michael AC Johnson, John A Paice, and Thomas J Maccarone. Gaia data release 2 distances and peculiar velocities for galactic black hole transients. *Monthly Notices of the Royal Astronomical Society*, 485(2):2642–2655, 2019.

Poshak Gandhi, Anjali Rao, Michael AC Johnson, John A Paice, and Tom J Maccarone. Gaia dr2 distances and peculiar velocities for galactic black hole transients. *arXiv preprint arXiv:1804.11349*, 2018.

Yolanda Gil, Ewa Deelman, Mark Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole Goble, Miron Livny, Luc Moreau, and Jim Myers. Examining the challenges of scientific workflows. *Computer*, 40(12):24–32, 2007.

Loukas Grafakos. *Classical fourier analysis*, volume 2. Springer, 2008.

J Greiner, JG Cuby, MJ McCaughrean, AJ Castro-Tirado, and RE Mennickent. Identification of the donor in the x-ray binary grs 1915+ 105. *Astronomy & Astrophysics*, 373(3):L37–L40, 2001.

H-J Grimm, M Gilfanov, and R Sunyaev. The milky way in x-rays for an outside observer-log (n)-log (s) and luminosity function of x-ray binaries from rxte/asm data. *Astronomy & Astrophysics*, 391(3):923–944, 2002.

Paul Groth, Ewa Deelman, Gideon Juve, Gaurang Mehta, and Bruce Berriman. Pipeline-centric provenance model. *arXiv preprint arXiv:1005.4457*, 2010.

Manuel Guizar-Sicairos, Samuel T Thurman, and James R Fienup. Efficient subpixel image registration algorithms. *Optics letters*, 33(2):156–158, 2008.

Dennis Halloran, Susan Manchester, John Moriarty, Robert Riley, James Rohrman, and Thomas Skramstad. Systems development quality control. *MIS Quarterly*, pages 1–13, 1978.

RJ Hanisch. Data standards for the international virtual observatory. *Data Science Journal*, 5:168–173, 2006.

James V Hansen. Audit considerations in distributed processing systems. *Communications of the ACM*, 26(8):562–569, 1983.

Marco Hetterscheidt, T Erben, P Schneider, R Maoli, L Van Waerbeke, and Y Mellier. Searching for galaxy clusters using the aperture mass statistics in 50 vlt fields. *Astronomy & Astrophysics*, 442(1):43–61, 2005.

Benne W Holwerda. Source extractor for dummies v5. *arXiv preprint astro-ph/0512139*, 2005.

Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Mathew R Pocock, Peter Li, and Tom Oinn. Taverna: a tool for building and running workflows of services. *Nucleic acids research*, 34(suppl_2):W729–W732, 2006.

Z Ivezic. Lsst: from science drivers to reference design and anticipated data products. Technical report, SLAC National Accelerator Lab., Menlo Park, CA (United States), 2014.

Ž Ivezić, RH Lupton, D Schlegel, B Boroski, J Adelman-McCarthy, B Yanny, S Kent, C Stoughton, Douglas Finkbeiner, N Padmanabhan, et al. Sdss data management and photometric quality assessment. *Astronomische Nachrichten: Astronomical Notes*, 325 (6-8):583–589, 2004.

Savannah Jacklin, Michael B Lund, Joshua Pepper, and Keivan G Stassun. Transiting planets with lsst. ii. period detection of planets orbiting 1 mâŁŹ hosts. *The Astronomical Journal*, 150(1):34, 2015.

Savannah R Jacklin, Michael B Lund, Joshua Pepper, and Keivan G Stassun. Transiting planets with lsst. iii. detection rate per year of operation. *The Astronomical Journal*, 153(4):186, 2017.

Sheldon H Jacobson, Laura A McLay, Shane N Hall, Darrall Henderson, and Diane E Vaughan. Optimal search strategies using simultaneous generalized hill climbing algorithms. *Mathematical and computer modelling*, 43(9-10):1061–1073, 2006.

Hans-Thomas Janka. Natal kicks of stellar mass black holes by asymmetric mass ejection in fallback supernovae. *Monthly Notices of the Royal Astronomical Society*, 434(2): 1355–1361, 2013.

