UiO : Institute of Theoretical Astrophysics
University of Oslo

Early Science with the Next Generation Transit Survey

Photometric performance and variable star identification

Vedad Hodžić

Master's Thesis, Spring 2017

Early Science with the Next Generation Transit Survey

Photometric performance and variable star identification

Thesis submitted for the degree of
Master of Science: Astronomy
March 2017

Author: Vedad Hodžić[1]
Supervisors: Professor Didier Queloz[2]
Dr. Edward Gillen[2]
Professor Hans Kristian Kamfjord Eriksen[1]

[1]Institute of Theoretical Astrophysics, P. O. Box 1029 Blindern, 0315 Oslo, Norway
[2]Astrophysics Group, Cavendish Laboratory, J J Thomson Avenue, Cambridge CB3 0HE, UK

ABSTRACT

The Next Generation Transit Survey (NGTS) is an ongoing wide-field exoplanet transit survey at ESO Paranal Observatory, Chile. It is searching for Neptune and super-Earth size planets around nearby stars using 12 robotic telescopes covering a total field of view of 88.8 deg^2 on the sky. The instruments are optimised for sensitivity to stars of spectral type K and early-M, and the experiment will provide prime targets for further characterisation with current and future instruments. We have analysed NGTS photometry from 30 fields observed during its first year in operation to assess the achieved photometric precision. We build a white noise model to estimate the level of systematic (red) noise in the photometry, and find a median red noise level of 0.82±0.10 mmag in our fields over their full observing season. We develop an automated way of detecting variable stars using our noise model, and find 1245 (1.31%) variable stars at high confidence. Their periods are found with an implementation of the generalised Lomb-Scargle periodogram, and we find a 71% matching rate with known variables in our fields from variable star catalogues. Excluding known variables, we discover ~900 new variable stars that include rapid pulsators (δ Scuti, RR Lyrae), short-period eclipsing binaries, Cepheids, and semiregular variables. We also find an overdensity of variables at periods of 10–30 days at amplitudes of less than 1%, which we find is consistent with main sequence rotation periods for our target stars. If confirmed in future studies, this would be the first time a wide-field survey has been able to detect these signals, opening up new studies on stellar evolution of low-mass stars from the ground.

Acknowledgements

This work was carried out at the Cambridge Exoplanet Research Centre at Cavendish Astrophysics, University of Cambridge. I extend my gratitude to Didier Queloz for hosting me, welcoming me to his group, and supporting me during the summer months. Furthermore, a warm thanks to both Didier and Ed Gillen for excellent guidance throughout my stay at Cambridge, and always making me feel welcome to knock on your doors and ask silly questions. I consider myself very lucky to have been able to learn from the very best in the field, and will always carry with me the experiences and lessons I learned during my stay.

Finishing a graduate degree in astrophysics is anything but straightforward, and towards the end it is easy to find yourself reflecting on how you got there, and who helped you along the way. In chronological order, I would like to start with thanking Ann, who so enthusiastically told me about the wonders of space and astronomy at a young and impressionable age. I also thank my former physics teacher, Marit, whose fascination for the wonders of physics and the Universe was so contagious that she awoke the same curiosity in me.

Thank you to Paul Magnus and Benedicte for keeping me sane through all the late nights together by our desks during our undergraduate years. Thank you to all the others students and academics at the Institute of Theoretical Astrophysics (ITA) for all our time together, making it quite a memorable experience. A heartfelt thank you to Hans Kristian at ITA for great teaching, guidance, and for giving me the opportunity to carry out my first research project at JPL.

Thank you to Johannes Sahlmann for kindly helping me find opportunities to get involved in exoplanet research for my Master's research project, and in the end introducing me to Didier in Cambridge.

A big thank you to ITA, and especially Viggo, for granting me financial means to attend summer schools and conferences to explore my interests, resulting in a unique experience as a Master's student.

Finally, and most importantly, an immeasurable amount of gratitude to my mother, without whom nothing of what I have achieved would have been possible. Thank you for your unconditional love and support in all aspects of my life. You have been my rock in a stormy ocean ever since the beginning, and I am the luckiest man in the world to have you as a mother.

Vedad Hodzic

Cambridge
February 2017

List of Figures

List of Tables

Contents

Thesis overview

This thesis is structured in the following way:

Chapter 1 We give a brief historical introduction to transiting exoplanet surveys in the context of NGTS. Next, we go through important considerations to take into account when designing a transit survey. We give an introduction to charge-coupled devices (CCDs) and noise sources in astronomical images. Finally, we introduce some types of variable stars that will be important for the remainder of this work, and show their example lightcurves.

Chapter 2 We introduce the Next Generation Transit Survey (NGTS), which is is the experiment that this thesis is built on. We present its characteristics, science goals, and important pipeline data reduction steps. Finally we give a brief summary of the status during its first year of operations.

Chapter 3 The data processing and analysis techniques that were used in this work are introduced and explained in detail. We present our noise model, and our method of using our noise model to identify variable stars. We further describe our method of finding the dominant period in these variables, and compare our results to catalogued variable stars with known periods.

Chapter 4 Here we present our results from analysing the noise levels from 30 fields observed with NGTS durings its first year and a half. We discuss noise-related issues , and present numbers and parameters on our variable star findings. Finally, we present some example lightcurves from our identifed variable stars to showcase the precision of NGTS photometry.

Chapter 5 We discuss the overall results and issues that we faced throughout this study. In particular, we review limitations of our methods and the data, and suggest solutions for improvements. We connect our noise findings in the context of planet yield simulations for NGTS. Finally, we suggest future avenues that can build on this work, before concluding with our findings.

CHAPTER 1

Wide-field astronomy

We are the middle children of history. Born too late to explore the Earth, born too early to explore the Galaxy.

Unknown

1.1 Introduction

Astronomy is truly a unique science, being one of few sciences that naturally sparks an interest and curiosity in the general population. One reason for this is the important role astronomy has played throughout human history. Needing only eyes to observe the heavens, humans have, for as long as we can remember, gazed upon the stars and found themselves contemplating our existence. Prompting such questions as: "How did we come to be?" and "Are we alone?". In addition to inspiring philosophical discussions, astronomy was a major driving force behind the scientific revolution in the 16^{th} century, fuelled in particular by historical people such as Copernicus, Kepler, Tycho Brahe, Newton, and Galileo. These scientists were behind deriving some of the most fundamental theories used in astronomy and astrophysics today which describe the interplay between planetary motion in elliptical orbits and the law of gravity. In particular, Galileo built a high-magnification refracting telescope with which he observed Venus. He discovered that the planet has phases similar to the Moon, which depend on Venus' orbit around the Sun, delivering a big blow to the geocentric model.

With his state-of-the-art telescope, Galileo also studied Jupiter and discovered four of its largest moons. Through systematic observations, he made great contributions to our knowledge of the Solar System, and most importantly laid the foundation for *observational astronomy*. The next several hundred years, humans developed better and bigger telescopes and raised observatories in most major cities in Europe. The invention of photographic plantes in the 19^{th} century made it possible for astronomers to record the images they were seeing. However, these early detectors suffered from low quantum efficiences (QE) of roughly 10%. A new revolution in observational astronomy began with the invention of charge-coupled devices (CCDs) in 1969. Unlike photographic plates, early CCDs could detect over 90% of near-infrared to ultraviolet light that reached the chip, which was a major improvement. Furthermore, the CCD converts measured voltages to digital numbers on a computer, which greatly simplified astronomical observations. Advancements made in technology and computers enabled astronomers to observe larger patches of the sky at once instead of doing targeted observations. These wide-field observations would generate much more data than was previously thought possible, and opened up the concept of time-domain astronomy.

These advancements would be important for a new research field in astronomy that was just about to emerge.

