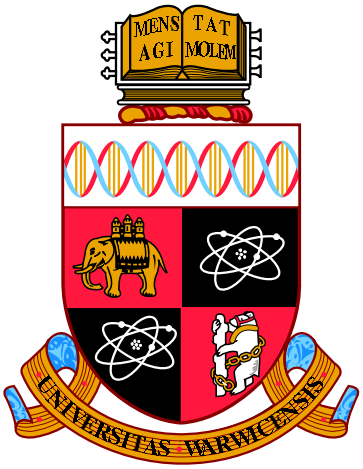

Tracing Exoplanets through Time with *TESS*

by

Matthew P. Battley

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Physics

May 2022

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

First and foremost I would like to thank my supervisor, Professor Don Pollacco, for his unerring support and guidance, valuable scientific discussion, and for his understanding when the stresses of the ongoing Coronavirus pandemic occasionally proved overwhelming. Your wide view of the exoplanet field and penchant for finding new scientific ideas (as discussed in many slightly off-topic chats during our meetings) has given me an important perspective on the biggest challenges to be solved in the field, and guided me to become a much better scientist than I was when I first contacted you from the other side of the world. I really appreciated both the freedom you gave me to pursue my own research direction and the insight you gave me on when to stop researching and start publishing something. A particularly big thanks also for helping me escape France half an hour before it closed its borders after an ill-timed observing trip.

Thank you to the Astronomy and Astrophysics group at Warwick for making me feel so welcome on the other side of the world from my home, and for the rich and exciting research environment fostered by all. In particular I'd like to thank Dr David Armstrong and Dr Dan Bayliss for your support (both scientific and personal), as well as the full Exoplanet Group for valuable scientific discussion and feedback in our weekly meetings. Thank you to all of the PhD students past and present (far too many to list here these days!) who have made my PhD so entertaining and for all of your wonderfully absurd conversations both in the Cave and on our ever-expanding Discord server.

As always, thank you Mum and Dad for your love and support both before and during this mammoth task. While I certainly didn't sign up for such a long

time without seeing you in person (thanks Corona), our weekly virtual catch-ups are always a highlight and I'm very pleased that you've kept me up to date with lots of Indy pictures. A big thanks to my brother Andrew for many a late-night chat (for one or other of us) about our usual topics of Lego and what media we should be distracting ourselves with this week.

I'd like to thank all of my friends on both sides of the world for keeping me sane through good times and bad, in particular Max (extra kudos for helping me proof-read...), Bobbi, Natasha, Theo, Hayley and EJ from NZ, and Sam, Ella, Alan, Josha, Ailish, Marlene, Oscar, Charlotte and Dan in the UK. At a wider scale, a massive thank you to everyone Warwick Mountains for giving me the chance to escape to the hills, embrace my inner mountain goat and stay up far too late at the Vs.

Finally, thank you to Sally. The last two years of this PhD have been two of the hardest of my life, but also two of the most wonderful thanks to meeting you. Thank you for your endless love and support, and for joining me on our adventures past, present and future.

I am grateful for the financial support supplied by the University of Warwick through the Chancellor's International Scholarship in order for me to complete this research.

Declarations

I declare that the work included in this thesis is my own work unless stated otherwise and that it has not been submitted to this or any other academic institution for any other degree or qualification. Chapters 4 to 6 are partial reproductions of my published first-author papers Battley et al. [2020], Battley et al. [2021] and Battley et al. [2022]. Sections in Chapters 1 and 2 also include paragraphs adapted from these works. Some paragraphs in Chapters 1 through 3 also draw on material submitted as part of the PhD 6 month report in March 2019. The only work completed by collaborators is presented in Sections 3.3, 2.8.2 and 6.4.3, the first of which is a large list of young stars built with Dr Ed Gillen and the latter two detailing a TTV analysis which was completed by Dr Michelle Kunimoto, a co-author to Battley et al. [2021]. Exact contributions from collaborators are discussed in the separate declarations at the beginning of each chapter. Work presented in Chapters 3 to 6 has also been presented at a public talks and posters as follows:

Oral Presentations

- **"Revisiting the Kepler Field with TESS"**, TESS Science Conference 2, MIT, USA (2021)
- **"Tracking Exoplanets through Time with TESS"**, UK Exoplanet Community Meeting, University of Birmingham, UK (2021)
- **"The Curious Case of Young Exoplanets: Unlocking the secrets of young exoplanets with multi-instrument observations"**, StScI Summer Symposium, STScI, USA (2021)

- **"Fantastic Planets and How to Find Them"**, ACORNS, University of Warwick, UK (2019)
- **"The Curious Case of Exoplanet Teenagers"**, Warwick AstroSoc Research Symposium, University of Warwick (2019)

Poster Presentations

- **"The YOUNGSTER Program: Extending the search for young transiting exoplanets"**, *Circumstellar Disks and Young Planets*, 2021 Sagan Exoplanet Summer Workshop, California Institute of Technology, USA (2021)
- **"A search for Young Exoplanets in the S1-5 TESS FFIs"**, Exoplanets III, University of Heidelberg, Germany (2020)
- **"Searching for Young Exoplanets with TESS"**, *Extreme Precision Radial Velocity*, 2020 Sagan Exoplanet Summer Workshop, California Institute of Technology, USA (2020)
- **"Searching for Young Planets with TESS"**, PLATO Science Meeting 2019, University of Warwick, UK (2019)

Abstract

In the thirty years since the discovery of the first exoplanet, over 5000 verified exoplanets have been discovered, unveiling a rich array of different exoplanetary architectures. However, there are still many unanswered questions regarding the formation and evolution pathways which have led to the observed population. To understand these processes it is imperative to trace exoplanets across time, both over galactic and human timescales. This thesis presents work in both of these areas, using data from the *Transiting Exoplanet Survey Satellite* (*TESS*).

The bulk of this thesis focuses on the challenge of discovering new young exoplanets (age <1 Gyr) and understanding the variability of their potential host stars. This begins with building an extended population of young stars around which to search for exoplanets, illustrating the kinematic power of the recently launched *Gaia* satellite and resulting in a target list of over three million young stars.

A dedicated new young star detrending pipeline is then presented, which is in turn used to search for new young exoplanets in stellar associations within *TESS* sectors 1-5. Although no new exoplanets were found, the pipeline's effectiveness is demonstrated by recovering the previously known young exoplanets DS Tuc Ab and AU Mic b, alongside all other 2 min Targets of Interest (TOIs) from the 30 min cadence data alone.

The completed young exoplanet search highlighted the challenging diversity of young stellar variability. To understand this variability, Kohonen Self-Organising Maps are used for the first time on a dedicated sample of young stars observed in the first year of *TESS*'s primary mission, in order to sort light-curves by topology and look for distinct variability classes. This analysis forms the first step in the YOUNGSTER programme, aiming to use knowledge of young star variability to inform more targeted detrending in future young exoplanet searches.

Finally, this thesis presents work on tracing exoplanets through time on *human* timescales, updating and improving the ephemerides for all previously known Kepler planets (22) and candidates (4) which were observed with sufficient signal to noise in *TESS*. It also explores any transit-timing variations seen for these objects, including intriguing results for HAT-P-7b, Kepler-411d, K00075.01 and K00076.01.

Abbreviations

BANYAN Bayesian Analysis for Nearby Young AssociatioNs

BLS Box Least Squared

BMU Best Matching Unit

CDIPS Cluster Difference Imaging Photometric Survey

CHEOPS CHaracterising ExOPlanet Satellite

CAMD Colour Absolute Magnitude Diagram

DIA Difference Imaging Analysis

EA Algol-type eclipsing binary

EB Beta-Lyrae type eclipsing binary

FFI Full Frame Image

GPI Gemini Planet Imager

JWST James Webb Space Telescope

MCMC Markov-Chain Monte Carlo

NGTS Next Generation Transit Survey

PDC Presearch Data Conditioning

PLATO PLAnetary Transits and Oscillations of stars

PSF Point Spread Function

QLP Quick-look Pipeline

RV Radial Velocity

SAP Simple Aperture Photometry

SOM Self-Organising Map

SPHERE Spectro-Polarimetric High contrast imager for Exoplanets REsearch

SPOC Science Processing Operation Center

TESS Transiting Exoplanet Survey Satellite

TIC TESS Input Catalog

TTV Transit Timing Variation

UVW Velocities in the Galactic coordinate system

XYZ Positions/directions in the Galactic coordinate system

Chapter 1

Introduction

"Everything starts somewhere,
though many physicists disagree"

— Terry Pratchett, *Hogfather*

The concept of other planets has fascinated humanity for millennia, with Epicurus (341-270 BC) stating that "there are infinite worlds both like and unlike this world of ours" over 2000 years ago. By the 20th century the idea had become so ubiquitous that it routinely featured in some of the most famous popular culture (especially in Science Fiction), from the desert planet Arrakis in Frank Herbert's *Dune* (1965), to the many worlds across the Galaxy in the original *Star Wars* Trilogy (1977-1983). Curiously however, the first definitive evidence that these worlds existed was only found at the very end of the 20th Century, with the discovery of the first 'exoplanets', or planets around stars other than the Sun, found in the 1990s. Since then exoplanets have become a key tool to explore our place in the Universe, including our physical place and our philosophical place: "Are we alone in the Universe?" At its widest, exoplanetary science aims to provide more clarity to both aspects of this question; while the discovery and characterisation of new exoplanets allows us to make sense of our physical place in the Universe, the continuing search on these planets for extraterrestrial life poses profound philosophical questions whether we find life or not. This introduction gives the background necessary to understand the methods, analysis and discussion underlying this research, from the history and current state of exoplanets to the challenge of 'young' exoplanet discovery. Background is also given for the main two sources of data used in this research (Section 1.5) and the field of machine learning in the context of stellar variability analyses (Section 1.6).

1.1 Exoplanet History

The first confirmed exoplanets to be discovered were PSR B1257+12 b & c, two super-Earth sized planets discovered via unusual timing of their host pulsar [Wolszczan and Frail, 1992]. This discovery was swiftly followed by the discovery of 51 Pegasi b [Mayor and Queloz, 1995], the first planet around a sun-like star. Although initially controversial, 51 Pegasi b's eventual confirmation represented the culmination of a steady increase in the instrumental capability of the radial velocity technique (measuring the reflex motion of its host star) which had been developing in the previous decades (e.g. Baranne et al. [1979]; Campbell et al. [1988]). For this discovery Mayor & Queloz were awarded a share of the Nobel Prize in Physics in 2019. The discovery of 51 Pegasi b changed the field of exoplanets forever, not only because it revealed an exoplanet so unlike those in our own solar system (a 'Hot-Jupiter'; a Jupiter-sized planet within the orbit of Mercury), but also because it showed that that exoplanet science was within reach with current technology.

The first transiting planet was found around the star HD 209458 in 1999 by two independent teams [Charbonneau et al., 2000; Henry et al., 1999], after radial velocity measurements suggested the presence of a planet. This same planet later became the first known to have its own atmosphere, using measurements from the Hubble Space Telescope [Charbonneau et al., 2002]. The discovery and abundance of these first hot Jupiters paved the way for more extensive studies of exoplanet statistics, as these types of planets are easier and cheaper to detect via the transit method than the more traditional radial velocity method [Haswell, 2010]. However, the faintness of many planet candidates detected via the transit method challenges spectroscopic confirmation and follow-up.

Following these earliest discoveries, additional discovery methods were found to be sensitive to other important exoplanets. The first microlensing planet OGLE 2003BLG235/MOA 2003BLG53 b was observed on July 2003 and later published by Bond et al. [2004]. Shortly thereafter the first directly imaged exoplanet 2M1207 b was found around a brown dwarf [Chauvin et al., 2004]. The first (and so far only) sub-stellar astrometric companion was found in 2013 around DENIS-P J082303.1-491201 [Sahlmann et al., 2013]. Because these different techniques are all sensitive to slightly different regions of exoplanetary parameter space, using them in parallel allows the community to explore the huge diversity of exoplanets.

A key step forward in the field of exoplanetary science was the beginning of large, wide-coverage (or 'all-sky') transit surveys, which surveyed far more stars than previous more directed searches. Particularly prolific ground-based photo-

metric surveys have included WASP [Pollacco et al., 2006], the Next Generation Transit Survey (NGTS) [Wheatley et al., 2018] and HAT/HATNet [Bakos et al., 2004, 2008]. These ground-based surveys discovered hundreds of exoplanets (and continue to discover more), paving the way for more extensive analysis of exoplanetary demographics and providing the proof of concept for larger space-based transit searches. Such detections were/are backed up by precision spectrographs such as ELODIE [Baranne et al., 1996], CORALIE [Queloz et al., 2000] and HARPS [Mayor et al., 2003], allowing indirect mass-measurements and confirmation of thousands of fascinating new exoplanetary systems. It should be noted however that given the discrepancy between data availability (and required brightness) for photometric and spectroscopic studies, many 'confirmed' transiting exoplanets do not yet have measured masses, and have instead had to be validated through statistical means [Morton et al., 2016].[1] Indeed, only 896 of the known 3759 transiting planets currently have measured planetary masses.[2]

The move to Space Telescopes proved to be revolutionary for exoplanet science. The CoRoT satellite [Barge et al., 2008; Auvergne et al., 2009], flown between 2007 and 2012, was the first space mission dedicated to exoplanet science, resulting in 30 confirmed exoplanet discoveries[2]. However, the true revolution began with the *Kepler* Space Telescope [Borucki et al., 2010; Koch et al., 2010], which surveyed an 100 square degree region of the sky in the vicinity of the Cygnus and Lyra constellations from 2009 to 2013. This was followed by the ecliptic-based K2 mission between 2013-2018 after the failure of two of the spacecraft's reaction wheels [Howell et al., 2014]. The *Kepler* Space Telescope verified almost 3000 exoplanets over the course of the main *Kepler* and secondary *K2* missions, with thousands more still undergoing follow-up. The extensive exoplanet population revealed by *Kepler* completely changed our view of exoplanetary science, revealing that exoplanets were far more common and in far more unusual configurations than originally thought (e.g. Borucki et al. [2010]; Howell et al. [2014]; Batalha et al. [2013]; Buchhave et al. [2012]; Lee [2018]; Thompson et al. [2018]). Unfortunately, with the *Kepler* satellite running out of fuel at the end of 2018, NASA officially retired the *Kepler* space telescope on 30 October 2018.[3] The next chapter of transiting exoplanet science is

[1] n.b. statistical validation typically consists of ruling out all other 'false-positive' signals: astrophysical/systematic signals which could mimic an exoplanet transit. Tests include visual or machine-learning based inspection to rule out obvious stellar activity, telescope systematics or eclipsing binary stars, high contrast imaging of the surrounding area and spectroscopic follow-up to rule out background/spectroscopic binaries. Typical automatic validation pipelines require a false positive probability of <1% before candidates can be validated as planets [Morton, 2012].

[2] NASA Exoplanet Archive - https://exoplanetarchive.ipac.caltech.edu; 15 Feb 2022

[3] https://www.nasa.gov/press-release/nasa-retires-kepler-space-telescope-passes-planet-hunting-torch

now beginning, following the launch of the Transiting Exoplanet Survey Satellite (*TESS*) in 2018 [Ricker et al., 2014]. The *TESS* mission is already extending *Kepler's* vision as it performs a nearly all-sky survey looking for transiting exoplanets around nearby bright stars. Given the importance of *TESS* as one of the primary data sources in this research it will be examined in more detail below.

Now, 30 years after the initial exoplanet discovery, over 5000 exoplanets have been validated, a dramatic increase largely driven by radial velocity searches until 2013 and dominated by transit surveys ever since. These two main discovery techniques will be discussed in detail next (alongside brief introductions to the other main discovery methods), before an overview of current exoplanet knowledge in Section 1.3.

1.2 Exoplanet Discovery Methods

A wide variety of methods have been used to detect exoplanets, typically based on variations in the brightness, frequency or timing of light arriving from planetary hosts. Here the two most important methods used and referred to in this thesis are discussed in depth (the Transit and Radial Velocity Methods), with brief overviews given for additional methods in Section 1.2.3. Both the transit and radial velocity methods are indirect methods of exoplanet detection, relying on measuring the effect potential exoplanets have on their host stars.

1.2.1 Transit Method

The most successful method of detecting exoplanets thus far is the transit method, which relies on measuring the characteristic dip in brightness of a star when an orbiting planet passes in front of it. This is the only method that allows for direct determination of a planet's radius, a crucial parameter for the characterisation of exoplanets.

As shown in Figure 1.1, when a planet passes in front of a star it causes some of the star's light to be blocked. This results in a characteristic transit shape from which important planetary parameters can be determined. At its simplest (assuming no excess flux from the planet itself and a stellar disk of uniform brightness) this change in flux (F) can be interpreted to uncover the planetary radius (R_{P}) as a function of stellar radius (R_*) using the relation

$$\frac{\Delta F}{F} = \frac{R_{\mathrm{P}}^2}{R_*^2} \tag{1.1}$$

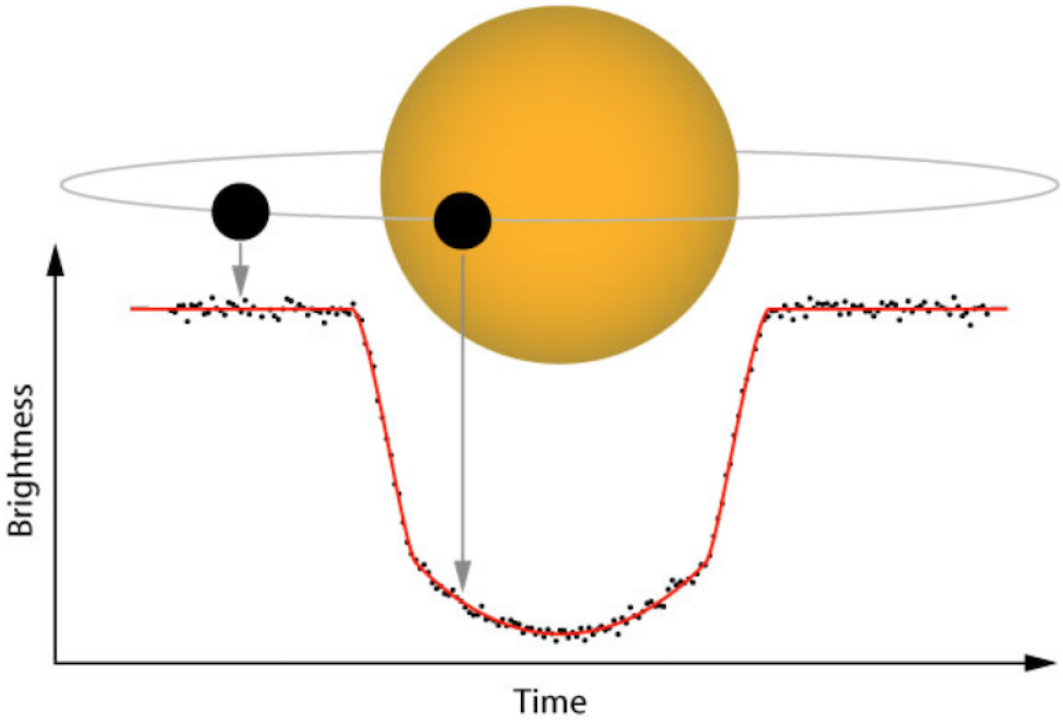

Figure 1.1: Schematic of the transit method for discovering exoplanets. As a planet passes in front of its host star is causes a characteristic drop in brightness which can tell us information about the planet's radius, period and inclination. Reproduced with permission from NASA - Source: https://heasarc.gsfc.nasa.gov/docs/tess/primary-science.html

Note however that determining the true radius of the planet relies on knowledge of the stellar radius, which must be derived from stellar models. This means that while the derivation of the flux ratio is model independent, the planet radius is not.

By observing the star over an extended period and analysing the constructed light-curve for this star, one can also determine such parameters as the period (P), semi-major axis (a) and inclination (i) of the planetary orbit. The period is typically obtained through constructing a Box-Least Squared periodogram [Kovács et al., 2002] generated from the original (unfolded) transit light-curve, but is simply the average time between the mid points of two adjacent transits. It is important to use such a box-shaped period search compared to a generalised Lomb-Scargle periodogram [Lomb, 1976; Scargle, 1982] typically used for finding stellar variability periods, as the signal of a transiting planet looks far more like a repeated dip (or 'box') in the light-curve instead of a sinusoidal signal which is fitted in a Lomb-Scargle search. Once the planetary period is known, the semi-major axis of the

orbit, a, can easily be obtained from Kepler's third law:

$$\frac{a^3}{P^2} = \frac{G(M_* + M_{\rm P})}{4\pi^2} \tag{1.2}$$

Given that typically $M_* + M_{\rm P} \approx M_*$ and an estimate of the stellar mass can be gained from its spectral type, the semi-major axis is thus:

$$a \approx \left(GM_* \left(\frac{P}{2\pi} \right)^2 \right)^{1/3} \tag{1.3}$$

Similar to the planetary radius, this is reliant on a single model dependent quantity - the stellar mass, M_*.

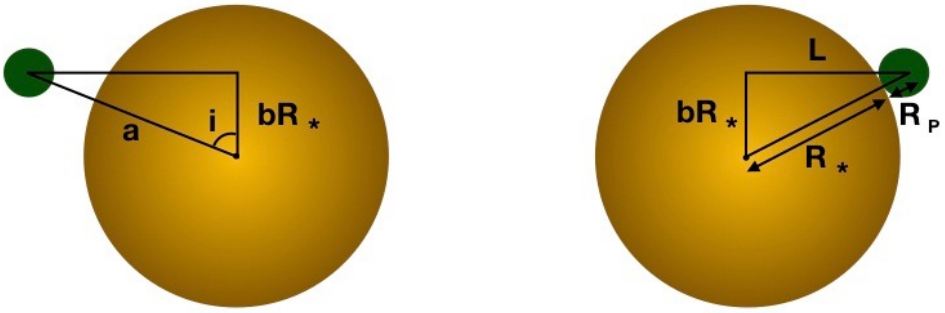

Figure 1.2: Geometry of the Transit Method. Here a denotes the planet's semi-major axis, i denotes the planet's inclination, b the impact parameter, L half the distance traversed by the planet across the disc of the star, and R_* and $R_{\rm P}$ the radii of the star and planet respectively.

Another key observation from the transit light-curve is the transit duration, $T_{\rm dur}$. The length of the transit duration is heavily dependent on the impact parameter, b, defined as:

$$b = \frac{a \cos i}{R_*} \tag{1.4}$$

This impact parameter is theoretically directly observable for a planet around a close enough star, but is often derived from the transit duration itself. Noting that the length of the transit duration is the total time the planet is even slightly in front of its star, the distance travelled in front of the star, 2L, can be defined with reference to Figure 1.2 as:

$$2L = 2\sqrt{(R_* + R_{\rm P})^2 - (bR_*)^2} \tag{1.5}$$

Finally, assuming that the planet is in a circular orbit (a similar analysis holds for eccentric planets, just with slightly more complex equations), and defining the angle traversed while transiting as $\alpha = 2 \arcsin \frac{L}{a}$, the transit duration is:

$$T_{\text{dur}} = P\frac{\alpha}{2\pi} = \frac{P}{\pi} \arcsin \frac{L}{a} = \frac{P}{\pi} \arcsin \frac{\sqrt{(R_* + R_{\text{P}})^2 - (bR_*)^2}}{a} \tag{1.6}$$

With knowledge of the planetary and stellar radii, period, and semi-major axis, one can easily obtain the impact parameter, and hence the orbital inclination. This parameter is crucial to determining an exoplanet's mass when coupled with radial velocity measurements, as discussed below.

The above equations show the power of the transit method for characterisation of exoplanets. What is particularly valuable about the transit method is that both radius and inclination are determined independent of any models of the exoplanet (only requiring knowledge of the stellar mass and radius), giving key measurements for these fundamental parameters. Nonetheless, the transit method is not without its caveats. Notably, using the transit method alone cannot give a mass for the exoplanet (unless notable transit timing variations are present - see below), so confirming that a signal is planetary in nature also requires follow-up radial velocity measurements [Lovis and Fischer, 2010]. Furthermore, the transit shape is typically distorted by limb-darkening of the star, since stars are not in fact of uniform brightness across their surfaces, and instead become increasingly darker and redder towards their edges [Haswell, 2010]. Additional care is needed to differentiate true planetary signals from astrophysical mimics such as blended eclipsing binaries, grazing eclipsing binaries and transits of planet-sized stars (such as brown or white dwarfs).

The probability of a planet transiting in the first place is also very dependent on the geometry of the system. It can be shown (e.g. Haswell [2010]) that assuming random orbit orientation the probability of a planet on a circular orbit transiting is

$$P_{\text{trans}} = \frac{R_* + R_{\text{P}}}{a} \approx \frac{R_*}{a} \tag{1.7}$$

This anti-correlation between transit probability and orbital semi-major axis, coupled with the fact that short-period planets will orbit more times over a given observation period (which is especially important for short-term observation windows such as the month-long observations of *TESS*) leads to the transit method being highly biased towards finding planets with close-in orbits. However, recent improvements to the transit method have included detecting longer-period planets

through individual 'monotransits' [Osborn et al., 2016] and unseen planets using transit timing variations [e.g. Leleu et al., 2021].

Key transiting exoplanet discovery instruments can be broken into two categories: ground-based and space-based. Ground-based observatories were crucial to confirming the effectiveness of the transit technique in early exoplanet discoveries, and continue to provide new exoplanet discoveries alongside important follow-up of candidates from larger space-based missions. First generation instruments include TReS [Alonso et al., 2004], HAT [Bakos et al., 2002], XO [McCullough et al., 2005], WASP [Pollacco et al., 2006], HAT-S [Bakos et al., 2008], KELT [Pepper et al., 2007] and MASCARA [Talens et al., 2017]. Of these, WASP (the Wide-Angle Search for Planets) was most prolific, detecting over 150 planets and paving the way for modern all-sky planet searches. More recent ground-based surveys such as MEarth [Nutzman and Charbonneau, 2008; Berta et al., 2012], TRAPPIST [Jehin et al., 2011] and NGTS [Wheatley et al., 2013] have built on these successes by implementing key improvements such as better removal of systematic noise, faster cadence observations, improved tracking and multi-telescope observations in order to chase smaller planets.

However, given the challenges presented by the atmosphere and limited time-window of observing transiting planets from the ground, in the 21st Century most transiting planet discovery has shifted to space. CoRoT, or the Convection Rotation et Transits mission [Baglin et al., 2006] was the first space-based mission designed to detect transiting planets, eventually finding 30 confirmed exoplanets over its six year mission, including the first transiting Super Earth CoRoT-7b [Léger et al., 2009]. This was followed by the paradigm-shifting *Kepler* satellite, harbouring both the initial *Kepler* [Borucki et al., 2010] and backup *K2* missions [Howell et al., 2014]. These missions resulted in over 3000 confirmed exoplanets between them[4], with further confirmation of candidates still ongoing. Further discussion of the *Kepler* mission and satellite can be found in 6.2.1. However, because many of the *Kepler/K2* candidates found were around stars from a relatively small area of sky and too dim for significant follow-up, the following *TESS* mission [Ricker et al., 2014] was designed to target brighter stars across almost the entire night sky (and is discussed further in Section 1.5.2). Several new and near-future exoplanet surveys will expand the discovery and characterisation of transiting exoplanets even further, including the James Webb Space Telescope (JWST - Gardner et al. [2006]) which will aid in-depth analysis of exoplanet atmospheres, CHEOPS [Broeg et al., 2013] which aims to precisely characterise the bulk composition of exoplanets and PLATO

[4]https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html

[Rauer et al., 2014] which will discover crucial new exoplanets, pushing down in size towards Earth twins. These missions will ensure that the transit method stays at the forefront of exoplanetary detection for many years to come.

Ephemerides and Transit Timing Variations

Exoplanet ephemerides are equations for calculating the expected transit time of each exoplanet in an exoplanetary system. Knowing the future transit times of an exoplanet is crucial to in-depth characterisation of such objects, as to fully constrain the mass, orbit, composition and atmosphere for an individual exoplanet one typically requires multiple follow-up measurements from several different instruments. Under normal circumstances, exoplanets should obey linear ephemerides, with the time of the next transit, T_N, following the simple equation

$$T_N = T_0 + NP \tag{1.8}$$

where T_0 is the initial transit time (typically the first recorded time of transit for a given transiting exoplanet, recorded from the centre of the transit[5]), N the number of orbits since the initial transit and P is the planet's orbital period. Hence knowing the time of an arbitrary transit is theoretically as simple as precisely knowing one transit time and the period of the exoplanet's orbit. However, because of the imperfect precision of these parameters and the multiplicative term in the ephemeris equation, the more time that passes since the last observed transit, the more imprecise the time of future transits becomes. For example, if the orbital period for a $P = 10$ day transiting exoplanet is only known to a precision of 0.001 days (see for example K2-167 from Mayo et al. [2018]), the 3σ error in its expected transit time will be over a day within just 10 years' time. This poses a serious problem for future characterisation, where instruments must juggle multiple important measurements in a single night, or make the most of their limited observing time. Indeed high-profile observatories such as JWST prefer timing precision of less than 30 min in order to maximise efficient follow-up. However because of the lead-time in the discovery and publication of new planets (and the vast array of previously know planets), it is often years between initial discovery and more in-depth follow-up. Hence to maintain up-to-date ephemerides for important exoplanets it is imperative to revisit known planets years after their discovery, ideally prioritising those most amenable to future follow-up.

[5]Note that in the case of asymmetric transits the true centre of the transit can be shifted, so obtaining T_0 needs to be completed more carefully, such as by fitting a well-defined transit model or by fitting the transit ephemeris equation to multiple transits and evaluating T_0 from there.

However, some planets have been found which do not appear to obey linear ephemerides. Instead, transits occur before and after the expected transit time, often evolving in time in a sinusoidal manner. This phenomenon is called Transit Timing Variations (TTVs), which are typically caused by changes in a planet's apparent period due to the gravitational pull of one or more other planets in the system [Holman and Murray, 2005; Lithwick et al., 2012]. The amplitude and period of these timing variations depend on the masses, periods, semi-major axes and eccentricities of the planetary orbits. For example, the presence of an inner planet in a multi-planet system effectively shifts the barycentre of the outer planet's orbit to the barycentre of inner-planet/star barycentre, which results in a new alignment where the outer planet transits slightly earlier or later than expected. For circular orbits, this results in TTV amplitudes of the outer planet (TTV_2) equal to

$$TTV_2 \approx \frac{P_2 a_1 \mu_1 \sin[2\pi(t - T_c)/P_1]}{2\pi a_2} \tag{1.9}$$

where subscripts 1 and 2 refer to the inner and outer planets respectively and $\mu_1 = M_1/M_{\text{Total}} \approx M_1/M_*$, the reduced mass of the inner planet. If the internal planet instead has an eccentric orbit, this effect can be bigger still. In either case, the magnitude of this effect scales with the mass of the perturbing (inner) planet, with more massive planets generating larger TTVs.

Meanwhile, if the inner planet is on a short circular orbit and the outer planet is on a wider, eccentric orbit, the gravitational perturbation of the outer planet causes a slight increase in the period of the orbit, with a magnitude that varies according to the instantaneous distance of the two planets from each other. The maximum amplitude of this TTV is given by the equation

$$TTV_1 \approx \mu_2 e_2 \left(\frac{a_1}{r_2}\right)^2 P_2 \tag{1.10}$$

where $\mu_2 \approx M_2/M_*$ and r_2 is the distance of the outer planet from the star. A full derivation of all of equations 1.9 and 1.10 are given in Haswell [2010]. Both of these above effects typically result in TTVs on the order of seconds.

If however the planets are in mean motion orbital resonance (where planetary periods are integer multiples of each other), much larger TTV amplitudes can be observed, on the order of minutes for orbital periods of a few days [Agol et al., 2005]. This occurs because when the resonant planets have successive conjuctions at the same section (or 'longitude') of the orbit, the strong gravitational interactions in these alignments increase the eccentricity of the lighter planet and change

its semi-major axis and period. This change in period also changes its longitude of conjunction, up to about 180^o from its initial longitude, at which point the eccentricity begins to decrease again, setting up a 'libration cycle'. Numerical simulations performed by Agol et al. [2005] show that a good fit for the maximum amplitude of this effect is

$$TTV_2 \approx \frac{P_2}{4.5j}\frac{M_1}{M_1 + M_2} \tag{1.11}$$

where j is the resonance order (where j+1 gives the resonance number; i.e. j = 1 implies a resonance of 2:1) and subscripts 1 and 2 represent the perturbing and transiting planets respectively. The corresponding period of this libration cycle is given by

$$P_{\text{lib}} \approx 0.5 j^{-4/3} \mu_2^{-2/3} P_2 \tag{1.12}$$

Fitting these TTVs allows for the masses of planets in multi-planet system to be found independently, giving confirmation of exoplanets without having to rely on typical radial velocity observations. This was first achieved in high-precision photometric data from the Kepler mission [Mazeh et al., 2013]. Careful modelling of such high precision TTVs can also reveal the presence of previously unseen or non-transiting exoplanets in the same system as known exoplanets, as was first discovered in the Kepler-19 system [Ballard et al., 2011]. Extended monitoring of known exoplanet systems can thus not only aid future characterisation methods, but also may lead to additional planetary discoveries. Furthermore, the increased precision of new and future instruments may allow the community t probe smaller TTV signals set up by other astrophysical bodies such as Exomoons [Kipping, 2009] or Exotrojans [Ford and Holman, 2007].

1.2.2 Radial Velocity

The radial velocity method is the second most prolific method of planetary detection and the best way to determine the mass of exoplanets to date. This method relies on detecting the Doppler shift of the spectral lines of a planet's host star as it moves towards and away from the observer due to its orbit around the barycentre of the star-planet system. This 'wobble' signature gives information about the masses and periods of any planets that exist around the host star.

In order to convert Doppler shifts of the star's spectral lines into a radial velocity (V_{RV}), one can simply use the Doppler equation

$$V_{RV} = c\frac{\lambda_{\text{obs}} - \lambda_{\text{rest}}}{\lambda_{\text{rest}}} \tag{1.13}$$

with the variation in this radial velocity over time for a star executing reflex motion due to a single planet following the form

$$V_{RV} = V_{RV,0} + K[\cos(v(t) + \omega_{OP}) + e\cos\omega_{OP}] \tag{1.14}$$

where $V_{RV,0}$ is the average radial velocity of the star with respect to Earth, K is the the radial velocity semi-amplitude, the 'true anomaly', $v(t)$, is the angular distance of the planet from its pericentre at any given time, ω_{OP} measures the orientation of the pericentre with respect to the intersection of the orbit with the plane of the observer and e denotes the planet's eccentricity. The true anomaly, $v(t)$, can be better understood by considering that the orbital path is described as an ellipse of the form

$$r(t) = \frac{a(1-e^2)}{1 + e\cos v(t)} \tag{1.15}$$

where a is the semi-major amplitude of the star from its barycentre and e is the planet's eccentricity. As the planet traverses this ellipse, the angle of the planet from its pericentre traces a path accroding to

$$\tan\frac{v(t)}{2} = \left(\frac{1+e}{1-e}\right)^{1/2}\tan\frac{E(t)}{2} \tag{1.16}$$

as shown by [Perryman, 2018], where $E(t)$ is the eccentric anomaly, which can be found by numerically solving Kepler's equation:

$$E(t) - e\sin E(t) = 2\pi(t - T_0)/P \tag{1.17}$$

Meanwhile the radial velocity semi-amplitude K is given by

$$K = \left(\frac{2\pi G}{P}\right)^{1/3}\frac{M_{\rm P}\sin i}{(M_{\rm P} + M_*)^{2/3}\sqrt{1-e^2}} \tag{1.18}$$

where P the planetary period, i is the planets' orbital inclination, e the orbital eccentricity, and $M_{\rm P}$ and M_* denote the planetary and stellar masses respectively. Given that P, e and a can be determined from the shape of the radial velocity curve itself, fitting the observed radial velocity measurements to the radial velocity equation and thus measuring the radial velocity amplitude can give a measurement for the planetary mass as a function of inclination and the stellar mass, i.e. $\frac{M_{\rm P}\sin i}{(M_{\rm P}+M_*)^{2/3}}$. If this is coupled with the inclination determined from the transit method and modelled stellar masses, one can in turn gain the true mass of the system. Typical magnitudes of these radial velocity amplitudes are on the order of 10s of m/s for

giant planets, <1m/s for super-Earths and Neptunes, and ≈10cm/s for Earth-like planets [Haswell, 2010].

The first exoplanet discovered via the radial velocity method was also the first exoplanet discovered around an sun-like star: 51 Pegasi b [Mayor and Queloz, 1995]. Since then over 900 exoplanets have been discovered via the radial velocity method,[6] and the method remains a crucial part of the confirmation of transiting exoplanet discoveries [Lovis and Fischer, 2010]

Key radial velocity instruments include CORAVEL [Baranne et al., 1979], ELODIE [Baranne et al., 1996], SOPHIE [Bouchy et al., 2009], CORALIE [Queloz et al., 2000] and HARPS [Mayor et al., 2003]. HARPS has achieved precision as low as 0.2-0.5m/s [Lovis and Fischer, 2010], putting the detection of super-Earths and Neptunes within reach. Recently built and future instruments intend to push this precision limit even further, with ESPRESSO capable of <10cm/s precision, and CODEX aiming as low as 1cm/s precision [Lovis and Fischer, 2010]. It is thus clear that like the transit method, the radial velocity method has planned technologies in place to probe increasingly small exoplanets as time progresses. However, measurements at such precision are also challenged by the inherent stellar activity of exoplanetary host stars, as discussed further in section 1.8.

1.2.3 Other methods

Astrometry

Similar to the radial velocity technique, the astrometric method of exoplanet discovery relies upon measuring the perturbation or 'wobble' of a host star caused by the presence of an orbiting planet [Perryman, 2018]. In the case of astrometry however, this is measured across the plane of the sky, rather than in the line of sight for radial velocity measurements. The path of the star across the sky follows an ellipse with a semi-major axis, a_s, of

$$a_s = \frac{M_{\rm P}}{M_* + M_{\rm P}} a \approx \frac{M_{\rm P}}{M_*} a \tag{1.19}$$

where a is the semi-major axis of the planetary orbit. It is this semi-major axis which is the observable *astrometric signature* of the planet in astrometric detection, defined as

$$\alpha = \left(\frac{M_{\rm P}}{M_*}\right)\left(\frac{a}{1au}\right)\left(\frac{d}{1pc}\right)^{-1} \text{arcsec}, \tag{1.20}$$

where d is the distance to the star in parsecs. As is clear from the equation, this technique is most sensitive to nearby, large mass planets in long-period orbits. How-

[6] https://exoplanetarchive.ipac.caltech.edu/index.html

ever, because of the distance term in this equation, astrometric detection is highly dependent on gaining high-precision distance measurements in order to discern the planetary signature. Indeed even for a star 50 pc from the Earth (e.g. in the relatively close-by Hyades star cluster), a Jupiter-analogue would have an astrometric signature of only $\alpha = 100$ μas, and an Earth-analogue only $\alpha = 0.06$ μas. Hence the detection of planets with this method requires an astrometric precision of less than a mas.

Because of these challenges, at the time of writing only one sub-stellar companion has been found using astrometry: a $28 M_{\mathrm{Jup}}$ object orbiting an L1.5 dwarf star [Sahlmann et al., 2013]. However, the extreme precision of the recently launched *Gaia* satellite (Gaia Collaboration et al. [2016b]; see Section 1.5.1) has the potential to revolutionise this detection method, with Perryman et al. [2014] estimating a yield of up to 70,000 1-15M_{Jup} planets will be found during the ten year *Gaia* mission. While the most recent data releases (e.g. [Gaia Collaboration et al., 2016a, 2018a]) have not yet been quite precise enough to chip away at this yield estimate, the upcoming Data Release 3 is expected to allow astrometric detection to begin in earnest.

Direct Imaging

Rather than relying on indirect signatures of a planet from observing the host star like the other detection methods discussed so far, Direct Imaging aims to image the planet itself, relying on either the reflected light from the planet or the planet's own thermal emission. Seeing the light from a planet itself is a considerable challenge, given the close angular separation of the planet from its star and the significant difference between the brightness of the star and planet. However, the discovery of planets through this method is crucial for discovering planets in very wide orbits (with periods too long to be reasonably detected via other methods) and for studying very young exoplanets as they are still forming.

The majority of the light from most directly imaged planets comes from reflected light. The ratio of planet to star brightness of a planet with semi-major axis α in reflected light of wavelength λ can be described as

$$\frac{f_p(\alpha, \lambda)}{f_*(\lambda)} = p(\lambda) \left(\frac{R_{\mathrm{P}}}{a} \right)^2 g(\alpha) \tag{1.21}$$

where $p(\lambda)$ is the geometric albedo of the planet and $g(\alpha)$ is a phase-dependent function [Perryman, 2018]. This ratio is typically exceedingly small, ranging from approximately 10^{-9} for a Jupiter-analogue to about 10^{-10} for an Earth analogue

when observed from as close as 10pc away. However, the contrast can be optimised by moving towards near- and mid-infrared wavelengths ($\sim$0.75-25μm), simultaneously reducing the observes brightness of the host star and increasing the brightness of thermal emission from the planet of interest. This allows contrast ratios to be improved to on the order of 10^{-4}. Further improvements can be gained by adding coronagraphs to block the main light from the star and by applying adaptive optics (AO) technology to contend with the Earth's atmosphere when observing from the ground.

Such improvements, linked to the development of key new observatories such as VLT/SPHERE [Beuzit et al., 2019] and the Gemini Planetary Imager [GPI; Macintosh et al., 2014] have now led to the discovery of 58 directly imaged exoplanets[7], including such key discoveries as a planet around the young debris disk β pic [Lagrange et al., 2009] and the four-planet system around HR 8799 [Marois et al., 2008]. These instruments both operate in the near-infrared region of the electromagnetic spectrum, with wavelengths on the order of 1μm. The recent launch of the James Webb Space Telescope (JWST, Gardner et al. [2006]) in December 2021 promises to expand these discoveries further, with its main instruments most sensitive to infrared light.

Microlensing

Microlensing is unusual as a detection method in that each discovered planet can only be observed once, compared to other techniques which benefit from repeated observations. Microlensing makes use of one of the effects of Einstein's theory of General Relativity, where massive objects distort space-time and cause electro-magnetic radiation (or light) to bend around the foreground massive object [Einstein, 1936]. Under the right circumstances this can result in multiple images of a distant object which are significantly amplified [Perryman, 2018]. As was originally shown by Mao and Paczyñski [1991], if the foreground object is a star with a planet around it (and that planet is also in the primary lensing plane), one can resolve a distinct sharp peak in the resulting light-curve when monitored over the course of the time-varying alignment change of the background object, lens and observer. Such a light-curve in shown here in Figure 1.3, showing the discovery of the first planet found via microlensing, OGLE 2003BLG235/MOA 2003BLG53 b [Bond et al., 2004].

Following Bond et al. [2004]'s initial planetary microlensing discovery, there have now been over 120 microlensing discoveries,[6] mostly from the MOA [Bond et al., 2001], OGLE [Udalski, 2003] and KMTNet [Kim et al., 2016] observatories.

[7] https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html

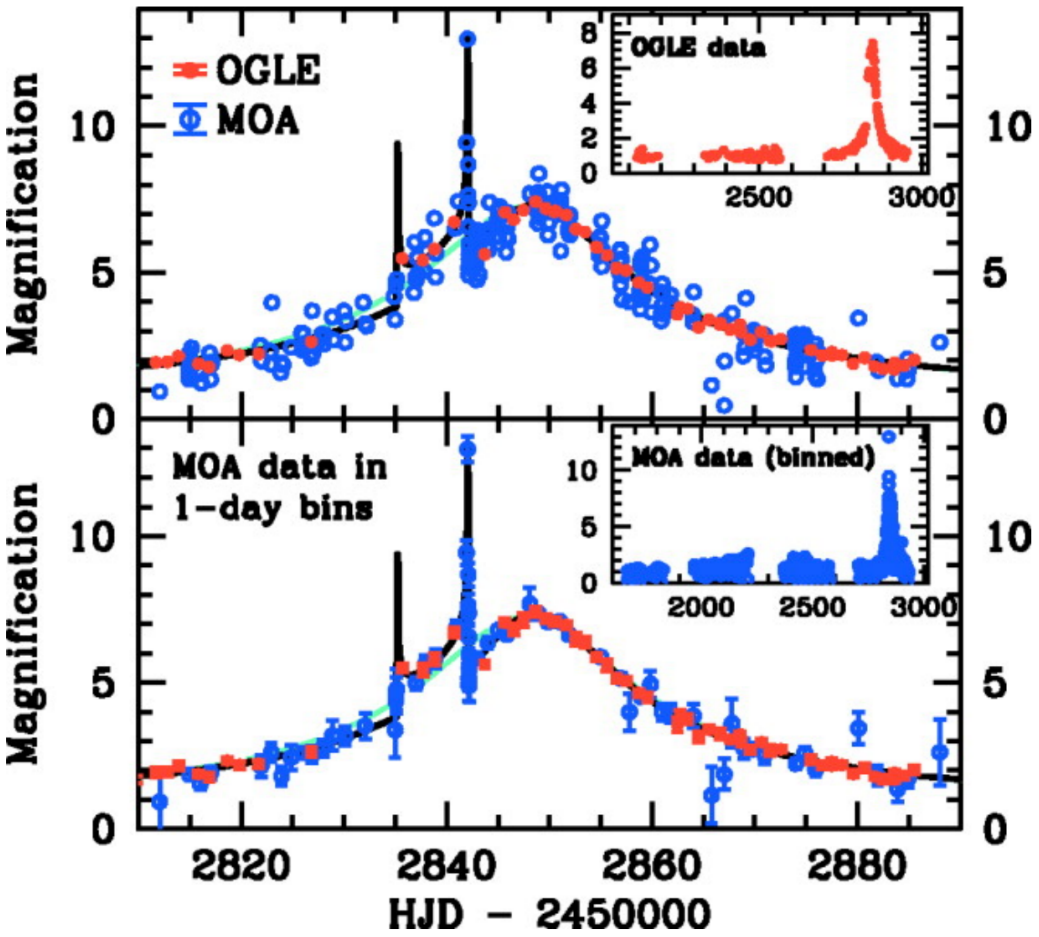

Figure 1.3: Example of microlensing light-curve, showing the discovery of OGLE 2003BLG235/MOA 2003BLG53 b. Data from the OGLE instrument is shown in red and data from MOA is shown in blue. Best-fitting microlensing model is plotted as a black line, with the unperturbed single-lens model (without the presence of a planet) over-plotted in cyan. Reproduced from Bond et al. [2004]

Microlensing has proved most sensitive to long-period, low-mass planets, meaning that it fills a crucial niche in the planetary discovery parameter space which more prolific methods such as the transit and radial velocity methods fail to populate. This means that the method is also able to probe planets more similar to those in our own solar system.

1.3 Current exoplanet knowledge

1.3.1 Overview of Exoplanet Demographics

Thanks to the wide variety of exoplanet discovery methods discussed above, almost 5000 validated exoplanets have now been discovered, with radii, masses and environments far more diverse than those planets in the Solar System. The current number of planets found via each method (alongside a number of less-common discovery methods) are included shown in Table 1.1, as well as their average sizes and periods. This vast population of exoplanets is now allowing the exoplanet field to move towards in-depth characterisation of exoplanetary systems alongside further discoveries.

Table 1.1: Overview of all Exoplanets discovered as of 23 Feb 2022. All values retrieved from the Exoplanet Archive. Note that because masses for many transiting planets are not yet known, the more representative radius measurement is used in the average size column instead of mass. Of the 4933 validated exoplanets known when this table was compiled, 3814 exoplanets had measured radii, 2053 had measured mass, and only 949 had both. The one astrometric "planet" from the Exoplanet archive is retained here for completeness, but may be a brown dwarf given its large mass.

Discovery Method	Number of Planets	Median Size	Median Period
Astrometry	1	$28.5 M_{\rm Jup}$	0.675 yrs
Direct Imaging	58	$11.3 M_{\rm Jup}$	327 yrs
Radial Velocity	915	$1.05 M_{\rm Jup}$	0.893 yrs
Transit Method	3778	$0.210 R_{\rm Jup}$	8.69 days
Microlensing	124	$0.640 M_{\rm Jup}$	7.30 yrs
Transit timing variations	22	$0.178 M_{\rm Jup}$	20.3 days
Eclipse timing variations	16	$6.18 M_{\rm Jup}$	10.0 yrs
Pulsar timing variations	7	$0.0135 M_{\rm Jup}$	45.9 days
Pulsation timing variations	2	$7.50 M_{\rm Jup}$	2.75 yrs
Orbital brightness modulations	9	$1.25 M_{\rm Jup}$	0.812 days
Disc Kinematics	1	$2.50 M_{\rm Jup}$	Unknown

From the earliest exoplanet discoveries like 51 Pegasi b, a Jupiter-sized planet orbiting within the equivalent orbit of Mercury, it quickly became clear that many exoplanetary systems are quite unlike those planets in our own system. Indeed because of the difficulty of detecting Earth-sized and/or long-period exoplanets, the vast majority of currently known exoplanets have no Solar-system analogue when both size and period are considered. However, many of the discovered exoplanets do have masses and radii similar to those in the Solar system, so in order to keep track of different types of new exoplanets, the exoplanet community has defined

Table 1.2: Approximate exoplanet size categories. Based on radius and mass definitions proposed by Borucki et al. [2011] and Stevens and Gaudi [2013] respectively. For reference, Jupiter itself is $11.2R_{\oplus}$ and $318M_{\oplus}$.

Planet Size	Radius	Mass
Earth-sized	$< 1.25R_{\oplus}$	$0.1 - 2M_{\oplus}$
Super-Earths	$1.25 - 2R_{\oplus}$	$2 - 10M_{\oplus}$
Neptune-sized	$2 - 6R_{\oplus}$	$10 - 100M_{\oplus}$
Jupiter-sized	$6 - 15R_{\oplus}$	$100 - 1000M_{\oplus}$

approximate radius and mass bins, as presented in Table 1.2. For the purpose of this research the radius limits are taken from those from Borucki et al. [2011] and the mass limits are based on those proposed by Stevens and Gaudi [2013].

The overall distribution of exoplanets bears some interesting features, as explored in Figures 1.4 and 1.5. The mass-period distribution period shown in Figure 1.4 gives the best overview of which areas of the exoplanetary distribution different discovery methods are sensitive to, and shows how combining these discovery methods is crucial to understanding the full extent of exoplanetary demographics. When coupled with discovery likelihoods the current distribution of planets suggests that super-Earths and sub-Neptunes account for the majority of all exoplanets [e.g. Mayor et al., 2011; Howard et al., 2012], especially as large exoplanets around the most common types of stars (M-dwarfs) appear to be rare [Dressing and Charbonneau, 2015].

Meanwhile, plotting radius against period instead (e.g. Figure 1.5) reveals an interesting paucity of planets in two regions: the first (and larger) for close-in giant planets (christened the 'evaporation desert') and the second for small planets with periods < 100 days and radii between 1.5-2.0$R_{\oplus}$. The secondary region of decreased planetary occurrence for small planets was originally christened the 'Fulton Gap', after the original discoverer [Fulton et al., 2017], however has more recently been renamed the 'radius/photoevaporation valley' as more planets have been found in this comparatively uncommon region of exoplanetary discovery space. Both of these regions are currently thought to be caused by atmospheric mass loss, with the presence of two different regions explained by there being two different mechanisms to achieve unstable mass-loss. These mechanisms are discussed in detail in Owen [2019], but in brief the evaporation desert is thought to be caused by rapid atmospheric loss of large exoplanets when their atmospheric radius dominates over the planet's radius (meaning that excess high-energy flux overcomes the gravitational force holding the atmosphere to the planet), with the photoevaporation continuing until these energies once again balance. Meanwhile, the evaporation valley sits be-

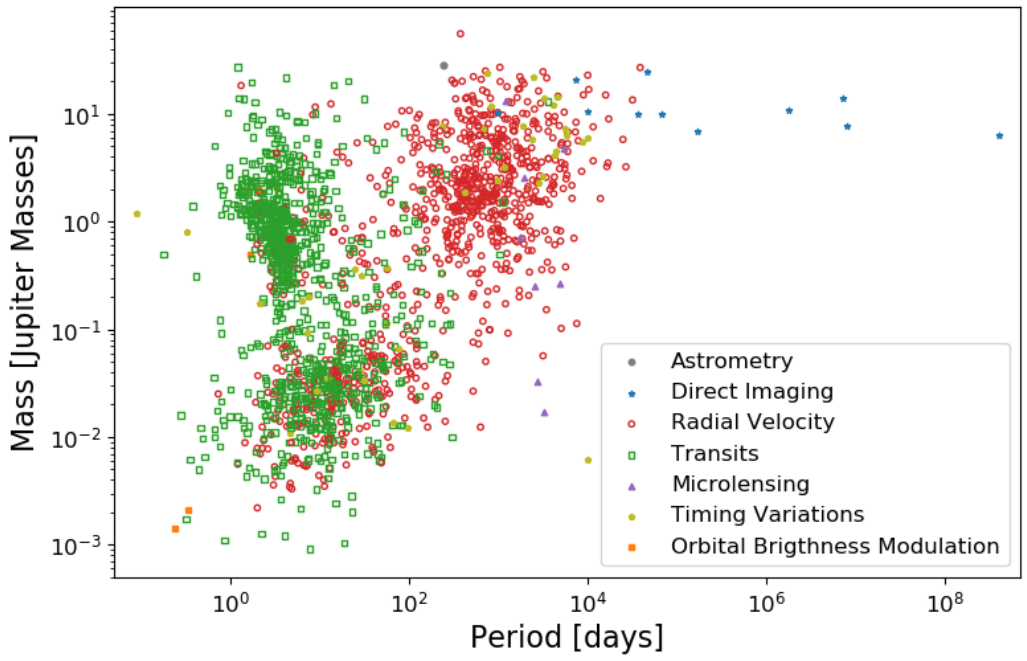

Figure 1.4: Mass-Period distribution for confirmed exoplanets with known masses. Mass, period and discovery method data retrieved from the Exoplanet Archive on 22nd Feb 2022.

tween those exoplanets with sufficient core mass to retain their atmospheres in the presence of high-energy fluxes, and those with sufficiently light cores to allow their atmospheres to be completely stripped (or at least reach a small enough atmospheric mass that it is stable to mass-loss). In this way, this valley is currently thought to represent the transition between rocky/terrestrial planets [e.g. Rogers, 2015] and planets with extended atmospheres [e.g. Wolfgang and Lopez, 2015], with planets in the valley still undergoing the loss of their atmosphere via photoevaporation [Owen, 2019].

Finally, although it is comparatively rare for exoplanets to have both precisely measured masses and radii, mass-radius diagrams for those rare systems (such as those shown in Figure 1.6) give insight into the potential composition of exoplanets, as discussed in section 1.3.3. Filling in this distribution further is one of the key goals of modern exoplanetary science, as it will give much greater clarity to the formation and evolution of exoplanets. Nonetheless, these general demographics are already allowing explanations for planetary evolution and composition to be devised, as discussed next.

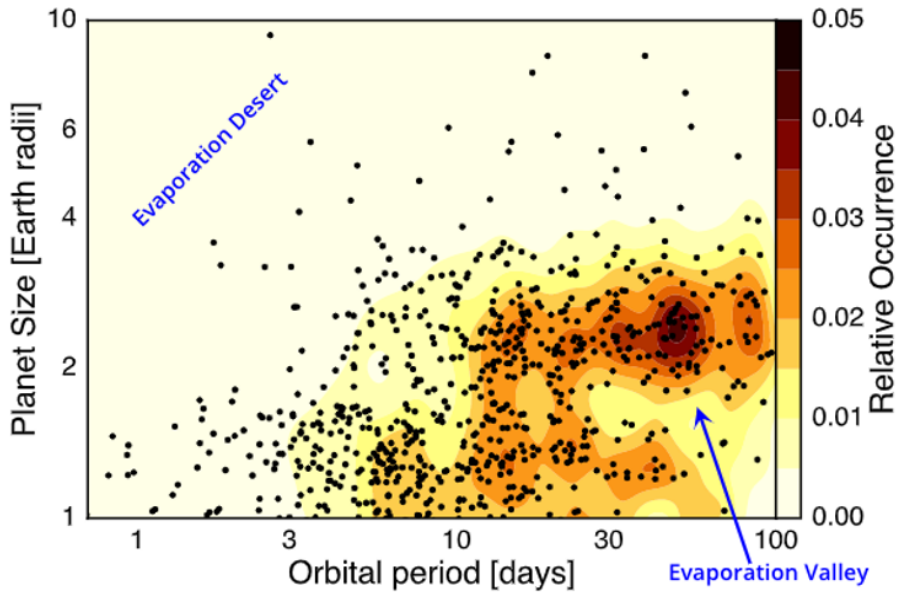

Figure 1.5: Radius-Period distribution for small exoplanets, showing the planetary occurrence rate corrected for observational bias in colour. The location of the Sub-Neptune desert and radius/photoevaporation valley are shown in blue, with planets studied in the California Kepler Survey [Petigura et al., 2017] shown as black points. Reproduced from [Owen, 2019].

1.3.2 Formation and Evolution

Planets begin their lives in clouds of dust and gas surrounding newly formed stars. Stars typically form within nebulae, in dense molecular clouds consisting of mostly hydrogen. When a portion of this cloud becomes dense enough, it begins to collapse under its own gravity, eventually ending up as a protostar fusing hydrogen to helium and spinning with the solar nebula's net angular momentum. This rotation, coupled with the law of conservation of momentum, causes the remaining cloud of dust and gas around the protostar to flatten out into the protoplanetary disk, typically containing a few percent of the system's mass [Herbig, 1962]. It is in these protoplanetary discs that planet formation can begin in earnest.

There are currently two competing theories for how planets form in these discs: gravitational instability [Boss, 1997] and core accretion [Safronov, 1972; Lissauer, 1993; Pollack et al., 1996; Matsuo et al., 2007]. The gravitational instability theory proposes that planets form from the breakup of the protoplanetary disc due to gravitational instabilities in the disc forming self-gravitating clumps of gas, which

eventually evolve into planets. This model relies heavily on the thermodynamic state of the disk allowing disk fragmentation to occur alongside a disc cooling timescale comparable to the local disc orbital period [Boss, 1997]. Such a method is particularly good at explaining the formation of massive planets at wide orbits, such as Jupiter and Saturn in our own solar system [Dodson-Robinson et al., 2009]. However, such exoplanets are rarely seen in current exoplanet discoveries (see Chauvin et al. [2015]; Lannier et al. [2016]).

The more commonly accepted model of exoplanetary formation (and that which is more effective at forming close-in exoplanets) is the core accretion method [Safronov, 1972; Lissauer, 1993; Pollack et al., 1996; Matsuo et al., 2007]. Under this method planets grow from the ground up, beginning as small particles about 1μm in size, and combining via adhesive collisions to mm-sized particles. The rate of growth in this regime is governed by aerodynamic considerations and material physics, with particles coalescing much more quickly outside the snow-line than within it [Kennedy and Kenyon, 2008]. Beyond the mm regime, particles begin to coalesce into larger m to km-sized planetesimals, which impact and eventually form terrestrial planets or the rocky cores of gas giants [Lissauer, 1993]. For more massive cores (over $\sim$10 $M_{\oplus}$), these cores are able to rapidly accrete remaining gas from the gas disc, growing into gas giants [Pollack et al., 1996]. Such accretion stops when all remaining gas and dust in the vicinity of the planet has been accreted or blown away by radiation pressure and activity from the host star. The difference in rate in grain growth in the early stages helps to explain the terrestrial planet formation in our own solar system, as planets within the snowline took longer to form, so less gas was left in the protoplanetary disc by the time they were large enough to accrete excess gas for their atmospheres.

However, while the core accretion method aligns much more with the observed population of exoplanets, the multitude of extremely close-in exoplanets like Hot-Jupiters and very short-period terrestrial planets still poses problems to this formation pathway, especially for smaller stars like M-dwarfs [e.g. Raymond et al., 2007; Miguel et al., 2019]. Because of the lack of planet-forming materials at such radii from the protostar, to form these planets significant planetary building blocks (or indeed planets themselves) must have migrated inwards at some point during the early system evolution [Rein, 2012]. Two main migration methods have been proposed to explain this: disc migration or gravitational disruption from another body in the system (typically another planet) after the disc has dissipated. In disc migration planets move inwards either because Linblad resonances excite spiral density waves in the gas within the protoplanetary disk and create a torque-mismatch caus-

ing planet to lose angular momentum (Type I migration; Ward [1986]), or because a massive planet opens a gap in the gaseous disc and transfers angular momentum to gas outside the planet, causing the planet to slowly spiral inwards following the accretion of gas onto the star (Type II migration; Lin and Papaloizou [1986]). Given the required planetary mass to open a gap in the disc, Type II migration is the dominant process for large mass planets [Lin and Papaloizou, 1986], while Type I migration is particularly important for smaller planets (e.g. $0.1M_{\oplus} < M_{\mathrm{P}} < 10M_{\oplus}$; Ogihara and Ida [2009]; Ormel et al. [2017].

If however the disc has already dissipated (as tends to occur within 10 Myr or so; Kennedy and Kenyon [2008]), migration can instead occur via exchanging angular momentum with another body in the system. For example, in a mechanism called the Kozai-Lidov effect [Kozai, 1962; Lidov, 1962], a massive outer perturber such as a gas giant or nearby star can cause an inner low-mass planet to librate between highly eccentric and highly inclined orbits. Should the disrupted planet's orbit take it close enough to the star during perihelion, it may exchange angular momentum with the star via strong tidal forces and gradually circularise at a much closer orbital radius [Rodríguez et al., 2011].

These migration pathways can be tested observationally, as disk migration would tend to dampen orbits, resulting in circular, 'well-behaved', uninclined orbits, while perturbation from interactions or exchange of angular momentum with another body in the system would tend to result in highly inclined, eccentric orbits (unless it was subsequently re-circularised). Common resonances of exoplanet chains around low-mass stars like TRAPPIST-1 [Luger et al., 2017] and TOI-178 [Leleu et al., 2021] strongly suggest that convergent migration via resonance capture may also play a significant part, but observations of more such systems are required before it can be determined how common this is.

Other important factors affecting the evolution of planets include thermodynamical cooling, tidal interactions with their host stars and atmospheric loss. Because exoplanets typically form through the collapse of gas and dust in planetary nebulae, much of their initial energy comes from transformed gravitational energy, which is particularly significant for gas-giant planets like Jupiter. Due to the thermal expansion properties of gases, the gradual loss of this initial energy through radiation leads to gradual cooling and contraction of exoplanetry atmospheres, in turn reducing the radius and luminosity of these planets [Guillot et al., 2014]. The rate at which exoplanets lose energy in this manner (and hence luminosity, L) can

be shown (e.g. by Perryman [2018]) to follow a modified Kelvin-Helmholtz relation:

$$L \approx \eta \frac{GM^2}{R\tau} \tag{1.22}$$

where τ is the age of the exoplanet, and η is a coefficient that depends on anumber of complicating factors such as the composition and compressibility of the atmosphere. In isolation then, exoplanets (especially large, gaseous ones) typically cool and contract as they age. Some extra care is required for planets in close orbits however, as exoplanet cooling rates also depend on irradiation from their host stars, and associated atmospheric loss (see below).

Meanwhile, for exoplanets which orbit particularly close to their host stars, tidal forces can have a significant effect on the long-term evolution of the planet. Indeed for planets within $\approx$ 0.2 au tidal effects typically lead to 'tidal locking', where a planet's rotation period and orbital periods become synchronised, alongside reducing the orbital eccentricity and semi-major axis, and subsequently heating the planet through conversion of energy from orbital energy [Perryman, 2018]. These processes, often referred to as 'tidal circularisation', typically occur up to the first 1Gyr of exoplanetary age [Beaugé and Nesvorný, 2012; Barnes, 2017].

As discussed in the previous section, the atmospheres of exoplanets also evolve over their lifetimes, particularly in the first 100Myr [Owen, 2019]. This is driven by a range of processes, such as photoevaporation or 'Jeans escape' where high-energy photons increase the temperature of molecules in the atmosphere, and may reach 'escape-velocity' from the atmosphere, hydrodynamic escape through heating of the planet itself (which can be externally or internally driven), photolytic dissociation where UV radiation breaks down individual molecules in the atmosphere, or from planetesimal erosion, where giant impacts from large bodies leftover in the planetary system (such as other planetesimals) can ablate significant portions of the atmosphere at once [Perryman, 2018]. This atmospheric evolution is reviewed in greater depth in Perryman [2018] and Owen [2019].

It should be noted however that planet formation and early evolution is still a rapidly evolving field, with many open questions remaining. Knowledge in this field is limited by a combination of factors such as a lack of bona-fide young stars, optically thick protoplanetary discs inhibiting photometric evaluation of early evolution and the fact that direct imaging is still only able to see structure with resolution on the order of AU. Hence the more confirmed young exoplanets across well-constrained ages the community gains, the more can be learnt about this early evolution.

1.3.3 Planetary Characterisation and Composition

A significant amount can be learnt about a planet just from a handful of measurable quantities: the stellar and planetary radii ($R_*; R_\mathrm{P}$), planetary mass (M_P), orbital distance (a) and stellar temperature (T_*). For example, due simply to balancing the incoming flux and outgoing radiation from a planet (assuming re-radiation from the day-side only and a planetary albedo, A, equal to the fraction of incident light reflected by the planet), one can derive a simple estimation of the planet's equilibrium temperature:

$$T_p = \left(\frac{\pi R_*^2 T_*^4 (1 - A)}{a^2} \right)^{1/4} \tag{1.23}$$

This temperature can then be used to characterise the planet (e.g. as a Hot-Jupiter), or to place it in or out of the 'Habitable Zone' - defined as the region around the host star where a planet would be capable of supporting liquid water [Kasting et al., 1993; Kopparapu et al., 2013]

Meanwhile, the precise masses and radii (and hence densities) from combined radial velocity and transit surveys can be used to give a first estimate of the composition or elemental make-up of such planets (though degeneracies can still exist between different compositions - see for example Figure 3 of Delrez et al. [2021]). Because of the great distances to most known exoplanets, most exoplanet interior characterisation is based on synthetic computer models and equations of state for typical planetary materials seen in the solar system [e.g. Guillot, 2005; SOTIN et al., 2007; Fortney et al., 2007; Dressing et al., 2015]. For example, very light planets such as Jupiter-density planets are thought to be dominated by hydrogen and helium (with minimal cores), while denser, terrestrial planets are thought to harbour iron cores and silicate-like mantles like Earth. These models (and associated mass-radius diagrams) typically align well with observations of Earth-like and gas-giant planets, except where planets are significantly inflated due to youth and/or XUV irradiation [e.g. Benatti et al., 2019; Eigmüller et al., 2019]. However, the most common planets so far found in exoplanet surveys (Super-Earths and sub-Neptunes) actually fall in parameter spaces that simply do not exist in the solar system (e.g. see Figure 1.6). Given the importance of these exoplanets and the degeneracies that exist between structural parameters for such objects, Dorn et al. [2017] have produced a separate generalised Bayesian inference method for constraining their interiors.

Very recently, a new method of directly probing the interior of exoplanets has also been discovered, which involves analysing the spectra of old planet-hosting stars at the end of their lives in the white dwarf phase [e.g. Gaensicke et al., 2019;

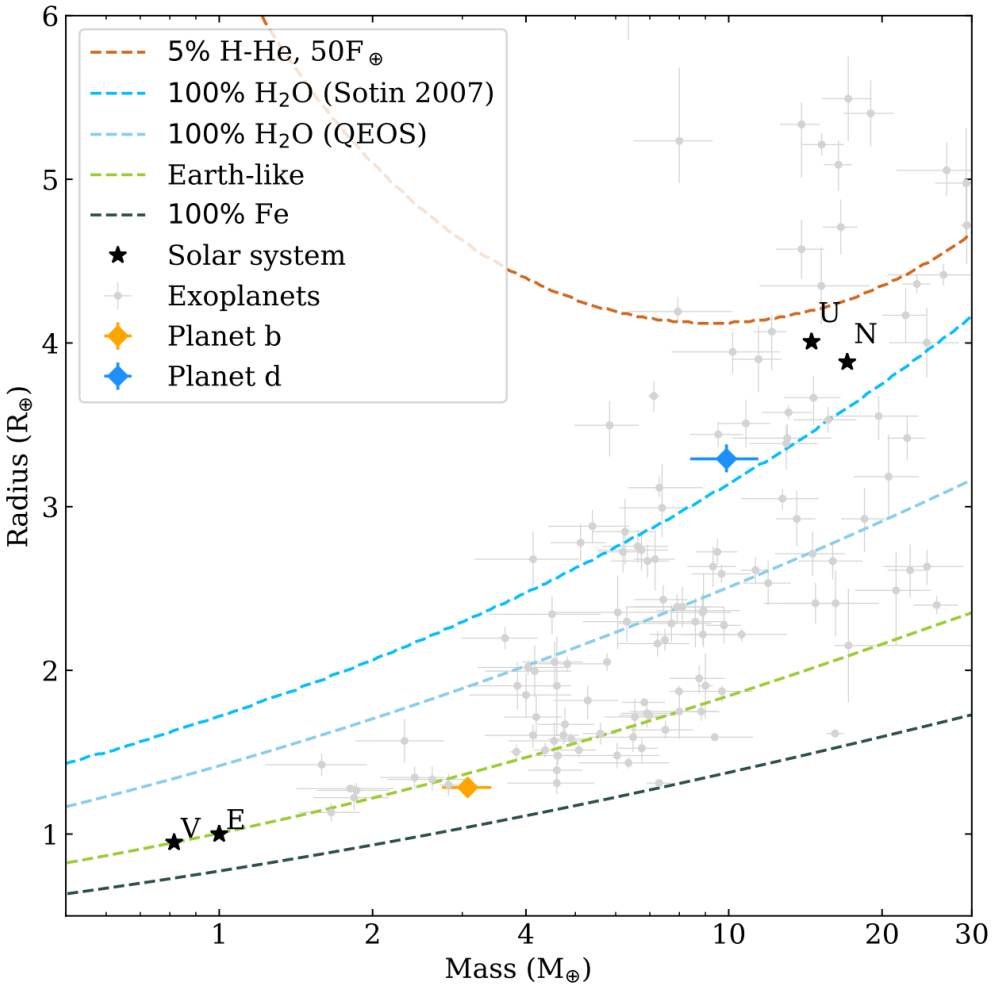

Figure 1.6: Mass-radius diagram of all known exoplanets with mass measurements more precise than 4σ from the NASA exoplanet archive, as of 22 Sept 2020. Composition lines from [SOTIN et al., 2007] are overplotted for iron (dark grey), Earth-like (green), water (blue) and a 5% H/He; 95% water mix in brown. The TOI-431 system is shown in orange (planet b) and blue (planet d) as examples of two planets in a single system which fall either side of the radius-period valley [Fulton et al., 2017]. Solar System planets are shown as black stars, labelled as V, E, U and N for Venus, Earth, Uranus and Neptune respectively. Reproduced from Osborn et al. [2021]

Cunningham et al., 2022]. Spectra of these white dwarfs have shown evidence of constituents like H_2O and H_2S accreting from an orbiting circumstellar gas disk, resembling predictions for the inner atmospheres of icy giant planets. Further observations of similar white dwarf systems thus may allow direct detection of exoplanetary materials to support the wider interior composition modelling.

Another crucial aspect of exoplanet composition is what their atmospheres are made of. Measuring the composition of exoplanet atmospheres not only offers clues to the formation and evolution of (especially larger) exoplanets, but also is suggestive of geophysical and potentially biological processes taking place on the exoplanet [Madhusudhan, 2019]. In this way the study of atmospheres offers a way to probe the question of "Are we alone in the Universe?" by searching for molecular species which could only be produced by life. Determining precisely which molecules are most likely as biosignatures remains a rich ongoing area of research, as summarised by Schwieterman et al. [2018].

The continued development of the highly successful transit technique over the last decade has proved transformational to exploring and understanding these atmospheres, particularly with the advent of transit spectroscopy [e.g. Charbonneau et al., 2002]. A transiting exoplanet offers three opportunities to measure its spectrum: i) a transmission spectrum during the transit, when light from a host star passes through the atmosphere of a transiting exoplanet; ii) an emission spectrum when the planet travels behind the star; and iii) a phase curve as the planet travels between its primary and secondary eclipse (and hence presents different sides of its atmosphere to an observer). The observed spectra can then be compared to known patterns of spectral lines (or 'line lists') for different chemical species in order to determine which species are present in the atmosphere (as summarised by Madhusudhan [2018]). Additional techniques of probing exoplanet atmospheres include high-resolution Doppler spectroscopy [Brogi et al., 2012; Birkby, 2018] and direct imaging spectroscopy [e.g. Barman et al., 2011, 2015], which give further information on the atmospheres of close-in giant exoplanets and young directly imaged exoplanets respectively. The combination of these three techniques have now allowed multiple tens of exoplanets to have their atmospheres measured, providing initial constraints on the presence of key molecular and atomic species such as H_2O, CO, CH_4, CO_2, HCN, TiO, VO, Na and K in exoplanetary atmospheres, as well as shedding light on such features as exoplanetary clouds and atmospheric circulation [Madhusudhan, 2019].

Measuring these atmospheres has also opened the window to understanding how exoplanetary atmospheres evolve with time, which may help to explain important features such as the period-radius gap in exoplanetary demographics first shown by Owen and Wu [2017]. Exoplanets are thought to lose their atmospheres by two main mechanisms: ionic pick-up caused by high energy charged particles in the stellar wind from their host stars [e.g. Khodachenko et al., 2015], or photoevaporation, where thermal heating of the upper layers of the atmosphere caused by

X-ray of extreme-UV causes hydrodynamic escape of atmospheric materials [Watson et al., 1981]). The latter is currently thought to be the dominant mechanism, as it can account for the radius gap by itself [Fujita et al., 2022], but this would be aided significantly by further observations of exoplanets with known ages either side of this gap.

Future prospects are bright for continued exoplanet characterisation in the coming decades, with the recent launch of the James Webb Space Telescope (JWST; Gardner et al. [2006]) and CHEOPS satellite [Broeg et al., 2013] extending both the wavelength coverage and sensitivity for both spectroscopic and transit-based measurements, alongside the upcoming launch of ESA's PLATO mission which is expected to dramatically increase the precision of known planetary radii for bright targets excellent for transit spectroscopy [Rauer et al., 2014].

1.4 Young Exoplanets

1.4.1 Young Exoplanet Opportunities

Despite the wealth of almost 5000 verified exoplanets known today, there are still many unanswered questions concerning the formation mechanisms and system evolution that has led to the observed distribution of these objects. In particular, planetary migration, dynamical interactions with nearby stars and atmospheric evolution all can have major effects on the distribution and architecture of the final exoplanetary population. This is especially important early in a planet's life (e.g. <1Gyr), where phenomena such as accretion [Marley et al., 2007; Manara et al., 2019], ionising radiation causing atmospheric loss [Baraffe et al., 2003; Owen, 2019] and dynamical interactions with other forming planetary system bodies [Ida and Lin, 2010; Schlichting et al., 2015] can significantly change the mass, radius and orbital parameters of early exoplanets. Thus while traditional exoplanet studies have been biased towards older exoplanets due to their relatively quiet host stars, there is a strong case to be made for the search for younger exoplanets, when planets are undergoing the majority of their evolution [Adams and Laughlin, 2006; Ida and Lin, 2010; Spiegel and Burrows, 2012; Ida et al., 2013].

Young planets are particularly crucial for testing planetary migration, with smooth planet migration through the planetary disk currently thought to operate on very short lifetimes (<10 Myrs/disk lifetime) [Baruteau et al., 2014] and long-term high-eccentricity migration taking place up to about 1 Gyr [Chatterjee et al., 2008]. However in dense environments the situation can be complicated further, with Parker [2020] conducting simulations of Jupiter-size exoplanets and showing

that dynamical interactions significantly disrupted the orbits of approximately 20% of the initial planets even after just 10 Myr, as illustrated in Figure 1.7. In order to observe these processes directly, it is crucial to build up an observational timeline of young exoplanets within the <1 Gyr age bracket. Early discoveries in this area are already challenging traditional planet formation and migration theories, with the discovery of the 11 Myr, P = 5.4 day planet K2-33b in Sco-Cen by Mann et al. [2016b] showing that at least some short-period planets migrate within the first 10 Myr or form in situ.

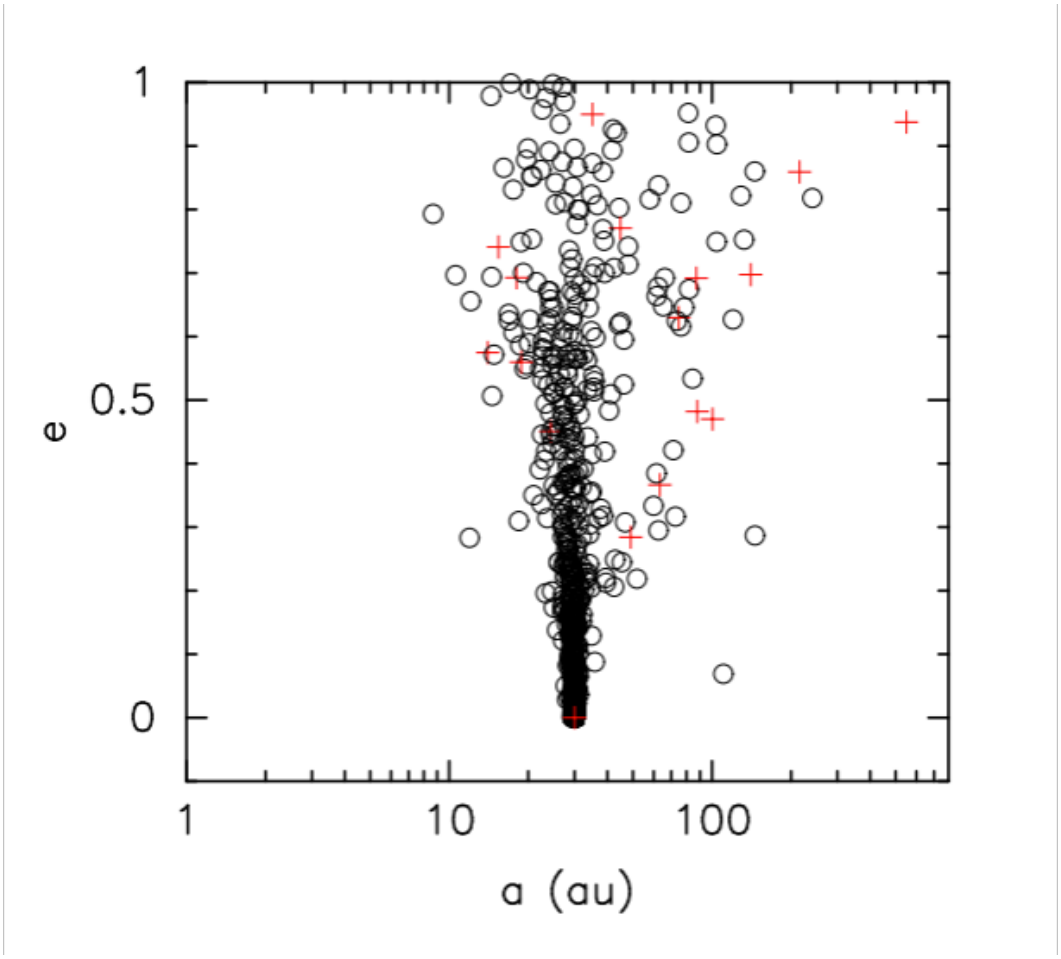

Figure 1.7: Orbital disruption for Jupiter-sized planets ($1 M_{\rm Jup}$; e=0; a=30AU) after 10 Myr in a dense ($\sim$1000 $M_{\rm star} pc^{-3}$) star forming region. Reproduced from Parker (2020)

Another key opportunity offered by young planets is the ability to directly probe atmospheric loss, either from UV/X-ray irradiation [Owen and Wu, 2017], or through internal heating [Lopez and Fortney, 2013; Ginzburg et al., 2018; Gupta

and Schlichting, 2019]. These processes are theorised to shape features in the mature population such as the sub-Neptune desert and radius valley [see Figure 1.5, reproduced from Owen, 2019], but the timeline of this evolution is still not fully constrained. Runaway evaporation may also occur for 2-4 R_{Earth} planets, leading to stripped cores [Armstrong et al., 2020b]. This, coupled with high-eccentricity migration may help to explain the dearth of close-in giant planets [Owen and Lai, 2018].

However, before these atmospheres are lost, the brightness, low density and large scale height of many young planets theoretically should be superb for atmospheric characterisation through transmission spectroscopy [Feinstein et al., 2022]. This has been tested already for such young planets as K2-25b, K2-100b and DS Tuc Ab, but features such as clouds, hazes and dust grains increase atmospheric opacity and have made these measurements far more challenging, leading to null detections in the past [e.g. Gaidos et al., 2020b,a; Benatti et al., 2021].

Young exoplanets may also offer the key to one of the key formation mysteries in exoplanet science: the formation of super-Earths and sub-Neptunes. Owen [2020b] showed that various planet formation models predict vastly different entropies at the end of the protoplanetary disc stage, meaning that if formation entropies can be measured, appropriate formation techniques can be assigned. In order to gain these entropies, precise masses, radii and ages are needed for planets of ages ∼20-60 Myr, with mass precision of <20% [Owen, 2020a]. Such measurements would allow the formation mechanisms for such important young irradiated planets such as DS Tuc Ab and V1298 Tau b to be obtained. Precise masses will also help the community to explore the dependency of the planetary mass-radius relation with the age, which is well-constrained for the older population but largely unexplored for sub-Gyr systems [Benatti et al., 2019]. Finding the precise masses of young planets is hence just as important as the initial discovery of these planets.

Discovering exoplanets <1 Gyr old is thus crucial to directly probe the main causes of planetary evolution and help to fill a key gap in the community's knowledge of exoplanet history.

1.4.2 Groups of young stars

A particularly promising place to look for these young exoplanets is within young clusters and stellar associations.[8] These are groups of stars which formed together and hence share a common history and similar composition. These commonalities

[8]Note that the definition of clusters and associations can be a little fluid in astrophysical literature, so occasionally are merged together or also named 'moving groups'.

mean that precise stellar properties can be determined for stars within the groups [Torres et al., 2006], which in turn provides a significant advantage for precise determination of exoplanet characteristics. Similarly the ages of stars within these groups are typically very well constrained, which can allow a more detailed timeline of planetary evolution to be assembled if planets are found within them.

Star clusters are large assemblies of gravitationally bound stars with similar composition, held together by their own gravity [Carroll and Ostlie, 2017]. These clusters can be broken down into two main groups: globular clusters and open clusters. Globular clusters are dense, populous, nearly spherical collections of stars, often containing some of the oldest stars in the galaxy. Open clusters are smaller, more irregular assemblies of stars, which are typically much younger than globular clusters. Open clusters are the most useful type of cluster for exoplanet searches, as they are far more diffuse than globular clusters. At present, over 1200 open clusters are known in the Milky Way, but many more are likely to exist [Cantat-Gaudin et al., 2018].