Jon M Jenkins, Joseph D Twicken, Sean McCauliff, Jennifer Campbell, Dwight Sanderfer, David Lung, Masoud Mansouri-Samani, Forrest Girouard, Peter Tenenbaum, Todd Klaus, et al. The tess science processing operations center. In *Software and Cyberinfrastructure for Astronomy IV*, volume 9913, page 99133E. International Society for Optics and Photonics, 2016.

Michael AC Johnson, Poshak Gandhi, Adriane P Chapman, Luc Moreau, Philip A Charles, William I Clarkson, and Adam B Hill. Prospecting for periods with lsst–low-mass x-ray binaries as a test case. *Monthly Notices of the Royal Astronomical Society*, 484(1):19–30, 2018a.

Michael AC Johnson, Luc Moreau, Adriane Chapman, Poshak Gandhi, and Carlos Sáenz-Adán. Using the provenance from astronomical workflows to increase processing efficiency. In *International Provenance and Annotation Workshop*, pages 101–112. Springer, 2018b.

Peter G Jonker and Gijs Nelemans. The distances to galactic low-mass x-ray binaries: consequences for black hole luminosities and kicks. *Monthly Notices of the Royal Astronomical Society*, 354(2):355–366, 2004.

Mario Jurić. Data products definition document, 2018.

Mario Juric, Jeffrey Kantor, K Lim, Robert H Lupton, Gregory Dubois-Felsmann, Tim Jenness, Tim S Axelrod, Jovan Aleksic, Roberta A Allsman, Yusra AlSayyad, et al. The lsst data management system, 2015.

N. Kaiser, H. Aussel, B. E. Burke, H. Boesgaard, K. Chambers, M. R. Chun, J. N. Heasley, K.-W. Hodapp, B. Hunt, R. Jedicke, D. Jewitt, R. Kudritzki, G. A. Luppino, M. Maberry, E. Magnier, D. G. Monet, P. M. Onaka, A. J. Pickles, P. H. H. Rhoads, T. Simon, A. Szalay, I. Szapudi, D. J. Tholen, J. L. Tonry, M. Waterson, and J. Wick. Pan-STARRS: A Large Synoptic Survey Telescope Array. In J. A. Tyson and S. Wolff, editors, *Survey and Other Telescope Technologies and Discoveries*, volume 4836 of , pages 154–164, December 2002.

Jeffrey P Kantor. Managing the evolution of the lsst data management system. In *Advanced Software and Control for Astronomy*, volume 6274, page 62740P. International Society for Optics and Photonics, 2006.

MG Kendall and A Stuart. The advanced theory of statistics, charles griffin & co. *Ltd.(London)*, 83:62013, 1977.

David G Koch, William J Borucki, Gibor Basri, Natalie M Batalha, Timothy M Brown, Douglas Caldwell, Jørgen Christensen-Dalsgaard, William D Cochran, Edna DeVore, Edward W Dunham, et al. Kepler mission design, realized photometric performance, and early science. *The Astrophysical Journal Letters*, 713(2):L79, 2010.

KII Koljonen, DM Russell, JM Corral-Santana, Montserrat Armas Padilla, Teo Muñoz-Darias, Fraser Lewis, M Coriat, and FE Bauer. A âĂŸhigh-hardâĂŹoutburst of the black hole x-ray binary gs 1354- 64. *Monthly Notices of the Royal Astronomical Society*, 460(1):942–955, 2016.

CH Kriebel et al. Evaluating the quality of information systems. *design and implementation of computer based information systems*, pages 29–43, 1979.

Akhil Kumar and Arie Segev. Cost and availability tradeoffs in replicated data concurrency control. 1988.

Gilbert Laporte and Yves Nobert. Generalized travelling salesman problem through n sets of nodes: An integer programming approach. *INFOR: Information Systems and Operational Research*, 21(1):61–75, 1983.

Inger Kristin Larsen, Milada Småstuen, Tom Børge Johannesen, Frøydis Langmark, Donald Maxwell Parkin, Freddie Bray, and Bjørn Møller. Data quality at the cancer

registry of norway: an overview of comparability, completeness, validity and timeliness. *European journal of cancer*, 45(7):1218–1231, 2009.