Approaching the mid 90s, a Swiss graduate student at Geneva Observatory was observing a sample of stars to measure their radial velocity (RV) from observing the Doppler shift of stellar absorption lines. For one particular star, Peg 51, the signal was varying periodically at low amplitudes. Convinced there was an error in his computer code or some systematic noise in the experiment, Didier Queloz set out to scrutinise every part of the CORALIE spectrograph and his computer code. Still after double and triple checking everything, the signal persisted. He had found the first extrasolar planet around a Sun-like star, with the signal being consistent with a Jupiter-mass object at a 4.2 day orbital period (M. Mayor and D. Queloz, 1995). The presence of such a planet unlike anything we have seen in our Solar System baffled planet formation theorists, and was difficult to accept in the scientific community. The findings were confirmed by a different group, and in the following years, radial velocity surveys were carried out to look for other worlds. At the turn of the millenium, dozens of giant planets in short orbits around their star had been found. Astronomers had become more accepting of the idea these strange worlds, and the research area of exoplanets (short for extrasolar planets) was born.

Towards the end of the 90s, astronomers had realised that if a giant planet was to transit the disc of the star as seen from the observer, it would slightly dim the brightness of the star, enough to be detected with current technology. They began the search for these transit signals around planet-hosting systems known from RV measurements, and the first transiting planet was found to be a $1.27R_{\rm Jup}$ around a $1.1R_{\rm Sun}$ star, which produced a 1.5% signal in the lightcurve. Knowing the transit detection method had worked, astronomers began to plan wide-field surveys that could detect them efficiently. From geometrical considerations, they knew that the probability of a transit in a planet-hosting system was around 5%, depending star-planet separation, which meant they had to monitor thousands of stars simultaneously for it to be efficient. Several pioneering transit surveys are worth mentioning in this regard. One was the Hungarian-made Automated Telescope (HAT; G. Bakos et al. 2004), which later extended to a network of telescopes (HATNet). By September 2016 it had discovered 60 giant planets in close orbits to their star. The Wide Angle Search for Planets (WASP; D. L. Pollacco et al. 2006), later extended to SuperWASP, followed in the same footsteps of detecting giant planets and has found 132 giant planets since the start of its science observations in the early 2000s. The first space telescope that was partly focused on exoplanets was the *CoRoT* mission, which made the first discovery of a super-Earth sized planet (A. Léger et al., 2009).

Across the Atlantic at NASA, astronomers wanted to build a space telescope to look for Earth's twin, and get a statistical census of the occurence rates of small planets, since observations from the ground were limiting detection of sub-Jupiter size planets. The *Kepler* spacecraft (W. J. Borucki et al., 2010) was launched in 2009, and has since then become the most successful exoplanet mission. By February 2017 it has 2330 confirmed planet detections, with another 4706 planet candidates. The *Kepler* mission found that giant planets are not that common in our Galaxy, while Neptune and super-Earth sized planets are found around more than 50% of stars. Furthermore, planets were preferentially found around smaller stars than our Sun. Despite its enormous success, one problem with the *Kepler* planets was that most of them were orbiting host stars that were several hundred and even thousand light years away, thus being very faint. This makes a large fraction of *Kepler* planets ineligible for radial velocity follow-up to get the mass of these planets. Most of these planets are also difficult to characterise by studying their atmosphere during transit (transmission spectroscopy), as the signal-to-noise ratio can be unforgiving.

After *Kepler*, the exoplanet community had realised that if they want to study these foreign planets in more detail and potentially address questions regarding habitability, they would have to focus on bright, nearby stars that were suited for follow-up observations. The Transiting Exoplanet Survey Satellite (TESS; G. R. Ricker et al. 2014), scheduled launch in early 2018, is an all-sky transit survey space

mission designed to find planets around nearby stars that make follow-up and atmospheric characterisation possible. However, some astronomers believed that this can be achieved from the ground, which would keep the costs at a fraction of what sending a telescope to space would cost. The scientific outcome from *Kepler* allowed astronomers to re-design transit surveys, optimised to look for smaller planets around smaller stars that are bright enough for follow-up radial velocity and atmospheric characterisation. These would be near-ubiquitous, given that their sub per-cent signal could be detected. Building on the experience from previous ground-based transit surveys, astronomers from universitites in the UK, Geneva, and Berlin, designed an exoplanet experiment that would be optimised for these targets, and would be capable of detecting signals less than $^1/_{1000}$ from the ground. With these principles in mind, the Next Generation Transit Survey (NGTS) was built. NGTS is currently in its second year of operations, observing with 12 telescopes from the Chilean Paranal desert. Each telescope is observing a field for 3–4 months, covering a total sky area of 88.8 deg^2. The survey has already gathered several terrabytes (TB) of end-product data, which has been the subject of study in this work.

Wide-field surveys with long-term observations of stars in search for exoplanets opens up possibilities for exciting secondary science, for example focused on variable stars or asteroseismology. The bulk of this work is focused on the photometric performance of NGTS during its first year of science observations, while a smaller part is focused on exploring some early science on variable stars observed with NGTS. This Chapter will give an overview of what drives the design of transit surveys, explain key challenges, and review common noise sources in photometric experiments. Finally we will briefly review the taxonomy of variable stars, which will provide a context for new variable star discoveries in this work. The bulk of the derivations and results in this Chapter follows the book *Transiting Exoplanets* by C. A. Haswell (2010).

1.2 Transit survey design

The idea behind the transit technique is rather simple; when a planet passes in front of the disc of a star along our line of sight, it will block out a fraction of the light of the star as seen from the observer. The search for transiting exoplanets thus refers to the search for periodic dips in the brightness of a star. As we will see later, there are several false alarms that can produce similar brightness dips, both astrophysical and noise-related, and the real challenge becomes vetting the transiting candidates for bona fide planets.

Although there is a surprising amount of science one can extract from transiting planets alone, a planet can never be fully characterised without follow-up observations through radial-velocity measurements of the host star. Wide-field observations also require targeted photometric follow-up to rule out false planetary signals. Transit surveys are also inherently biased towards larger planets in close orbits around their host star, as will be shown below.

1.2.1 General considerations

Transit depth

One of the fundamental properties one can extract from a transiting planet is an estimate of its size relative to its host star. The dip in brightness of the star during the transit can be estimated as the fraction of the disc covered by the planet:

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

where F is the out-of-transit flux from the star, ΔF is the change in flux during transit, and R_p and R_* are the radii of the planet and star, respectively. Let us for a moment imagine aliens observing our own Sun, looking for transiting planets. Earth transiting the Sun would cause a dip

$$\frac{\Delta F}{F} = \frac{R_\oplus^2}{R_\odot^2} = \left(\frac{6.4 \times 10^3\,\mathrm{km}}{7.0 \times 10^5\,\mathrm{km}}\right)^2 = 8 \times 10^{-5},$$

a signal less than 0.1%. A transit of Jupiter, on the other hand, would produce a dip

$$\frac{\Delta F}{F} = \frac{R_\mathrm{J}^2}{R_\odot^2} = \left(\frac{7.0 \times 10^4\,\mathrm{km}}{7.0 \times 10^5\,\mathrm{km}}\right)^2 = 1 \times 10^{-2},$$

a signal of 1%. Signals such as the latter are easy to detect with current technology. In fact, the first transiting exoplanet, around the star HD 209458, was observed already in 1999 (D. Charbonneau et al. 2000; G. W. Henry et al. 2000); a gas giant with $R_\mathrm{p} = 1.27\,\mathrm{R_J}$ (at the time) around a Sun-like star with $\{R_*, M_*\} = \{1.1\,\mathrm{R_\odot}, 1.1\,\mathrm{M_\odot}\}$, producing a transit signal of about 1.5% in an approximate Johnson R band. As a rule of thumb, giant planets will cause a $\sim$1% signal, while a terrestrial planet will cause a $\sim$0.1% signal.