Stellar associations on the other hand are groups of gravitationally *unbound* stars which have a common origin, so still retain a common proper motion across the sky. These stars typically have similar ages and compositions, and likely resulted from star clusters which have steadily moved apart and become gravitationally unbound [de Zeeuw et al., 1999; Gagné et al., 2018b]. The first stellar associations were discovered by Victor Ambartsumian (1947), who categorised them in two types – OB and T – depending whether their most obvious stars were O/B stars or T-Tauri stars respectively [Ambartsumian, 1947]. Now almost 30 stellar associations are known, including the Ursa Major moving group and the separate parts of the larger Upper-Scorpius OB association. [Gagné et al., 2018b]. The relatively diffuse nature of stellar associations compared to similarly young star clusters makes them far more suited to study with *TESS* given the relatively large pixels of the satellite [angular resolution ∼21", Vanderspek et al., 2018].

Traditionally stellar associations and open clusters have been discovered from bright and more obvious stellar types, however by considering the Initial Mass Function for stars (Salpeter [1955] to Chabrier [2003]) it is expected that these massive stars should be associated with large numbers of less-massive stars. Determination of these lower-mass members is however made considerably more difficult given the lower-precision galactic position and velocity measurements for these naturally dimmer stars. Recent missions such as *Hipparcos* and the all-sky *Gaia* Mission (described in more detail below) have begun to combat this problem, collecting very high precision measurements of stars down to magnitude 20. This should pave the

way for both significantly expanding the membership lists of known stellar clusters and associations, as well as potentially discovering new ones. New groups of stars dominated by brighter F to M-type stars would be particularly beneficial for exoplanetary science, especially given the prospects of follow-up from TESS.

The most extensive census of 'bona-fide' members of these stellar associations was assembled by Gagné et al. [2018b] in the development of their BANYAN Σ Bayesian membership tool, but has recently been further expanded by works such as Gagné et al. [2018c] and Esplin and Luhman [2019] thanks to the release of *Gaia* DR2 [Gaia Collaboration et al., 2018a]. The unprecedented astrometric precision offered by the second data release from the *Gaia* satellite [Gaia Collaboration et al., 2016b, 2018a] has also allowed extensive membership lists to be built for all known open clusters [e.g. see Cantat-Gaudin et al., 2018], as well as revealing a wide array of associated 'stellar streams' [Kounkel and Covey, 2019]. Bouma et al. [2019] recently assembled many of these larger membership lists into the extensive CDIPS catalogue of young stars in clusters and associations (of which the BANYAN sample is a subset) by considering a wide array of sources in the literature. For the initial design of the pipeline constructed in this work (and described in Section 2.6) the homogeneous BANYAN selection criteria focused solely on stellar associations was preferred, however the full CDIPS target list was used to choose stars from the first year of the *TESS* mission in Chapter 5. The methods used to assemble such lists are described in section 2.1.

1.4.3 Stellar Activity

Unfortunately the host stars of young exoplanets provide significant challenges for discovery due to their typically increased activity, rotation rates and relative proximity to neighbouring stars [e.g. Koeltzsch et al., 2009; Sergison, 2015; Rivilla et al., 2015; Mascareño et al., 2016; Briceno et al., 2019]. Most stellar activity is set up by the complex interplay between convection and magnetic fields of the stars [Cegla et al., 2019]. The main stellar features that cause problems for exoplanetary surveys are surface granulation, rotating star spots, convection-induced pulsations and flares. Surface granulation arises from hot bright bubbles of gas rising to the surface which then cool and fall back down into the inner regions of the star at the cooler edges of each convection cell [Cegla et al., 2019]. This sets up a patchwork of small lighter and darker regions on the majority of the star (e.g. see bottom right, Figure 1.8), which vary in brightness and net red/blue shift when averaged across the whole stellar surface. If however there is a concentration of magnetic flux in a specific region of a star, this can act to inhibit convection at this point, setting

up a large darker spot on the surface of the star [Biermann, 1941] (e.g. see top right, Figure 1.8). This is what is referred to as a star spot. As a star rotates and stellar spots move across the surface of the star, the observed average brightness of the star varies, often with significant amplitude in young stars (e.g. see Figure 1.9). Similarly the observed radial velocity will change (see Figure 1.10) as light is blocked from the blue-shifted side of a rotating star coming towards the observer and then blocked from the red-shifted side moving away from the observer. Particularly active stars such as young stars and/or fast rotators often harbour many star-spots, which can set up very complex patterns in observed photometry and spectroscopy. A closely related phenomenon to star spots are faculae and plages, which are bright spots/regions on the solar photosphere that are produced by concentrations of magnetic field lines [Solanki et al., 2006], and have the opposite effect to the dark star spots. These are pictured as the brighter regions of the star shown in Figure 1.8. Regions of intense magnetic flux can also result in stellar flares (e.g. see left of Figure 1.8), when the reconnection of magnetic field lines causes an intense release of electromagnetic radiation [Benz and Güdel, 2010]. Typically these are short in duration but can result in dramatic brightening of the observed star. Finally, convection within the star also sets up pressure-mode oscillations, which can change the shape of the star with time. This effect typically dominates over stellar granulation, but is far less of an effect than star spots for young stars [Cegla, 2020].

The combination of all of these phenomena can result in periodic stellar variability that can be similar in period but much larger in amplitude than planetary signals around these stars [Armstrong et al., 2015; Cody et al., 2018]. Furthermore, the increased rotation rates and corresponding increased magnetic activity of such young stars can result in swiftly evolving stellar activity signals (especially from the changing geometry of large star-spots), making predicting the exact effect of stellar activity at a given time considerably harder. Such challenges are illustrated in Figures 1.9 and 1.10 in real photometric and radial velocity time-series data respectively.

In photometric transit searches, granulation can often be handled by binning/averaging over the 5-10min granulation evolution timescale, and the short-duration intense brightening caused by stellar flares can often be easily found and removed before transit searches. This being the case, the main stellar activity nuisance signals are caused by stellar pulsations and evolving star spots, alongside additional non-activity types of young star variability (e.g. dust and gas in the way of the planets and other variability modes). In order to discover new young exoplanets, it is imperative to dissociate true transit signals from this stellar activity

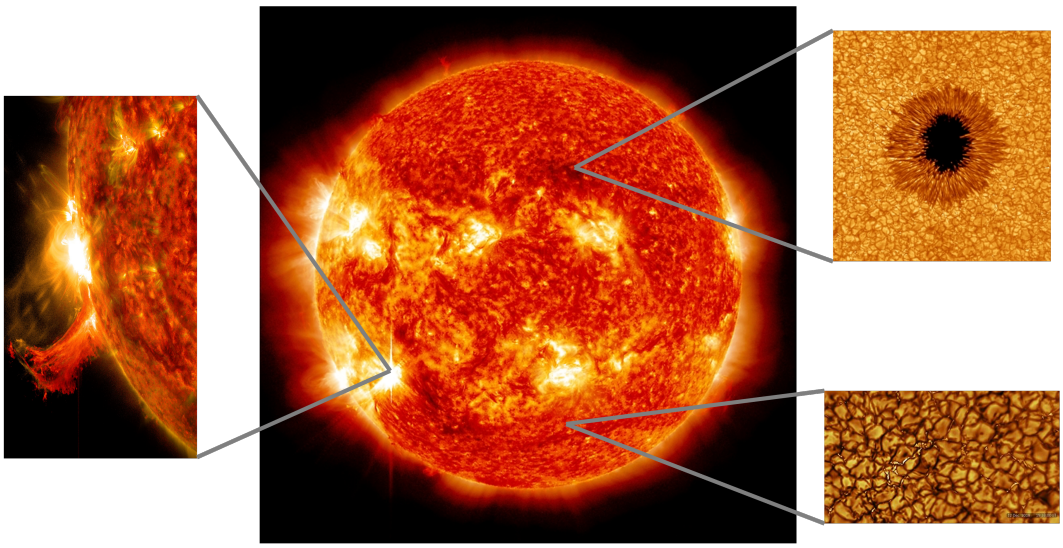

Figure 1.8: Overview of real stellar activity observed on the Sun, illustrating the main stellar features which can mimic or obscure signals of exoplanets in photometric and spectroscopic planet searches. *Left:* Example of stellar flare; *Top right:* Stellar starspot; *Bottom right:* Granulation on the surface of the sun. Images reproduced with permission from NASA, SDO and the Big Bear Solar Observatory.

and variability. While instrumental trends are commonly removed using techniques like cotrending basis vectors [Thompson et al., 2016a], spacecraft pointing-based decorrelation [Vanderburg and Johnson, 2014; Aigrain et al., 2016] or subtracting trends shared by simultaneously observed nearby stars [Kovács et al., 2005; Kim et al., 2009], the problem of dissociating stellar noise from transit signals remains a challenging one. A large array of different methods have been developed in an attempt to solve this problem, which are summarised in Section 2.5. New methods developed in this work to disassociate transiting exoplanetary signals from stellar activity are then introduced in section 2.6.

As an aside, it is also worth noting that while such stellar activity is challenging in photometric transit searches, the presence of this activity is perhaps even more confusing in radial velocity surveys [Saar and Donahue, 1997; Korhonen et al., 2015]. While stellar granulation and pulsations can result in net radial velocity shifts on the order of 1 m/s [Dumusque, 2018], the increased star spot activity for young stars can result in radial velocity shifts of multiple 10 m/s. Although spot modelling codes such as `SOAP` [Dumusque et al., 2014] and `starry` [Luger et al., 2019] are aiding the community to understand the effect of these spots on both spectroscopic and photometric observations, the greatly increased stellar activity of young stars severely hampers radial velocity searches and follow-up, meaning that

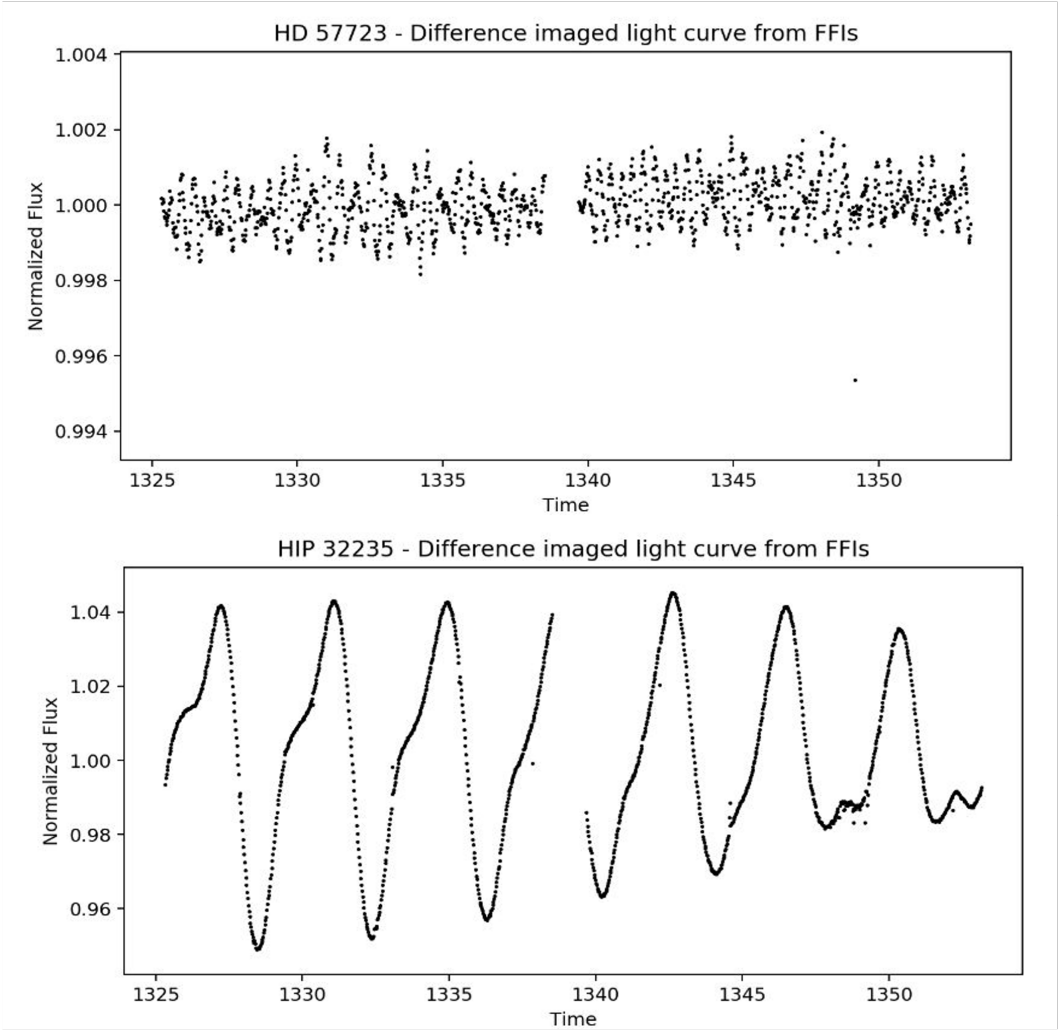

Figure 1.9: Example of challenging stellar activity in photometric light-curves from *TESS*. The upper panel, HD 57723, illustrates pulsation-like stellar variability, while the lower panel, HIP 32235, shows an example of the photometric signal from evolving star spots.

the masses of most young exoplanets are very poorly constrained. Indeed the first mass measurement of an exoplanet from a young open cluster (K2-100 b) was only achieved in 2019 [Barragán et al., 2019b], using techniques similar to the `PYANETI` code [Barragán et al., 2019a], where activity indicators are used to constrain the variability of stars before modelling the spectroscopic data with a latent GP model. Additional techniques to unravel exoplanet radial velocities from such active stars are ongoing [e.g. Rajpaul et al., 2015; Dumusque, 2018; Cretignier et al., 2020]). Given the great importance of mass measurements both to the confirmation and

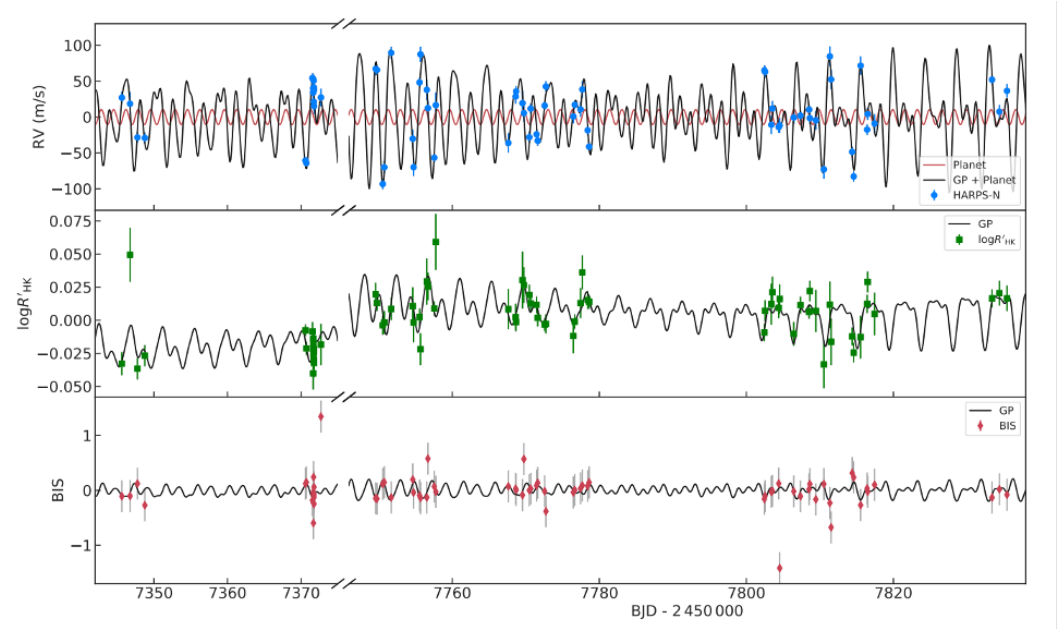

Figure 1.10: Example of challenging RV stellar activity for the young exoplanet-harbouring K2-100 system. Reproduced from Barragan et al. (2019). Shown in the top column of this figure is the Radial Velocity (RV) data (in blue), with the GP + planet fit shown in black and the planetary model shown as a red sinusoid. The lower two columns show two well-known activity indicators, logR'$_{HK}$ (as green dots) and the bisector inverse slope (BIS) of the radial velocity cross correlation function (shown as red dots). GP models for all for these two indicators are shown as black lines. Time is shown on the lower axis, in units of days.

evolutionary analysis of young exoplanets, it is crucial that these stellar activity mitigation techniques in radial velocity observations develop alongside the photometric detrending techniques.

1.4.4 Currently known young exoplanets

Despite these challenges, a small number of young exoplanets have been found. In line with early exoplanet discoveries, the first exoplanets around young stars were found using the radial velocity method, including four Hot Jupiters found in the Hyades [Sato et al., 2007; Quinn et al., 2014] and Praesepe [Quinn et al., 2012] open clusters. However, for planets less massive than Hot Jupiters, the radial velocity method was found to be severely hampered by the inherent stellar variability and radial velocity jitter of the young host stars [Saar and Donahue, 1997; Paulson et al., 2004; Brems et al., 2019]. Recent methods such as those employed by Korhonen et al. [2015]; Rajpaul et al. [2015] have pushed this limit down to approximately Neptune-

sized planets for some solar-like stars, however Earth-sized planets are currently still beyond the capabilities of these methods for young active stars. Because of the intricacies of direct imaging detection (e.g. comparative brightness in the infra-red of young dusty systems), the vast majority of direct imaging detections have also been made around young stars, now totalling almost 60 planets. However, because the youngest planets are typically the brightest in thermal emission, these planets are typically only up to 50 Myr old, leaving 950 Myr of the first billion years of exoplanet evolution largely unexplored. Furthermore, direct imaging targets just a subset of the planetary population (focusing on super-bright nearby star), and finds only those planets in very wide orbits.

The high photometric precision offered by the launch of the *Kepler* satellite proved crucial in increasing this sample of young exoplanets, especially when deliberately pointed at young open clusters in the *K2* extended mission. Most discoveries in this era were made by the *Zodiacal Exoplanets In Time (ZEIT)* team, who have so far found 18 planets since the beginning of their search in 2016 [e.g. Mann et al., 2016a; Rizzuto et al., 2017, 2018]. Other interesting discoveries of young exoplanets in K2 have included K2-33b - a Neptune-sized planet in the 5-10 Myr Upper Scorpius stellar association [David et al., 2016a], K2-136A c in the ∼650 Myr Hyades - the first Neptune-sized planet orbiting a binary system in an open-cluster [Ciardi et al., 2018], and the closely-packed system of four planets around a ∼23 Myr pre-main sequence star in the Taurus-Auriga star forming region [David et al., 2019b,c].

The launch of the *TESS* is now ushering in a new era of exoplanet discovery, including the discovery of several new young exoplanets. The first young exoplanets to be found using *TESS* were the Neptune sized planets DS Tuc Ab in the 45 Myr Tucana-Horologium association [Benatti et al., 2019; Newton et al., 2019], and AU Mic b in the 24 Myr Beta Pictoris Moving Group [Plavchan et al., 2020]. Further *TESS* discoveries include the very young (∼17 Myr) hot Jupiter HIP 67522 [Rizzuto et al., 2020], a two-planet system in the 400 Myr Ursa Major group [Mann et al., 2020] and two young planets around field stars of ages 20 and 320 Myrs [Zhou et al., 2021].

Largely driven by *TESS*, the pace of young exoplanet discovery has accelerated greatly over the course of this research, especially for young exoplanets around field stars. Given the importance of these young exoplanets, all those known at the time of writing are summarised in Tables 1.3 and 1.4, split into those around stars in clusters/associations and around field stars respectively. With the majority of these planets having been found since the start of this research (indeed 57% of the 72 currently known young exoplanets were discovered after this research began, with

Table 1.3: Overview of currently known young transiting exoplanets in **clusters and associations** as of August 2022. Note that exoplanets were only included if their quoted age plus maximum error fell below 1Gyr. All data was retrieved from the NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu; retrieved 14 August 2022), except for the most recently validated young exoplanet (TOI-2048), which was received from its discovery paper [Newton et al., 2022]. At present, only 7 planets in clusters and associations have measured planetary masses, so these were omitted in this instance.

Planet Name	Cluster	Stellar age Myr	Stellar Mass M_{Sun}	Planet Radius $R_{\oplus}$	Period Days	Discovery Ref.
AU Mic b	Beta-Pic	22^{+3}_{-3}	0.5	4.2	8.463	Plavchan et al. [2020]
AU Mic c	Beta-Pic	22^{+3}_{-3}	0.5	2.79	18.859	Gilbert et al. [2022]
DS Tuc A b	Tuc-Hor	40^{+5}_{-5}	0.96	5.63	8.139	Benatti et al. [2019]
EPIC 211822797 b	Praesepe	790^{+30}_{-30}	0.61	2.2	21.170	Mann et al. [2017a]
HD 63433 b	Ursa Major	414^{+23}_{-23}	0.99	2.15	7.108	Mann et al. [2020]
HD 63433 c	Ursa Major	414^{+23}_{-23}	0.99	2.67	20.545	Mann et al. [2020]
HD 110082 b	MELANGE-1	250^{+50}_{-70}	1.21	3.2	10.183	Tofflemire et al. [2021]
HIP 67522 b	Sco-Cen	17^{+2}_{-2}	1.22	10.07	6.960	Rizzuto et al. [2020]
HIP 94235 b	AB Dor	118^{+18}_{-15}	1.09	3	7.713	Zhou et al. [2022]
K2-25 b	Hyades	730^{+50}_{-52}	0.26	3.44	3.484	Mann et al. [2016a]
K2-33 b	Upper Sco	9^{+1}_{-1}	0.56	5.04	5.425	David et al. [2016b]
K2-95 b	Praesepe	790^{+30}_{-30}	0.43	3.7	10.135	Mann et al. [2017a]
K2-100 b	Praesepe	790^{+30}_{-30}	1.18	3.5	1.674	Mann et al. [2017a]
K2-101 b	Praesepe	790^{+30}_{-30}	0.8	2	14.677	Mann et al. [2017a]
K2-102 b	Praesepe	790^{+30}_{-30}	0.77	1.3	9.916	Mann et al. [2017a]
K2-103 b	Praesepe	790^{+30}_{-30}	0.59	2.2	21.170	Mann et al. [2017a]
K2-104 b	Praesepe	790^{+30}_{-30}	0.51	1.9	1.974	Mann et al. [2017a]
K2-136 b	Hyades	688^{+63}_{-63}	0.71	0.99	7.975	Mann et al. [2017b]
K2-136 c	Hyades	688^{+63}_{-63}	0.71	3.03	17.308	Ciardi et al. [2018]
K2-136 d	Hyades	688^{+63}_{-63}	0.71	1.45	25.575	Mann et al. [2017b]
K2-264 b	Praesepe	790^{+30}_{-30}	0.471	2.231	5.840	Rizzuto et al. [2018]
K2-264 c	Praesepe	790^{+30}_{-30}	0.471	2.668	19.660	Rizzuto et al. [2018]
K2-1627 A b	Delta Lyra	38^{+6}_{-5}	0.95	0.338	7.203	Bouma et al. [2022]
Kelt-25 b	Theia 449	162	2.18	317.78	4.401	Martínez et al. [2020]
Kepler-970 b	MELANGE-3	105^{+10}_{-10}	0.67	2.609	16.737	Barber et al. [2022]
Kepler-1928 b	MELANGE-3	105^{+10}_{-10}	1.01	1.992	19.578	Barber et al. [2022]
TOI-451 b	Pisces-Eridanus	120	0.95	1.91	1.859	Newton et al. [2021]
TOI-451 c	Pisces-Eridanus	120	0.95	3.1	9.193	Newton et al. [2021]
TOI-451 d	Pisces-Eridanus	120	0.95	4.07	16.365	Newton et al. [2021]
TOI-837 b	IC 2602	35^{+11}_{-5}	1.12	8.631	8.325	Bouma et al. [2020]
TOI-1227 b	Lower Cen	11^{+2}_{-2}	0.17	9.572	27.364	Mann et al. [2022]
V1298 Tau b	Taurus	23^{+4}_{-4}	1.1	10.22	24.139	David et al. [2019a]
V1298 Tau c	Taurus	23^{+4}_{-4}	1.1	5.59	8.250	David et al. [2019c]
V1298 Tau d	Taurus	23^{+4}_{-4}	1.1	6.41	12.403	David et al. [2019c]
V1298 Tau e	Taurus	23^{+4}_{-4}	1.1	8.74	60	David et al. [2019c]

Table 1.4: Overview of known young transiting exoplanets around **field stars** as of August 2022. Once again, exoplanets were only included if their quoted age plus maximum error reliably fell below 1Gyr. Note however that the ages of field stars are generally considered less reliable as their ages are based on less-precise gyrochronology and stellar models rather than clusters/associations with well-defined ages (as can be seen from the larger errors in age compared to the cluster/association table). Because of this, planets also required an independent age assessment in their literature to be included here. All data retrieved from the NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu; retrieved 14 August 2022).

Planet Name	Stellar Age	Stellar Mass	Planet Radius	Planet Mass	Period	Discovery Ref.
	Myr	M_{Sun}	$R_{\oplus}$	$M_{\oplus}$	Days	
CoRoT-18 b	600^{+0}_{-600}	0.95	14.684	1099.6918	1.900	Hébrard et al. [2011]
HAT-P-70 b	600^{+380}_{-200}	1.89	20.961	2154.8874	2.744	Zhou et al. [2019]
HATS-58 A b	310^{+330}_{-200}	1.46	12.274	327.3649	4.218	Espinoza et al. [2019]
HATS-67 b	510^{+240}_{-240}	1.44	18.887	460.8535	1.609	Hartman et al. [2019]
HD 73583 b	480^{+190}_{-190}	0.73	2.79	10.2	6.398	Barragán et al. [2022]
HD 73583 c	480^{+190}_{-190}	0.73	2.39	9.7	18.880	Barragán et al. [2022]
K2-233 b	360^{+490}_{-140}	0.80	1.398	-	2.467	David et al. [2018b]
K2-233 c	360^{+490}_{-140}	0.80	1.335	-	7.061	David et al. [2018b]
K2-233 d	360^{+490}_{-140}	0.80	2.64	-	24.366	David et al. [2018b]
K2-284 b	120^{+640}_{-20}	0.63	2.78	-	4.795	David et al. [2018a]
KELT-17 b	650^{+150}_{-150}	1.64	17.094	416.3573	3.080	Zhou et al. [2016]
Kepler-51 b	500^{+250}_{-250}	0.98	6.89	3.69	45.154	Steffen et al. [2013b]
Kepler-51 c	500^{+250}_{-250}	0.98	8.98	4.43	85.313	Steffen et al. [2013b]
Kepler-51 d	500^{+250}_{-250}	0.98	9.46	5.7	130.185	Masuda [2014]
Kepler-78 b	625^{+150}_{-150}	0.83	1.2	1.87	0.355	Sanchis-Ojeda et al. [2013b]
Kepler-411 b	212^{+31}_{-31}	0.87	2.401	25.6	3.005	Wang et al. [2014]
Kepler-411 c	212^{+31}_{-31}	0.87	4.421	26.4	7.834	Morton et al. [2016]
Kepler-411 d	212^{+31}_{-31}	0.87	3.319	15.2	58.020	Sun et al. [2019]
Kepler-462 b	500^{+460}_{-460}	1.59	3	53.9	84.687	Morton et al. [2016]
Kepler-462 c	500^{+460}_{-460}	1.59	4	6	207.620	Masuda and Tamayo [2020]
Qatar-3 b	310^{+1}_{-1}	1.15	12.285	1369.8473	2.508	Alsubai et al. [2017]
Qatar-4 b	170^{+10}_{-10}	0.90	12.720	1939	1.805	Alsubai et al. [2017]
Qatar-5 b	530^{+4}_{-4}	1.13	12.408	1373.0256	2.879	Alsubai et al. [2017]
TOI-251 b	180^{+140}_{-140}	1.04	2.74	317.83	4.938	Zhou et al. [2021]
TOI-942 b	50^{+30}_{-20}	0.88	4.242	16	4.326	Carleo et al. [2021]
TOI-942 c	50^{+30}_{-20}	0.88	4.793	37	10.161	Carleo et al. [2021]
TOI-1268 b	245^{+135}_{-135}	0.96	9.1	96.4	8.158	Dong et al. [2022]
TOI-1431 b	290^{+320}_{-190}	1.90	16.701	991.62463	2.650	Addison et al. [2021]
TOI-1807 b	300^{+80}_{-80}	0.76	1.37	2.57	0.549	Hedges et al. [2021]
TOI-1860 b	133^{+26}_{-26}	0.99	1.31	-	1.066	Giacalone et al. [2022]
TOI-2046 b	450^{+430}_{-21}	1.13	16.141	731.00534	1.497	Kabáth et al. [2022]
TOI-2076 b	204^{+53}_{-50}	0.85	2.5	-	10.356	Hedges et al. [2021]
TOI-2076 c	204^{+53}_{-50}	0.85	3.38	-	-	Hedges et al. [2021]
TOI-2076 d	204^{+53}_{-50}	0.85	3.16	-	-	Hedges et al. [2021]
TOI-2260 b	321^{+96}_{-96}	0.99	1.62	-	0.352	Giacalone et al. [2022]
WASP-178 b	430^{+310}_{-250}	1.93	21.745	448.1403	3.345	Hellier et al. [2019]
WASP-189 b	730^{+130}_{-130}	2.03	18.147	632.47853	2.724	Lendl et al. [2020]

10 found this year alone), it is clear that it is now possible to find planets across a wide range of the early evolution of exoplanetary systems. However, considerably more planets are needed at all ages less than 1 Gyr before reliable statistics can be generated for each epoch and the full early evolution of exoplanets can be understood. To find these planets the community is now turning to two new satellites driving the all-sky space telescope revolution: *Gaia* and *TESS*.

1.5 The all-sky space telescope revolution

1.5.1 *Gaia*

The European Space Agency's (ESA) *Gaia* Mission [Gaia Collaboration et al., 2016b] aims at generating a 3D map of the Milky Way by measuring the position, parallax and velocities of over a billion stars (and some other astrophysical objects) within the galaxy. This represents approximately 1% of the entire galactic stellar population, and promises to provide significantly increased understanding about the structure, formation and evolution of our galaxy. For stars with magnitudes of approximately 5.7-20, it is anticipated that full astro- and photometric data will be measured, as well as radial velocities for objects up to 17th magnitude [Gaia Collaboration et al., 2016a]. The *Gaia* Spacecraft launched on 19th December 2013, and began its nominal 5-year science operations phase on 25 July 2014. Situated in a Lissajous-type orbit around L2, the spacecraft itself has two identical telescopes utilised by three primary instruments (one each for astrometry, high resolution spectrography in the region 845-72nm and photometry) which share a common focal plane on a 0.5 x 1.0m CCD array [Gaia Collaboration et al., 2016a].

Initial data from the mission has been made public already in two separate data releases controlled by the *Gaia* Data Processing and Analysis Consortium (DPAC) - DR1 (Gaia Collaboration et al., 2016a) and DR2 (Gaia Collaboration et al., 2018). DR2 is particularly of note for this research, as it provided positions and g-magnitudes for 1.7 billion stars, high precision proper motions, parallaxes and BP-RP colours for over 1.3 billion stars, and even radial velocities for 7 million of them. Such unprecedented precision over such a large part of our galaxy is already resulting in more in-depth testing of stellar association membership and is greatly increasing the number of known members, especially towards later classification stars.

There are however some important considerations for the use of *Gaia* data. First of all, Bailer-Jones et al. [2018] found that simply inverting the measured parallaxes yielded unreliable distances in some cases, due mostly to the non-linearity

in the parallax-to-distance transformation and the requirement that distance must be positive. Furthermore, many *Gaia* parallaxes have relatively low signal to noise, especially for dimmer stars or those in more crowded fields. Bailer-Jones instead provides a Bayesian framework for determining more reliable distances and has released a catalogue (Vizier catalogue number *I/347*) for these distances for every star in DR2. It is these distances that are used in this work. On the other end of the scale, brighter sources (G < 12) require the removal of the *Gaia* DR2 proper motion inertial spin, which is basically the precession set up in the Gaia satellite's orbit due to the interaction between its rotation and changing inertial tensor [Lindegren et al., 2018]. Finally, recent work by groups like Damiani et al. [2019a] and Faherty et al. [2017] recommend only using stars with signal to noise levels above 10 in terms of parallax in order to remove stars with poor kinematics. Paying attention to these considerations improves the quality of the *Gaia* data used, and improves the accuracy of any conclusions drawn.

1.5.2 TESS

The Transiting Exoplanet Survey Satellite (*TESS*), is a new space-based satellite searching for transiting exoplanets around nearby bright stars, whose primary aim is to identify hundreds of transiting planets smaller than Neptune around host stars bright enough to allow effective characterisation of planetary masses and atmospheres [Ricker et al., 2014]. TESS began science operations on 25 July 2018 [Huang et al., 2018], and over the course of its 2 year primary mission (ending July 4th 2020) surveyed >75% of the sky, broken into 13 sectors in each of the Northern and Southern Hemispheres. These pointings are illustrated in Figure 1.11. The observing time for each star varies by latitude, but ranges between 27.4 days to 1 year. TESS has four cameras, each with 2048 x 2048 pixel imaging array and a field of view of 24^o yielding an angular resolution of 21 arcseconds [Ricker et al., 2014]. Following the success of this primary mission, *TESS* began its extended mission, for which it is currently in Year 4. In this mission *TESS* is aiming to cover almost the entire sky, with Year 3 returning to the Southern Hemisphere and shifted somewhat to cover gaps in the Year 1 observations, and Year 4 completing a mix of ecliptic latitudes and half the Northern Hemisphere again. Year 5 will complete the second scan of the Northern Hemisphere, before beginning another scan of the South.

Two main data products were produced by the *TESS* primary mission: 2min cadence light-curves (detrended and raw) and 30min cadence Full-Frame-Images (FFIs). 2min cadence light-curves are generated by Science Processing and Operations Center (SPOC) pipeline [Jenkins et al., 2016] for approximately 300,000 of

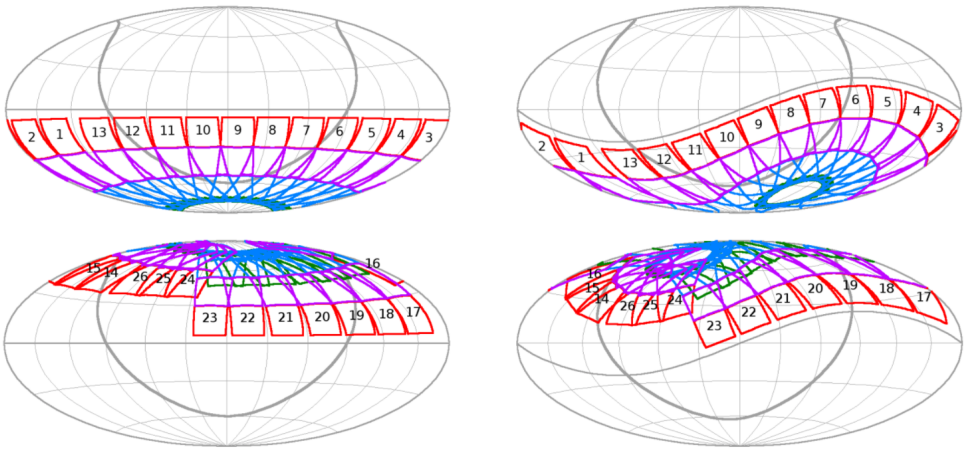

Figure 1.11: *TESS* pointings during the primary mission, in ecliptic (left) and celestial coordinates (right). Year 1 pointings and shown above and Years 2 pointings below. Note that the shift of sectors 14-16 and 24-26 in the Northern Hemisphere was enacted to avoid excessive scattered light contamination from the Earth and the Moon. Reproduced from https://heasarc.gsfc.nasa.gov/docs/tess/primary.html

the most promising stars per cycle (prioritised by the smallest transiting planets that can be detected) from the *TESS* Candidate Target List [CTL, Stassun et al., 2018, 2019], and are then made accessible to the public via the Mikulski Archive for Space Telescopes (MAST[9]). Different methods for extracting 30min-cadence light-curves from the FFIs are discussed in detail in Section 2.2. In the extended mission (beginning from Cycle 3 observations), Full Frame Images were updated to a 10min cadence, with up to 1000 individual targets per sector additionally observed with 20s cadence. By the end of February 2022, *TESS* had produced 197 confirmed exoplanets and over 5000 exoplanet candidates.

1.6 Machine Learning

One of the biggest challenges of these all-sky surveys is the large volume of data which needs to be analysed. Such large volumes make human eyeballing and target-by-target analysis unwieldy, lending themselves to more automated approaches like Machine Learning instead. Here a brief introduction to the use of machine learning in astronomy is given, before a more focused look at variability classification methods and Self-Organising Maps which are particularly important for this work.

[9]https://mast.stsci.edu/portal/Mashup/Clients/Mast/Portal.html

1.6.1 Machine Learning in Astronomy

Over the last two decades machine learning has become increasingly popular among astronomers, driven by the rapid expansion in data availability from such surveys as the Sloan Digital Sky Survey [SDSS; York et al., 2000], the Zwicky Transient Facility [Bellm, 2014] and the current *Gaia* all-sky survey [Gaia Collaboration et al., 2016b]. Machine learning offers tools and techniques to complete a wide range of analyses from classification or clustering to regression, dimensionality reduction and outlier detection. There are two main types of machine learning: 'supervised' and 'unsupervised'. Supervised machine learning relies on pre-classified classes (with known features) and training data, hence is very useful for classification of data into known groups (e.g. galaxy classification), but introduces human bias into the analysis. On the other hand, unsupervised machine learning methods look for patterns and relationships within an unseen dataset, without providing pre-defined labels from a human expert. Baron [2019] argues that unsupervised machine learning methods are particularly useful for astronomy as they can uncover new information in existing datasets and lead to new discoveries.

Supervised methods typically consist of three stages: training, validation and testing. Under this framework a human user typically splits an input dataset into a training set, validation set and testing set, each pre-sorted into known groups. To begin with, model hyper-parameters are set (such as the number of features used to sort data into different groups), before the model and model parameters are learnt through application to the training dataset. During the validation stage the model hyper-parameters are then tuned and optimised according to a set 'cost-function', which defines how well the model is fitting the data. This stage is typically carried out iteratively (for potentially thousands of iterations of parameters) until the optimum model performance is reached for the validation dataset. Finally, the chosen model is tested by applying it to the unseen training dataset in order to check the final model's performance compared to the known ground truth. Providing this testing step is successful, the model can then be applied to completely new datasets, for example for classifying new light-curves into known variability classes. Common supervised methods used within astronomy include Support Vector Machines [e.g. Qu et al., 2003; Pashchenko et al., 2018], Decision Trees/Random Forest methods [e.g. Breiman, 2001; Ball et al., 2006] and Artificial Neural Networks [e.g. Storrie-Lombardi et al., 1992; Cantat-Gaudin et al., 2020].

Meanwhile unsupervised machine learning methods are more varied in process, encompassing a wide array of different algorithms. This means that they have found use in applications as varied as studying spectra from asteroids [Rizos et al.,

2021] to characterising instrumental trends in photometric transit studies [Jenkins et al., 2010] within astronomy. However because unsupervised algorithms do not start with pre-labelled groups and may be affected by chosen external parameters, some caution is needed in analysing the final outputs of these algorithms [Baron, 2019]. Common types of unsupervised machine learning methods in astronomy include clustering algorithms (e.g. K-means [MacQueen, 1967] and Hierarchical clustering [Ward, 1963]), dimensionality reduction algorithms (e.g. Principal Component Analysis [e.g. Deeming, 1964; Bailer-Jones et al., 1998], t-Distributed Stochastic Neighbour Embedding (tSNE; van der Maaten and Hinton [2008]), Autoencoders [e.g. Gianniotis et al., 2015; Schawinski et al., 2018] and Self Organising Maps [Kohonen, 1982, 2001]) and Anomaly Detection algorithms [e.g. Meusinger et al., 2012; Giles and Walkowicz, 2019]. A detailed overview of unsupervised methods used in astronomy can be found in Baron [2019].

Given this wide range of machine learning techniques, a full review of machine learning applications within astronomy is beyond the scope of this brief introduction. Instead the reader is directed to recent wider reviews on the subject by Ivezic et al. [2019] and Baron [2019] or the collection of machine learning tools offered within the Python-based sci-kit learn package [Pedregosa et al., 2011].[10] Of particular interest to this work however is the challenge of using machine learning to analyse and classify photometric variability, which is the focus of the remainder of this introduction.

1.6.2 Variability classification

Large-scale stellar variability classification has been a subject of interest since the first large-scale surveys such as the Hipparcos mission[11], Optical Gravitational Lensing Experiment (OGLE[12]), and All-Sky Automated Survey (ASAS[13]) began making traditional human eyeballing untenable. Most of the earliest variability classifiers focused on supervised methods, labelling distinct variability groups and 'features' before applying versions of neural networks [Pojmanski, 2002] or decision trees [Ball et al., 2006]. Similarly, Debosscher et al. [2007], Sarro et al. [2009] and Debosscher et al. [2011] built a pipeline based on Gaussian Mixtures and Bayesian networks to conduct global variability classification (and training) on data from CoRoT [Auvergne et al., 2009], OGLE and Kepler Quarter 1. Richards et al. [2011] returned to the decision-tree (or Random Forest) method to classify stars from OGLE and Hipparcos, eventually using these as a training set to construct the large Machine-

[10] https://scikit-learn.org/stable/index.html
[11] https://www.cosmos.esa.int/web/hipparcos
[12] http://ogle.astrouw.edu.pl/
[13] http://www.astrouw.edu.pl/asas/

learned ASAS Classification Catalogue [MACC Richards et al., 2012]. The versatility of retraining supervised machine learning methods have also allowed them to be applied to slightly more unusual types of variability, including for the discovery of transients in synoptic survey imaging data [Brink et al., 2013], detecting anomalous light curves in massive astronomical catalogues [Nun et al., 2014] and even on MACHO (MAssive Compact Halo Objects) and LINEAR (LincoIn Near-Earth Asteroid Research) data Kim and Bailer-Jones [2016].

Unsupervised methods on the other hand allow their algorithms to discover information and clusters in the data without pre-decided labels or groups. This can lead to new data groups emerging and is useful for situations in which not all potential solutions are known. However, these methods are comparatively rare in stellar classification analyses. One of the earliest applications of such a method was completed by Eyer and Blake [2005], using an unsupervised Bayesian classifier on features from Fourier-space to classify variable stars in ASAS data releases 1-2. In another direction, Valenzuela and Pichara [2018] created a 'Variability Tree' which allowed simultaneous feature generation and variability search to be completed concurrently on OGLE, MACHO and *Kepler* data. Others such as Modak et al. [2020] and Armstrong et al. [2016] have turned to clustering methods instead, allowing similar variability shapes or light-curves to be placed in similar clusters. This work introduces a new wider approach, applying Self-Organising Maps (discussed further in section 1.6.3) for the first time to a dedicated sample of young star data.

More recently, there has been a drive to combine the different benefits provided by both unsupervised and supervised methods into a single classification. Naul et al. [2018] for example combined an unsupervised autoencoded recurrent neural network to find features within their dataset with a supervised random forest to make the final classifications for optical variable star catalogues. Meanwhile the large *TESS* Data for Asteroseismology (T'DA) collaboration is currently working on a large stellar variability classification pipeline [Audenaert et al., 2021], or 'meta-classifier', which combines the Multiclass Solar-like Oscillation Shape Hunter [multiSLOSH Hon et al., 2018] neural network, Random Forest General Classification [RFGC, a hybrid unsupervised Self-Organising Map and Random Forest implementation developed by Armstrong et al., 2016], Supervised Random Forest Variability Classifier Using High-resolution Photometry Attributes in TESS Data [SORTING-HAT Audenaert et al., 2021] and Gradient Boosting General Classification [GBGC Friedman, 2001] methods. Furthermore, the *Gaia* mission collaboration applies a wide variety of classification pipelines depending on the type of variability, with classifications continuing to be improved with the ongoing mission [Gaia Collaboration

et al., 2016b, 2018a; Eyer et al., 2019]

However, despite these widespread approaches to general stellar variability analyses, machine-learning classification of variability in young stars remains in its comparative infancy. Due to challenges like swiftly evolving and unusually shaped activity, coupled with a traditional dearth of bona-fide young stars [Gagné et al., 2018b], most classification of young star light-curves has been based on human eye-balling [e.g. Cody and Hillenbrand, 2018]. However, as young stars exhibit a breadth and depth of stellar variability often not seen in older populations, there is significant value in completing variability classification independent of the older population. Thankfully the improved performance and coverage of the new all-sky *TESS* and *Gaia* missions is now beginning to supply the data necessary for young-star specific analyses, making this a useful time to begin such analyses. Indeed Hedges et al. [2018] have recently shown the power of young-star specific classification in a search for for dippers/bursters found from supervised machine learning.

1.6.3 Self-Organising Maps

Self-Organising Maps (SOMs; aka Self-Organizing Maps) are an unsupervised machine-learning technique originally developed by Kohonen [1982, 2001] which allows for sorting of data based on topology, or similar shapes. A form of artificial neural-network birthed from research into neural networks in the human brain [Kohonen, 1982], SOMs use comparisons between the shape of input data and the shape of individual data-arrays across the SOM structure to group similar input data into similar regions of a 'map'. At each iteration the input data is moved towards the 'pixel' on the map which contains a data array most similar to its own shape, before the data-array at that pixel is changed to be an average of all input arrays which ended up there. As the iterations continue, the area which input data shapes are compared to is decreased, reducing movement with time and collecting those most similar together. The final map, usually 1-2 dimensions for clarity, can then be used to explore groups of similar data or to more simply investigate the evolution of a complex data-set in a form of dimensionality reduction [Armstrong et al., 2016]. One key advantage of Self-Organising Maps is their unsupervised nature, meaning that underlying trends and groups in the input data can be explored without the need for pre-allocated classes of variability or significant knowledge about the target stars. This makes them particularly useful for young star studies, where new classes of photometric variability are still being discovered and understood [e.g. Zieba et al., 2019; Tajiri et al., 2020]. A useful extension to this behaviour is that not only are those data which are most similar placed closest together in the final map, but those

which are most different are placed furthest apart [Mostert et al., 2021]. However, an important proviso with the use of the SOM is that far more importance should be given to the relative position of the analysed data compared to its overall distribution, since, as noted by Geach [2012], without a seed the randomness of the initial SOM setup results in a different final distribution on each new run, even for precisely the same input data-set. This means that while the same input data should always end up closest to those data which are most similar, the overall shape of the SOM may be different in each run depending on how the SOM was initialised.

SOMs are particularly popular in computer science, quantitative biology and statistics [e.g. Rahmani et al., 2019; Burnap et al., 2018], however until recently have been under-utilised in astronomy. Brett et al. [2004] were the first to use this technique for astrophysical data, using self-organising maps to classify a set of synthetic light-curves and 1026 real light-curves from the ROTSE experiment [Akerlof et al., 2000]. As well as providing an overview of how the SOM evolves over different numbers of iterations, Brett et al. [2004] illustrated that the performance of the SOM was relatively insensitive to changes in the value and decay shape of the two main SOM parameters: the learning rate α and effective neighbourhood size σ. This makes it a versatile and easy approach to use for a wide range of data-sets.

More recently Armstrong et al. [2016] extended this work by using a SOM alongside a supervised Random Forest implementation to classify the variability (or lack thereof) of all stars in Campaigns 0-4 of the K2 mission. In this work Armstrong et al. [2016] used a 2D 40x40 Kohonen layer with an initial learning rate of $\alpha_0 = 0.1$ $\sigma_0 = 40$ (the width of the Kohonen layer) to clearly demonstrate the utility of the SOM for splitting types of photometric variability by training the SOM on a set of 4047 known variables from campaigns 0-2 of the *K2* mission. The trained SOM with known variables placed on it is reproduced in Figure 1.12. An improved version of this same SOM setup has recently been folded into the wider *TESS* Asteroseismic Science Consortium (TASOC) stellar variability classifier, combined with a small selection of other supervised and unsupervised classification methods [Audenaert et al., 2021], however it has never previously been tested on the unique groups of variability seen in young stars. In related works Armstrong et al. [2016, 2020a] have also used SOMs to classify transit shapes in an effort to aid automated vetting of planetary transits.

In wider astronomy SOMs have been used for a wide array of analyses, from Photometric Redshift calibration [Geach, 2012; Carrasco Kind and Brunner, 2014; Wright et al., 2020b,a], Radio astronomy [Galvin et al., 2020; Mostert et al., 2021] and spectroscopic image segmentation [Fustes et al., 2013; Schilliro and Romano,

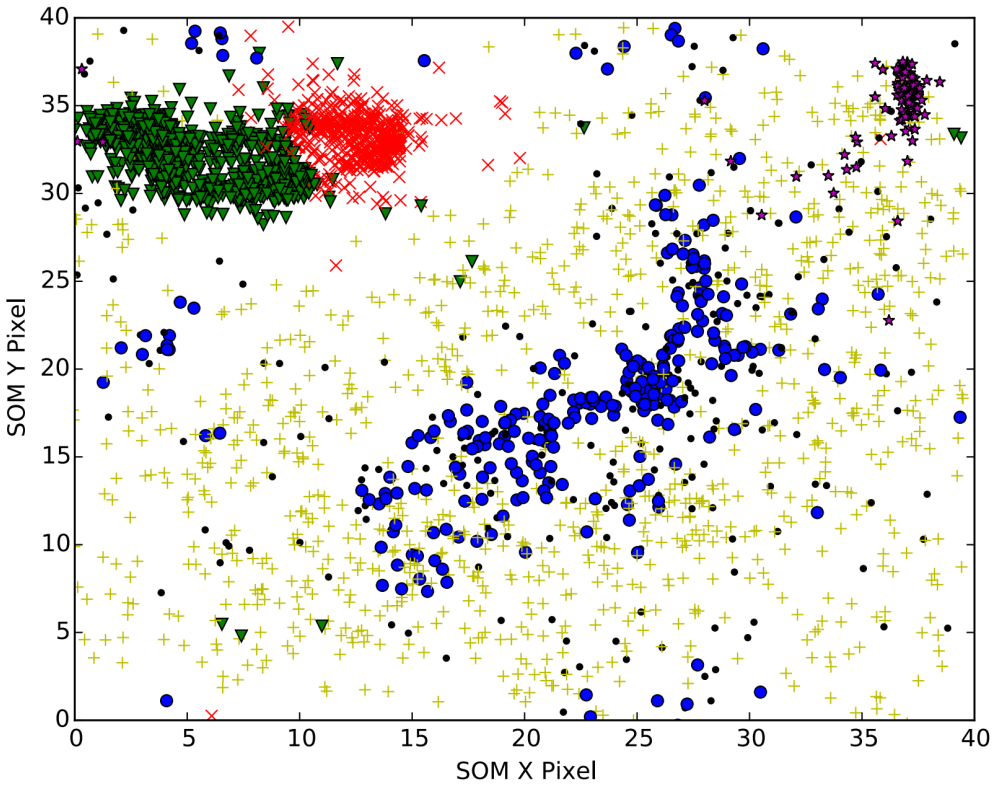

Figure 1.12: Self-Organised Map trained on known variables from *Kepler* and the initial K2 variability catalogue of [Armstrong et al., 2015]. Green triangles = Algol-type eclipsing binaries (EAs); red crosses = Beta-lyrae eclipsing binaries (EBs), pink stars = ab-type RR Lyraes; blue circles = Delta-Scuti variables; black dots = Gamma Doradus variables & yellow pulses = Other periodic/quasi-periodic objects. Reproduced from Figure 2 of Armstrong et al. [2016].

2021] to classification of Galaxies [Naim et al., 1997; Johnston et al., 2021], AGNs [Faisst et al., 2019] and unusual quasars [Meusinger et al., 2012]. Of particular note were the discoveries of Johnston et al. [2021] who showed how effective SOMs can be at identifying systematics alongside the main data-groups (in their case in a study of galaxy clustering in the Kilo-Degree Survey), and Khacef et al. [2020]'s suggestion that the efficiency of SOMs could be improved for very large data-sets by applying the SOM to extracted features rather than the raw data. Although many of these wider applications are less relevant to the current work, the wide range of uses highlights the versatility of the SOM for complex data, alongside the ability to classify a wide array of interesting behaviour into groups. This versatility, coupled with the unsupervised clustering algorithm, intuitive results format, and promising

track record for photometric variability classification, make the Self-Organising Map an appealing technique for exploring and classifying young star variability.

1.7 Thesis Outline

The remainder of this thesis details work tracing exoplanets through time with *TESS*, both through variability analyses and planet-searches within light-curves of young stars (Chapters 3 to 5) and through examining how the ephemerides of known exoplanets from the *Kepler* mission have evolved since they were last observed (Chapter 6). This begins with an examination of the methods used when conducting this research in Chapter 2, before moving on to the main science chapters in Chapters 3 to 6. Chapter 3 details work completed to extend the known population of young stars, both through new kinematic analyses and through assembly of young star target lists from the community. Chapter 4 presents the application of a new young star detrending pipeline (presented in Section 2.6) to stars in stellar associations from the first five sectors of *TESS* data. Chapter 5 extends this work by beginning the YOUNGSTER program, using Self-Organising Maps to analyse the variability of young stars from the first full year of *TESS* data, with the view to use this information to construct targeted young exoplanet detrending/searches in the future. Finally, Chapter 6 presents a different tack, updating the ephemerides for all *Kepler* planets reobserved by *TESS* during the second year of its primary mission, including a small number of known young exoplanets. This thesis concludes with a thesis summary, future outlook and closing statements in Chapter 7.

Chapter 2

Methods

"Though this be madness,
yet there is method in't"

— William Shakespeare, *Hamlet*,
Act 2, Scene 2

This chapter describes the main methods and data sources used to complete this research. It also describes in detail the new young star detrending and exoplanet search pipeline developed in 'A search for young exoplanets in Sectors 1-5 of the TESS Full-Frame-Images' [Battley et al., 2020] in Section 2.6.

Declaration - The following chapter includes partial reproduction of the methods sections within Battley et al. [2020], Battley et al. [2021] and [Battley et al., 2022].

2.1 Assembling Young Star Target Lists

2.1.1 Kinematic analyses

As explained in Section 1.4.2, a key place to find galactic stars is in young clusters, moving groups and stellar associations. In many cases the higher-mass membership of these groups have been know for decades [e.g. de Zeeuw et al., 1999] due to their comparative brightness, but the impressive astrometric precision of the *Gaia* satellite is now allowing the population of the lower-mass members of these groups to be explored as well. A number of different methods have been used to discover and extend the membership of these groups/associations, considering factors ranging from the proper motion of objects to approximate ages of members.

The most crucial parameter for membership tests is the common motion of

stellar members, where stars appear to move across the sky with similar velocities. Early extensions to known stellar associations such as the census of nearby OB associations performed by de Zeeuw et al. [1999] focused on the 2D proper motions in the galactic (l,b) coordinate system based on *Hipparcos* data, a technique still used as part of modern membership tests [Damiani et al., 2019b; Gagné et al., 2018b]. An example of using this method to narrow down potential candidates is illustrated in Figure 2.1, showing how limits on the 2D proper motions of stars can be used to find new stars associated with the known members of the Sco OB2 association.

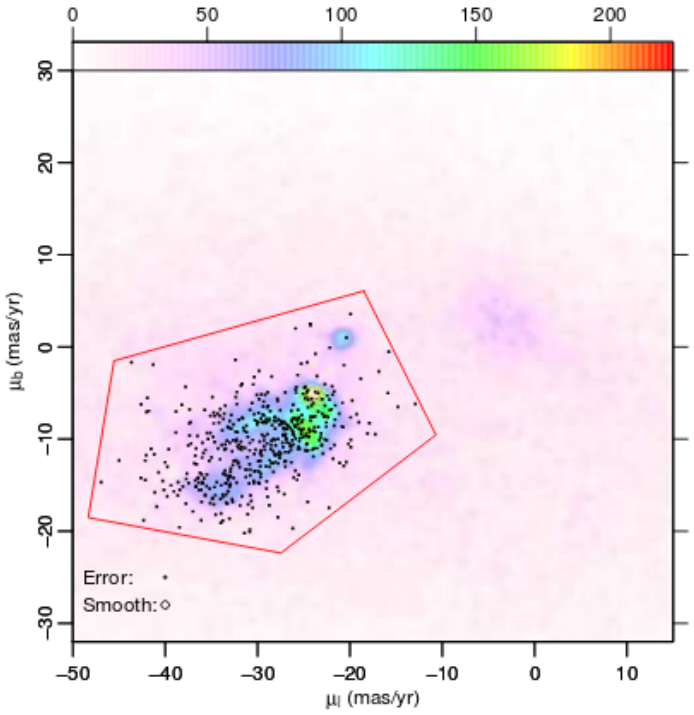

Figure 2.1: Example of imposing proper motion limits to narrow down membership candidate selection for the Sco OB2 stellar association, reproduced from Damiani et al. [2019b]. Density of stars in the region is represented by colour, in units of sources/(mas/yr)^2. Black dots represent known Hipparcos Sco OB2 members from de Zeeuw et al. [1999]

However, such a method used in isolation can suffer from introducing interlopers at quite different distances or radial velocities from the true association members. It is thus also critically important to consider (where possible) the parallax and radial velocity of the candidate members, in order to test for stellar cluster/association membership. One method of achieving this is through simple parallax (or distance) limits applied to the raw data, as used by de Zeeuw et al. [1999]

and Damiani et al. [2019b] for the Sco-OB2 association. Following the release of *Gaia* DR2 [Gaia Collaboration et al., 2018a,b], Bailer-Jones et al. [2018] released a robust catalogue of distances for all objects in the data release which is particularly useful for imposing these limits.

Where radial velocity data is available for objects, an even more instructive method can be converting the proper motions to velocities within the galaxy itself. These velocities are denoted as U, V and W, representing velocities in the X, Y and Z directions of the heliocentric galactic coordinate system. In this reference frame young stellar associations have frequently been observed to have internal velocity dispersions of less than 3km/s [Gagné et al., 2018b], which provides another useful tool for testing true association membership. However, Kuhn et al. [2019] noted that even within gravitationally bound star clusters there is still a surprising range in kinematic properties for member stars (over 1-3km/s in places), so it is clear that care is needed when determining the galactic motion limits. Indeed Gagné et al. [2018c] choose to use a 5km/s velocity dispersion cut-off to ensure greater capture of candidate members with more unusual velocities.

An alternative, geometric method for testing membership to an association or cluster, called the 'Convergent Point Method', was originally developed by Brown [1950]; Jones [1971]. This method allows for the determination of the convergent point of a related group of stars and as an extension a probabilistic measure for how likely each star is likely to belong to this group. Figure 2.2 shows the theory behind this method - stars with common space motions (so likely belonging to the same cluster/association) will in equatorial space appear to converge towards a common 'convergent point' when their proper motions are overlaid. By solving for the most likely convergent point and calculating the sum of squares for each star's tangential component of their motion, one can gain a probabilistic value for the likelihood that each of the stars belongs to the association. de Zeeuw et al. [1999] used this method to complete a census of the stellar content of 14 nearby OB associations based of Hipparcos data. In the pre-*Gaia* era this method was used quite commonly for studies of clusters and associations, including by Mamajek [2005]; Torres et al. [2006]; Rizzuto et al. [2011].

Positional limits can be another useful tool to narrow down candidate members further, however require caution when used close associations, where sky coverage in excess of 100 deg^2 has been observed [Armstrong et al., 2018], as illustrated in Figure 2.3. Interestingly, Armstrong et al. [2018] observed that the position of OB stars within associations also does not appear to preferentially align with the position of lower mass association members, and so these positions should not be

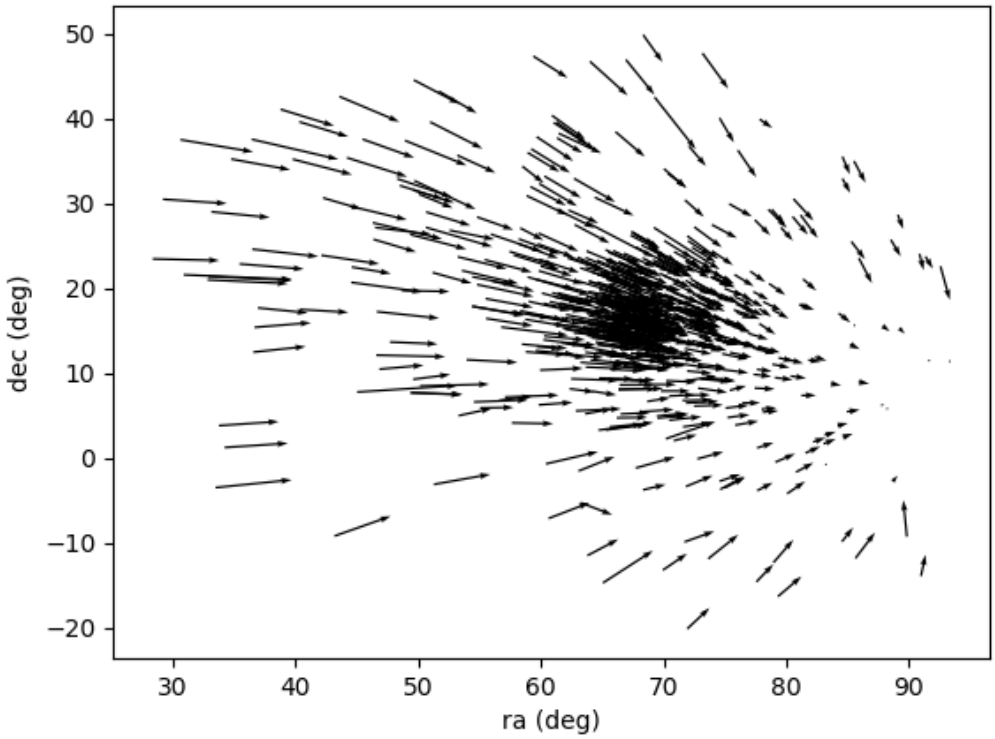

Figure 2.2: Equatorial (RA,Dec) position and velocity plot for all members of the Pleiades Cluster with membership probabilities ⩾99% from the census completed by Bouy et al. [2015]. The overlaid equatorial velocities of members of the Pleiades cluster appear to converge when reversed. The Convergent Point method allows membership probability of individual stars within larger clusters to be calculated by evaluating the true convergent point and testing how well individual cluster members fit the pattern of this convergence.

given too much weight in initial candidate selection. Nonetheless, the sheer number of stars surveyed by all-sky surveys such as *Gaia* necessitates the use of at least wide positional limits in order to focus the candidate search and gain manageable data sizes.

Another method of testing association membership is by examining the ages of candidate members, either by internal means or by comparison to previously known bona-fide association members. This can be achieved in a variety of ways. One of the most widely-used methods is plotting candidate members on colour-absolute magnitude diagrams (CAMDs) and thereby observing pre-main sequence and early-type stars, which was recently used powerfully by Zari et al. [2018] to separate young stars in *Gaia* DR2. Similar methods were used to define members of open clusters in Gaia Collaboration et al. [2018b]. Other key methods include observing the lithium depletion boundary in young clusters, using pulsations and

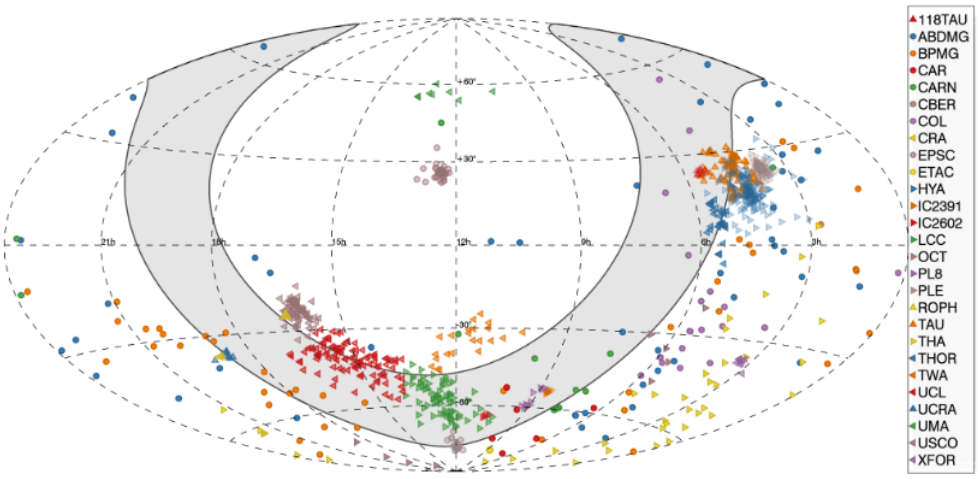

Figure 2.3: Position of young associations used in the development of the BANYAN Σ tool [Gagné et al., 2018b]. The Galactic plane (b < 15^o) is highlighted in grey. Reproduced from [Gagné et al., 2018b].

seismology to measure stellar age, measuring stellar surface gravity, looking for certain types of rotation and activity and observing lithium abundance [Soderblom et al., 2014].

Additional less-widely used methods for narrowing down and testing candidate members include the weighted median method for parallax and proper motion [Kuhn et al., 2019], parallax-proper motion diagrams [Damiani et al., 2019b], density maps [Armstrong et al., 2018; de Zeeuw et al., 1999], the 'Spaghetti' method featuring interconnecting Gaussian cylinders [de Zeeuw et al., 1999] and finder charts [Gagné and Faherty, 2018].

In order to combine some of the above methods and define a standardised method of determining the membership probability for candidate association members, two different groups have developed Bayesian membership tools. Firstly, Rizzuto et al. [2011] developed a Bayesian membership criterion based on proper motions, parallax, radial velocity and galactic latitude in order to look for further high-mass membership of the Sco OB2 association (though it was labelled a 'moving group' in this paper). Applying positional limits of $285^o \leqslant l \leqslant 360^o$ and $-10^o \leqslant b \leqslant 60^o$ and a colour limit of B-V $\leqslant$ 0.6 (which doubled as a mass limit), they compared six physical parameters (UVW velocities, distance and Galactic latitude/longitude) to the expected average value for this association, and obtained a membership probability for each candidate. BANYAN, the Bayesian Analysis for Nearby Young AssociatioNs tool first developed by Malo et al. [2013] presents

an alternative approach, choosing to model stellar associations/moving groups as single-dimensional Gaussian distributions in XYZ UVW space. This has since been further developed into the BANYAN Σ tool,[1] which models associations by six-dimensional multivariate gaussians in XYZ UVW space, and includes 27 known young associations (with ages 1 - 800Myr) within 150pc based on *Gaia* DR1 data. Using a Bayesian inference method and built on a training set of bona-fide or high likelihood stellar association members, BANYAN Σ requires only sky position and proper motion to calculate membership probability for candidate members, but can give more accurate probabilities with the inclusion of radial velocity and parallax measurements. This tool has already been used to identify 32 new bona-fide F0-M3 members of associations within 150pc, and 219 additional high-likelihood members [Gagné et al., 2018c]. Gagné et al. [2018c] consider stars to be bona fide members of an association if they have full kinematic measurement allowing them to be placed in XYZ UVW space, display a $> 90\%$ membership probability to an existing association (using the BANYAN Σ tool [Gagné et al., 2018b]), and if they display "signs of youth" consistent with the approximate known age of the association. They used a wide rage of signs of youth, including G-J vs GALEX NUV - G colours, X-ray emission with HR1 $\geqslant$ -0.15, mid-infrared excess, lithium absorption above 100 mÅ, a luminosity class or young isochronal age consistent with youth, or Ca II infrared triplet age consistent with the proposed association.

Given the large number of young groups, clusters and stellar associations already known, current work with *Gaia* DR2 is focusing on populating the lower-mass regions of these associations [Gagné et al., 2018b; Zuckerman, 2019; Faherty et al., 2017; Gagné et al., 2018a], however the high precision of the *Gaia* DR2 data (and upcoming further data releases) also offers the opportunity to discover additional new associations. For example, Kounkel and Covey [2019] used an unsupervised clustering algorithm to find hundreds of new filamentary 'strings' of stars with very similar kinematics, which may be of similar ages. Similarly Cantat-Gaudin et al. [2018] applied the unsupervised UPMASK classification scheme of Krone-Martins and Moitinho [2014] to *Gaia* DR2 data to define membership probabilities for both new and known open clusters. With continual improvement in astrometric precision as the *Gaia* mission continues, it is considered highly likely that more groups of young stars will be found.

In this work the primary methods used are the BANYAN-Σ method and kinematic common proper motion/location cuts, however essentially all methods are called upon in some manner in the assembly of the extensive young star target

[1] http://www.exoplanetes.umontreal.ca/banyan/banyansigma.php

list in Section 3.3.

2.1.2 Catalogue matching

Given the significant interest in groups of young stars, considerable lists of known young stars already exist in the literature. In order to make the most of these lists (and to combine them together into an overall young star list - see Section 3.3) it is useful to be able to cross-match these lists with large catalogues such as the *Gaia* DR2 data [Gaia Collaboration et al., 2016b, 2018a] and the *TESS* Input catalogue [TIC v8; Stassun et al., 2019]. To do so, the *TOPCAT* table handling software [Taylor, 2005] was used, which is a very powerful piece of software which allows for table manipulation and built-in sky-based cross-matching with large astronomical catalogues. Before assembly into larger lists, all target lists of interest were cross-matched with *Gaia* DR2 (by using a sky-crossmatch to find the best nearby object within a radius of 5 arcseconds) to obtain precise astrometric data on each object, before cross-matching each object with the *TESS* Input Catalog based on the *Gaia* DR2 identifier. This was achieved using TOPCAT's in-built interface with the Centre de Données astronomiques de Strasbourg (CDS) crossmatch service. In particularly crowded regions, the closest-matching star was chosen by TOPCAT (using TOPCAT's 'Best' find mode), with a uniqueness cut performed at the end to remove any targets which had inadvertently been assigned to the same source. Finally the sector in which each object had been (or would be) observed in was found using the `tess-point` software [Burke et al., 2020]. Similar methods were used when preparing the BANYAN target list for use in Chapter 4 and the CDIPS target list in Chapter 5.

2.2 TESS data extraction methods

While *Gaia* supplies the data necessary to expand the population of known young stars, *TESS* supplies the photometric data with which to search for transits. During *TESS*'s primary mission it generated two primary sources of light-curve data: 2 min cadence light-curves and 30 min cadence Full-Frame-Images (FFIs), each totalling approximately 27 days of observations for each *TESS* sector. While 2 min light-curves represent the highest-cadence and most reliable source of light-curves for a pre-defined subset of targets observed by *TESS*, the 30 min Full Frame Images allow light-curves to be generated for any target that fell on one of the *TESS* detectors, greatly extending the number of targets for which data is available.

The main 2 min light-curve extraction is performed by the *TESS* Science

Processing Operation Center (SPOC) pipeline [Jenkins et al., 2016] and based on the highly successful *Kepler* Science Operations pipeline of Jenkins et al. [2010]. This pipeline includes the removal of systematic instrumental effects, computing the optimal photometric aperture for each star, removing background flux, computing the centroid for each object and using PSF-fitting for 200 nearby bright target stars on each CCD to establish the camera pointing and focus. At this point the Simple Aperture Photometry (SAP) light-curves are saved, before an additional Presearch Data Conditioning (PDC) step performs a series of important corrections to find and remove principal components related to time-correlated instrumental systematics, combat changes in camera pointing and focus, and remove outliers. It also solves problems regarding the photometric aperture capturing too much (or too little) flux in the case of crowded regions or very bright stars. The final light-curve products include both the SAP and PDC light-curves. 2 min targets are chosen from the *TESS* Candidate Target List [CTL, Stassun et al., 2018], which prioritises stars by the smallest transiting planets which could be found around them. Alongside this, approximately 10,000 targets are included from the Guest Investigator program per cycle, for a total of ∼200,000 light-curves per cycle. These light-curves are freely available on MAST[2] shortly after observation of each sector each month.

The main downside of the 2 min cadence data is that it is only available for approximately 20,000 targets per sector. The 30 min cadence FFI data is therefore crucial to extend the search for exoplanets further. In late 2020, SPOC began releasing TESS-SPOC FFI light-curves [Caldwell et al., 2020a], applying the main SPOC pipeline [Jenkins et al., 2016] to up to 160,000 targets per sector, pre-selected from the *TESS* Input Catalog [Stassun et al., 2019]. Selection criteria include all targets with 2 min data, targets bright in near-infrared (H mag $\leqslant$ 10) but with minimal crowding, targets within 100 parsecs and targets with *TESS* magnitude $\leqslant$13.5, up to at most 10,000 targets per *TESS* CCD. While these light-curves took longer to appear originally, they are now one of the most extensive sources of FFI light-curves, with light-curves currently available for sectors 1-39.

However, because the TESS-SPOC FFI light-curves have only become available recently (and for a select sample of ∼160,000 targets per sector), several alternative methods have been created for extracting light-curves from the *TESS* FFIs by the community. The simplest approach is to perform aperture photometry on the raw FFIs by overlaying an aperture over a cut-out around the object of interest, and summing up the flux under the target aperture for each cadence. This

[2]https://mast.stsci.edu/portal/Mashup/Clients/Mast/Portal.html

can be performed easily using the prebuilt *lightkurve*[3] `Python` package [Lightkurve Collaboration et al., 2018], and observing the resulting target-pixel-file provides an instructive look at the area around the star of interest. However, the simplistic nature of this analysis yields relatively noisy light-curves uncorrected for issues such as spacecraft pointing, jitter and localised scattered light. A greatly improved simple aperture photometry pipeline which accounts for these issues has been built in `Python` by Feinstein et al. [2019]. Christened *eleanor*[4], this package performs background subtraction, removal of spacecraft systematics such as jitter and pointing drift, and aperture/psf photometry. In addition, it provides tools to complete further systematics-removal via principal component analysis or psf-modelling. Light-curves from *eleanor* are steadily being ingested into *MAST* (see Montet et al. [2020], but at the time of writing was only reliably available as an open-source tool designed to work for *TESS* Sectors 1 to 13.

An alternative difference imaging approach for 30 min light-curve generation has been pursued by Oelkers and Stassun [2018], Bouma et al. [2019] and Huang et al. [2020a]. Attempting to overcome the challenges posed to aperture photometry by *TESS*'s large pixel sizes, these approaches extract light-curves via difference imaging. The first public difference-imaging 30 min cadence data was generated by the Difference Imaging Analysis (DIA) pipeline of Oelkers and Stassun [2018, 2019]. In this method, a 'master frame' is generated by combining all images from the telescope for each sector, and then evaluating the difference in flux in each individual image by subtracting the master frame from it. In this manner only the variation in flux of stars between images is retained. Highlighting stars of interest within this process aids removal of contaminating stars in crowded regions and hence improves the light-curve extraction compared to standard aperture photometry methods. A similar technique has been used in ground-based surveys such as the Kilodegree Extremely Little Telescope [Pepper et al., 2007; Siverd et al., 2012], though with a simpler Gaussian kernel than the Dirac δ-function kernel (where the function is equal to zero except at exactly the wavelength of interest) used here. A full description of the DIA technique as applied to *TESS* light-curves in sectors 1 and beyond can be found in Oelkers and Stassun [2018, 2019]. Oelkers and Stassun [2019] use this pipeline to extract light-curves for all stars in the *TESS* Input Catalog [Stassun et al., 2018], but unfortunately only for sectors 1-5. Light-curves extracted via this pathway are accessible from the Filtergraph data visualization service.[5].

Meanwhile the CDIPS pipeline [Bouma et al., 2019] uses a slightly different

[3]https://github.com/KeplerGO/lightkurve
[4]https://github.com/afeinstein20/eleanor
[5]https://filtergraph.com/tess_ffi

difference-imaging approach to extract light-curves from the *TESS* FFIs. Based largely on the FITSH package of Pál [2012], the pipeline begins with the calibrated *TESS* SPOC FFIs, performs large-scale background subtraction and then builds an astrometric reference frame based on a series of the brightest stars in each image, verified with the World Coordinate system (WCS) and the *Gaia* DR2 catalogue [Gaia Collaboration et al., 2016b, 2018a]. All calibrated images are then transformed to this frame before a photometric reference frame is constructed from 50 images with low background noise. In order to get the final light-curves, each target frame is then subtracted from the photometric reference frame and aperture photometry applied to each source. An in-depth description of the full CDIPS pipeline can be found in Bouma et al. [2019]. The CDIPS target list is unusual in that its selection criteria focuses explicitly on those stars which have evidence of youth in the literature, making it particularly useful for this work. At present this accounts for over 650,000 light-curves from the first half (S1-13) of *TESS*'s primary mission.

More recently, the Quick-look Pipeline (QLP) of Huang et al. [2020a,b] has been used to extract light-curves for all stars in the *TESS* input catalog with *TESS* magnitudes brighter than 13.5 in the *TESS* Primary Mission (and a smaller sample of stars of brightness between T mag 13.5-15 with proper motions >200 mas/yr). Totalling ∼14,770,000 light-curves from the Southern Hemisphere and ∼9,600,000 light-curves from the Northern Hemisphere, this represents the largest collection of *TESS* photometry available to date. To extract these light-curves Huang et al. [2020a,b] use the TICA software [Fausnaugh et al., 2020] to calibrate the Full-Frame Images and the `nebuliser` software of Irwin [1985] to perform a global background subtraction, before determining an astrometric solution for each image and constructing a reference image from a combination of 40 good quality images. Five different circular apertures are used to measure the differential brightness, which are then used for the second step of background subtraction. Finally the measured difference fluxes are converted to absolute fluxes by adding the expected flux from each source from the TIC and the time adjusted to the *TESS* Barycentric Julian date. This results in what Huang et al. [2020a] deem the "raw light curve", which typically also includes low-frequency variability from instrumental effects or stellar activity. In order to remove these signals before planetary searches, Huang et al. [2020a] perform an additional post-processing step which includes cleaning of poor-quality data points based on the engineering quaternion data[6], flattening using a basis spline [Vanderburg and Johnson, 2014], an additional bad data-quality check and normalisation of the light-curves. The final light-curve files include detrended

[6]https://archive.stsci.edu/missions/tess/engineering/

light-curves using three different apertures, alongside the original 'raw' light-curve. This pipeline continues to be run on data from the extended mission, as described by Kunimoto et al. [2021].

One further 30-min pipeline which will be particularly useful for young star studies is the TASOC pipeline, under development by the *TESS Asteroseismic Science Consortium* (*TASC*). Heavily based on the K2P^2 pipeline [Lund et al., 2015], this pipeline aims to supply *TESS* photometry data for use in asteroseismology and stellar variability analyses. Data from this source can be accessed online[7] after joining the consortium. So far raw light-curves extracted from the FFIs are available for sectors 1 and 2 [Handberg and Lund, 2019], while the open-source code used to extract these images from the FFIs can be found on github.[8] Although this pipeline is still currently under development, a useful overview of its design and aims has been written by Handberg et al. [2020].

Additional public light-curve extraction methods for *TESS* have been built by Nardiello et al. [2019] (PATHOS), which uses a PSF-based approach to extract high-precision light-curves for members of stellar clusters, and by Montalto et al. [2020] (DIAmante), which produces systematics-corrected multi-sector light-curves, but these are not used as data-sources for this research, so not discussed in depth in this work.

Given the young star focus of much of this research, the difference-imaging approach of 30min-cadence light-curve extraction was preferred in this work, meaning that the main data-extraction methods used were the DIA [Oelkers and Stassun, 2018], CDIPS [Bouma et al., 2019] and QLP [Huang et al., 2020a] pipelines. However, as available pipelines evolved considerably over the course of this research, the specific pipeline used varied depending on what light-curves were available at the time that analysis was completed. For clarity, the final data-source choices made are discussed within each science chapter, along with a discussion of any intricacies associated with the use of those specific data sources. The one exception to this is Chapter 6, which uses data extracted from the main *Kepler* and *TESS* 2 min-cadence pipelines. These are explained further in Section 6.2.

2.3 Planetary transit detection

As discussed in 1.2.1, planetary signals present themselves in photometric data as distinctive dips in the flux from their hosts stars as they transit between their star

[7] http://tasoc.dk
[8] https://github.com/tasoc

and an observer. However, given the relatively low probability of transit per star and the correspondingly large number of stars which require searching through, an automated algorithm is necessary to speed up the search process. The most common exoplanet detection algorithm is the Box Least Squared (BLS) method. This method, originally developed by Kovács et al. [2002], fits a series of (strictly periodic) box-shaped dips at a range of periods in order to generate a periodogram, comparing the relative strengths of the different period hypotheses. Specifically, the BLS algorithm minimises the chi-squared value for a five parameter fit of period P_0; epoch of transit, t_0; fraction of time spent in transit, q, and depth in (L) and out (H) of transit, reporting the most likely period as that which minimises the chi-squared value. Some care must be taken when choosing the final period however, as the BLS periodograms also contains peaks at integer multiples or fractions of P_0 (period aliases, e.g. $P_0/3$, $P_0/2$, $2P_0$, $3P_0$). This method has been the archetypal transit discovery algorithm in the community for decades now due to its speed and signal detection power, so is the main algorithm implemented in transit searches and pipeline development in this work.

A recent improvement made to the BLS algorithm (released partway through this research) is the Transit Least Squares (TLS) algorithm designed by Hippke and Heller [2019]. Similar in vein to BLS, TLS fits a periodic transit to search for planetary transits. However, unlike the simple box-shape of BLS, TLS fits a much more realistic transit shape to the data, based off an average of thousands of real planetary transits from the Exoplanet Archive.[9] Hippke and Heller [2019] show that this new algorithm is of comparable speed to BLS but "substantially more efficient and reliable in finding small planets than BLS." In the future this algorithm will be folded into the main young star transit-search pipeline designed in this work.

2.4 Variability period searching algorithms

Due to the increased activity and variability of young stars, determining the main variability period of young star light-curves is often almost as important as searching for periodic transits, especially where this source of 'red noise' needs to be removed before the main planet search can take place. In this work the community-standard Lomb-Scargle (L-S) method [Lomb, 1976; Scargle, 1982] was used in order to find the main variability periods for all targets. The L-S method is a form of Discrete Fourier Transform specifically designed to handle the unevenly space observations often seen in astronomical data. It works by breaking a given time-series down into a linear

[9] https://exoplanetarchive.ipac.caltech.edu/index.html

combination of sinusoidal functions, the basis of which effectively transforming the data from the time domain to the frequency domain. Finding the strongest frequency in the resulting power spectrum (or periodogram) thus gives the inverse of the strongest period in the data-set, which is typically the chief variability period of the light-curve. However, when using this technique it should be realised that it is most sensitive to sinusoidal-like signals, so signals which cannot be easily decomposed into sinusoids (such as the signals of transiting exoplanets) are better found using other methods. Other methods of measuring stellar variability periods include the autocorrelation function [ACF; e.g. McQuillan et al., 2014; Briegal et al., 2022], which measures how correlated (or similar) different regions are of the same light-curve, Wavelet analysis [e.g Torrence and Compo, 1998; Martins et al., 2020], which fits a generalised basis function (or 'mother wavelet') which can be stretched in both frequency and time, and Gaussian Processes [GPs; e.g Rasmussen and Williams, 2006; Angus et al., 2018], which models the covariance between data points and improves the ability to find periods in light-curves which evolve in time. Nonetheless, because of the algorithms' speed, the ease of recovering multiple frequencies in the same light-curve and the fact that the majority of stellar variability in young stars is from approximately sinusoidal spot modulation and pulsations, the Lomb-Scargle method was chosen as the primary method to obtain variability periods in this work.