Jean-Pierre Lasota. The disc instability model of dwarf novae and low-mass x-ray binary transients. *New Astronomy Reviews*, 45(7):449–508, 2001.

Nicholas M Law, Shrinivas R Kulkarni, Richard G Dekany, Eran O Ofek, Robert M Quimby, Peter E Nugent, Jason Surace, Carl C Grillmair, Joshua S Bloom, Mansi M Kasliwal, et al. The palomar transient factory: system overview, performance, and first results. *Publications of the Astronomical Society of the Pacific*, 121(886):1395, 2009.

QZ Liu, J Van Paradijs, and EPJ Van Den Heuvel. A catalogue of low-mass x-ray binaries in the galaxy, lmc, and smc. *Astronomy & Astrophysics*, 469(2):807–810, 2007.

Yingbo Liu, Feng Wang, Kaifan Ji, Hui Deng, Wei Dai, and Bo Liang. Nvst data archiving system based on fastbit nosql database. *arXiv preprint arXiv:1612.07587*, 2016.

Zhong Liu, Jun Xu, Bo-Zhong Gu, Sen Wang, Jian-Qi You, Long-Xiang Shen, Ru-Wei Lu, Zhen-Yu Jin, Lin-Fei Chen, Ke Lou, et al. New vacuum solar telescope and observations with high resolution. *Research in Astronomy and Astrophysics*, 14(6): 705, 2014.

Nicholas R Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and space science*, 39(2):447–462, 1976.

Kevin Loney. *Oracle database 10g: the complete reference*. McGraw-Hill/Osborne London, 2004.

Michael B Lund, Joshua Pepper, and Keivan G Stassun. Transiting planets with lsst. i. potential for lsst exoplanet detection. *The Astronomical Journal*, 149(1):16, 2014.

Michael B Lund, Joshua A Pepper, Avi Shporer, and Keivan G Stassun. Transiting planets with lsst iv: Detecting planets around white dwarfs. *arXiv preprint arXiv:1809.10900*, 2018.

Phil Marshall, Timo Anguita, Federica B Bianco, Eric C Bellm, Niel Brandt, Will Clarkson, Andy Connolly, Eric Gawiser, Zeljko Ivezic, Lynne Jones, et al. Science-driven optimization of the lsst observing strategy. *arXiv preprint arXiv:1708.04058*, 2017.

Paul J McMillan. Mass models of the milky way. *Monthly Notices of the Royal Astronomical Society*, 414(3):2446–2457, 2011.

Timothy McPhillips, Tianhong Song, Tyler Kolisnik, Steve Aulenbach, Khalid Belhajjame, Kyle Bocinsky, Yang Cao, Fernando Chirigati, Saumen Dey, Juliana Freire, et al. Yesworkflow: a user-oriented, language-independent tool for recovering workflow information from scripts. *arXiv preprint arXiv:1502.02403*, 2015.

S Mineshige, M Takeuchi, and H Nishimori. Is a black hole accretion disk in a self-organized critical state? *The Astrophysical Journal*, 435:L125–L128, 1994.

Paolo Missier, Stian Soiland-Reyes, Stuart Owen, Wei Tan, Alexandra Nenadic, Ian Dunlop, Alan Williams, Tom Oinn, and Carole Goble. Taverna, reloaded. In *International conference on scientific and statistical database management*, pages 471–481. Springer, 2010.

Bruce Momjian. *PostgreSQL: introduction and concepts*, volume 192. Addison-Wesley New York, 2001.

Benjamin T Montet, Guadalupe Tovar, and Daniel Foreman-Mackey. Long-term photometric variability in kepler full-frame images: Magnetic cycles of sun–like stars. *The Astrophysical Journal*, 851(2):116, 2017.

Luc Moreau, Belfrit Batlajery, Trung Dong Huynh, Danius Michaelides, and Heather Packer. A templating system to generate provenance. *IEEE Transactions on Software Engineering*, 2017.

Leonardo Murta, Vanessa Braganholo, Fernando Chirigati, David Koop, and Juliana Freire. noworkflow: capturing and analyzing provenance of scripts. In *International Provenance and Annotation Workshop*, pages 71–83. Springer, 2014.

AB MySQL. Mysql, 2001.