It is clear then that in order to search for Earth-like planets around Sun-like stars, one needs a higher precision than what is achievable from the ground. Even at the best observatories in the world with the most favourable seeing, the varying transparency of the atmosphere will simply smear out the signal before reaching the needed precision. Thus the rationale behind the *Kepler* mission; an attempt to probe the occurence rates of Earth-size planets from space, escaping the problem of Earth's atmosphere. Perhaps the most significant result from the *Kepler* mission was occurence rates of exoplanets of various size, which helped shape the design of current and future transit surveys (F. Fressin et al., 2013).

From Eq. 1.1 we also notice that we can decrease the size of the host star in order to create a larger, detectable signal for smaller planets. Indeed, the limitation of detecting smaller planets from the ground can be circumvented by observing stars of spectral type K and early M. This is the type of stars that NGTS is targeting to search for Neptune-sized planets and smaller. Focusing on these targets allows NGTS to reach a precision of <1 mmag; the first wide-field survey to reach that kind of precision from the ground. Thinking further along these lines sparked the idea of observing even lower-mass stars for transiting planets, such as ultra-cool dwarf stars of spectral type late M and L, found at the bottom of the main sequence. This sparked the SPECULOOS experiment (**S**earch for Habitable **P**lanets **Ec**lipsing **Ul**tra-C**oo**l **S**tars; M. Gillon et al. 2013), which is scheduled to start science observations in the second half of 2017. The TRAPPIST-South telescope (**Tra**nsiting **P**lanets and **P**lanetes**i**mals **S**mall **T**elescope) at La Silla observatory in the Chilean Atacama Desert has observed a few dozens such stars, and already from a very small sample found a planetary system with seven Earth-size planets around the M8 type star TRAPPIST-1 with mass $0.08\,\mathrm{M_\odot}$ (M. Gillon et al. 2016; M. Gillon et al. 2017, in press). The discovery led to the first atmospheric characterisation of Earth-sized exoplanets in the habitable zone (J. de Wit et al. 2016), and opened up the study of a whole new population of exoplanets around ultra-low mass stars.

Transit probability

Another aspect that shapes the design of transit surveys is the probability of observing a transit. For a transit to occur from an observer's point of view, the orbital inclination, i, must be close to 90°, as shown in Fig. 1.1. From Fig. 1.2 one can see that for the planet to transit the stellar disc the orbital inclination, i,

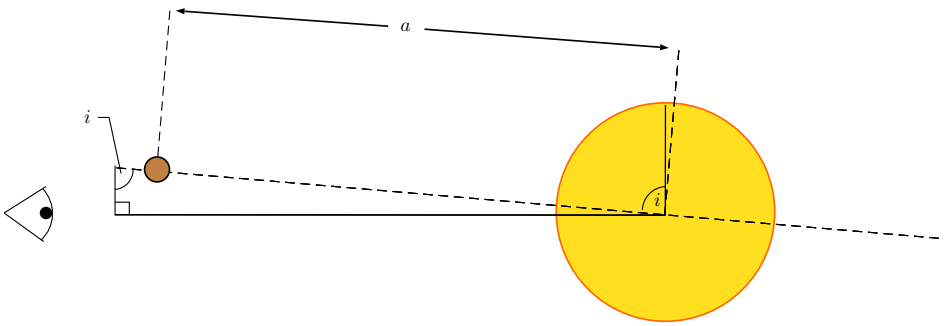


Figure 1.1: The distant observer views a transit with orbital inclination i. The distance a is the semi-major axis of the orbit, which we assume to be circular.

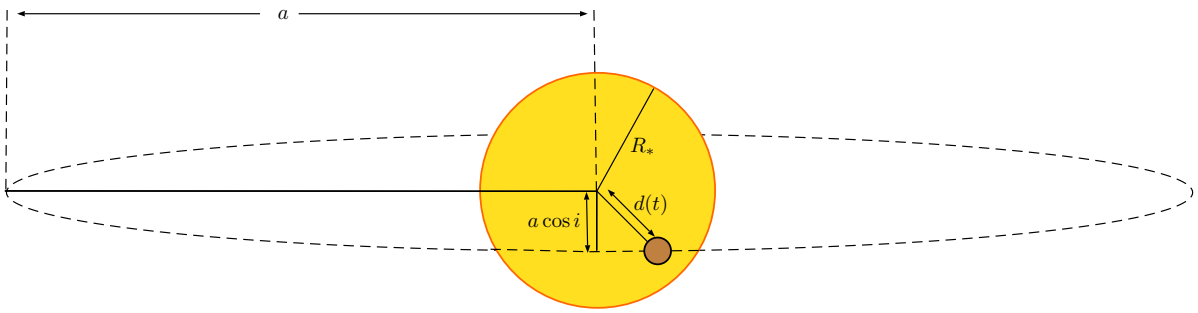


Figure 1.2: Transit geometry from an observer's point of view.

must satisfy

$$a\cos i \leq R_* + R_\mathrm{p}. \tag{1.2}$$

In the case of the unit vector normal to the orbital plane, we notice that its projection onto the sky plane is $\cos i$, and is equally likely to take any value between 0 and 1. To find an expression for the transit probability we need to integrate over all orbital inclinations that satisfy Eq. 1.2, divided by all possible values of $\cos i$. We substitute $x = \cos i$ and have that the geometric probability of transit is

$$\begin{aligned} P(\text{transit}) &= \frac{\int_0^{(R_*+R_p)/a} \mathrm{d}x}{\int_0^1 \mathrm{d}x} \\ &= \frac{R_* + R_\mathrm{p}}{a} \\ &\approx \frac{R_*}{a}. \end{aligned} \tag{1.3}$$

Equation 1.3 thus highlights the strong bias of close-in planets around large host stars in transit surveys. Note that expression above is the *geometric* transit probability, and does not take into account the frequency of planets of a certain size. If we imagine an alien civilisation observing our Sun, applying Eq. 1.3 to our own Solar System planets we have that the geometric transit probability is less than 1% for

all our planets except Mercury (1.2%).

1.2.2 Scaling laws for the discovery of transiting exoplanets

At the end of this sub-section we would like to arrive at an expression that describes how the expected planet yield scales with various factors that depends on the survey, planet, and star. The derivation of these calculations are not important for the current work, but their result provides a relevant context to the design of NGTS. Therefore we only include the important results, and refer the reader to *Transiting Exoplanets* (C. A. Haswell, 2010) for details of the calculations.

Number of sources brighter than limiting flux

Astronomical equipment, such as the telescope design in transit surves, will have a limiting flux, S, below which sources are undetectable. We will then generally detect sources in the field of view with a flux, $F \geq S$. Considering a uniformly distributed population of stars with number density n_0, all with luminosity, L, there is a limiting distance, $d_{\max}$, out to which the sources are detectable, given by

$$d_{\max} = \left(\frac{L}{4\pi S}\right)^{1/2}. \tag{1.4}$$

The number of sources brighter than the limiting flux, S, is then

$$N(S) = \frac{4\pi n_0}{3}\left(\frac{L}{4\pi S}\right)^{3/2} \propto S^{-3/2}. \tag{1.5}$$

Transit duration

We need an expression for the transit duration, which can be calculated by considering the orbital period and the fraction of the planet's orbit for which the planet is in front of the star. We define two points, V and W, in the planet's orbit where the planet is passing in front of the disc, and leaving the disc, respectively. Assuming that the radius of the planet's orbit, a, is much larger than the radius of the star, R_*, the arc from V to W, can be approximated as the distance between the two points, which is just $2R_*$. Further assuming a circular orbit, the transit duration is

$$T_{\rm dur} = P \times \frac{D_{\rm arc}^{V\rightarrow W}}{2\pi a} \approx \frac{P \times 2R_*}{2\pi a} = \frac{PR_*}{\pi a}, \tag{1.6}$$

where $D_{\rm arc}^{V\rightarrow W}$ is the length of the arc from V to W, and P is the planet's orbital period.