2.5 Traditional light-curve detrending Methods

The main aim of light-curve detrending is to remove various aspects of red noise from the light-curve (such as instrumental systematics, flux discontinuities and often stellar variability) while leaving behind the signal of interest. In searches for exoplanets this often is likened to 'flattening' the light-curve, removing all non-planetary signals to leave behind a flat light-curve with added transit dips. It is important to note here the difference between two different aspects of detrending: i) removal of instrumental systematics/trends, and ii) removal of astrophysical signals for more targeted searches. The first aspect of this detrending is required for all types of survey, as different instruments have different systematic noise profiles in their data, from terrestrial nuisances - such as clouds or atmospheric absorption - to spacecraft roll and scattered light signals on orbiting satellites. In order to search for any astrophysical signals, these systematic trends need to be removed from the data first (or sometimes, concurrently). Meanwhile, for searches which focus on one type of signal (e.g. planet searches), other astrophysical signals could be considered 'noise' which needs detrending. For example, in a standard planet search, stellar

rotation of a spotty star or pulsation of a potential host both leave behind signals in a photometric light-curve which can hide planetary transits. The detrending and removal of both instrumental and astrophysical 'noise' is thus crucial to exoplanetary surveys, whether these are performed piece-wise (removing instrumental systematics first and subsequently removing astrophysical signals), or concurrently, using more aggressive detrending method. While instrumental noise is increasingly removed in the main pipelines for these instruments, significant astrophysical noise typically still remains, which is the main type of detrending focused on in the remainder of this section.

Given its importance to the community, a huge variety of different detrending methods have been trialled to detrend light-curves, from the common Savitzky-Golay filter [Savitzky and Golay, 1964] to more involved Gaussian processes [e.g. Aigrain et al., 2015]. Following the formalism of Hippke et al. [2019], the most common methods can be roughly split into three main types: sliding filters, splines and polynomials, and Gaussian processes. Sliding filters [e.g. Savitzky and Golay, 1964; Wheatley et al., 2010] work by effectively moving a 'window' or 'box' across the data, estimating (and removing) a trend within a given preset time-frame or number of points (cadences). Changing the length of the window adjusts whether the algorithm removes long-term or short-term trends, while changing the weighting of the windows (e.g. using robust estimators such as Tukey's biweight [Mosteller and Tukey, 1977]) can give more or less value to those data-points closer to the one currently being fit. On the other hand, polynomial or spline detrending methods [e.g. Giles et al., 2018; Vanderburg et al., 2016] work by fitting smooth, piecewise polynomial approximations to the data, usually with equally-spaced knots along the function. Some methods such as LOWESS smoothing [Cleveland, 1979a] combine these two methods, effectively fitting a locally-weighted polynomial regression to a specific window of the data. Finally, Gaussian process methods [e.g. Aigrain et al., 2015] are a non-parametric regression method, effectively setting up a probability distribution over functions, which assumes that nearby observations are correlated. This can allow complex unknown detrending functions to be found for individual data-sets without having to guess the approximate form beforehand. A full explanation of this family of methods can be found in Rasmussen and Williams [2006]. Gaussian processes are particularly powerful due to their ability to simultaneously fit stellar activity and planetary transits within planetary searches, however can struggle in larger surveys due to greatly increased computational costs compared to simpler algorithms. An excellent recent review of all common detrending methods can be found in [Hippke et al., 2019], where they conclude that the optimum simple

detrending algorithm for normal transit searches is a time-windowed sliding filter with a Tukey's biweight location estimator [Mosteller and Tukey, 1977].

Automated detrending is especially important for large surveys, where the sheer volume of data produced is simply too much for human eyeballing. The earliest wide-field survey, HATNet [Bakos et al., 2004], used the Trend Fitting Algorithm [Kovács et al., 2005] and External Parameter Decorrelation [Bakos et al., 2010] to detrend its light-curves (both of which focused more on instrumental noise, but also removed some astrophysical signals), while WASP [Pollacco et al., 2006] shortly after preferred the `sysrem` algorithm of Tamuz et al. [2005] to remove instrumental effects – which removes common time- and position-dependent trends by taking into consideration the weighted average magnitude residuals for all stars – coupled with a boxcar-smoothing technique to handle some of the more variable stars observed [Collier Cameron et al., 2006]. More recent ground-based surveys such as KELT [Pepper et al., 2007] and NGTS [Wheatley et al., 2018] have developed these techniques further, and even introduced more advanced detrending methods such as gaussian processes as they target more active stars [e.g. Costes et al., 2020; Gillen et al., 2020]. However, the largest source of light-curves in the search for exoplanets has been space-based surveys, the volume and variety of which has required the development of more complex and robust detrending techniques. The most prolific transiting exoplanet surveys thus far have been the *Kepler* [Borucki et al., 2010] and *K2* [Howell et al., 2014] missions and the ongoing *TESS* mission [Ricker et al., 2014], all of which share broadly similar detrending techniques which first remove known instrumental systematic errors before applying data-smoothing in their pre-search data conditioning (PDC) steps. The *Kepler* and *TESS* missions all use the wavelet-matched method originally developed by Jenkins [2002] to simultaneously detrend intrinsic variability of potential host stars and search for transiting planets. This method uses an overcomplete wavelet-transform of the data, which is effectively a more complex version of a Fourier transform where the basis functions are concentrated in time as well as frequency. This is efficient for handling simple stellar variability, but is less-well optimised for particularly active stars, especially when their variability evolves with time. Overviews of the full *Kepler*, *K2* and *TESS* SPOC pipelines can be found in Jenkins et al. [2010], at https://keplerscience.arc.nasa.gov/k2-pipeline-release-notes.html and in Jenkins et al. [2016] respectively. Gaussian Process detrending is also an important detrending method for space-based surveys, having been used to simultaneously model instrumental systematics and astrophysical variability while extracting K2 light-curves in [Aigrain et al., 2016]'s K2SC pipeline.

However, despite the success shown by these traditional detrending methods, they fall short when it comes to young exoplanet discovery. Indeed, Hippke et al. [2019] show that none of the methods tested recover more than 37.2% of injected $0.5R_{\rm Jup}$ sized planets in young star light-curves, even when recovery rates of the same planets are near 100% in older, less noisy stars. Interestingly all methods returned similar performance of ∼35% recovery, despite large variations in detrending method. The poor performance of these traditional methods is largely because the typical window lengths chosen for typical 'one size fits most' detrending processes are too long to remove much of the observed large-amplitude, short-period young stellar activity, especially when it may evolve over the course of a single 27 day *TESS* sector. Furthermore, the presence of stellar activity may often distort exoplanetary transit signals, leaving non-transit like shapes after detrending. More specialised pipelines designed by Rizzuto et al. [2017] have shown some more promise, with the 'notch-filter' and LoCoR pipelines shown to be capable of reaching down to Earth-sized planets in some older clusters and associations, as demonstrated with continued discoveries in the ZEIT and THYME programs [e.g. Mann et al., 2017c; Vanderburg et al., 2018; Rizzuto et al., 2018, 2020]. However, even these pipelines struggle with very fast and/or evolving stellar variability of many young hosts, so it is clear that more work is needed in order to push down to smaller exoplanets around such active young stars.

2.6 LOWESS detrending pipeline

2.6.1 Choice of base-detrending method

Given the challenging range of activity-related and intrinsic stellar variability seen in the light-curves of young stars (and the associated difficulty of modelling so many individual light-curves in an initial transit search), it was considered wise to approach the astrophysical detrending problem from the ground up, rather than necessarily relying on more traditional methods such as the Savitzky-Golay filter [Savitzky and Golay, 1964]. A number of different detrending techniques were trialled early in the process, including simple low-order polynomial fitting, sinusoidal modelling, Savitky-Golay filtering and general smoothing over a range of window sizes. The most successful of these are illustrated in the 2nd to 4th panel of Figure 2.4. Of these, low-order polynomial fitting and the smoothing methods proved most successful at recovering transit signals injected into young star light-curves, as shown by the relative log likelihood of the periodogram on the left of Figure 2.4. Note that the utilised methods for injecting and searching for transiting planets are outlined

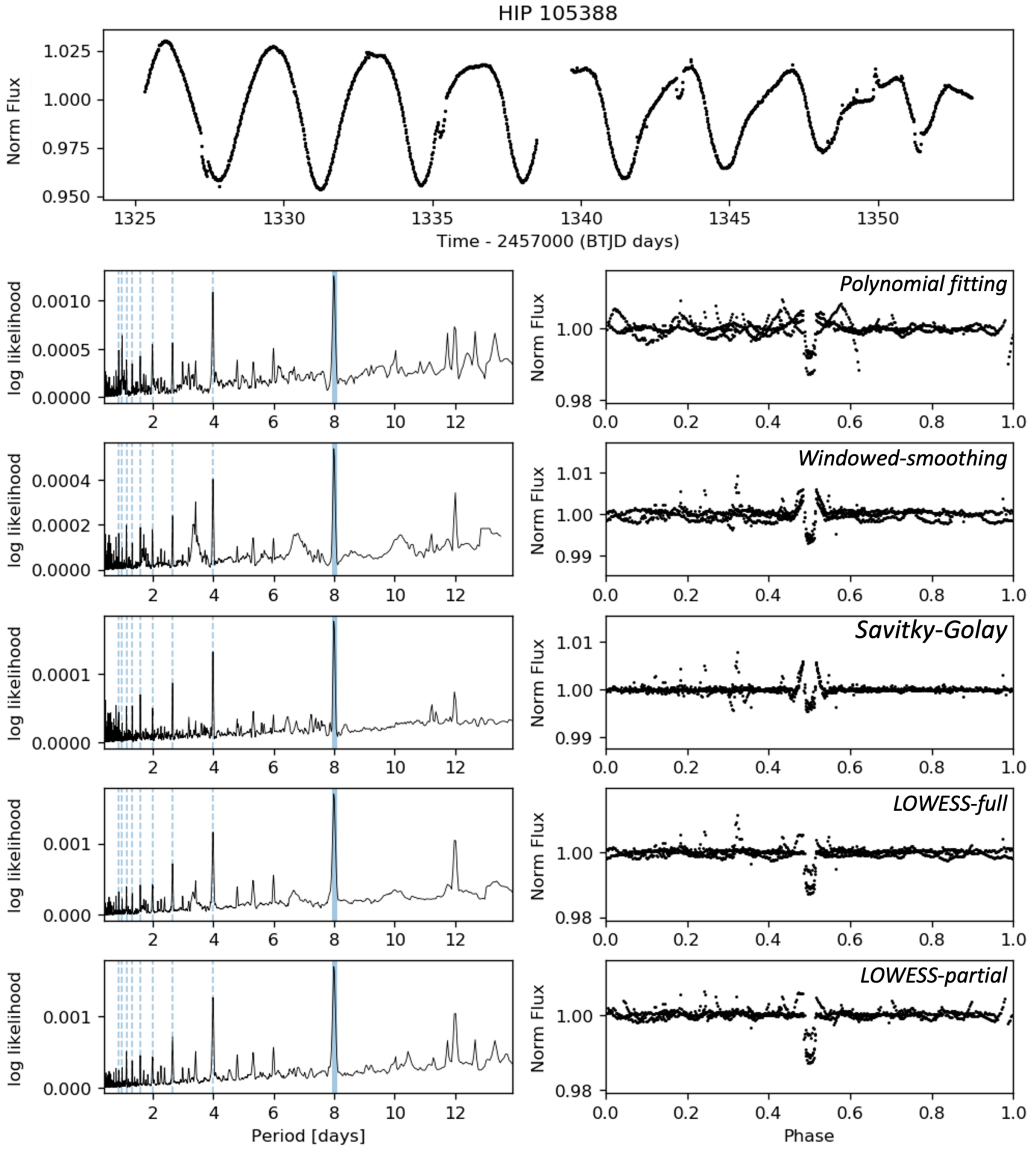

Figure 2.4: Comparison of detrending methods for an 8-day period, Hot-Jupiter sized planet (0.1 $R_{\rm P}/R_*$) injected into the light-curve of the young star HIP 105338. The top panel shows the original 30min light-curve for the star, including the injected planet transits. This is followed by a periodogram from a BLS search (with the strongest period highlighted in blue, and aliases of this as dotted blue lines), and the detrended data on the right folded by this strongest period. All detrending methods are labelled in the top right of the phase-fold box. For all methods requiring a defined window, a 15hr window size was used. Peak cutting was applied for poly and LOWESS-partial methods, as these are when these techniques are most useful.

in section 2.6.5 and 2.6.4 respectively.

A search for a method to combine and improve both of these methods led to LOWESS smoothing, or Locally Weighted Scatterplot Smoothing. This method, developed originally by Cleveland [1979b], is a local polynomial regression method which works by fitting a low-order polynomial to a subset of the data (the width of which is set by a user-defined window) at each point along the x-axis using weighted least-squares regression. Under the weighted least-squares regression method, points nearer to the data-point being estimated are given more weight than those further away in the window. This weighting is one of the key differences between this method and the more commonly used Savitzky-Golay filter [Savitzky and Golay, 1964], and is particularly important for this application given the often swift evolution of young star light-curves. Indeed following the choice of this LOWESS-smoothing method, it was independently highlighted to be one of the best-performing detrending methods for the young-star sample tested by Hippke et al. [2019]. In exoplanet literature however no other mention of LOWESS smoothing for exoplanet searches was found, with even the related LOESS smoothing method appearing quite rarely, despite having been used in detrending the TRAPPIST-1 system [Luger et al., 2017] and in the Autoregressive Planet Search of Caceres et al. [2019].

More specifically, the LOWESS method works by fitting a simple polynomial, $g(x_i)$, to a window (or 'fraction') of the data-set, such that the 'i-th' point, (x_i, y_i), in the window is represented by the equation

$$y_i = g(x_i) + \epsilon_i, \tag{2.1}$$

where ϵ_i is a random variable with a mean of zero and a constant scale. However, where LOWESS diverges from traditional smoothing/windowing methods is that standard least-squares regression method of polynomial fitting is replaced by a 'locally weighted least-squares regression' method, where points nearer the value currently being evaluated are weighted as more important than though further away. The specific weighting method employed in the LOWESS method is defined as

$$W(x) = \begin{cases} (1 - |x|^3)^3 & \text{within the window,} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

as recommended by Cleveland [1979b]. This method also allows for multiple iterations of residual-based re-weightings in order to remove any significant outliers; in this case the number of residual-based re-weightings retained as the default value of 3. Through experimentation a window size of 30 FFI data points (15hrs) was found

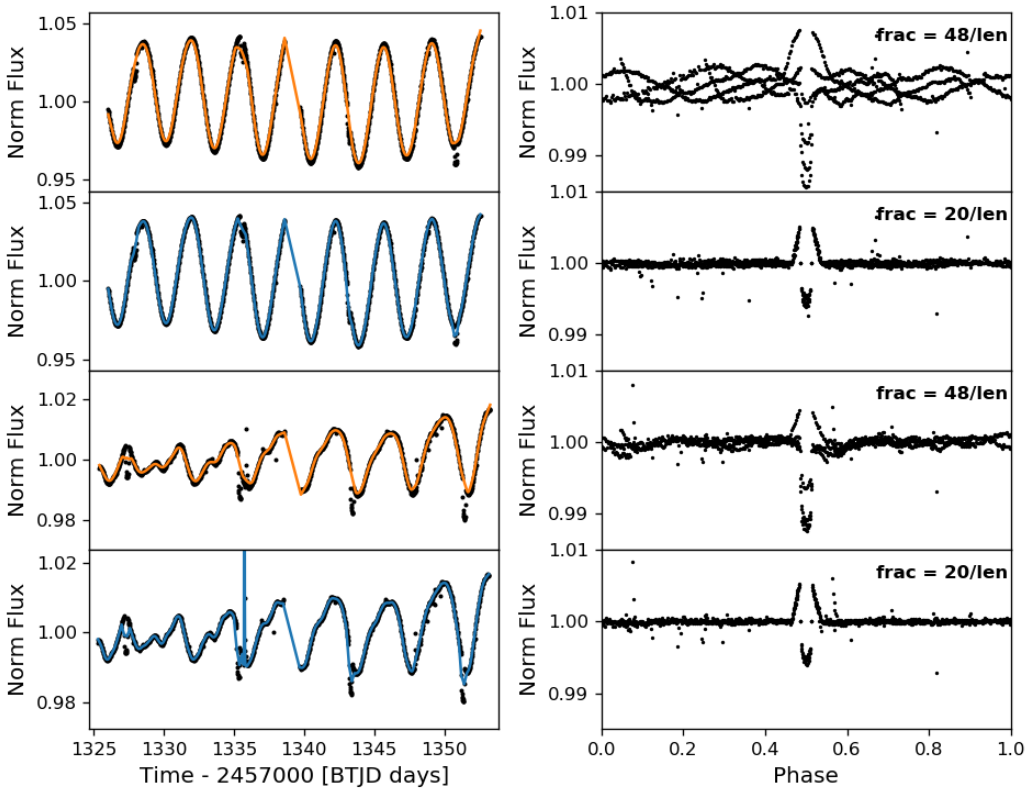

Figure 2.5: Effect of changing the LOWESS 'frac' parameter for HIP 1113 and AB Pic, two active young stars, including an injected Hot-Jupiter sized planet. Two values of the frac parameter are compared: 48 bin and 20 bin, corresponding to effective window lengths of one day and 10hrs respectively. It can be seen that choosing this parameter is a fine balance between leaving behind too much of the star's variability/activity, and beginning to smooth out (or decrease in amplitude) of the planetary signal.

to yield a good compromise between preserving the shape of injected transits and smoothing the stellar activity and variability of the host stars, except in more rapidly evolving light-curves, where a window size of 20 data points (10hrs) was found to be more appropriate. As such the 'frac' parameter (representing the fraction of the full data to fit a polynomial to at each step) was set as `frac = 30/length(data)` or `frac = 20/length(data) respectively`, which allowed for consistent window sizes even when there were gaps in the dataset. For stars where 2min light-curve data was available, these were proportionally increased to windows of 450 and 300 respectively. The effect of changing this 'frac' parameter is illustrated in Figure 2.5. Note that there are two slightly different ways in which LOWESS can be run through the data: running over the entire light-curve at once, deemed 'LOWESS-full', and restarting the detrending any time there is a significant gap in the data (here treated as 0.1days), which is deemed 'LOWESS-partial'. Both of these methods were used here, and are illustrated in the bottom two panels of Figure 2.4. One final important parameter for the specific `statsmodels.nonparametric.lowess` implementation used in this work is the delta parameter, which can be used in the case of particularly large data-sets to speed up computation by using linear-interpolation for a given distance (delta) from the point of interest. In this analysis this was retained as $delta = 0.0$ as the sector-by-sector data-sets were not overly large, but this can be adjusted if further speed enhancements are desired. Furthermore, for very high-cadence data-sets, the author recommends binning the original data in time to save on computation load.

2.6.2 Removal of Peaks and Troughs

One of the other key challenges present in using a Box Least Squares (BLS) search for light-curves from young active stars is that the troughs of regular stellar activity (e.g. from the rotation of star-spots) are often picked up as the largest peaks in the BLS periodograms, even after LOWESS detrending. This is particularly the case for rapidly rotating stars, or those with short-period intrinsic variability. Such sources typically exhibit sharp peaks and troughs in the extracted 30min light-curve, which is often highlighted in the BLS search. In an attempt to combat this problem for light-curves with particularly sharp stellar activity, the effect of cutting the peaks and troughs of this stellar activity was tested. To begin with, the peaks were located using the `find_peaks` function from the `scipy.signal` library[10] [Virtanen et al., 2019] using a required prominence of 0.001 and width of 15 data points (7.5hrs). This helped to ensure that only wider peaks generally associated

[10]https://docs.scipy.org/doc/scipy/reference/signal.html

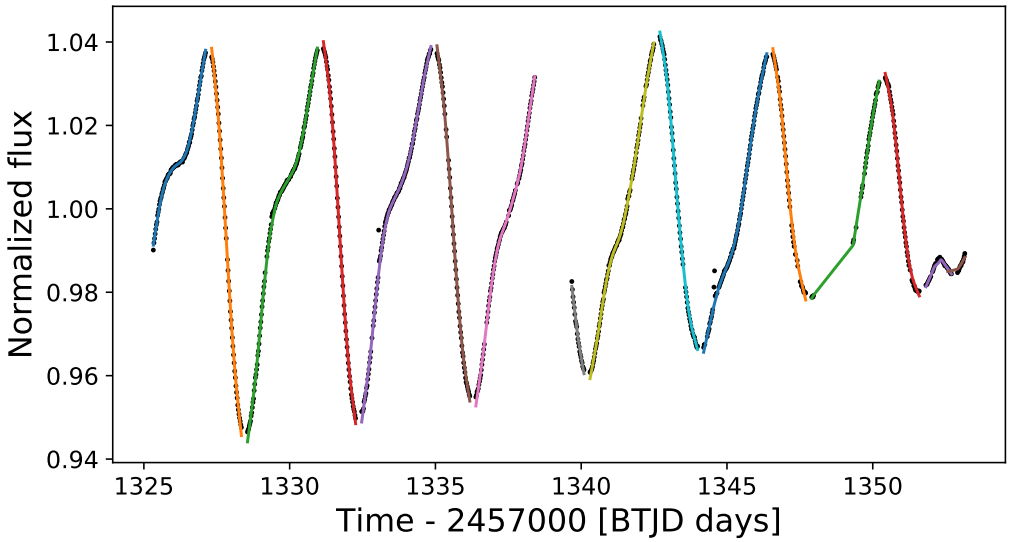

Figure 2.6: Example of the peak cutting technique in use on the rapidly evolving light-curve for HIP 32235. The applied 20-box partial LOWESS-smoothing model used for detrending the light-curve is also over-plotted in colour. Note that the change in colour between subsequent sections of the LOWESS-detrending shows where the light-curve has been split after any 0.1d gap in the data. This typically occurs after each peak or trough, or where significant detrimental scatter has been removed.

with stellar activity/variability were flagged as peaks or troughs, but the settings may need adjusting for some more complex light-curves. The activity/variability troughs were located using a similar `find_peaks` search on a negative version of the light-curve flux. Data points within 0.1d either side of each peak and trough were then cut from the light-curve before LOWESS-detrending was applied. An example of this in practice is shown in Fig 2.6 for HIP 32235, a rotationally variable G6V star in the Carina stellar association.

The power of this peak-cutting method is illustrated for the same source in Fig 2.7. When a 0.04 $R_{\rm P}/R_*$ radius ratio planet is injected into the light-curve (using the method described in Section 2.6.5) and the light-curve is detrended using the LOWESS-based method described above (boxsize = 20), any signal of the planet is clearly overwhelmed by the 3.84d period rotational variability of the star, as is shown in the top of Fig 2.7. However, applying the described peak-cutting technique to the data easily recovers the signal of the injected 8-day 0.04 $R_{\rm P}/R_*$ planet, as illustrated in the lower half of Fig 2.7. This technique is thus another powerful tool for pushing down to lower radii in the search for exoplanets around young active stars. In

general, this peak-cutting method was found particularly effective for light-curves with sharp oscillations/rotations, or ones which were strongly periodic. However, it is important to note that this technique did not always offer improvements over the non-peak-cut method. A further discussion of when this technique is most effective can be found in section 4.3.5.

2.6.3 Transit masking and light-curve interpolation

Transit masking is a commonly used method in planetary and eclipsing binary science to independently detrend long-term variability in stellar light-curves without changing the shape of the transit curve, and is implemented into most common light-curve manipulation tools (e.g. Luger et al. [2016]; Lightkurve Collaboration et al. [2018]). However, due to the often rapid evolution of young star light-curves, simply cutting out the data near a suspected transit before detrending can lead to spurious variability signals over the duration of the transit. In order to combat this issue in the developed detrending method a two-step transit masking approach was used which accounts for the brightness variation of the host star during the transit. Firstly, a mask is generated based on a period, epoch and duration for the suspected transit (either user-defined or from a previous BLS search), and the new light-curve generated by removing any points within the selected transit duration. Then, in order to account for the stellar flux variability over the course of the transit, the new holes in this light-curve are refilled using a quadratic interpolation (using `scipy.interpolate.interp1d`) between the cut points. It is this new flux array featuring interpolated sections over the transit mask that is then used in the LOWESS-detrending step, and in turn divided out to form the final detrended light-curve. The importance of using such a method is illustrated using the source DS Tuc A in Fig 2.8. It is immediately obvious that not taking into consideration the variation in stellar flux around the epoch of the first cut transit (approx TJD 1132) would lead to a considerably different transit shape after the main LOWESS-based detrending was is undertaken. A full discussion of the full detrending of DS Tuc A and the resulting recovery of the young exoplanet DS Tuc Ab can be found below in Section 4.3.1. Note that a basic version of this detrending code can be found online at https://github.com/mbattley/YSD.

2.6.4 Transit searching algorithm

The standard Box Least Squares method was used to search for periodic transit-like signals in the detrended light-curves. The specific implementation used in this code

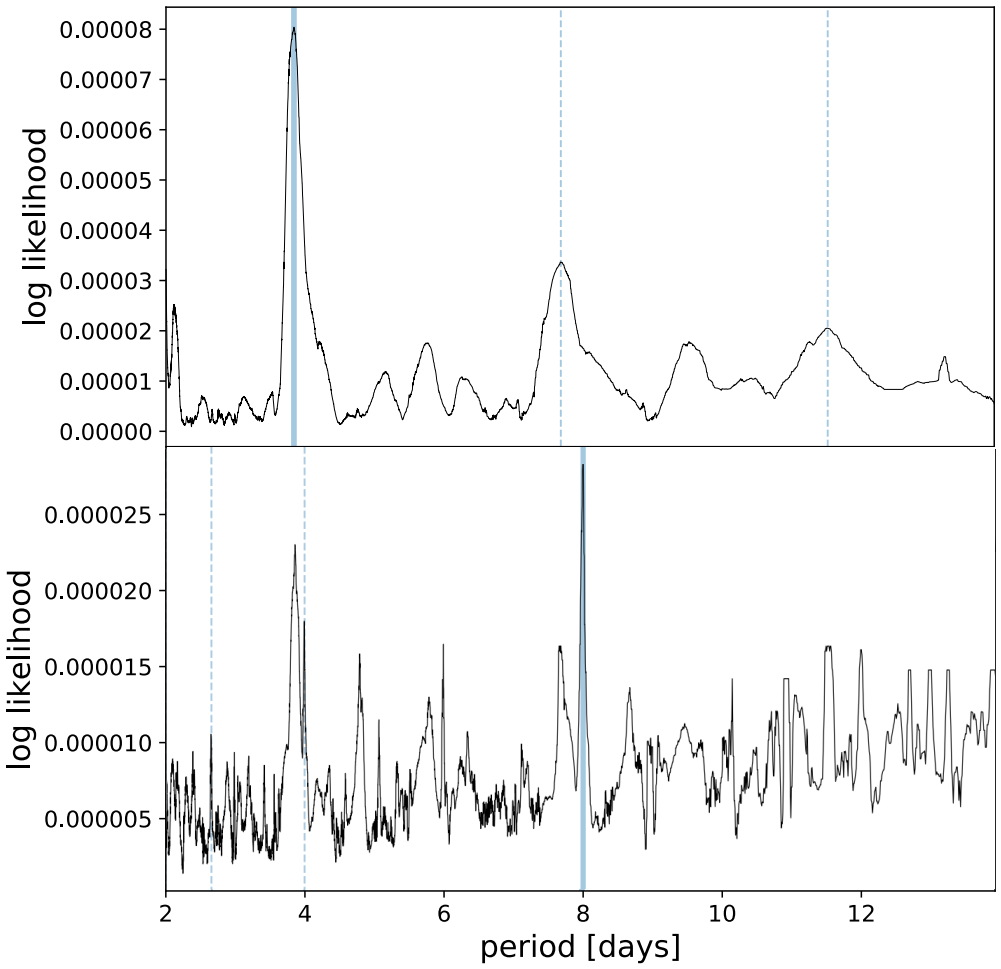

Figure 2.7: BLS periodograms for HIP 32235 following 20-bin LOWESS detrending without (top) and with (bottom) peak cutting implemented. In this case peak cutting displays a clear benefit to the recovery of an injected $0.04R_{\rm P}/R_*$ 8d planet in the presence of rapidly evolving (3.84d period) stellar activity. The highest power period is highlighted in blue, with aliases of the same period shown by dotted blue lines.

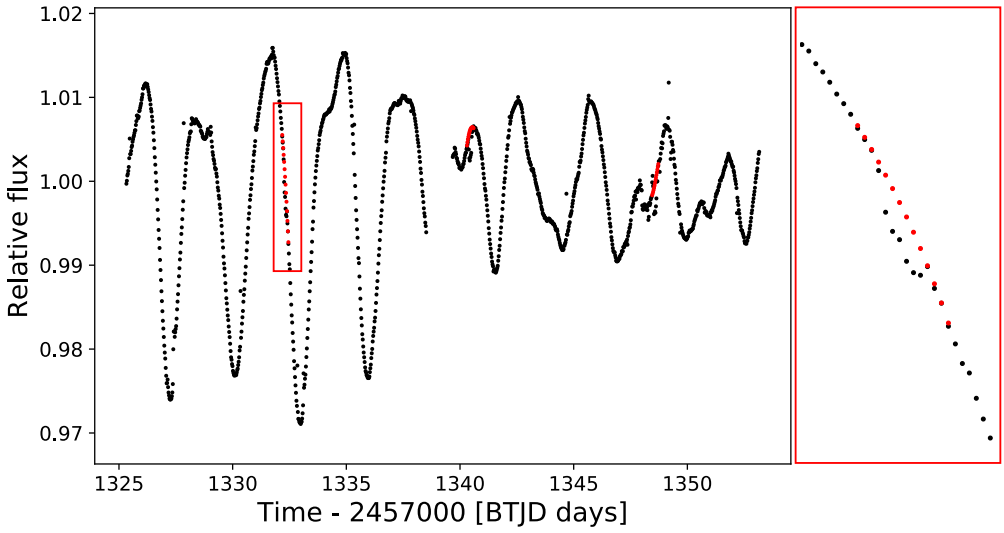

Figure 2.8: Example of interpolation over transit masked areas of the light-curve for DS Tuc A in order to best preserve the stellar activity signature through these segments. Original data points are shown in black, while interpolation over the removed transits is shown in red. The red box on the right shows detail of the area in the vicinity of the first transit.

is the `BoxLeastSquares` method within `astropy`[11], which is a `python` implementation of the computational method described by Hartman and Bakos [2016]. In this work the strongest BLS peak and the two next strongest non-harmonic peaks were investigated. In the future, the intention is to swap out this method to the more sophisticated Transit Least Squares method [TLS, Heller et al., 2019] to utilise its improved performance for finding smaller exoplanetary signals.

2.6.5 Injected transits

In order to test the sensitivity of this new detrending method to finding planets around young stars, a series of model planet transits were injected into the light-curves before detrending. For this analysis eight different orbital periods were tested (1.0, 2.0, 4.0, 6.0, 8.0, 10.0, 12.0 and 14.0 days) for each star, with the epoch chosen randomly for each injection. Note that unlike ground-based surveys, *TESS* does not exhibit 1-day systematic errors due to the Earth's rotation which could adversely affect these period choices. Rather than injecting specific planet sizes, five planet to star ratios R_P/R_* were tested for every star in this sensitivity analysis: 0.1, 0.075, 0.05, 0.04 and 0.03, beyond which recovery was observed to be quite rare.

[11] https://docs.astropy.org/en/stable/api/astropy.timeseries.BoxLeastSquares.html

For reference, around a sun-like star a radius ratio of 0.1 corresponds to an approximately Jupiter-sized planet and $R_P/R_* = 0.03$ corresponds to a sub-Neptune-sized planet. Stellar parameters for each star were retrieved from TIC v8 [Stassun et al., 2019]. Where possible, orbital separation for each planet was then derived from Kepler's Third Law. If information on a star's mass or radius was not available, the corresponding planet was assigned an orbital separation of $17.0R_*$, representing the average orbital separation for planets with 8d periods (in the middle of the period range) on the NASA Exoplanet Archive.[12] Planet transits were generated using Kreidberg [2015]'s `batman` python implementation of Mandel and Agol [2002]'s transit model, assuming non-linear limb-darkening with coefficients [0.5, 0.1, 0.1, -0.1]. After injection, each light-curve was detrended using a standard 30-bin run of the LOWESS-smoothing pipeline (without peak-cutting implemented). Injected signals were considered to have been 'recovered' if they appeared as one of the three highest peaks in the BLS periodogram, ignoring harmonics of the maximum peak.

2.7 Self-Organising Maps

As discussed in Section 1.6, analysing the large volumes of data from all-sky surveys such as *TESS* lends itself towards automated Machine Learning methods. This research made use of one particular unsupervised machine-learning method, Self Organising Maps (SOMs), which were introduced in Section 1.6.3. The SOM implementation used in this work is similar to the one used by Brett et al. [2004] and Armstrong et al. [2016], but is briefly explained here for context.

The SOM groups input data based on topology, placing data which are most similar in shape close together and separating those which are dissimilar. In this work the input data are photometric phase-curves formed by folding young star light-curves by their primary period. In order to balance computing efficiency with the preservation of complex phase-shapes, these phase curves were binned into 64 bins and normalised between zero and one (as is discussed further in section 5.2.3). Meanwhile the SOM is initialised as a square of 'pixels' comprised of vectors of the same length as the phase-curves (64 bins), but with elements with random values between zero and one. The collection of these vectors is referred to as the 'Kohonen layer'. To group the data, the input phase-curves are compared to the data at each pixel in the Kohonen layer and the pixel which contains the vector most similar to the input phase curve (named the 'Best Matching Unit', or BMU) is found by minimising the Euclidean distance between each input phase curve and the vector at

[12] https://exoplanetarchive.ipac.caltech.edu/

this pixel. Once the BMU has been located for each phase-curve, the phase-curves are effectively moved to this pixel, and the vector at each pixel is updated according to the equation

$$m_{xy,k,\text{new}} = m_{xy,k,\text{old}} + \alpha e^{-d_{xy}^2/2\sigma^2}(s_k - m_{xy,k,\text{old}}) \tag{2.3}$$

following the formalism of Armstrong et al. [2016], where $m_{xy,k}$ is the value m of the element at pixel coordinates x, y, element k in the phase curve, d_{xy} is the Euclidean distance of that pixel from the BMU in the layer, α and σ are coefficients called the 'learning rate' and 'learning radius' respectively, and s_k is the kth element of the input phase curve currently being compared. This process is repeated for each iteration, with learning rate and learning radius evolving as follows

$$\alpha = \alpha_0(1 - \frac{i}{n_{\text{iter}}}) \tag{2.4}$$

$$\sigma = \sigma_0 e^{(\frac{-ilog(r)}{n_{\text{iter}}})} \tag{2.5}$$

where i is the iteration number, r is the radius of the SOM (equivalent to the length of one dimension of the SOM) and n_{iter} is the total number of iterations (chosen as 1000 in this work). Note however that the SOM performance is relatively insensitive to the precise functional form of how α and σ evolve, as explored by Brett et al. [2004].

As the iterations continue, similar phase-curves steadily move closer together as the vectors at each pixel gradually become better representations of the main data shape at that point. The final result of the algorithm is the Kohonen map, a representation of the overall variation in the data-set. In this case each pixel effectively represents an average of those phase arrays which ended up at it after the final iteration. Querying individual pixels in the Kohonen map thus gives an approximation of each phase-curve at that pixel.

Three main parameters are important for the setup of the SOM: the initial learning rate, α_0, initial learning radius, σ_0 and the SOM shape. The learning rate controls how quickly pixels in the Kohonen layer are changed, while the learning radius controls how large the neighbourhood of comparison is during each iteration. Brett et al. [2004] recommended learning rate values of $\alpha_0 \leqslant 0.2$ in order to prevent undesirable movement and constant restructuring with further iterations, however the specific value of the learning radius was not found to have a significant effect on SOM performance. Similar to Armstrong et al. [2016], these parameters were

eventually chosen as $\alpha_0 = 0.1$ and $\sigma_0 = r$, where r is the SOM radius. Note that a 'radius' is defined here instead of a length, because the boundaries of the SOM are cyclic, wrapping around at each edge. An interesting feature of the SOM is that it can be arranged in any number of N dimensions, however following Armstrong et al. [2016], the nominal SOM shape was set to a 2D square of 40x40 pixels for a good balance of detail and required human eyeballing time. Variations on this shape are also explored in Section 5.5.2.

The specific software implementation used in this work was adapted from the TransitSOM code of Armstrong et al. [2017][13], which in turn was based on the SOM code from the PyMVPA Python package (Hanke et al., 2009).[14]

2.8 Exoplanet ephemeris maintenance

Discovery of exoplanets is merely the start of the process in exoplanet science. In order to fully understand the properties and evolution of exoplanetary systems, it is crucial to characterise and model individual exoplanets and observe if/how their properties change with time. As discussed in section 1.2.1, one of the most important pieces of information for the ongoing characterisation of a transiting planet is its ephemeris, which tracks the timing of a future transits of a known exoplanet. This is particularly important for future characterisation as many time-sensitive characterisation observations (such as atmospheric characterisation, orbital obliquity measurements and transit timing analyses) rely on precisely knowing the time of next transit. However, as discussed, the more time since the planet was last observed, the more the error in the expected time of transit grows, reducing the precision of this ephemeris. This timing challenge is occasionally complicated further by additional planets in the system which interact with the planet of interest, causing 'transit timing variations', where a planet transits earlier or later than expected from the linear ephemeris. Here the methods used to update exoplanet ephemerides and calculate transit timing variations (TTVs) are explained, forming the base of the ephemeris updates of all Kepler planets reobserved by *TESS* in Chapter 6.

2.8.1 Ephemeris updates and construction of O-C diagrams

The techniques used to determine individual transit times and update linear ephemerides in this work are based on those used by Gajdos et al. [2019] and Holczer et al. [2016] to analyse the full Kepler Q1-17 data. As a first step, transit epochs for plan-

[13] https://github.com/DJArmstrong/TransitSOM
[14] http://www.pymvpa.org

ets/candidates other than the one being currently analysed were masked from the light-curve. In some cases with large TTVs (such as Kepler-396 b - Xie [2014]), the masking window was widened to 3-6 hrs either side of the expected transit time to catch all of the transits. Because of the differing dilution characteristics and data cadence of the *Kepler* and *TESS* instruments, the *Kepler* transits were found first and then the results from this analysis were used to inform the initial *TESS* model.

To achieve consistent times for the two satellites, the time data for both datasets was converted to Barycentric Julian Date using the following conversions:

- $Kepler_{BJD} = BKJD + 2454833$ days
- $TESS_{BJD} = BTJD + 2457000$ days

where BKJD is the Barycentric *Kepler* Julian Date (the native time for the Kepler satellite) and the BTJD is the similar Barycentric *TESS* Julian Date.

For the first step in the *Kepler* analysis, a Mandel and Agol [2002] transit model was constructed for the object of interest using Kreidberg [2015]'s `batman` software, based off the archival parameters assembled in Section 6.3. These models were used to search for the observed transit time by minimising the chi-squared statistic in a grid of 1 min resolution around the expected transit. This method, similar to that used by Holczer et al. [2016], ensured that the approximate initial transit times could be obtained automatically even when there were significant TTVs present. These 1 min resolution transit times were used to construct a 'stacked' light-curve for each planet/candidate by aligning the obtained transit centres, similar to the method used by Gajdoš et al. [2017]; Gajdos et al. [2019] (see Figure 2.9). This stacked transit was compared to the transit curve obtained simply by folding the *Kepler* data by the known planetary period, with the cleanest transit curve taken forward for further analysis. The dual approach ensured that accurate transit curves could be obtained for targets both with and without transit timing variations. The chosen transit curve was fitted with a new `batman` model using three iterations of the in-built `optimize` routine within `exoplanet` [Foreman-Mackey et al., 2020]. Because the individual transit times here are considered more important than the overall system parameters, the planetary radius and inclination were allowed to vary slightly in order to fit the phase-folded transit most effectively. Note that the priors for both parameters were modelled as normal distributions with means equal to value reported in literature (or 90^o for systems without published inclinations). The mean value outside each transit was also allowed to vary in case it was skewed by outliers or detrending artefacts. An example of the final transit model can be seen for Kepler-68 b [Gilliland et al., 2013] in Figure 2.9.

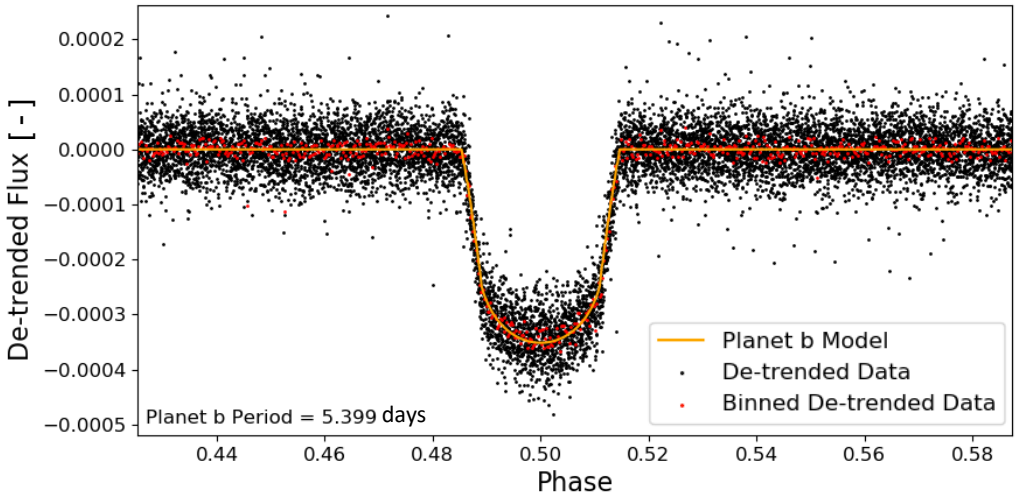

Figure 2.9: Stacked transit curve for Kepler-68 b overplotted with final `batman` model in orange.

The generated transit model was used to find precise transit times for individual transits in the *Kepler* data. This was achieved by cutting out short-duration windows around each expected transit (2 days either side of the transit, unless the planet period was $\leqslant$2 days) and finding the best-fit of the transit model within that interval. For consistency only the transit time and mean flux were allowed to vary in this step, the priors for both of which were set as wide Normal distributions centred on the values found from the chi-squared fit. To begin with, three iterations of `exoplanet`'s inbuilt `optimize` routine were used to hone in on the true transit times, before PyMC3 [Salvatier et al., 2016] was used to fit the final planet model to each individual *Kepler* transit. This was achieved using a two-chain Markov-Chain Monte Carlo (MCMC) analysis with 1000 tuning steps and 5000 draws for each object. Longer chains were tested briefly, but increasing the length further for these relatively simple fits was not found to change the results significantly. Convergence was assessed using the Gelman-Rubin statistic and a visual examination of the trace. Statistically significant values for the final time and error for each individual transit were then found from the mean and standard deviation of the MCMC trace.

Because of the comparative dearth of transits observed by *TESS*, broadly the same model built from the *Kepler* data was used to find the times of the individual *TESS* transits, however the depth of the transit was adjusted according to the local dilution characteristics of the *TESS* environments. Once again the model was fit to each transit in turn (in a 4-day data interval centred on the expected transit time) using three iterations of `exoplanet`'s `optimize` function and a `pyMC3` MCMC

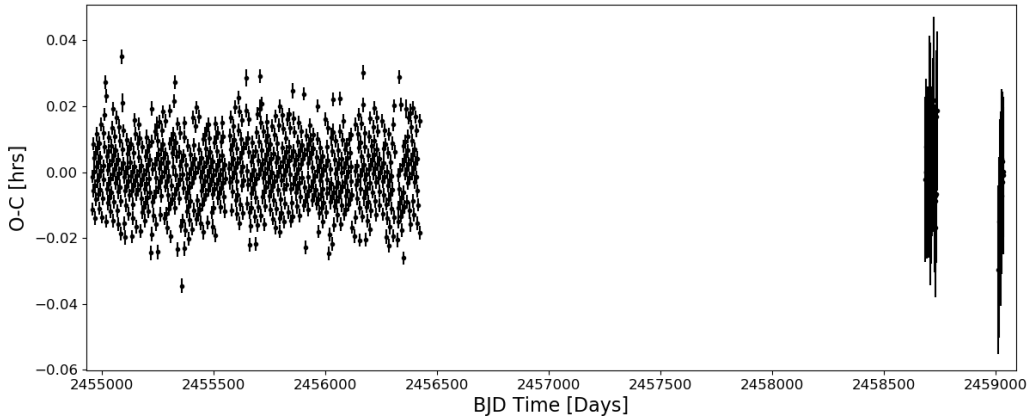

Figure 2.10: Example O-C plot for Kepler-13/KOI-13 b. K0I-13 b received *TESS* 2 min data in Sectors 14, 15 and 26, so represents one of the targets which received the most new data from *TESS*'s primary mission.

analysis with 10000 tuning steps and 10000 draws in order to gain statistically significant values and errors for each mid-transit time.

After the time of each individual *TESS* and *Kepler* transit was determined, the linear ephemeris based on the combined datasets was calculated. An uncertainty-based requirement of $\sigma_{transit} < 0.2$hrs was found to remove the majority of the questionable transits prior to further analysis (typically those which fell in data gaps or were incomplete), however the remainder were viewed by eye to catch any other transits which clearly had been fitted incorrectly. The remaining data points were fit using three iterations of a weighted linear least-squares fit while varying the initial transit time, T_0, and period P. The resulting ephemerides were used to generate an Observed-Calculated (O-C) plot for each object, providing a visual check for odd behaviour, significant TTV signals or significant outliers. An example O-C plot is shown in Figure 2.10 for KOI-13 b.

Because of the long data gap between the *Kepler* and *TESS* observation windows, one slight concern was the introduction of cycle count errors, i.e. miscalculating the number of transits between the *Kepler* and *TESS* transits. To check whether this was significant, the archival error in period derived from the *Kepler* data alone was multiplied by the number of missing transit cycles between the final *Kepler* transit and first *TESS* transit. In all cases this was found to result in a value at least two orders of magnitude smaller than a whole transit cycle. As a result, the ephemerides derived in this work are considered free of cycle count errors.

For any objects which did not display significant TTVs in their O-C diagrams (labelled with '0' in the TTV flag column of Table 6.1), an additional simultaneous

Kepler/TESS MCMC fit was carried out on the entire light-curve in order to narrow down the precision for these planets/candidates further. This was achieved by fitting the model based on the stacked transit parameters to the entire dataset and allowing t0, the mean out-of-transit value and the planetary period to vary. Once again, 1000 tuning steps and 5000 draws were used in this MCMC analysis, using pyMC3 [Salvatier et al., 2016] within `exoplanet` [Foreman-Mackey et al., 2020], following three `optimize` iterations.

2.8.2 TTV modelling

The updated systems in this work included a number of *Kepler* multi-planet systems known to exhibit significant TTVs, namely Kepler-18, Kepler-25, Kepler-51, Kepler-89, and Kepler-396 [Cochran et al., 2011; Steffen et al., 2012, 2013a; Masuda, 2014; Weiss et al., 2013; Xie, 2014]. Most of these systems were independently highlighted as worthwhile systems for further TTV analysis by Goldberg et al. [2018] and/or Jontof-Hutter et al. [2021]. In an attempt to update the TTV masses for planets in these systems, additional TTV fitting was carried out. Midtransit times for each planet in a system were simulated using `TTVFast`, a numerical integrator for computing transit times given a set of planet masses and orbital parameters [Deck et al., 2014]. With these simulations, an MCMC analysis explored the parameter space to find the best-fit masses and orbits describing the observed transit times, both before and after the addition of *TESS* datapoints.

Each transiting planet in a system was fit for mass, μ_i (in units of solar masses), orbital period, P_i, orbital eccentricity and argument of pericenter (via eccentricity vectors $h_i = \sqrt{e_i}\cos\omega_i$ and $k_i = \sqrt{e_i}\sin\omega_i$), and time of first transit, T_i, where $i = 1, 2, ..., N$ and N is the number of planets. Coplanar orbits were assumed. Guess parameters were estimated using a Levenberg-Marquardt least-squares algorithm, and datapoints more than 4σ from this initial best-fit solution were marked as outliers and removed from the data. MCMC analysis was then performed using the affine invariant ensemble sampler `emcee` [Foreman-Mackey et al., 2012], with 100 walkers initialised in a tight ball around the guess parameters. Each walker was run for 200,000 steps, and the first 50,000 steps were discarded as burn-in.

Following the procedure of Hadden and Lithwick [2016, 2017], both "default" and "highmass" priors were considered, recognising that TTV fits are often challenged by mass-eccentricity degeneracies. Under the default run, mass was assigned a logarithmic prior

$$p(\mu) \propto \begin{cases} (\mu + \mu_0) & \mu \geqslant 0, \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

with $\mu_0 = 3 \times 10^{-7}$ to prevent divergence at $\mu \to 0$, while eccentricity was assigned a uniform prior

$$p(h,k) \propto \begin{cases} (h^2 + k^2)^{-1/2} & (h^2 + k^2)^{1/2} < 0.9, \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

with an upper cutoff at 0.9 to avoid the need for extremely small time steps in the `TTVFast` integrations. For the high-mass run, mass was assigned a uniform prior

$$p(\mu) \propto \begin{cases} \text{constant}, \mu \geqslant 0, \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{2.8}$$

while eccentricity was assigned a logarithmic prior

$$p(h,k) \propto \begin{cases} (h^2 + k^2)^{-1/2} (\sqrt{h^2 + k^2} + e_0)^{-1} & (h^2 + k^2)^{1/2} < 0.9, \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

with $e_0 = 10^{-3}$ to prevent divergence at $e \to 0$. In both runs, period and initial transit time were assigned uniform priors.

Chapter 3

Expanding the population of young stars with *Gaia*

"The [stars] hung in the sky in much the same way that bricks don't"

— Douglas Adams, Adapted from *The Hitchhiker's Guide to the Galaxy*

In order to search for new young transiting exoplanets, it is critically important to have sufficient young stars to search around. Until very recently this has been challenged by the lack of precise kinematic data for all but the brightest stars. However, the launch of the *Gaia* satellite in 2018 has now allowed the community to determine high precision positions and velocities for over a billion stars in the Milky Way. This new kinematic data has revealed hundreds of thousands of additional stars which are associated with young stellar groups. This chapter presents two related pieces of work focused on expanding the known population of young stars, both through interacting directly with *Gaia* data to extend the lower-mass population of several young stellar associations and through the assembly of a large catalogue of known young stars from the literature.

Declaration - This chapter includes a partial reproduction of work reported in the PhD 6-month report, submitted for examination in April 2019. However, none of it has previously been published in the public domain. The analysis in section 3.2 was completed solely by the author, however the construction of young star list discussed in section 3.3 was built in collaboration with Dr Ed Gillen at Queen Mary University of London. In this case both Dr Gillen and the author were responsible

for adding approximately half of the catalogues to the full list of young stars.

3.1 Motivation

A key challenge in the search for a significant population of young exoplanets is the historical lack of a large enough sample of 'bona-fide' young stars around which to search. Basic transit statistics suggest that only ∼0.47% of sun-like stars are expected to host Hot Jupiters [Haswell, 2010], so considering for example that only 3076 stars exist in the extended BANYAN sample of stellar association members (some of which were not even observed in *TESS*'s primary mission), at most 15 such planets could be expected to be found around these young stars. Coupling this with the increased stellar activity of host stars at such young ages, finding even single planets within these associations becomes a significant challenge. Bouma et al. [2019] recently attempted to remedy this situation by concatenating 13 different catalogues of young stars and cluster members from literature, yielding 1,061,447 individual target stars. However, because of the mix of catalogues used and the non-homogeneous membership criterion applied, this may include some stars that are not truly young. Another promising method for expanding these target lists is expanding cluster and association membership through proper- and galactic-motion relationships. This has recently become much more reliable with the release of a significant amount of high precision astrometric data from the *Gaia* satellite [Gaia Collaboration et al., 2016b, 2018a]. Groups such as Damiani et al. [2019b] and Lodieu et al. [2019] have led the way in such analyses, using astrometric measurements from *Gaia*'s second data release (DR2; Gaia Collaboration et al. [2018a]) to expand the Sco-OB2 and Hyades associations. These successes motivated the work described here to expand the population of young stars around which to look for young exoplanets, both through new kinematic analyses and through assembling previous catalogues of young stars from literature.

3.2 *Gaia* Kinematic analyses

As stars in stellar associations and moving groups are typically closer, brighter and in less dense environments than those in open clusters, they are particularly valuable stars to search for planets around with *TESS*. While previous stellar association analyses have been restricted to the brightest members observed to move across the sky together, the high precision astrometry of the *Gaia* mission allows for studying the potential membership of lower mass members too.

Gaia data was accessed using the *Gaia* Archive[1], using a series of Astronomy Data Query Language [ADQL] searches to obtain data for all stars within certain rectangular sky areas of interest. In order to only obtain stars with reliable three-dimensional position, an additional condition that the ratio of parallax to its error was >10 was imposed (thus ensuring that the relative error in distance was less than 10%). In total, four stellar associations were analysed in this work, beginning with the 10-15 Myr Sco-OB2 association to validate the astrometric methods, before moving on to the β Pictoris Moving Group (BPMG; 20-26 Myr) and the Hyades (∼650 Myr) and Octans (30-40 Myr) stellar associations.

3.2.1 Method validation using Sco-OB2

The initial sky area surveyed was that around the Scorpius OB2 (Sco-OB2) stellar association, chosen to allow direct comparison to the work of Damiani et al. [2019b] and validate the expansion methods. Following the exact spatial selections of Damiani et al. [2019b], three overlapping (RA,Dec) boxes were selected with diagonal vertices of ([144.8^o,221.1^o],[-70.8^o,-38.8^o]),([203.2^o,259.9^o],[-61.6^o,-20.9^o]) & ([229.6^o,257.1^o],[-36.3^o,-11.0^o]), with an imposed coarse parallax (π) limit of $\pi > 5$ mas (nominal distance $d < 200pc$) to weed out stars further than any known members of the Sco-OB2 association. This resulted in 271,395 unique *Gaia* DR2 sources. Plotting a density plot for these sources in the proper motion plane (using a Gaussian kernel density estimator with 100 boxes in each axis) yielded the interesting result shown in Figure 3.1. Note that because most stellar associations are very close to Earth and hence cover a large range of the sky, proper motion in galactic coordinates (pm_l_cosb and pm_b, measured from the sun as the origin)[2] was found to be be more reliable than *Gaia*'s natural equatorial coordinates (pm_ra and pm_dec), which is a geocentric coordinate system with positions and velocities effectively based on the how the stars fall on the night sky. Hence because stellar associations cover much smaller ranges of galactic space than they do of equatorial space, the heliocentric galactic coordinate system allows different stellar associations to be separated more clearly into different groups of proper motion space compared to the equatorial coordinate system.

Two distinct regions of increased proper-motion density are immediately obvious in Figure 3.1, as well as some underlying structure in these regions. While the clustering centred around [pm_l_cosb,pm_b] = [0,0] can be explained by the motion of field stars, the second region centred around [-25, -10] is suggestive of a stellar asso-

[1] http://gea.esac.esa.int/archive/

[2] n.b. 'l' and 'b' here denote galactic longitude and latitude respectively

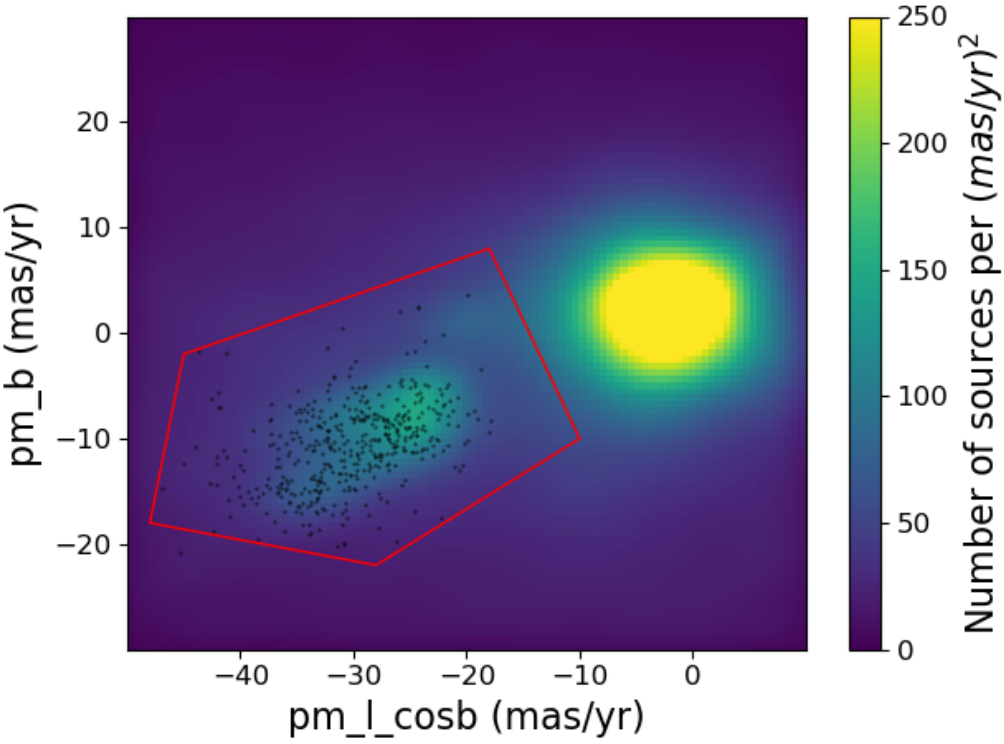

Figure 3.1: Density plot of *Gaia* DR2 sources in the vicinity of the Sco-OB2 association, presented in the galactic proper motion plane. Overplotted black dots represent accepted Sco-OB2 members from de Zeeuw et al. [1999]'s Hipparcos examination of this association. The red selection polygon shows the region of proper motion space selected for further narrowing of the sample. Note that for clarity the top of the colour-map is capped at 250, but the true density at the centre of the field star sample at [0,0] is over 400 (mas/yr)^2.

ciation: a set of stars with common proper motion across the sky. Over-plotting the proper motion plot with previously accepted Sco-OB2 members from the de Zeeuw et al. [1999] survey suggests that this association may be the source of the second over-density. In order to investigate this further, the sources within the red polygon in Figure 3.1 (enclosing the likely Sco-OB2 members) were selected and a colour-absolute magnitude diagram (CAMD) was constructed in density form, as shown in Figure 3.2. Note that the vertices of the red polygon were chosen to encompass all previously known members of the Sco-OB2 association (from de Zeeuw et al. [1999]), following the method of Damiani et al. [2019b]. In order to retain data consistency, the CAMD plotted in Figure 3.2 was based on *Gaia* G-band magnitudes and *Gaia* BP - RP colours from *Gaia* DR2.

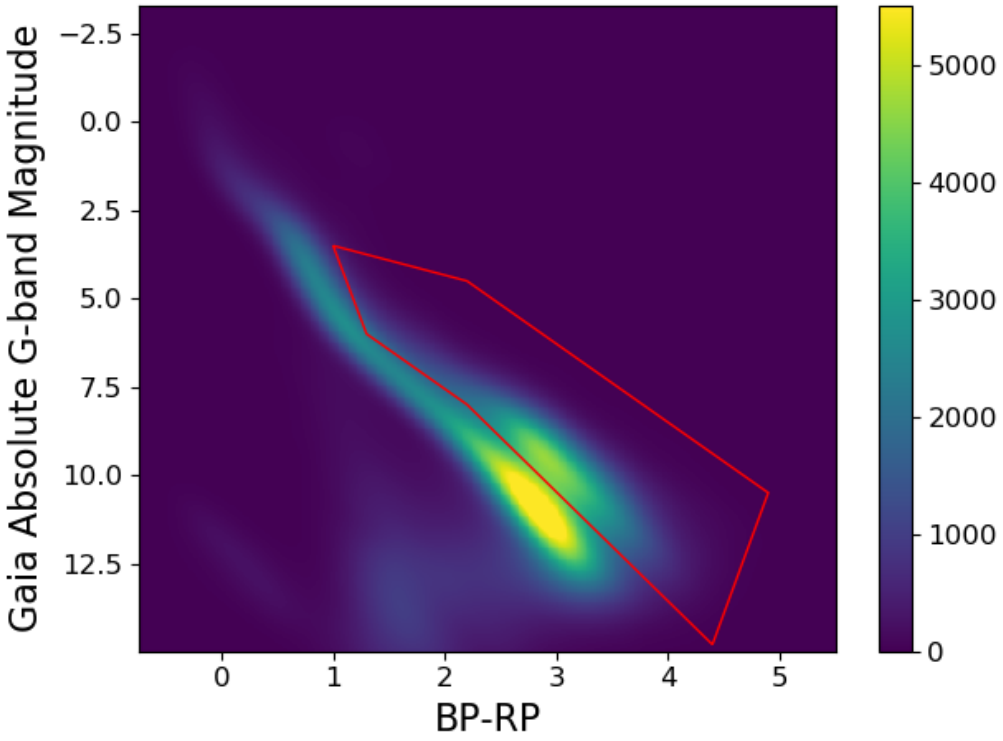

Figure 3.2: Colour absolute magnitude diagram for the area around Sco-OB2. Of particular interest is the significant bulge to the top right of the main sequence, suggestive of stars on the pre-main sequence. The red polygon shows those stars selected as the pre-main sequence sample for further analysis, tracing the area above the main sequence at its base and extending far out above this to catch even the youngest stars. Note that the colour-scale represents stellar density in units of number of stars per mag^2.

Colour-magnitude diagrams are useful probes of stellar evolution. Stars begin their life as young protostars to the upper right of the main sequence, before dropping onto the main sequence at the 'zero-age main-sequence' around the bright yellow region of Figure 3.2. These stars evolve along the main sequence (up and to the left along the CAMD) until they reach what is known as the 'main-sequence turn-off' after having burnt the majority of their hydrogen. Observing the shape of the overall CAMD can thus be a useful first-look at the age of a group of stars. What is particularly interesting about this CAMD is the significant over-density (or bulbous section) to the top right of the lower main sequence. This is suggestive of a significant number of young, pre-main-sequence stars which are yet to begin their lives on the main sequence. One can select the youngest stars by applying a sec-

ond polygon selection, this time selecting the stars falling in this younger region to the upper right of the main sequence. Because lower-mass stars evolve more slowly than high-mass stars, most of these stars consist of smaller, lower-mass stars. Such stars are much easier to find planets around than the large O and B-type stars of the original stellar association members, largely because of the increased planet to star mass/radius ratios and narrower, more common spectral lines. The result of applying the applied additional cut is shown in Figure 3.3, in the form of a final density plot in proper motion space. Following this cut, a total of 10,989 sources remained, highlighting the wealth of potential lower-mass members within this stellar association.

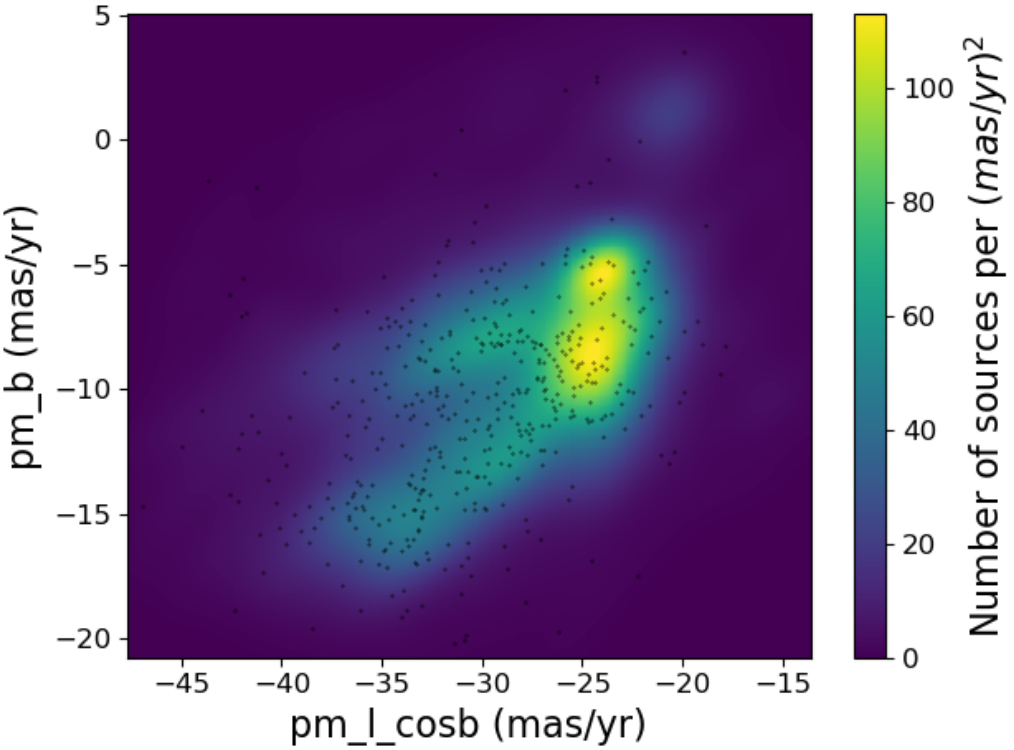

Figure 3.3: Final proper-motion diagram for Sco-OB2 candidate members, with likely Sco-OB2 members from the de Zeeuw et al. [1999] survey overplotted as black dots. Notable sub-structure is present in the lower-mass members, suggestive of a denser region around [-24,-10] and a more diffuse cloud towards the bottom left.

There are several important observations that can be drawn from Figure 3.3. Firstly, there appears to be a finer structure to the distribution of stars in this association in proper-motion space, suggesting that there may be denser regions of star formation within the association. Damiani et al. [2019b] observed similar

structures in their analysis and suggested the presence of a compact core with a diffuse halo towards the bottom left of the association. Interestingly, a second observation is that the previously accepted 'bona-fide' members of Sco-OB2 from de Zeeuw et al. [1999] (overplotted as black points) do not appear to directly align with the over-densities in the overall population, instead demonstrating a relatively even distribution. Finally, and most importantly for this work, there appear to be considerably more *Gaia* DR2 sources with similar proper motions than are accounted for by the previously accepted de Zeeuw et al. [1999] members in this association (10,989 stars in this sample, compared to 400 members in the de Zeeuw et al. [1999] paper). These observations suggest that there is a lot to be learnt from further analysis of stellar associations in the era of *Gaia* DR2, and that this dataset is likely to allow the community to find considerable numbers of additional stars associated with these groups (especially lower-mass stars). This has the potential to result in much larger target lists for studies of young stars and their associated planets. Meanwhile, the sample of 10,989 pre-main-sequence stars found in this work aligns well with the ∼11,000 pre-main-sequence stars found by [Damiani et al., 2019b], validating this method as an efficient and effective method of expanding the low-mass population of stellar associations with *Gaia*.

Additional checks can be applied for membership of some stars by moving from 2D proper-motion space to 3D Galactic velocity space. While the above proper-motion cuts ensure that all selected stars have similar positions and motions across the sky in two dimensions, it does not say anything about the respective distances and radial velocities of these objects. As a simple first check distances were observed for all *Gaia* DR2 sources using the Bailer-Jones et al. [2018] catalogue of *Gaia* DR2 distances in order to cut out any particularly obvious outliers. However, for objects with radial velocity measurements (and thus full 6D kinematics) it was found more instructive to convert all spatial and velocity measurements into the galactic coordinate system, thus gaining a homogeneous system of XYZ galactic positions and corresponding UVW galactic velocities. It was then easy to plot the objects in positional space with overlain galactic velocities as arrows in order to immediately see any spurious signals. An example for this in galactic (Y,Z) coordinates is shown in Figure 3.4 for Sco-OB2 members from the compiled list with known radial velocities. As a first cut those members with positions or galactic velocities greatly different to those around them can be removed as candidates, while if more confidence is desired one can construct gaussian distributions for each of the XYZUVW parameters and remove those over 3σ from the mean (similar to the analysis completed by Gagné et al. [2018b]). However, this full method required

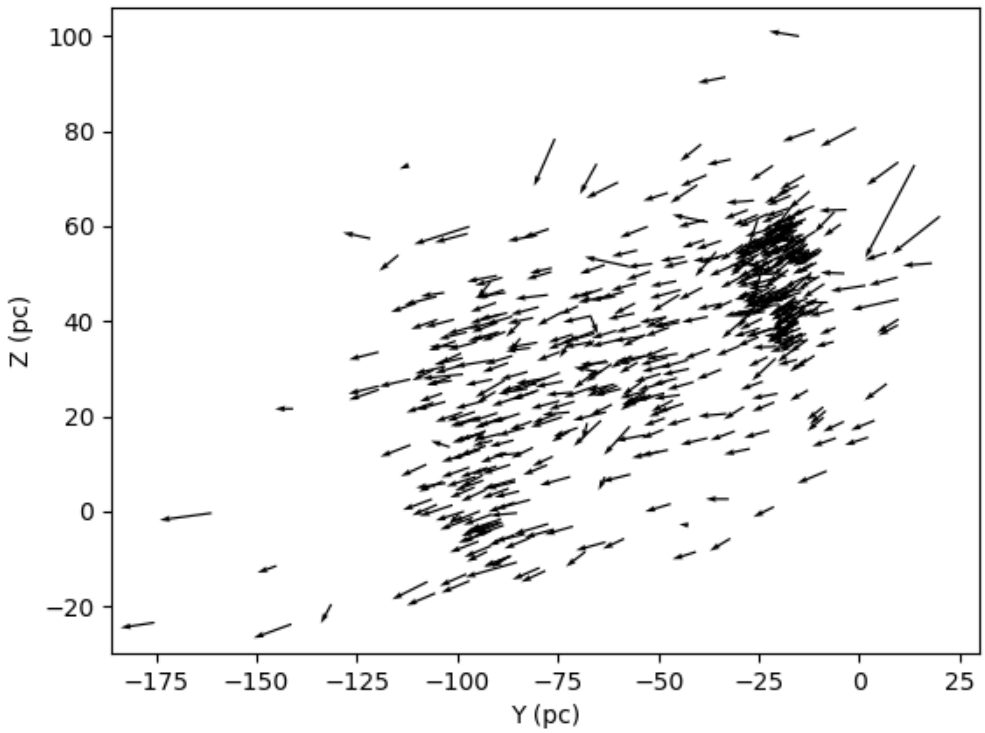

Figure 3.4: Galactic position plot (YZ plane) of potential Sco-OB2 members with overlaid VW galactic velocities. This diagram can be used to visually inspect how well individual members fit within the kinematics of the wider population of the Sco-OB2 association.

high-precision radial velocities to be known for all objects, which are simply not known for many dimmer *Gaia* sources at present. Hence the simpler method of 2D proper motion cuts is currently preferred for constructing wider target lists, at least until more radial velocities become available in future data releases from the *Gaia* Mission.

3.2.2 Application to Hyades, BPMG and Octans

Three further stellar groups were examined using the techniques described above in this work: the Hyades open cluster, Beta Pictoris Moving Group (BPMG) and Octans stellar association. All of these systems have cores with previously known proper motions, distances and approximate positions, and hence in theory should be able to be analysed like Sco-OB2 above.

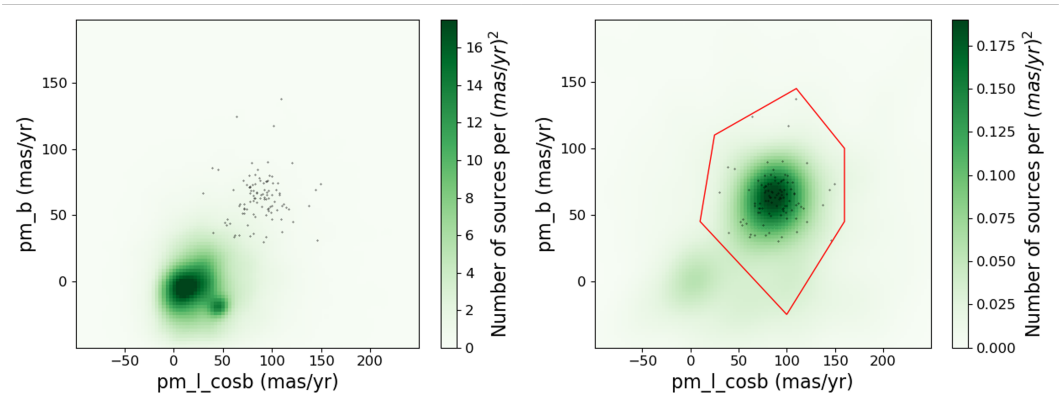

Figure 3.5: Proper motion density plots for the area around the Hyades open cluster without distance limits imposed (left) and with distance limits of 37-57 pc (right). Accepted Hyades members from the Röser et al. [2011] deep sky census of the Hyades are overplotted in black, with the selection for the final Hyades sample shown with the red polygon.

Hyades

Initial data for the Hyades was drawn from the *Gaia* archive by selecting stars with right ascensions (RAs) between 40^o and 80^o and declinations (Decs) between -5^o and 35^o. The resulting galactic proper motion plot is shown in the left of Figure 3.5. Unfortunately given that the Hyades is much more diffuse than younger associations like Sco-OB2, the field star population in this region greatly outweighs the strength of the over-densities near the Hyades cluster. In order to remove the bulk of the surrounding field star population and focus on those members closest to (and most likely to be associated with) the Hyades, distance limits were applied to the dataset at $\pm$5, 10, and 20 pc either side of the literature distance between the Earth and the Hyades ($\sim$47 $\pm$ 4 pc) [Perryman et al., 1997; Gagné et al., 2018b]. Testing different distance increments from this mean distance revealed that a distance cut of 10 pc (i.e. d = 37-57 pc or 47 pc $\pm$ 2.5σ) either side of this value made the population around the Hyades association clearest whilst reducing the effect of nearby field stars. Applying these cuts and plotting the proper-motion diagram in galactic coordinates as before revealed the distribution of proper motions shown in the right side of Figure 3.5. In this figure a defined offset region of increased density is clearly evident at about [pm_l_cos_b,pm_b] = [100, 70] (clearly associated with the known Hyades members from Röser et al. [2011] which are overplotted as black dots), alongside the usual field star sample at [0,0]. Applying the proper-motion cuts shown by the red polygon (essentially encompassing all previously known members and the over-

density seen around [100,0]) resulted in the final sample of stars shown in proper motion space in Figure 3.6. Alike to Sco-OB2, this final distribution clearly shows additional substructure in the overall distribution of stars, not all of which is directly related to the previously known population of the Hyades.

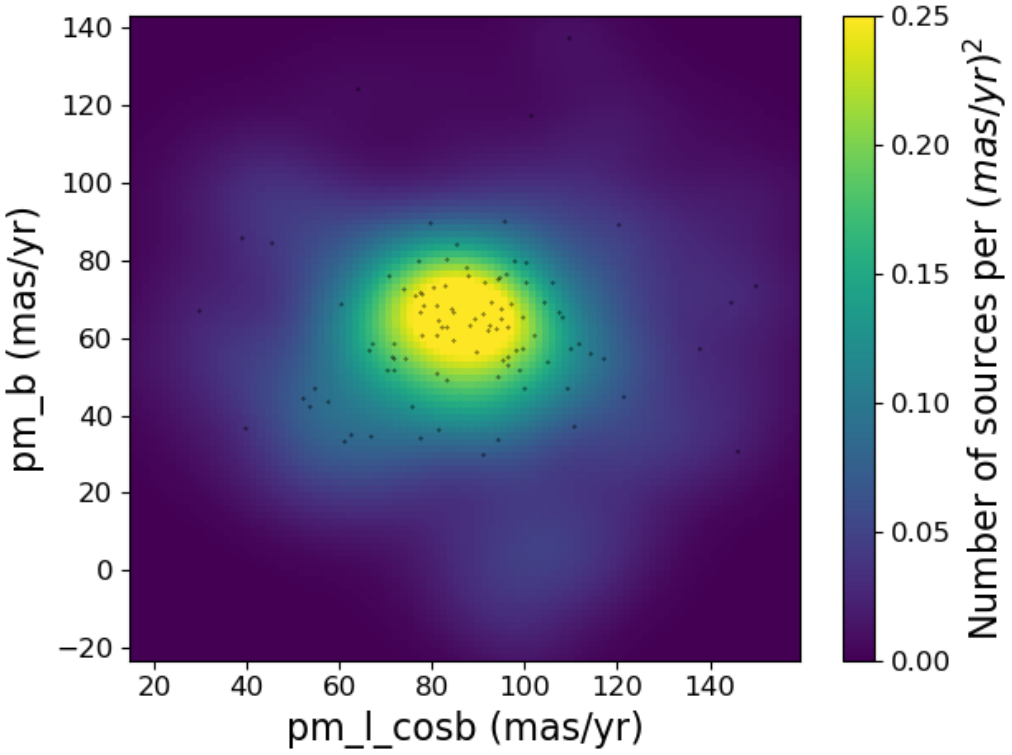

Figure 3.6: Proper-motion plot for final selection for Hyades members in galactic coordinates. As before, the density of candidate members is represented in colour, with the previously known Hyades members from [Röser et al., 2011] plotted as black dots. Similar to Sco-OB2, a slight substructure can be seen in the lower-mass members of the Hyades, not directly related to the distribution of previously confirmed members. Note however the decreased density overall (and correspondingly large spread in proper motion) compared to the other younger associations studied here.

Unfortunately, unlike the young (∼10-15 Myr) Sco-OB2 association, the comparatively old age (∼650 Myr) of the Hyades association means that the shape of the association's colour-magnitude diagram (see Figure 3.7) is no longer useful to siphon off the youngest stars, as almost all stars in the association now exist on the main sequence of their evolution. This means that other methods are required to determine which of the stars in the association are youngest and smallest, and that more care may be needed to weed out additional old interlopers with similar

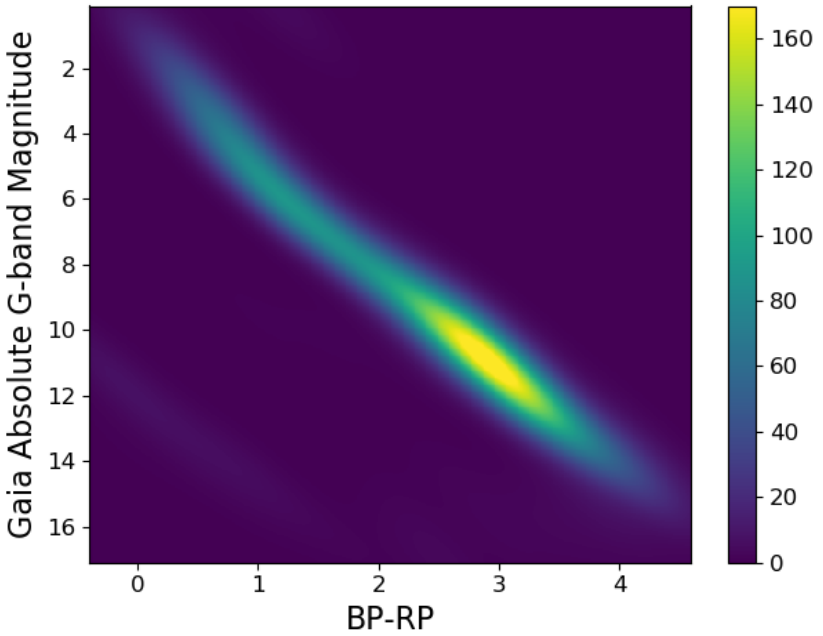

Figure 3.7: Final colour-absolute magnitude diagram for Hyades cluster. As in Figure 3.2, the colour-bar has units of sources per mag^2.

kinematics. Note that since the completion of this work a more in-depth census of the Hyades association has been completed by Lodieu et al. [2019], so a full target list from the present analysis is not included here.

BPMG and 3D structure of stellar associations

In the case of the Beta-Pic Moving Group (BPMG) the results were less promising. The particularly young age of this association initially made this group a promising candidate for further analysis. However, because of the proximity of this moving group and its correspondingly large spatial extent on the sky, it was not possible to constrain it to a small enough area of the sky (or indeed area of proper motion space) to allow significant cuts to be made through the above graphical membership method. This challenge is well illustrated by Figure 3.8, showing the spread in galactic proper motion for the BPMG association. Particularly challenging is the fact that the current members of this association actually encircle the field star population, making it implausible to remove them without further information. For this cluster a membership method based on the full galactic XYZ positions and UVW velocities is recommended instead.

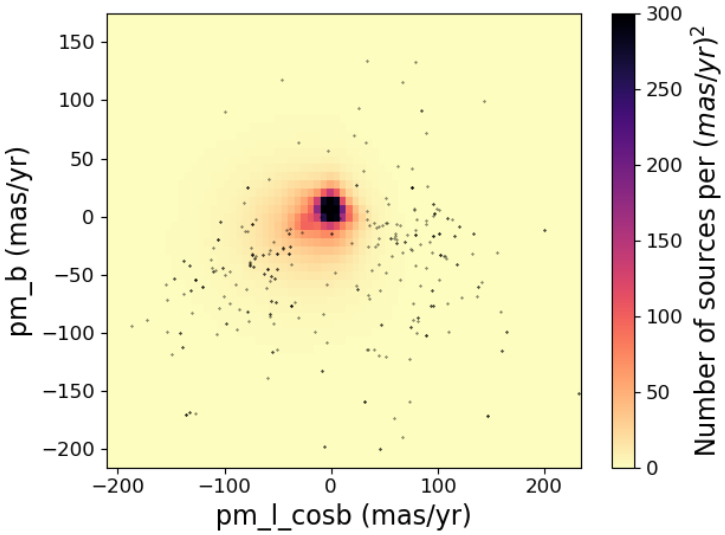

Figure 3.8: Proper-motion plot for area around Beta-Pic Moving Group (BPMG). Previously accepted members of BPMG from Gagné et al. [2018b] are plotted as black dots, with the number of sources per $(mas/yr)^2$ represented in colour. In this case the spread of the group is so large and diffuse (and situated either side of the main field star population at [0,0]) that the graphical methods used elsewhere are no longer appropriate for use.