Charles E Noon and James C Bean. A lagrangian based approach for the asymmetric generalized traveling salesman problem. *Operations Research*, 39(4):623–632, 1991.

François Ochsenbein, Patricia Bauer, and James Marcout. The vizier database of astronomical catalogues. *Astronomy and Astrophysics Supplement Series*, 143(1):23–32, 2000.

Mike Owens. *The definitive guide to SQLite*. Apress, 2006.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Sebastien Picaud and AC Robin. 3d outer bulge structure from near infrared star counts. *Astronomy & Astrophysics*, 428(3):891–903, 2004.

Camelia-M Pintea, Petrica C Pop, and Camelia Chira. The generalized traveling salesman problem solved with ant algorithms. *Journal of Universal Computer Science*, 13 (7):1065–1075, 2007.

G Pojmanski. The all sky automated survey. *arXiv preprint astro-ph/9712146*, 1997.

Jacques Renaud and Fayez F Boctor. An efficient composite heuristic for the symmetric generalized traveling salesman problem. *European Journal of Operational Research*, 108(3):571–584, 1998.

Serena Repetto, Andrei P Igoshev, and Gijs Nelemans. The galactic distribution of x-ray binaries and its implications for compact object formation and natal kicks. *Monthly Notices of the Royal Astronomical Society*, 467(1):298–310, 2017.

George R Ricker, Joshua N Winn, Roland Vanderspek, David W Latham, et al. Transiting exoplanet survey satellite. *Journal of Astronomical Telescopes, Instruments, and Systems*, 1(1):014003, 2014.

RE Ryan Jr. igalfit: an interactive tool for galfit. *arXiv preprint arXiv:1110.1090*, 2011.

Carlos Sáenz-Adán, Beatriz Pérez, Trung Dong Huynh, and Luc Moreau. UML2PROV: Automating provenance capture in software engineering. In *International Conference on Current Trends in Theory and Practice of Informatics*, pages 667–681. Springer, 2018.

Jignesh N Sarvaiya, Suprava Patnaik, and Salman Bombaywala. Image registration by template matching using normalized cross correlation. In *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*, pages 819–822. IEEE, 2009.

Jeffrey D Scargle. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835–853, 1982.

Edward F Schlafly and Douglas P Finkbeiner. Measuring reddening with sloan digital sky survey stellar spectra and recalibrating sfd. *The Astrophysical Journal*, 737(2): 103, 2011.

David J Schlegel, Douglas P Finkbeiner, and Marc Davis. Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *The Astrophysical Journal*, 500(2):525, 1998.

James Anthony Seibert, John M Boone, and Karen K Lindfors. Flat-field correction technique for digital detectors. In *Medical Imaging 1998: Physics of Medical Imaging*, volume 3336, pages 348–354. International Society for Optics and Photonics, 1998.

John Silberholz and Bruce Golden. The generalized traveling salesman problem: A new genetic algorithm approach. In *Extending the horizons: advances in computing, optimization, and decision technologies*, pages 165–181. Springer, 2007.

L Stella, W Priedhorsky, and NE White. The discovery of a 685 second orbital period from the x-ray source 4u 1820-30 in the globular cluster ngc 6624. *The Astrophysical Journal*, 312:L17–L21, 1987.

Peter B Stetson. Daophot: A computer program for crowded-field stellar photometry. *Publications of the Astronomical Society of the Pacific*, 99(613):191, 1987.

Chris Stoughton, Robert H Lupton, Mariangela Bernardi, Michael R Blanton, Scott Burles, Francisco J Castander, AJ Connolly, Daniel J Eisenstein, Joshua A Frieman, GS Hennessy, et al. Sloan digital sky survey: early data release. *The Astronomical Journal*, 123(1):485, 2002.

Jay Strader, Elias Aydi, Christopher Britt, Adam Burgasser, Laura Chomiuk, Will Clarkson, Brian D Fields, Poshak Gandhi, Leo Girardi, John Gizis, et al. The plane's the thing: The case for wide-fast-deep coverage of the galactic plane and bulge. *arXiv preprint arXiv:1811.12433*, 2018.