The signal-to-noise ratio

Whether detecting a transit or not is possible for an instrument depends mainly on the signal-to-noise ratio (SNR) of the data. For faint star, the dominant source of noise is from the sky background. The number of sky photons can written as

$$n_{\rm sky} = AQ\,\Delta\lambda\,\sigma_{\rm FWHM}^2\,l_{\rm sky}\,\Delta t,$$

where Δt is the duration of the observation; A is the light collecting area of the telescope (mirror area); Q is the quantum efficiency (fraction of photons that actually excites electrons); $\Delta\lambda$ is the bandpass of the instrument; $\sigma_{\rm FWHM}$ is the the point spread function (PSF) at FWHM; $l_{\rm sky}$ is a property of the sky at the

telescope location that describes the number of sky photons per pixel per unit area per unit wavelength interval.

The PSF is a property of the instrument and describes how detected photons from a point source are spread out on the CCD. Generally, a cross-section through the PSF is close to Gaussian, and it is therefore quantified by the width of the best-fitting Gaussian function, σ_{FWHM}. The width of the PSF strongly depends on how much the light is refracted on its path through the atmosphere. How much the light is refracted is referred to as the seeing, which is usually less than $1''$. For transit surveys in particular, it is more important to have a large sky coverage than to fully resolve the PSF of the seeing disc.

The statistical uncertainty in the number of photons, N, is determined by Poisson statistics, with statistical fluctuation $\sqrt{N}$ (see Section 1.3). The in-transit flux will dominate the uncertainty, as most of the time is spent observing the flux out of transit. The noise is then

$$\begin{aligned} \mathrm{N} &\approx \sqrt{n_{\mathrm{sky}}} \\ &\approx \sqrt{AQ\,\Delta\lambda\,\sigma^2_{\mathrm{FWHM}}\,l_{\mathrm{sky}}\,\Delta t} \\ &\approx \sqrt{AQ\,\Delta\lambda\,\sigma^2_{\mathrm{FWHM}}\,l_{\mathrm{sky}}\,\frac{PR_*}{\pi a}}. \end{aligned} \tag{1.7}$$

The signal we receive from the star during transit can be expressed as

$$\mathrm{S} = \frac{R_{\mathrm{p}}^2}{R_*^2} AQ\,\Delta\lambda\eta \frac{L_*}{4\pi d^2} \frac{\overline{\lambda}}{hc} \frac{PR_*}{\pi a} \exp\left(-Kd\right). \tag{1.8}$$

Combining Eqs. 1.7 and 1.8 we have that the signal-to-noise ratio for *one* transit is

$$\begin{aligned} \frac{\mathrm{S}}{\mathrm{N}} &\approx \frac{R_{\mathrm{p}}^2}{R_*} \frac{AQ\,\Delta\lambda \frac{\eta\overline{\lambda}}{hc} \frac{PL_*}{4\pi^2 a d^2} \exp\left(-Kd\right)}{\sqrt{AQ\,\Delta\lambda\,\sigma^2_{\mathrm{FWHM}}\,l_{\mathrm{sky}}\,\frac{PR_*}{\pi a}}} \\ &\approx \frac{R_{\mathrm{p}}^2}{4(\pi R_*)^{3/2}} \left(\frac{AQ\,\Delta\lambda\,P}{l_{\mathrm{sky}} a}\right)^{1/2} \frac{\eta\overline{\lambda} L_* \exp\left(-Kd\right)}{hcd^2\sigma_{\mathrm{FWHM}}}. \end{aligned} \tag{1.9}$$

Here $\overline{\lambda}$ is the mean photon wavelength; K is the interstellar extinction coefficient; d is the distance to the star; η is an efficiency parameter that takes into account that some of the star's flux falls outside the bandpass, $\Delta\lambda$. However, a requirement for a transiting planet candidate is repeated transits. If N_{t} transits are observed, the noise is reduced by a factor $\sqrt{N_{\mathrm{t}}}$. For ground-based observations, a minimum of three transits are required to determine the orbital period due to being able to observe only during night. The number of transit observed can be expressed as

$$N_{\mathrm{t}} = \frac{\xi t}{P},$$

where ξ is the duty cycle: the fraction of time the survey is collecting data, and t is the elapsed time. Finally we can write the signal-to-noise ratio for a given instrument and star with N_{t} transits:

$$\frac{\mathrm{S}}{\mathrm{N}} \approx \frac{R_{\mathrm{p}}^2}{4(\pi R_*)^{3/2}} \left(\frac{AQ\,\Delta\lambda\,\xi t}{l_{\mathrm{sky}} a}\right)^{1/2} \frac{\eta\overline{\lambda} L_* \exp\left(-Kd\right)}{hcd^2\sigma^2_{\mathrm{FWHM}}}. \tag{1.10}$$

Survey volume and number of stars searched

In Eq. 1.4 we found the distance, $d_{\rm max}$, to which a survey can detect sources. We only search a fraction of this sphere, which is our field of view. Our field of view is the solid angle θ^2, thus the volume we observe is

$$V = \frac{\theta^2}{4\pi}\frac{4\pi d_{\rm max}^3}{3} = \frac{\theta^2 d_{\rm max}^3}{3}.$$

The sum of inidivdual shells within this volume is

$$V = \int_0^{d_{\rm max}} {\rm d}V = \int_0^{d_{\rm max}} \frac{\theta^2}{4\pi} \times 4\pi r^2 \, {\rm d}r = \theta^2 \int_0^{d_{\rm max}} r^2 \, {\rm d}r. \tag{1.11}$$

In Eq. 1.5 we assumed that the density of sources was uniform. In our Galaxy, the stellar density varies with Galactic longitude and latitude. The Galactic plane is at latitude $b = 0$. If we assume that the stellar density is n_0 in the Galactic plane, we can express the spatial variation in density as

$$\begin{aligned} n &= n_0 \exp\left(-\frac{h}{H}\right) \\ &= n_0 \exp\left(-\frac{d\,|\sin b|}{H}\right), \end{aligned} \tag{1.12}$$

where h is the distance above the Galactic plane, and H is the scale-height of the density distribution of stars. Bringing Eqs. 1.11 and 1.12 we have that the total number of stars surveyed is

$$\begin{aligned} N_{\rm star} &= \theta^2 \int_0^{d_{\rm max}} n(r)\, r^2 \, {\rm d}r \\ &= \theta^2 \int_0^{d_{\rm max}} n_0 \exp\left(-\frac{r\,|\sin b|}{H}\right) r^2 \, {\rm d}r. \end{aligned} \tag{1.13}$$

The number of transiting planets per star

One of the most important results from the *Kepler* mission was a statistical cencus of the occurence rates of planets around stars. In particular, the mission sought to find $\eta_{\rm Earth}$; the number of Earth-like planets in the habitable zone per star in our Galaxy. From *Kepler* and other successful exoplanet surveys we thus have an approximate function, $\alpha_{\rm p}(a, M_{\rm p})$, that describes the distribution of planet frequency as function of semi-major axis, a, and mass, $M_{\rm p}$. The average number will depend on the properties of the star. For example, Jupiter-mass planets are found to be preferentially around metal-rich stars. The total number of planets per star, $\eta_{\rm p}$, will be an intergral over the two-dimensional $(a, M_{\rm p})$ space:

$$\eta_{\rm p} = \iint \alpha_{\rm p}(a, M_{\rm p}) \, {\rm d}a \, {\rm d}M_{\rm p}. \tag{1.14}$$

Note that only a fraction of these planets will transit their host star from our line of sight, as described in Eq. 1.3.