Nonetheless, the challenge of BPMG spurred an interesting look into the 3D galactic structure of known associations, at least for those members with measured radial velocity data. Beginning with target lists from the BANYAN sample of stellar associations [Gagné et al., 2018b], all bona-fide members were cross-matched with their *Gaia* counterparts in order to find locations, proper motions, distances (from the Bailer-Jones et al. [2018] catalogue) and, where possible, radial velocities for these targets. All those with full information were then converted to 6-dimensional galactic XYZ UVW space in order to investigate the comparative positions and velocities of these stellar associations. An example plot of the distribution of galactic velocities (UVW) is shown in Figure 3.9. This analysis suggested that particularly well-separated associations to expand would be the Ursa Major Cluster, Coma Berenices, Octans and Taurus-Auriga associations (labelled respectively as UMA, CBER, OCT and TAU in Figure 3.9), as well as the Hyades which has already been examined.

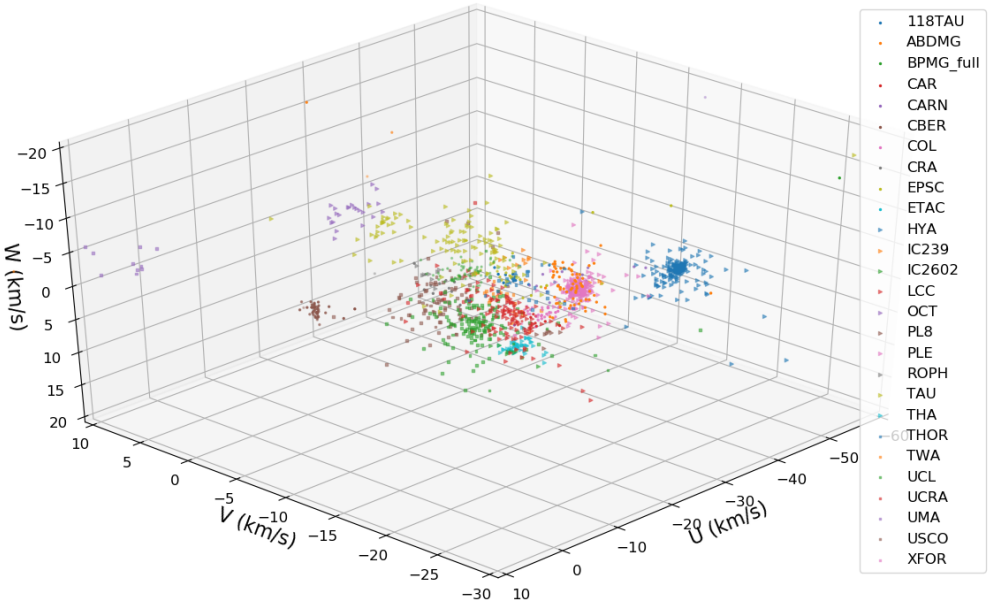

Figure 3.9: 3D velocity overview (in galactic UVW coordinates) of all 'bona-fide' stellar association members from Gagné et al. [2018b] with radial velocities measured by *Gaia*. Individual clusters are represented by different colours and shapes, according to the legend on the right. The abbreviated names for each cluster are taken directly from Gagné et al. [2018b] and explained in Table 4.1.

Octans

Of these associations, Octans is considered the most interesting for planet searches, being 35±5 Myr in age [Gagné et al., 2018b], right in the middle of a highly dynamic period of exoplanet evolution (0-100 Myr). In order to capture stars in the region of Octans, initial sky position boxes with diagonal vertices of ([50^o,-56^o] to [111.25^o,-20^o]) and ([0^o,-90^0],[360^o,-56^o]) were queried from the *Gaia* Archive, selecting all stars with parallaxes greater than $\pi > 4$. These stars were then further restricted to those nearest in proper-motion by imposing cuts of -40 $\leqslant$ pmra $<$ 40 and -30 $\leqslant$ pmdec $\leqslant$ 50. The remaining stars were plotted on a galactic proper-motion diagram, yielding the distribution of stars shown in Figure 3.10.

Similar to Sco-OB2 and the Hyades, the known Octans members displayed a distinct offset from the field star population at [0,0], centred around approximately [pm_l_cosb, pm_b] = [-20,-10] mas/yr. Performing the selection cut shown by the red polygon on Figure 3.10 resulted in the overall proper motion distribution pictured

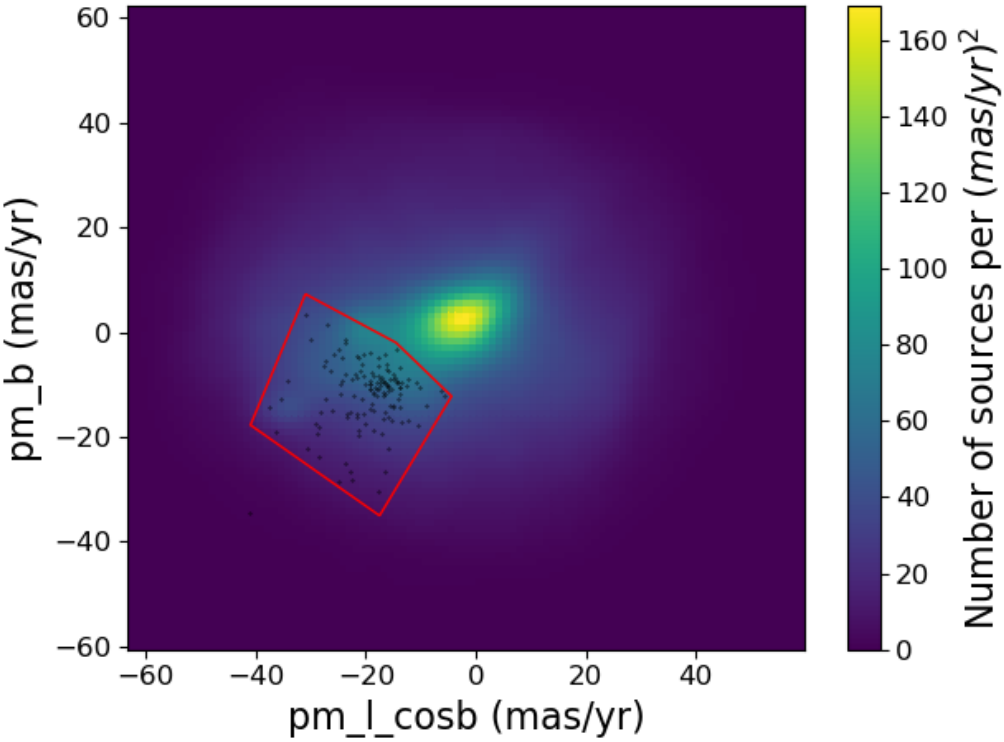

Figure 3.10: Galactic proper-motion plot for stars in the region of the Octans stellar association. Previously accepted members of the association from Gagné et al. [2018b] are plotted as black dots, with the density of stars illustrated in colour. The red polygon illustrates the region of proper motion space taken forward for further analysis, situated around the previously known members.

in Figure 3.11. However, this distribution appeared to have a suspiciously large number of distinct populations within it. This peculiarity was explored further by converting the data to Galactic (XYZ) coordinates for position, and overplotting the galactic position for not only the Octans association, but also other nearby stellar associations from the BANYAN catalogue [Gagné et al., 2018b], as pictured in Figure 3.12.

As can clearly be seen in Figures 3.9 and 3.12, the nearby IC239, IC2602, EPSC and LCC associations all have members near Octans both in galactic position and velocity, providing explanation for some of the densest regions in this vicinity. Nonetheless, there still appears to be a significant region of the galactic positional space filled primarily with known Octans members. Selecting these members explicitly by capturing all stars within the area of the red polygon in Figure 3.12 resulted in 14,050 individual stars, with the proper-motion distribution shown

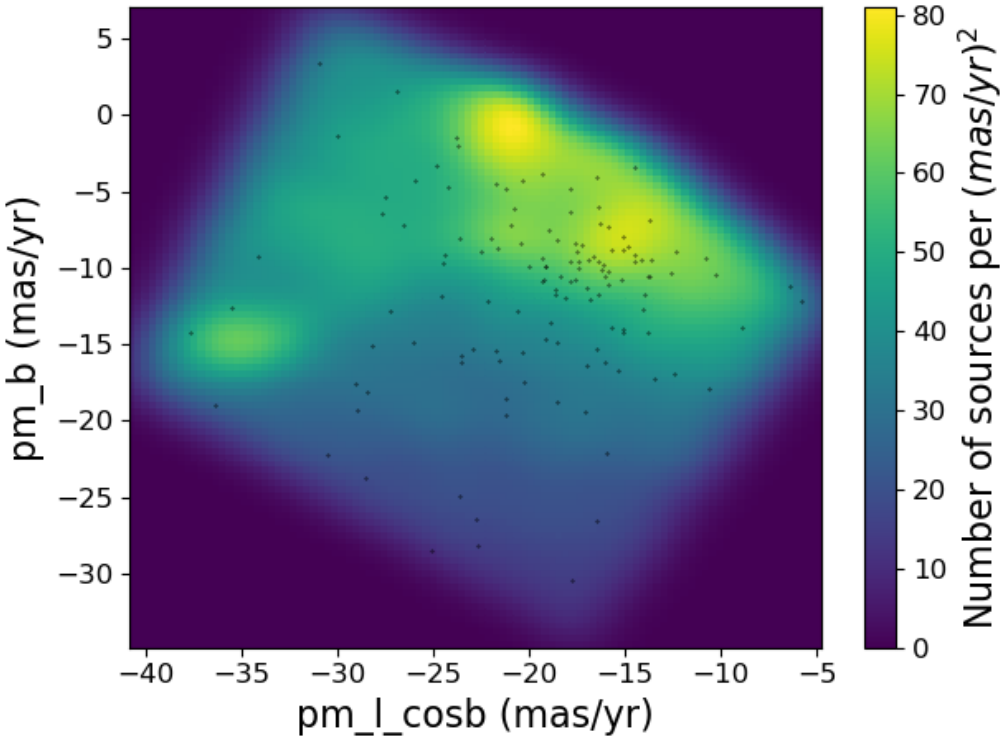

Figure 3.11: Proper motion plot for the region within the selection polygon. Colours and black dots have the same meaning as in Figure 3.10. Multiple over-densities appear to be present in proper-motion space, warranting further investigation.

in Figure 3.13. Once again, previously known members of the Octans association (from Gagné et al. [2018b] and [Murphy and Lawson, 2015]) are overplotted as black dots. Similar to the case of the Hyades and Sco-OB2, the new *Gaia* data shows increased substructure within the prospective association members which is not seen nearly as significantly in the previously known population. The colour-absolute magnitude diagram for the remaining Octans candidates is pictured in Figure 3.14. Unfortunately because Octans is older than Sco-OB2, clear pre-main sequence stars cannot easily be seen in this case, so other independent signs of youth will be required to confirm the youth of these members. Additional statistical cuts with limits of mean $\pm$ 3σ from the average galactic (XYZ) position and (UVW) velocities of the association were briefly trialled as an additional method to cut down this list further, but given the odd substructure and potential extended filamentary strings of the association, the graphical method is considered more reliable than arbitrary mean/standard deviation cuts in this instance.

The construction of this target list thus increases the number of stars asso-

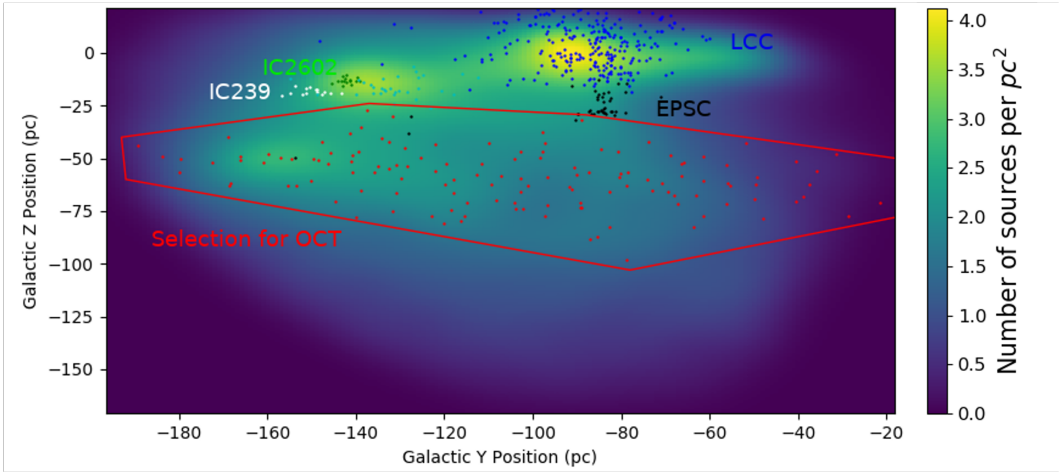

Figure 3.12: Galactic YZ position plot for stars within selection polygon. Stellar density is represented as the background colour, with previously known members of Lower-Centaurs Crux (LCC), ϵ Chamaeleontis (EPSC), IC2602, IC239 and Octans (OCT) overplotted as blue, black, green, white and red dots respectively. The final positional selection for Octans members is shown by the red polygon, chosen to encompass all previously known members of Octans from Gagné et al. [2018b].

ciated with the Octans association by a hundred-fold (from 138 to 14,050), which evidences the power of *Gaia* data for significantly expanding the lower-mass population of stars associated with stellar associations. Such an expansion of members represents a very important step forward in the search for young exoplanets. As can be seen in Figure 3.15, the luminosity distribution of the final sample of 14,050 targets is more uniform than the original distribution and much more closely resembles that expected from the known luminosity and initial mass functions of stars [Salpeter, 1955; Chabrier, 2003]. The full catalogue of these candidate members of the Octans association is available online.[3]

The effectiveness of these relatively simple graphical methods is promising for expanding the population of young exoplanets in coming years. However, each of these young star candidates still needs an independent sign of youth before it can be accepted as a 'bona-fide' member of a stellar association. While the position of stars on the colour-absolute magnitude diagram can give this in the youngest clusters (as was seen for Sco-OB2), for most stars between 50-1000 Myr, this independent sign of youth is currently the biggest bottle-neck in expanding the known sample of young stars (especially since many confirmation methods require individual spectroscopic follow-up). Given this bottle-neck it is currently considered wise to retain all high-

[3]https://warwick.ac.uk/fac/sci/physics/research/astro/people/matthewbattley/Final_Octans_Data.csv

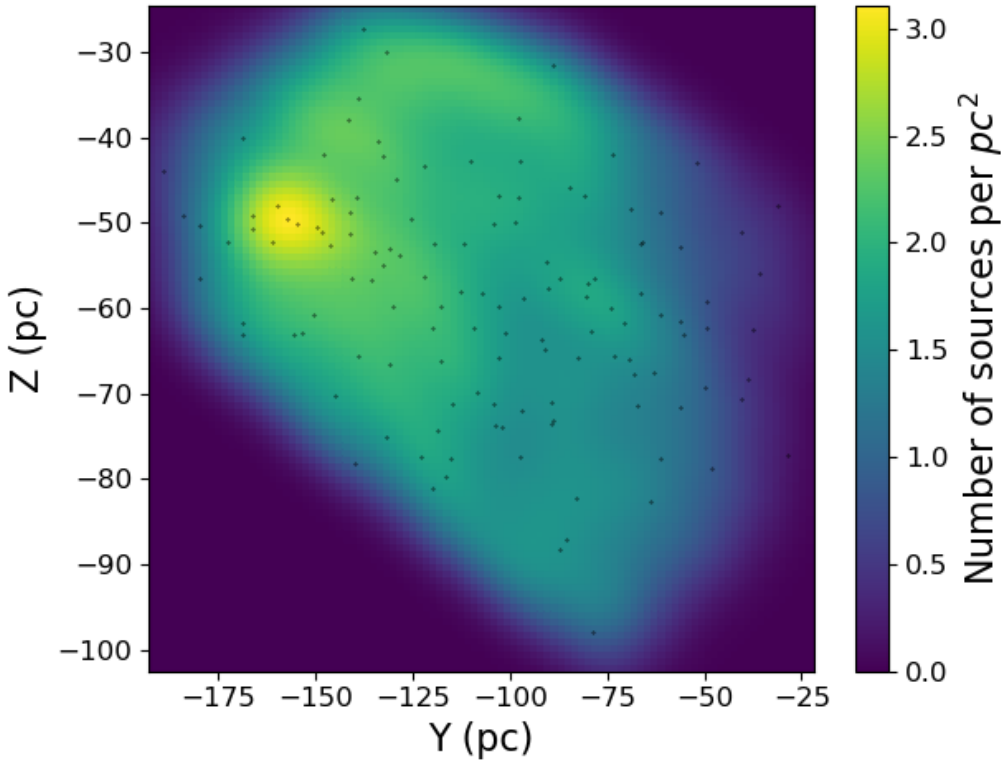

Figure 3.13: Final distribution of new candidate members of the Octans association in galactic proper-motion space. Black dots and background colours are as explained in Figure 3.10. Clear substructure is seen in the density of stars over the region of the association.

likelihood young candidates for initial exoplanet searches, as confirming the youth of any potential exoplanet signals on a case-by-case basis is significantly faster than waiting on independent signs of youth for every individual candidate member of a stellar association. Should any of these members host exoplanets but be proven old in follow-up analyses, they are still considered interesting given their co-location (and co-moving velocities) with young birth environments like stellar associations. One particularly promising avenue would be if the community can find new signs of youth in large surveys such as all-sky photometric or astrometric space-missions, as this has the potential to rapidly increase the number of bona-fide young stars. Such signs of youth will also be useful to determine the true membership of individual cluster members which reside near to multiple stellar associations like those at the upper edge of the Octans selection polygon in Figure 3.12.

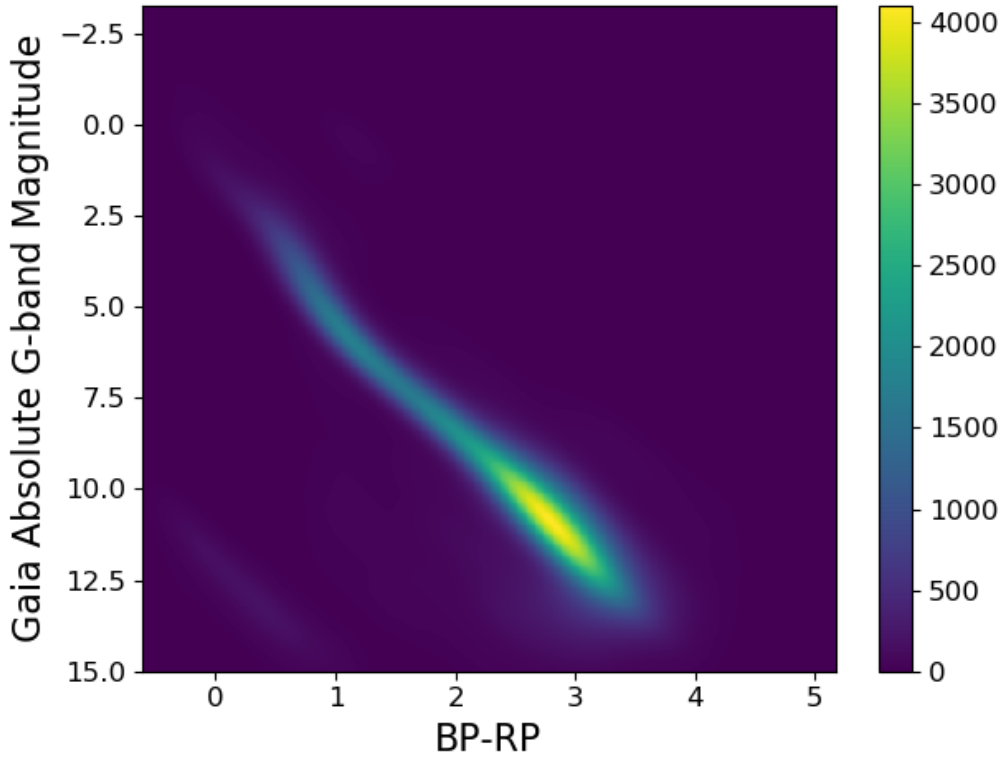

Figure 3.14: Final colour-absolute magnitude diagram for candidate members of the Octans association found in this analysis. Colour represents the density of stars, in units of number of sources per mag^2.

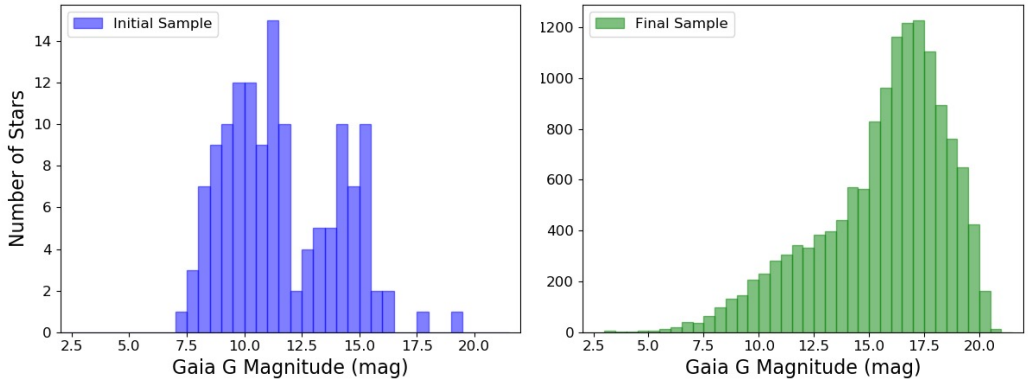

Figure 3.15: Distribution of Gaia absolute G-band magnitudes for the previously known members of the Octans association from de Zeeuw et al. [1999] (left) and the final members assembled in this work (right).

3.3 Construction of Young Star Target List

Of course, stellar associations are only one possible environment where young stars can be found. Large populations of young stars have also been found in young open clusters, dense star-forming regions and newly discovered stellar streams. The importance of these young stars to exoplanet formation and system evolution has prompted a significant drive by the community to extend the known population of young stars in recent years. Indeed shortly after the above exploratory research was undertaken, the CDIPS catalogue of Bouma et al. [2019] was released, collecting a wide assembly of past studies of clusters and associations in order to search for exoplanets around them. Based on a combination of large homogeneous catalogues of co-moving stars and smaller catalogues detailing individual cluster and association membership, this list highlighted the benefit of compiling existing catalogues of young stars in order to search for new young exoplanets. However, given the rate at which new cluster analyses are expanding, this list was already out of date by the time it was published, prompting the author alongside Dr Ed Gillen at Queen Mary University of London (formerly of the University of Cambridge), to begin the construction of a new, constantly updating, young star list.

3.3.1 Included catalogues

This list used the CDIPS target list compiled by Bouma et al. [2019] at its base, collecting four large catalogues [Cantat-Gaudin et al., 2018; Gaia Collaboration et al., 2018b; Kharchenko et al., 2013; Dias et al., 2014] and nine smaller catalogues of moving groups and stellar associations [Gagné et al., 2018b,c; Gagné and Faherty, 2018; Kraus et al., 2014; Röser et al., 2011; Bell et al., 2017; Rizzuto et al., 2011; Oh et al., 2017; Zari et al., 2019]. These catalogues included stars found through the *Gaia* DR2 data via an unsupervised membership assignment algorithm [Cantat-Gaudin et al., 2018], membership of well-studied open clusters [Gaia Collaboration et al., 2018b] and the full BANYAN catalogue of stellar associations [Gagné et al., 2018b,c; Gagné and Faherty, 2018] which was used to test the performance of the main light-curve detrending pipeline created in this work (see Section 2.6 and Chapter 4). An in-depth overview of these catalogues is given in Bouma et al. [2019]. In total, the CDIPS catalogue consists of 1,061,447 unique stars (with the sky distribution pictured in Figure 3.16), representing a significant population upon which to build a wider target list.

A further 19 catalogues were added to the young star catalogue by Dr Gillen and the author, yielding a total of 32 catalogues in the young stars list's current form.

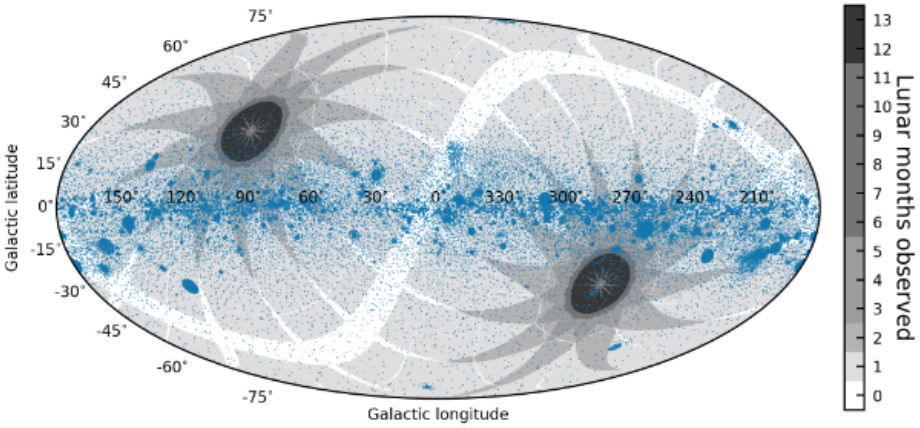

Figure 3.16: CDIPS catalogue sky distribution with overplotted pointing for the *TESS* primary mission. Note that the eventual pointings for sectors 14-16 and 24-26 in the Northern Hemisphere were eventually shifted further north to avoid scattered light effects from the Earth and Moon. Updated *TESS* pointings for the primary mission can be seen in Figure 1.11. Figure reproduced from Bouma et al. [2019].

These catalogues represent a wide range of analyses from new and updated large stellar cluster surveys [e.g. Cantat-Gaudin et al., 2020] to updated membership lists for individual clusters/associations [e.g. Damiani et al., 2019b; Dickson-Vandervelde et al., 2021], so a brief description of each is given here.

Six new catalogues of wider cluster surveys were added during this update. The largest of these was completed by Cantat-Gaudin et al. [2020], who used a set of objects with well-defined parameters to train an artificial neural network and 'paint a portrait' of the structure of the galactic disk out to ∼4 kpc. This resulted in reliable parameters and fully updated membership lists for 1867 clusters. Meanwhile Kounkel and Covey [2019] used an alternative convolutional neural network trained on the cluster populations from Cantat-Gaudin et al. [2018] to reveal new stellar streams and small clusters to a much wider extent in the Milky Way, representing 288,370 sources in total. Meingast et al. [2021] found similar 'coronae' of coeval stars linked to ten nearby open clusters (α Per, Blanco 1, IC 2602, IC 2391, Messier 39, NGC 2451A, NGC 2516, NGC 2547, Platais 9 and the Pleiades) by considering bulk cluster velocities and deconvolving the spatial distribution of each cluster with a Gaussian mixture model. The final list from this source consists of 7925 associated young stars (∼30-300 Myr) which extend over 100pc from the cores of the clusters. On the other hand, Kuhn et al. [2021] presented ∼120,000 candidate young stellar objects in their Spitzer/IRAC Candidate YSO (SPICY) cat-

alogue, using a random forest implementation with careful removal of background contaminants based mostly on IRAC photometry. The two final larger lists consist of TESS-specific catalogues, one assembled by Andrew Mann of Columbia University entitled "Searching For Infant Exoplanets In Young Moving Groups" in the first TESS investigator program[4], and the other composed of all *TESS* short-cadence targets from *TESS*'s primary mission (S1-26) which have >50% likelihood of being young cluster/association members according to the BANYAN Σ tool [Gagné et al., 2018b], as accessed through the tess-infos tool[5] created by Dr Max Guenther and Adina Feinstein.

The remaining catalogues all focus on single clusters, associations or regions of the sky. The largest of these was constructed by Cantat-Gaudin et al. [2019a] and Cantat-Gaudin et al. [2019b], which together provide membership lists for 48 new open clusters and distinct kinematic groups in the direction of Vela Puppis and Perseus respectively. In order to build these lists they used a "coarse to fine" selection method, beginning with density-based partitioning of the sky in these regions and then moving on to a gaussian mixture model in proper motion space to separate individual open clusters. Finally they applied an unsupervised assignment method to select new candidates, which are validated using colour-absolute magnitude diagrams to probe the youth of stars. The new open clusters found in these works challenge previous claims by groups like Kharchenko et al. [2013] that the community already had a complete census of clusters out to 1.8 kpc, and further demonstrate how powerful the new astrometric data from *Gaia* is for mapping the galaxy.

Meanwhile Damiani et al. [2019b] completed an extensive reanalysis of the Sco-OB2 stellar association, using a combination of similar proper motions, parallaxes and colour absolute magnitude diagrams, as discussed in chapter 2 and section 3.2. This resulted in them finding extended membership lists of nearly 11,000 pre-main sequence and ∼3600 main sequence stars with minimal field-star contamination, with the sky distribution illustrated in Figure 3.17.

Three further catalogues related to the BANYAN tool of Gagné et al. [2018c] were also added at this stage. In the first of these Gagné et al. [2018a] built a model within the BANYAN tool for the 90Myr Volans Carina association originally identified by Oh et al. [2017] and used this to add 46 candidate members to the association. Gagné et al. [2020] then used similar techniques to identify 500 new members in the new μ Tau association, allowing them to define the age as 62±7

[4]https://heasarc.gsfc.nasa.gov/docs/tess/data/approved-programs/cycle1/G011175.txt
[5]https://github.com/MNGuenther/tess_infos

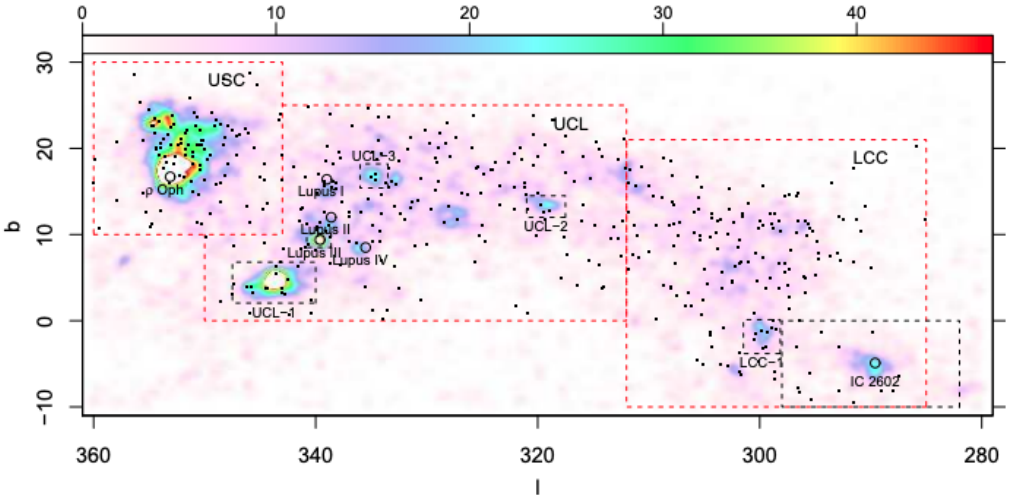

Figure 3.17: Spatial density overview of Sco-OB2 from Damiani 2019. The colour-bar across the top axis represents the number of stars per square degree. Similar to Figure 3.3, black dots represent 'bona-fide' members of the association from de Zeeuw et al. [1999]. The large red dashed rectangles represent the boundaries of the Upper Sco-Cen (USC), Upper Centaurus-Lupus (UCL) and Lower Centaurus-Crux (LCC) regions.

Myr. Meanwhile Zuckerman [2019] compiled members for the 40-50 Myr Argus association from literature and made final membership cuts made by considering UVW velocities and youth indicators such as lithium abundance and the position of stars on a colour-magnitude diagram. The Volans Carina and Argus associations are now included within the official BANYAN tool.[6]

The final set of catalogues represent recent updates to very young star forming regions, with the release of *Gaia* DR2 allowing significant updates to the population of known stars with ages $<\sim$10 Myr. Specifically, Galli et al. [2020] and Luhman [2020] independently completed new censuses of the 1-3 Myr Lupus Clouds, finding 137 and 178 high-probability members respectively, based on a combination of *Gaia* kinematic and photometric classifications. Luhman [2020] claims that this catalogue is complete for all stars with spectral types earlier than M7. Meanwhile Dickson-Vandervelde et al. [2021] and Luhman [2018] used similar kinematic techniques and colour-magnitude diagram cuts to update and extend the membership of the 3-5 Myr η Cha (36 members; 20 candidates) and 3-4 Myr Taurus star forming regions (438 members) respectively. Luhman and Esplin [2020] went one step further in their updated census of Upper Scorpius (part of the larger Sco-OB2 association with an

[6]http://www.exoplanetes.umontreal.ca/banyan/banyansigma.php

age of ∼10 Myr), gaining spectra of stars and brown dwarfs within this region to independently confirm the youth and membership of 1761 members. Finally, Kounkel et al. [2018] presented two catalogues (one using APOGEE-2 data and one using *Gaia* DR2 data) formed from using a hierarchical clustering algorithm to look for 6D structure in Orion and found several kinematically distinct groups within Orion. This further illustrates the fact that *Gaia* DR2 is not only giving the opportunity to add more members to these associations, but also define the internal structure of star forming regions and associations.

All of the catalogues discussed above have now been added to the young stars list compiled by Dr Gillen and the author. In order to combine all of the above catalogues into a single catalogue, each catalogue was cross-matched with *Gaia* DR2 and version 8 of the *TESS* Input catalogue (using the methods described in section 2.1.2) and then concatenated into a uniform list. Where stars were present in multiple catalogues, only one unique entry was retained, but information on all catalogues it is present in was retained for later analysis.

3.3.2 Overview of current young star list

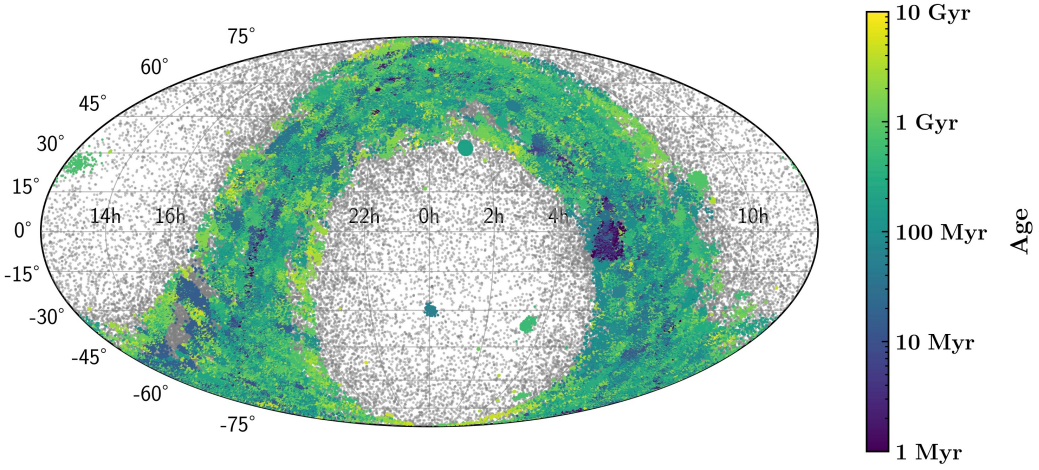

Figure 3.18: Overview of ages and sky distribution for assembled young stars list. Ages can be seen to vary from very young (<10 Myr) to over 1 Gyr for some of the older clusters in the largest catalogues. Stars without well-defined ages (but associated with young regions) are plotted in grey. Image Credit: Dr Ed Gillen.

The catalogue in its current form consists of **3,036,992 unique objects**, representing an unprecedented population of young stars around which to search for planets. Information included in the catalogue includes TIC and *Gaia* DR2 IDs

for all stars (except in rare cases when one or other of these are not known), full kinematic measurements, *TESS* and *Gaia* magnitudes and membership probabilities for each cluster/association. An overview of the sky and age distribution of the current catalogue can be seen in Figure 3.18, with the magnitudes plotted in Figure 3.19. While the skew towards relatively dim stars (naturally common for such young stars) makes it most useful for photometric searches and statistical validation, there still remain 400,753 stars with luminosities of $Gmag \leqslant 13.5$ which would be ideal candidates for radial velocity searches/confirmation. Meanwhile the stellar radii and mass distributions for young stars in this sample are shown in Figure 3.20. Based on the population of previous exoplanetary discoveries and typical exoplanet sizes,[7] the most useful stars will be those with radii less than five stellar radii ($R_* < 5R_\odot$) and/or masses less than approximately 3 stellar masses ($M_* < 3M_\odot$), which forms the bulk of this young star list.

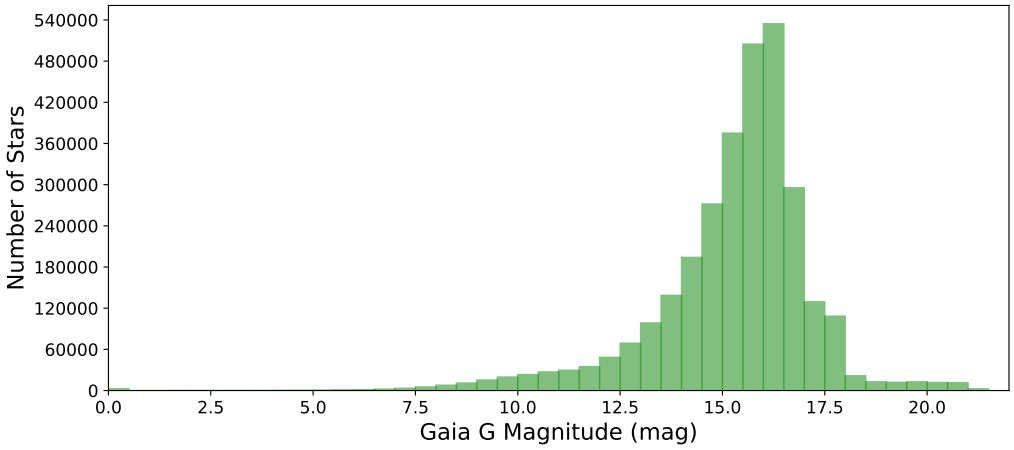

Figure 3.19: Overview of Gaia G magnitudes in all 3,036,992 stars in the assembled young star list.

This catalogue is already serving as the base for ongoing photometric searches for young exoplanets by academic and citizen scientists, as well as forming the base of *TESS* Cycle 4 & 5 proposals led by Dr Ed Gillen. Note that because of these ongoing searches the full catalogue is not reproduced in full here, but can be made available in small collaborations by contacting the author.

[7] According to data from the NASA Exoplanet Archive; https://exoplanetarchive.ipac.caltech.edu/

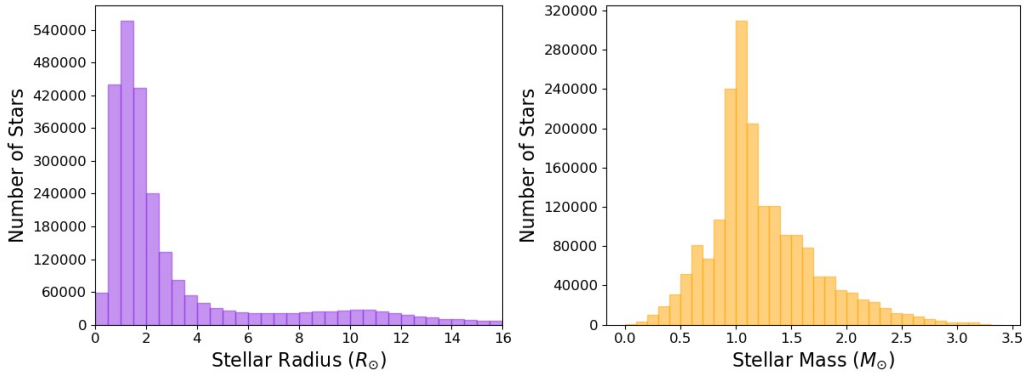

Figure 3.20: Overview of stellar radii and masses for all 3,036,992 stars in the assembled young star list.

3.4 Summary

One significant challenge in the search for new young exoplanet has been the traditional lack of significant numbers of bona-fide young stars. In an attempt to address this challenge, in this chapter two independent methods have been used to assemble new large lists of young stars around which to search for young exoplanets. Both of these methods draw upon the excellent kinematic precision of the *Gaia* satellite in order to determine which stars are associated with known groups of young stars, whether they are stellar associations, young open clusters or diffuse stellar streams.

The first of these methods involved interacting with *Gaia* data directly in order to search for smaller stars which are associated with the previously known massive star populations in stellar associations. The utilised graphical kinematic methods were validated in an analysis of the Sco-OB2 association by comparison to similar work completed by Damiani et al. [2019b] before being used to expand the population of candidate young stars in the Hyades and Octans associations. In the case of Octans this increased the number of candidate members of this association by ten-fold (to 14,050), greatly improving the likelihood that planets could be found in this association. A similar analysis was attempted for the Beta-Pic moving group, but found to be implausible due to the large spatial extent of the nearby association.

The second method focused on using catalogue-matching techniques to make use of the wide array of young star analyses recently completed by the community, collecting a total of 32 catalogues of young stars in stellar associations, clusters, stellar streams and star forming regions into one homogeneous catalogue. The current catalogue consists of 3,036,992 stars and is forming the base of significant young exoplanet searches in *TESS* data and beyond. Continuing kinematic analyses being

undertaken by the community will allow this list to continue to grow in the future.

The combination of these two methods has resulted in a sufficiently large population of young stars to make dedicated searches for new young exoplanets worthwhile, which forms the focus of the majority of the rest of this thesis.

Chapter 4

A search for young exoplanets in sectors 1-5 of the TESS Full-Frame Images

> "Once confined to fantasy and science fiction, time travel is now simply an engineering problem"
>
> — Michio Kaku

Despite the wealth of knowledge humanity has gained about planets in the solar system and beyond, there are still serious questions to be answered about how they formed in the first place. As human time travel is unfortunately constrained to the forwards direction, learning more about the past is best achieved by observing other planetary systems still in their infancy. With the *Gaia* satellite providing a dramatic increase in the number of known young stars (as discussed in Chapter 3) the time is now ripe for wider searches for new young exoplanets. However, finding such young exoplanets is significantly complicated by the fast rotation and complex activity of their young host stars, which are often not well handled by state-of-the-art automatic pipelines. In order to circumvent this problem, a new young-star specific detrending method has been developed by the author, which was described in detail in Section 2.6. This chapter presents the application of this pipeline to young stars in the first five sectors of young star data from the *TESS* mission. While no new exoplanet candidate signals were found in this initial search, interesting results included the recovery of all 2min TOIs known around young stars in sectors 1-5 (based solely on the 30min data), the recovery of the young exoplanets

DS Tuc A b and AU Mic b, a number of young eclipsing binaries and a wide array of interesting rotation. A sensitivity analysis is also presented for the entire sample, showing how recovery of injected planets varied with both depth and period for each individual target. This work represents a pilot study for wider young planet searches still being undertaken by the author.

A brief motivation for this work is given in Section 4.1, placing the work in context of the wider search for young exoplanets. Section 4.2 discusses the young star target selection, observations and the choice of a FFI pipeline, before describing the methods used to clean additional systematics from the *TESS* data, detrend stellar variability, search for transits and inject model transits for a sensitivity analysis. Section 4.3 then discusses important results seen as a result of the developed detrending techniques, including the recovery of known young exoplanets, TOIs, and eclipsing binaries, followed by an overview of the sorts of rotation and activity seen in this sample and the results of the conducted sensitivity analysis. The implications of these results for the future of exoplanetary searches around young stars are then discussed in detail in Section 4.4, before a summary and conclusion in Section 4.5.

Declaration - The following chapter is a partial reproduction of the paper 'A search for young exoplanets in Sectors 1-5 of the TESS Full-Frame-Images' [Battley et al., 2020], published in the *Monthly Notices of the Royal Astronomical Society* in August 2020. All work in this chapter was carried out by the author, with co-authors to the paper providing high level guidance and simple proof-reading.

4.1 Motivation

As discussed in section 1.4, discovering young exoplanets with well-defined ages is crucial to answer key questions about the formation and early evolution of exoplanetary systems. Increasing the population of young exoplanets not only improves our knowledge of their demographics, but also gives important new systems to probe such evolutionary processes as atmospheric loss [Owen, 2019], disk migration [Nelson, 2018] and dynamical interactions [Ida and Lin, 2010]. However, while state-of-the-art detrending methods such as those discussed in section 2.5 have been extremely effective at discovering new exoplanets around older stars, they still struggle to disassociate true transiting signals from the complex activity of young host stars. This was particularly well illustrated by Hippke et al. [2019] in the construction of their WOTAN tool, where only 35% of $0.5R_{\rm Jup}$ exoplanet signals injected into a sample of 316 young stars were recovered (irrespective of the detrending method chosen), compared to an almost 100% recovery rate for similar planets around less

noisy stars. Interestingly however, each of the methods tested recovered slightly different populations of the injected signals, so combining all of the tested methods increased the percentage of recovered planets to 43.8%. This discrepancy was further highlighted by Rodenbeck et al. [2018], who showed that different detrending methods resulted in quite different conclusions when considering the potential super-moon around *Kepler-1625 b*. It is thus clear that not only is there a need for new detrending methods specifically focused on young stars, but also that there is a distinct benefit of using multiple detrending methods on the same data-set, thus providing dual motivations for the construction of an alternative detrending method in this work. The development and current design of such a detrending pipeline is described in detail in section 2.6, with its first application to the *TESS* data discussed for the remainder of this chapter.

4.2 Data Preparation

4.2.1 Target Selection

While stellar association membership is now being expanded on a cluster by cluster basis thanks to the increased astrometric precision of the *Gaia* satellite [Gaia Collaboration et al., 2016b, 2018a] (see for example Kuhn et al. [2019]; Damiani et al. [2019b]; Zari et al. [2019] alongside the wider literature review in section 1.4.2), a key list of 'bona-fide' stellar association members were assembled by Gagné et al. [2018b] during preparation of the BANYAN Σ Bayesian cluster membership tool. Gagné et al. [2018b] consider a star to be a 'bona-fide' member if it has galactic XYZ UVW values consistent with those known for a given stellar association and exhibits an independent sign of youth. Such signs can include mid-infrared excess, G-J vs GALEX NUV-G colours consistent with youth, X-ray emission with HR1 $\geqslant$ -0.15, lithium absorption above 100 mArmstrongs or a compatible luminosity class [Gagné et al., 2018c]. Combining Gagné et al. [2018b]'s initial census of bona-fide/high probability stellar association members with new high-probability members added in the two following BANYAN Σ papers [Gagné et al., 2018c; Gagné and Faherty, 2018] yielded a total of 2977 objects spread over the 27 nearest known young stellar associations. Gagné et al. [2018a] later expanded the BANYAN Σ tool to also include the two new Argus [Zuckerman, 2019] and Volans-Carina associations using the same membership criteria, so these were also added to the initial target list, to give the final distribution of 3076 young stars in stellar associations illustrated in Figure 4.1. For clarity an association by association breakdown of targets in the original BANYAN sample and the final sample of the sector 1-5 targets analysed in

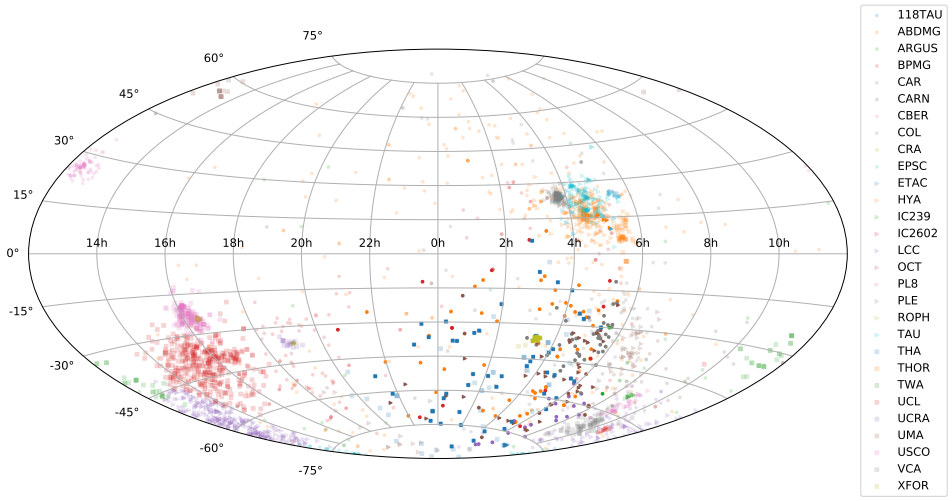

Figure 4.1: Equatorial sky distribution of initial young star sample, based on the bona-fide and high probability members of the 29 nearest stellar associations built into the BANYAN Σ tool. The entire sample is shown in the background with the 256 sector 1-5 targets analysed in this work presented in solid colour. Abbreviations after those chosen in Gagné et al. [2018b], which are also explained in Table 4.1.

this work is presented in Table 4.1. This census is the most extensive assemblage of bona-fide and high probability stellar association members with a common membership criterion at the time of this study, and hence represents a valuable starting place for the search for exoplanets around young stars in this work.

The various BANYAN survey results were combined into a single tabular target list using the *TOPCAT* table handling software [Taylor, 2005]. This list was then cross-matched with version 8 of the *TESS* input catalogue (TIC) [Stassun et al., 2019] using a 3 arcsecond radius. The Web *TESS* Viewing Tool[1] on the NASA *TESS* website was used to determine which sector each target would be observed in and in turn compile target lists for each individual sector. Of the original 3076 objects, 1832 of them were forecast to be observed in *TESS*'s first year of observations, with the breakdown of sources to be viewed in each Southern Hemisphere sector outlined in Table 4.2. For this work, only those viewed in Sectors 1 to 5 were considered.

[1]https://heasarc.gsfc.nasa.gov/cgi-bin/tess/webtess/wtv.py

Table 4.1: Breakdown of association distribution for the complete BANYAN sample (column 3) and final set of targets analysed in this work (column 4). The final targets consist those targets in Sectors 1-5 with available 30min-cadence light-curves from the Oelkers and Stassun [2018] pipeline. Association abbreviations are based on those presented in Gagné et al. [2018b] with the addition of ARGUS for the Argus association and VCA for the Volans-Carina Association.

Association	Abbreviation	BANYAN Targets	S1-S5 Analysed Targets
118 Tau	118TAU	12	0
AB Doradus	ABDMG	298	55
Argus	ARGUS	40	0
β Pictoris	BPMG	135	13
Carina	CAR	85	19
Carina-Near	CARN	110	18
Coma Berenices	CBER	76	0
Columba	COL	107	43
Corona Australis	CRA	14	0
ϵ Chamaeleontis	EPSC	42	0
η Chamaeleontis	ETAC	16	0
Hyades	HYA	241	7
IC2391	IC2391	14	0
IC2602	IC2602	16	0
Lower Centaurus Crux	LCC	310	0
Octans	OCT	101	40
Platais 8	PL8	32	0
Pleiades	PLE	204	0
ρ Ophiuchi	ROPH	182	0
Taurus-Auriga	TAU	178	1
Tucana-Horologium	THA	92	49
32 Orionis	THOR	46	0
TW Hya	TWA	50	0
Upper Centaurus Lupus	UCL	410	0
Upper Corona Australis	UCRA	16	0
Ursa major	UMA	13	0
Upper Scorpius	USCO	161	0
Volans-Carina	VCA	59	0
χ Fornacis	XFOR	16	11
Total	3076	256	

Table 4.2: Overview of BANYAN bona-fide members of stellar associations viewed in the first year of *TESS* observations. Note that sources in the overlapping regions of the *TESS* sectors, such as those in the continuous viewing zone, are counted multiple times.

Sector	Number of Sources observed
1	118
2	130
3	124
4	186
5	336
6	258
7	163
8	166
9	272
10	399
11	750
12	396
13	200

4.2.2 Observations

Sectors 1-5 studied in this work were observed between 25 July and 11 December 2018. For these sectors, two main data products are available: 2min-cadence primary light-curves and 30min cadence light-curves extracted from the *TESS* Full Frame Image data. While the primary 2min data products are very powerful for the main *TESS* mission, because of the limited data transfer rates of the *TESS* primary mission far from all of the stars in *TESS*'s field of view will have 2min light-curves generated. In addition, because of the increased activity and rotation rate of young stars discussed above, the standard detrending methods built for the 2min light-curves have difficulty flattening the light-curves for transit searches, and may even introduce confusing additional artefacts as a result of detrending [Hippke et al., 2019]. Because of the greater coverage offered by the 30min cadence data and concerns about detrending artefacts, the 30min cadence data were thus chosen for the initial transit search, with the simple aperture photometery (SAP) 2min cadence light-curves (where available) used as a secondary check. As discussed in Section 2.2, a number of different sources of 30min cadence data extraction currently exist, however given the increased availability and comparatively clean light-curves provided by the DIA pipeline of Oelkers and Stassun [Oelkers and Stassun, 2018, 2019] at the time of writing, this pipeline was chosen to extract light-curves from the *TESS* FFIs in this work, resulting in light-curves for 256 individual objects.

An additional challenge common to both the 2min and 30min cadence data is the reasonably large pixels of the *TESS* focal plane [Vanderspek et al., 2018], which leads to significant blending of stars in crowded regions. This precludes the use of *TESS* photometry alone for confirming transits in the dense centres of young clusters without confirmation from higher resolution instruments such as the *Spitzer* Space Telescope [Werner et al., 2004] or the cameras of the *Next Generation Transit Survey - NGTS* [Wheatley et al., 2013]. The power of *NGTS*'s comparative resolution for identifying the true photometric source for *TESS* objects is well-illustrated by Jackman et al. [2019] in the discovery of NGTS-7Ab. However, for initial searches, the volume, consistent sky coverage and quality the *TESS* light-curves make them the optimum solution for this work.

4.2.3 Removal of additional systematics from 30min light-curves

Unlike the 2min PDCSAP light-curves retrieved from MAST, the 30min light-curves supplied by Oelkers and Stassun [2019] have not undergone the in-depth quality analysis completed by the SPOC pipeline, and as such still include some less trustworthy epochs of increased pointing jitter, regular spacecraft momentum dumps and known data anomalies. It was thus necessary to remove these systematics before activity detrending and transit searches could begin.

The first step undertaken was to cut any epochs where fine pointing was known to have been lost, or other spacecraft anomalies affected the data. This was achieved by consulting the *TESS* data release notes[2] for each sector and primarily affected sectors 1, 3 and 4. In Sector 1 a period of anomalously high pointing jitter was seen between approximately TJD 1347-1349 due to problems with the fine-pointing calibration. This was observed to be particularly bad between TJD 1348-1349.29, so all epochs between these times were masked from the analysis. Similarly in sector 3 a few experiments on the attitude control system (ACS) were undertaken by the *TESS* team, dramatically increasing the scatter at these times. As a result, only data between TJD 1385.90-1395.48 in orbit 13 and TJD 1396.61-1406.29 in orbit 14 are scientifically useful, and all data at other epochs in this sector was cut. Sector 4 on the other hand was plagued by an instrument anomaly between TJD 1418.54 and TJD 1421.21, where communication was lost between the instrument and satellite. As a result, no data or telemetry was collected for this period, and some systematic trends were introduced following activation of the on-board heaters. Additional strong glints between TJD 1422.23 - 1423.50 (orbit 15) and TJD 1436.10 - 1436.83 (orbit 16) also plague some of the light-curves in

[2]Available at https://archive.stsci.edu/tess/tess_drn.html

sector 4 (particularly those on camera 4), however the amplitude and duration of these appear to vary between different targets, so may be better examined on a case by case basis.

Following the removal of these sections of unreliable data, a more automated method was required to identify and remove additional epochs of increased spacecraft scatter on an epoch-by-epoch time-frame, such as those around the regular spacecraft momentum dumps. This was completed by generating scattering quality masks based on the engineering "quaternion" data, using a similar method to that described by Vanderburg et al. [2019] to prepare Sector 3 30min cadence data in the discovery of multiple super-Earths around HR 858. The *TESS* quaternion data[3] consists of 2s-cadence time-series data for each sector, describing attitude changes in three primary vectors (Q1, Q2, Q3) based on deviations from a selection of local guide stars. This provides a sector-specific overview of the spacecraft attitude and thus allows the generation of scatter-based quality masks for all targets in the *TESS* aperture. An example of this data (both raw and binned into 30min bins) can be seen in Fig 4.2. To identify epochs with excessive scatter, the standard deviation for each vector was calculated, and any epoch with pointing scatter $\geqslant$5 standard deviations from the mean was flagged for removal. By combining the results from all three vectors into a single mask, all epochs within 0.01d of these points were removed from the data-set. This step efficiently removed all spurious signals relating to momentum dumps, as well as any remaining short periods of overly large scatter that were not picked up in the initial wider cuts.

For sources of particular interest which also possess *TESS* 2min quality flags, it may also be advisable to implement the quality flags from the SPOC pipeline to the 30min data (perhaps in a similar manner to the conservative approach taken by Bouma et al. [2019] in building their difference-imaging extraction pipeline for sectors 6 and 7) in order to remove the effects of additional systematics such as cosmic rays. However, given the lack of 2min light-curve availability for many of the objects in this sample, this feature was not implemented in the development of the initial pipeline.

Following the removal of such sector-specific effects, the data was split into separate sections wherever there was a data gap of more than 0.1 days (to reduce the effect of flux jumps often seen after data gaps), except where doing so would result in fewer data points than the prescribed detrending window length discussed below.

[3]Available at https://archive.stsci.edu/missions/tess/engineering/

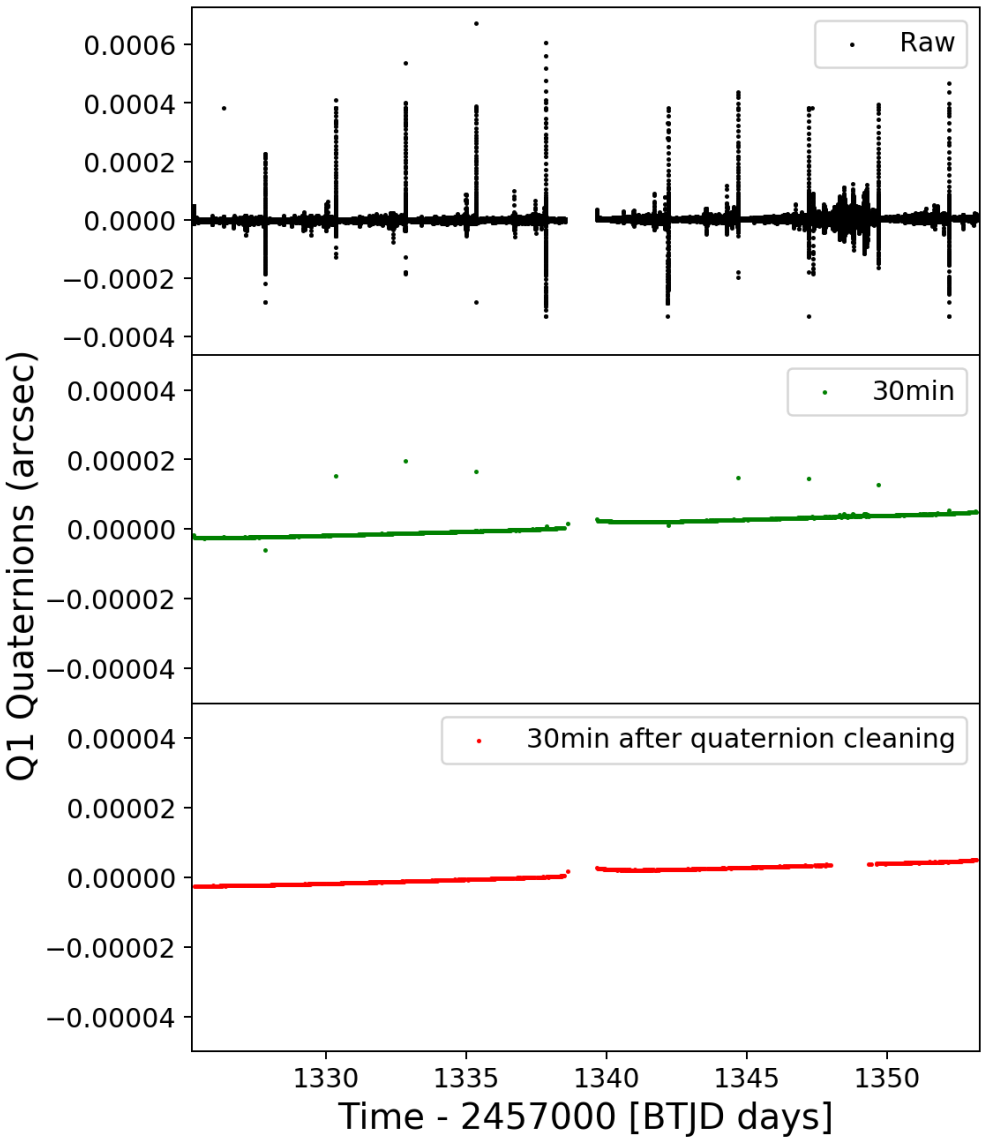

Figure 4.2: Q1 Engineering quaternion data for Sector 1, Camera 1 showing a clear increase in scatter around the 2.5d momentum dumps and the loss of fine pointing between TJD 1347-1349. Top: Raw 2s Engineering Quaternion data. Middle: Quaternion data binned into 30min bins to show typical number of affected data points. Bottom: Final quaternions after the removal of known periods of increased scatter and quaternion-based cleaning.

4.2.4 Detrending and search pipeline overview

The full detrending pipeline developed by the author is discussed in depth in Section 2.6, but is briefly reviewed here in context. The base detrending method is Locally-Weighted Scatterplot Smoothing [LOWESS; Cleveland, 1979a], a local polynomial regression method which fits a low-order polynomial to a pre-defined 'window' of the data (typically 10-15hrs, depending on light-curve evolution). This allows it to handle local variations in the data more effectively than the commonly adopted Savitzky-Golay smoothing method [Savitzky and Golay, 1964]. Smoothing can either be applied to a light-curve as a whole or automatically broken into sections every time there is a gap of 0.1days. This smoothing is backed up by the optional removal of peaks and troughs in the light-curve, given that the sharpest troughs can often be mistaken for transits after detrending. If a signal of interest is highlighted, additional transit-masking is available to be implemented into the pipeline in order to best preserve the shape of the candidate transit signals.

Following detrending, each light-curve was searched using the Box-Least Squares (BLS) method [Kovács et al., 2002] and the three strongest unrelated BLS peaks from the resulting periodogram investigated for evidence of planetary transits using hand-built eyeballing plots. This full pipeline was applied to all targets in *TESS* Sectors 1-5 with 30min cadence data. To gauge its performance on the complicated young star light-curves, an additional injection and recovery test was also conducted, testing planetary periods between 1 and 14 days and planet radius ratios of $R_{\rm P}/R_{*} = 0.1$ to 0.03, roughly corresponding to Jupiter and sub-Neptune-sized planets around a sun-like star.

4.3 Results from Sectors 1-5

Despite 30min DIA light-curves only being available for 256 of the BANYAN objects within the first five sectors of *TESS* data, a wide range of interesting activity was observed, from the recovery of known young exoplanets and *TESS* objects of interest (TOIs) to an eccentric eclipsing binary system and a large variety of unusual rotation and activity profiles. Even this relatively small sample clearly demonstrates the unusually large variation in light-curves of young stars compared to their older counterparts, and consequently helps to explain why far fewer planets have thus far been found around stars of these ages. The conducted sensitivity analysis goes one step further, investigating the comparative recovery rates for different combinations of injected period and planetary radius. The most interesting results from this search are summarised below in this section.

4.3.1 Recovery of confirmed young exoplanets

One of the most promising initial results from the application of this pipeline on the 30min data was the recovery of both of the known transiting exoplanets found around young stars in sectors 1-5: DS Tuc A b [Benatti et al., 2019; Newton et al., 2019] and AU Mic b [Plavchan et al., 2020]. The recovery of DS Tuc A b is described in detail here as an example of the full pipeline in use.

DS Tuc A (TIC 410214986/TOI 200.01/HIP 116748 A) was observed in Sector 1 of the *TESS* observations, carried out between 25th July - 22nd August 2018. It is a G6V type star known to be associated with the 45 Myr Tucana-Horologium association [Zuckerman and Webb, 2000]. DS Tuc A fell on camera 3 of the instrument, and yielded approximately 27 days of photometry. Interestingly, Newton et al. [2019] admit in their work that the candidate signal was initially found by human eyeballing around a spurious activity-induced periodicity peak flagged by a run of the SPOC Transiting Planet Search (TPS) module on the 2min PDC-SAP data. However, a later archival TPS run of the SPOC pipeline (after the planet candidate was announced to the community as TOI 200.01) was seen to detect a periodic transit crossing event which passed all of Newton et al. [2019]'s false positive tests. The planetary candidate was later confirmed using additional photometry, spectroscopic methods and high contrast imaging.

Benatti et al. [2019] independently reprocessed the TESS data for this object using improved stellar parameters (including crucially accounting for dilution from DS Tuc B) and fitted two different models to determine planetary parameters. The first model involved modelling only the first two transits (on account of the large pointing jitter around the 3rd transit) using `PyOrbit` [Malavolta et al., 2016] to complete a simultaneous fit for modulation and a transit signal, along with `emcee` [Foreman-Mackey et al., 2012] and `PyDE` [Storn and Price, 1997] to establish the most likely planetary parameters. Alongside this they tested modelling all three transits with the `batman` package [Kreidberg, 2015] after applying a 0.55d 3rd order running polynomial to flatten the light-curve and analysed the posterior with `emcee`. While both methods yielded consistent results, they eventually adopted the planetary parameters from the first method and found a best-fit solution of a 0.50 ± 0.02 $R_{\rm Jup}$ planet. Benatti et al. [2019] confirmed the planetary nature of this object using radial velocities from the HARPS spectrograph.

In this work, the planetary signal was highlighted as the highest peak of the BLS periodogram in a 20-bin LOWESS-partial run of the standard detrending pipeline, however the transits were also clearly visible by eye after the 30-bin LOWESS-partial run. Noting that the third transit fell in the period of height-

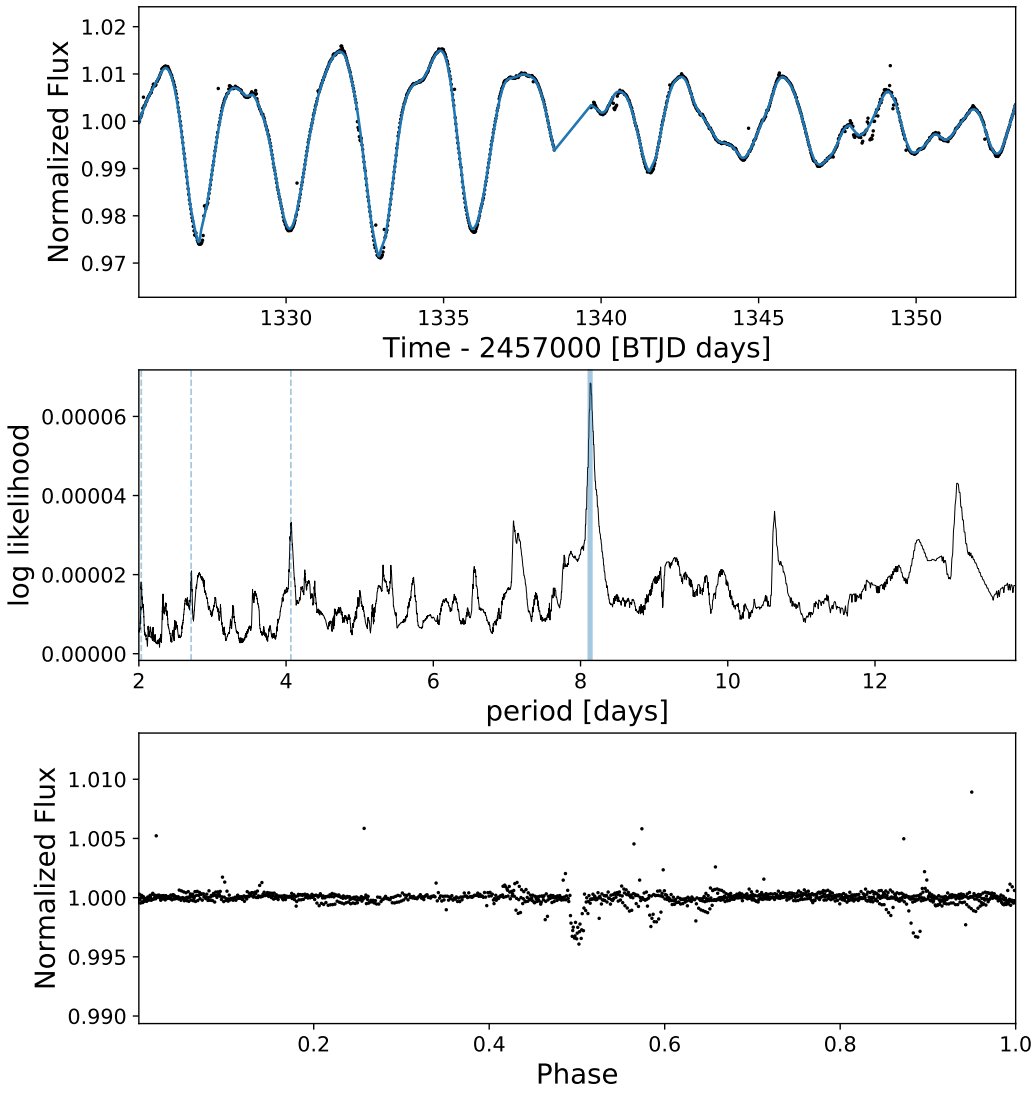

Figure 4.3: Example of the developed detrending pipeline in use, showing the recovery of the 45Myr exoplanet DS Tuc A b. Top: The original sector 1 light-curve for DS Tuc A with overplotted 20-bin LOWESS detrending in blue, including interpolation over suspected transits. Middle: BLS Periodogram for the the light-curve after detrending, with the peak period of 8.138d highlighted in blue (n.b. aliases of this period are shown by the blue dotted lines). Bottom: the resulting light-curve after the detrending pipeline has been applied, folded by the maximum peak of the BLS Periodogram

ened pointing jitter between TJD 1347-1349, this section of the data was unmasked for this target. The recovered period was 8.14d, in agreement with the accepted value from Newton et al. [2019]. Using the period and epoch derived from the BLS search, the transit masking and light-curve interpolation technique described in section 2.6.3 was applied, completing the clear detrending and transit recovery of DS Tuc A b presented in Fig 4.3. Note that in the bottom panel the light-curve has been folded by 8.138d, the accepted planet period from Newton et al. [2019]. Interestingly, the use of the peak-cutting technique for this object was found to increase the significance of the true transit period in the 30-bin case (changing the 8.14d period-peak in the BLS periodogram from insignificant to the 6th strongest after peak-cutting), but decreased its significance for the 20-bin LOWESS-detrend. The clear recovery of this Neptune to Saturn-sized planet using the 30min data alone bodes well for future exoplanet candidate discoveries from the Full Frame Images, thus demonstrating the wealth of knowledge to be gained from these images.

Unfortunately data for AU Mic was not extracted by Oelkers and Stassun [2018]'s 30min pipeline, however application of this pipeline to the 2min data for AU Mic easily recovered the signal of the 8.46d planet proposed by Plavchan et al. [2020], as shown in Fig 4.4. In this case a 20bin-LOWESS smoothing run revealed the 8.46d period as the third highest peak in the periodogram. In this particular case the peak-cutting option did not aid recovery of the planet substantially since the improvements of cutting the sharp troughs were balanced by the inadvertent cutting of the first transit. Nonetheless, the 8.5d signal remained the third highest peak after 20bin-LOWESS smoothing was applied to the peak-cut light-curve.

4.3.2 Retrieval of other *TESS* objects of interest

A number of other interesting signals were independently found using this pipeline, including a pair of additional *TESS* Objects of Interest (TOIs). In the interest of independence these signals were found with no prior knowledge of the TOIs in these sectors, but instead candidate signals highlighted in the BLS search in this work were cross-matched with the TOI list at a later date. This resulted in the recovery of two known TOIs: TOI 447.01 and TOI 450.01.

For TOI 447.01, a repeated dip of approximately 15 mmag was highlighted by the BLS search in sector 5 with a period of 5.528 days around the F5/6V star HD 33512/TIC 14091633 (see Fig 4.5). This signal was uncovered using the a standard 30-bin LOWESS run of the pipeline, without peak-cutting. HD 33512 is a relatively bright (T=8.8) F5/6V star in the Octans association, giving it an approximate age of 35±5 Myr [Gagné et al., 2018b], which would make this an interesting target

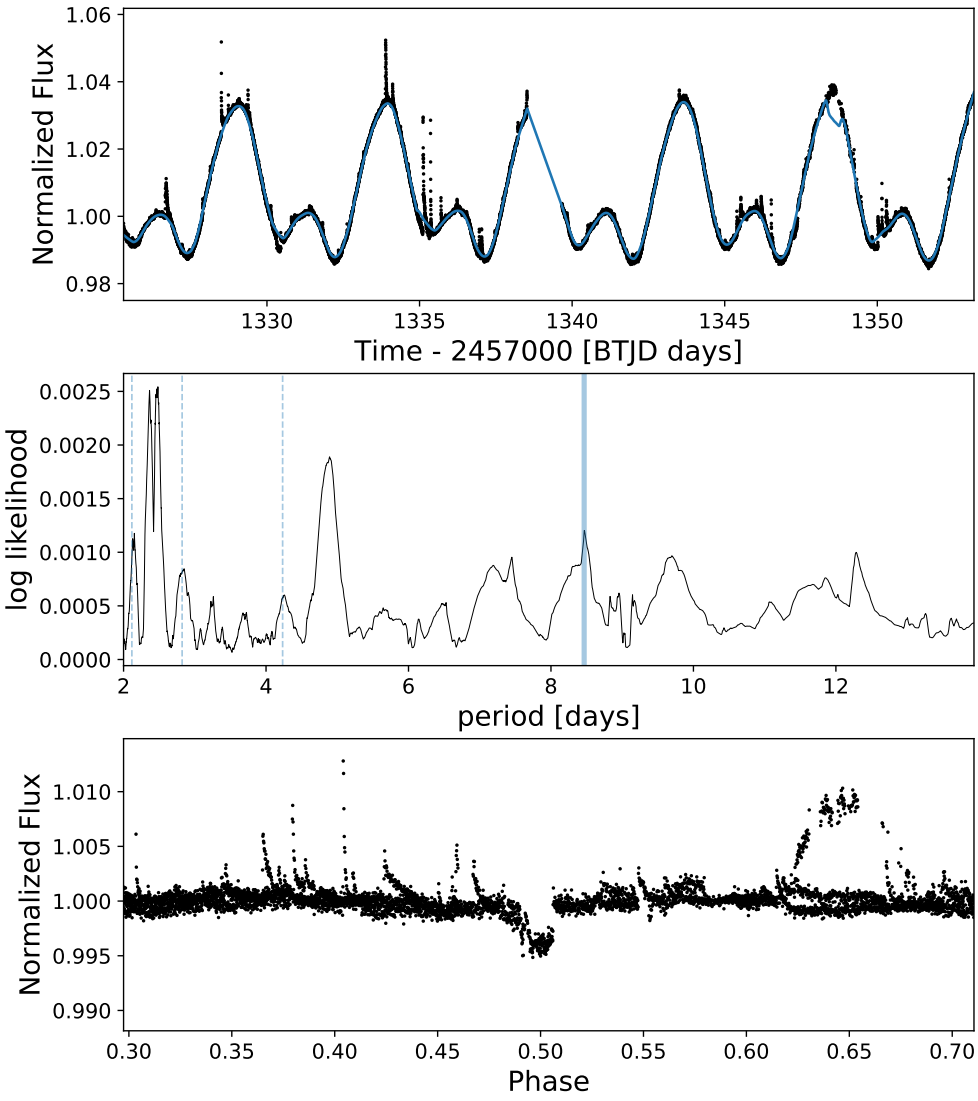

Figure 4.4: A 20bin-LOWESS smoothing run on the *TESS* 2min cadence data for AU Mic, showing the recovery of the 8.46d planetary signal first identified by Plavchan et al. [2020]. Format similar to that explained for Figure 4.3, except that the final light-curve is folded by the third highest period, as highlighted in blue in the BLS periodogram.

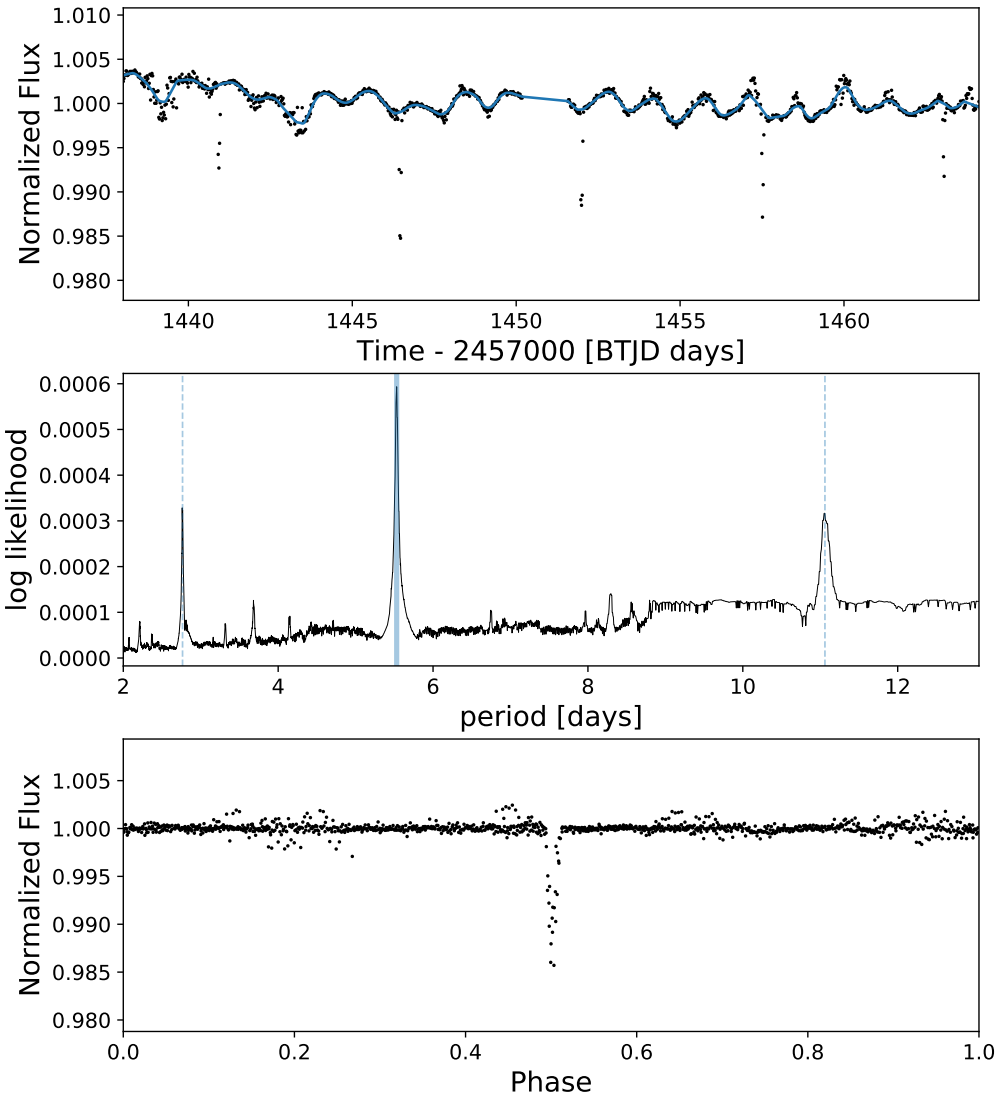

Figure 4.5: Recovery of TOI 447.01/HD 33512, a likely long-period binary in the 35Myr Octans Association. In this case the maximum period was found to be 5.528d. Format after that explained in Fig 4.3.

for follow-up of young evolving systems if the planetary hypothesis was correct. As a TOI this target underwent follow-up by the TFOP team, and though initially seemed positive from the *TESS* data alone, it was eventually flagged as a False Positive (FP) by the TFOP working group due to the observed linear drift in RV measurements, relatively large radius, and changing width of FWHM.[4] Nonetheless, this object is a likely long-period binary, so may still be an interesting system for the study of young binary systems.

It should also be noted that the recovery of this signal was originally complicated by the fact that the DIA 30min light-curve for HD 33512 and TYC 7053-832-1 appeared near-identical due to their close proximity. Indeed, as these sources are only 30.43" apart they fall onto very similar regions of the *TESS* CCD array, and thus suffer from blending on the spacecraft's relatively large 21" pixels. As such these sources cannot easily be distinguished using the *TESS* data alone and require higher resolution photometry to unravel the true origin of this signal. Fortuitously both of these objects were observed by the *Next Generation Transit Survey*, which has much greater resolution of approximately 5" per pixel [Wheatley et al., 2013]. By folding the *NGTS* light-curves for these two stars by the recovered 5.528d period, the candidate signal was clearly determined to have come from the source HD 33512.