Peter W Sullivan, Joshua N Winn, Zachory K Berta-Thompson, David Charbonneau, Drake Deming, Courtney D Dressing, David W Latham, Alan M Levine, Peter R McCullough, Timothy Morton, et al. The transiting exoplanet survey satellite: simulations of planet detections and astrophysical false positives. *The Astrophysical Journal*, 809(1):77, 2015.

T. M. Tauris and E. P. J. van den Heuvel. *Formation and evolution of compact stellar X-ray sources*, page 623âĂŞ666. Cambridge Astrophysics. Cambridge University Press, 2006.

The Lynx Team. The lynx mission concept study interim report. *arXiv preprint arXiv:1809.09642*, 2018.

J Timmer and M König. On generating power law noise. *Astronomy and Astrophysics*, 300:707, 1995.

Doug Tody. The iraf data reduction and analysis system. *Instrumentation in astronomy VI*, 627:733–748, 1986.

Austin B Tomaney and Arlin PS Crotts. Expanding the realm of microlensing surveys with difference image photometry. *arXiv preprint astro-ph/9610066*, 1996.

J Anthony Tyson. Large synoptic survey telescope: overview. In *Survey and Other Telescope Technologies and Discoveries*, volume 4836, pages 10–21. International Society for Optics and Photonics, 2002.

J Van Paradijs and N White. The galactic distribution of low-mass x-ray binaries. *The Astrophysical Journal Letters*, 447(1):L33, 1995.

J. T. VanderPlas and Ž. Ivezić. Periodograms for Multiband Astronomical Time Series. , 812:18, October 2015.

Jacob T VanderPlas and Željko Ivezic. Periodograms for multiband astronomical time series. *The Astrophysical Journal*, 812(1):18, 2015.

Yair Wand and Richard Y Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95, 1996.

ChangYing Wang, Min Lin, YiWen Zhong, and Hui Zhang. Solving travelling salesman problem using multiagent simulated annealing algorithm with instance-based sampling. *International Journal of Computing Science and Mathematics*, 6(4):336–353, 2015.

Yng-Yuh Richard Wang, Veda Storey, and Chris Firth. *Data quality research: a framework, survey, and analysis.* Total Data Quality Management Research Program, Sloan School of Management âĂę, 1993.

MC Weisskopf, B Brinkman, C Canizares, G Garmire, S Murray, and LP Van Speybroeck. An overview of the performance and scientific results from the chandra x-ray observatory. *Publications of the Astronomical Society of the Pacific*, 114(791):1, 2002.

Mark Wells, Andrej PrsÌĘa, Lynne Jones, and Peter Yoachim. Initial estimates on the performance of the lsst on the detection of eclipsing binaries. *Publications of the Astronomical Society of the Pacific*, 129(976):065003, 2017.

Marc Wenger, François Ochsenbein, Daniel Egret, Pascal Dubois, François Bonnarel, Suzanne Borde, Françoise Genova, Gérard Jasniewicz, Suzanne Laloë, Soizick Lesteven, et al. The simbad astronomical database-the cds reference database for astronomical objects. *Astronomy and Astrophysics Supplement Series*, 143(1):9–22, 2000.

Ralf Widenhorn, Armin Rest, Morley M Blouke, Richard L Berry, and Erik Bodegom. Computation of dark frames in digital imagers. In *Sensors, Cameras, and Systems for Scientific/Industrial Applications VIII*, volume 6501, page 650103. International Society for Optics and Photonics, 2007.

WANG Xin. Design and implementation of cneost image database based on nosql system. *Chinese Astronomy and Astrophysics*, 38(2):211–221, 2014.

Donald G York, J Adelman, John E Anderson Jr, Scott F Anderson, James Annis, Neta A Bahcall, JA Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579, 2000.

C Zurita, J Casares, and T Shahbaz. Evidence for optical flares in quiescent soft x-ray transients. *The Astrophysical Journal*, 582(1):369, 2003.

Martin A Zwaan, Martin J Meyer, Rachel L Webster, Lister Staveley-Smith, Michael J Drinkwater, David G Barnes, Raghir Bhathal, WJG De Blok, Michael J Disney, Ron D Ekers, et al. The hipass catalogue–ii. completeness, reliability and parameter accuracy. *Monthly Notices of the Royal Astronomical Society*, 350(4):1210–1219, 2004.