Bringing it all together

Now we can bring together the results from this subsection and look at how survey design affects the number of planet discoveries. We have that the number of planets within our survey volume for stars of a particular type is

$$N_{\rm p} = \theta^2 \eta_{\rm p} \eta_0 \int_0^{d_{\rm max}} \exp\left(-\frac{r\,|\sin b|}{H}\right) r^2\,{\rm d}r.$$

To obtain the number of transiting planets from these, we have to factor in the dependence of the geometric transit probability on the semi-major axis. We therefore use the distribution of planet frequency, $\alpha_{\rm p}(a, M_{\rm p})$ instead of the total number of planets per star. To simplify, we ignore the density variation and just assume an average stellar density of a particular type, n. We then have

$$\frac{{\rm d}N_{\rm p,transit}}{{\rm d}a\,{\rm d}M_{\rm p}} = \frac{\theta^2 d_{\rm max}^3 n}{3} \frac{R_*}{a} \alpha_{\rm p}(a, M_{\rm p}). \tag{1.15}$$

Here we have assumed that the distribution of planet frequency is independent of distance.

We can define a limiting signal-to-noise ratio for which transits can be detected. The SNR decreases with distance, so we are limited by a maximum distance, $d_{\rm max}$, for stars of each spectral type. If we denote the limiting SNR by $L_{\rm SN}$, we have

$$L_{\rm SN} \approx \frac{R_{\rm p}^2}{4(\pi R_*)^{3/2}} \left(\frac{AQ\,\Delta\lambda\,\xi t}{l_{\rm sky}\,a}\right)^{1/2} \frac{\eta\bar{\lambda} L_* \exp\left(-K d_{\rm max}\right)}{hc d_{\rm max}^2 \sigma_{\rm FWHM}}. \tag{1.16}$$

From Eq. 1.16 we can get an expression for $d_{\rm max}^3$:

$$d_{\rm max}^3 \approx \frac{R_{\rm p}^3}{8(\pi R_*)^{9/4}} \left(\frac{AQ\,\Delta\lambda\,\xi t}{l_{\rm sky}\,a}\right)^{3/4} \left(\frac{\eta\bar{\lambda} L_*}{hc\sigma_{\rm FWHM} L_{\rm SN}}\right)^{3/2} \exp\left(-\frac{3K d_{\rm max}}{2}\right).$$

Substituting for $d_{\rm max}^3$ in Eq. 1.15 we finally have

$$\frac{{\rm d}N_{\rm p,transit}}{{\rm d}a\,{\rm d}M_{\rm p}} \approx \frac{\theta^2}{24\pi^{9/4}} \left(\frac{AQ\,\Delta\lambda\,\xi t}{l_{\rm sky}}\right)^{3/4} \left(\frac{\eta\bar{\lambda}}{hc\,\sigma_{\rm FWHM}\,L_{\rm SN}}\right)^{3/2} \times \frac{R_{\rm p}^3}{a^{7/4}} \alpha_{\rm p}(a, M_{\rm p}) \times \frac{n L_*^{3/2} \exp\left(-3K d_{\rm max}/2\right)}{R_*^{5/4}}, \tag{1.17}$$

where the terms are dependent on the survey, the planet, and the star, in order. This equation can be used to quantify the trade-offs between the various equipment choices. On the survey side, we notice that the number of detected planets depends on $\theta^2 A^{3/4}$, meaning it is more important to have a larger sky coverage than mirror size. On the other hand, high photometric precision is harder to achieve with wider sky coverage. The WASP survey (D. L. Pollacco et al., 2006) was built for wide sky coverage since the required photometric precision was just 1%, while NGTS is focused on high precision at the cost of smaller sky coverage. The terms for the dependence on the planet and star also reinforce what we discussed in Section 1.2.1; that we find more planets in close-in orbits around the host star, and preferentially around smaller stars. These principles drove the design of NGTS.

Planet haul

However, when evaulating the expected number of planets that a survey will discover, it is easier to take a different approach. One such approach is to work from the limiting magnitude of a particular survey. This further corresponds to a limiting signal-to-noise ratio for detecting a typical transit of the planet sizes one is interested in. One can then simulate then simulate the expected planet yield by populating a Galaxy with stars, following stellar models and statistics on spectral type distributions and binary systems. These stars can further be populated with planets, following planet statistics mostly uncovered by *Kepler* (F. Fressin et al. 2013; C. J. Burke et al. 2015; and references therein). Such yield simulations for NGTS were carried out by M. N. Günther et al. (2017), finding that the small-planet yield of NGTS over its 4 year mission depends strongly on the level of systematic (red) noise that NGTS would ultimately be able to reach. Another important factor in planet yield simulations is the detection of false positive planet candidates. These are signals that mimick transit lightcurves, but originate from other astrophysical or noise-related phenomena. Proper vetting criteria to identify bona fide planets are then important to not waste telescope time on follow-up observations. As we will see in Section 1.4.3, there are a number of astrophysical phenomena that can mimick a planetary transit.

1.3 Noise

1.3.1 CCD images

A charge-coupled device (CCD) is the most commonly used detector in visible and near-infrared light astronomy. A CCD is made up of an insulator and silicon semiconductor. During an exposure, where the chip is illuminated, a photon is absorbed by the silicon and will excite an electron to the conduction band, which is localised in what we refer to as a "pixel". The number of electrons accumulated in each pixel is therefore (theoretically) linear with the number of photons multiplied by the quantum efficiency of the CCD, which is usually very close to unity. Fig. 1.3 shows the structure of a CCD pixel viewed from the side.

The electrons are captured in potential wells during exposure. After the exposure has ended, the voltage in the pixels along a row of the CCD are varied using the three gates, which has the effect of moving the electrons from one pixel to the other until it reaches the readout area, where it goes through a preamplifier, amplifier, and finally an analogue-to-digital converter which converts the number of electrons to analogue-to-digital units (ADU), related by the gain, $\mathcal{G}$, which is in units of ADU/e^-. The charge transfer process is illustrated in Fig. 1.4.

By tracking the timing of reading out each pixel, the two-dimensional array can be reconstructed after each row has been read, effectively forming an image of the sky that was observed. The raw image, however, needs to be processed before any science can be done from it. There are several sources of noise that will introduce electrons in the pixels that are not from the source itself, affecting the overall signal. These are described further in Section 1.3.3. Many of these signals can be partially removed or mitigated. Bias frames are taken at short exposure and subtracted from the science image. The science image is also corrected for pixel sensitivity differences across the chip by dividing by an image taken while the CCD is uniformly illuminated by white light, referred to as flatfield correction. Once the data has been properly reduced, one can extract the flux from each individual star, referred to as *aperture photometry*. Circular apertures of a few pixels in radius are placed at the centre of the star. The point-spread function (PSF) of the system describes how photons from a point source are spread on the detector. The PSF is then used to measure the total number of electrons that fall within the circular aperture, essentially giving the flux of the star in the photometric band of the detector. A natural next step is to translate the positions

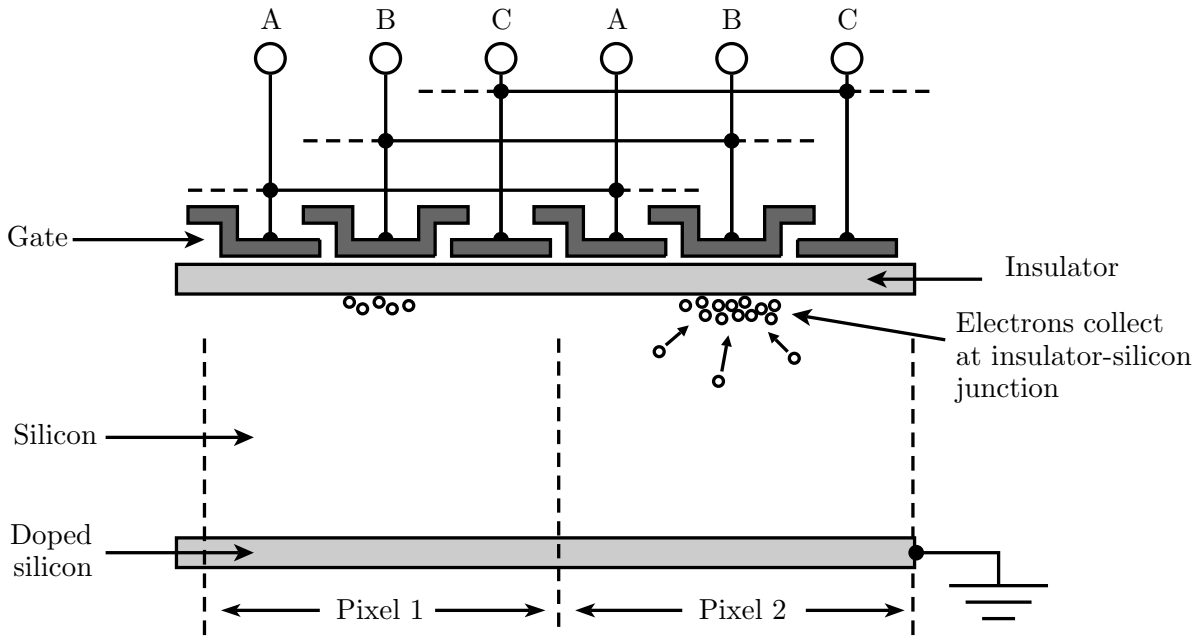


Figure 1.3: An overview of the structure of a CCD pixel and the accumulated charge after an exposure.