The second TOI independently recovered was TOI 450.01, around the source 2MASS J05160118-3124457/TIC 77951245. This source was observed in sectors 5 and 6 of *TESS* observations, however alike to TOI 447.01 only sector 5 light-curves were available for it from the DIA 30min pipeline. In this work the candidate signal was initially revealed by a 30-bin LOWESS-partial run of the main detrending pipeline on sector 5. Two approximately 40mmag transits were observed in the flattened light-curve, with a recovered period of 10.74 days and initial epoch of 1443.2d, as shown in Fig 4.6. The parameters derived from this 30min data are slightly different to those derived from the 2min data by the SPOC TOI analysis (P = 10.7148$\pm$0.0001, depth = 63.6$\pm$0.8, epoch = 2458443.1686 $\pm$ 0.0003), however this is not surprising given the much lower density of points in each transit signal (15 times fewer). TIC 77951245 is an M4 star of T = 12.375 in the 42 Myr Columba association. Based on the depth of the observed signals and the radius of the star, this signal would correspond to an approximately Jupiter-sized planet if the planetary hypothesis was correct. However, follow-up of this object by the TFOP working group eventually revealed it to be a near-equal mass spectroscopic binary, based on HRS measurements on the 10m SALT telescope on 10th August

[4]https://exofop.ipac.caltech.edu/tess/target.php?id=14091633

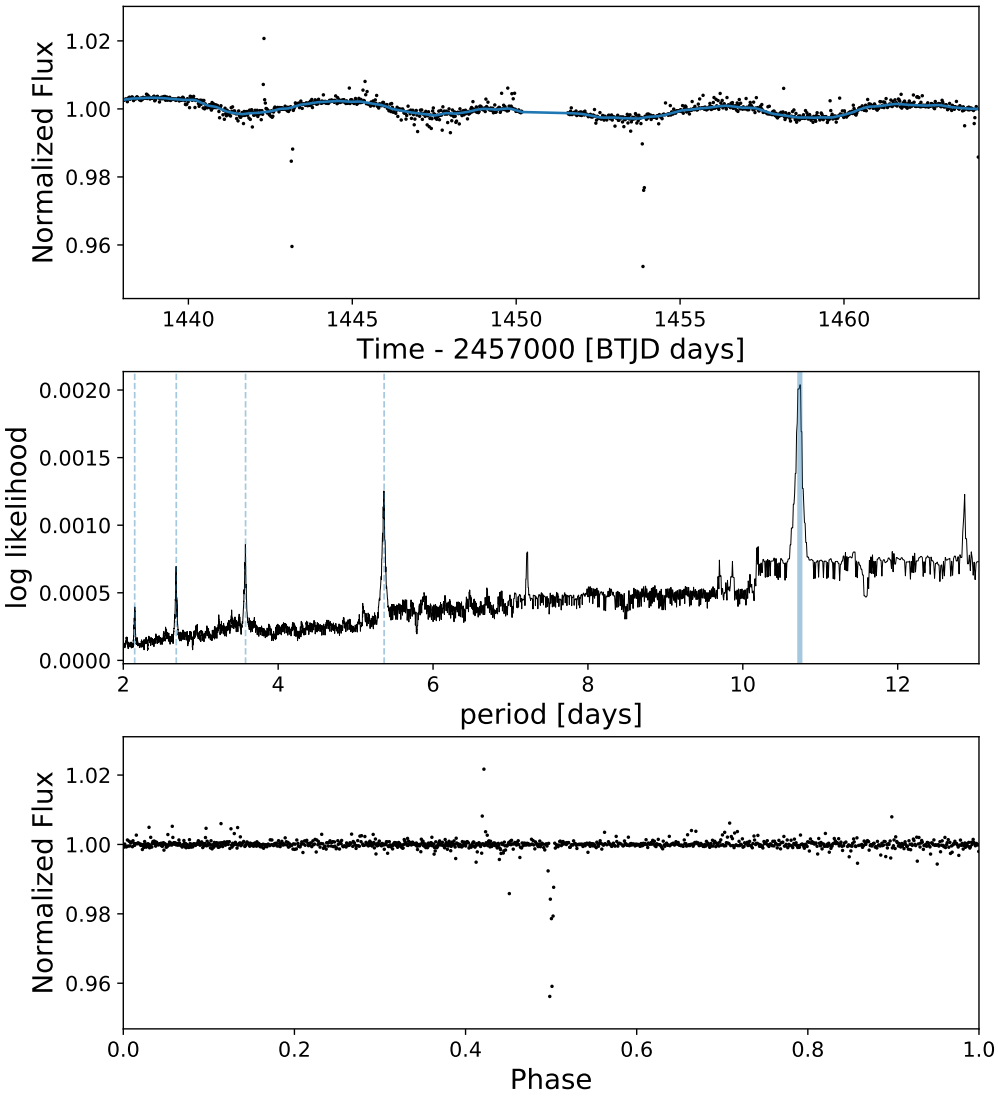

Figure 4.6: Recovery of TOI 450.01/J0516-3124, a spectroscopic binary in the 42 Myr Columba Association. In this case the maximum period was found to be 10.7d. Format after that explained in Fig 4.3, except that over-plotted LOWESS-detrending in the top panel has bins of 30 rather than 20.

2019.[5]

These two TOIs along with the new confirmed exoplanet DS Tuc A b discussed in Section 4.3.1 represent all three of the TOIs highlighted by the main 2min SPOC pipeline for this selection of young stars. This illustrates the fact that this pipeline is working at least as effectively as the main SPOC 2min pipeline for these young objects, further emphasising its potential for finding candidate signals around stars present only in the 30min cadence data.

4.3.3 Eclipsing Binaries

A number of clear young eclipsing binaries were also revealed by this survey, the most interesting of which being HD 28982. This highly eccentric eclipsing binary system was revealed in this work by a 30-bin run of the LOWESS-based detrending (without peak cutting) applied to sector 5 of the *TESS* 30min DIA data (see Fig 4.7, panel a). A period of 5.97 days was clearly highlighted as the maximum power period in the BLS periodogram, and two distinctly different duration and depth transit signals were visible. HD 28982 is associated with the AB Doradus Moving Group (ABDMG), which is approximately 150Myr old [Gagné et al., 2018b].

Because it has long been known to be a 'bona-fide' member of this moving group it was included in the initial *TESS* CTL [Stassun et al., 2019] and received 2min coverage in the main *TESS* survey. This provides an interesting opportunity to compare the 2min and 30min data for this young object. After retrieval from MAST, the 2min data was subjected to the same detrending as the 30min data by expanding the number of bins for each LOWESS-based detrending step to 450, 15 times as many as in the 30min analysis. This revealed the detrended light-curve shown in panel b of Fig 4.7. Aside from the obvious increase in signal to noise and visually more obvious signal provided by the 2min data, the 30min data clearly provides sufficient information to constrain the EB period and durations. However alike to the TOIs, the 30min data suffers in terms of depth accuracy, likely due to exposure smearing from the longer 30min cadence data. In addition, the 30min cadence data from the DIA pipeline is more affected by scatter than the 2min data, likely due to a combination of pointing scatter and improperly corrected scattered light. Care must therefore be taken when evaluating planetary radii for planet candidates based on 30min data alone, and may be better left to higher-cadence photometry in follow-up observations.

[5] https://exofop.ipac.caltech.edu/tess/target.php?id=77951245

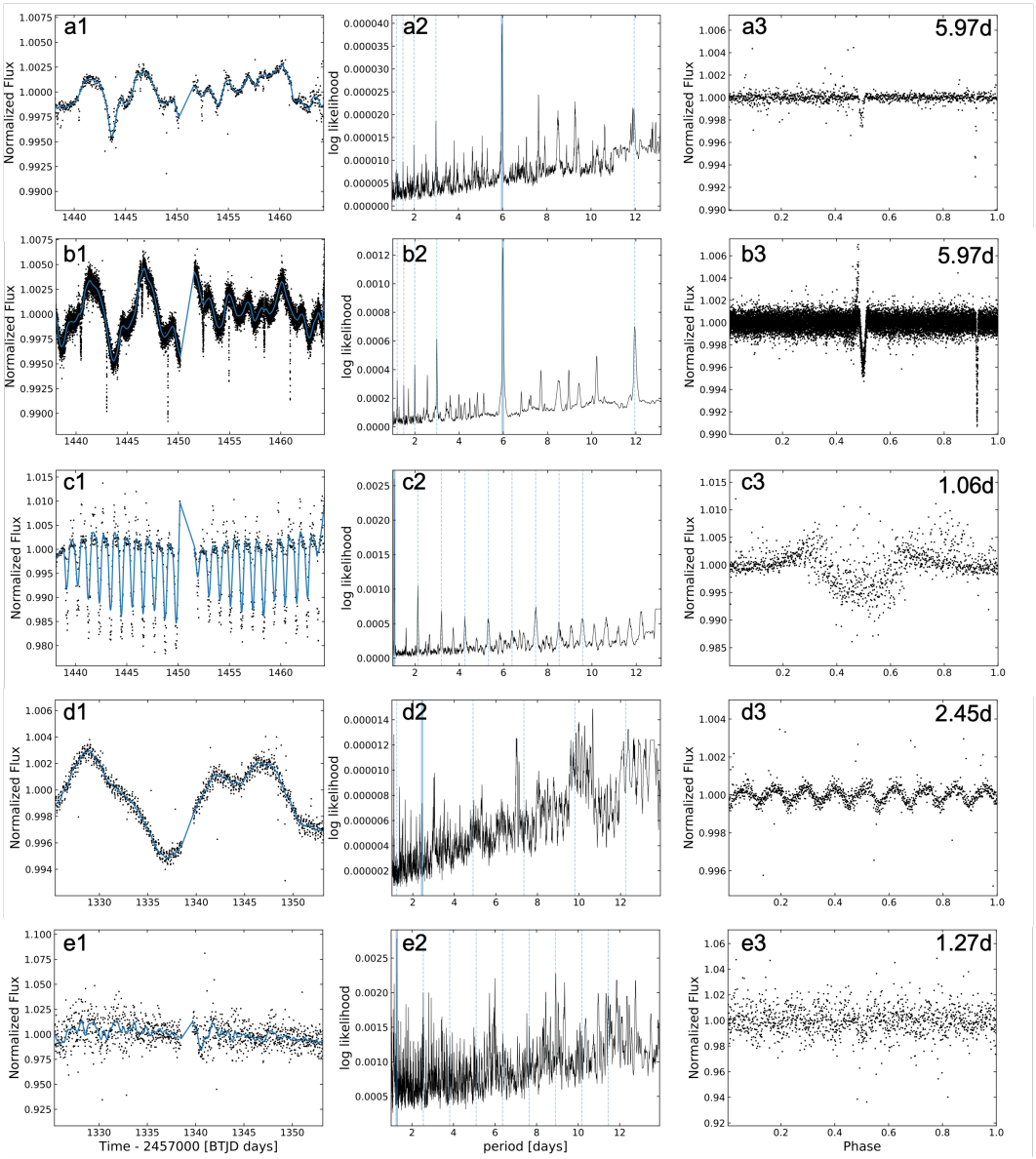

Figure 4.7: Sector 1-5 light-curve zoo. Column 1: original light-curve for each star with a 20-bin LOWESS-smoothed fit over-plotted in blue. Column 2: BLS periodogram after LOWESS-detrending, with the highest power period highlighted in blue (except for J0635-5737 where a shorter alias is chosen for clarity), and aliases shown with dotted blue lines. Column 3: flux after 30-bin LOWESS-detrending, folded by the highest power period. Individual sources from top to bottom: Panel a: HD 28982, an eccentric EB in ABDMG (approx 150Myr) in 30min data, Sector 5; Panel b: HD 28982 in *TESS* 2min data; Panel c: J0529-2852/TIC 31281820 in COL (approx 42 Myr), Sector 5; Panel d: J0635-5737/TIC 348839788 in ABDMG (approx 150 Myr), demonstrating 'double variation' in its rotation data, Sector 1; Panel e: J0552-5929/TIC 350712873 in THA (approx 45 Myr), Sector 1.

4.3.4 Rotation and activity

Rotation and spurious stellar activity were the cause of the strongest peaks in many of the stars viewed in this sample, which helps to explain why searching for planets around young stars is so much harder than around many older stellar hosts. As a first step in characterisation of this rotation and activity, the period of each star's main stellar variability/activity was also recorded using the generalised Lomb-Scargle method [Lomb, 1976] implemented into `python` as the `LombScargle` function in the `python astropy library`[6] [The Astropy Collaboration et al., 2013; Astropy Collaboration et al., 2018]. Furthermore, the main amplitude of the primary variability was determined in the peak-cutting step discussed in Section 2.6.2, by measuring half the difference between the first peak and first trough identified for removal. This avoids measuring the amplitude from flares or noisy data, but will be unreliable for light-curves which evolve in amplitude over the course of a sector. A wide variety of rotation and activity curves were observed in this work, broadly separated into four categories: near-uniform periodic, periodic but evolving, aperiodic and fast rotators. Amplitudes of oscillation varied from 0.02% to 9%, while primary periods of flux variation varied from 0.122d (e.g. J0552-5929 - panel e of Fig 4.7) to non-variable over the 27-day time-period.

The first of these categories, those light-curves which exhibit near-uniform variation in amplitude over a set period, are well illustrated by the sinusoidal curve of sources such as HIP 1993 and HIP 1113 (Fig 4.8), alongside strongly periodic drops in sources like 2MASS 05292529-2852274 (see panel c of Fig 4.7). These types of sources are theoretically ideal for detrending, as they can be easily modelled by quasi-periodic smoothing functions, especially when the period of flux variation is over periods of two or more days. Furthermore, the peak-cutting method discussed in Section 2.6.2 was often observed to aid the retrieval of injected transits for these periodic sources. However, very fast rotation of this type still provides difficulties, as discussed below.

Another very important category of variation seen are those light-curves with strong periods but obvious variation in flux amplitude. This type of light-curve is well illustrated by sources such as HIP 105388 (observed in *TESS* sector 1) and AB Pic (observed in all southern hemisphere sectors - 1-13) - see Fig 4.9. The oscillation amplitudes of these sources are both observed to considerably change over the course of a single sector. Indeed, in the case of AB Pic, the amplitude of the primary oscillations changes from 0.01% to >0.03% over the course of the five

[6] https://docs.astropy.org/en/stable/api/astropy.timeseries.LombScargle.html

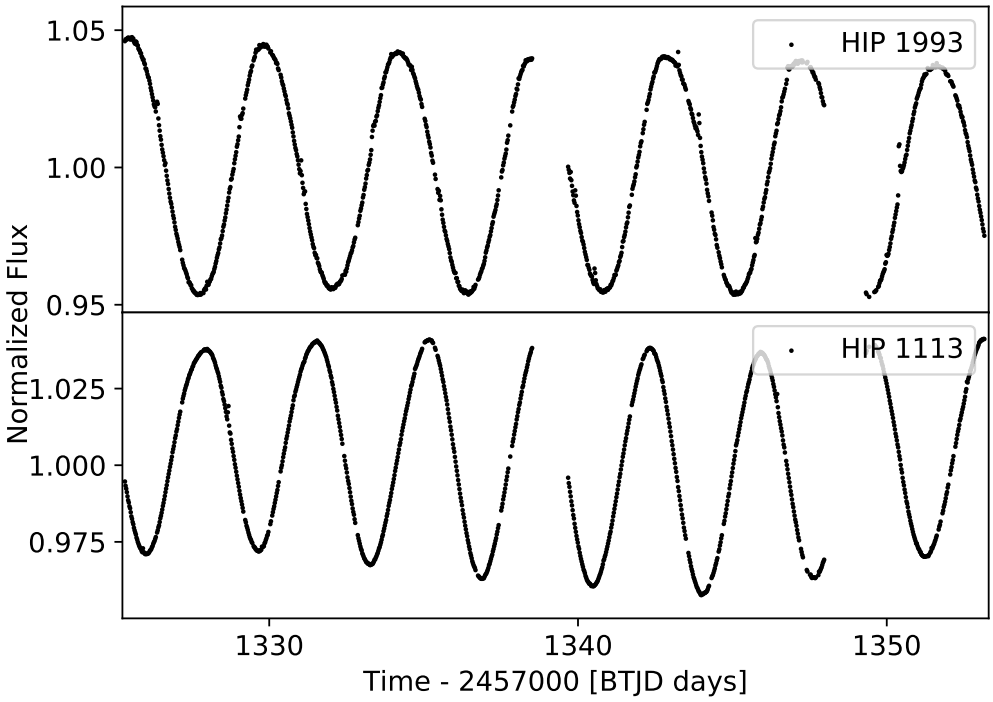

Figure 4.8: Examples of rotation/activity type 1: Near-uniform sinusoidal variation - HIP 1993 (top) and HIP 1113 (bottom) in Sector 1

sectors of data for which the 30min DIA data exists. Other sources such as HIP 1481 appear to exhibit 'beating'-like behaviour in their light-curves, going through periods of more and less intense oscillations yet with similar periods throughout. Thus for these objects, while the strong periodic nature of their oscillations makes the periods easy to identify, blind removal of these varying oscillations is difficult on a wide-scale basis. Nonetheless, the peak-cutting technique discussed in Section 2.6.2 is particularly beneficial for sources of this type, such as HIP 32235 (Fig 2.6) and HIP 105388 (top, Fig 4.9).

Alongside these periodic light-curves, a smaller number of aperiodic variations (or perhaps those with rotation or variation periods much longer than the 27 day observation time) were observed. Most of these variations were sufficiently long-period to be easily removed by the base 30-bin LOWESS-smoothing method (as was the case for J0449-5741 and HD 35289 (Fig 4.10)), or are sufficiently aperiodic that any activity would not overcome true periodic transit signals on the BLS periodogram. However, some more complex aperiodic cases exist, such as TIC 348839788 (Panel d, Fig 4.7), where two separate rotation and activity profiles seem to be apparent - one aperiodic larger amplitude evolution clearly evident in the

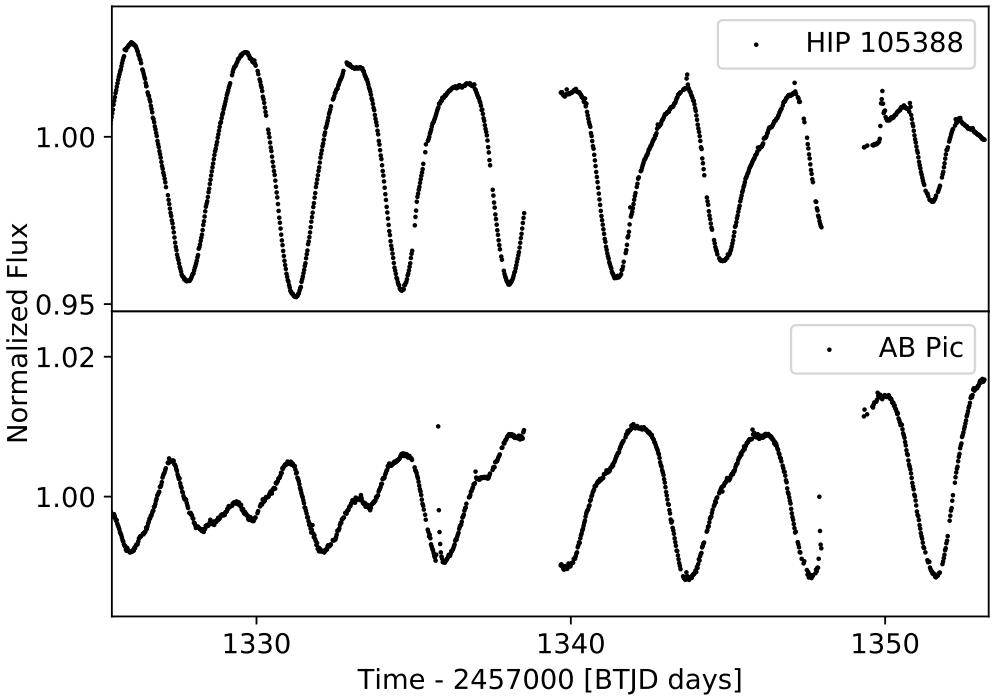

Figure 4.9: Examples of rotation/activity type 2: Periodic with rapidly evolving amplitudes - HIP 105388 (top) and AB Pictoris (bottom) in Sector 1

original light-curve, and one much faster rotation-based signal with a period of approximately 0.27 days (see panel d, Fig 4.7. This 'double variation' is present in a small number of other sources too, so should be considered carefully in future detrending efforts. For such sources this will likely require at least two detrending steps (a wider smoothing to remove the large-scale activity, followed by smaller-scale smoothing/modelling of the rapid rotation) in order to yield a flat pre-BLS search light-curve.

Perhaps the greatest challenge in the search for young exoplanets is the case of rapidly rotating stars such as HIP 22295 (Panel e, Fig 4.7) and CD-46 287 (bottom panel, Fig 4.13. These fast rotators are unfortunately quite common in stars of such young ages (indeed at 52 stars (20.3%) in this sample had rotation periods of less than 1 day), likely due to leftover angular momentum from their formation [Prialnik, 2009]. For the fastest rotators, 30min data alone struggles to untangle rotation from any potential transit signals, especially when rotation periods of less than 1 day are coupled with significant changes in amplitude. In order to more effectively dissociate such rapid rotation from transit signals, using more sensitive 2min cadence data is preferred, likely coupled with more intelligent modelling than

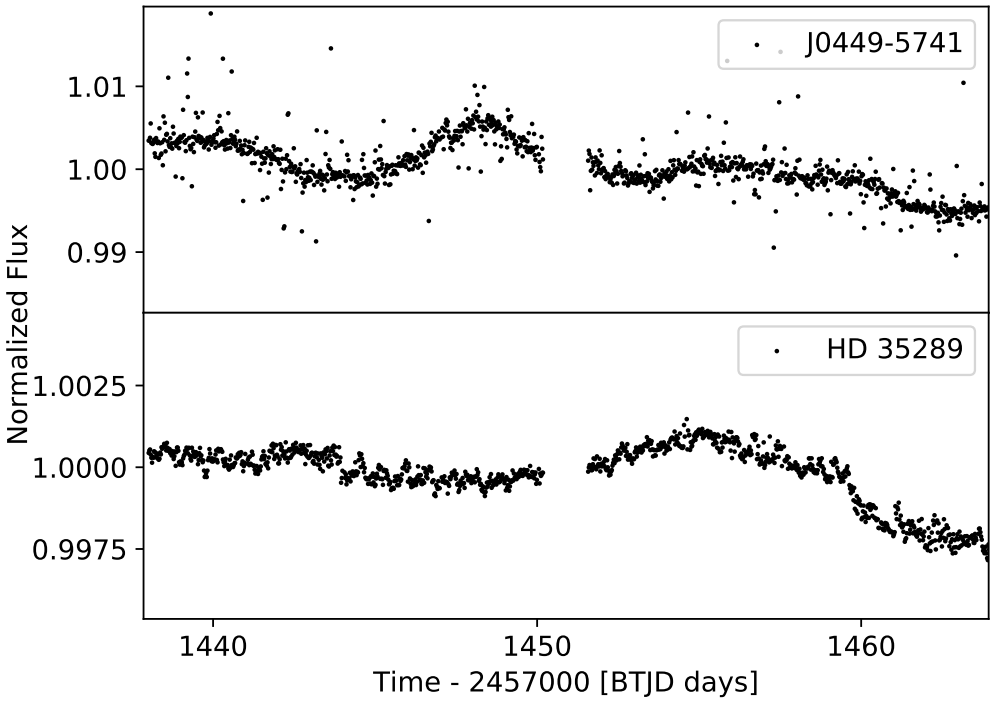

Figure 4.10: Examples of rotation/activity type 3: Aperiodic variations. DIA 30min-cadence light-curves for J0449-5741 (TIC 220425740, top) and HD 35289 (bottom) in *TESS* Sector 5

simple smoothing methods. The planned *TESS* 20s cadence data would also aid this effort, as rotation profiles will become more carefully defined and more detail of transit ingresses/egresses may become apparent. However, techniques to overcome this fast rotation challenge still need to be developed.

4.3.5 Sensitivity analysis results

Overall results

The conducted sensitivity analysis (discussed in section 2.6.5) revealed a number of interesting results, for which an overall summary is presented in Table 4.3 and Figure 4.11. What is immediately apparent is the steep drop-off in recovery as the $R_{\rm P}/R_*$ radius ratio decreases, falling from 77.6% at a radius ratio of 0.1 down to 20.4% at 0.03. This is to be expected as the inherent scatter and leftover variability amplitude of the light-curves steadily overcomes the signal of the smaller injected planets.

Interestingly the recovery of injected signals was not entirely consistent across

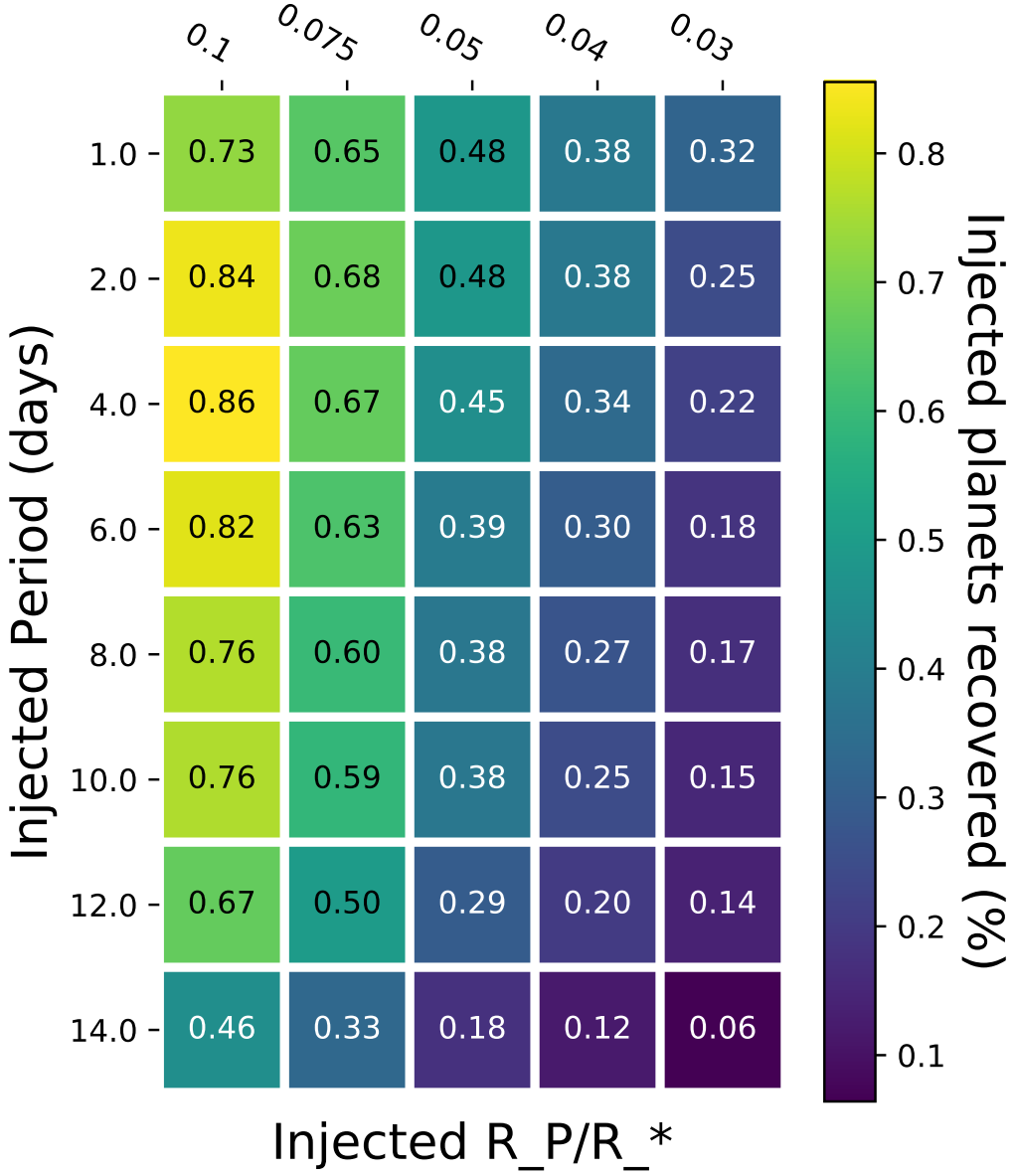

Figure 4.11: Overall sensitivity analysis results (Sectors 1-5) by injected planetary period and radius ratio. The numerical contents of each box corresponds to the overall percentage recovery of injected planets for each particular combination, with colour-gradient of the same added for clarity.

Table 4.3: Results from complete Sector 1-5 sensitivity analysis for the 256 stars in the BANYAN young star sample with DIA FFI data. Percentage recovery is presented both overall and in each individual sector. Note that results from the 14d injections were excluded from the percentage recoveries due single transits frequently being present at this period. Total number of individual sources for each sector are: Sector 1: 74; Sector 2: 77; Sector 3: 75; Sector 4: 120; Sector 5: 138.

R_P/R_*	Overall (%)	S1 (%)	S2 (%)	S3 (%)	S4 (%)	S5 (%)
0.1	77.6	79.7	80.3	73.1	69.8	84.3
0.075	61.7	61.6	64.0	56.8	51.4	72.2
0.05	40.9	43.8	36.9	39.6	31.6	50.31
0.04	30.2	31.3	30.4	31.1	20.0	38.0
0.03	20.4	24.3	18.4	22.1	11.1	26.6

the different sectors viewed in this analysis. This discrepancy appears to result from two major factors: comparative sector systematics and evolving light-curve amplitudes for a small subset of stars. Regarding the first factor, Sector 4 was the worst affected, with recovery rates for all radius ratios consistently far below the average. This is likely due to the strong reflected light glints seen at the end of each orbit for some sources (particularly those on camera 4), which are harder to systematically remove without also removing useful data from other light-curves. Conversely, recovery was significantly better in Sector 5 compared to the other sectors, with 26.6% of injected planets still retrieved down to radius ratios of 0.03. This inter-sector variation is well illustrated by the light-curves for J0455-6051/TIC 55651278 (an M5 star in the AB Doradus Moving Group), where injected planetary signals were recovered down to radius ratios of 0.075, 0.05, 0.075, 0.1 and 0.03 in sectors 1, 2, 3, 4 and 5 respectively (for planetary periods of 2-6 days). In this case such variation was caused by a combination of evolving flaring-type stellar variability coupled with extra noise in Sector 4. On the other hand, AB Pictoris demonstrates the second source of sector-dependent sensitivity, with injected planetary signals recovered down to steadily larger radius ratios of 0.04, 0.04, 0.075, 0.075 and 0.1 in sectors 1-5 due to the evolving stellar variability increasing in amplitude over time. While differences in the systematics of each sector can reasonably be expected to reduce over time as the *TESS* satellite pointing is refined, the evolution of individually active sources will remain a challenge not only for *TESS* but also for future missions such as *PLATO*.

A slightly more complex trend was observed when varying the injected planetary periods, as is highlighted by the overall heat-map in Figure 4.11, and the sector-by-sector heat-maps in the Appendix. For all sectors except Sector 2 (where

Table 4.4: Full sensitivity analysis table for each of the 256 stars with DIA light-curves. Includes information on the highest likelihood period recovered for each star in every sector it appears, with injected 1-14 day period planets and R_P/R_* radius ratios from 0.1 to 0.03. Full table available online.

Target ID	RA (deg)	Dec (deg)	Sector	Injected Period (d)	Radius Ratio	...	Log likelihood	Max Period (d)	Recovered?	Notes
2M0123-6921	20.79	-69.36	1	1.0	0.1	...	0.00121	2.65	TRUE	Alias
2M0123-6921	20.79	-69.36	1	1.0	0.075	...	0.00125	3.98	TRUE	Alias
...	...	...	...	...	...	...	...	...	...	...
WX Col B	84.30	-42.72	5	14.0	0.04	...	2.57e-5	12.52	FALSE	-
WX Col B	84.30	-42.72	5	14.0	0.03	...	1.95e-5	11.22	FALSE	-

a 2-day period was preferred), the recovery peaked around a 4-day injected period for the deepest ($0.1R_P/R_*$) transits, unlike the shortest 1.0 day period that may be initially expected. However as the injected planet radii decreased a more standard drop in recovery from 1-14d was observed. Such behaviour suggests that this form of analysis is most sensitive to larger planets in the 2-6 day period regime, and most sensitive to closer-orbiting planets as radii decrease further. In case the observed drop for 1-day planets was due to 1-day Earth-related systematics such as Earth-shine, the injection/recovery analysis was repeated with a 1.1d period planet, however similar recovery rates were observed to the 1-day case. Instead this discrepancy may be related to a combination of the increased activity of young stars coupled with the small number of data points per transit for such short-period planets. The primary reason for the much lower recovery of planets with injected periods of 14-days was that data gaps and randomly injected epochs frequently led to single transits appearing in the data-set. This same effect was also observed for some 12-day injections. Some interesting sector-to-sector variations are also clear in the sector-specific heat-maps (see Figures A.1-A.5), with Sector 4 exhibiting the lowest recovery rates (especially for the smallest radius planets), and Sector 5 the highest.

Another useful product of this sensitivity analysis is the individual sensitivity to planet detection for each star. An overview of this table is presented in Table 4.4, with a full version available online.

Rotation period vs recovery

As has been seen in Section 4.3.4, a wide range of rotation periods was observed for stars in this sample. This provides an interesting opportunity to test the relationship between rotation rate and recovery, which is especially important given concerns about finding planets around swiftly rotating stars and those with rotation periods near the injected planetary period. In order to investigate this relationship, an injection/recovery analysis of rotation rate vs depth recovered for each target was conducted for each injected period. A typical example of one of these plots is shown

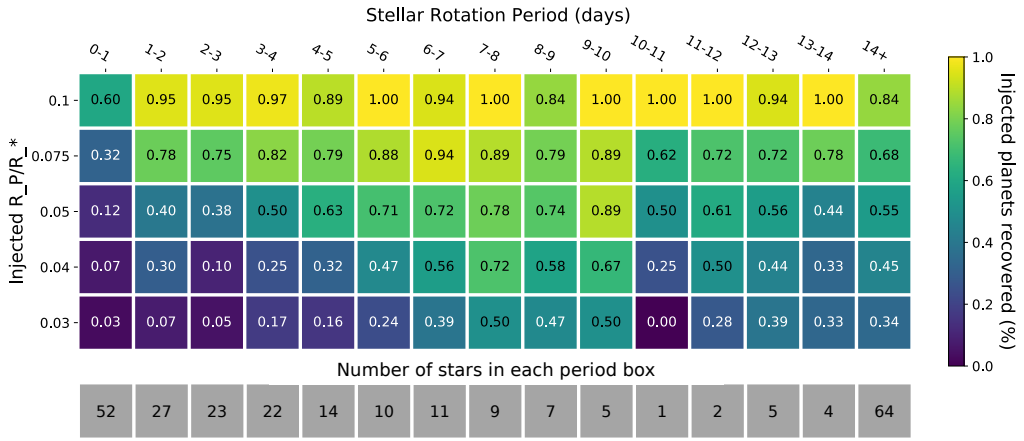

Figure 4.12: Stellar rotation period vs recovery depth for a 4-day injected period planet across all targets and sectors. Design similar to Figure 4.11. The number of stars in each rotation bin tested is given in grey below the main sensitivity analysis.

in Figure 4.12 with an injected period of 4.0 days. Somewhat counter-intuitively, while there is a slight skew towards larger radius ratios below rotation periods of 5 days, there is no significant evidence that the overall recovery depth of injected transits is a function of rotation period, with recoveries down to radius ratios of 0.03 (and complete non-recoveries) across the entire period range. In addition, recovery depth does not seem to be detrimentally affected by being close to the injected period (4 days in this case), as similar recovery was observed for every injected period. Note that many variations between individual period ranges can be explained by the relatively small number of targets in this sample, especially in the 10-11d period case where only eight targets were present with periods in this range. However, it should be noted that 60% of all unrecovered signals were from targets with rotation periods of less than 1 day. This constituted 40% of all injected signals into targets with these rotation periods. Furthermore the vast majority (79%) of the recovered light-curves with rotation periods <1 day were only found down to radius ratios of $0.075 R_{\rm P}/R_*$). This suggests that searching for small planets ($<0.075 R_{\rm P}/R_*$) around stars with rotation periods less than 1 day is likely futile until better techniques are developed to detrend this fast rotation. However, since failures for all other rotation periods were consistent at 0-3 light-curves per 1-day period interval (e.g. 4-5d), searching for planets around stars with rotation periods longer than 1 day shows promise, even down to planets with radius ratios of $0.03 R_{\rm P}/R_*$.

Table 4.5: Comparison between percentage recovery of injected planets in Sector 1 for both in the original and peak-cut light-curves.

Radius Ratio	Number of lcs	Original (%)	Peak-Cut (%)
0.1	61	90.2	82.0
0.075	59	62.7	64.4
0.05	59	39.0	40.7
0.04	60	31.7	35.0
0.03	60	25.0	25.0

Effectiveness of the peak-cutting technique

Since the peak-cutting technique exhibited variable effectiveness according to the shape of the light-curves in building the pipeline, much of the prior analysis was undertaken without the peak-cutting option applied. However, to test the wider effectiveness of this option and evaluate where its application was most useful, a comparison study was undertaken for all targets in Sector 1, both with and without peak-cutting. For this comparison test planets with a set period of 8.0d and radius ratios of $R_P/R_* = 0.1$ to 0.03 were injected into each of the light-curves in Sector 1. Light-curves were then detrended using the standard 30-bin LOWESS-detrending method described in section 2.6 and searched through using the standard BLS method.

Of the 74 Sector 1 targets in the sample with light-curves available from Oelkers and Stassun [2018]'s DIA FFI pipeline, the basic peak-cutting technique failed for 13-15 of the objects (depending on the injected radius ratio). In all cases this was due to peaks and/or troughs not being located by the automatic `find_peaks` function, typically because the light-curves were simply too flat to exhibit any significant peaks or troughs. Of the remaining 59-61 light-curves, recovery with peak-cutting was in general comparable, or slightly better than, the recovery of planets when peak-cutting wasn't applied (as is summarised in Table 4.5). The one exception to this was the $0.1R_P/R_*$ case, which was caused by the peak-cut analysis failing for three of the original 0.1 radius ratio light-curves.

However, it is when looking at individual targets that the power of the peak-cutting technique is most evident. For seven of the targets in Sector 1, the use of the peak-cutting technique yielded significant improvements in the recovery of smaller injected radius ratios. The most significant improvements were seen for the object J0247-6808, where injected planets were recovered down to a radius ratio of $0.03R_P/R_*$, despite not being recovered for any radius ratio in the non-peak-cutting case. Similarly in the case of HIP 32235 the depth recovered dropped to

0.03 from 0.075 in the non-peak-cut analysis. Similar improvements (though with less significant drops in recovered depth) were observed for RBS 38, TYC 8895-112-1, J0346-6246, J0414-7025 and J2231-5709.

There were two different reasons why recovery was improved for these objects - one showing the technique working as designed and the other a fortuitous side-effect. The former can be seen in HIP 32235, RBS 38, TYC 8895-112-1 and J2231-5709 where each light-curve exhibits sharp variability peaks with periods of order 3-8 days and amplitudes greater than 2.5%. In this case the technique aids recovery of the planets by successfully cutting the sharp turning points of these light-curves which were previously leftover as false-transits after the LOWESS-smoothing step. It is these types of light-curve variability (approximately $3 < P_{rot} < 8$ days; Amplitude $> 2.5\%$) which are best-handled by using the peak-cutting technique. All light-curves with rotation periods of 3-8 days which weren't improved by peak-cutting were later found to have had at least one of their transits cut, however their depths reached were not affected. The other three light-curves improved by the use of this technique (J0247-6808, J0346-6246 and J0414-7025) were aided accidentally, having sections of increased scatter masked as a result of the applied cuts. This is a convenient side-effect of searching for planets among light-curves with intrinsically higher scatter, but less scientifically interesting.

The limitations of this technique were identified by investigating the small selection of stars detrimentally affected by the peak-cutting. The three main failure modes were:

1. Effectively flat light-curves dominated by scatter
2. Light-curves with activity/variability periods of $\sim$ 2 days
3. HIP 33737 - a star with a flat-bottomed rotation activity

In the first case, all peak-cut does is remove useful data, since no significant activity-based peaks and troughs were present. Meanwhile in the 2-day activity/variability period case, a significant portion of the data is cut, with the remaining intervals between each cut too short to be effectively flattened by the 20-30bin (10-15hr) LOWESS-smoothing. These two failure modes effectively represent the two limits of usefulness for this technique in a single *TESS* sector, of order 15 days and 2 days respectively. The final case (HIP 33737) was a unique one, where the light-curve exhibited unusually flat-bottomed troughs well-handled by the LOWESS-smoothing technique, and thus the peak-cutting only removed useful data, alike to the first failure mode.

Crucially however, for all light-curves which did not fall into one of the three failure modes identified above, the use of the peak-cutting technique was not found to affect the shallowest depth of transit recovered. This alleviates the chief concern about the use of the peak-cutting technique: that the recovery depth may be reduced if transits near peaks and troughs are inadvertently cut. Overall then, this peak-cutting technique shows the greatest effectiveness for stars exhibiting activity/variability of periods 3-8 days, but may be applied to all stars with periods of approximately 3-15 days without significant detriment, unless those light-curves are particularly flat.

4.4 Discussion

As this work has shown, young host stars present many extra challenges in comparison to the generally older, less active stars previous exoplanetary searches have been biased towards. Quicker rotation, increased amplitude activity and other strong stellar-based periodicities wash out candidate planetary signals in BLS searches and make finding transiting signals harder using the traditional automated exoplanet-searching tools. Furthermore, by eye the large amplitude variation of many sources such as HIP 105388 effectively hide real transit signals unless careful detrending is applied first.

Nonetheless, the methods presented in this work have shown promise at pushing down to lower-radius planets around young stars. The base LOWESS-detrending method provides a useful combination of smoothing and polynomial fitting and generally outperforms more traditional exoplanet smoothing methods such as Savitzky-Golay filters for younger stars due to its weighted approach to smoothing. Furthermore, for sharp but evolving oscillations such as those seen for HIP 33235, cutting the peaks and troughs of these oscillations yields a significant improvement in the recovery of injected planets, especially for periods of 3-8 days. After initial recovery, the shape and depth of the retrieved transit is then greatly improved by incorporating the developed activity interpolation over transit gaps.

Overall, two young exoplanets were recovered in this work (DS Tuc A b and AU Mic b), both of which were previously known from searches in the short-cadence (2min) *TESS* data. If the planet occurrence rate of young exoplanets is assumed similar to that for older exoplanets, it is possible to compare this number to the expected exoplanet yield from this sample of 256 young stars. For planets with periods between 0.8-17 days (or approximately those which would result in two or more transits in a single sector of *TESS*), Fressin et al. [2013] calculated that the

average cumulative number of planets per star was 25.3 $\pm$ 1.6%, or roughly one planet per every 4 stars. Combining this with the geometric transit probability ($P_{\rm trans} \approx \frac{R_*}{a}$; taking the average a/R_* of transiting planets from the exoplanet archive), this suggests an overall yield of 3.81 $\pm$ 0.24 planets from the searched sample of 256 young stars. However, the conducted sensitivity analysis revealed that only 46.2% of planets injected into these young stars were recovered by this pipeline, largely due to the increased stellar activity of these hosts. Taking this into consideration, the final expected planetary yield from this search is 1.76 $\pm$ 0.11 planets. This agrees within 3σ with the two young planets recovered here.

As demonstrated by the sensitivity analysis, the large range in activity type, period and amplitude coupled with the variation in scatter for each source results in significantly different recovery rates for injected planets overall, ranging from sources like CD-46 287 (a K6Ve star in the 35 Myr Octans association, viewed in sector 2) where no planet was recovered even for the largest radius ratio, to HD 202969 (an F8/G0V star also from Octans, viewed in sector 1), where the injected planet was easily recovered across the whole radius ratio range. Comparing these two light-curves (shown in Fig 4.13), the reason for this discrepancy is immediately clear; while the period of the rotational variation in the flux from HD 202969 is much longer than that of the transit duration (and easy to smooth with the 30-bin LOWESS-detrending technique), the rotation period of CD-46 287 is 0.37d, well within the realm of the duration of a planetary transit. Since this pipeline is designed to only remove variation with periods longer than a transit it struggles to remove such short period variations, and as such cannot recover injected planets from such light-curves.

Observing other sources for which no injected planet sizes were ever recovered, the major challenges appear to be rotation with periods <1 day, excess scatter, large amplitude flares or other spurious outlying points in the light-curve. These latter two problems present a problem to automatic detrending due to their unique light-curve by light-curve behaviour. Nonetheless these could conceivably be flagged by shape or sigma-clipping and handled on a case-by-case basis. On the other hand, scatter falls into two categories: time-localised scatter such as the scattered light glints seen at the end of some sector 4 light-curves, and non-localised scatter which affects the entirety of the light-curve. The first of these problems is soluble, and is likely best handled in a similar way to flares and instrumental scatter, by flagging and removing specific time periods which are affected. However, excess scatter which extends to the full extent of the light-curve (which can be in excess of 2%, as in the case of J0122-2566) likely precludes the use of these light-curves for planet

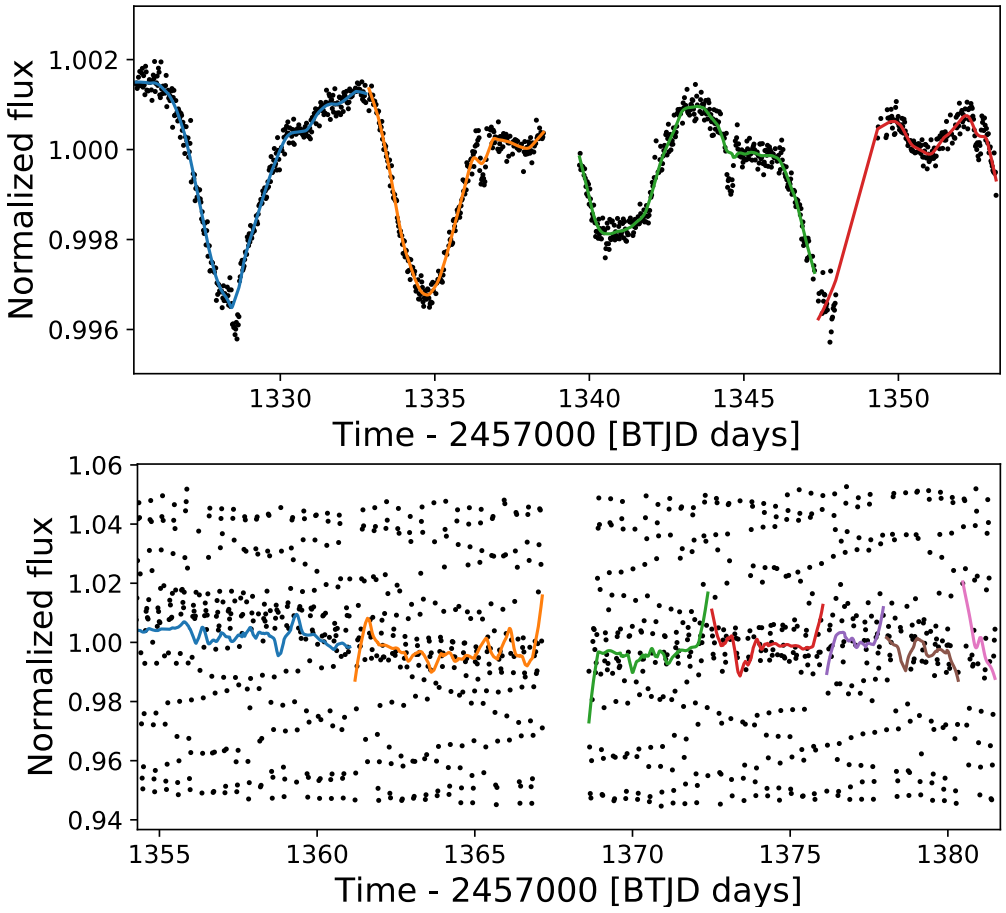

Figure 4.13: Light-curve comparison for HD 202969 (above) and CD-46 287 (below) with an 8.0d, 0.03 R_P/R_* radius ratio injected planet. While the LOWESS-detrending pipeline has no trouble modelling HD 202969 due to its longer duration variation, it struggles to detrend the sharp 0.37d period variability of CD-46 287.

searches. This latter type of scatter was most common for very dim stars in the sample, largely with *TESS* T magnitudes in excess of 13.

The problem of fast rotation is a more systematic issue and is arguably the most important detrending-related challenge presented by young stars. In this work the detrending pipeline struggled to recover injected planetary signals in light-curves with large amplitude flux changes coupled with rotation periods of less than 1 day, largely due to the width of the LOWESS-smoothing being at least 20-bins, or 10hrs wide. Since any variability with a period of 1 day or less will involve at the very least one turning point, such widely-spaced windows can not reasonably be expected to accurately trace and smooth such activity/rotation. Indeed, as discussed in section

4.3.4, some of the fastest rotation observed had periods of less than 0.4 days, or less than 10hrs. However, it is not unheard of for verified planets such as Kepler-1283 b [Morton et al., 2016] to have transit durations as long as 0.4 days, so dropping the LOWESS-smoothing window to any shorter than this could result in significant distortion of potential transit shapes along with removal of stellar activity signals. It is thus clear that a more targeted method is required for such fast rotators, likely with more intelligent modelling-based methods such as Gaussian Processes [e.g. Gillen et al., 2020]. Unfortunately however GP methods are much more intensive than a simple smoothing or polynomial based methods, and require a more informed knowledge of a star's characteristics. This makes GP-based techniques very well suited to in-depth followup of candidates, but perhaps applying such methods to the entire stellar sample would be less efficient than simply flagging them in an initial BLS periodogram and analysing them separately. At these early stages it is still worth considering other modelling methods as well (such as fitting a sum of sinusuoids as in Gillen et al. [2020]), especially since Hippke et al. [2019] found such a range in effectiveness in young star detrending methods. If a simpler algorithmic method can be found to model or detrend these very fast rotators then it could help to speed up wide-field transit searches around young, rapidly rotating stars.

Another interesting challenge highlighted by the sensitivity analysis is the rapid evolution of some young star light-curves. This evolution makes modelling the light-curves more difficult and also means that detrending is more effective in some regions of the light-curve than others. As demonstrated for AB Pictoris, this changes how easy it is to recover injected (or no doubt real) exoplanet signals depending on when in the activity cycle one views the star. If however one gains a better understanding of a star's activity cycle through long-term monitoring and asteroseismology, this evolving activity can become an opportunity for increasing the effectiveness of planet searches by targeting quiet sections of the stellar activity cycle. This knowledge would also be crucial for radial velocity follow-up of young host stars, since as Oshagh et al. [2017] have shown, for very active stars radial-velocity jitter is correlated with photometric variation in stellar light-curves. Hence being able to predict epochs of low stellar activity based on knowledge gained from photometric monitoring of these young stars will prove crucial for characterising any discovered planets through radial-velocity followup.

One thing that was made increasingly clear from this initial survey of young star light-curves is that the 'one size fits most' approach of large-scale photometric surveys such as *Kepler*, *K2* and the main *TESS* SPOC pipeline is often not appropriate for young stars given the large range in shape, amplitudes and periods

of light-curve variability observed. Indeed, as discussed in section 4.3.4, even in this relatively small number of young stars surveyed periods were observed to vary between 0.27d to non-variable over the 27d observation window. Meanwhile activity and rotation behaviour varied from near-uniform to constantly evolving, and variability shape changed from beating sinusoids to almost flat aside from significant flaring activity. In order to more comprehensively search for planets around such active stars, future detrending pipelines need to focus on more effectively matching detrending techniques to the primary type of light-curve variability observed. This could be achieved through an initial automated variability-type assignment, similar to - but more advanced than - the current 'variable' vs 'non-variable' assignment implemented into the *Kepler/K2/TESS* pipelines. By creating defined groups of similarly shaped light-curves and types of variability/activity, machine-learning techniques could then be used to assign light-curves the most effective detrending technique based on their perceived 'group' of variability. Particularly important will be separating known types of intrinsic variability, quick rotators and rapidly evolving light-curves, however many other important groups may become obvious with time. A more in-depth look into different types of variability in young stars over all sectors would thus be very beneficial, and may inform future detrending methods in missions like *PLATO* [Rauer et al., 2014]. This provided the primary motivation for beginning the YOUNGSTER programme discussed in chapetr 5.

Attempting to understand these stars in more detail raises the important question of whether 30min data is enough to characterise any discovered variability or exoplanet candidate signals, or whether 2min cadence data is required. In this work 30min cadence light-curves were shown to effectively find the period of primary variability (and often second and third variability periods), and to recover all of the currently known TOIs identified through the 2min data, so it is undeniably a very powerful data source for the general stellar sample. However, for those TOIs recovered, exposure smearing resulted in smaller transit depths and thus less accurate transit parameters compared to the 2min data. This highlights the importance of the 2min data (or alternative follow-up photometry) for accurate characterisation of any discovered signals. Furthermore, while the 30min data was sufficient for characterisation of longer-period variability and activity in the light-curves, as the community pushes towards shorter period rotation (especially that with periods of less than 1 day), dissociating transit signals from stellar activity and rotation signals becomes increasingly difficult. This is the realm where 2min data may be crucial in the search for young exoplanets. Furthermore, as the community attempts to understand the causes and evolution of activity in young stars, 2min cadence data

would significantly aid asteroseismic efforts for these young stars. However, a full comparison between 2min and 30min data-sets still needs to be undertaken before significant conclusions can be made.

4.5 Summary

In this work, techniques have been developed to aid the search for transiting exoplanets around young, active stars in the 30min cadence *TESS* FFI data. Young exoplanets (<1Gyr in age) inhabit a very important part of the exoplanet evolutionary timescale, where formation mechanisms, accretion, migration and dynamical interactions can significantly change the shape of observed planetary systems. However, they are also typically situated around young, active and often rapidly rotating host stars, severely hampering the discovery of new planets using the transit method. The developed method attempts to detrend these spurious stellar activity signatures using a 20-30 bin LOWESS method of Cleveland [1979a] at its base, combined with automated peak-cutting and activity interpolation over transit gaps to more effectively differentiate activity from transit signals and preserve the transit shape. A basic version of this pipeline is made available online.[7] It is hoped that using this method in tandem with other detrending/transit-search pipelines such as the main *TESS* SPOC pipeline [Jenkins et al., 2016] and Gaussian-Process based methods [e.g. Gillen et al., 2020] will expand the number of young planets that can be found.

These techniques were applied to young stars in stellar associations from the extended BANYAN sample [Gagné et al., 2018b,c; Gagné and Faherty, 2018], using the *TESS* Sector 1-5 light-curves derived from the Difference Imaging Approach (DIA) pipeline of Oelkers and Stassun [2018]. Lacking the data quality-flags of the *TESS* 2min cadence data, periods of excess pointing scatter were instead removed by considering the *TESS* data release notes and the engineering quaternion data.

While no new exoplanet candidates were revealed in this work, results from this initial survey revealed a variety of interesting objects, including the retrieval of the new young exoplanet DS Tuc Ab, TOI 447.01, TOI 450.01, a number of eclipsing binaries and a large array of interesting rotation and activity. In order to test the sensitivity of the developed detrending techniques to different planetary sizes, model `batman` transits [Kreidberg, 2015] were injected into each of the young star light-curves at a range of $R_{\rm P}/R_{*}$ radius ratios from 0.1 to 0.03 and periods from 1-14 days. The percentage of recovered transit signals from the injected planets dropped from 77.6% at a radius ratio of $0.1R_{\rm P}/R_{*}$ to 20.4% at a ratio of 0.03, however was seen to

[7] https://github.com/mbattley/YSD

vary considerably between different targets and sectors. Meanwhile while increasing the injected planet period was seen to result in a decreasing recovery rate for smaller planets, the recovery rate was actually observed to peak around periods of 2-6 days for larger planets. An investigation into the relationship between rotation period and recovery depth did not suggest that the two were significantly correlated, aside from a slight skew towards larger planets at short rotation periods and the known difficulty of very short-period (<1day) rotation.

These results alongside deeper examination of light-curves in this sample lead to a number of interesting conclusions. The sensitivity of specific young star light-curves to transit searches appears to be most limited by fast rotation (<1 day rotation periods), excess scatter, scattered light glints and significant flaring activity. Meanwhile the rapid evolution of many young star light-curves offers both a detrending challenge and a potential opportunity to search more efficiently for exoplanets at less active times. Given the vast array of different types of young star light-curves seen in this initial survey, in the future a multi-faceted detrending approach which first classifies light-curves according to their broad activity/variability is perceived as beneficial. It is clear that the 30min cadence data shows particular promise for additional detections to the main SPOC 2min pipeline, as it is capable of retrieving all of the TOIs highlighted in the sectors analysed. However, planet parameters derived from the 30min light-curves can be less reliable due to phenomena such as exposure smearing. Thus the acquisition of 2min light-curves for young stars remains desirable, especially since this will also aid in-depth astroseismic characterisation of these sources.

Hence overall, although searching for planets around young, active stars undeniably presents several extra challenges compared to searching around older host stars, the techniques developed in this work are beginning to delve more effectively into this age range. However, significant challenges such as very fast rotation and rapidly evolving activity remain, standing in the way more extensive young exoplanet discoveries. In order to extend the search to these more challenging cases, it is imperative to both develop techniques to automatically classify and understand such types of young star variability and also to develop targeted detrending techniques to handle these in future planet searches.

Chapter 5

The YOUNGSTER Program

> "People assume that time [like flux] is a strict progression of cause to effect, but *actually* from a linear, non-subjective viewpoint - it's more like a big ball of wibbly wobbly... timey-wimey... stuff"
>
> — The Tenth Doctor, *Blink*, *Doctor Who*.
>
> Square brackets are the author's.

As discussed in Chapter 4, the increased activity and variability of young stars results in complex photometric light-curves which are significantly harder to search for transiting exoplanets in. Although young-star specific methods such as those developed in the previous chapter are extending the search further compared to traditional planet searches, in order to target planets around more challenging stars, it is important to understand the variability of the young stellar hosts alongside the detrending and planetary search. With that in mind, this chapter presents the beginning of the YOUNGSTER program, or "YOUNG Star detrending for Transiting Exoplanet Recovery", which aims to classify young stellar variability prior to planetary searches and then use these classifications to inform targeted detrending methods. This chapter presents the first step in this process, using Kohonen Self Organising Maps (SOMs) to explore young star variability present in the first year of observations from the *Transiting Exoplanet Survey Satellite (TESS)*, with such knowledge valuable to understand how these may be split into different groups in future searches for young exoplanets. This technique was found to be particularly effective at separating the signals of young eclipsing binaries and potential transit-

ing objects from stellar variability, a list of which are provided in Table 5.2. The effect of pre-training the Self-Organising Maps on known variability classes was also tested, but found to be challenging for dimmer objects without a significant training set from *TESS*. SOMs are also found to be useful as an intuitive and informative overview of leftover systematics in the *TESS* data, providing an important new way to characterise troublesome systematics in photometric data-sets. The results from this analysis provide a crucial building-block towards more targeted detrending of young stars and will extend the search for young exoplanets to smaller planets around more active hosts.

This chapter begins with an introduction to the YOUNGSTER programme in Section 5.1 before Section 5.2 discusses the data preparation ahead of the main Self-Organising Map analysis. A review of setting up, training and interpreting Self-Organising Maps in the context of this work is presented in Section 5.3, which is followed by an examination in Section 5.4 of how the SOM technique can aid systematics detection and removal in *TESS* data. The results of the overall Sector 1-13 SOM analysis are presented in Section 5.5, followed by a wider discussion of the results in Section 5.6. This chapter concludes with a summary in Section 5.7.

Declaration - The following chapter is a partial reproduction of the paper 'YOUNG Star detrending for Transiting Exoplanet Recovery (YOUNGSTER) I: Using Self-Organising Maps to explore young star variability in Sectors 1-13 of TESS data' [Battley et al., 2022]. All work in this chapter was carried out by the author, with co-authors to the paper it is based on providing high level guidance and simple proof-reading.

5.1 Overview of YOUNGSTER Program

The aim of the "YOUNG Star detrending for Transiting Exoplanet Recovery" (YOUNGSTER) programme is to develop a targeted detrending approach in young exoplanet searches by using information about the stellar variability of potential young host stars to inform detrending of different categories of variability. Unlike a traditional 'one size-fits-most' detrending approach, this program relies on first classifying different types of young star variability which can then be treated independently. By considering the variability of potential host stars during the analysis (rather than simply eliminating it), this approach helps to reduce the prevalence of leftover signals induced by improper detrending and should aid the search for shallower exoplanets around young stars considerably. Furthermore, it provides additional information about the individual systems, thus aiding further analysis of

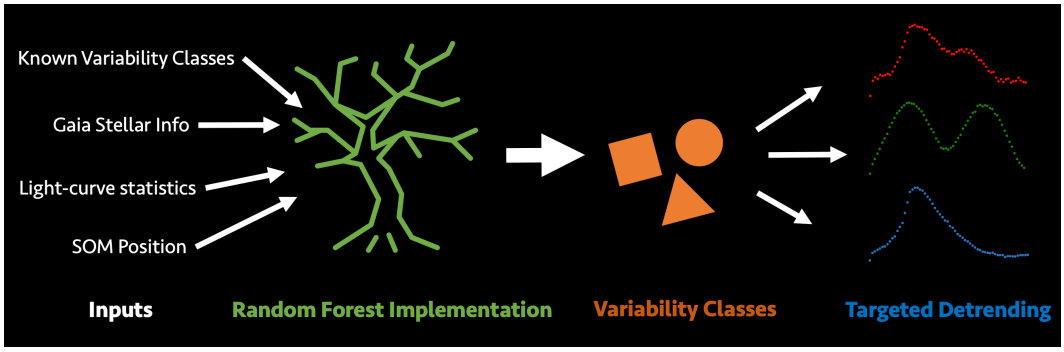

Figure 5.1: Overview of YOUNGSTER program

the system as a whole.

The YOUNGSTER programme was instigated by the results and analysis presented in [Battley et al., 2020, hereafter YOUNGSTER I] and discussed in Chapter 4. This chapter presents the results from the use of an unsupervised Self-Organising Map [SOM, Kohonen, 1982, 2001] as a tool to explore young star variability further, and acts as a first step into more targeted characterisation and sorting of young star variability for exoplanet searches. It also provides a useful additional method of recognising and removing leftover systematics in young star light-curves and gives an efficient method to recover example light-curves of unusual variability types on which to train future classification methods. Future work in this program will extend this classification using a supervised random-forest implementation which combines additional light-curve statistics with the results from the SOM in order to give final classifications for each star. Specifically, it will combine the SOM position and Euclidean distance of each light-curve from its closest SOM pixel with noise statistics and (a)periodic information about the light-curves in order to give a final classification for each light-curve. These classifications will then be used as test-beds to develop new targeted detrending methods and apply these to each variability group in order to search for smaller transiting exoplanets. A pictorial overview of the full YOUNGSTER program is shown in Figure 5.1.

5.2 Data Preparation

5.2.1 Light-curve extraction

For machine-learning approaches like Self-Organising Maps, two factors are crucial: consistency and data volume. Now, in the era of *TESS* [Ricker et al., 2014], the community has access to a uniform source of millions of light-curves from an ever-

increasing proportion of the sky, including hundreds of thousands of young star light-curves extracted from the Full-Frame-Image (FFI) data. Numerous different sources of FFI light-curves are now available, including a number of very large data-sources like the *TESS* SPOC [Caldwell et al., 2020a], MIT Quick-Look [QLP, Huang et al., 2020a] and *eleanor* pipelines [Feinstein et al., 2019]. Crucially for this analysis, there are now also a selection of dedicated young star light-curve extraction programs such as the CDIPS [Bouma et al., 2019], PATHOS [Nardiello, 2020], and Oelkers and Stassun [2018]'s Difference Imaging Analysis light-curves. These specialised extraction pipelines are particularly useful for dense regions where young stars are commonly found due to their difference imaging (n.b. also used to extract Huang et al. [2020a]'s QLP light-curves) or PSF-based approaches.

In this work, light-curves extracted from the Cluster Difference Imaging Photometric Survey (CDIPS) pipeline [Bouma et al., 2019; Bhatti et al., 2019][1] were chosen as the primary data source due to their dedicated young-star extraction approach and the availability of a large number of light-curves for the entire first year of *TESS*'s Primary Mission (Sectors 1-13). They also provide a 'lightly detrended' PCA form of light-curve particularly useful for the SOM technique, as discussed below. Through comparison to different year-1 light-curve sources currently available to the community, this form of light-curve was found to best preserve signals of stellar activity whilst removing additional spacecraft systematics. Note however that because no light-curves from the second year of *TESS*'s observations were available from the CDIPS pipeline at the time of writing, the more widely-available QLP light-curves of Huang et al. [2020a] were chosen as a backup data source for exploring scattered light systematics in section 5.4.

It should be noted that while times of known momentum dumps and coarse pointing were removed using the *TESS* data release notes[2] early in the CDIPS pipeline, a number of additional systematics from the spacecraft motion and scattered light still exist in the data after the main difference-imaging extraction. In order to combat this, Bouma et al. [2019] present two partial-detrending approaches: Principal Component Analysis (PCA) and the Trend-Filtering Algorithm (TFA). Both approaches rely on ensemble-detrending of the light-curve of interest using up to 200 nearby 'template' stars without significant periodicity. Principal Component Analysis works by deriving 10-15 principal components which comprise the systematic trends in each individual CCD and then building a model for each light-curve using a subset of these components, weighted by consideration of linear least

[1] http://doi.org/10.5281/zenodo.3370324
[2] https:archive.stsci.edutesstess_drn.html

squares. This removes the largest systematics on each CCD while avoiding overfitting of most stellar variability. The specific implementation used here was that included in the SCIKIT-LEARN package [Pedregosa et al., 2011]. Meanwhile TFA works by deriving a filter function to remove long-period trends and thus is well-suited to transit studies, but too severe for analysis of stellar signals, as it often removes/distorts interesting stellar variability alongside true systematics.

Due to its ability to remove spacecraft systematics whilst maintaining most stellar variability, the PCA detrended light-curves were eventually chosen as the primary brightness data-source for this analysis. For consistency, aperture 2 (1.5 pixels in radius) was used for all light-curves, chosen as a balance of noise and the relative crowding of many young stars. The final data vectors used for each object were thus time = *'TMID_BJD'*, magnitude = *'PCA2'* and mag_err = *'IRE2'*. Finally, these magnitudes were converted to normalised flux to prepare them for further analysis, based on the median flux for each source.

5.2.2 Target Selection

An additional benefit of using the CDIPS light-curves is that all light-curves extracted by the pipeline show evidence of youth, either through candidate membership of young clusters/associations/moving groups or due to external photometric indications. The CDIPS target list was constructed from four large archival catalogues of open clusters [Cantat-Gaudin et al., 2018; Gaia Collaboration et al., 2018a,b; Kharchenko et al., 2013; Dias et al., 2014] as well as nine smaller catalogs of moving groups and stellar associations [Gagné et al., 2018b,c; Gagné and Faherty, 2018; Kraus et al., 2014; Röser et al., 2011; Bell et al., 2017; Rizzuto et al., 2011; Oh et al., 2017; Zari et al., 2019]. This resulted in an overall target list of 1,061,447 stars, for which 671,894 light-curves are available over the first year of *TESS* observations.[3] All of these light-curves were included in the present analysis. For a full description of the target list and its construction see Bouma et al. [2019]. However, some additional care is needed when using this list, as Bouma et al. [2019] admit that the it was compiled for "completeness, not accuracy", so the true youth of some of the included objects should be treated with caution. Nonetheless, since the overall aim of the YOUNGSTER program is to understand and detrend different types of variability in the wider sample of young stars, the inclusion of some older, likely less-active sources is of little concern.

[3] https://archive.stsci.edu/hlsp/cdips

5.2.3 Period-finding and SOM array preparation

The first step in characterising the variability of the young star sample is deriving the variability period of each source. Although most instrumental systematic effects were removed in the CDIPS extraction and PCA steps, in order to combat the common systematic period of 13.5 days (half the length of a typical *TESS* sector), a 13.5 day period ($f = 1/14$) sinusoid of the form

$$y = a \sin 2\pi f t + b \cos 2\pi f t + c \tag{5.1}$$

was fitted to the data and removed by division from the flux data before the main period search. Primary, secondary and tertiary periods for each light-curve were then found using the common Lomb-Scargle method [Lomb, 1976; Scargle, 1982], as implemented into `python` as the LombScargle function in the `python` astropy library[4] [The Astropy Collaboration et al., 2013; Astropy Collaboration et al., 2018]. These periods were found iteratively, with the primary period in each iteration obtained using the Lomb-Scargle method and then removed by fitting a sinusoid of the same form used for the systematic 13.5d period. As the present work is most interested in the overall shape of the variability rather than the absolute period for each target, the periods were allowed to vary from 0.25 days (4 d^{-1}) to 25 days (0.04 d^{-1}), allowing the non-variable or aperiodic light-curves to remain largely unfolded whilst not risking the introduction of more noise-related frequencies at much longer frequencies. Given that most periods found were on the order of multiple days, extending the upper frequency to the Nyquist frequency ($\sim$1h period) is not expected to change this analysis significantly. Furthermore, attempting to remove stellar activity with periods shorter than a quarter of a day would be exceedingly difficult in the 30min data.

To prepare for the SOM analysis, the extracted CDIPS PCA light-curves were folded by their strongest period and normalised between 0 and 1 to allow direct comparisons of different shapes of variability. Similar to Armstrong et al. [2016], these phase-curves were binned into 64 bins in order to save computational time whilst maintaining the overall shape of the variability. The phased arrays were then aligned such that the minimum point always occurred as the first element in the array, before all phase-curves were arranged into per-sector lists as the main input to the SOM analysis. Where gaps existed in the light-curves of over 0.5 days, these were filled in with linear interpolation (after the Lomb-Scargle period search) in order to avoid processing issues. In most sectors this was only required across the

[4] https://docs.astropy.org/en/stable/api/astropy.timeseries.LombScargle.html

1.1 day inter-orbital gap, and thus was not found to significantly affect the shape of the binned and folded phase curve. An additional test confirming this more systematically can be found in the discussion. The final phase arrays were then arranged into a full input array for each sector, which could be input directly into the SOM algorithm.

5.3 SOM setup, training and interpretation

As discussed in depth in Sections 1.6.3 and 2.7, Self-Organising Maps (SOMs) are an unsupervised machine-learning technique (originally developed by Kohonen [1982]) which work by clustering data with similar shapes close to each other. This is based on minimising the Euclidean distance between data arrays at each point in a constantly updating 'map'. In order to complete the SOM analysis, phase arrays for each sector were used to train individual maps using methods adapted from the TransitSOM code of Armstrong et al. [2017]. In this work a 2-dimensional SOM was chosen for ease of visualisation, but the SOM could conceivably be used in as many (or as few) dimensions as desired. A deeper discussion of the specific algorithm upon which the SOM training is based can be found in Section 2.7.

The are two primary ways to train the SOM, both of which have different benefits to understanding the variability in the young star population. These methods are here defined as 'Self-trained', where the SOM is trained directly on the young star light-curves analysed in this work, and 'Pre-trained', where light-curves from known variability groups are used to train the SOM. In the latter case the young star light-curves are then mapped onto the trained SOM in order to see where they fall in regards to the known classes of variability. The former training method is used for the majority of this analysis, as it is useful to explore the intrinsic variability in the young star light-curves specifically. However, the 'pre-trained' SOM may be useful for future machine-learning classification of these light-curves, so is also explored here.

Self-training the SOM is the easiest option for training the SOM, and is particularly useful as it does not require any prior knowledge about the expected shape of the light-curves, or indeed any extra data-set. In this method the SOM is trained by inputting the young star SOM arrays prepared in Section 5.2.3 directly into the SOM algorithm. The algorithm then moves these light-curves into similar regions when they share similar shapes, as explained in Section 5.2. The final Kohonen map in this instance is thus trained on the shapes present only in the young star light-curves, so can be useful to explore the overall variability trends

and any clusters of interesting young star variability. Results from this training method are explored through most of Section 5.5, and in the systematics discussion in Section 5.4. Using this method SOMs were built for each sector individually (aside from Sectors 1-5 which were combined due to a comparative lack of young stars), as well as for the entire first year of *TESS* light-curves (Sectors 1-13) at once.

The alternative approach of pre-training the SOM is most useful for classification/grouping of light-curves into pre-known classes of variability, but can also be used as a valuable way to compare different data-sets on a common and repeatable structure. This method involves training the SOM on a pre-selected set of light-curves (often pre-classified into variability classes) and then mapping all new light-curves onto the same pre-trained Kohonen layer. This is helpful because it breaks the randomness of each newly trained SOM, and means that the Kohonen shape at each pixel of the SOM will remain unchanged, even when new light-curves are mapped onto it. Such behaviour makes it very useful for comparing data-sets (e.g. different *TESS* sectors), as each region of the final SOM map will correspond to the same shape, so only the number of light-curves at each spot will vary between runs. Note that Kohonen [1982] discussed an alternative approach to compare data-sets whereby a known 'seed' is used in the SOM initialisation to ensure that the distribution of shapes on the SOM is repeatable between SOM runs, but this was not attempted in this work.

In this case light-curves from the pre-classified K2 Variability Catalog [Armstrong et al., 2016] were chosen to pre-train the SOM, as these would allow specific SOM regions to correspond to known variability classes after training. Stars in the K2 Variability Catalog were classified using a combination of a Self-Organising Map and a Random Forest classifier, making them well-suited to use in further SOM analysis. In total the catalog consists of 68910 classified light-curves, including 154 ab-type RR (RRab) Lyraes, 377 δ Scuti stars (DSCUT), 133 γ Doradus variables (GDOR), 183 Algol-type eclipsing binaries (EA), 290 contact eclipsing binaries (EB) and 9399 other periodic variables (OTHERPER) of high-probability (Prob>0.7) in campaigns 0-4 of the K2 mission. Ideally *TESS* light-curves would be used for each of these stars to maximise similarity to the young star data-set, however at the time of writing only 8903 of these stars had been observed by *TESS*, as the K2 mission took place primarily on the ecliptic plane, where *TESS* observations are still ongoing. Instead the Warwick detrended K2 light-curves from Armstrong et al. [2016] were chosen as the primary data-source, allowing the primary rotation periods derived in this source to be used directly. SOM arrays were then built in the same manner as outlined in Section 5.2.3, using the periods presented in Table

5 of Armstrong et al. [2016]. Two separate approaches were then used to train the SOM: firstly the SOM was trained using all light-curves except those classified as 'Noise' (24,947 light-curves) and secondly using all 68,910 light-curves at once. The resulting SOMs are shown in Figures 5.23 and 5.24 respectively and are discussed in detail in Section 5.5.4.

However, building these SOMs is only half of the challenge, as it is the interpretation of the final SOM (e.g. Figure 5.2) that makes this technique so powerful. As discussed in Section 5.2, in this work the main data-input is young star light-curves, folded by their primary variability period. After the SOM has been trained, each Kohonen 'pixel' on the map thus corresponds to the average shape of the phase-curves which settled on the same pixel during training, which should all be very similar shapes. After training either the same or new phase-curves can be applied to the map, with all phase-curves ending up by the shape which is most similar to their own. Plotting individual pixels from the map thus shows the approximate shape of all phase-curves at that point. For deeper analysis, individual phase-curves and light-curves assigned to these points can also be queried.

In all figures in this work, the final SOMs have all been plotted with colour and size of the plotted pixel corresponding to the number of phase-curves which ended up at that point, with a colour-bar shown to the right of the plot giving explicit numbers to these colours. Those pixels where the most light-curves ended up thus appear brightest and largest. Individual pixels are referred to using the format [x, y], which also correspond to x and y in equation 2.3. For example, the most common pixel in Figure 5.2 is pixel [x,y] = [16,11], appearing bright yellow and large on the map. Furthermore, in larger regions, most popular areas can also be recognised by the densest clusters, as can be seen around [30,27] in Figure 5.6.

5.4 Identification and removal of systematics

One of the key benefits of the unsupervised clustering method used by the SOM is the easy identification of leftover systematics in the *TESS* light-curves. Indeed Geach [2012] refer to the SOM as a form of "non-linear principal component analysis" due to its ability to find patterns and common features in data. The intuitive and visual results interface offered by the SOM makes understanding these systematics simple and efficient. To explore this aspect of SOM performance, SOMs generated for the raw flux ('IFL2') and PCA ('PCA2') CDIPS light-curves for Sectors 1-5 (shown in Figures 5.2 and 5.3 respectively) are compared here.

While the SOM based on PCA flux (Figure 5.3) exhibited a relatively even

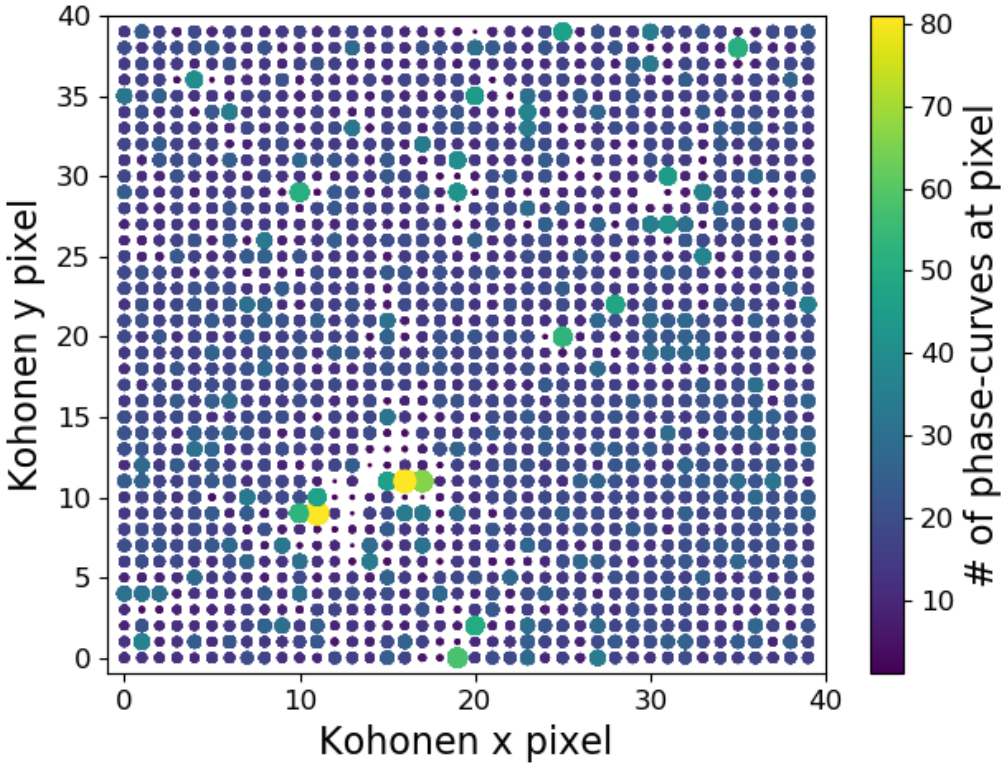

Figure 5.2: Original S1-5 SOM, constructed from CDIPS 'IFL2' flux light-curves. The Kohonen x and y pixels correspond to x and y in Equation 2.3, with individual pixels referred to in the form [x, y]. Both colour (low = purple; high = yellow) and pixel radius are a function of the number of phase-curves situated at that pixel in the final SOM map. The larger, brighter pixels thus represent the pixels which the greatest number of phase-arrays ended up at, corresponding to the most common shapes in the input light-curve data.

distribution of phase-curves aside from some 'hot-spots' of large numbers of eclipsing-binary-like light-curves near pixels [3,5] and [5,8] (as is common for all sectors in the final SOM - see main discussion in Section 5.5), the SOM based on 'IFL2' fluxes (Figure 5.2) extracted straight from the difference images exhibited considerably more hot-spots, spread over much of the 40x40 grid. The most common pixels for phase curves to end up at in the 'IFL2' SOM were pixels [11,9] and [16,11]. When these pixels were queried in the Kohonen layer, all of the phase curves clustered at these points were found to be from Sector 4, with a very distinctive SOM array shown in Figure 5.4. At first glance this phase-array appears to represent interesting outburst-like activity, however investigating this feature further by checking the raw light-curves upon which the phase-curves were based (e.g. for object *Gaia* DR2

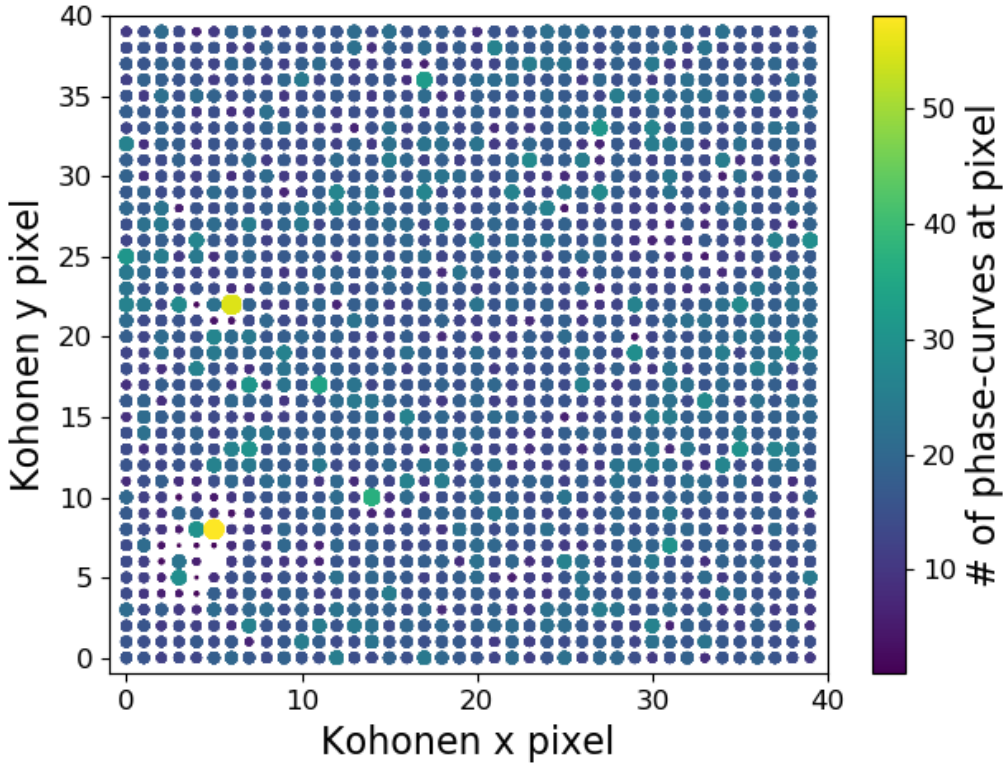

Figure 5.3: Final S1-5 SOM using PCA flux. Colours and pixel size share same meaning as in Figure 5.2. In this case a more even distribution is present than in Figure 1, but clear hot-spots can be seen at pixels [x,y] = [3,5] and [5,8], representing Algol and Beta-Lyrae type eclipsing binary signals.

5275841004963097728 shown in Figure 5.5) revealed that the cause of this unusual shape was simply a common flux-jump present in the first part of many dimmer Sector 4 light-curves. Variations in the phase of this flux-jump in the final phase-curves accounted for a large number of the other 'hot-pixels', especially in the regions immediately adjacent to [11,9] and [16,11]. Closer inspection of the wider IFL2-based SOM revealed that other individual hot-pixels could be explained by additional systematics, such as flux ramps (e.g. [10,29], [20,2] and [25,39]) and sudden flux dips, mimicking Algol-type eclipsing binaries (e.g. [31,30] and [31,27]).

While the significance of the leftover systematics in the raw IFL2 flux light-curves varies from sector to sector,[5] the presence of such effects even in a single sector make it concerning for variability analyses, as telling the difference between true variability and leftover systematics shared with adjacent stars is made considerably

[5]Similar to in Chapter 4, Sector 4 was found to most commonly include such systematics, with Sectors 1 and 2 relatively clear.

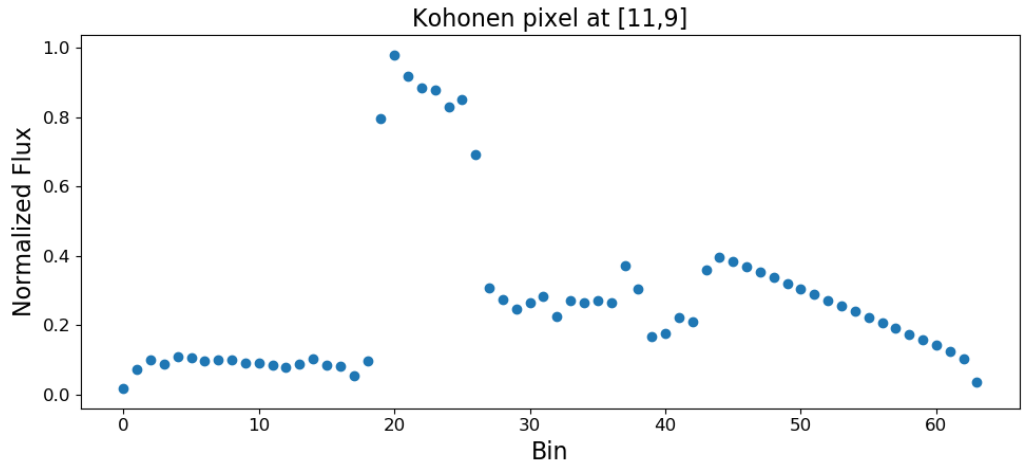

Figure 5.4: The most common Kohonen pixel [11,9] in the CDIPS S1-5 'IFL2' flux run, showing clear flux-jump around bin 20 in the phase-curve. Note that 'Bin' here is synonymous with phase, with bin 0 at phase = 0 and bin 63 at phase = 1.