of the pixels on the CCD to sky coordinates, which is done by fitting the image to a reference image of the sky with astrometric position standards. The image needs to be corrected for the sky background level by implementing an algorithm that estimates the photon count from the sky itself. Once a smooth sky background level is found, it is typically subtracted from the science image. Several tools have been developed to do this automatically, one such example being CASUTOOLS from the Cambridge Astronomy Survey Unit.

However, noise values that are subtracted from the science images in the data reduction are mean values, and will have an uncertainty. This uncertainty is not removed in the data reduction, and is propagated to the science image. These will contribute to the noise level of the data.

1.3.2 Statistical noise

After an observation, our data is in its simplest form

$$\text{data} = \text{true signal} + \text{noise},$$

where the level of noise depends on various factors and originates from various sources. The statistical fluctuation (error) from the counting of discrete independent events, such as the arrival of individual photons on the CCD, or the number of electrons they excite, is determined by Poisson statistics. The standard deviation from this counting of events is just

$$\sigma = \sqrt{N},$$

where N is the number of detected photo-electrons.

Statistical (white) noise is typically characterised as following Gaussian statistics and will bin down with the square root of the number of frames (observations) averaged together. White noise is typically well understood and can originate from the object, instrument, sky background, and atmosphere. In

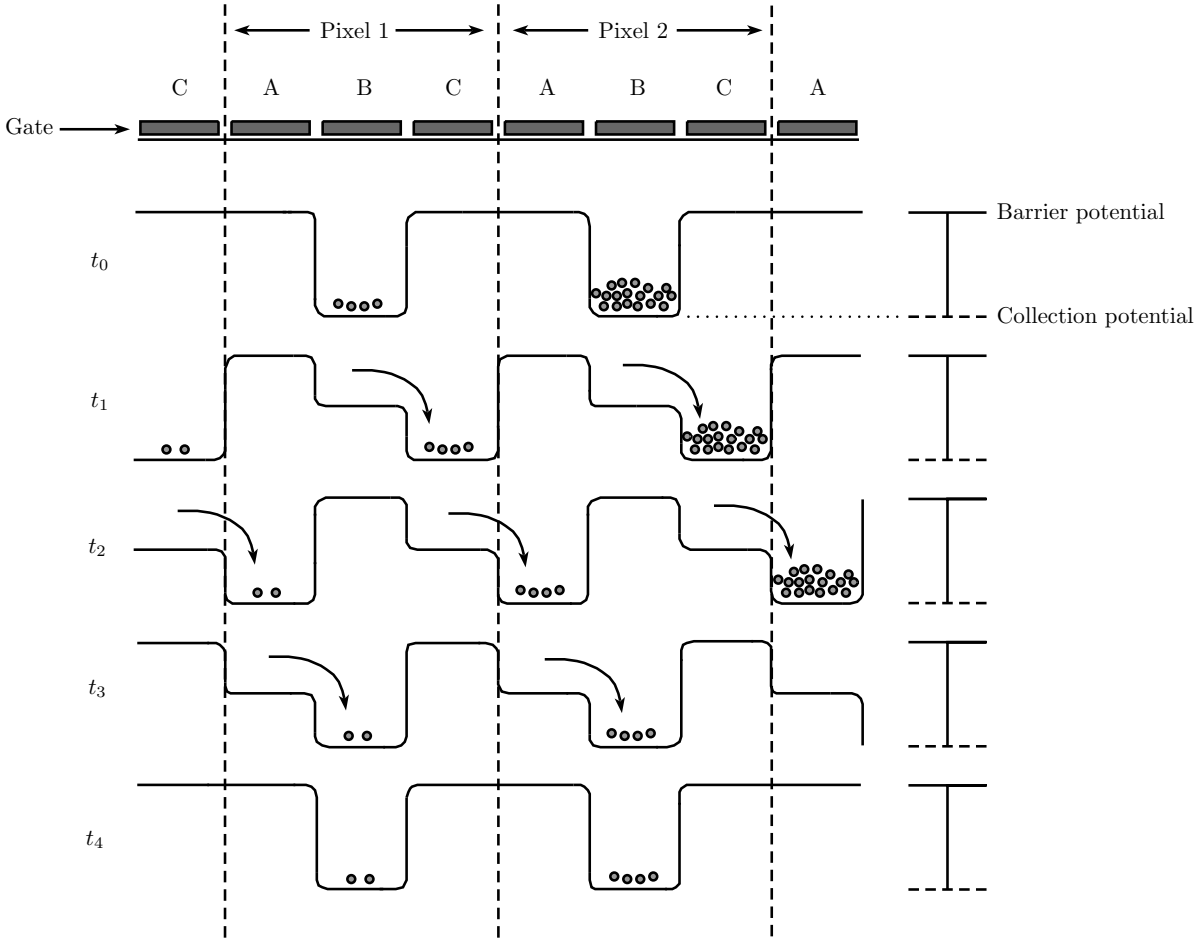


Figure 1.4: Illustration of the charge transfer process during readout.

addition to white noise, any experiment may suffer from systematic noise with unknown origin. This is commonly referred to as red noise, and can come from various systematic errors from the software, instrument, or environment.

1.3.3 Statistical noise sources in CCD images

Read noise

Read noise is introduced into the image during the readout process after an exposure. The read noise is usually found from bias images, where several averaged bias frames are subtracted from a target bias frame, and the remaining noise is taken to be the readout noise. CCD manufacturers usually quote the read noise level as a number of e^- RMS, effectively giving the standard deviation around the mean pixel value. Given a rms value R^2_{read}, and aperture radius r, the read noise is

$$\sigma_{\text{read}} = \sqrt{\pi r^2 \times R^2_{\text{read}}} = \sqrt{n_{\text{pix}} R^2_{\text{read}}}, \tag{1.18}$$

with n_{pix} being the number of pixels within the aperture.

Dark current

Dark current frames are taken at the same cadence as the observations, but without the CCD being illuminated. The temperature in the instruments themselves will sporadically introduce electrons into the pixels. Dark current noise is typically negligible in most modern experiments as the detectors are kept in a cold environment. However, for CCDs operating at near-IR wavelengths, dark current can still be significant.