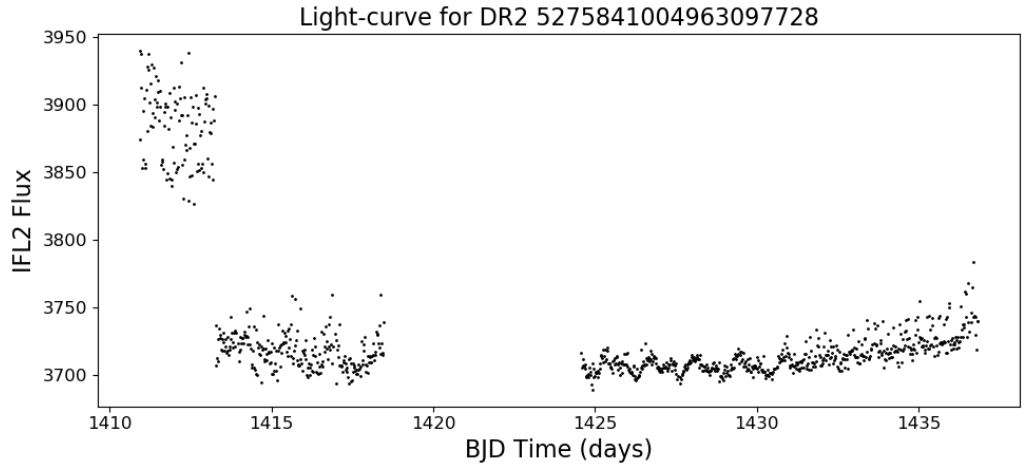

Figure 5.5: IFL2 light-curve for young star Gaia DR2 5275841004963097728, representative of all those light-curves with SOM arrays ending up at pixel [11,9] in the IFL2 SOM for CDIPS S1-5. The clear systematic seen in Figure 5.4 is explained by a flux jump present in the first part of the first orbit, a common systematic in Sector 4.

more difficult. This suggested that some ensemble post-processing of the light-curves was wise, such as that performed for the 'lightly-detrended' PCA light-curves. Indeed, when the PCA light-curves were used instead of the IFL2 fluxes, the final result (Figure 5.3) was no longer systematics-dominated, with the hottest pixels corresponding to eclipsing binaries [5,8] and 'noisy' sinusoids ([6,22]; sinusoids with significant scatter), alongside the typical evolution of activity discussed in Sector 5.5. This suggests that these light-curves are far more suited to analyses of overall variability compared to the raw 'IFL2' ones, backing up their choice as the primary light-curve in the full year SOM analysis.

Another challenging systematic known in *TESS* data is scattered light, which is particularly prevalent in early Year II of the *TESS* data. Because CDIPS light-curves were only available for Year I at the time of writing, this systematic effect was explored using QLP light-curves instead [Huang et al., 2020a]. In this case young star targets were chosen from the catalogues upon which the CDIPS target list is based, by performing a cross-match between them and the QLP light-curves. This resulted in 24,495 light-curves at the time of writing. Light-curves were extracted from the QLP light-curves using the 'SAP_FLUX' keyword, before being cleaned by removing any data points with QLP 'QUALITY' values >4095 from the light-curve. Moments of known poor telescope pointing (as detailed in the *TESS* data release notes[6]) were also removed where these were not caught by the quality cut. This cleaning method resulted in removing similar sections of the data removed in the initial steps of the CDIPS pipeline.

A SOM was then built using all 24,495 light-curves using the same setup outlined in Section 5.2, resulting in the final SOM map shown in Figure 5.6. What is immediately obvious is the unusually-shaped clusters around [16,12] and [30,27]. These structures display a much wider spatial extent on the final map compared to those clusters in the raw IFL2 CDIPS SOM shown in Figure 5.2, appearing as large collections of light-curves spread over many pixels, instead of single 'hot-pixels' at defined points on the SOM. This implies that they are variations on a shape theme, rather than exactly the same shape each time, hence spread out further on the SOM. Closer inspection of these regions reveal SOM pixels with shapes like that presented in Figure 5.7, featuring a steep climb and a slow descent in a triangular pattern. Querying typical light-curves at this pixel makes it clear that this is caused by steep flux-ramps at the end of each orbit. According to the *TESS* data release notes for S17, this was typically caused by scattered light from the Earth and Moon glinting across the detector at these epochs, and can be clearly seen in the background flux

[6]https://archive.stsci.edu/tess/tess_drn.html

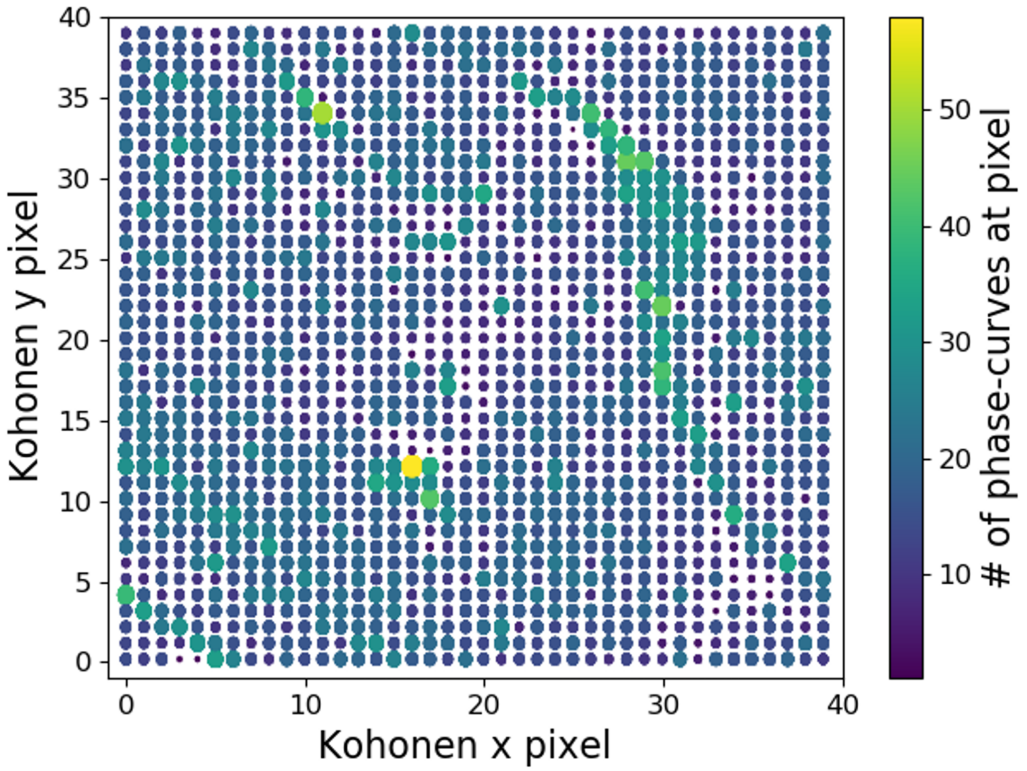

Figure 5.6: Original S17 QLP SOM (after cleaning with the quality flag only) showing scattered light systematics, particularly in the vicinity of [16,12] and [30,27].

(e.g. panel 2 of Figure 5.10).

Removing the effect of this scattered light is complicated, as it varies in significance depending on a wide range of factors, including position on the ccd/detectors, brightness and crowding on the source star and the angle of the space-craft. These scattered-light systematics were still retained (though slightly reduced in severity) when the most affected detector (Camera 3, Detector 1) was removed from the analysis. Background-based and initial ensemble detrending methods based on nearby stars were also trialled, but found to have variable success. The reason for this is best explained by Figure 5.9, showing the raw light-curves for TIC 64940788 and the 9 closest sources with QLP light-curves from *TESS*. It is immediately clear that removing an average of the activity will not entirely remove the scattered light signal of the original light-curve, and may actually distort it further in some cases. This is a product of the scattered light effect evolving in a complex manner, affected by such factors as the detector position and magnitude of each source. A simple epoch-cut was also found to be challenging, as the epoch where scattered light sys-

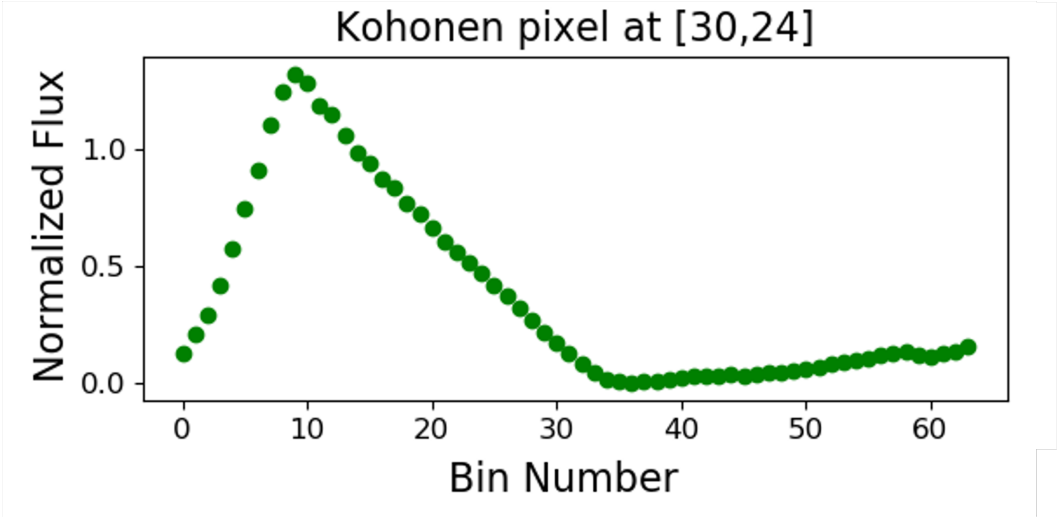

Figure 5.7: Typical scattered light pixel in QLP quality clean light-curves.

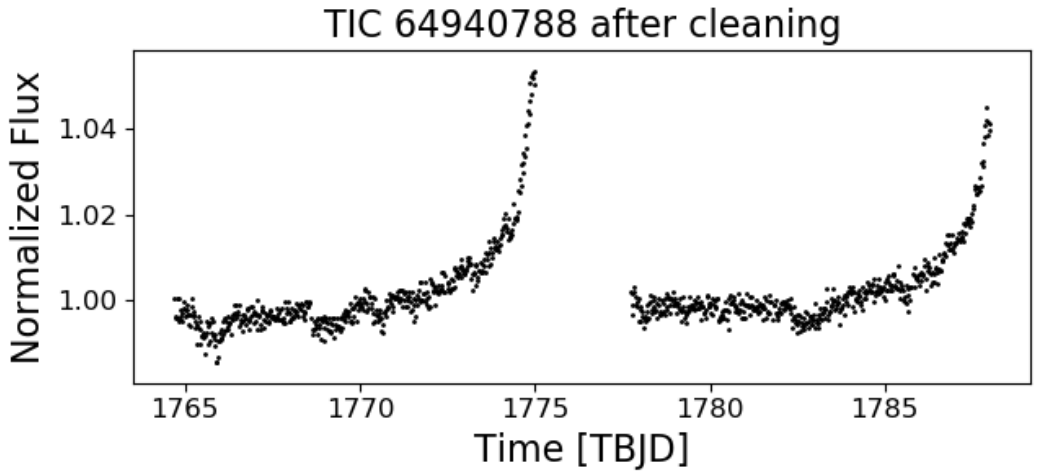

Figure 5.8: Example of scattered light systematics leftover in 'cleaned' QLP light-curve.

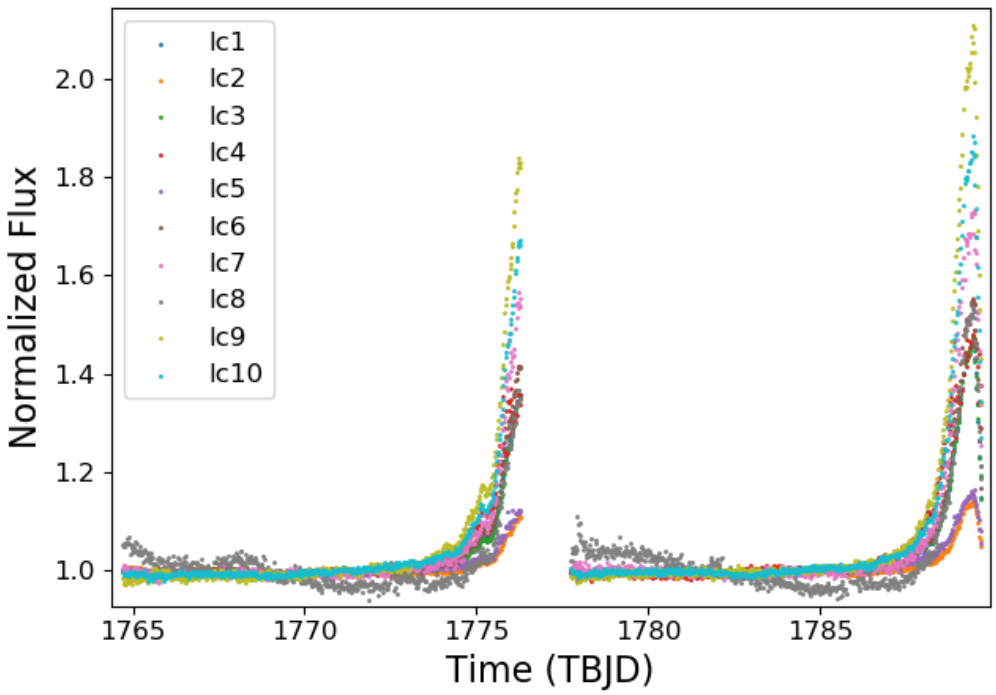

Figure 5.9: Scattered Light variation on 10 adjacent QLP light-curves, illustrating the challenge of removing this systematic effect through simple co-trending techniques.

tematics began to dominate varied between source to source and, in some brighter cases, never to occur at all.

Eventually however, a hybrid approach was found to have some success by considering the background error data row 'SAP_BKG_ERR'. Through comparison of the SAP flux ('SAP_FLUX'), SAP background flux ('SAP_BKG') and SAP background error ('SAP_BKG_ERR'), the SAP flux was found to become scattered-light dominated beyond the point in each orbit when the background error exceeded twice the median background err. This relationship is demonstrated visually in Figure 5.10. Cutting the light-curve at this point was found to remove the majority of the scattered light systematic whilst minimising loss of useful data at the same epochs in objects less affected by scattered light. After cleaning all light-curves based on this background error cut approach, a new SOM was built for the S17 QLP light-curves, which is shown in Figure 5.11. This SOM no longer displays the clear systematic scattered light effects seen in the original S17 QLP SOM (Figure 5.6) and instead results in a SOM similar to the final CDIPS SOM for Sectors

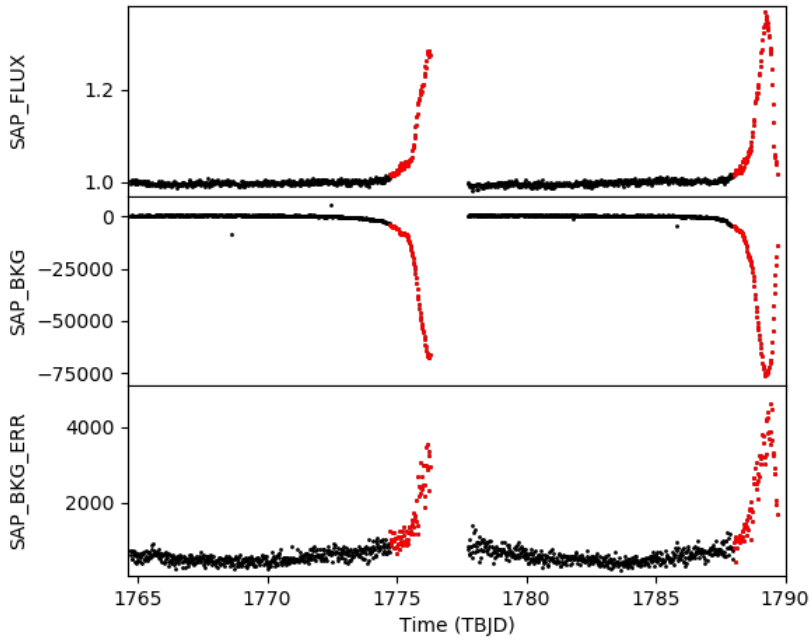

Figure 5.10: Example of the background error cut being applied. Flux data is cut in regions where the background error exceeded two times the median background error, as highlighted in red.

1-13 (Figure 5.12), with steadily evolving stellar variability over the whole 40x40 SOM and hot-spots of eclipsing-binary-like activity around [8,4] and [12,1]. The use of such a cleaning technique thus extends the potential of SOM analysis to the prevalent QLP light-curves as well.

Hence by clustering similar shapes of phase-curve close together, the SOM effectively gives a visual overview of any common systematics in the data-set which may otherwise be missed. It also gives the user a useful set of trial light-curves to test techniques of removing these troublesome systematics, without having to check each by eye. Furthermore, the clustering gives an overview of which systematics appear to be related in shape, and can more easily allow the user to view what is being removed in any related principal component analysis. Identifying and removing these systematics one by one can significantly aid searches for planetary transits, whether the stars are young or otherwise. The SOM's ability to identify these automatically and visually thus make it attractive as an additional tool for narrowing down troublesome systematics. This analysis also illustrates the necessity of carefully checking individual light-curves when applying wider machine-learning techniques to the *TESS* data, as some phase arrays can be suggestive of real activity

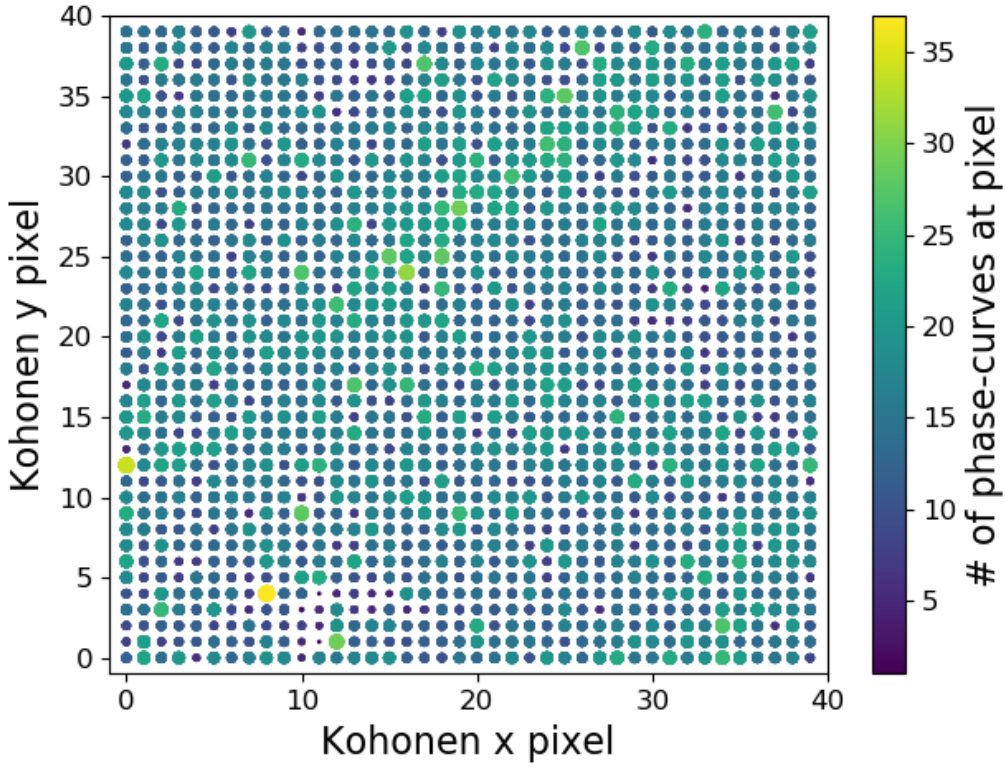

Figure 5.11: S17 QLP SOM after applying background error clean. This displays a much cleaner distribution across the SOM map and traditional EA/EB hot-pixels, compared to the systematics-dominated original S17 SOM in Figure 5.6.

when the extra information given in the original light-curves is simplified.

5.5 Results

5.5.1 Overall results

The overall SOM built and trained on all young star CDIPS light-curves from the first year of *TESS* data is shown in Figure 5.12, with an overview of the main features presented in Table 5.1. The raw result from the SOM is somewhat subtle, as the most important information is hidden in the pixels themselves (a sample of which are shown alongside). However, there are still some important features to note in the overall SOM shape. To reiterate, pixels have been colour-coded and given radii in the overall SOM plot based on the number of light-curves (or their respective phase arrays) which ended up at that pixel after training on the CDIPS young star data-set. This aids the eye to see clusters of larger numbers of light-

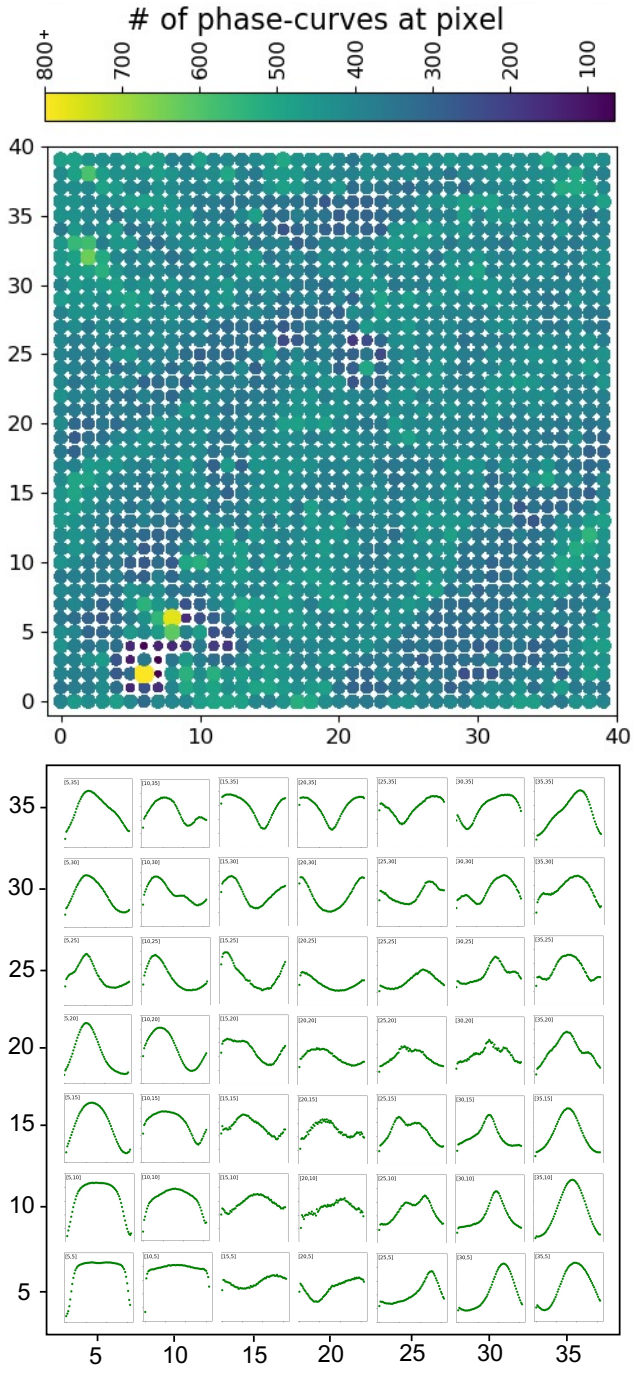

Figure 5.12: S1-13 Final CDIPS SOM. *Top*: Final SOM for all CDIPS light-curves. The number of individual light-curve arrays at each point is represented both by colour and by pixel radius. The most common pixels are lighter in colour and in regions of the map with less white space. *Bottom*: Plots of individual SOM pixels from the trained SOM, at 5-pixel intervals over the SOM, to show the approximate evolution of light-curve phase arrays over the SOM map.

Table 5.1: Overview Table of full S1-13 Self-Organising Map

Variability Type	Main Position
Algol type eclipsing binaries (EAs)	[8,6]
Beta Lyrae type eclipsing binaries (EBs)	[6,2]
Flares	[22,24]
Dipper-like behaviour	Regions around [10,30], [25,16], [35,1] and [35,29]
Main sinusoidal modulation	Distinct diagonal band North-East from [34,7] to [13,24]
Asymmetric sinusoids	Region around [2,32]
Other periodic signals/shapes	Large region around [30,30]
Flat/Noise-dominated	Large region around [18,10]

curves in lighter, denser regions. It is also important to remember that the SOM boundaries are cyclic, such that pixel [0,0] is diagonally adjacent to [39,39]. This can clearly be observed in the continued 'ridge' from [0,13] to [39,13] and in the specific phase-curves at the edges of Figure 5.12.

The most striking features in this SOM are the hot spots at [6,2] and [8,6], represented as the yellow spots on the SOM, alongside the slower evolution of light-curve shape across the remainder of the map. The SOM pixels for these two 'hot-pixels' at [6,2] and [8,6] are plotted in Figures 5.13 and 5.14 respectively, and appear to represent two very well-known classes of photometric variability: Beta Lyrae type contact eclipsing binaries (EB; [6,2]) and Algol-type detached eclipsing binaries (EA; [8,6]). Querying specific light-curves at these pixels confirms this classification, with clear examples of both types of binary evident. This suggests that this clustering method is very effective at finding eclipsing binaries, and further gives an efficient and intuitive view of those present in the data-set. A closer examination of the EA-like signals is also useful given that transiting planets could give very similar eclipse shapes, especially for those pixels where secondary eclipses are not obvious in the phase-folded light-curves.

The regions immediately around these 'hot-pixels' are also interesting to examine, representing shapes similar to the most distinct EA and EB eclipsing binaries. The region surrounding pixel [6,2] is comparatively sparse, suggesting that it is a very distinctive shape, with few variations from the shape present. On the other hand, there are a number of 'warm pixels' to the left of [8,6], which when queried show evidence of a second eclipse, for which a hint can be seen on the bottom of Figure 5.12, at pixel [5,5]. This comparison suggests that there is considerably more shape variation in EA light-curves viewed in these young stars compared to the variation observed in EB light-curves, much of which may be down to the period of the detached EAs. To the right of both of these hot pixels is a steep drop-off, which makes sense given the distinct change in shape to noisy sinusoids which occurs towards pixel [15,5] (also pictured in the bottom panel of Figure 5.12).

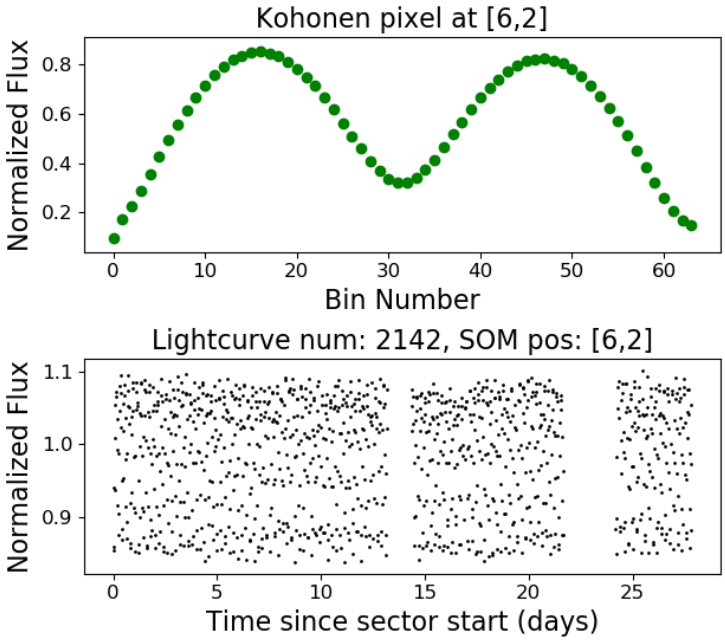

Figure 5.13: Phase curve and example light-curve for Kohonen pixel [6,2] in the final S1-13 SOM, representing the position of most EB signals.

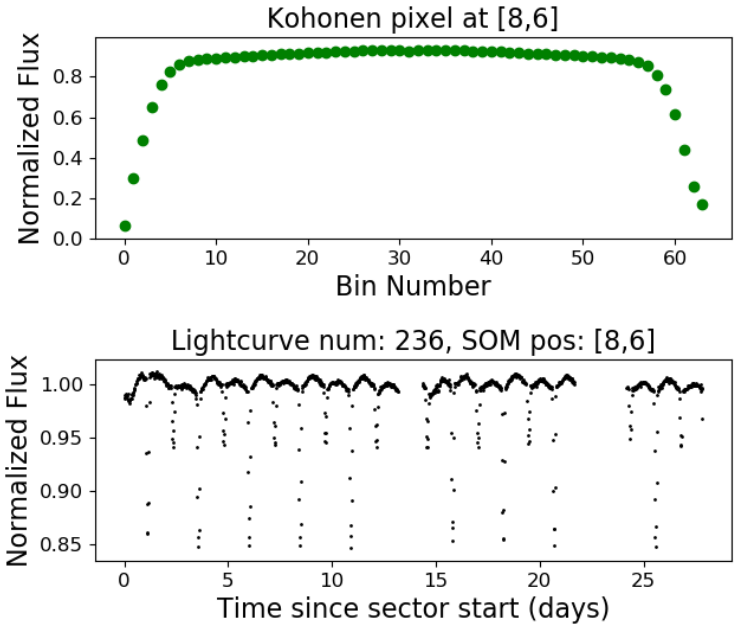

Figure 5.14: Phase curve and example light-curve for Kohonen pixel [8,6] in the final S1-13 SOM, representing the position of most EA signals.

A first glance at the full SOM distribution through individual pixels reveals the rich array of variability present in young star light-curves. While sinusoidal pixels are certainly common (as is expected given the common spot modulation in young stars), it is not hard to find alternative pixels where simple sinusoidal detrending would be completely inappropriate (e.g. [0,30], [10,30] and [30,20]). It is also obvious that in some regions of the SOM the shape trend varies slowly (e.g. the diagonal from [0,10] to [20,30]), but in others the changes in shape are rapid (e.g. the swift change between pixels [10,30] and [15,35]). When analysing the wider SOM distribution further, two key things to look out for are light clusters of the most common light-curves (e.g. [2,32] and [15,5]) and dark regions (e.g. [21,26] and adjacent to [6,2]) separating quite distinct shapes of variability. As discussed previously, it is these individual clusters and not the overall SOM distribution that is of greatest importance, especially as the specific locations of clusters will change with every run without pre-training. The most common phase-curve shapes aside from the EA/EBs discussed previously appear to be variations on the asymmetric sinusoids at [2,32] (similar to [5,30] in Figure 5.12) and negative sines/cosines (e.g. the distinct diagonal cluster from [34,7] to [13,24] in Figure 5.12). However one particularly unusual one is [22,24], falling in the middle of another less-dense region, like the EA and EB pixels. Querying the individual phase-curves and corresponding light-curves at this point (e.g. Figure 5.15) curiously reveals most to contain flares, suggesting that this extended peak in Figure 5.15 is caused by averaging the different phases of maximum flare brightness.

One variability shape that proved harder to find than originally expected was the Dipper/burster class. As these light-curve shapes are thought to be caused by occultations by wraps of accretion columns in circumstellar disks [Hedges et al., 2018], they could theoretically take quite unique and unusual shapes, meaning that they may not all fall on one defined region of the final map. However, through examination of individual light-curves at pixels with unusual Kohonen phase arrays, some dipper and burster-like light-curves were found in the vicinity of [10,30], [25,16], [35,1] and [35,29]. Lists and specific examples of these will be discussed in YOUNGSTER III (Battley et al, in prep) after additional supervised machine-learning methods are combined with the SOM to more automatically weed out the large number of noisy light-curves also present at these pixels. The co-location of the noisy and dipper-like light-curves at these pixels is an unfortunate side-effect of swiftly-evolving dippers and some heavily noise-dominated light-curves appearing similar when condensed down to 64 flux bins. Considering non-periodic statistics alongside the SOM shapes in the future should help to separate these signals much

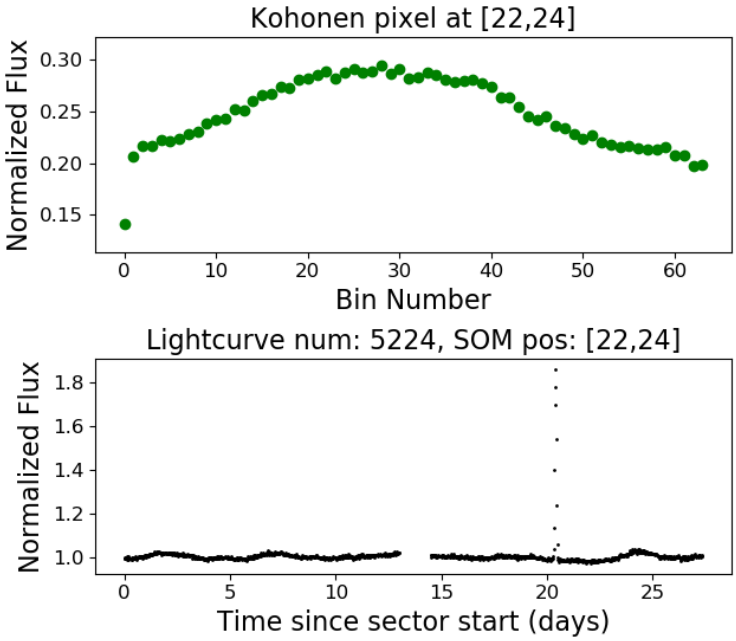

Figure 5.15: Phase curve and example light-curve for Kohonen pixel [22,24] in the final S1-13 SOM, another site of unusually distinct variability.

more effectively.

The remainder of the final map is mostly split into two main large regions, centred on [30,30] and [18,10] respectively. The region around [30, 30] typically contains phase-/light-curves with other periodic variability (similar in vein to the 'OTHERPER' class defined by Armstrong et al. [2016]), appearing as repeated signals with shapes unlike typical sinusoids. Such profiles may be caused by such processes as stellar oscillations or rotating star spots. Some of these also harbour significant evolution even along the duration of a single *TESS* sector. Meanwhile the region around [18,10] consists primarily of flat or noise-dominated light-curves, with no obvious variability. This would be the bulk of a SOM completed for an older population, yet is comparatively rare for such young active light-curves.

A final interesting feature of some of the phase-curves pictured in Figure 5.12 is that there is some variation in the range of flux, from 0 to 1 for the strongest signals (e.g. [8,6]), down to as low as 0.15 to 0.3 for some of the noisier light-curves (see [22,24], Figure 5.15). This is somewhat surprising given that all phase-arrays were normalised between zero and one in their construction. Such a discrepancy is explained through examining individual phase-curves at this pixel, showing that many of the light-curves grouped at these pixels are less similar in shape to each other, hence average out to smaller ranges.

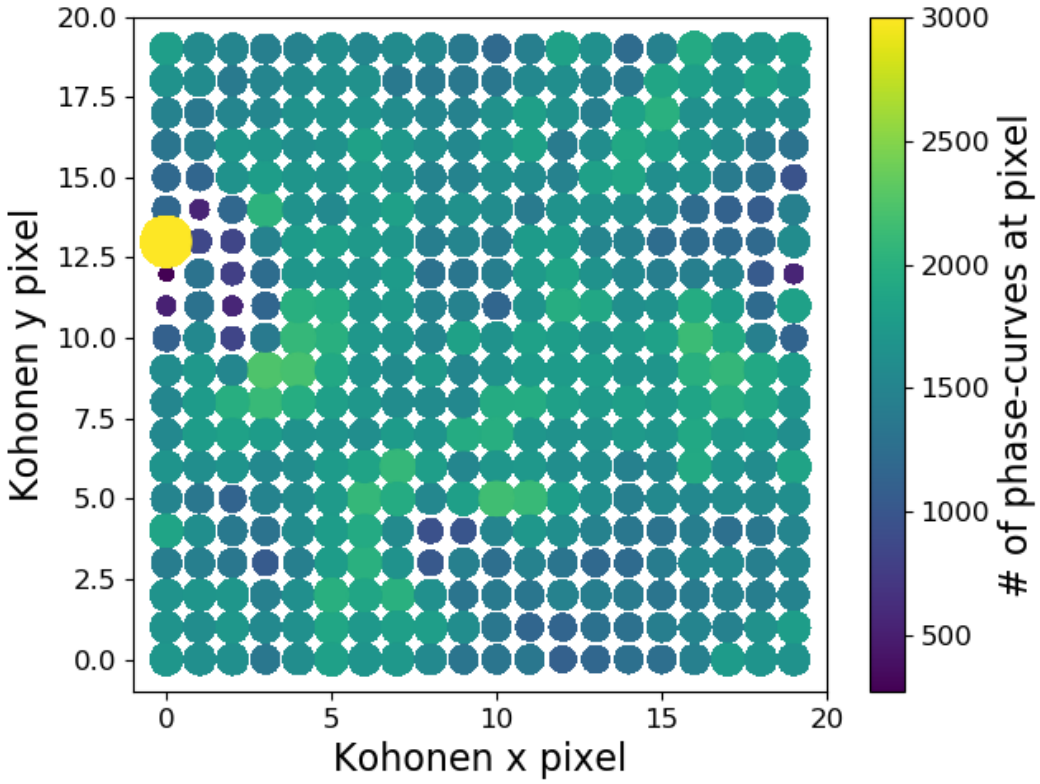

Figure 5.16: SOM built using all young star CDIPS light-curves from S1-13 but with only 20 pixels in each direction. Colours and pixel size share the same meaning as in Figure 5.2. Here the large-scale variation over the full map is clearer, but at the expense of merging some similar pixels together. In particular the EA and EB pixels have now been merged into one region by pixel [0,13].

5.5.2 Variation in SOM size

Two additional sizes of SOM were trialled to yield more information from the SOM: 20x20 and 60x60 pixels, both of which highlight different aspects of the young star data.

The 20x20 pixel SOM (Figure 5.16) provides an easier way to see overall trends in the data-set, including the main evolution from sinusoids (e.g. [3,9]) to more complex shapes (e.g. [10,0]) and the 'hot-pixels' at [0,13] and [19,11] for the EAs and EBs respectively. This makes it a useful first overview to the data-set, and can capture any significant systematics present. However, given the much smaller mapping space to spread things out, it is prone to combining slightly different types of variability into a single pixel. Such a combination typically makes the individual SOM pixels less representative of all those light-curves within it, which was seen to

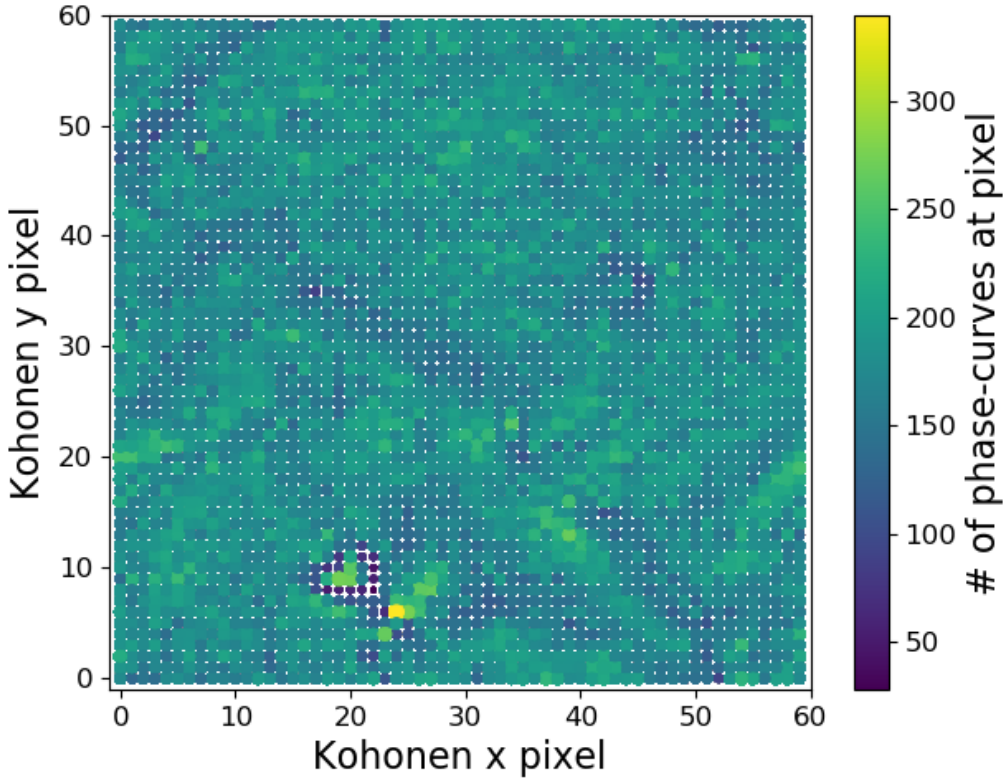

Figure 5.17: SOM built using all young star CDIPS light-curves from S1-13 but with 60 pixels in each direction. Colours and pixel size share the same meaning as in Figure 5.2. In this case the 60x60 SOM is better at highlighting individual small clusters of unusual variability but makes the overall trends in variability harder to recognise.

distort the main EB pixel shape compared to the 40x40 run. This size of SOM is thus most useful for getting an idea of the overall larger variability classes present in the model by eye, but is considered slightly too simple for representing the wide variation in light-curves shapes present in young star light-curves.

On the other hand, the 60x60 SOM (Figure 5.17) exacerbates individual variability types while sacrificing the ease of understanding overall trends by human eyeballing. Examining all 3600 pixels by eye is unwieldy, and unlikely to give much of an understanding of the large-scale variations in the data-set. It is the hot-pixels where this larger SOM really excels however, showing considerably more small-scale minima and maxima than the 40x40 SOM. As in the 40x40 case, the key features to look for are bright pixels surrounded by darker pixels, as these are comprised of arrays very different to the surrounding data. The clearest example of this is

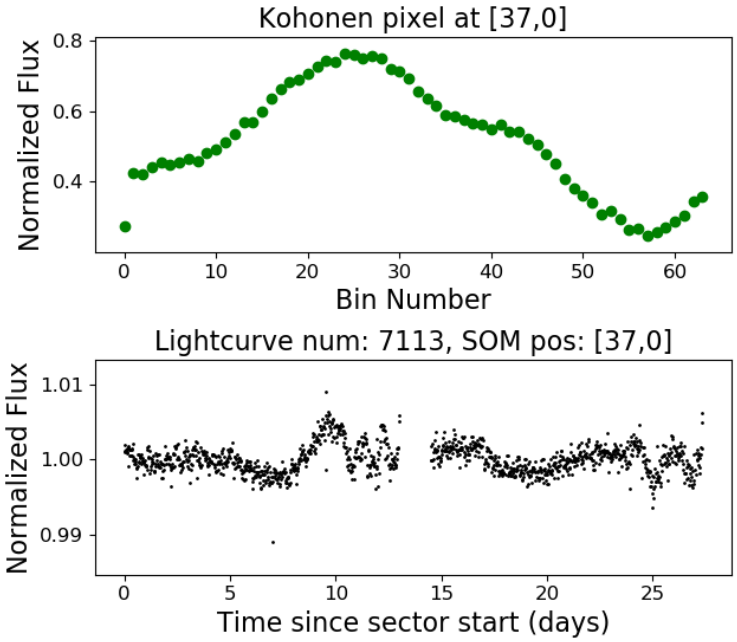

Figure 5.18: Example 1 of interesting Kohonen-pixel shape and corresponding example light-curve from 60x60 SOM run - pixel [37,0]

around the pixel [20,9] in Figure 5.17, which once again represents the distinctive group of Beta-Lyrae type Eclipsing Binaries. Slightly subtler interesting groups can be found at pixels such as [37,0] and [40,30], pictured in Figures 5.18 and 5.19 respectively. While some phase-curves ending up at these pixels are simply due to noise scatter, many others are worth investigating further for behaviour such as dippers and bursters.

5.5.3 Sector by sector analysis

As well as generating an overall SOM for the complete Year 1 CDIPS light-curves, SOMs were generated for each *TESS* Sector individually. Because the constructed phase arrays are of a constant size and normalised between zero and one, direct comparisons between sectors are easy, despite variations in data gaps and periods of extra scatter. Note however that as most young stars are concentrated along the galactic equator, the number of targets per *TESS* sector was found to vary considerably, ranging from as few as 3120 in Sector 3 to 146,557 in Sector 11. Because of this, building a SOM out of Sectors 1-5 individually was not found to be very instructive, so instead these were combined into a single S1-5 run totalling 26,153 targets.

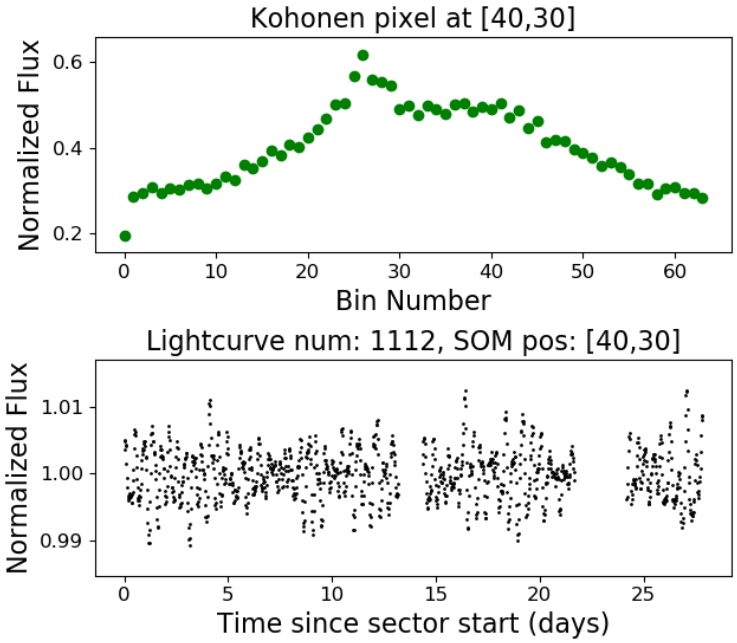

Figure 5.19: Example 2 of interesting Kohonen-pixel shape and corresponding example light-curve in 60x60 SOM run - pixel [40,30]

Alike to the full Year-1 SOM run, the self-training method of building the SOM was used for each sector, with the results for Sectors 1-5 shown in Figure 5.3 and the remaining sectors presented in Figures B.1 to B.8. Although specific positions vary because of the random initialisation of the SOM, the major characteristics of the final maps for each sector are very similar, with hot-pixels for the EA/EB eclipsing binaries surrounded by the lowest density regions, alongside more steady evolution over the rest of the map. Some discrepancies do however exist with regards to additional hot pixels, such as pixels [6,22] in the S1-5 SOM and [7,14] in Sector 9. In the case of pixel [6,22] of the Sector 1-5 SOM, this pixel represents a noisy sinusoid (plotted in Figure 5.20). This is thought to be caused by the extra pointing noise present in the early *TESS* sectors, which were corrected in Sector 6 onward. Because these light-curves represent a reasonably small number in the overall Year 1 data, these light-curves do not form a large enough group to be a significant hot pixel in the full Year 1 run.

Meanwhile, the additional two hot pixels in the Sector 9 plot at [7,14] and [35,10] have the geometry plotted in Figures 5.21 and 5.22 respectively. These are less easy to characterise using the SOM pixels alone, and hence are better understood by checking some of the individual light-curves at these points. Closer inspection of

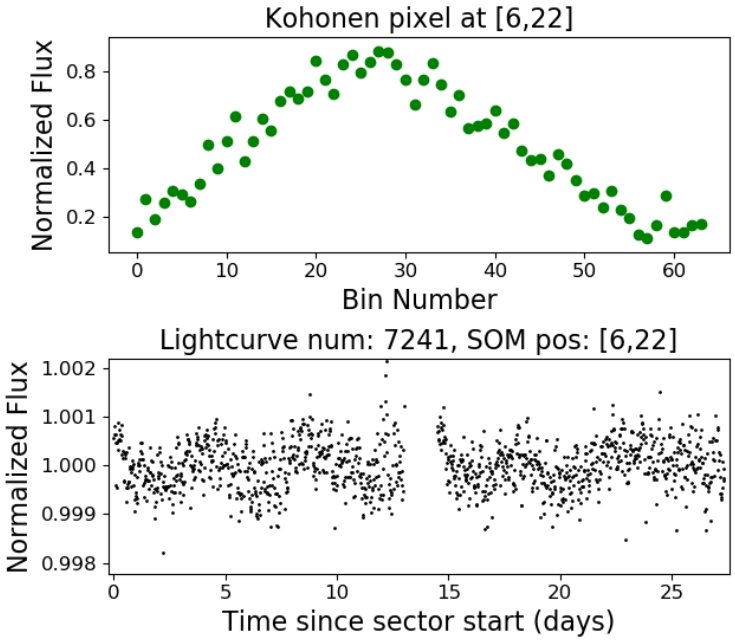

Figure 5.20: Phase-curve and example light-curve for additional 'Hot-Pixel' [6,22] in S1-5 SOM, showing noisy sinusoidal geometry.

these light-curves revealed that the targets at both pixels were dominated by excess noise, coupled with a distinctive background signal consisting of V-shaped dips at the beginning of both orbits. An additional flux ramp before these V-shaped dips in the light-curves at [7,14] changed the overall shape at this pixel further. All of the light-curves grouped to these pixels came from suspiciously similar light-curve numbers (45,000-50,000 for pixel [7,14] and around 1400-4000 for pixel [35,10]), suggesting that they come from similar regions of the CCDs originally. Sector 9 also contains some very crowded regions of sky near the ecliptic plane, which may explain why so many light-curves were adversely affected.

Despite these minor discrepancies, the similarity in SOM clustering and overall trends present between sectors show that the SOM remains effective across all sectors, with the one proviso that enough light-curves are available to allow significant clustering to occur. This bodes well for using SOM positions as a useful metric for further machine learning classification of young stars, as useful results can be obtained even when only running a single month of data.

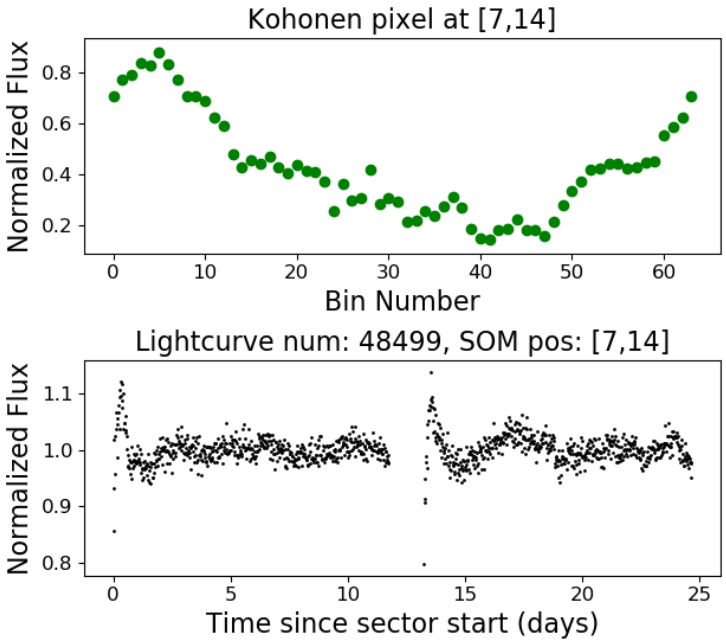

Figure 5.21: Phase-curve and example light-curve for additional 'Hot-Pixel' [7,14] in S9 SOM. Inspection of individual light-curves revealed this to be caused by the distinctive background signal at the start of each orbit in the light-curve.

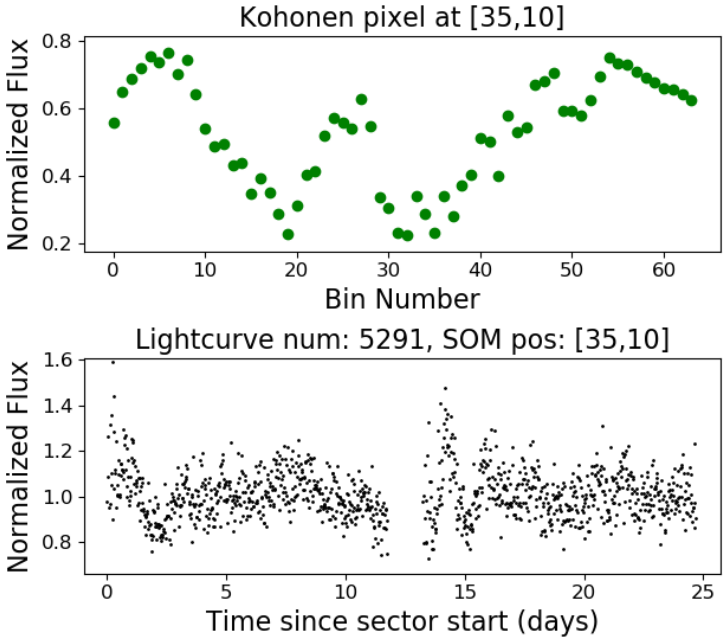

Figure 5.22: Phase-curve and example light-curve for additional 'Hot-Pixel' [35,10] in S9 SOM. Similar to [7,14] this was caused by additional noise and a distinctive background signal.

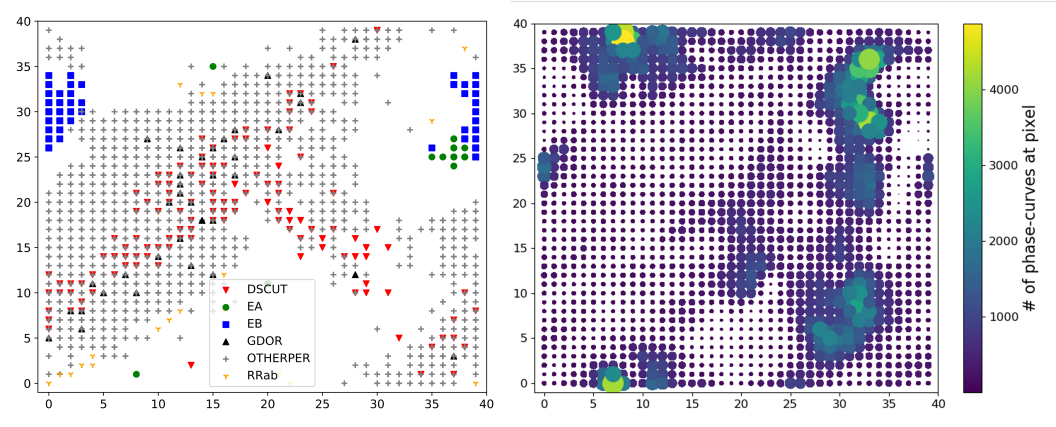

Figure 5.23: SOM trained on the 24,442 light-curves from the K2 Variability Catalogue which had classifications indicating that they were not noise. As usual, axis numbers correspond to Kohonen x and y pixel numbers, similar to Figure 1. Note that only targets with >90% class probability in the catalogue are plotted for clarity. *Left:* The 24,442 non-'noise' light-curves mapped onto the SOM, showing distribution of known variability types. *Right:* All young star CDIPS light-curves from the first year of the *TESS* mission (Sectors 1-13) mapped onto the same K2 Variability Catalogue-trained SOM. The bulk of the CDIPS light-curves can be observed to fall into gaps in the K2 Variability Catalogue map.

5.5.4 Pre-trained SOM

Results from pre-training the SOM on the K2 Variable Catalogue data [Armstrong et al., 2016] are shown in Figures 5.23 and 5.24, excluding and including those light-curves classified as 'noise' respectively. The same trained SOM underlies both the left and right hand side of each figure, with the K2 Variability Catalogue light-curves mapped onto the SOM on the left and the full *TESS* S1-13 young star light-curves mapped onto the SOM on the right. This means that the shape of each pixel in the K2 Variability Catalogue map on the left corresponds to the same shape at the corresponding pixel on the right, allowing direct comparisons between expected variability classes and the position of young star phase-curves.

In the case of the 'no-noise' SOM (Figure 5.23), the plotted variability classes were observed to cover most of the 40 x 40 grid, with a wide diagonal band of primarily other periodic (OTHERPER), delta-scuti (DSCUT) and Gamma-Dor (GDOR) variables. Meanwhile eclipsing binaries of both types (EA/EB) congregated around [39, 30] (remembering that the SOM wraps around each edge), distinctly off-set from the other types of variability. In this case the white space left on the plot is mostly full of targets with class probabilities of <90%, representing those with less

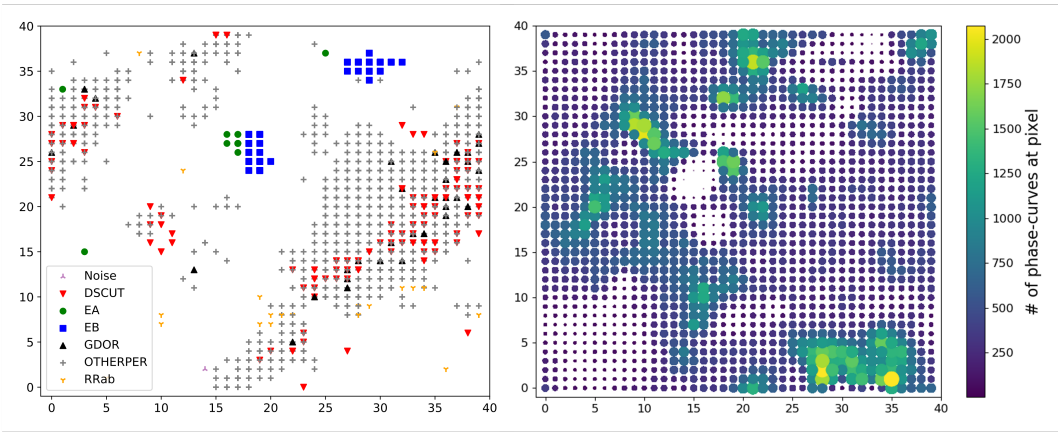

Figure 5.24: Same as Figure 5.23, but considering all 68,910 in the K2 variability Catalogue, including those classified as 'noise'. Axis numbers represent Kohonen x and y pixels, alike to Figure 1. While overall classification performance is improved (see for example the clear EA/EB over-densities at [19,25] and [26,33]), many light-curves still fall in gaps in the K2 Variability Catalogue trained SOM, likely due to excess noise in the *TESS* light-curves and additional types of young variability not present in the K2 Variability Catalog.

clean-cut classifications from the K2 Variability Catalogue. Interestingly, when the *TESS* S1-13 CDIPS light-curves analysed in this work were applied to this same map, the vast majority of them fell on regions of the map with low densities of K2 variability catalogue targets. This suggests most of the shapes seen in the young *TESS* light-curves (likely from alternative types of variability or noise) are different to the specific variability classes defined in the K2 variability catalogue. Such poor classification performance stresses the importance of conducting young-star specific classification rather than simply relying on previously known variability classes. It should be noted however that some of the young star light-curves do still fall on, or nearby, known variability classes on the 'no-noise' SOM, particularly over the centre of the main OTHERPER block of targets and adjacent to the main group of eclipsing binary light-curves. Querying CDIPS young star light-curves which appeared in this EB-adjacent region does reveal them to be eclipsing-binary-like signals in the *TESS* data as well, suggesting that these types of signals are the ones which are best handled by (and most similar to) the K2 Variability Catalogue trained SOM.

Examining the SOM trained on the full K2 Variability Catalogue (Figure 5.24) provides more promising results. Because training of the SOM spreads variability across the entire map, there is considerably more white space in this latter case, where most of the 'Noise' light-curves fell. However, the OTHERPER sources are still spread out over a large proportion of the SOM, suggesting that there is a lot

of variation in their specific light-curves, compared to the much more localised EA and EB light-curves. In this particular run the eclipsing binary signals were split into two main groups rather than a single region, but while there is some difference in the shape of the phase-curves between these groups (with the top right group having wider dips), this is considered more to do with the random initialisation of the SOM than significant differences in shape, as a previous run with the same K2 Variable Catalogue data placed all EAs/EBs in a single joined group. The inclusion of the 'noise' class appears to aid the SOM performance considerably, both by grouping light-curves similar to known variability classes more obviously and by allowing internal structure and specific groups of the light-curves which fall onto quieter regions of the plot to be viewed. Of particular interest is the main EA/EB group at [19,25] on the K2 Variability Catalogue map, which is clearly reproduced as an area of high density in the young star map on the right. As is expected, querying individual phase- and light-curves ending up at this point reveals true EA/EB-like signals, showing that the SOM is working effectively at finding these. A similar region of high density can be seen at [26,33] in the young star map on the right, just slightly below and to the left of the expected position ([28,35]) in the K2 Variability Catalogue map.

Meanwhile there are a handful of clear over-densities in the quieter regions of the original K2 Variability Catalogue map, specifically around [5,20], [10,28], [16,10], [21,36] and [35,1]. While the shape of the average phase curve at each of these pixels varies somewhat, all of these phase-curves have y-ranges of about 0.3-0.7 rather than 0-1 as they were initialised with, suggesting that they contain more variation between each phase curve at these pixels. This is common where initial light-curves are noisy, causing large variation between each corresponding phase-curve. Indeed querying individual *TESS* phase- and light-curves which congregated at these pixels (e.g. see Figure 5.25) reveals them to be noise-dominated, significantly more noisy than is usually seen in light-curves extracted from the *K2* mission. This was not entirely unexpected given *TESS*'s reduced photometric precision compared to the *Kepler* spacecraft, as indeed similarly large noise differences were also seen by Battley et al. [2021] when comparing a number of *Kepler* planets reobserved by *TESS* (see also section 6.2.3). Such a problem is particularly challenging for dimmer *TESS* targets, where the instrumental white noise far outweighs any potential variability signals. In order to explore this aspect further, a brighter subset of the CDIPS targets were selected by choosing those stars brighter than Tmag = 12 in the TESS Input Catalogue [TIC v8; Stassun et al., 2019]. Applying the SOM trained on the full K2 Variability Catalogue (including the 'Noise' class) to the phase-curves for these

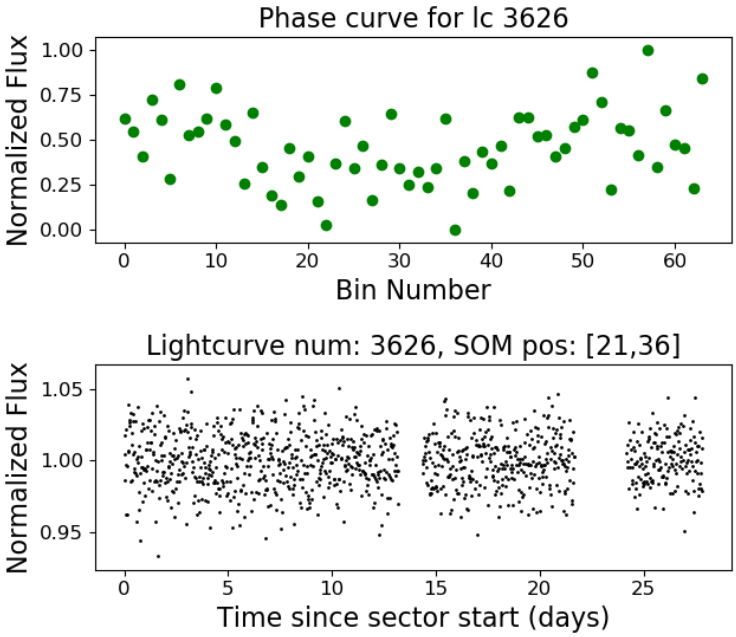

Figure 5.25: Example of one of the noise-dominated *TESS* phase-curves (top) and light-curves (bottom) placed into gaps in the SOM pre-trained on the full K2 Variability Catalogue.

targets resulted in the final map pictured in Figure 5.26. Through comparison to Figure 5.24, it is clear that these brighter TESS targets align much better with the previously known variability groups from the K2 Variability Catalogue (plotted on the left of Figure 5.24) than the full CDIPS light-curves did, with the bulk of the targets falling on the main band of OTHERPER/DSCUT variables or EA/EB clusters. Furthermore, the distinct over-densities observed around [5,20], [10,28], [16,10], [21,36] and [35,1] have all reduced significantly in relative strength compared to the main clusters of known variability type.

Nonetheless, some of the light-curves falling in the gaps cannot be explained by the excess noise in *TESS*. Alongside these phase-curves exist a number of unusual variability types, including dipper-like behaviour and unusual rotation-like signals. The reason for these falling in gaps could perhaps be partially explained by the included young stars having a much wider sample of young ages compared to the often older K2 stars. This would mean that some K2 light-curves may have types of variability not seen in young stars, and some particularly young stars may have types of variability not present in bulk in the K2 Variability Catalogue. Additional differences in shapes could also be explained by the differences in *TESS* sector lengths (roughly 27d), compared to K2 campaign length (∼80d for the analysed

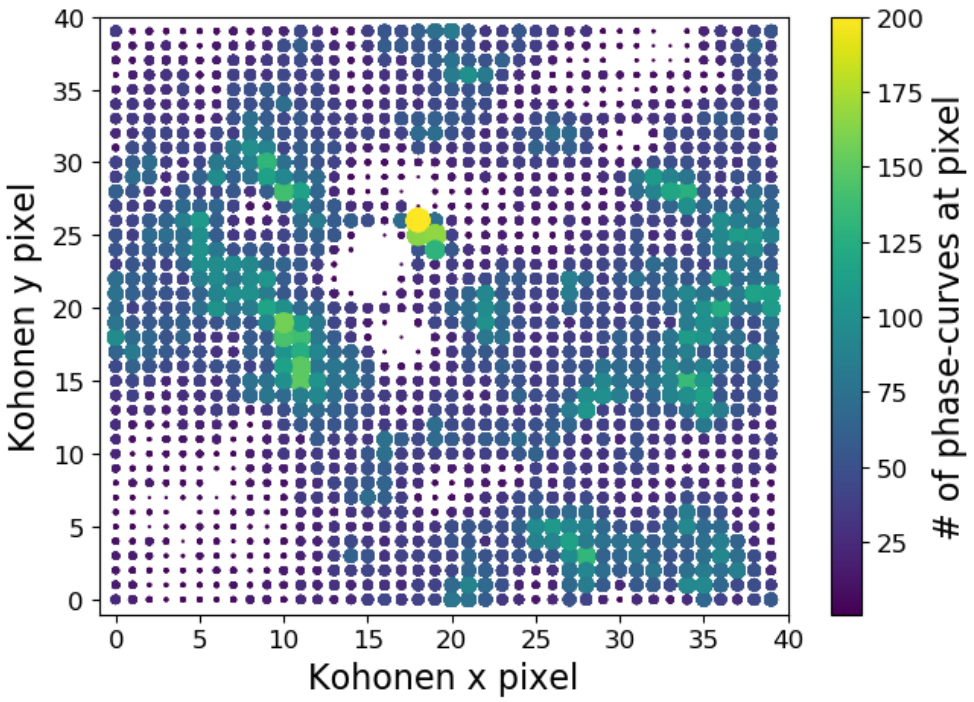

Figure 5.26: Bright (Tmag<12) CDIPS targets from Sectors 1-13 mapped onto the full K2 Variability Catalogue trained SOM shown in Figure 23. Although a few smaller areas of higher density still exist in gaps in the original K2 Variability Catalogue SOM (e.g. on the lower right of the plot, around pixel [30,3]), the majority of the data now falls into regions of previously known variability classes.

0-4th campaigns), meaning that K2 targets may evolve more in phase over the time period. However, such interesting types of variability will be much more easily explored and explained when not overwhelmed by the large number of noisier *TESS* light-curves also constrained to these pixels.

These challenges show that although pre-training the SOM shows some promise in classifying new targets into previously known variability groups, the performance of this classification would benefit significantly from using light-curves from the same survey/instrument compared to archival data. In future analyses in the YOUNGSTER program, *TESS* light-curves will be used instead, based on objects with previously classified variability classes. Unfortunately at the time of writing not enough of these existed to test such an analysis, however with *TESS* now observing the ecliptic plane, many of these pre-classified targets from the K2 Variability Cataloguer should soon receive *TESS* data.

5.5.5 Catalog of Eclipsing binaries in TESS Year 1

As discussed in Section 5.5.1, the SOM method appears to be very effective at highlighting and separating the distinctive shapes of eclipsing binaries from other variability present in the young star light-curves, appearing as distinctive hot-pixels in all the self-trained SOMs generated in this work. As young eclipsing binaries are very important to understanding the early evolution of star systems, those identified in this work are presented here in the hope that they will be useful for the wider young star community. Although the bulk of these light-curves ended up at the hot-pixels at [6,2] and [8,6] in the complete Year 1 SOM (Figure 5.12), there are additional binary-like signals in pixels adjacent to these, especially to the left of the main detached eclipsing binary (EA) pixel at [8,6]. Because of this, a more conservative method of selecting light-curves was used, selecting phase-curves by using a cut based on the Euclidean distance from the archetypal phase-curve at the strongest pixels. The specific cuts used were $d_{xy} < 1.56$ from pixel [8,6] for the EA candidates and $d_{xy} < 1.37$ from pixel [6,2] for the EB candidates. This allowed the most similar light-curves from the surrounding area to be selected whilst reducing the number of false-positives that would be selected by simply choosing all pixels which harbour some eclipsing binary-like signals. These cuts resulted in a total of 13,366 candidate EA signals and 2448 candidate eclipsing binary signals, including many systems which were picked up in multiple sectors. All selected light-curves were then checked by eye in order to catch any remaining obvious false positives. Such false positives typically consisted of excess noise, sharp stellar activity or obvious blends with other nearby targets. Following this step and a uniqueness cut on TIC ID, 8103 EA and 1297 EB candidate systems remained, all of which are presented in Table 5.2, with a full version of the table online.[7] This significant drop is largely not due to mis-classifications by the algorithm, but instead light-curves with the EA/EB shape which were subsequently found to be blends during human eyeballing, or those which appear in multiple sectors. As such, some care is advised when interacting with this catalogue, as further less obvious blends may still exist due to the large *TESS* pixels. This catalogue is not exhaustive, but represents the cleanest sample of young eclipsing binaries identified by the SOM analysis of these young stars. The catalogue will be updated with less distinctive cases using future supervised machine learning methods in the continuing YOUNGSTER program.

[7] https://warwick.ac.uk/fac/sci/physics/research/astro/people/matthewbattley/ea_and_eb_table_youngster2.csv

Table 5.2: Candidate eclipsing binary signals found in Year 1 of the CDIPS *TESS* light-curves. Light-curves with only single eclipses are labelled as 'Single' in the Period column. A machine-readable version of the full table (9400 targets) is included alongside the thesis online.[8] Periods shown are the primary period found in section 5.2.3.

Gaia DR2 ID	TIC ID	Period [Days]	Type
6585253422567909760	197601231	Single	EA
6619768913628810624	32197062	0.331	EA
...	...	...	...
6630166342263764992	365770318	0.289	EB
5777534485542148608	310104320	1.17	EB

5.6 Discussion

With these promising results, it is clear how powerful the Self-Organising Map technique is for exploring the variability of young stars, both in terms of viewing the broad distribution in variability and for isolating individual groups of distinct variability such as EAs and EBs. The final Kohonen map provides an intuitive and useful overview of the entire first year of young star data in *TESS* (especially when individual pixels are queried) and can even speed up identification and characterisation of systematic errors in the data. This is the first time that such a technique has been applied to a dedicated sample of young stars, as well as the first time such a technique has been applied in bulk to *TESS* data. Applying this technique to young stars specifically allows for a deeper view of unusual variability types which would easily be drowned out by some of the distinctive variability classes presented in older samples of stars (e.g. Cepheids, RR Lyraes). This is especially important as their spot modulation and comparatively short-period variability can result in significantly different shapes to the standard accepted variability classes. Indeed one particularly interesting feature of the full S1-13 SOM (Figure 5.12) is how uncommon basic sinusoidal shapes are, suggesting that simple spot modulation with single large spots dominating the photosphere are not the dominant variability type. This wide array of different shapes helps to explain why transit surveys struggle to find planets around young stars with traditional 'one-size-fits-most' detrending methods.

However, despite the useful results offered by the SOM analysis, some key provisos remain. Most importantly, as the SOM technique works by folding light-curves and binning phase-curves by the strongest periodic signal, it effectively obscures or reduces any short-term or aperiodic activity, including important features such as flares and aperiodic dipping activity. Similarly, because all phase-curves are normalised, some noisy light-curves can have their slight variability greatly exagger-

ated. These two factors show the need for extra factors like period, peak-to-peak scatter statistics and amplitude measurements when attempting wider classification of light-curves. This is corroborated by Brett et al. [2004] who point out that because real-world examples will have a lot of similar shapes at the edges of each class, additional diagnostic attributes are required for safer, more accurate classification. In order to combat this in the continuing YOUNGSTER program, information from the generated SOMs (in particular the final SOM position and Euclidean distance from final pixel) will be combined with known stellar information, light-curve statistics and known variability classes in a wider random-forest-based machine-learning method in future (see YOUNGSTER III, Battley et al., in prep). While biases and varying membership likelihoods in the CDIPS target-list mean that reliable statistics are not easily obtained from this current sample, using these SOM-based techniques with more accurate target-lists (such as in individual moving groups/associations) have the potential to provide statistical overviews of the shapes present in a specific sample.

One additional concern was that interpolating over the data gaps in the analysis was changing the phase curves sufficiently to affect the final position of the light-curve on the overall map. As a test to confirm this was not the case, phase arrays for all light-curves from Sector 2 (except those with periods resulting in gaps in the phase-curve) were reconstructed with the gap retained instead of interpolation and mapped onto the final SOMs trained on Sectors 1-5 (Figure 5.3 and 1-13 (Figure 5.12 respectively. All light-curves in well-defined variability clusters (such as EA/EB eclipsing binaries) were found to fall precisely on their original positions in both test cases, with 85% of all light-curves falling within 10 pixels of their original position (and 70% overall within 2 pixels). As the shape associated with each Kohonen pixel in the more amorphous region of the SOM was not observed to evolve quickly over the course of 10 pixels, the imposed linear interpolation does not appear to result in significant mis-classifications of shape.

Because of *TESS*'s observing strategy, most targets analysed in this study only had a single sector of *TESS* photometry, totalling approximately 27days of observations. This means that young stars in this sample with dominant activity periods of longer than this window will not be folded by their correct period, instead likely appearing as aperiodic. One way to solve this challenge would be to focus on stars in the *TESS* continuous viewing zone (CVZ), however because the majority of young stars fall near the galactic plane, only on the order of 500 stars in the CDIPS sample analysed here fell in this area of the sky, far fewer than necessary to generate a reliable and informative Self-Organising Map. Using new data for these targets

from *TESS*'s extended mission will help to extend the observational baseline for these objects, however at the time of writing no light-curves beyond Year 1 of the *TESS* mission were available from the utilised CDIPS pipeline. None the less, it is clear that even using a single sector of data allows a wide distribution of variability shapes to be observed. Indeed, one of the key benefits of using binned phase-curves instead of light-curves as the primary SOM data-input is that light-curves from different sectors and data-lengths can be easily compared. Theoretically this should allow data from different instruments to be compared also, however this was shown to be more challenging than originally expected when comparing to the K2 Variable catalogue [Armstrong et al., 2016] in Section 5.5.4 above. In the future this pre-training will be improved by using *TESS* light-curves for objects pre-classified into known variability classes, alongside potential new groups found through deeper SOM analysis.

One additional aspect of the SOM that was made clear by the sector by sector analysis was the required number of light-curves before notable patterns and groups could be seen on the map. In this case, primarily using a 40x40 SOM, 10,000 light-curves were found to be the point at which clustering became significant, which is why the first five sectors were eventually run as one (Sectors 1-5 had 4772, 3535, 3059, 5435 and 9353 young star light-curves respectively). Naturally smaller SOMs could be used to aid clustering in the case of fewer light-curves, but the 40x40 SOM was found to be a good balance of allowing understanding of the global pattern while not merging too many interesting types of individual variability. Thus here it is recommended that at least 10,000 light-curves are utilised for optimum SOM performance.

A challenge of completing this analysis with *TESS* light-curves compared to previous *Kepler/K2* analyses is the increased blending from neighbouring stars due to *TESS*'s larger pixels. This challenge fundamentally means that some objects nearby to targets with strong variability may be miss-classified due picking up some of this adjacent signal. This effect is made even worse for young stars compared to the general stellar population due to many young stars existing in relatively dense birth environments. Such a challenge is well illustrated by the number of supposed EA/EB light-curves which had to be removed in the eyeballing stage of section 5.5.5, so classifications from *TESS* data alone should be taken with a grain of salt. However, since one of the primary goals of YOUNGSTER program is to detrend the variability of these young stars to search for new young exoplanets, knowing the dominant variability shape of each target is still very important, regardless if this is the true variability class of that target.

One final consideration of using SOMs is their computational cost. While applying data to a previously trained SOM is simple and efficient (requiring only a few minutes provided that it is presented in the same 'SOM_array' format), the initial preparation of phase-folded SOM_arrays and the training of the SOM from these arrays can take on the order of hours to one day for a single sector, depending on the size of a sector. Indeed when carried out on a 2012 Macbook Pro (2.5GHz Intel Core i5 processor; 8GB memory), the SOM_array preparation step was found to take 44 hrs for Sector 7 (89,221 light-curves), or 1.80 seconds per light-curve, while training the SOM using the 'CreateSOM' routine was found to take 75 hrs (3.1 days) to train the full S1-13 SOM (based on 635,107 light-curves), or 0.43 seconds per light-curve.[8] However, applying new data to this trained SOM and plotting the final distribution was typically complete in just a few minutes, so if all data is compared to the same SOM for classification, the main computational consideration is simply the time take to create SOM_arrays for each lightcurve. Further time enhancements can be made by parallelising the process (e.g. completing different sets of light-curves on parallel cores), or analysing only one sector at a time, especially as individual sectors were shown to be very effective at sorting types of variability by themselves (see section 5.5.3).

In Summary, Self-Organising Maps are a powerful technique to explore and characterise young star variability in large surveys such as *TESS*, allowing both an intuitive understanding of the overall variability evolution across the entire sample and highlighting groups of distinct variability. For this reason it will be folded into the wider supervised machine learning methods of the continuing YOUNGSTER program, combined with knowledge of the light-curve statistics and stellar parameters. The final variability classifications from this program can then be used to inform future detrending during searches for new young exoplanets, for example by building variability models or using the knowledge of the light-curve shapes to tune smoothing parameters. Although this analysis primarily focused on the CDIPS light-curves, the methods discussed in this paper are well-suited to other, often more extensive, light-curve sources such as those generated by TESS-SPOC pipeline [Caldwell et al., 2020b], the Quick Loop Pipeline [QLP Huang et al., 2020a] and eleanor pipeline [Feinstein et al., 2019], provided that typical systematics are well characterised for these sources. Furthermore, because the SOM analysis is based on binned phase-curves, it can easily be extended to the main 2min *TESS* data and planned 10min and 300s FFI observations.

[8] Note that this averages out at 24hrs for the SO_array generation and 5.8hrs for the CreateSOM step per sector, which is more manageable

5.7 Summary

Here Kohonen Self-Organising Maps [Kohonen, 1982, 2001] have been used for the first time on a dedicated selection of young stars from the *TESS* survey, in order to explore the wide-scale variability of these interesting objects. This unsupervised machine learning technique allows light-curves to be grouped according to their dominant shapes, without needing prior knowledge of expected groups and clusters. This is particularly important for young stars, where the community it still discovering new types of stellar activity. All currently available light-curves from the Cluster Difference Imaging Photometric Survey [CDIPS, Bouma et al., 2019] are analysed here, representing over 600,000 young star light-curves from the first year of *TESS*'s observations.

Overall Self-Organising Maps reveal young star activity shape distribution to be more varied than expected, with a significant spread over the whole parameter space of shapes, from archetypal sinusoidal signals to more varied 'dipper' or pulsation-like activity. The SOMs are also very proficient at separating eclipsing binary and potential transiting signals from the data, typically appearing as significant hot-spots in the final Kohonen maps. Given the importance of these objects, a target list of those identified in this work is supplied in this work and online. SOMs generated for individual sectors retained similar distributions to the overall Year 1 version, provided that at least 10,000 light-curves were present in that sector (i.e. for all but sectors 1-5, which were combined into a single SOM). This illustrates the SOM's effectiveness in characterising the variability of young stars in a single sector alone, without having to run a full year of light-curve data. While most analysis was completed using a 40x40 SOM, 20x20 and 60x60 versions were also tested for the full sample of light-curves, revealing the strongest groups of variability in the first case and a deeper sub-structure of variability in the latter. However, the 20x20 SOM was found to merge too many dissimilar light-curves, while the 60x60 SOM is somewhat unwieldy for in-depth analyses, so the 40x40 version is recommended for future analyses.

Pre-training the SOM was attempted using light-curves with pre-defined variability classes from the K2 Variability Catalog of Armstrong et al. [2016], however this was found to be challenging due to differences between the characteristics of light-curves from *TESS* and the *K2* mission. Nonetheless, such a method shows sufficient promise to be worth repeating when sufficient *TESS* data is available for these objects from the *TESS* extended mission.

The self-organising map method was also revealed to give valuable and intu-

itive insight into leftover systematics in the *TESS* data (regardless of which pipeline is used), especially where there is variation in the severity of the effects. This will be particularly useful for diagnosing and understanding the significant scattered light systematic present in *TESS* light-curves for dimmer stars.

The increased understanding garnered about young star variability from this analysis will be used to guide future light-curve classification and targeted detrending methods in order to search for new young exoplanets in the continuing YOUNGSTER program.

Chapter 6

Tracking known exoplanets though time: Revisiting the Kepler Field with TESS

> "A straight line may be the shortest distance between two points, but it is by no means the most interesting"
>
> — The Third Doctor, *The Time Warrior*, *Doctor Who*

As exoplanet science moves from detecting exoplanets to characterising their architectures and atmospheres in depth, up to date planet ephemerides are becoming increasingly important. However, with the *Kepler* mission finishing in 2013, many of the important planets and candidates discovered during this mission went essentially unobserved until the second year of the *TESS* mission, meaning that the ephemerides for these objects became increasingly stale. This chapter presents work combining new observations from the *TESS* mission with observations from the *Kepler* Space Telescope in order to update the ephemerides for 22 *Kepler* planets and 4 *Kepler* planet candidates, constituting all *Kepler* planets and candidates with sufficient signal to noise in the *TESS* 2min dataset. This includes a number of young exoplanetary systems such as Kepler-63 ($\sim$ 210Myr), Kepler-289 ($\sim$ 650Myr) and Kepler-411 ($\sim$ 210Myr). A purely photometric method is utilised here to allow ephemeris updates for planets even when they do not posses significant radial velocity data. The obtained ephemerides are of very high precision and at least seven years 'fresher' than archival ephemerides. In particular, signifi-

cantly reduced period uncertainties for Kepler-411d, Kepler-538b and the candidates K00075.01/K00076.01 are reported. O-C diagrams were generated for all objects, with the most interesting ones discussed here. This analysis represents another important method of tracing exoplanets through time, giving evolution of individual planetary systems instead of analysing snapshots of planets at different ages.

This chapter begins with a wider motivation for this research in Section 6.1, before reporting the observations for each object and instrument in Section 6.2. This is followed by a brief reminder of the methods used to update ephemerides (essentially a quick summary of Section 2.8.1), before the main results are presented in Section 6.4. The chapter ends with a discussion and conclusion in sections 6.5 and 6.6.

Declaration - The following chapter is a partial reproduction of the paper 'Revisiting the Kepler field with TESS: Improved ephemerides using TESS 2min data' [Battley et al., 2021], published the *Monthly Notices of the Royal Astronomical Society* in May 2021. The vast majority of the work presented in this chapter was carried out by the author, with co-authors to the paper mostly providing high level guidance and simple proof-reading. The one exception is the transit timing analysis presented in section 5.4.3, which was carried out by Dr Michelle Kunimoto.

6.1 Motivation

Maintenance of planet ephemerides is crucial to further characterisation of known planets. This is especially true for follow-up with high-profile observatories such as the upcoming *James Webb Space Telescope* [*JWST*, Gardner et al., 2006], where timing uncertainties of less than 30 minutes are desirable [Dragomir et al., 2020]. Furthermore, ensuring that ephemerides of known planets and planet candidates remain fresh secures the legacy of large-scale planet-finding missions such as those performed by the *Kepler* satellite [Borucki et al., 2010] and the *Transiting Exoplanet Survey Satellite* [*TESS*, Ricker et al., 2014].

The *Kepler* satellite dramatically changed the field of exoplanet science, discovering almost 3000 validated exoplanets and thousands more planet candidates over the course of its main mission and following *K2* mission [Howell et al., 2014]. However, the original *Kepler* mission came to a close on May 11 2013 after the failure of two of the satellite's reaction wheels[1], meaning that it has now been almost ten years since most of these planets have been observed. This observational gap,

[1]https://archive.stsci.edu/missions/kepler/docs/drn/release_notes25/KSCI-19065-002DRN25.pdf

coupled with uncertainties in the periods and epochs of the *Kepler* transits, has led to many of the planet/candidate ephemerides becoming imprecise or 'stale'.

A number of different surveys have attempted to solve this problem, both for *Kepler* and *K2* planets [e.g. Livingston et al., 2019; Ikwut-Ukwa et al., 2020; Edwards et al., 2020], however the sheer number of planets and candidates discovered by the satellite makes this a significant challenge. Promising follow-up solutions include re-observing the entire *Kepler* field or using large-scale citizen-science approaches [e.g. Zellem et al., 2020; Kokori et al., 2020]. The launch of *TESS* in 2018 provided a new opportunity to address this challenge, surveying ˜75% of the night sky over the course of its two-year primary mission. During this mission, *TESS* returned to the original *Kepler* field, providing new data for the *Kepler* field in Sector 14 (and to a lesser extent, Sectors 15 and 26) of its observations.

Ikwut-Ukwa et al. [2020] have previously demonstrated the significant impact re-observing data from the *Kepler* satellite with *TESS* can have, completing full ephemeris and wider system parameter updates for K2-114, K2-167, K2-237 and K2-261 from the *K2* mission by utilising a combined photometric (*TESS* and *Kepler* data) and spectroscopic (using archival radial velocities) measurements. In the case of K2-114 these new measurements were shown to reduce the uncertainty in the planet's period by a factor of 66 compared to its discovery period. This promising result heralds the way for a wider treatment of the entire *TESS/Kepler* crossover sample in order to have ephemerides ready for the launch of *JWST*.

Furthermore, Christ et al. [2018] point out that re-observing the *Kepler* field with *TESS* provides a host of additional benefits, providing a window into long-term trends such as tidal decay of Hot Jupiters. In particular, Christ et al. [2018] suggest that the Hot Jupiters Kepler-2b/HAT-P-7 and KOI-13b will be particularly interesting objects to follow up with *TESS* data. They also suggest that *TESS* may be very powerful for investigating longer-term transit timing variations. These thoughts are echoed by the modelling of Goldberg et al. [2018] where they conclude that mass uncertainties could be improved for 6-14 planets with the new *TESS* 2min data, depending on the eventual measurement uncertainties.

While the combined photometric/spectroscopic approach of Ikwut-Ukwa et al. [2020] allowed for new global models of the systems to be created, many of the lower-mass Kepler planets and candidates do not currently possess significant radial-velocity data. This motivated the construction of a homogeneous method of updating new ephemerides based on *TESS/Kepler* photometry alone, as was presented in Section 2.8. Here the developed method is used to update the ephemerides for every *Kepler* planet and candidate which was re-observed in the *TESS* 2min data.

6.2 Observations

6.2.1 *Kepler*

Kepler's primary mission ran for approximately four years, or 17 'quarters', from first light on May 2nd 2009 until the loss of the satellite's second reaction wheel in May 2013. Observing the same rich patch of sky in the vicinity of the Cygnus and Lyra constellations for its entire primary mission, the *Kepler* satellite yielded an unprecedented volume of high quality, long-duration photometry, and was the first telescope capable of finding Earth-size planets around Sun-like stars in year-long orbits [Borucki et al., 2010]. The majority of the targets in the *Kepler* mission's primary field were observed in *Kepler's* 29.4min (30min) long-cadence mode, however approximately 512 objects per quarter received 1min short-cadence light-curves covering a month in time [Thompson et al., 2016a].

In this work, *Kepler* long-cadence PDCSAP light-curves were retrieved from the public Mikulski Archive for Space Telescopes (MAST) using the Lightkurve python package [Lightkurve Collaboration et al., 2018]. These light-curves were prepared by the standard *Kepler* science processing pipeline [Jenkins et al., 2010], which involves pixel-level calibration, smear and background removal, optimal aperture selection and modelling of systematic errors introduced by the spacecraft. Due to the stellar variability of most sources over the long *Kepler* observation timeline, an additional detrending step was necessary before planetary transits could be used to update ephemerides. This was achieved by applying a simple 24hr window LOWESS filter [Cleveland, 1979a; Battley et al., 2020] to the out-of-transit light-curve, except in the cases such as Kepler-9 [Holman et al., 2010], where shorter 12-15hr windows were required to handle the shorter-period activity cycles. In order to preserve the form of each individual transit, the transit epochs were masked from the light-curves during this detrending step and replaced with linear interpolations until the detrending was complete.

6.2.2 *TESS*

TESS re-observed the *Kepler* field in the second year of its primary mission. The majority of the Kepler field was observed in Sector 14 (between 18 July to 15 August 2019), but due to overlap in adjacent *TESS* observations, some of these stars were also re-observed in Sectors 15 (15 August - 11th September 2019) and 26 (8 June - 4th July 2020). However, although the entire *Kepler* field was re-observed in Sector 14 of the *TESS* primary mission, only 962 Kepler stars were pre-selected to receive 2min data in *TESS*'s primary mission. These stars are plotted in Figure 6.1,

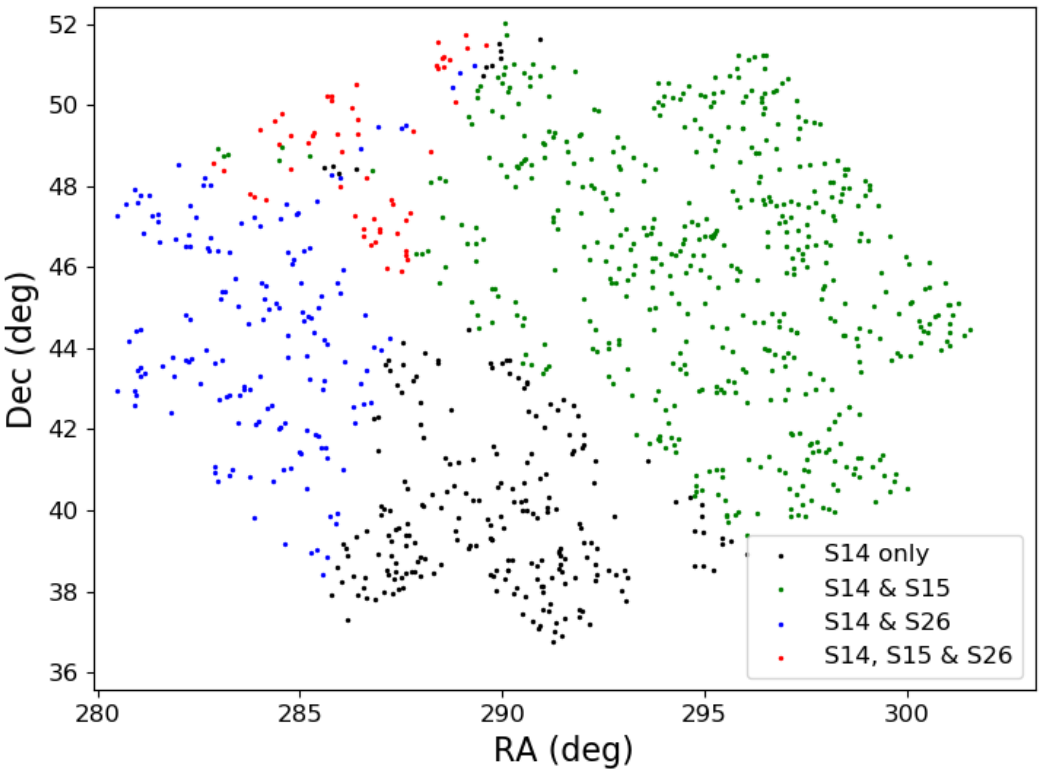

Figure 6.1: Overview of *Kepler* stars reobserved by *TESS* with 2min cadence. Individual stars are colour-coded according to which *TESS* sectors they were observed in: Black: Sector 14 only; Green: Sectors 14 and 15; Blue: Sectors 14 and 26; Red: Sectors 14, 15 and 26.

colour-coded by which sectors they were re-observed in.

All *TESS* light-curves used in this work were generated by NASA's Science Processing Operations Centre (SPOC) and, similarly to the *Kepler* data, accessed using the Lightkurve python package [Lightkurve Collaboration et al., 2018]. This data was extracted from the raw images using the standard SPOC pipeline [Jenkins et al., 2016], which includes pixel-level decorrelation, centroiding and aperture optimisation. For all light-curves the PDCSAP flux data was analysed, which had previously had systematic errors such as times of poor pointing or excessive scattered light removed [Jenkins et al., 2016]. Similar to the *Kepler* data, after extraction via the Lightkurve package, long-term stellar variability outside the transit was removed from the *TESS* 2min light-curves using a 24hr window LOWESS filter.

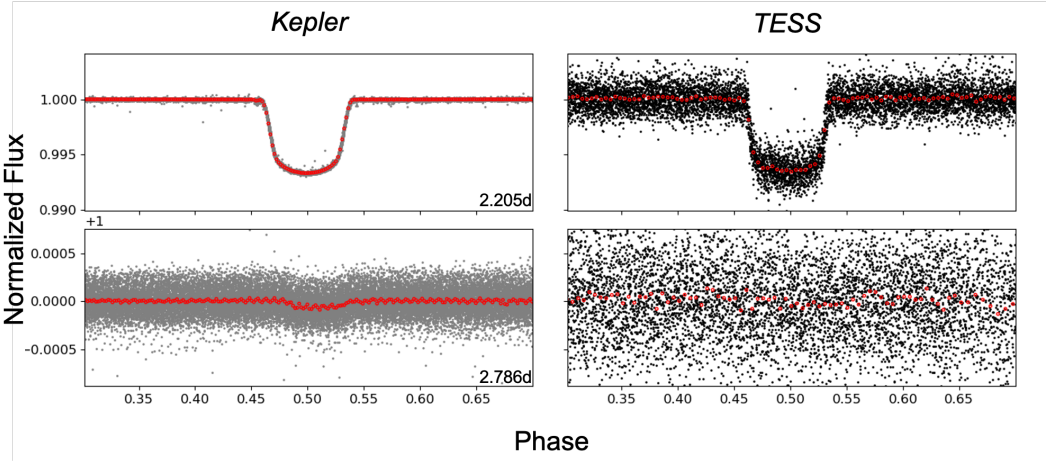

Figure 6.2: Photometric data comparison between *Kepler* (left) and *TESS* (right). Top: HAT-P-7 b/*Kepler*-2 b, a 1.51 $M_{\rm Jup}$ Hot Jupiter is clearly recovered by both *Kepler* and *TESS*. Bottom: *Kepler-21* b, a 1.6$M_{\oplus}$ super-Earth is clear in the folded *Kepler* data but lost in scatter in the *TESS* data.

6.2.3 Photometry Comparison

The *Kepler* and *TESS* missions were designed with two very different survey strategies in mind, and hence differ in their photometric performance. While the *Kepler* mission focused on a relatively small (100 deg^2, or 0.25% of the sky) northern hemisphere section of the sky near the Cygnus, Lyra and Draco constellations [Borucki et al., 2010], the *TESS* primary mission focused on maximising the amount of sky viewed, achieving almost 75% coverage of the sky[2] in its two year primary mission, updated from the originally planned 85% due to issues with scattered light [Ricker et al., 2014]. To achieve this coverage with *TESS*, a slight sacrifice in sensitivity was required, with combined differential photometric precision (CDPP) dropping from an average of 30ppm for the *Kepler* mission [Gilliland et al., 2011] to approximately 100ppm for Sector 14 of the *TESS* mission.[3] This reduced sensitivity comes from a variety of sources, but most importantly the reduced effective lens size in *TESS* due to reducing from a single large mirror (effective lens diameter of 95cm) on the *Kepler* satellite to smaller 10.5cm lens-based telescopes in *TESS*. This reduction in effective lens diameter results in the resolution of the *TESS* instrument decreasing to 21 arcseconds per pixel compared to Kepler's 4 arcseconds per pixel and yields increased blending from nearby stars compared to the *Kepler* satellite. The com-

[2]https://www.nasa.gov/feature/goddard2020/nasa-s-planet-hunter-completes-its-primary-mission

[3]https://archive.stsci.edu/missions/tess/doc/tess_drn/tess_sector_14_drn19_v02.pdf

bination of these effects, coupled with the shorter observational timescale in *TESS* (27 days to 1 year in *TESS*; 4 years in *Kepler*), means that *TESS* is less sensitive to small planets than *Kepler* was, with Christ et al. [2018] predicting that only 277 *Kepler* planets have a $\geqslant 50\%$ likelihood of being recovered by *TESS* to a 3σ level in the wider-extent 30-min cadence data.

This issue is well illustrated in Figure 6.2, showing the comparison between *Kepler* and *TESS* data for a large Jupiter-sized planet (HAT-P-7 b; Pál et al. [2008]) and a small Super-Earth sized planet (Kepler-21 b; Howell et al. [2012]). While the planetary signal for both systems is clear in the *Kepler* data, the smaller planet is entirely indiscernible in the 2min *TESS* light-curve. This perhaps helps to explain why only seven known *Kepler* planets reobserved with *TESS* 2min cadence were identified as *TESS* Objects of Interest (TOIs).[4]

For the purpose of this study we focus on all *Kepler* confirmed planets and candidates with transits visible in the *TESS* 2min data, including many that were missed as *TESS* TOIs.

6.2.4 Overall Target Selection

In order to find all *Kepler* planets and candidates which received 2min *TESS* data, the 962 *Kepler* stars with short-cadence data were cross-matched with all transiting planets and *Kepler* planet candidates from the NASA Exoplanet Archive.[5] This revealed a population of 49 *Kepler* planetary host stars (harbouring 93 transiting planets) and 28 planet candidate systems which were re-observed by *TESS* in its short-cadence mode. However, due to the comparatively short (28-day) *TESS* observation window compared to some long-period *Kepler* planets, the expected transit times were missed for 19 planets and 8 planet candidates, making them unusable for ephemeris updates. The planetary system of Kepler-34 (TIC 164457525) was also removed at this stage, as its planet is not expected to transit again until Nov 18 2066 due to its complex circumbinary orbit [Welsh et al., 2015; Martin, 2017].

The *TESS* 2min light-curves for the remaining 73 transiting planets and 20 planet candidates were searched for new transits both by eye and systematically with a box-least-squares search [Kovács et al., 2002; Hartman and Bakos, 2016], as implemented as the `BoxLeastSquares` function in the `astropy` python package [Astropy Collaboration et al., 2013, 2018]. Because of the reduced sensitivity of the *TESS* instrument, many of the smallest *Kepler* planets were indiscernible from noise in the *TESS* data alone, so these systems were also cut from the analysis. Overall

[4] https://exofop.ipac.caltech.edu/tess/view_toi.php
[5] https:exoplanetarchive.ipac.caltech.edu [Accessed 12 August 2020]

there remained 22 planets (in 18 planetary systems) and 4 planet candidates where the ephemerides could be reasonably improved using the *TESS* 2min data. It is worth noting however that for many of these only a single transit was observed. Table 6.1 gives an overview of the final systems which have been updated in this study, while a system by system summary of this target selection is included for each planet host in Table C.1.

6.3 Overview of Methods

The full description of the methods used to update planet ephemerides was presented in Section 2.8, however for context an overview is given here, along with specifics of how planetary and stellar parameters were sourced for the 22 planets and 4 candidates updated in this work.

Based on the methods of Gajdos et al. [2019] and Holczer et al. [2016] to analyse the full Kepler Q1-17 data, the first step consisted of building a transit model for the planet of interest based on archival system parameters, and then using this model to estimate the transit times of individual transits by minimising the chi-squared statistic in a grid of 1min resolution around the expected transit.

As a starting point for the models, planetary and stellar parameters were retrieved from the NASA Exoplanet Archive for all objects. In cases where the data highlighted on the exoplanet archive was out of date, these parameters were updated to reflect the most recent literature. In order to collect the most recent radii information, the entire sample was cross-matched with Berger et al. [2018]'s revised radii of *Kepler* stars and planets based on *Gaia* DR2 data [Lindegren et al., 2018; Gaia Collaboration et al., 2018a]. This list was then cross-matched with the updated linear ephemerides found by Gajdos et al. [2019], who used the entire length of *Kepler* data (Q1-17) to update ephemerides for 1977 exoplanets. As a first check for significant transit timing variations, the TTV flag was checked on the Exoplanet Archive, and all systems cross-matched with TTV data from surveys completed by Holczer et al. [2016] and Gajdos et al. [2019].

Similar to the *Kepler* planets, original data for the *Kepler* planet candidates was collected from the *Kepler* Objects of Interest (KOI) list on the Exoplanet Archive, taking care to disregard any KOIs previously downgraded to false positives. These data were then cross-matched with data from the most recent TESS Input Catalog [TICv8, Stassun et al., 2019] and Berger et al. [2018]'s revised radii of *Kepler* stars and planets to obtain up to date information for the stellar hosts.

Table 6.1: Overview of final Kepler planetary and candidate systems for which ephemerides were updated in this study. A TTV flag of '1' denotes any systems where unusual/non-linear behaviour was observed in the O-C diagrams constructed in this work.

Planets

System Name	Planets updated	TTV flag	Discovery Paper
Kepler-2/HAT-P-7	b	1	Pál et al. [2008]
Kepler-10	c	1	Fressin et al. [2011]
Kepler-13/KOI-13	b	0	Borucki et al. [2011]
Kepler-14	b	0	Buchhave et al. [2011]
Kepler-18	d	1	Cochran et al. [2011]
Kepler-25	b, c	1	Steffen et al. [2012]
Kepler-51	b	1	Steffen et al. [2013a]
Kepler-51	d	1	Masuda [2014]
Kepler-63	b	0	Sanchis-Ojeda et al. [2013a]
Kepler-68	b	0	Gilliland et al. [2013]
Kepler-89/KOI-94	d	1	Weiss et al. [2013]
Kepler-96	b	0	Marcy et al. [2014]
Kepler-289	c	1	Rowe et al. [2014]
Kepler-396	b, c	1	Xie [2014]
Kepler-411	c	1	Morton et al. [2016]
Kepler-411	d	1	Sun et al. [2019]
Kepler-412	b	1	Deleuil et al. [2014]
Kepler-448/KOI-12	b	1	Bourrier et al. [2015]
Kepler-538	b	0	Morton et al. [2016]
Kepler-1517	b	0	Morton et al. [2016]

Candidates

System Name	Candidates updated	TTV flag
KIC 7199397	K00075.01	1
KIC 8554498	K00005.01	0
KIC 9955262	K00076.01	1
KIC 9418619	K06068.01	0

Once estimated transit times have been found based on the theoretical models, individual transits were aligned and stacked into an overall transit from the *Kepler* or *TESS* data individually, in order to gain a 'master' transit from the data alone. This master transit was then compared to the one found by simply folding by the light-curve by its primary period, with the cleanest phase curve taken forward in the analysis. This dual approach meant that accurate transit curves could be obtained for all systems, regardless of whether the individual systems had transit timing variations. A new `batman` [Kreidberg, 2015] transit model was built from this clean phase curve and then in turn used to refine the time of individual transits further, using a combination of the `optimize` function within Foreman-Mackey et al. [2020]'s `exoplanet` package, and a `pyMC3` MCMC analysis [Salvatier et al., 2016]. These final transit times were fit with a linear ephemeris using three iterations of a weight-linear least-squares fit, which in turn enabled Observed-Calculated (O-C) diagrams to be generated for each planet. The resulting ephemerides and most interesting O-C diagrams drawn from this analysis are discussed next.

6.4 Results

6.4.1 Known Planets

The final updated ephemerides are presented in Table 6.2, with known planets in the top section of the table and the four planet candidates presented underneath. Overall good agreement was found for all updated planet ephemerides when compared to those from *Kepler* alone. With one exception (Kepler-51 d - Masuda [2014]), all new periods were within 3σ of their archival values, despite the very precise measurement errors. Any differences between new and archival T_0s can easily be explained by different individual transits being chosen as the zero point. The difference in period for Kepler-51 d is perhaps not surprising given the known transit timing variations of this planet coupled with its long period (130.2 days). Indeed, only nine reasonable transits were captured in the *Kepler* data, and only a single transit in the data from *TESS*'s primary mission. Because of the >8 year gap between *Kepler* and *TESS* observations and the inclusion of the new *TESS* data increasing the available number of data-points by 11%, the new ephemeris found in this study is favoured. The transit timing variations for this planet are plotted in Figure 6.3 and explored further in Section 6.4.3 below.

The overall precision for each ephemeris was found similar to those derived by Gajdos et al. [2019], likely because both their study and this one used the entirety of the *Kepler* data to define the planet ephemerides. The new *TESS* data was thus

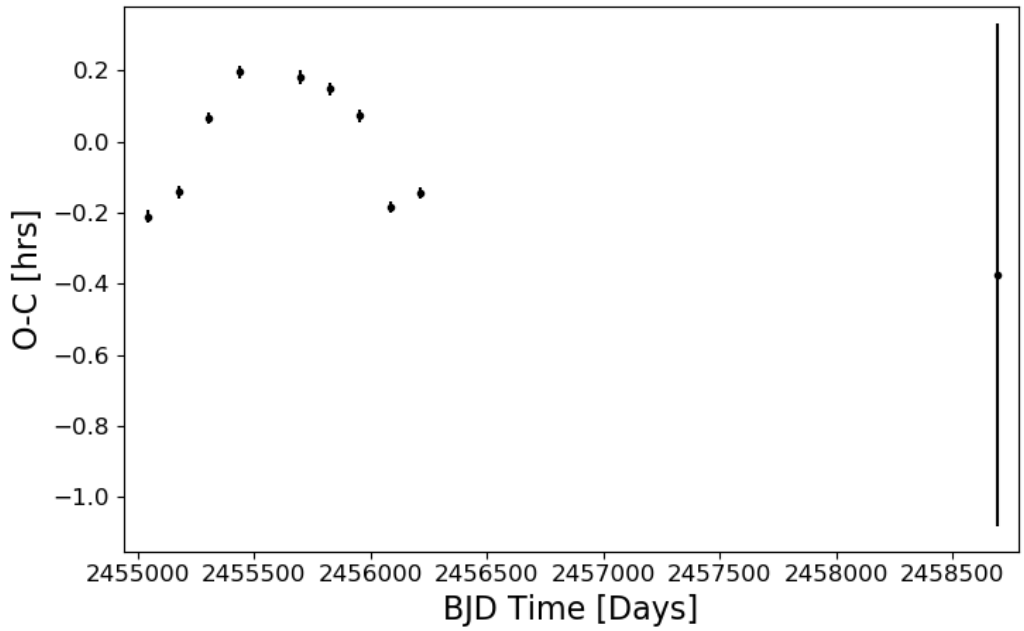

Figure 6.3: O-C plot for Kepler-51 d, showing evidence of transit timing variations after a linear ephemeris fit.

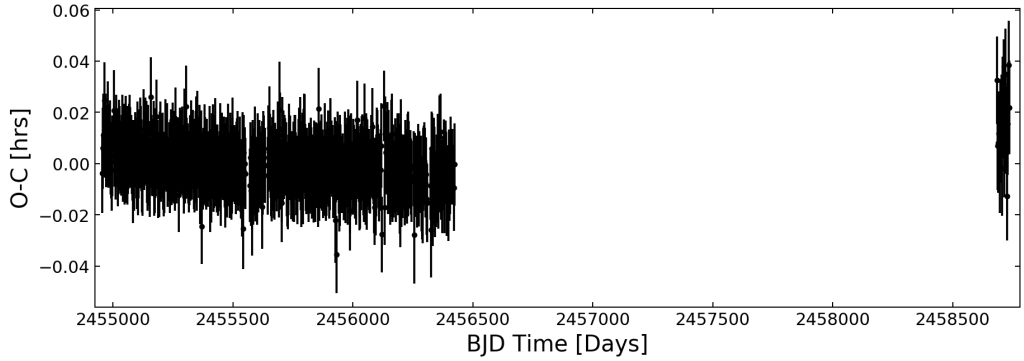

Figure 6.4: O-C plot for HAT-P-7b, showing offset between *Kepler* (BJD $<$ 24565000) and *TESS* (BJD $>$ 2458500) datasets.

most useful for improving the precision for planets whose most recent ephemerides came from different sources to Gajdos et al. [2019], namely Kepler-2/HAT-P-7 b, Kepler-13 b, Kepler-18 d, Kepler-289 c, Kepler-396 b & c, Kepler-411 c & d and Kepler-538 b [Pál et al., 2008; Borucki et al., 2011; Cochran et al., 2011; Rowe et al., 2014; Xie, 2014; Morton et al., 2016; Sun et al., 2019, see Table 6.2 for most recent ephemeris references]. It should be noted that due to the typically shallower transits recovered in the *TESS* data, the uncertainties in each *TESS* transit are considerably higher than those of each *Kepler* transit, which helps to explain why the new *TESS* data is less constraining than initially expected.

Of the Kepler planets studied in this work, only KOI-13/Kepler-13 b (Borucki et al. [2010], Figure 2.10) has previously been updated with the newly available *TESS* data. This is presented by Szabó et al. [2020], who find a period of 1.76358760 $\pm$ 0.00000003 days, and a T_0 of 2455101.707254 $\pm$ 0.000012 BJD, in clear agreement with the values obtained in this work (see Table 6.2). Similar to Szabó et al. [2020], no evidence of TTVs or tidal decay of this Hot Jupiter was seen in this analysis. It should be noted that though the eventual ephemeris reported for Kepler-13b in this work was drawn from a simultaneous *TESS/Kepler* fit of the entire light-curve, the ephemeris obtained using the initial TTV-focussed (transit-by-transit) fit was also in clear agreement, with a value of T_0 = 2454955.32946735 $\pm$ 0.000006153 and Period = 1.76358760059 $\pm$ 0.00000001281. This system thus provides a useful test-case for the overall methods used in the wider analysis carried out in this work.

One planetary O-C diagram of particular note in this work is that for Kepler-2b/HAT-P-7b, a bright (Gmag = 10.365), well-studied Hot Jupiter (2.0 R_J) originally discovered by Pál et al. [2008]. This was previously identified as being a prime target for testing orbital decay by Christ et al. [2018]. As can be clearly seen in Figure 6.4, a slight (2.5min) offset was found to exist between the original *Kepler* data zero-point and new *TESS* data. While extending the errors of the *TESS* data points to two or three standard deviations would put them in agreement with the general *Kepler* trend, the observed shift is still considered odd given that the excellent timing agreement for similar Hot Jupiter systems such as KOI-13/Kepler-13b and Kepler-412b. In an attempt to solve this discrepancy, additional transits for HAT-P-7 b were sought from the Exoplanet Transit Database[6] [Poddaný et al., 2010], but given the even larger timing uncertainties from the ground based data, no trend was evident. However, the original discovery T_0 [Pál et al., 2008] also aligns with a slight timing offset of 2.5min above this *Kepler* zero-point, which further supports the hypothesis that the *Kepler* data represents the flat turning point

[6] http://var2.astro.cz/ETD/etd.php?STARNAME=HAT-P-7&PLANET=b

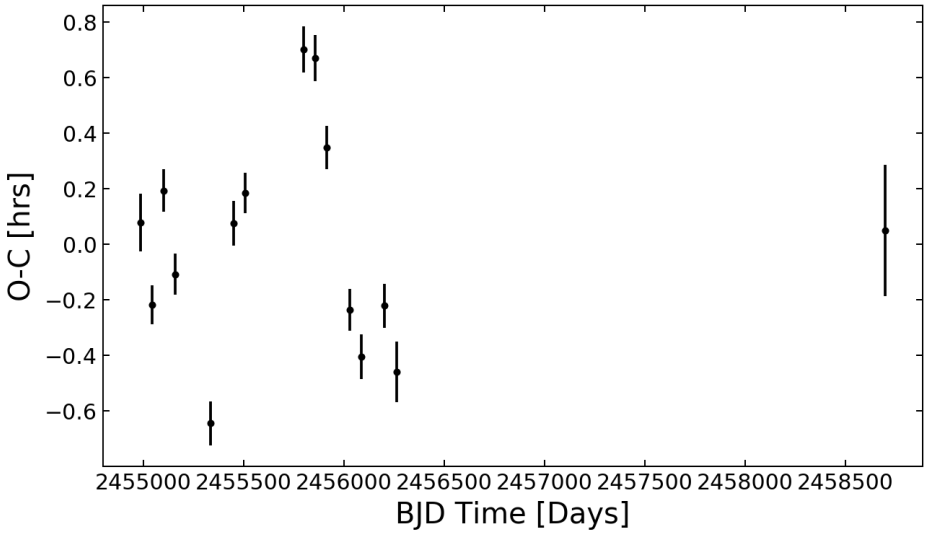

Figure 6.5: O-C plot for Kepler-411d, illustrating interesting Transit Timing Variations with semi-amplitudes of almost an hour.

of longer-period transit timing variations. It is hoped that further data from space telescopes (such as when *TESS* returns to the *Kepler* field in year four of its mission) will help to test whether this trend is real.

Another interesting system updated here is Kepler-411 d (Figure 6.5). While only one transit was caught in the *TESS* data (due to the relatively long 58d period), there is clear variation in the O-C diagram with a semi-amplitude of almost one hour. These variations are not explained by the inner planets in the system (Kepler-411 b & c), with Sun et al. [2019] analysing these same TTVs (see Figure 8 from Sun et al. [2019]) and deducing that they were best modelled by considering an extra non-transiting planet (31.5 day period) in the Kepler-411 system, Kepler-411 e. More new *TESS* data will be valuable in exploring these interesting TTVs further and hopefully constrain the parameters of this planetary system. Similarly large amplitude O-C variations were seen for Kepler-396c, so it is imperative that transit timing variations are taken into account when forecasting future observations in the Kepler-396 system (Figure 6.9).

Other than those systems discussed above, the confirmed planetary systems which received updated linear ephemerides in this work had transits that aligned well with the previous *Kepler* observations and thus should be trustworthy for future observations. However, some care should be taken when forecasting individual transits within these systems, as approximately half of the updated systems exhibited

non-linear behaviour in their O-C diagrams.

6.4.2 Planet Candidates

While issues such as transit epochs falling outside the *TESS* observation window and the increased scatter of *TESS* data precluded updating the ephemerides for the majority of the 28 *Kepler* candidates, the *TESS* data provided sufficient sensitivity to update four systems: KOI-5, KOI-75, KOI-76 and KOI-6068. The new ephemerides for the candidates in these systems can be found in the second part of Table 6.2. The inclusion of the *TESS* data significantly decreased the uncertainty in both the initial transit time (T_0) and the period for these candidates. This new data was most helpful for the two longer-period candidates, K00075.01 (105.9d period) and K00076.01 (77.5d period), reducing their uncertainties by almost an order of magnitude despite only single transits being present in the *TESS* observations. Care must be taken when using these new linear ephemerides however, as some evidence of transit timing variations can be seen in the O-C diagrams for both candidates (K00075.01 - Figure 6.6; K00076.01 - Figure 6.7).

In the case of K00075.01, the O-C plot resulting from the weighted linear fit is clearly not linear, suggesting that it would be better fit with a non-linear ephemeris. Van Eylen et al. [2019] previously fit a sinusoidal model to the O-C plot for K00075.01, which suggested a period of 1892 days and a TTV amplitude of 22min. However, the new *TESS* data point suggests a considerably larger amplitude is more appropriate. It is hoped that further transits of this candidate system from the extended *TESS* mission will help to constrain the long-term TTV behaviour further.

Meanwhile, although the long-term trend of the K00076.01 system's ephemeris appears linear, there is evidence of shorter-term transit timing variations, with an amplitude of approximately 0.2hrs. This may be partially due to the stellar variability clear in the *Kepler* light-curve, which has peak and trough-like features on a similar time-scale to the transit duration. Perhaps because of this variability, K00076.01 has been variously labelled both a candidate and false positive by different *Kepler* KOI data releases, though most recently a candidate. More in-depth analysis of this individual system is recommended in future to test its planetary nature.

It should also be noted that while the O-C diagram for K06068.01 is flat and well-constrained, it has been labelled as a false positive in some past catalogues (e.g. q1_q16_koi [Mullally et al., 2015], q1_q17_dr25_koi [Thompson et al., 2018]). Nonetheless, since the *TESS* transit appears planet-like and the system

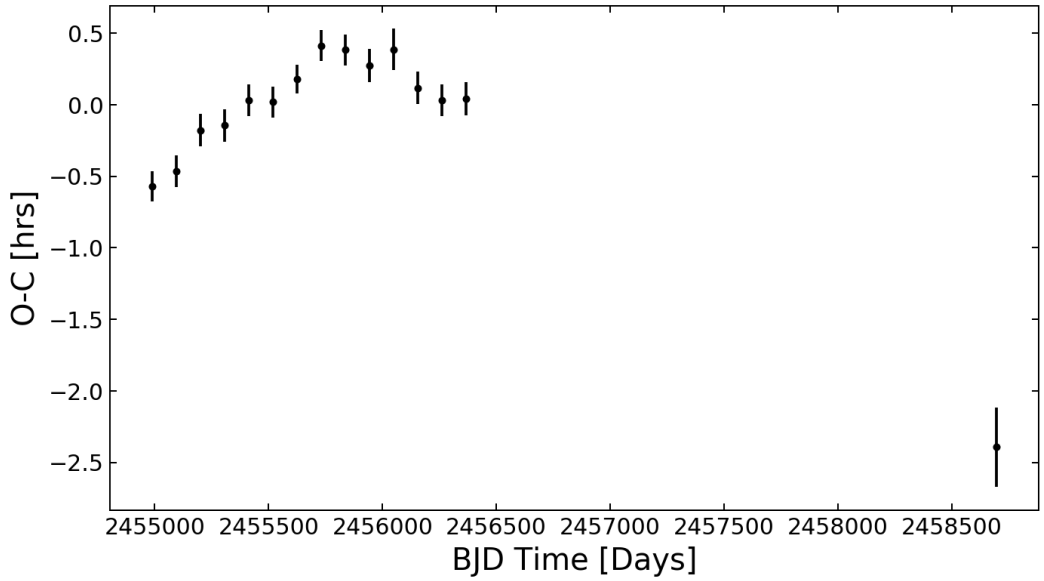

Figure 6.6: O-C plot for *Kepler* candidate K00075.01 after fitting the linear ephemeris derived in this work. A non-linear ephemeris is considered more appropriate for this object, however is left for the future when more *TESS* datapoints are available.

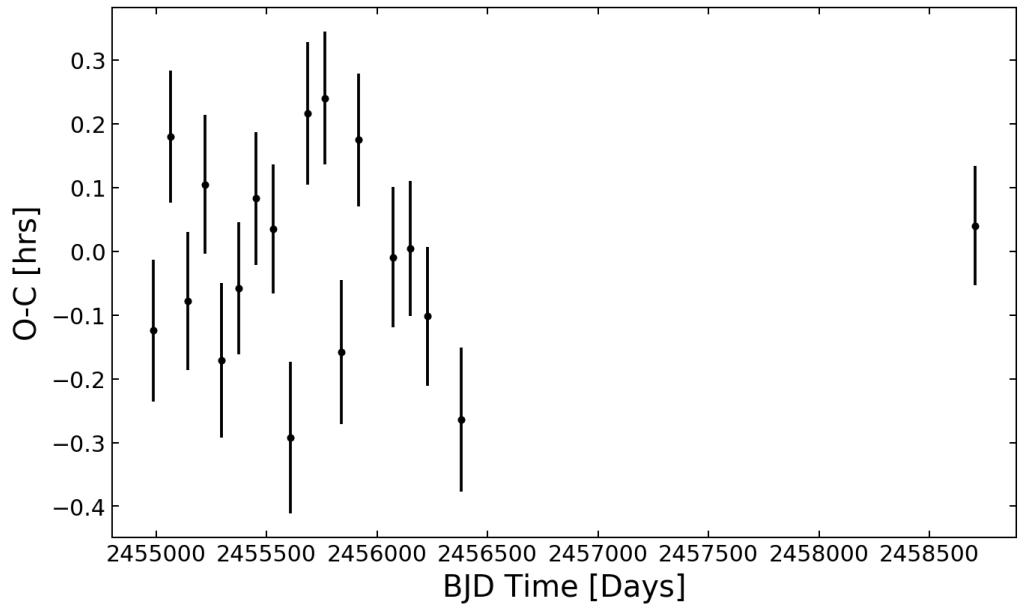

Figure 6.7: O-C plot for *Kepler* candidate K00076.01. Some evidence of short-term transit-timing variations are apparent.

Table 6.2: Overview of all ephemerides updated in this work. Planet ephemerides are presented first, followed by the four planet candidates. Presented uncertainties are the 1σ uncertainty for each value. n.b. 'Ref' refers to the reference each component of the archival ephemerides are drawn from, typically the most recent for each system: 1. Bonomo et al. [2017]; 2. Gajdos et al. [2019]; 3. Szabó et al. [2020] 4. Holczer et al. [2016]; 5. Su et al. [2020]; 6. Mayo et al. [2019]. All archival information for the four candidates was retrieved from the $q1\_q17\_dr25\_koi$ KOI data release [Thompson et al., 2018]. A machine-readable copy of this table can be found in the online supplementary material.

Planets

Planet Name	Archival T_0 [BJD]	Archival Period [days]	Ref.	Updated T_0 [BJD]	Updated Period [days]
Kepler-2b	2454954.357462 ± 0.000005	2.204737 ± 0.000017	1	2454954.3585572 ± 0.0000063	2.20473539167 ± 0.00000001654
Kepler-10c	2455062.2665100 ± 0.0004297	45.29430079 ± 0.00003051	2	2454971.6772661 ± 0.0006847	45.29426146 ± 0.00003783
Kepler-13b	2455101.707254 ± 0.000012	1.76358760 ± 0.00000003	3	2454953.56595833 ± 0.00003827	1.76358750002 ± 0.00000002147
Kepler-14b	2454971.08821000± 0.00006808	6.7901211029 ± 0.0000005613	2	2454957.50815829 ± 0.00008051	6.7901236131 ± 0.0000003985
Kepler-18d	2454961.155156 ± 0.000661	14.85891873 ± 0.00000074	4	2454961.1542075 ± 0.0002423	14.858908757 ± 0.0000004364
Kepler-25b	2455004.7108100 ± 0.0001007	6.2385326915 ± 0.0000007349	2	2454954.7979391 ± 0.0002168	6.2385347882 ± 0.0000001619
Kepler-25c	2455011.52792000 ± 0.00007383	12.720374906 ± 0.000001158	2	2454960.6467450 ± 0.0001144	12.720370495 ± 0.000001703
Kepler-51b	2455714.5917200 ± 0.0001741	45.15530956 ± 0.00001897	2	2454992.10682769 ± 0.0003851	45.15529233 ± 0.00002211
Kepler-51d	2455695.9210000 ± 0.0002442	130.17662541 ± 0.00007274	2	2455045.0339014 ± 0.0004519	130.17784455 ± 0.00008134
Kepler-63b	2455010.84340000 ± 0.00002768	9.4341503479 ± 0.0000003339	2	2454954.23899519 ± 0.00003794	9.4341522797 ± 0.0000004323
Kepler-68b	2455006.85878000 ± 0.00007639	5.3987525913 ± 0.0000005231	2	2454958.2700925 ± 0.0009446	5.398752420 ± 0.000003113
Kepler-89d	2454965.7417600 ± 0.0001014	22.342971172 ± 0.000002603	2	2454965.7413033 ± 0.0001169	22.342982351 ± 0.000002982
Kepler-96b	2455004.02020000 ± 0.00008997	16.238459306 ± 0.000001893	2	2454955.3013031 ± 0.0001767	16.238459945 ± 0.000005665
Kepler-289c	2455069.661672 ± 0.002848	125.86526539 ± 0.00000325	4	2455069.6605154 ± 0.0002487	125.86521071 ± 0.00004151
Kepler-396b	2454995.495267 ± 0.005940	42.99292187 ± 0.00000635	4	2454995.4942951 ± 0.0004076	42.99292140 ± 0.00002072
Kepler-396c	2455015.677038 ± 0.019460	88.51067812 ± 0.00002174	4	2455104.1858280 ± 0.0002890	88.51097554 ± 0.00003917
Kepler-411c	2454968.2224 ± 0.0002	7.834435 ± 0.000002	5	2454960.3876333 ± 0.0001131	7.834436247 ± 0.000001137
Kepler-411d	2454984.8484 ± 0.0061	58.02035 ± 0.00056	5	2454984.8478013 ± 0.0005473	58.02023116 ± 0.00004203
Kepler-412b	2454966.02102000 ± 0.00002357	1.7208612825 ± 0.0000000491	2	2454966.02101665 ± 0.00002771	1.72086125797 ± 0.00000005710
Kepler-448b	2454979.59635000 ± 0.00002880	17.8552258437 ± 0.0000006438	2	2454961.74166158 ± 0.00002718	17.8552273080 ± 0.0000005863
Kepler-538b	2455044.6789 ± 0.0010	81.73778 ± 0.00013	6	2454962.9402449 ± 0.0007438	81.73797957 ± 0.00007330
Kepler-1517b	2454966.50342000 ± 0.00006786	5.5460843094 ± 0.0000004568	2	2454955.4112371 ± 0.0000758	5.5460845139 ± 0.0000005019

Planet Candidates

Candidate Name	Archival T_0 [BJD]	Archival Period [days]	Updated T_0 [BJD]	Updated Period [days]
K00005.01	2454965.974086 ± 0.000148	4.780327581 ± 0.000000852	2454956.4131442 ± 0.0003056	4.7803297849 ± 0.0000006835
K00075.01	2454989.97935 ± 0.00101	105.8817667 ± 0.0001312	2454989.9831552 ± 0.0003309	105.88145696 ± 0.00004402
K00076.01	2454987.70734 ± 0.00230	77.4794704 ± 0.0002456	2454987.7051992 ± 0.0001855	77.47983018 ± 0.00001939
K06068.01	2454967.4228253 ± 0.0000879	6.15025138 ± 0.00000070	2454955.12232977 ± 0.00003749	6.1502525063 ± 0.0000003761

was considered a candidate in the most recent cumulative and q1_q17_dr25_sup_koi KOI deliveries, it is still retained here. In contrast, candidate transits were also plausibly recovered for K00971.01, however eyeballing of this object in the higher-cadence *TESS* data suggested that host star variability is a more likely hypothesis (see Figure 6.8).

Thus while follow-up of *Kepler* candidates with data from *TESS* is somewhat hampered by the short observation windows and decreased sensitivity, this new source of data provides significant opportunities for improved characterisation of *Kepler* candidates, both in transit timing and shape. This is particularly true for long-period *Kepler* candidates, where relatively few data points exist in the *Kepler* data alone. Furthermore, the shorter-cadence *TESS* data is already revealing further details in the photometry to help to separate true planets/candidates from alternate false-positive scenarios.

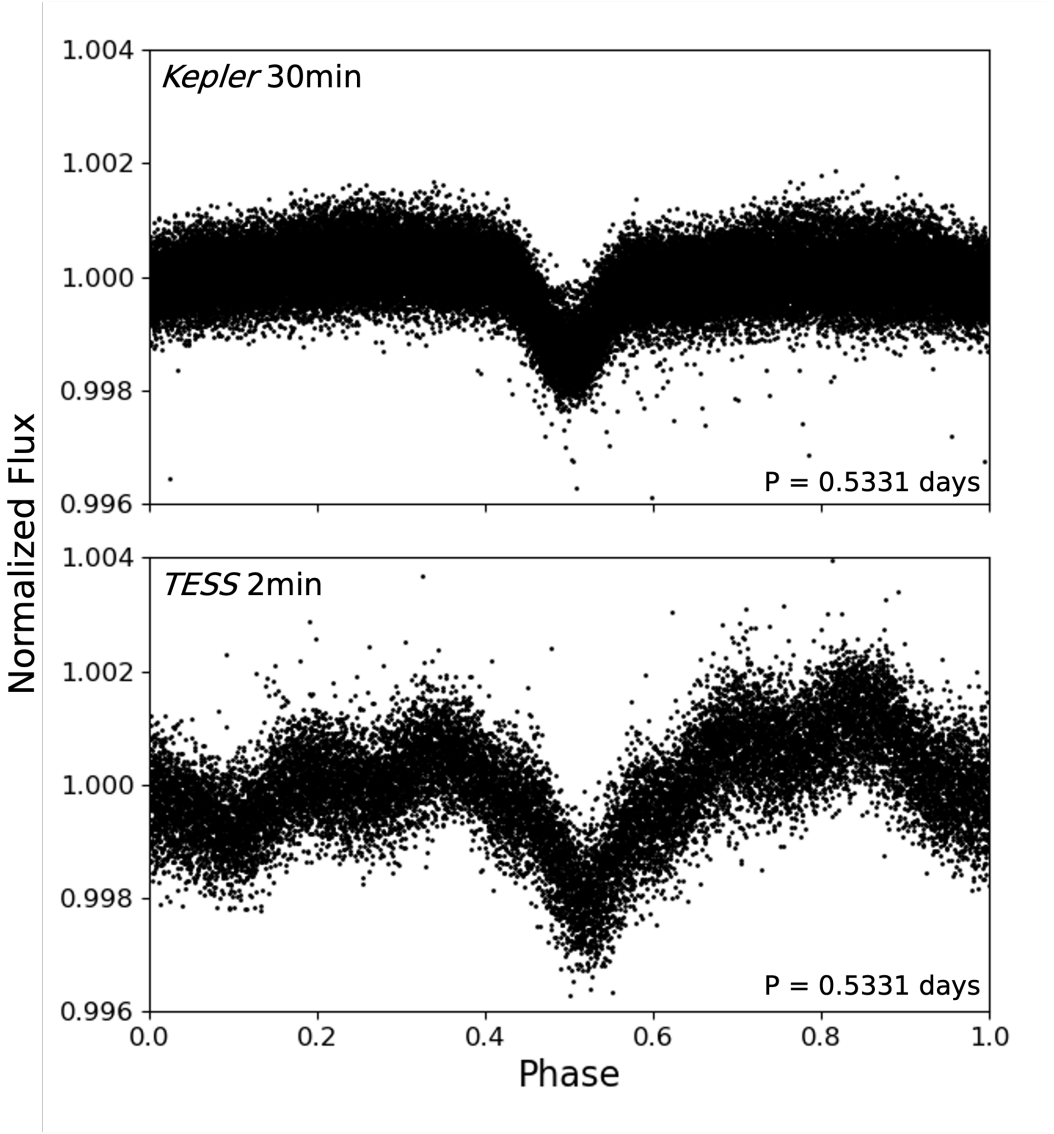

Figure 6.8: *Kepler* 30min (above) and *TESS* 2min data (below) for *Kepler* candidate K00971.01, phase folded by KOI period. The faster cadence *TESS* 2min data reveals significant evidence of stellar activity instead of a true planetary signal, also evidenced by the slight change of phase when folded by the same epoch.

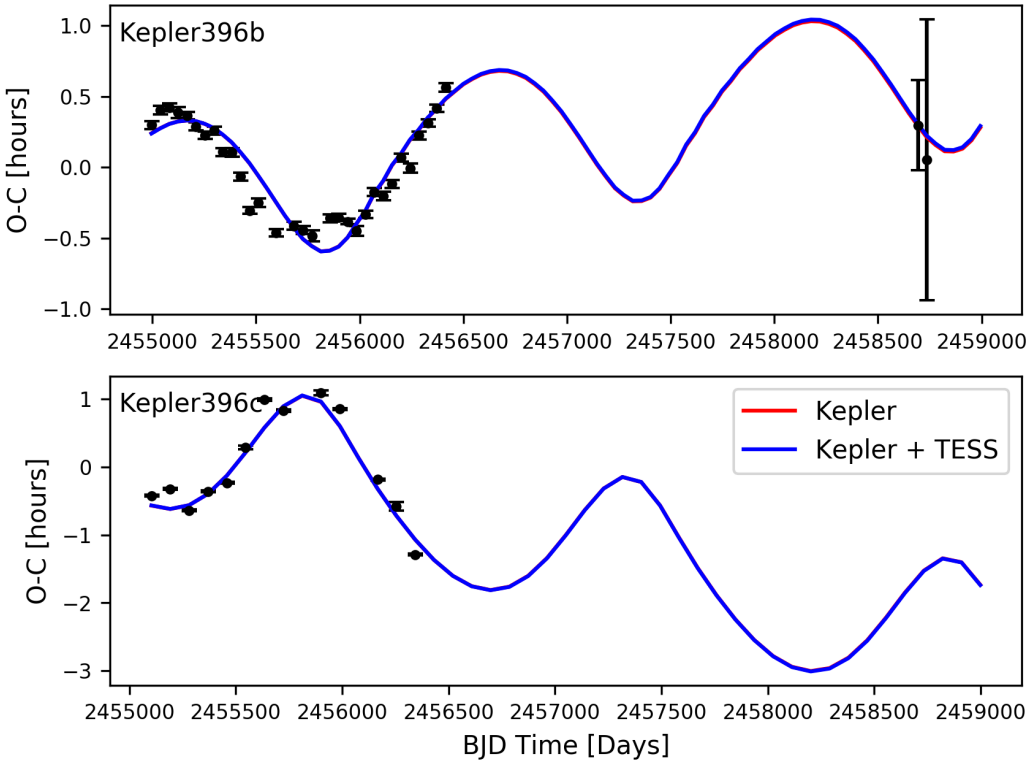

Figure 6.9: O-C diagrams for the Kepler-396 system (b: top; c: bottom) showing clear evidence of transit timing variations for each planet. The best-fit TTV results assuming a default prior, using only *Kepler* data (red) and using the additional *TESS* data for Kepler-396b (blue), are plotted. The solutions lie almost directly on top of one another.

6.4.3 TTV analysis

TTV fits were attempted for *Kepler* multi-planet systems known to exhibit significant TTVs that were re-observed by *TESS*. Five of the systems analysed had at least one planet with ephemerides recoverable from *TESS* data: Kepler-18, Kepler-25, Kepler-51, Kepler-89, and Kepler-396. Inferred masses for the planets in these systems are shown in Table 6.3, giving the median and 68.3% credible region of the MCMC posteriors for both default and high mass runs. Figure 6.9 demonstrates the TTV results for the two-planet Kepler-396 system using the default prior as an example.

While continued transit timing variations were clear in most systems, any differences between the mass results before and after the *TESS* data-points are well within error, and uncertainties on masses did not consistently decrease with the additional data. Hence the TTV fits did not improve with the *TESS* data-points,

which is attributed to the large uncertainties associated with each new transit time, often larger than the TTV amplitudes themselves. Overall, the analysis is challenged by the faintness of the *Kepler* targets, and the difficulty to see individual transits in the *TESS* data.

6.5 Discussion

Precise and accurate ephemerides are crucial to the success and efficiency of future planet characterisation missions such as JWST. Without regular updates to their ephemerides, increasing timing uncertainties in known planets and planet candidates lead to increased observation costs and lost time, especially for longer-period systems.

In this chapter the ephemerides of 22 *Kepler* planets and 4 planet candidates have been updated by analysing new transits in the 2min-cadence data from *TESS*'s primary mission. However, the extent to which ephemeris updates were possible was less than was originally anticipated. The analysis carried out by Christ et al. [2018] prior to the observations of the *Kepler* field by *TESS* was found to be excellent for determining recovery, however was perhaps a little optimistic for individual transit fitting. This was particularly the case for targets such as Kepler-10 b, Kepler-93 b, Kepler-138 c and Kepler-411 b [Batalha et al., 2011; Marcy et al., 2014; Kipping et al., 2014; Wang et al., 2014], which were recovered when all transits were considered, but too shallow to allow individual transits to be fit reliably. For these objects (and other objects of similarly low signal to noise), an alternative method of ephemeris update is advised, perhaps by averaging all *TESS* observations by sector/year.

On a similar note, the null-result of the updated TTV analysis is an important, yet slightly disheartening one, as it suggests that updating inferred planet masses from TTVs with *TESS* may be more challenging than the community hoped. Both the Kepler-396 and Kepler-51 systems were highlighted as high priority systems for improvement with *TESS* by Goldberg et al. [2018], yet neither was significantly improved by the data from the *TESS* primary mission. The biggest challenge faced in updating these TTV masses is the faintness of most *Kepler* systems with significant transit timing variations in the *TESS* photometry, leading to very large uncertainties in individual transit times. When coupled with the fact that the systems analysed typically only had 1-3 transits in the *TESS* data, the new data is not yet very constraining. It is hoped that this situation will improve as the *TESS* extended mission continues.

Table 6.3: Overview of all TTV masses inferred in this work. Central values are the median of the MCMC posteriors, while lower and upper uncertainties are calculated from the 15.9th and 84.1th percentiles, representing the 68.3% credible region.

Default prior

Planet Name	Period [days]	Mass [$M_\oplus$] (*Kepler*)	Mass [$M_\oplus$] (*Kepler + TESS*)
Kepler-18c	7.64	$8.6^{+5.4}_{-3.5}$	$8.4^{+5.6}_{-3.3}$
Kepler-18d	14.86	$8.5^{+2.5}_{-2.5}$	$8.4^{+2.6}_{-2.3}$
Kepler-25b	6.24	$1.0^{+1.1}_{-0.5}$	$1.2^{+1.5}_{-0.5}$
Kepler-25c	12.72	$4.5^{+3.2}_{-1.9}$	$5.0^{+3.7}_{-1.9}$
Kepler-51b	45.16	$0.9^{+1.3}_{-0.8}$	$1.1^{+1.2}_{-0.9}$
Kepler-51c	85.32	$2.8^{+0.5}_{-0.4}$	$2.8^{+0.5}_{-0.4}$
Kepler-51d	130.18	$4.4^{+1.1}_{-0.9}$	$4.4^{+1.0}_{-0.9}$
Kepler-89c	10.42	$5.3^{+2.0}_{-1.5}$	$5.2^{+2.1}_{-1.5}$
Kepler-89d	22.34	$39.4^{+9.0}_{-8.1}$	$39.3^{+8.9}_{-9.0}$
Kepler-396b	42.99	$1.4^{+0.1}_{-0.1}$	$1.4^{+0.1}_{-0.1}$
Kepler-396c	88.51	$1.1^{+0.1}_{-0.1}$	$1.1^{+0.1}_{-0.1}$

High-mass prior

Planet Name	Period [days]	Mass [$M_\oplus$] (*Kepler*)	Mass [$M_\oplus$] (*Kepler + TESS*)
Kepler-18c	7.64	$14.9^{+3.8}_{-3.9}$	$15.1^{+3.8}_{-3.7}$
Kepler-18d	14.86	$11.3^{+1.3}_{-1.6}$	$11.4^{+1.3}_{-1.5}$
Kepler-25b	6.24	$4.8^{+3.8}_{-2.2}$	$5.0^{+3.9}_{-2.2}$
Kepler-25c	12.72	$11.7^{+2.8}_{-3.2}$	$11.9^{+2.7}_{-3.0}$
Kepler-51b	45.16	$2.2^{+1.3}_{-1.1}$	$2.3^{+1.3}_{-1.1}$
Kepler-51c	85.32	$3.3^{+0.4}_{-0.4}$	$3.4^{+0.4}_{-0.4}$
Kepler-51d	130.18	$5.2^{+1.0}_{-1.0}$	$5.2^{+1.0}_{-1.0}$
Kepler-89c	10.42	$7.6^{+2.5}_{-1.9}$	$7.5^{+2.4}_{-1.9}$
Kepler-89d	22.34	$42.6^{+7.7}_{-7.3}$	$42.1^{+7.8}_{-7.4}$
Kepler-396b	42.99	$1.4^{+0.1}_{-0.1}$	$1.4^{+0.2}_{-0.1}$
Kepler-396c	88.51	$1.1^{+0.1}_{-0.1}$	$1.2^{+0.2}_{-0.1}$

These results provide an interesting comparison of *Kepler* and *TESS* photometry in practice and illustrate the differences between optimisation of *TESS* and *Kepler* instruments. One thing that is immediately clear is how impressive the *Kepler* satellite was for analysis of ephemerides and TTVs for fainter stars. These differences are particularly striking in Figure 6.2, clearly illustrating the difference in noise in the *Kepler* and *TESS* observations. The fact that the new *TESS* data did not significantly change the uncertainties in the periods for those analysed by Gajdos et al. [2019] is testament both to their analysis and the very high quality of *Kepler data*. For short-period *Kepler* planets with many transits in the *Kepler* data, it is perhaps unsurprising that only 1-3 sectors of *TESS* data would change the already very precise ephemerides. This is especially the case for such dim stars as make up the bulk of the *Kepler* field, since *TESS* is deliberately designed to search for planets around brighter stars [Ricker et al., 2014]. Nonetheless, for those planets without Gajdos et al. [2019] ephemerides, *TESS* has once again proven itself a very powerful tool for improving the precision of planet and candidate ephemerides, decreasing period uncertainty by orders of magnitude in cases such as HAT-P-7/Kepler-2b, Kepler-411d, Kepler-538b and the candidates K00075.01/K00076.01.

One limitation of this work is that it focused only on *Kepler* targets which received 2min data from the *TESS* mission. While this data-set represents the shortest-cadence data so far released by the *TESS* mission, it suffers from the choices made for which systems were put forward for 2min data collection. Given that the *TESS* 2min catalogue prioritises bright dwarf stars in its search for smaller planets [Stassun et al., 2019], many of the dim stars in the *Kepler* field did not receive 2min data. The analysis completed by Christ et al. [2018] demonstrates the significant promise of extending this analysis to the *TESS* Full Frame Images (FFIs), especially when the *Kepler* field is revisited in extended mission. Such an analysis has the potential to increase the number of updated ephemerides from the approximately 25 systems updated here to hundreds of systems. This analysis is made significantly easier with the recent release of the *TESS* SPOC [Jenkins et al., 2016; Caldwell et al., 2020a] and Quick Look Pipeline [Huang et al., 2020a] FFI light-curves on MAST. The planned 10min cadence data will also help considerably with this effort.

However, given the challenges faced in this study due to the dimness of many *Kepler* planets and candidates, it is imperative that we also find alternative methods/data-sources to maintain the ephemerides of dim or small *Kepler* planets/candidates. Failing to do so risks letting many interesting systems fade into timing uncertainty oblivion.

6.6 Summary

In this chapter ephemerides have been updated for 22 *Kepler* planets and 4 planet candidates using short-cadence data from the *Transiting Exoplanet Survey Satellite*. This represents all *Kepler* planets and candidates which so far have received *TESS* 2min data with sufficient signal-to-noise to allow ephemeris updates to be carried out. The primary challenges to updating more systems were long period planets/candidates falling outside the *TESS* observation window, and systems being too dim in the *TESS* data.

Because of the dimness of these objects, a purely photometric method was used to update these ephemerides, using transits from *Kepler* and *TESS* data only. Transit times for individual transits were recovered by using an MCMC fit of a transit model based on stacked *Kepler* transits. These transit times were then fit a weighted-linear fit in order to obtain the final ephemerides for each object. Any systems which did not appear to have significant transit timing variations were additionally fit with a simultaneous *TESS/Kepler* fit to improve the ephemerides further.

The resulting ephemerides were in good agreement with archival ephemerides and drastically reduced the uncertainty in period for Kepler-411d, Kepler-538b and the candidates K00075.01/K00076.01. Residuals to the linear ephemeris fits gave an important window into transit timing variations for these objects. TTV fits were attempted for five multi-planet systems known to exhibit significant TTVs (Kepler-18, Kepler-25, Kepler-51, Kepler-89, and Kepler-396), however these were challenged by the relative scarcity of reasonable *TESS* transits and the dimness of these systems in *TESS*. In the end there were no significant differences between the TTV masses before and after the inclusion of the *TESS* data. More data is required (preferably of higher signal to noise) in order to constrain the TTV masses more effectively. Interesting TTV behaviour was observed also in the O-C diagrams for HAT-P-7b/Kepler-2b, Kepler-411d, K00075.01 and K00076.01, which warrants further analysis when new data becomes available.

Thus while ephemeris updates and TTV analysis of *Kepler* systems reobserved by *TESS* proved less constraining than originally anticipated, *TESS* has once again proved itself as an important follow-up instrument in addition to its primary planet-finding and asteroseismic aims. Overall, the ephemerides improved and updated in this study extend the life of these *Kepler* systems, and improve prospects for future characterisation. Significant additional benefits are likely if the methods outlined in this paper are extended to the much larger sample of planets and can-

didates reobserved in the *TESS* 30min cadence data, especially as *TESS* returns to the *Kepler* field in its extended mission. Care must be taken however to ensure that the community develops equipment and methods to maintain the ephemerides of dimmer *Kepler* systems, as many of these are simply too dim to be followed up by *TESS*. The challenges faced in this study provide a window into the wider challenge of exoplanet characterisation in the era of all-sky surveys. Carefully deciding how to prioritise targets and maximise scientific return using limited follow-up resources is going to be one of the key challenges in exoplanet science going forward.

Chapter 7

Conclusions

> "It's not enough to know what the future is. You have to know what it means."
>
> — Terry Pratchett & Neil Gaiman, *Good Omens*

7.1 Thesis Summary

Since 1992, the exoplanet community has discovered over 5000 verified exoplanets, with a diversity which far outstrips that seen for planets within the Solar System. This diverse population has revealed that planets are far more dynamic in their lifetimes than was originally anticipated, and may change dramatically between their formation and their final location and composition. This is particularly true in the first billion years of planetary evolution, where some planets have been seen to be rapidly losing their atmospheres or migrating to significantly different orbits. In order to fully explore such evolution, it is critically important to both gain new young exoplanets with well-defined ages to act as snap-shots of exoplanet evolution, and to track known exoplanets in real-time to keep their ephemerides fresh enough for future characterisation. This thesis presented new work on both aspects of this puzzle, based on new observations from the *Transiting Exoplanet Survey Satellite* (*TESS*).

Following an introduction to previous work in this area (chapter 1) and an overview of the methods used throughout (chapter 2), chapters 3-5 presented work towards finding new young exoplanets. Such a task is made significantly more challenging than finding older exoplanets by a comparative lack of known young stars

and the dramatically increased stellar activity of these hosts. Chapter 3 tackled the first part of this problem, using two independent methods to overcome the historical lack of known young stars around which to search for planets, made possible by the high astrometric precision of the *Gaia* satellite launched in 2013. The first method was based on new kinematic analyses of young stellar associations, using *Gaia* data directly to expand the lower mass populations around each association's previously known larger, brighter members. The utilised kinematic methods were validated using the Scorpius OB2 association through comparison to the work of Damiani et al. [2019b], before being used to explore the lower mass membership of the Hyades and Octans associations. A particularly valuable result of this analysis was the selection of 14,050 candidate members of the Octans association, a 100-fold increase in candidate members compared to the bona-fide Octans membership lists compiled previously by Murphy and Lawson [2015] and Gagné et al. [2018b]. A target list detailing these candidates is presented online alongside this thesis. The success of this relatively simple method evidences the transformational power of the *Gaia* satellite for the young star population analyses. However, such graphical kinematic analyses were found to be untenable where the spatial or kinematic extent of the association was too large, such as in the case of the nearby Beta-Pictoris moving group.

Of course, graphical kinematic methods are but one technique used to study the membership of young star-forming regions. The second approach used here made use of the extensive array of young star analyses in literature, using a catalogue-match of recent analyses of open clusters, stellar associations, galactic streams, moving groups and other star forming regions (most of which were similarly based on data from *Gaia* DR2). This resulted in a homogeneous catalogue of 3,036,992 stars from 32 individual catalogues, which continues to expand as new catalogues become available. This list forms the base of present large young exoplanet searches, including several *TESS* data cycle proposals. While continued data releases from the extended *Gaia* mission are likely to extend the lower-mass membership of these groups of young stars further, the large extent of the current catalogue means that a lack of young stars is no longer the significant bottleneck that it was before the *Gaia* era. Consequently the young exoplanet community can now redirect its focus to the larger challenge of disassociating young exoplanetary signals from the increased stellar activity of such young host stars.

Due to rapid leftover rotation from formation and the increased magnetic activity of young stars, photometric time-series of young stars can exhibit stellar activity signals greater in amplitude but similar in period to most transiting

exoplanets, which has significantly challenged traditional transit surveys. In order address this challenge, a new photometric detrending pipeline explicitly designed for young star searches has been assembled in this work and presented in detail in section 2.6. The new pipeline uses locally-weighted scatterplot smoothing (LOWESS) at its base, combined with peak-cutting and interpolation over masked transits to efficiently search for new young exoplanets in the 30 min cadence *TESS* Full-Frame Image (FFI) light-curves. Application of the pipeline to the first 5 sectors of *TESS* data was presented in chapter 4, illustrating the pipeline's excellent performance by detecting all 2 min cadence *TESS* objects of interest at the time of writing from the 30 min data alone, as well as automatically detecting the two previously known young exoplanets DS Tuc Ab and AU Mic b which were previously identified through more targeted eyeballing. A full sensitivity analysis was also undertaken for all young members of BANYAN young associations within these sectors, revealing how recovery of injected planets varied with depth, planetary period and stellar rotation period. Overall the biggest challenges were found to be types of stellar activity with rotation periods of <1 day, exoplanet sizes resulting in radius ratios of $R_P/R_* \leqslant 0.04$, and stellar targets which suffered from significant scattered light on the *TESS* detectors. This pipeline is now being used for extensive searches for new young exoplanet candidates in the continuing *TESS* mission.

In order to detect smaller planets around such young active stars (or those embedded in activity with periods of < 1 day), it is imperative to know more about the stellar activity of each host star. This was the focus of chapter 5, which introduced the wider YOUNGSTER program ("Young Star detrending for Transiting Exoplanet Recovery") and used a Kohonen Self Organising Map (SOM; Kohonen [1982]) to explore the stellar activity and variability of young stars in the in the first year of the *TESS* mission. The Self Organising Map is an unsupervised machine learning method which clusters data based on topology. This algorithm was applied to all young stars from the CDIPS catalogue [Bouma et al., 2019] which received 30 min cadence FFI data in the first year of *TESS*'s primary mission, representing the first time such a technique has been used on a dedicated sample of young stars. This method was found to be particularly effective at separating the signals of young eclipsing binaries and potential transiting objects from other forms of photometric variability. Given the importance of young eclipsing binaries to stellar evolution, a list of those discovered in this work are presented in machine readable format online.[1] Beyond these most obvious points (which appeared as bright yellow dots in

[1] https://warwick.ac.uk/fac/sci/physics/research/astro/people/matthewbattley/ea_and_eb_table_youngster2.csv

the overall SOM), the overall distribution of young star activity topology was found to be much more complex than expected, varying from strictly periodic sinusoids to completely aperiodic dipper and burster-like activity. 'Pre-training' the SOM based on light-curves with known variability classes from the K2 Variability Catalogue of [Armstrong et al., 2016] was also attempted, proving reasonably effective for bright *TESS* targets. However classification of dimmer targets was challenged by the difference in noise properties between *TESS* and Kepler light-curves, so repeating this pre-training step with *TESS* light-curves for each of the pre-classified objects is recommended in the future. An unexpected benefit of the SOM analysis was the effectiveness of the SOM in finding and diagnosing systematic issues in photometric analyses. As explored in section 5.4, because the SOM groups similar shapes together, it provides a fast visual overview of significant leftover systematics and provides a key selection of light-curves with which to test removal methods. The author highly recommends this as a time-saving method in cleaning up data from large-scale planet searches. Overall, the SOM analysis completed here forms the first stage of a wider young star activity/variability classification pipeline, with the view to use this information in the continuing YOUNGSTER programme to perform targeted detrending in exoplanet searches in the future. Such knowledge is anticipated to allow the signals of smaller exoplanets to be detected around noisier young star light-curves.

Detection of exoplanets is only the first part of understanding exoplanetary systems however. The launch of *TESS* also offers the opportunity to trace the evolution of exoplanets many years after their discovery. This is particularly important for planets which are of interest for atmospheric follow-up, or those which display transit timing variations suggestive of additional dynamically interacting planets. The final science chapter of this thesis, chapter 6, revisited the productive *Kepler* field, updating ephemerides for all Kepler planets and candidates which were re-observed with 2min cadence in *TESS*'s primary mission. Because of the reduced resolution of the *TESS* cameras compared to those on the *Kepler* telescope, *TESS* observations were plagued with increased scatter compared to the *Kepler* data, which slightly reduced its ability to follow-up the shallowest planets. Nonetheless, 22 *Kepler* planets and 4 candidates were re-observed with sufficient signal to noise in *TESS* to allow their ephemerides to be updated in this work. The resulting ephemerides are at least seven years 'fresher' than those from *Kepler*, with significantly reduced period uncertainties reported for Kepler-411d, Kepler-538b and candidates K00075.01/K00076.01 in particular. Potential Transit Timing Variations (TTVs) were explored for all targets by building observed minus calculated (O-C) diagrams,

with new TTV fits attempted for five multi-planet systems (Kepler-18, Kepler-25, Kepler-51, Kepler-89 and Kepler-396), however mass updates were hampered by the comparative scarcity and noise of these systems within *TESS*. Meanwhile HAT-P-7b, Kepler-411d, K00075.01 and K00076.01 were all found to exhibit particularly interesting TTV behaviour, suggesting the need for further analysis of these systems as the *TESS* mission continues. The ephemerides updated here thus provide valuable insights into how these systems have evolved (or not) in the past decade, and also improve the prospects of future follow-up by reducing the uncertainty in future transit timing. Now that data is available from *TESS*'s extended mission for much of the ecliptic plane, it would be beneficial to extend the techniques presented in this chapter to all exoplanets found in the K2 mission, as well as those which are present only in the 30 min cadence data.

All of these chapters taken together give the community valuable tools and results to help trace exoplanets through time, as well as providing new glimpses into the environments in which young exoplanets exist. It is anticipated that they will especially aid the discovery and characterisation of young exoplanets. However, much work still remains to elucidate how planets form and evolve through time, which is discussed in the next section.

7.2 Future Outlook

Building a precise timeline of planetary evolution remains one of the biggest challenges in modern exoplanet science. To fully understand the processes causing this evolution, the community needs a population of well characterised planets across precise ages, with a particular focus on the first billion years of planetary evolution.

The key first step in this is the discovery of new young exoplanets, which chapters 3-5 focused on in this thesis. Photometric planet searches have so far proven most prolific at discovering these planets, largely due to the extensive data supplied by the *Kepler* and *TESS* space missions. The photometric detrending pipeline developed here is now being used to search for new transiting young planets, alongside other searches being completed by groups like the THYME team (e.g. Newton et al. [2019]; Rizzuto et al. [2020]; Tofflemire et al. [2021]), NGTS Clusters Survey [Gillen et al., 2020], *TESS* young field stars survey [Zhou et al., 2021] and PATHOS programme [Nardiello et al., 2019, 2021]. The range of different data extraction and detrending methods now employed by the community greatly increases the likelihood of finding young exoplanets compared to at the beginning of this research when only one or two groups had dedicated young star detrending pipelines.

This is particularly important given Hippke et al. [2019]'s analysis which demonstrated that different detrending pipelines were sensitive to slightly different $\sim 30\%$ shares of planetary signals injected into young star light-curves. For this reason new detrending pipelines are always welcome to ensure that any exciting young planets are not missed. It is anticipated that these searches will uncover hundreds of new young planets in the coming decades, especially larger planets.

However, while these pipelines have already shown some success in discovering new young exoplanets, it would be remiss to assume that the photometric detrending and discovery problem has completely been solved. Modern detrending pipelines (including that introduced here) still fail where stellar activity is particularly short-period, swiftly-evolving or hampered by stochastic signals such as granulation. For this reason it is crucial to gain a better understanding of stellar activity and other sources of photometric variability, and use this to work towards more informed removal of red noise in planet searches. This is expected to be the only way most Earth and Super-Earth sized planets will be recoverable in young star light-curves (or indeed in any systems with significant stellar activity signals). The YOUNGSTER programme introduced in chapter 6 represents a step towards this more targeted approach, by first classifying the zoo of young stellar activity and variability into classes with similar topology, noise statistics and evolutionary state. The intention with the continuing programme is to tailor targeted detrending approaches most effective for each individual class of light-curve (ideally based on physical phenomena), such that smaller deviations caused by Earth to Super-Earth sized planets can be recovered. This replaces the traditional 'one size fits most' approach of traditional photometric detrending pipelines, which simply fail for the most challenging cases. In order to build the most effective detrending methods, it will be crucial to understand the physical reasons which cause the observed variability (e.g. star-spots, leftover orbiting debris or granulation). It is thus imperative to foster growing communication and collaboration between solar/stellar physicists, theoretical modellers and exoplanet observers in order to understand the physics behind stellar noise, rather than simply removing it as a nuisance signal. Promising avenues currently being explored include packages such as SOAP [Dumusque et al., 2014] and `starry` [Luger et al., 2019] which allow the user to evaluate the effect of star spots on their observations, multi-colour observations of exoplanets/host stars [Günther et al., 2022] and observing the transit of solar system planets across the solar disc (Doyle, in prep). Crucially, because exoplanetary evolution is tied to the activity of their stellar hosts (especially at early ages), considerably more knowledge could be gained by focusing on exoplanets in difficult, active environments, instead

of the traditional focus on those in less active systems. The knowledge gained from finding these most challenging planets will in turn allow the community to recover smaller planets around less active stars.

Of course the photometric transit method is only one of the methods used to find exoplanets. Both Direct Imaging and the Radial Velocity (RV) method are also key to discovering and characterising young exoplanets. While direct imaging discoveries have mostly been constrained to the earliest ∼50 Myr of stellar evolution due to brightness constraints, RV measurements have the potential to fill in crucial rungs of the exoplanet evolution ladder, alongside providing vital mass measurements for previously discovered transiting exoplanets. However, just as stellar activity plagues the discovery of young exoplanets in photometric data, considerable further work is needed to disassociate young planetary signals from stellar activity in spectroscopic measurements. Indeed due to this activity the first RV mass measurement for a transiting planet in an open cluster (K2-100b) was only managed in 2019 [Barragán et al., 2019b]. Similar to the photometric detrending problem, reliable removal of stellar activity signals from spectroscopic data requires greater knowledge of the physics behind the stellar activity, as highlighted in the final report of the NASA/NSF Extreme Precision Radial Velocity Working Group [Crass, 2021]. The most important physical phenomena suggested by this work were "understanding the Sun in connection to EPRV", "Spectral line formation and behaviour in the stellar atmosphere" and the effect of "faculae/plage[s]". Evaluating the full effect of these phenomena alongside some of the more widely known phenomena such as rotating spots, (super-)granulation and flares will not only allow smaller planets to be found around active young stars, but likely also allow the community to progress past the current challenge of the 1 m/s sensitivity barrier in radial velocity measurements. Indeed, given the recent advances in stability of radial velocity instruments such as ESPRESSO [Pepe et al., 2021] the community is now in the unusual position where the dominant noise for even the quietest stars is now stellar rather than instrumental. Breaking this 1 m/s barrier is fundamental to characterising Earth-like exoplanets, with an Earth analogue imparting a radial velocity amplitude of only 9 cm/s on its host star [Crass, 2021]. Particularly promising methods under development include using spectroscopic activity indicators to model the activity (e.g. Rajpaul et al. [2015]; Barragán et al. [2021]) and mitigating the stellar signal during RV extraction by modelling the CCF and/or selecting individual lines (e.g. Dumusque [2018]; Rajpaul et al. [2020]; Collier Cameron et al. [2021]; Cretignier et al. [2021]). This is supported by 'sun as a star' observations by groups such as Haywood et al. [2016]; Thompson et al. [2020] and Cegla et al. [2016, 2019].

With the ever-expanding population of known exoplanets in the 21st century, the community is steadily shifting towards deeper characterisation of these planets, analysing aspects of their composition, atmospheres and dynamics. To do so requires accurate measurements of their expected transit times and phases, once again highlighting the importance of ephemeris updates such as that completed in chapter 6. As more young exoplanets are discovered, the focus for these planets is also anticipated to shift from discovery-dominated to characterisation-dominated. Overcoming stellar activity in photometric and spectroscopic measurements is absolutely crucial to this effort, as these methods are the best way to obtain precise radii and masses for individual exoplanets. As discussed in section 1.3.3, these two fundamental parameters give a first look at the density, temperature, mass-radius relation and approximate composition of exoplanets, as well as paving the way for key measurements of their atmospheres. Furthermore, these parameters provide key observables on which to base formation and evolution models. As it stands, the in-depth characterisation of young exoplanets is very much in its fledgling stages, largely due to the increased stellar activity of their host stars. While XUV observations and measurements of the Rossiter-McLaughlin effect are starting to give better ideas of the environments and orbits of young exoplanets (e.g. Poppenhaeger et al. [2020]; Zhou et al. [2020]), measurements of their atmospheres have so far proved comparatively fruitless. For example, recent analyses completed by Thao et al. [2020] and Benatti et al. [2021] failed to detect any individual chemical species in the atmospheres of the young planets K2-25b and DS Tuc Ab, despite high quality data from Spitzer and ESPRESSO. There is thus clearly still much work to be done in understanding how stellar activity affects the observed spectra of young exoplanets, and long-term monitoring of interesting planetary systems. One of the key bottle-necks in this extended characterisation is the comparative lack of high precision radial velocity instruments with which to confirm, measure masses for and follow up all discovered planets, with observing time on high-precision instruments like HARPS typically over-subscribed by 5-10 times. In order to make the most of the plethora of ongoing discoveries it is crucial that the community develop additional new radial velocity instruments. Such instruments, combined with ongoing photometric follow-up, will also allow planetary system evolution to be tracked over the course of human timescales like in chapter 6, filling in another important part of the exoplanet timeline.

Another discovery method expected to rapidly come to the forefront in the coming decade is astrometric detection. As has been discussed earlier in this thesis, the extreme-precision astrometric measurements from the first few data releases of

the *Gaia* mission have already transformed the community's knowledge of stellar parameters, positions and kinematics, and has led to the discovery of countless previously unseen groups of stars. However, future data releases (in particular the anticipated Data Releases 3 and 4) are forecast to give astrometric displacement measurements with sufficient precision to detect 21,000 ($\pm$6000) new planets over its nominal 5 year mission [Perryman et al., 2014]. Such a vast number of new planets will allow the community to gain important new insights into planetary formation, exoplanet demographics and dynamic interactions. Particularly powerful will be coupling the astrometric planetary detections with the new groups of stars found in stellar clusters and associations, as this will allow close examination of how planetary dynamics evolves in denser young environments. To prepare for this, it is important that the community as a whole becomes used to interacting with *Gaia* data and more familiar with the potential signals and false-positives of astrometric planetary signals.

One comparatively under-researched aspect of the building the exoplanet timeline is accurately ageing each star system, which is part of what has led to the previous dearth of known young exoplanets discussed in chapter 3. While this thesis followed the approach of most previous young exoplanet discoveries by primarily focusing searches and analyses of known young star forming regions such as stellar clusters and associations (which are easier to age due to the ability to perform full cluster analyses and obtain expansion ages), there is highly likely to be a significant population of young *field* stars which are also young. Because traditional isochronal ageing methods are unreliable at early ages [Soderblom et al., 2014], the community must instead turn to alternate ageing methods such as a measurement of the lithium depletion boundary, surface gravity or rotation rates. An excellent overview of all promising ageing methods for young stars is presented in Soderblom et al. [2014]. Of these, gyrochronolgy is slowly taking the lead in young exoplanet studies because it uses readily available information from transit surveys (the main activity period) and is more constraining than other methods like lithium depletion and Ca II HK emission. Indeed Zhou et al. [2019] recently used the rotation rates of their stars to give the most constraining ages on two new exoplanetary systems around young field stars. However, even this method presents a range of likely ages of several hundreds of Myrs, so more modelling and calibration of stellar ageing methods is needed to build a reliable exoplanet evolution timeline which also includes field stars.

Despite these challenges, the future of exoplanet evolution is bright, with a large number of recently launched and upcoming missions to discover and characterise exoplanets. In particular, the new space satellites CHEOPS [Benz et al., 2021]

and JWST [Gardner et al., 2006] provide much-needed high-precision photometric and spectroscopic follow-up for the most interesting exoplanets, allowing a closer examination of everything from planetary phase-curves to atmospheres. Meanwhile PLATO [Rauer et al., 2014] will couple exoplanetary and stellar analyses to measure planetary radii, stellar age and planet mass to a precision of 3%, 10% and 10% respectively (when combined with spectroscopic follow-up). These are supported by ground-based photometric instruments such as MEarth [Berta et al., 2012] and NGTS [Wheatley et al., 2013], alongside new and upcoming extreme-precision radial velocity instruments like ESPRESSO [Pepe et al., 2010, 2021] and HARPS3 [Thompson et al., 2016b]. Meanwhile direct imaging surveys such as GPI [Graham et al., 2007] and SPHERE [Beuzit et al., 2019] will continue to fill in the youngest end of exoplanetary evolution and future data releases from the *Gaia* mission will soon allow planets to be discovered from their astrometric wobbles alone. The combination of all of these instruments will push down the precision of planetary radii, masses and orbital parameters, alongside providing valuable new data with which to understand and handle stellar activity, both key parts of building a more complete picture of exoplanet evolution.