If the dark noise level is D, in units of e^-/pixel/s, then the noise due to dark current is

$$\sigma_{\text{dark}} = \sqrt{\pi r^2 \times D \times t_{\text{exp}}} = \sqrt{n_{\text{pix}} \times D \times t}. \tag{1.19}$$

Shot noise

The uncertainty in the detected photons from the source itself is referred to as shot noise, or source noise. If the source provides R_{star} photo-electrons per second, then the total number of electrons collected during an exposure is $R_{\text{star}} \times t_{\text{exp}}$. The error is then

$$\sigma_{\text{star}} = \sqrt{R_{\text{star}} \times t_{\text{exp}}}. \tag{1.20}$$

Scintillation noise

Scintillation, or twinkling, is the term for variations in the brightness of a star because the light is passing through different layers of a turbulent atmosphere. The rms error due to the low-frequency component of scintillation can be approximated as (D. Dravins et al., 1998)

$$\sigma_{\text{scint}} = 0.09\, D^{-2/3}\, (\sec Z)^{7/4}\, \frac{\exp\left(-h/h_0\right)}{(2T)^{1/2}}, \tag{1.21}$$

where D is the aperture diameter in centimeters; $\sec Z$ is the airmass at the site; h is the observatory's height above sea level; h_0 is the atmospheric scale height, $h_0 \simeq 8\,000$ m; T is the exposure time during an observation. However, Eq. 1.21 was found after doing experiments in La Palma. J. Osborn et al. (2015a) found correction factors for various locations of astronomical observatories.

Sky background noise

The sky background has an intrinsic brightness that will contribute to the image. The sky background is typically subtracted from the final image, but the noise of the sky estimate will remain in the reduced image. Given the photo-electron rate from the sky background, R_{sky} in units of e^-/pixel/s, the statistical uncertainty in the background level is

$$\sigma_{\text{sky}} = \sqrt{\pi r^2 \times R_{\text{sky}} \times t_{\text{exp}}} = \sqrt{n_{\text{pix}} \times R_{\text{sky}}}. \tag{1.22}$$

1.4 Variable stars

Variable stars are stars that undergo brightness changes over time. Fig. 1.5 shows a non-exhaustive variability tree for variable stars. There are generally two groups of variable stars, referred to by whether the reason behind their variability is intrinsic to the star itself, or due to extrinsic factors.

Pulsating variable stars change their brightness periodically due to continuous expansion and contraction of the star. These classes of variable stars mainly inhabit the instability strip on the Hertzsprung-Russell (HR) diagram, as shown in Fig. 1.6. While pulsating, they will move up and down the strip, changing spectral types periodically. The pulsation comes from variations in the rate at which the radiation from the core of the star can escape. These stars are harmonic oscillators and can have more than one period at which they oscillate. Eruptive variables are transient are typically massive stars with high activity.

For extrinsic variables, eclipsing binary stars are a big class. The orbital planes of these double-star systems can be oriented in such a way that they will periodically eclipse each other as seen from the observer. Rotating stars generally encompass stars with high magnetic activity and strong magnetic fields. Here we can also talk about regular main sequence stars that have spot-covered surfaces that lead to low-amplitude brightness variations, which with NGTS' precision, we will be able to pick up.

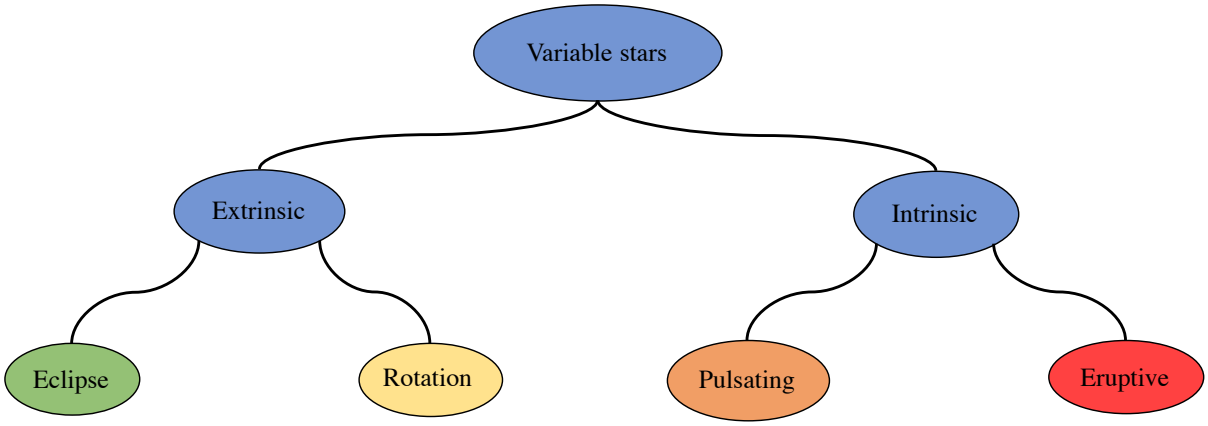


Figure 1.5: Variability tree showing the general groups of variable stars.

Here we present a few classes of variables that will be useful for this work. Their types follow classifications from the General Catalogue of Variable Stars[1].

1.4.1 Pulsating variables

Cepheids (CEP)

Cepheids are a type of radially pulsating, high luminosity variables. Their periods range from 1–135 days, and amplitudes from 1–10 per-cent. While pulsating, their spectral type range from F at minimum to G-K at maximum. Cepheids follow a strict relationship between the period of pulsation and its luminosity. Because of this reason, Cepheids and other similar variables are used as distance indicators in our Galaxy, as well as in other nearby galaxies. Classical Cepheids are young, massive supergiants with periods from a few days to months.

RR Lyrae (RR)

These stars are less luminous than Cepheids, and have periods ranging from several hours to about a day, with high amplitudes similar to Cepheids. These stars can have variable lightcurve shapes and periods, referred to as the Blazhko effect. RR Lyrae stars are generally subclassed into RR(B), RRAB,

[1] http://www.sai.msu.su/gcvs/gcvs/vartype.htm

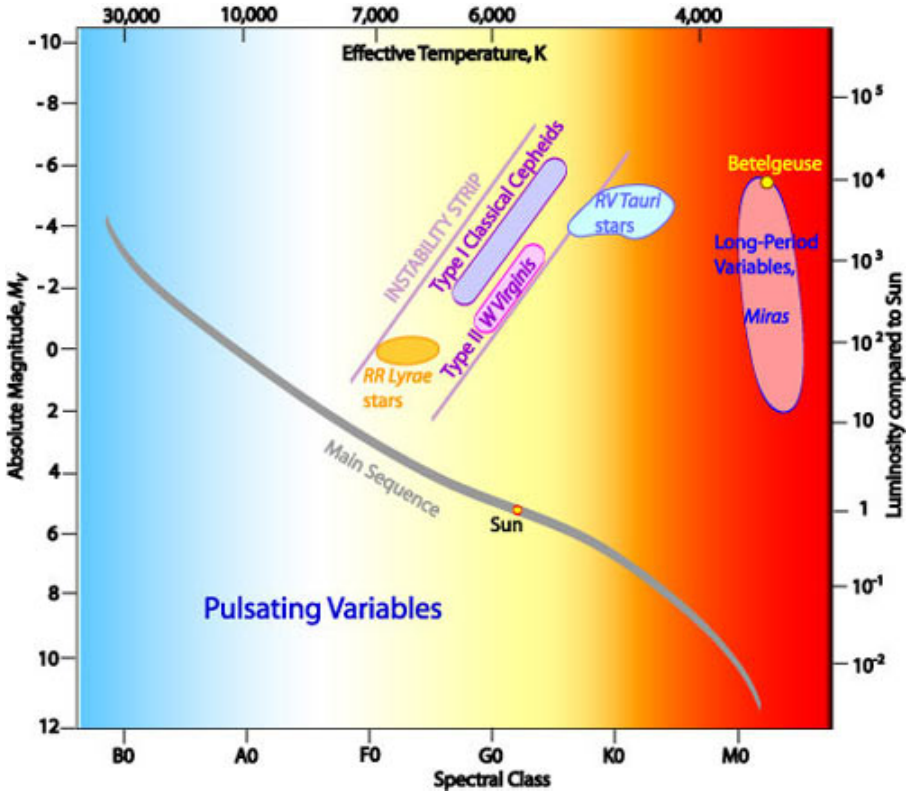


Figure 1.6: The Hertzsprung-Russel diagram showing the locations of common pulsating variables. Credit: Australia Telescope National Facility.

and RRC. The first showing two pulsation modes with the dominant period P_0 and its first harmonic P_1. RRAB have asymmetric light curves with steep ascencion to maximum and longer recovery descent, with periods from 0.3–1.2 days. RRC have close to symmetric lightcurves resembling a sinusoidal pattern, with periods from 0.2–0.5 days and lower amplitude than RRAB. See Fig. 1.7 for reference.