7.3 Closing Statements

Thirty years ago, exoplanets were merely a concept, with no confirmed signals of planets outside the Solar System. This year the community passed the significant milestone of 5000 verified exoplanets. This paradigm shift has been accompanied by many surprising discoveries, from Jupiter-sized planets orbiting within the equivalent orbit of Mercury, to dense cores of planets with completely stripped atmospheres. Such a plethora of exoplanets coupled with a steady increase in instrumental precision is now allowing the community to explore key questions about how planets form, what they are made of and how they may evolve over their lifetimes. In order to answer these questions, it is imperative to observe exoplanets both through galactic time (through observations of exoplanets at different stages of their evolution) and over human timescales (through repeat observations of known exoplanets). This thesis has presented new work on both of these fronts, including techniques to aid the search for young exoplanets in their period of swiftest evolution, an exploration of the activity of potential young exoplanet hosts for future detrending, and ephemeris updates for known planets to ensure that key discoveries from the Kepler space mission are prepared for follow-up characterisation. The two most important next steps towards a more complete picture of exoplanet evolution

are increasing the community's understanding of stellar activity and extending the number of instruments available for spectroscopic follow-up. Overcoming these two challenges promises not only allow the community to build a timeline of how exoplanets evolve through time, but also offer key new insights into our place in the universe.

Appendix A

Sensitivity Analysis for *TESS* Sectors 1-5

A sector by sector breakdown for the conducted sensitivity analysis is given below in Figures A.1-A.5. Figure format after that explained in Figure 4.11.

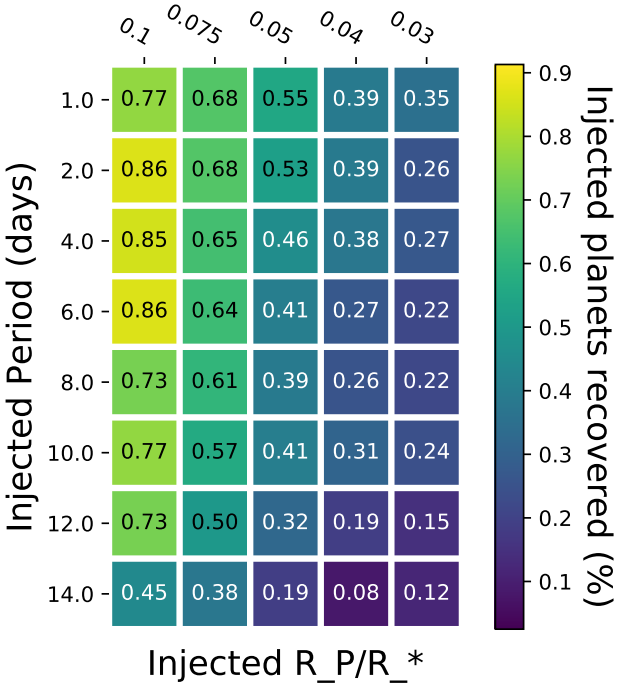

Figure A.1: Sensitivity Analysis for Sector 1. Total number of sources = 74

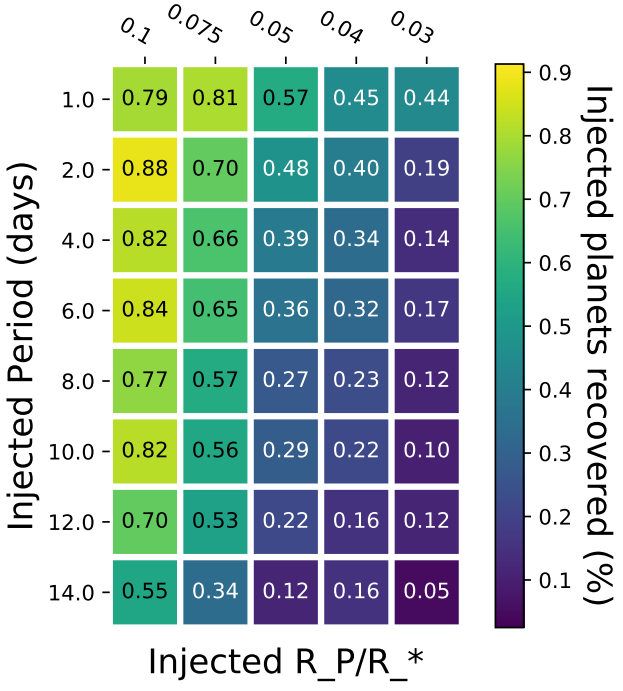

Figure A.2: Sensitivity Analysis for Sector 2. Total number of sources = 77

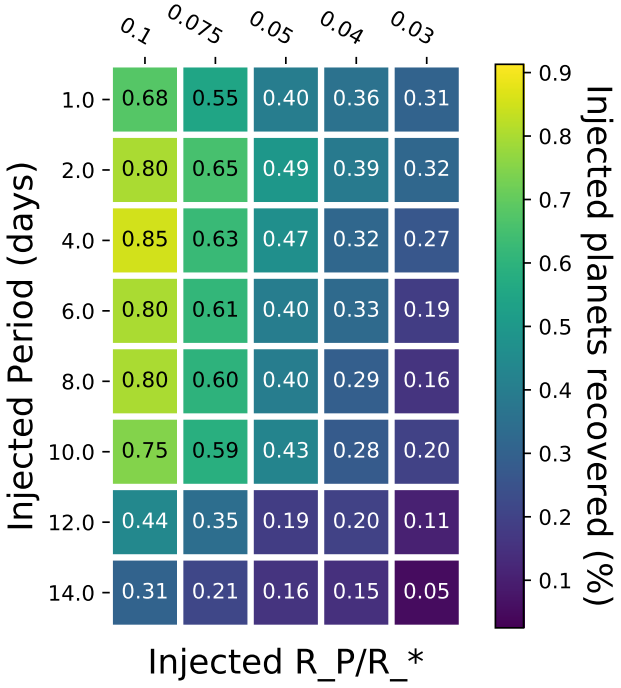

Figure A.3: Sensitivity Analysis for Sector 3. Total number of sources = 75

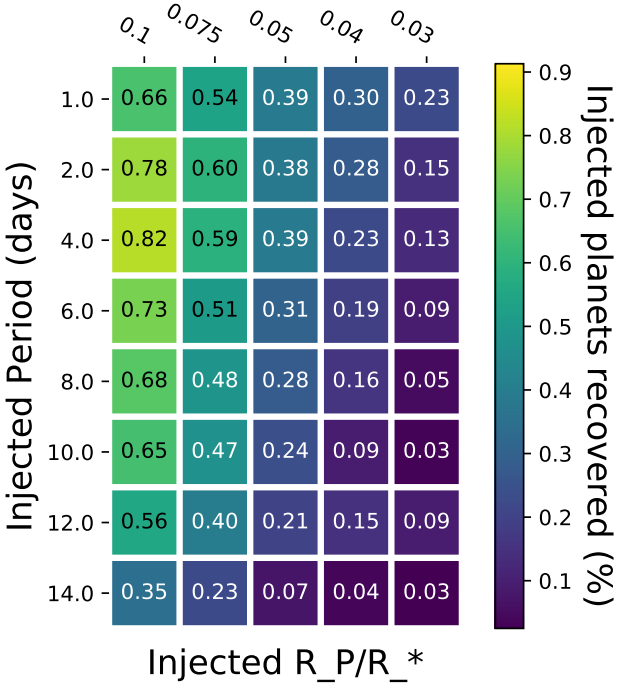

Figure A.4: Sensitivity Analysis for Sector 4. Total number of sources = 120

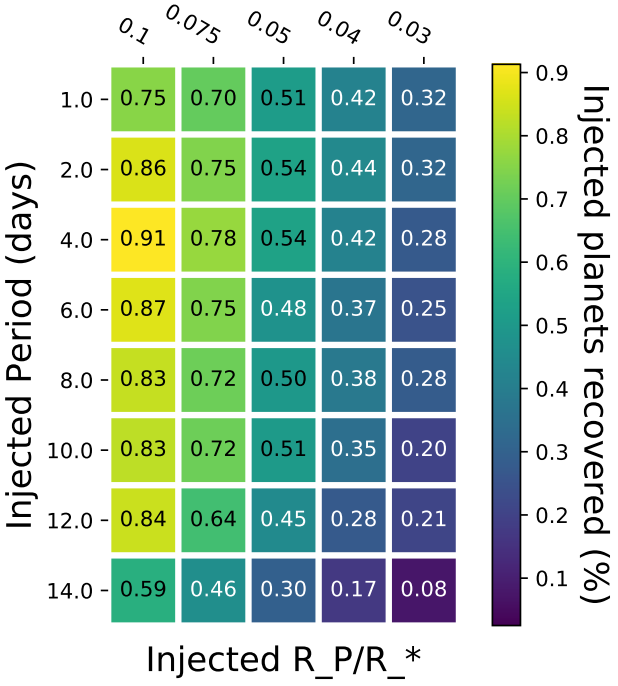

Figure A.5: Sensitivity Analysis for Sector 5. Total number of sources = 138

Appendix B

SOM plots for Sectors 6-13 of the CDIPS light-curves

Self-Organising Maps for sectors 6-13 follow for reference. Notes that the SOM created for Sectors 1-5 (which had too few light-curves to be analysed independently) is shown in Figure 5.3.

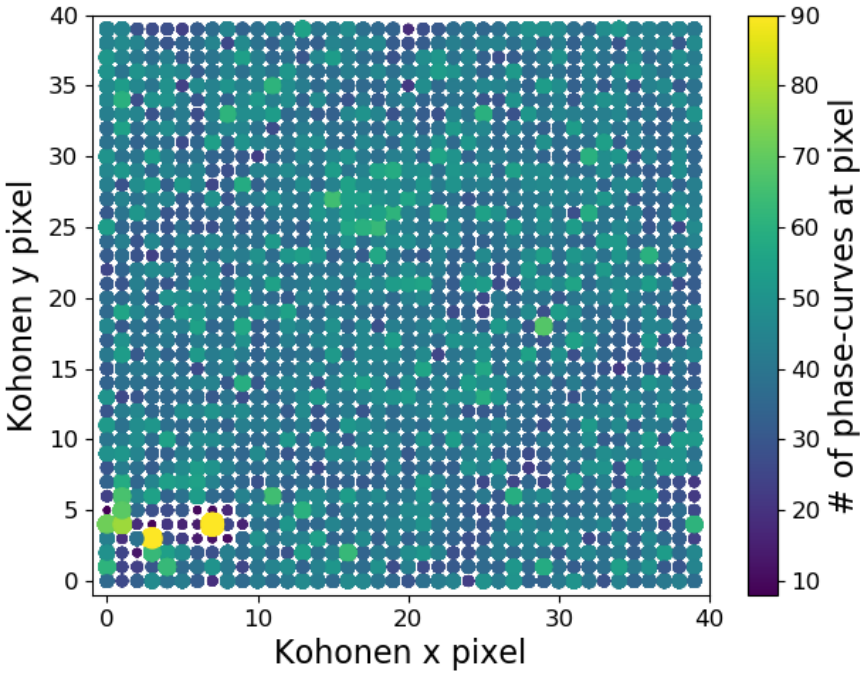

Figure B.1: Sector 6 SOM

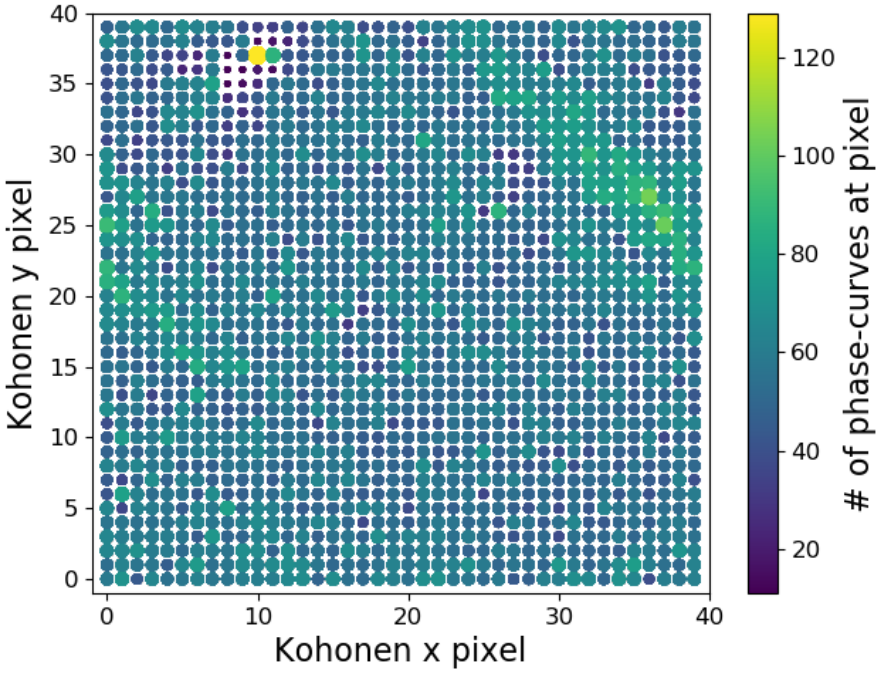

Figure B.2: Sector 7 SOM

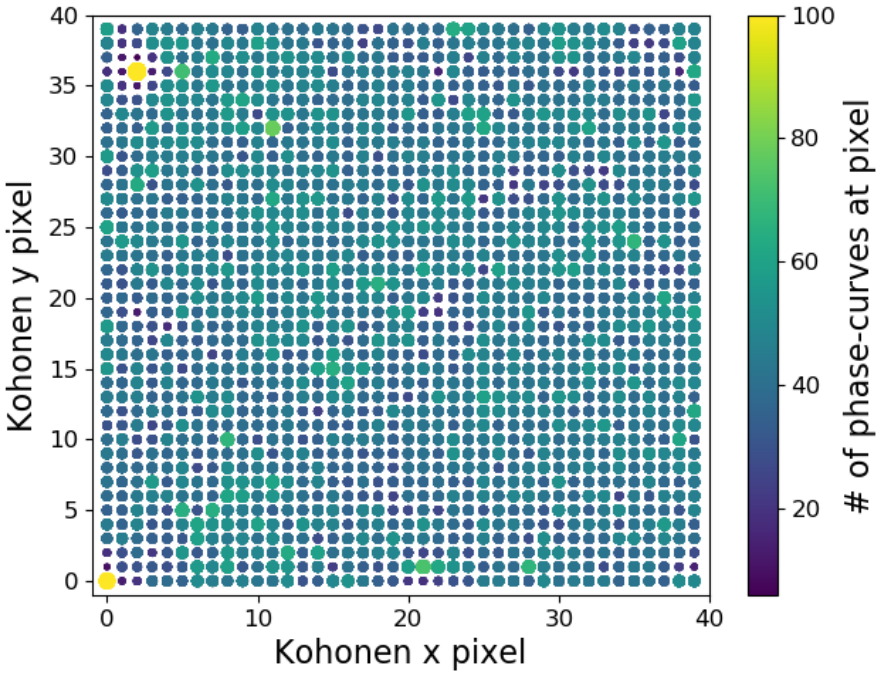

Figure B.3: Sector 8 SOM

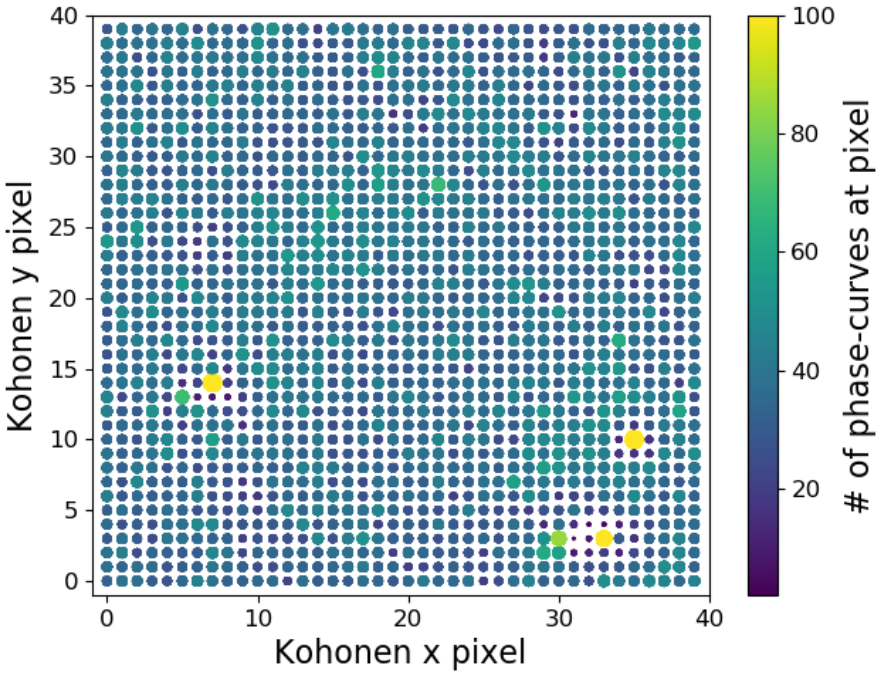

Figure B.4: Sector 9 SOM

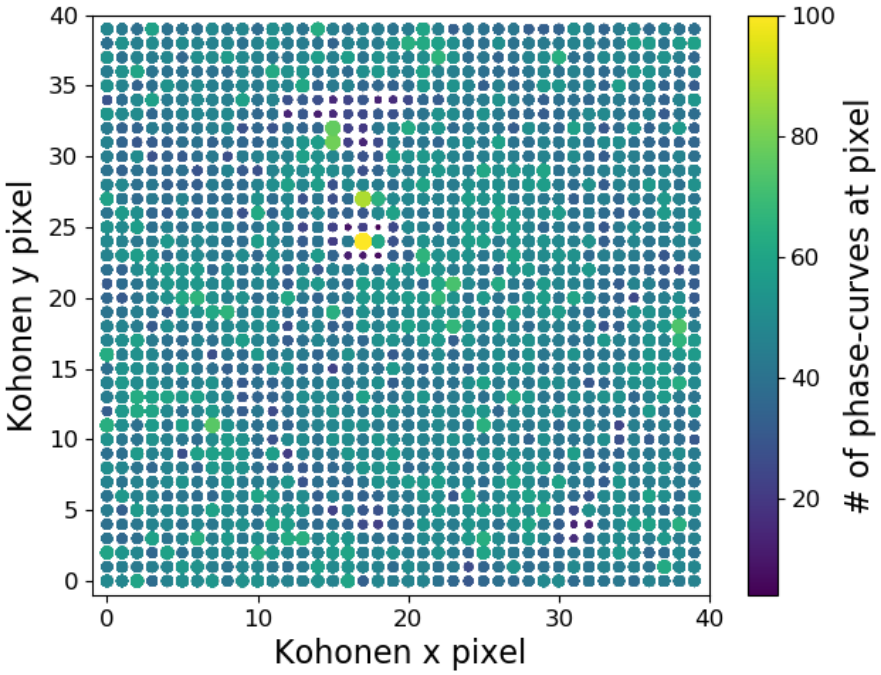

Figure B.5: Sector 10 SOM

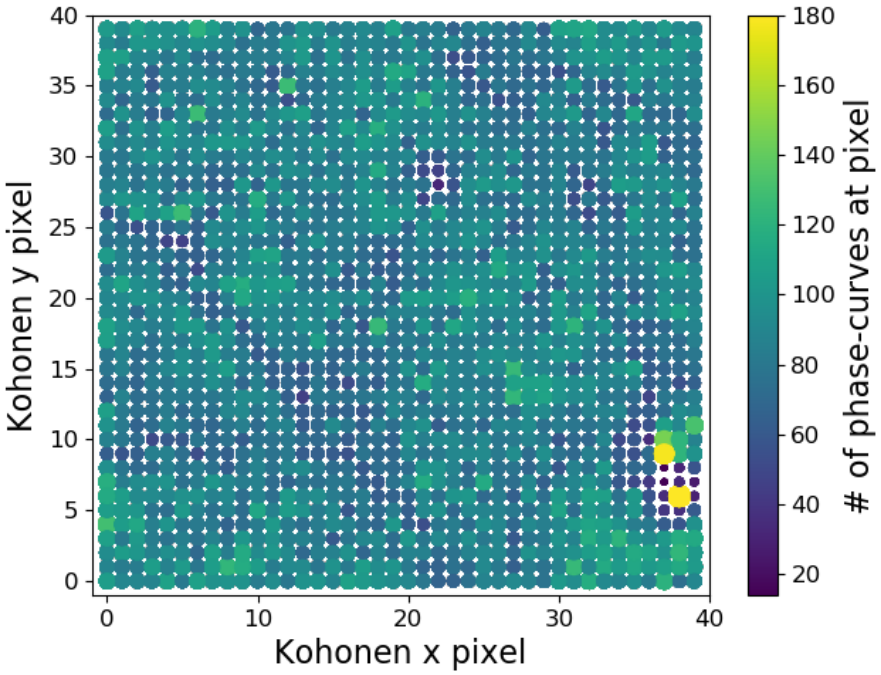

Figure B.6: Sector 11 SOM

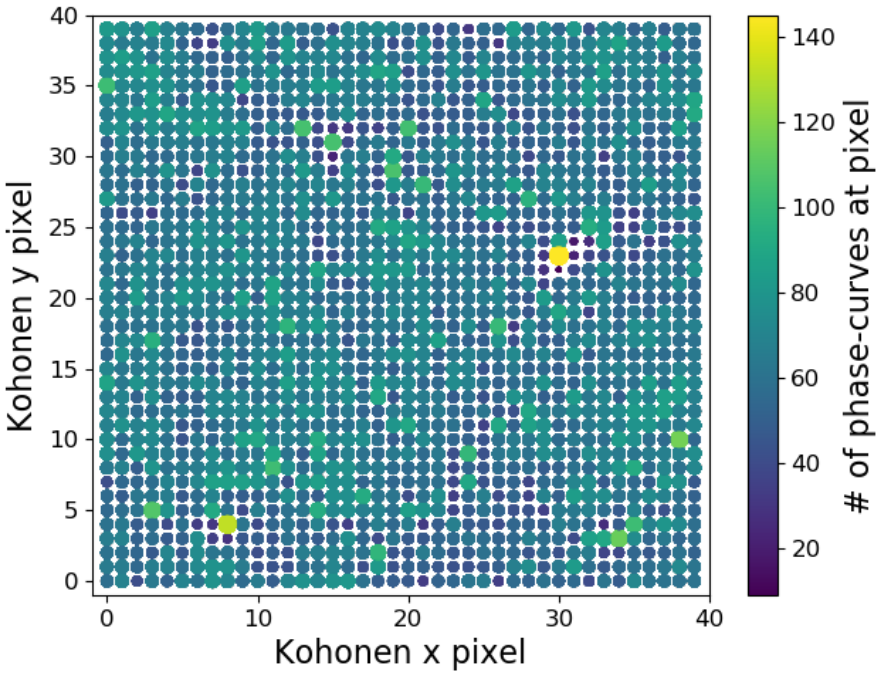

Figure B.7: Sector 12 SOM

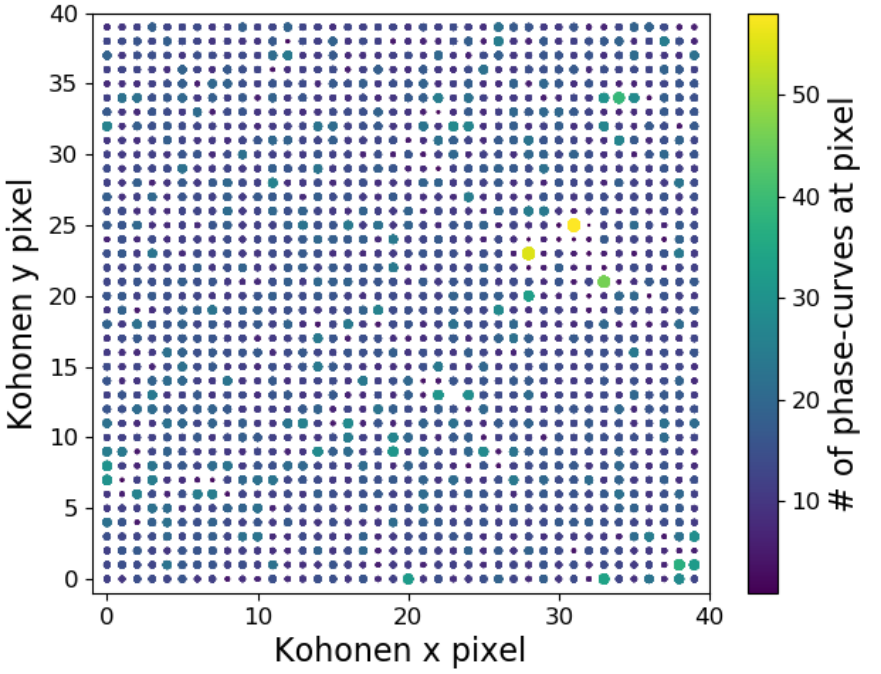

Figure B.8: Sector 13 SOM

Appendix C

Overview of all *Kepler* systems with *TESS* 2min data.

Table C.1 summarises all *Kepler* systems with known transiting planets which received short-cadence (2min) data in *TESS*'s Primary Mission. For each system all known planets are listed, along with whether their ephemerides were updated. For planets which were theoretically reobserved but not updated in this work, a brief explanation is supplied for their exclusion. Most commonly this is because their expected transits fell outside the eventual *TESS* observation window, or they were shallow enough that any potential transits were lost in the increased noise of the *TESS* data.

Table C.1: Overview of all *Kepler* planetary systems which received *TESS* short cadence data during *TESS*'s Primary Mission. N.b. 'outside' = any planets whose expected transit epochs fell outside times of *TESS* observations; 'shallow' = planets with transits which were sufficiently shallow for individual transits to be indiscernible from noise in the *TESS* data.

Kepler Name	TIC ID	Ephemeris Updated?	Reason for exclusion
Kepler-2 b	TIC 424865156	**Yes**	-
Kepler-9 b,c,d	TIC 120571842	No	c outside; b,d shallow
Kepler-10 b,c	TIC 377780790	**Yes** - c, No - b	b shallow
Kepler-11 b,c,d,e,f,g	TIC 169175503	No	All shallow
Kepler-13 b	TIC 158324245	**Yes**	-
Kepler-14 b	TIC 158561566	**Yes**	-
Kepler-16 b	TIC 299096355	No	Outside
Kepler-18 b,c,d	TIC 273690178	**Yes** - d, No b,c	b,c shallow
Kepler-21 b	TIC 121214185	No	Shallow
Kepler-25 b,c,d	TIC 120960812	**Yes** - b,c; No - d	d outside
Kepler-30 b,c,d	TIC 399794329	No	b shallow; c,d outside
Kepler-34 b	TIC 272369124	No	Outside
Kepler-35 b	TIC 271040768	No	Outside
Kepler-36 b,c	TIC 350810590	No	Shallow
Kepler-38 b	TIC 158316612	No	Outside
Kepler-47 b,c,d	TIC 271548206	No	b shallow; c,d outside
Kepler-51 b,c,d	TIC 27846348	**Yes** - b,d; No - c	c shallow
Kepler-63 b	TIC 299158887	**Yes**	-
Kepler-65 b,c,d,e	TIC 121731834	No	b,c,d shallow; e outside
Kepler-68 b,c	TIC 417676622	**Yes** -b; No c	c shallow
Kepler-78 b	TIC 270701667	No	Shallow
Kepler-79 b,c,d,e	TIC 239306681	No	All shallow
Kepler-83 b,c,d	TIC 123416515	No	All shallow
Kepler-89 b,c,d,e	TIC 273231214	**Yes** - d; No - b,c,e	b,c shallow; e outside
Kepler-91 b	TIC 352011875	No	Shallow
Kepler-93 b,c	TIC 137151335	No	b shallow; c outside
Kepler-96 b	TIC 169081296	**Yes**	-
Kepler-122 b,c,d,e	TIC 122714267	No	b,c,d shallow; e outside
Kepler-138 b,c,d	TIC 159376971	No	b,c shallow; d outside
Kepler-289 b,c,d	TIC 273234825	**Yes** - d; No - b,c	b shallow; c outside
Kepler-297 b,c	TIC 48304302	No	Both shallow
Kepler-381 b,c	TIC 164884235	No	Both shallow
Kepler-396 b,c	TIC 27769688	**Yes**	-
Kepler-408 b	TIC 48450369	No	Shallow
Kepler-409 b	TIC 270619260	No	Shallow
Kepler-411 b,c,d	TIC 399954349	**Yes** - c,d; No b	b shallow
Kepler-412 b	TIC 158170594	**Yes**	-
Kepler-413 b	TIC 298969838	No	Outside
Kepler-448 b,c	TIC 169461816	**Yes** - b; No - c	c outside
Kepler-453 b	TIC 164457525	No	No longer transiting
Kepler-462 b	TIC 269263577	No	Shallow
Kepler-508 b	TIC 271671025	No	Shallow
Kepler-538 b	TIC 28227113	**Yes**	-
Kepler-1084 b	TIC 267749737	No	Shallow
Kepler-1244 b	TIC 123447592	No	Shallow
Kepler-1517 b	TIC 158555987	**Yes**	-
Kepler-1647 b	TIC 170344769	No	Outside
Kepler-1661 b	TIC 164886585	No	Outside
PH1 b	TIC 170348142	No	Outside

Bibliography

Adams F.C. and Laughlin G., 2006. *The Astrophysical Journal*, 649(2):1004–1009.

Addison B.C. et al., 2021. *The Astronomical Journal*, 162(6):292. ISSN 0004-6256.

Agol E. et al., 2005. *Monthly Notices of the Royal Astronomical Society*, 359(2):567–579. ISSN 0035-8711.

Aigrain S., Parviainen H., and Pope B.J., 2016. *Monthly Notices of the Royal Astronomical Society*, 459(3):2408–2419. ISSN 13652966.

Aigrain S. et al., 2015. *Monthly Notices of the Royal Astronomical Society*, 447(3):2880–2893. ISSN 13652966.

Akerlof C. et al., 2000. *Astronomical Journal*, 542:251–256.

Alonso R. et al., 2004. *The Astrophysical Journal*, 613(2):L153–L156. ISSN 0004-637X.

Alsubai K. et al., 2017. *The Astronomical Journal*, 153(4):200. ISSN 1538-3881.

Ambartsumian V., 1947. *Stellar Evolution and Astrophysics*. Academy of Sciences of the Armenian SSR, Yerevan, 1 edition.

Angus R. et al., 2018. *Monthly Notices of the Royal Astronomical Society*, 474(2):2094–2108. ISSN 0035-8711.

Armstrong D.J., Gamper J., and Damoulas T., 2020a. *MNRAS*, 504(4):5327–5344.

Armstrong D.J., Pollacco D., and Santerne A., 2017. *Monthly Notices of the Royal Astronomical Society*, 465(3):2634–2642.

Armstrong D.J. et al., 2015. *Astronomy & Astrophysics*, 579(A19).

Armstrong D.J. et al., 2016. *Monthly Notices of the Royal Astronomical Society*, 456(2):2260–2272. ISSN 13652966.

Armstrong D.J. et al., 2020b. *Nature*, 583(7814):39–42.

Armstrong J.J., Wright N.J., and Jeffries R.D., 2018. *Mon. Not. R. Astron. Soc: Letters*, 480(1):L121–L125.

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

Astropy Collaboration et al., 2018. *The Astronomical Journal*, 156(3):123.

Audenaert J. et al., 2021. *Astronomical Journal*, 162(5):209.

Auvergne M. et al., 2009. *Astronomy & Astrophysics*, 506(1):411–424. ISSN 0004-6361.

Baglin A. et al., 2006. In M. Fridlund, A. Baglin, J. Lochard, and L. Conroy, editors, *The CoRoT Mission Pre-Launch Status - Stellar Seismology and Planet Finding*, volume 1306 of *ESA Special Publication*, 33.

Bailer-Jones C.A.L., Irwin M., and Von Hippel T., 1998. *Monthly Notices of the Royal Astronomical Society*, 298(2):361–377. ISSN 0035-8711.

Bailer-Jones C.A.L. et al., 2018. *AJ*, 158:58.

Bakos G. et al., 2002. *Publications of the Astronomical Society of the Pacific*, 114(799):974–987. ISSN 0004-6280.

Bakos G. et al., 2004. *Publications of the Astronomical Society of the Pacific*, 116(817):266–277.

Bakos G. et al., 2010. *Astrophysical Journal*, 710(2):1724–1745. ISSN 15384357.

Bakos G.A. et al., 2008. *Proceedings of the International Astronomical Union*, 4(S253):21–27. ISSN 1743-9213.

Ball N.M. et al., 2006. *The Astrophysical Journal*, 650(1):497–509. ISSN 0004-637X.

Ballard S. et al., 2011. *The Astrophysical Journal*, 743(2):200. ISSN 0004-637X.

Baraffe I. et al., 2003. *Astronomy & Astrophysics*, 402(2):701–712.

Baranne A., Mayor M., and Poncet J.L., 1979. *Vistas in Astronomy*, 23:279–316.

Baranne A. et al., 1996. *Astronomy and Astrophysics Supplement*, 119:373–390.

Barber M.G. et al., 2022. *The Astronomical Journal*, 164(3):id.88.

Barge P. et al., 2008. In Y.S. Sun, S. Ferraz-Mello, and J.L. Zhou, editors, *Exoplanets: Detection, Formation and Dynamics*, volume 249 of *IAU Symposium*, 3–16.

Barman T.S. et al., 2011. *Astrophysical Journal*, 733(1). ISSN 15384357.

Barman T.S. et al., 2015. *The Astrophysical Journal*, 804(1):61. ISSN 1538-4357.

Barnes R., 2017. *Celestial Mechanics and Dynamical Astronomy*, 129(4):509–536. ISSN 0923-2958.

Baron D., 2019. *arXiv e-prints*, arXiv:1904.07248.

Barragán O., Gandolfi D., and Antoniciello G., 2019a. *Monthly Notices of the Royal Astronomical Society*, 482(1). ISSN 13652966.

Barragán O. et al., 2019b. *Monthly Notices of the Royal Astronomical Society*, 490(1):698–708.

Barragán O. et al., 2021. *Research Notes of the AAS*, 5(3).

Barragán O. et al., 2022. *Monthly Notices of the Royal Astronomical Society*, 514(2):1606–1627. ISSN 0035-8711.

Baruteau C. et al., 2014. In *Protostars and Planets VI*. University of Arizona Press.

Batalha N.M. et al., 2011. *Astrophysical Journal*, 729(1):27.

Batalha N.M. et al., 2013. *The Astrophysical Journal Supplement Series*, 204(2):24.

Battley M.P., Armstrong D.J., and Pollacco D., 2022. *Monthly Notices of the Royal Astronomical Society*, 511(3):4285–4304. ISSN 0035-8711.

Battley M.P., Pollacco D., and Armstrong D.J., 2020. *Monthly Notices of the Royal Astronomical Society*, 496(2):1197–1216. ISSN 0035-8711.

Battley M.P. et al., 2021. *Monthly Notices of the Royal Astronomical Society*, 503(3):4092–4104. ISSN 0035-8711.

Beaugé C. and Nesvorný D., 2012. *The Astrophysical Journal*, 751(2):119. ISSN 0004-637X.

Bell C.P., Murphy S.J., and Mamajek E.E., 2017. *Monthly Notices of the Royal Astronomical Society: Letters*, 468(1):1198–1220. ISSN 17453933.

Bellm E.C., 2014. In *Proceedings of the Third Hot-Wiring the Transient Universe Workshop*, 27–33.

Benatti S. et al., 2019. *Astronomy & Astrophysics*, 630:A81.

Benatti S. et al., 2021. *Astronomy & Astrophysics*, 650:A66.

Benz A.O. and Güdel M., 2010. *Annual Review of Astronomy and Astrophysics*, 48(1):241–287. ISSN 0066-4146.

Benz W. et al., 2021. *Experimental Astronomy*, 51(1):109–151. ISSN 0922-6435.

Berger T.A. et al., 2018. *Astrophysical Journal*, 866(2):99.

Berta Z.K. et al., 2012. *The Astronomical Journal*, 144(5):145. ISSN 0004-6256.

Beuzit J.L. et al., 2019. *Astronomy & Astrophysics*, 631:A155. ISSN 0004-6361.

Bhatti W., Bouma L., and Yee S., 2019. waqasbhatti/cdips-pipeline: cdips-pipeline v0.1.0.

Biermann L., 1941. *Vierteljahresschrift der Astronomischen Gesellschaft*, 76:194.

Birkby J.L., 2018. *arXiv e-prints*, arXiv:1806.04617.

Bond I.A. et al., 2001. *Monthly Notices of the Royal Astronomical Society*, 327(3):868–880. ISSN 00358711.

Bond I.A. et al., 2004. *The Astrophysical Journal*, 606(2):L155–L158. ISSN 0004-637X.

Bonomo A.S. et al., 2017. *Astronomy and Astrophysics*, 602:A107. ISSN 14320746.

Borucki W.J. et al., 2010. *Science*, 327(5968):977–980.

Borucki W.J. et al., 2011. *Astrophysical Journal*, 736(1):19.

Boss A.P., 1997. *Science*, 276(5320):1836–1839. ISSN 0036-8075.

Bouchy F. et al., 2009. *Astronomy and Astrophysics*, 505(2):853.

Bouma L.G. et al., 2019. *The Astrophysical Journal Supplement Series*, 245(1):13.

Bouma L.G. et al., 2020. *The Astronomical Journal*, 160(5):239.

Bouma L.G. et al., 2022. *The Astronomical Journal*, 163(3):121. ISSN 0004-6256.

Bourrier V. et al., 2015. *Astronomy & Astrophysics*, 579:A55.

Bouy H. et al., 2015. *Astronomy & Astrophysics*, 577:A148. ISSN 0004-6361.

Breiman L., 2001. *Machine Learning*, 45(1):5–32. ISSN 08856125.

Brems S.S. et al., 2019. *Astronomy & Astrophysics*, 632:A37.

Brett D.R., West R.G., and Wheatley P.J., 2004. *MNRAS*, 353(2):369–376.

Briceno C. et al., 2019. *Astronomical Journal*, 157(2):85. ISSN 1538-3881.

Briegal J.T. et al., 2022. *Monthly Notices of the Royal Astronomical Society*, 513(1):420–438. ISSN 0035-8711.

Brink H. et al., 2013. *Monthly Notices of the Royal Astronomical Society*, 435(2):1047–1060. ISSN 1365-2966.

Broeg C. et al., 2013. *EPJ Web of Conferences*, 47:03005.

Brogi M. et al., 2012. *Nature*, 486(7404):502–504. ISSN 0028-0836.

Brown A., 1950. *The Astrophysical Journal*, 112:225.

Buchhave L.A. et al., 2011. *Astrophysical Journal Supplement*, 197(1):3.

Buchhave L.A. et al., 2012. *Nature*, 486(7403):375–377.

Burke C.J. et al., 2020. TESS-Point: High precision TESS pointing tool.

Burnap P. et al., 2018. *Computers & Security*, 73. ISSN 01674048.

Caceres G.A. et al., 2019. *The Astronomical Journal*, 158(2):57. ISSN 0004-6256.

Caldwell D.A. et al., 2020a. *Research Notes of the American Astronomical Society*, 4(11):201.

Caldwell D.A. et al., 2020b. *Research Notes of the American Astronomical Society*, 4(11):201.

Campbell B., Walker G.A.H., and Yang S., 1988. *The Astrophysical Journal*, 331:902. ISSN 0004-637X.

Cantat-Gaudin T. et al., 2018. *A&A*, 618:A93.

Cantat-Gaudin T. et al., 2019a. *Astronomy & Astrophysics*, 626:A17. ISSN 0004-6361.

Cantat-Gaudin T. et al., 2019b. *Astronomy & Astrophysics*, 624:A126. ISSN 0004-6361.

Cantat-Gaudin T. et al., 2020. *Astronomy & Astrophysics*, 640:A1. ISSN 0004-6361.

Carleo I. et al., 2021. *Astronomy & Astrophysics*, 645:A71. ISSN 0004-6361.

Carrasco Kind M. and Brunner R.J., 2014. *Monthly Notices of the Royal Astronomical Society*, 438(4):3409–3421. ISSN 00358711.

Carroll B.W. and Ostlie D.A., 2017. *An Introduction to Modern Astrophysics*. Cambridge University Press, 2 edition. ISBN 9781108422161.

Cegla H., 2020. Physics Responsible for Various Sources of Stellar Variability.

Cegla H.M. et al., 2016. *Astronomy & Astrophysics*, 588:A127. ISSN 0004-6361.

Cegla H.M. et al., 2019. *The Astrophysical Journal*, 763(2):95–102.

Chabrier G., 2003. *Publications of the Astronomical Society of the Pacific*, 115(809):763–795. ISSN 0004-6280.

Charbonneau D. et al., 2000. *The Astrophysical Journal*, 529(1):L45–L48. ISSN 0004-637X.

Charbonneau D. et al., 2002. *The Astrophysical Journal*, 568(1):377–384. ISSN 0004-637X.

Chatterjee S. et al., 2008. *The Astrophysical Journal*, 686(1):580–602. ISSN 0004-637X.

Chauvin G. et al., 2004. *Astronomy & Astrophysics*, 425:L29–L32.

Chauvin G. et al., 2015. *Astronomy & Astrophysics*, 573:A127. ISSN 0004-6361.

Christ C.N., Montet B.T., and Fabrycky D.C., 2018. *Astronomical Journal*, 157(6):235.

Ciardi D.R. et al., 2018. *The Astronomical Journal*, 155(1):10.

Cleveland W.S., 1979a. *Journal of the American Statistical Association*, 74(368):829. ISSN 01621459.

Cleveland W.S., 1979b. *Journal of the American Statistical Association*, 74(368):829–836. ISSN 1537274X.

Cochran W.D. et al., 2011. *Astrophysical Journal Supplement*, 197(1):7.

Cody A.M. and Hillenbrand L.A., 2018. *The Astronomical Journal*, 156(2):71. ISSN 1538-3881.

Cody A.M. et al., 2018. *Research Notes of the AAS*, 2(4):199.

Collier Cameron A. et al., 2006. *Monthly Notices of the Royal Astronomical Society*, 373(2):799–810. ISSN 00358711.

Collier Cameron A. et al., 2021. *Monthly Notices of the Royal Astronomical Society*, 505(2):1699–1717. ISSN 0035-8711.

Costes J.C. et al., 2020. *Monthly Notices of the Royal Astronomical Society*, 491(2):2834–2844.

Crass J., 2021. Extreme Precision Radial Velocity Working Group Final Report. Technical report, NASA-NSF.

Cretignier M. et al., 2020. *Astronomy & Astrophysics*, 633:A76.

Cretignier M. et al., 2021. *Astronomy & Astrophysics*, 653:A43.

Cunningham T. et al., 2022. *Nature*, 602(7896):219–222. ISSN 0028-0836.

Damiani F. et al., 2019a. *A&A*, 623:A25.

Damiani F. et al., 2019b. *A&A*, 623:A112.

David T.J. et al., 2016a. *Nature*, 534(7609):658–661. ISSN 14764687.

David T.J. et al., 2016b. *Nature*, 534(7609):658–661. ISSN 0028-0836.

David T.J. et al., 2018a. *The Astronomical Journal*, 156(6):302. ISSN 1538-3881.

David T.J. et al., 2018b. *The Astronomical Journal*, 155(5):222. ISSN 1538-3881.

David T.J. et al., 2019a. *The Astronomical Journal*, 158(2):79. ISSN 0004-6256.

David T.J. et al., 2019b. *The Astronomical Journal*, 158(2):79. ISSN 0004-6256.

David T.J. et al., 2019c. *Astrophysical Journal Letters*, 885(1). ISSN 20418213.

de Zeeuw P.T. et al., 1999. *The Astronomical Journal*, 117(1):354–399.

Debosscher J. et al., 2007. *Astronomy & Astrophysics*, 475(3):1159–1183. ISSN 0004-6361.

Debosscher J. et al., 2011. *Astronomy & Astrophysics*, 529(A89):A89. ISSN 0004-6361.

Deck K. et al., 2014. TTVFast: Transit timing inversion.

Deeming T.J., 1964. *Monthly Notices of the Royal Astronomical Society*, 127(6):493–516. ISSN 0035-8711.

Deleuil M. et al., 2014. *Astronomy & Astrophysics*, 564:A56–undefined.

Delrez L. et al., 2021. *Nature Astronomy*, 5(8):775–787. ISSN 2397-3366.

Dias W.S. et al., 2014. *Astronomy and Astrophysics*, 564:A79. ISSN 14320746.

Dickson-Vandervelde D.A., Wilson E.C., and Kastner J.H., 2021. *The Astronomical Journal*, 161(2):87. ISSN 0004-6256.

Dodson-Robinson S.E. et al., 2009. *The Astrophysical Journal*, 707(1):79–88. ISSN 0004-637X.

Dong J. et al., 2022. *The Astrophysical Journal Letters*, 926(2):L7. ISSN 2041-8205.

Dorn C. et al., 2017. *Astronomy & Astrophysics*, 597:A37. ISSN 0004-6361.

Dragomir D. et al., 2020. *Astronomical Journal*, 159(5).

Dressing C.D. and Charbonneau D., 2015. *Astrophysical Journal*, 807(1). ISSN 15384357.

Dressing C.D. et al., 2015. *The Astrophysical Journal*, 800(2):135. ISSN 1538-4357.

Dumusque X., 2018. *Astronomy and Astrophysics*, 620. ISSN 14320746.

Dumusque X., Boisse I., and Santos N.C., 2014. *Astrophysical Journal*, 796(2). ISSN 15384357.

Edwards B. et al., 2020. *MNRAS*, staa1245.

Eigmüller P. et al., 2019. *Astronomy & Astrophysics*, 625:A142.

Einstein A., 1936. *Science*, 84(2188):506–507.

Espinoza N. et al., 2019. *The Astronomical Journal*, 158(2):63. ISSN 1538-3881.

Esplin T.L. and Luhman K.L., 2019. *The Astronomical Journal*, 158(2):54.

Eyer L. and Blake C., 2005. *Monthly Notices of the Royal Astronomical Society*, 358(1):30–38. ISSN 0035-8711.

Eyer L. et al., 2019. *Astronomy and Astrophysics*, 623(A110). ISSN 14320746.

Faherty J.K. et al., 2017. *A&A*, 598:A48.

Faisst A.L. et al., 2019. *Astrophysical Journal Letters*, 881(1):L9.

Fausnaugh M.M. et al., 2020. *Research Notes of the AAS*, 4(12):251.

Feinstein A.D. et al., 2019. *PASP*, 131(1003):094502.

Feinstein A.D. et al., 2022. *The Astrophysical Journal Letters*, 925:L9.

Ford E.B. and Holman M.J., 2007. *The Astrophysical Journal*, 664(1):L51–L54. ISSN 0004-637X.

Foreman-Mackey D. et al., 2012. *PASP*, 125(925):306. ISSN 00046280.

Foreman-Mackey D. et al., 2020. exoplanet-dev/exoplanet v0.4.3.

Fortney J.J., Marley M.S., and Barnes J.W., 2007. *The Astrophysical Journal*, 659(2):1661–1672. ISSN 0004-637X.

Fressin F. et al., 2011. *Astrophysical Journal Supplement*, 197(1):5.

Fressin F. et al., 2013. *Astrophysical Journal*, 766(2). ISSN 15384357.

Friedman J.H., 2001. *The Annals of Statistics*, 29(5). ISSN 0090-5364.

Fujita N., Hori Y., and Sasaki T., 2022. *arXiv e-prints*, arXiv:2202.06585.

Fulton B.J. et al., 2017. *The Astronomical Journal*, 154(3):109. ISSN 1538-3881.

Fustes D. et al., 2013. *Astronomy & Astrophysics*, 559:A7. ISSN 0004-6361.

Gaensicke B.T. et al., 2019. *Nature*, 576(7785):61–64.

Gagné J. and Faherty J.K., 2018. *The Astrophysical Journal*, 862(2):138.

Gagné J., Faherty J.K., and Mamajek E.E., 2018a. *The Astrophysical Journal*, 865(2):136.

Gagné J. et al., 2018b. *ApJS*, 856(1).

Gagné J. et al., 2018c. *The Astrophysical Journal*, 860(1):43.

Gagné J. et al., 2020. *The Astrophysical Journal*, 903(2):96. ISSN 0004-637X.

Gaia Collaboration et al., 2016a. *A&A*, 595(A2):A2.

Gaia Collaboration et al., 2016b. *Astronomy & Astrophysics*, 595:A1.

Gaia Collaboration et al., 2018a. *Astronomy and Astrophysics*, 616(A1):A1.

Gaia Collaboration et al., 2018b. *Astronomy and Astrophysics*, 616:A10. ISSN 14320746.

Gaidos E. et al., 2020a. *Monthly Notices of the Royal Astronomical Society*, 495(1):650–662.

Gaidos E. et al., 2020b. *Monthly Notices of the Royal Astronomical Society*, 498(1):L119–L124.

Gajdos P., Vaňko M., and Parimucha S., 2019. *RAA*, 19(3):41.

Gajdoš P. et al., 2017. *Monthly Notices of the Royal Astronomical Society: Letters*, 469(3):2907–2912. ISSN 17453933.

Galli P.A.B. et al., 2020. *Astronomy & Astrophysics*, 643:A148. ISSN 0004-6361.

Galvin T.J. et al., 2020. *MNRAS*, 497(3):2730–2758.

Gardner J.P. et al., 2006. *Space Science Reviews*, 123(4):485–606. ISSN 1572-9672.

Geach J.E., 2012. *Monthly Notices of the Royal Astronomical Society*, 419(3):2633–2645. ISSN 00358711.

Giacalone S. et al., 2022. *The Astronomical Journal*, 163(2):99. ISSN 0004-6256.

Gianniotis N. et al., 2015. *arXiv eprints*, arXiv:1505.00936.

Gilbert E.A. et al., 2022. *The Astronomical Journal*, 163(4):147. ISSN 0004-6256.

Giles D. and Walkowicz L., 2019. *Monthly Notices of the Royal Astronomical Society*, 484(1):834–849. ISSN 0035-8711.

Giles H.A.C. et al., 2018. *Monthly Notices of the Royal Astronomical Society*, 475(2):1809–1818. ISSN 0035-8711.

Gillen E. et al., 2020. *Monthly Notices of the Royal Astronomical Society*, 492(1):1008–1024. ISSN 13652966.

Gilliland R.L. et al., 2011. *Astrophysical Journal, Supplement Series*, 197(1). ISSN 00670049.

Gilliland R.L. et al., 2013. *Astrophysical Journal*, 766(1):40.

Ginzburg S., Schlichting H.E., and Sari R., 2018. *Monthly Notices of the Royal Astronomical Society*, 476(1):759–765. ISSN 0035-8711.

Goldberg M. et al., 2018. *Astronomical Journal*, 157(4):152.

Graham J.R. et al., 2007. *arXiv eprint*, arXiv:0704.1454.

Guillot T., 2005. *Annual Review of Earth and Planetary Sciences*, 33(1):493–530. ISSN 0084-6597.

Guillot T. et al., 2014. In *EAS Publications Series*. ISBN 9782759816842. ISSN 16381963.

Günther M.N. et al., 2022. *The Astronomical Journal*, 163:144.

Gupta A. and Schlichting H.E., 2019. *Monthly Notices of the Royal Astronomical Society*, 487(1):24–33. ISSN 0035-8711.

Hadden S. and Lithwick Y., 2016. *Astrophysical Journal*, 828(1):44.

Hadden S. and Lithwick Y., 2017. *Astronomical Journal*, 154(1):5.

Handberg R. and Lund M.N., 2019. T'DA Data Release Notes: Data Release 4 for TESS Sectors 1+2. Technical report, TASOC.

Handberg R. et al., 2020. TESS Data for Asteroseismology: Photometry. Technical report, TASOC.

Hartman J.D. and Bakos G., 2016. *Astronomy and Computing*, 17:1–72. ISSN 22131337.

Hartman J.D. et al., 2019. *The Astronomical Journal*, 157(2):55. ISSN 1538-3881.

Haswell C.A., 2010. *Transiting Exoplanets*. Cambridge University Press, Cambridge, 1 edition. ISBN 978-0-521-13938-0.

Haywood R.D. et al., 2016. *Monthly Notices of the Royal Astronomical Society*, 457(4):3637–3651. ISSN 0035-8711.

Hébrard G. et al., 2011. *Astronomy & Astrophysics*, 533:A130. ISSN 0004-6361.

Hedges C., Hodgkin S., and Kennedy G., 2018. *MNRAS*, 476(3):2968–2998.

Hedges C. et al., 2021. *The Astronomical Journal*, 162(2):54. ISSN 0004-6256.

Heller R., Rodenbeck K., and Hippke M., 2019. *Astronomy & Astrophysics*, 627:A66.

Hellier C. et al., 2019. *Monthly Notices of the Royal Astronomical Society*, 490(1):1479–1487. ISSN 0035-8711.

Henry G.W. et al., 1999. In *IAU Cicrcular, No. 7307*, volume 7307, 1.

Herbig G., 1962. In *Advances in Astronomy and Astrophysics*, volume 1, 47–103. Lick Observatory, University of California, California.

Hippke M. and Heller R., 2019. *Astronomy & Astrophysics*, 623:A39.

Hippke M. et al., 2019. *The Astronomical Journal*, 158(4):143. ISSN 1538-3881.

Holczer T. et al., 2016. *The Astrophysical Journal Supplement Series*, 225(1):9.

Holman M.J. and Murray N.W., 2005. *Science*, 307(5713):1288–1291. ISSN 0036-8075.

Holman M.J. et al., 2010. *Science*, 330(6000):51.

Hon M., Stello D., and Zinn J.C., 2018. *The Astrophysical Journal*, 859(1):64. ISSN 1538-4357.

Howard A.W. et al., 2012. *The Astrophysical Journal Supplement Series*, 201(2):15. ISSN 0067-0049.

Howell S.B. et al., 2012. *Astrophysical Journal*, 746(2):123.

Howell S.B. et al., 2014. *PASP*, 126:938.

Huang C.X. et al., 2018. *The Astrophysical Journal Letters*, 868(2):L39.

Huang C.X. et al., 2020a. *Research Notes of the AAS*, 4(11):204.

Huang C.X. et al., 2020b. *Research Notes of the AAS*, 4(11):206. ISSN 2515-5172.

Ida S. and Lin D.N., 2010. *Astrophysical Journal*, 719(1):810–830. ISSN 15384357.

Ida S., Lin D.N., and Nagasawa M., 2013. *Astrophysical Journal*, 775(1). ISSN 15384357.

Ikwut-Ukwa M. et al., 2020. *AAS Journals*, 160:209.

Irwin M.J., 1985. *Monthly Notices of the Royal Astronomical Society*, 214(4):575–604. ISSN 0035-8711.

Ivezic Z. et al., 2019. *Statistics, Data Mining, and Machine Learning in Astronomy*. Princeton University Press.

Jackman J.A.G. et al., 2019. *Monthly Notices of the Royal Astronomical Society*. ISSN 0035-8711.

Jehin E. et al., 2011. *The Messenger*, 145:2–6.

Jenkins J.M., 2002. *The Astrophysical Journal*, 575(1):493–505. ISSN 0004-637X.

Jenkins J.M. et al., 2010. *Astrophysical Journal Letters*, 713(2 PART 2). ISSN 20418213.

Jenkins J.M. et al., 2016. In *Software and Cyberinfrastructure for Astronomy IV*, volume 9913, 99133E. ISBN 9781510602052. ISSN 1996756X.

Johnston H. et al., 2021. *Astronomy & Astrophysics*, 648. ISSN 0004-6361.

Jones D.H.P., 1971. *Monthly Notices of the Royal Astronomical Society*, 152:231–259.

Jontof-Hutter D. et al., 2021. *arXiv eprints*, arXiv:2101.01202.

Kabáth P. et al., 2022. *Monthly Notices of the Royal Astronomical Society*. ISSN 0035-8711.

Kasting J.F., Whitmire D.P., and Reynolds R.T., 1993. *Icarus*, 101(1):108–128. ISSN 00191035.

Kennedy G.M. and Kenyon S.J., 2008. *The Astrophysical Journal*, 682(2):1264–1276. ISSN 0004-637X.

Khacef L., Gripon V., and Miramond B., 2020. *preprint (arXiv: 2009.03665)*, arXiv:2009.03665.

Kharchenko N.V. et al., 2013. *Astronomy and Astrophysics*, 558:A53. ISSN 00046361.

Khodachenko M.L. et al., 2015. *The Astrophysical Journal*, 813(1):50. ISSN 1538-4357.

Kim D.W. and Bailer-Jones C.A.L., 2016. *Astronomy & Astrophysics*, 587:A18. ISSN 0004-6361.

Kim D.W. et al., 2009. *Monthly Notices of the Royal Astronomical Society*, 397(1):558–568. ISSN 00358711.

Kim S.L. et al., 2016. *Journal of Korean Astronomical Society*, 49(1):37–44.

Kipping D. et al., 2014. *Astrophysical Journal*, 784(1):28.

Kipping D.M., 2009. *Monthly Notices of the Royal Astronomical Society*, 392(1):181–189. ISSN 00358711.

Koch D.G. et al., 2010. *The Astrophysical Journal*, 713(2):L79–L86.

Koeltzsch A. et al., 2009. *Astronomische Nachrichten*, 330:482.

Kohonen T., 1982. *Biol. Cybern*, 43:59–69.

Kohonen T., 2001. *Self-Organizing Maps*. Springer, New York, 3rd edition. ISBN 3540679219.

Kokori A. et al., 2020. *arXiv e-prints*, arxiv: 2012.07478.

Kopparapu R.K. et al., 2013. *The Astrophysical Journal*, 765(2):131. ISSN 0004-637X.

Korhonen H. et al., 2015. *Monthly Notices of the Royal Astronomical Society*, 448(4):3038–3052.

Kounkel M. and Covey K., 2019. *The Astronomical Journal*, 158(3):122. ISSN 1538-3881.

Kounkel M. et al., 2018. *The Astronomical Journal*, 156(3):84. ISSN 1538-3881.

Kovács G., Bakos G., and Noyes R.W., 2005. A trend filtering algorithm for wide-field variability surveys.

Kovács G., Zucker S., and Mazeh T., 2002. *A&A*, 391(1).

Kozai Y., 1962. *The Astronomical Journal*, 67:591. ISSN 00046256.

Kraus A.L. et al., 2014. *Astronomical Journal*, 147(6):146. ISSN 00046256.

Kreidberg L., 2015. *Publications of the Astronomical Society of the Pacific*, 127:1161–1165.

Krone-Martins A. and Moitinho A., 2014. *Astronomy & Astrophysics*, 561:A57. ISSN 0004-6361.

Kuhn M.A. et al., 2019. *The Astrophysical Journal*, 870(1):32. ISSN 1538-4357.

Kuhn M.A. et al., 2021. *The Astrophysical Journal Supplement Series*, 254(2):33. ISSN 0067-0049.

Kunimoto M. et al., 2021. *Research Notes of the AAS*, 5(10):234. ISSN 2515-5172.

Lagrange A.M. et al., 2009. *Astronomy & Astrophysics*, 493(2):L21–L25. ISSN 0004-6361.

Lannier J. et al., 2016. *Astronomy & Astrophysics*, 596:A83. ISSN 0004-6361.

Lee C.H., 2018. *Galaxies*, 6:51.

Léger A. et al., 2009. *Astronomy & Astrophysics*, 506(1):287–302. ISSN 0004-6361.

Leleu A. et al., 2021. *Astronomy & Astrophysics*, 649:A26. ISSN 0004-6361.

Lendl M. et al., 2020. *Astronomy & Astrophysics*, 643:A94. ISSN 0004-6361.

Lidov M., 1962. *Planetary and Space Science*, 9(10):719–759. ISSN 00320633.

Lightkurve Collaboration et al., 2018. Lightkurve: Kepler and TESS time series analysis in Python. Astrophysics Source Code Library.

Lin D.N.C. and Papaloizou J., 1986. *The Astrophysical Journal*, 309:846. ISSN 0004-637X.

Lindegren L. et al., 2018. *Astronomy & Astrophysics*, 616:A2. ISSN 0004-6361.

Lissauer J.J., 1993. *Annual Review of Astronomy and Astrophysics*, 31(1):129–172. ISSN 0066-4146.

Lithwick Y., Xie J., and Wu Y., 2012. *The Astrophysical Journal*, 761(2):122. ISSN 0004-637X.

Livingston J.H. et al., 2019. *AJ*, 157(3):102.

Lodieu N. et al., 2019. *Astronomy & Astrophysics*, 623:A35.

Lomb N., 1976. *Astrophysics and Space Science*, 39:447–462.

Lopez E.D. and Fortney J.J., 2013. *The Astrophysical Journal*, 776(1):2. ISSN 0004-637X.

Lovis C. and Fischer D.A., 2010. In Sara Seager, editor, *Exoplanets*, chapter Radial Velocity, 27–53. University of Arizona Press, Tucson, 1 edition.

Luger R. et al., 2016. *The Astronomical Journal*, 152(4):100.

Luger R. et al., 2017. *Nature Astronomy*, 1:0129.

Luger R. et al., 2019. *The Astronomical Journal*, 157(2):64.

Luhman K.L., 2018. *The Astronomical Journal*, 156(6):271. ISSN 1538-3881.

Luhman K.L., 2020. *The Astronomical Journal*, 160(4):186. ISSN 0004-6256.

Luhman K.L. and Esplin T.L., 2020. *The Astronomical Journal*, 160(1):44. ISSN 1538-3881.

Lund M.N. et al., 2015. *The Astrophysical Journal*, 806(1):30.

Macintosh B. et al., 2014. *Proceedings of the National Academy of Sciences*, 111(35):12661–12666. ISSN 0027-8424.

MacQueen J., 1967. In *Proceedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability*, 281–297. University of California Press.

Madhusudhan N., 2018. In *Handbook of Exoplanets*, 2153–2182. Springer International Publishing, Cham.

Madhusudhan N., 2019. *Annual Review of Astronomy and Astrophysics*, 57(1):617–663. ISSN 0066-4146.

Malavolta L. et al., 2016. *Astronomy and Astrophysics*, 588. ISSN 14320746.

Malo L. et al., 2013. *Astrophysical Journal*, 762(2):88.

Mamajek E.E., 2005. *ApJ*, 634:1385–1394.

Manara C.F. et al., 2019. *Astronomy & Astrophysics*, 631:L2. ISSN 0004-6361.

Mandel K. and Agol E., 2002. *Astrophysical Journal*, 580:L171–L176.

Mann A.W. et al., 2016a. *ApJ*, 818(1):46.

Mann A.W. et al., 2016b. *The Astronomical Journal*, 152(3):61.

Mann A.W. et al., 2017a. *Astronomical Journal*, 153(2). ISSN 00046256.

Mann A.W. et al., 2017b. *The Astronomical Journal*, 155(1):4. ISSN 1538-3881.

Mann A.W. et al., 2017c. *Astronomical Journal*, 155(1):4.

Mann A.W. et al., 2020. *Astronomical Journal*, 160(4):179.

Mann A.W. et al., 2022. *The Astronomical Journal*, 163(4):156. ISSN 0004-6256.

Mao S. and Paczyñski B., 1991. *ApJ*, 374:L37–L40.

Marcy G.W. et al., 2014. *Astrophysical Journal Supplement*, 210(2):20.

Marley M.S. et al., 2007. *The Astrophysical Journal*, 655(1):541–549. ISSN 0004-637X.

Marois C. et al., 2008. *Science*, 322(5906):1348–1352. ISSN 0036-8075.

Martin D.V., 2017. *Monthly Notices of the Royal Astronomical Society*, 465(3):3235–3253. ISSN 13652966.

Martínez R.R. et al., 2020. *The Astronomical Journal*, 160(3):111. ISSN 0004-6256.

Martins B.L.C. et al., 2020. *The Astrophysical Journal Supplement Series*, 250(1):20. ISSN 1538-4365.

Mascareño A.S., Rebolo R., and Hernández J.I.G., 2016. *Astronomy & Astrophysics*, 595:A12.

Masuda K., 2014. *Astrophysical Journal*, 783(1):53.

Masuda K. and Tamayo D., 2020. *The Astronomical Journal*, 160(5):224. ISSN 1538-3881.

Matsuo T. et al., 2007. *The Astrophysical Journal*, 662(2):1282–1292. ISSN 0004-637X.

Mayo A.W. et al., 2018. *The Astronomical Journal*, 155(3):136. ISSN 1538-3881.

Mayo A.W. et al., 2019. *Astronomical Journal*, 158(4):165. ISSN 0004-6256.

Mayor M. and Queloz D., 1995. *Nature*, 378(6555):355–359.

Mayor M. et al., 2003. *The Messenger*, 114:20–24.

Mayor M. et al., 2011. *arXiv e-prints*, arXiv:1109.2497.

Mazeh T. et al., 2013. *The Astrophysical Journal Supplement Series*, 208(2):16. ISSN 0067-0049.

McCullough P. et al., 2005. *Publications of the Astronomical Society of the Pacific*, 117(834):783–795. ISSN 0004-6280.

McQuillan A., Mazeh T., and Aigrain S., 2014. *The Astrophysical Journal Supplement Series*, 211(2):24. ISSN 0067-0049.

Meingast S., Alves J., and Rottensteiner A., 2021. *Astronomy & Astrophysics*, 645:A84. ISSN 0004-6361.

Meusinger H. et al., 2012. *Astronomy & Astrophysics*, 541:A77. ISSN 0004-6361.

Miguel Y. et al., 2019. *Monthly Notices of the Royal Astronomical Society*. ISSN 0035-8711.

Modak S., Chattopadhyay T., and Chattopadhyay A.K., 2020. *Journal of Applied Statistics*, 47(2):376–392. ISSN 0266-4763.

Montalto M. et al., 2020. *Monthly Notices of the Royal Astronomical Society*, 498(2):1726–1749.

Montet B.T. et al., 2020. *The Astronomical Journal*, 159(3):112.

Morton T.D., 2012. *Astrophysical Journal*, 761(1). ISSN 15384357.

Morton T.D. et al., 2016. *The Astrophysical Journal*, 822(2):86. ISSN 1538-4357.

Mosteller F. and Tukey J.W., 1977. *Data Analysis and Regression: A Second Course in Statistics*. Pearson, London.

Mostert R.I.J. et al., 2021. *Astronomy & Astrophysics*, 645:A89. ISSN 0004-6361.

Mullally F. et al., 2015. *Astrophysical Journal Supplement Series*, 217(2):31.

Murphy S.J. and Lawson W.A., 2015. *Monthly Notices of the Royal Astronomical Society*, 447(2):1267–1281. ISSN 13652966.

Naim A., Ratnatunga K.U., and Griffiths R.E., 1997. *ApJSS*, 111(2):357–367.

Nardiello D., 2020. *MNRAS*, 498(4):5972.

Nardiello D. et al., 2019. *MNRAS*, 490(3):3806–3823.

Nardiello D. et al., 2021. *MNRAS*, 505(3):3767–3784.

Naul B. et al., 2018. *Nature Astronomy*, 2(2):151–155. ISSN 2397-3366.

Nelson R.P., 2018. In H.J. Deeg and J.A. Belmonte, editors, *Handbook of Exoplanets*, 2287–2317. Springer International Publishing, Cham.

Newton E.R. et al., 2019. *The Astrophysical Journal*, 880(1):L17. ISSN 2041-8213.

Newton E.R. et al., 2021. *The Astronomical Journal*, 161(2):65. ISSN 1538-3881.

Newton E.R. et al., 2022. *arXiv eprints*, arXiv:2206.06254.

Nun I. et al., 2014. *The Astrophysical Journal*, 793(1):23. ISSN 1538-4357.

Nutzman P. and Charbonneau D., 2008. *Publications of the Astronomical Society of the Pacific*, 120(865):317–327. ISSN 0004-6280.

Oelkers R.J. and Stassun K.G., 2018. *ApJ*, 156(3):132.

Oelkers R.J. and Stassun K.G., 2019. *Research Notes of the AAS*, 3(1):8.

Ogihara M. and Ida S., 2009. *The Astrophysical Journal*, 699(1):824–838. ISSN 0004-637X.

Oh S. et al., 2017. *Astronomical Journal*, 153(6):257.

Ormel C.W., Liu B., and Schoonenberg D., 2017. *Astronomy & Astrophysics*, 604:A1. ISSN 0004-6361.

Osborn A. et al., 2021. *Monthly Notices of the Royal Astronomical Society*, 507(2):2782–2803.

Osborn H.P. et al., 2016. *Monthly Notices of the Royal Astronomical Society*, 457(3):2273–2286. ISSN 0035-8711.

Oshagh M. et al., 2017. *Astronomy & Astrophysics*, 606:A107.

Owen J.E., 2019. *Annual Review of Earth and Planetary Sciences*, 47(1):67–90. ISSN 0084-6597.

Owen J.E., 2020a. *Monthly Notices of the Royal Astronomical Society*, 498(4):5030–5040. ISSN 0035-8711.

Owen J.E., 2020b. *Monthly Notices of the Royal Astronomical Society*, 498(4):5030–5040.

Owen J.E. and Lai D., 2018. *Monthly Notices of the Royal Astronomical Society*, 479(4):5012–5021. ISSN 0035-8711.

Owen J.E. and Wu Y., 2017. *The Astrophysical Journal*, 847(1):29.

Pál A., 2012. *MNRAS*, 421(3):1825–1837.

Pál A. et al., 2008. *Astrophysical Journal*, 680(2):1450–1456.

Parker R.J., 2020. *Royal Society Open Science*, 7(11):201271.

Pashchenko I.N., Sokolovsky K.V., and Gavras P., 2018. *Monthly Notices of the Royal Astronomical Society*, 475(2):2326–2343. ISSN 0035-8711.

Paulson D.B., Cochran W.D., and Hatzes A.P., 2004. *The Astronomical Journal*, 127(6):3579–3586.

Pedregosa F. et al., 2011. *Journal of Machine Learning Research*, 12:2825–2830.

Pepe F. et al., 2021. *Astronomy & Astrophysics*, 645:A96.

Pepe F.A. et al., 2010. In I.S. McLean, S.K. Ramsay, and H. Takami, editors, *Proceedings of the SPIE*, 77350F.

Pepper J. et al., 2007. *Publications of the Astronomical Society of the Pacific*, 119(858):923–935. ISSN 0004-6280.

Perryman M., 2018. *The Exoplanet Handbook*. Cambridge University Press, Cambridge, 2nd edition. ISBN 9781108419772.

Perryman M. et al., 2014. *Astrophysical Journal*, 797(1).

Perryman M.A.C. et al., 1997. *ASTRONOMY AND ASTROPHYSICS*, 323:L49–L52.

Petigura E.A. et al., 2017. *The Astronomical Journal*, 154(3):107. ISSN 1538-3881.

Plavchan P. et al., 2020. *Nature*, 582(7813):497–500. ISSN 0028-0836.

Poddaný S., Brát L., and Pejcha O., 2010. *New Astronomy*, 15(3):297 – 301. ISSN 1384-1076.

Pojmanski G., 2002. *Acta Astronomica*, 52:397–427.

Pollacco D.L. et al., 2006. *Publications of the Astronomical Society of the Pacific*, 118(848):1407–1418.

Pollack J.B. et al., 1996. *Icarus*, 124(1):62–85. ISSN 00191035.

Poppenhaeger K., Ketzer L., and Mallonn M., 2020. *Monthly Notices of the Royal Astronomical Society*, 500(4):4560–4572. ISSN 0035-8711.

Prialnik D., 2009. *An introduction to the Theory of Stellar Structure and Evolution*. Cambridge University Press, Cambridge, 2 edition.

Qu M. et al., 2003. *Solar Physics*, 217(1):157–172. ISSN 00380938.

Queloz D. et al., 2000. *Astronomy and Astrophysics*, 354:99–102.

Quinn S.N. et al., 2012. *The Astrophysical Journal*, 756(2):L33.

Quinn S.N. et al., 2014. *The Astrophysical Journal*, 787(1):27.

Rahmani D., Fay D., and Brodzki J., 2019. *preprint (arXiv: 1905.05540)*.

Rajpaul V. et al., 2015. *Monthly Notices of the Royal Astronomical Society*, 452(3):2269–2291.

Rajpaul V.M., Aigrain S., and Buchhave L.A., 2020. *Monthly Notices of the Royal Astronomical Society*, 492(3):3960–3983.

Rasmussen C.E. and Williams C.K.I., 2006. *Gaussian Processes for Machine Learning*, volume 2. MIT Press. ISBN 978-0-262-18253-9.

Rauer H. et al., 2014. *Experimental Astronomy*, 38:249–330.

Raymond S.N., Scalo J., and Meadows V.S., 2007. *The Astrophysical Journal*, 669(1):606–614. ISSN 0004-637X.

Rein H., 2012. *Monthly Notices of the Royal Astronomical Society: Letters*, 427(1):L21–L24. ISSN 17453925.

Richards J.W. et al., 2011. *Astrophysical Journal*, 733(1). ISSN 15384357.

Richards J.W. et al., 2012. *Astrophysical Journal Supplement*, 203(2):32.

Ricker et al., 2014. *JATIS*, 1(1):014003.

Rivilla V.M. et al., 2015. *The Astrophysical Journal*, 808(2):146.

Rizos J. et al., 2021. *Icarus*, 364:114467. ISSN 00191035.

Rizzuto A.C., Ireland M.J., and Robertson J.G., 2011. *Monthly Notices of the Royal Astronomical Society*, 416(4):3108–3117. ISSN 00358711.

Rizzuto A.C. et al., 2017. *Astronomical Journal*, 154(6):224.

Rizzuto A.C. et al., 2018. *Astronomical Journal*, 156(5):195.

Rizzuto A.C. et al., 2020. *Astronomical Journal*, 160(1):33.

Rodenbeck K. et al., 2018. *Astronomy and Astrophysics*, 617. ISSN 14320746.

Rodríguez A. et al., 2011. *Monthly Notices of the Royal Astronomical Society*, 415(3):2349–2358. ISSN 00358711.

Rogers L.A., 2015. *The Astrophysical Journal*, 801(1):41. ISSN 1538-4357.

Röser S. et al., 2011. *Astronomy & Astrophysics*, 531(A92):1–15.

Rowe J.F. et al., 2014. *Astrophysical Journal*, 784(1):45.

Saar S.H. and Donahue R.A., 1997. *The Astrophysical Journal*, 485(1):319–327.

Safronov V.S., 1972. *Evolution of the protoplanetary cloud and formation of the earth and planets.* Israel Program for Scientific Translations, Jerusalem.

Sahlmann J. et al., 2013. *Astronomy & Astrophysics*, 556:A133.

Salpeter E.E., 1955. *Astrophysical Journal*, 121:161.

Salvatier J., Wiecki T.V., and Fonnesbeck C., 2016. *PeerJ Computer Science*, 2:e55.

Sanchis-Ojeda R. et al., 2013a. *Astrophysical Journal*, 775(1):54.

Sanchis-Ojeda R. et al., 2013b. *The Astrophysical Journal*, 774(1):54. ISSN 0004-637X.

Sarro L.M. et al., 2009. *Astronomy & Astrophysics*, 494(2):739–768. ISSN 0004-6361.

Sato B. et al., 2007. *The Astrophysical Journal*, 661(1):527–531.

Savitzky A. and Golay M.J.E., 1964. *Analytical Chemistry*, 36(8):1627–1639. ISSN 0003-2700.

Scargle J., 1982. *Astrophysical Journal*, 263:835–853.

Schawinski K., Turp M.D., and Zhang C., 2018. *Astronomy & Astrophysics*, 616:L16. ISSN 0004-6361.

Schilliro F. and Romano P., 2021. *MNRAS*, 503(2):2676–2687.

Schlichting H., Sari R., and Yalinewich A., 2015. *Icarus*, 247:81–94.

Schwieterman E.W. et al., 2018. *Astrobiology*, 18(6):663–708. ISSN 1531-1074.

Sergison D.J., 2015. Untangling the signals: Investigating accretion and photometric variability in young stars An observational analysis. Technical report, University of Exeter, Exeter.

Siverd R.J. et al., 2012. *Astrophysical Journal*, 761(2). ISSN 15384357.

Soderblom D.R. et al., 2014. In C.P.D. Henrik Beuther Ralf S. Klessen and T.K. Henning, editors, *Protostars and Planets VI*, chapter 1, 219–241. University of Arizona, Tuscon.

Solanki S.K., Inhester B., and Schüssler M., 2006. *Reports on Progress in Physics*, 69(3):563–668. ISSN 0034-4885.

SOTIN C., GRASSET O., and MOCQUET A., 2007. *Icarus*, 191(1):337–351. ISSN 00191035.

Spiegel D.S. and Burrows A., 2012. *Astrophysical Journal*, 745(2). ISSN 15384357.

Stassun K.G. et al., 2018. *The Astronomical Journal*, 156(3):102. ISSN 0004-6256.

Stassun K.G. et al., 2019. *The Astronomical Journal*, 158(4):138. ISSN 1538-3881.

Steffen J.H. et al., 2012. *Monthly Notices of the Royal Astronomical Society*, 421(3):2342–2354.

Steffen J.H. et al., 2013a. *Monthly Notices of the Royal Astronomical Society*, 428(2):1077–1087.

Steffen J.H. et al., 2013b. *Monthly Notices of the Royal Astronomical Society*, 428(2):1077–1087. ISSN 1365-2966.

Stevens D.J. and Gaudi B.S., 2013. *Publications of the Astronomical Society of the Pacific*, 125(930):933–950. ISSN 00046280.

Storn R. and Price K., 1997. *Journal of Global Optimization*, 11:341–359.

Storrie-Lombardi M.C. et al., 1992. *Monthly Notices of the Royal Astronomical Society*, 259(1):8P–12P. ISSN 0035-8711.

Su K.Y.L. et al., 2020. *The Astrophysical Journal*, 898(1):21.

Sun L. et al., 2019. *Astronomy & Astrophysics*, 624:A15.

Szabó G.M. et al., 2020. *MNRAS Letters*, 492(1):L17–L21.

Tajiri T. et al., 2020. *ApJSS*, 251(2):18.

Talens G.J.J. et al., 2017. *Astronomy & Astrophysics*, 601:A11. ISSN 0004-6361.

Tamuz O., Mazeh T., and Zucker S., 2005. *Monthly Notices of the Royal Astronomical Society*, 356(4):1466–1470.

Taylor M., 2005. In P. Shopbell, M. Britton, and R. Ebert, editors, *Astronomical Data Analysis Software and Systems XIV*.

Thao P.C. et al., 2020. *The Astronomical Journal*, 159(1):32. ISSN 1538-3881.

The Astropy Collaboration et al., 2013. *A&A*, 558:A33.

Thompson A.P.G. et al., 2020. *Monthly Notices of the Royal Astronomical Society*, 494(3):4279–4290.

Thompson S.E. et al., 2016a. Kepler Archive Manual. Technical report, NASA.

Thompson S.E. et al., 2018. *Astrophysical Journal Supplement Series*, 235(2):38.

Thompson S.J. et al., 2016b. In C.J. Evans, L. Simard, and H. Takami, editors, *Proceedings of the SPIE*, 99086F.

Tofflemire B.M. et al., 2021. *Astronomical Journal*, 161(4):171.

Torrence C. and Compo G.P., 1998. *Bulletin of the American Meteorological Society*, 79(1):61–78. ISSN 0003-0007.

Torres C.A.O. et al., 2006. *A&A*, 460:695–708.

Udalski A., 2003. *Acta Astronomica*, 53:291–305.

Valenzuela L. and Pichara K., 2018. *Monthly Notices of the Royal Astronomical Society*, 474(3):3259–3272. ISSN 0035-8711.

van der Maaten L. and Hinton G., 2008. *Journal of Machine Learning Research*, 9(86):2579–2605.

Van Eylen V. et al., 2019. *The Astronomical Journal*, 157(2):61. ISSN 1538-3881.

Vanderburg A. and Johnson J.A., 2014. *Publications of the Astronomical Society of the Pacific*, 126(944):948–958. ISSN 00046280.

Vanderburg A. et al., 2016. *The Astrophysical Journal Supplement Series*, 222(1):14. ISSN 0067-0049.

Vanderburg A. et al., 2018. *Astronomical Journal*, 156(2):46.

Vanderburg A. et al., 2019. *The Astrophysical Journal*, 881(1):L19. ISSN 2041-8213.

Vanderspek R. et al., 2018. TESS Instrument Handbook. Technical report, NASA.

Virtanen P. et al., 2019. *arXiv e-prints*, arXiv:1907.10121.

Wang J. et al., 2014. *The Astrophysical Journal*, 783(1):4. ISSN 0004-637X.

Ward J.H., 1963. *Journal of the American Statistical Association*, 58(301):236–244. ISSN 0162-1459.

Ward W.R., 1986. *Icarus*, 67(1):164–180. ISSN 00191035.

Watson A.J., Donahue T.M., and Walker J.C., 1981. *Icarus*, 48(2):150–166. ISSN 00191035.

Weiss L.M. et al., 2013. *Astrophysical Journal*, 768(1):14.

Welsh W.F. et al., 2015. *Astrophysical Journal*, 809(1):26–undefined. ISSN 15384357.

Werner M.W. et al., 2004. *The Astrophysical Journal Supplement Series*, 154(1):1–9.

Wheatley P.J. et al., 2010. *arXiv e-prints*, arXiv:1004.0836.

Wheatley P.J. et al., 2013. In *Proceedings of the International Astronomical Union*. ISBN 9781107045200. ISSN 17439221.

Wheatley P.J. et al., 2018. *Monthly Notices of the Royal Astronomical Society*, 475(4):4476–4493.

Wolfgang A. and Lopez E., 2015. *The Astrophysical Journal*, 806(2):183. ISSN 1538-4357.

Wright A.H. et al., 2020a. *Astronomy & Astrophysics*, 640:L14. ISSN 0004-6361.

Wright A.H. et al., 2020b. *Astronomy & Astrophysics*, 637:A100. ISSN 0004-6361.

Xie J.W., 2014. *Astrophysical Journal Supplement*, 210(2):25.

York D.G. et al., 2000. *The Astronomical Journal*, 120(3):1579–1587. ISSN 00046256.

Zari E., Brown A.G., and De Zeeuw P.T., 2019. *Astronomy and Astrophysics*, 628:A123. ISSN 14320746.

Zari E. et al., 2018. *Astronomy and Astrophysics*, 620:A172. ISSN 14320746.

Zellem R.T. et al., 2020. *Publications of the Astronomical Society of the Pacific*, 132(1011):054401–undefined.

Zhou G. et al., 2016. *The Astronomical Journal*, 152(5):136. ISSN 1538-3881.

Zhou G. et al., 2019. *The Astronomical Journal*, 158(4):141.

Zhou G. et al., 2020. *The Astrophysical Journal Letters*, 892(2):L21.

Zhou G. et al., 2021. *The Astronomical Journal*, 161(1):2.

Zhou G. et al., 2022. *The Astronomical Journal*, 163(6):289. ISSN 0004-6256.

Zieba S. et al., 2019. *Astronomy & Astrophysics*, 625:L13.

Zuckerman B., 2019. *ApJ*, 870(1):i.d. 27.

Zuckerman B. and Webb R.A., 2000. *The Astrophysical Journal*, 535(2):959–964. ISSN 0004-637X.