δ Scuti (DSCT)

Delta Scuti type are pulsating variables of spectral types A0-F5 with short periods and small amplitudes. Periods are generally from 0.01–0.2 days, with amplitudes ranging from 1–80 per-cent. Their shape, amplitude, and periods usually vary greatly, making them difficult to classify. See Fig. 1.8, upper panel, for reference.

Semiregular variables (SR)

Semiregular variables are giants or supergiants up to late spectral types showing periodicity interrupted by irregularities. Periods are generally ≥ 20 days. These are divided in several subclasses based on whether their periodicity is persistent (SRA) or poorly defined (SRB), and amplitude ranges. See Fig. **??**, lower panel, for reference.

1.4.2 Eclipsing binary stars (E)

Eclipsing binary stars (EBs) are a type of extrinsic variables. They are binary systems with orbital planes close to our line of sight, i.e. the inclination i is close to 90°, so we view the system edge-on. In this configuration the two components will periodically eclipse each other, blocking a fraction of the light of the other every half period, producing dimming in the lightcurves. EBs can be found at a range of periods, from shorter than one day, to several years. Eclipsing binary stars are generally subclassed into EA, EB and EW systems. The following classifications are based on lightcurve characteristics alone, while other classifications can be based on the level of Roche lobe filling. See Fig. 1.7 for reference.

Detached binaries (EA)

Algol-type binaries that are detached systems with their Roche lobe not significantly filled. They can have spherical or slightly ellipsoidal components, but it is possible from their lightcurves to specify the ingress and egress of the eclipse. Periods can range from 0.2–10 000 days, and secondary minima can be absent.

Contact/semi-detached binaries (EB)

These are Beta Lyrae-type eclipsing systems that have significant ellipsoidal components that makes it impossible to specify the exact time of beginning and end of eclipses due to a continuous, smoothly varying change in brightness throughout the lightcurve. Secondary eclipses are always observed, and periods are generally larger than 1 day. Not to be confused with a general eclipsing binary (EB).

Contact binaries (EW)

Binary systems with periods shorter than 1 day with ellipsoidal components almonst in contact. These systems have their Roche lobe completely filled, and the depths of the primary and secondary minima are very similar. Their depths are usually less than 10 per-cent.

1.4.3 False positive planet candidates

Some variability types will have lightcurves that can mimick planetary transits. Most notable of these are eclipsing binaries, as they produce transit signals the same way that a planet would. Often these signals are much deeper than one would expect from a transiting planet. If the difference between the primary and secondary eclipse is also large, it is easy to distinguish an eclipsing binary system from a transiting planet. However, the situation gets slightly more complicated when the two binary stars are of similar size, as they both produce close to identical eclipse depths. Furthermore, if their orbital plane not viewed edge on, but slightly tilted, the eclipses will not be full and the transit depth will be closer to a planetary signal. The transit signal from these *grazing* eclipses will be V-shaped (Fig. 1.7, top), and in many cases it is easy to distinguish the V-nature of the transit from the traditional U-shape with varying degrees of flat bottom due to limb-darkening effects of the star.

The situation worsens if there is an eclipsing binary system within the aperture of the target star that is not spatially resolved. These are referred to as background eclipsing binaries (BEBs). The aperture receives the combined light from both systems, and if the eclipsing binary in the background undergoes an eclipse, the eclipse signal will be diluted by the foreground star. Because of the dilution, it will manifest as a potentially U-shaped transit signal at a depth consistent with a planet. These systems would often be uncovered in the follow-up radial velocity measurements of the star. However, in order to minimise

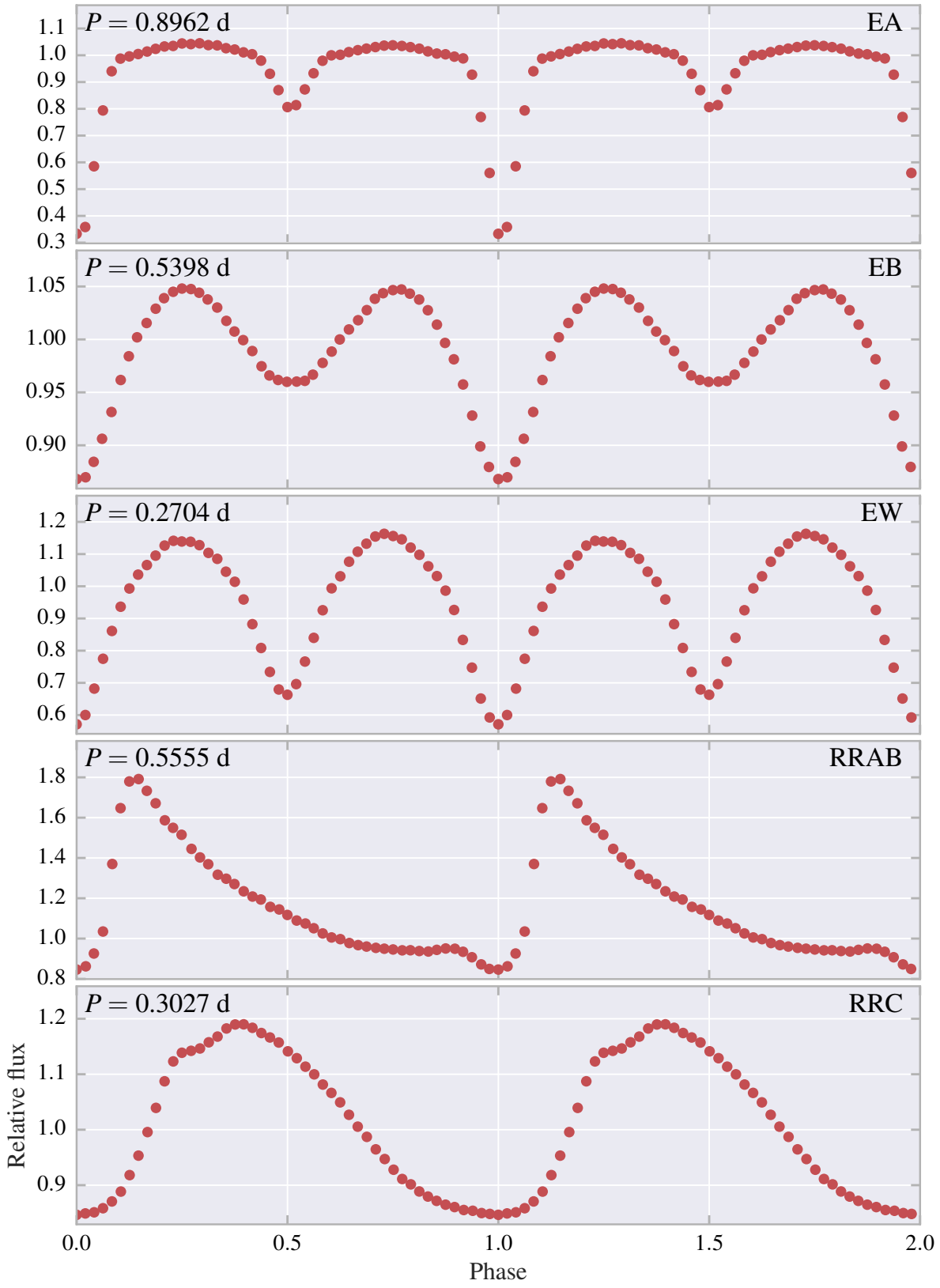


Figure 1.7: 5 short-period variable lightcurves from the AAVSO catalogue with NGTS data. From top to bottom: EA, EB, EW, RRAB, RRC.

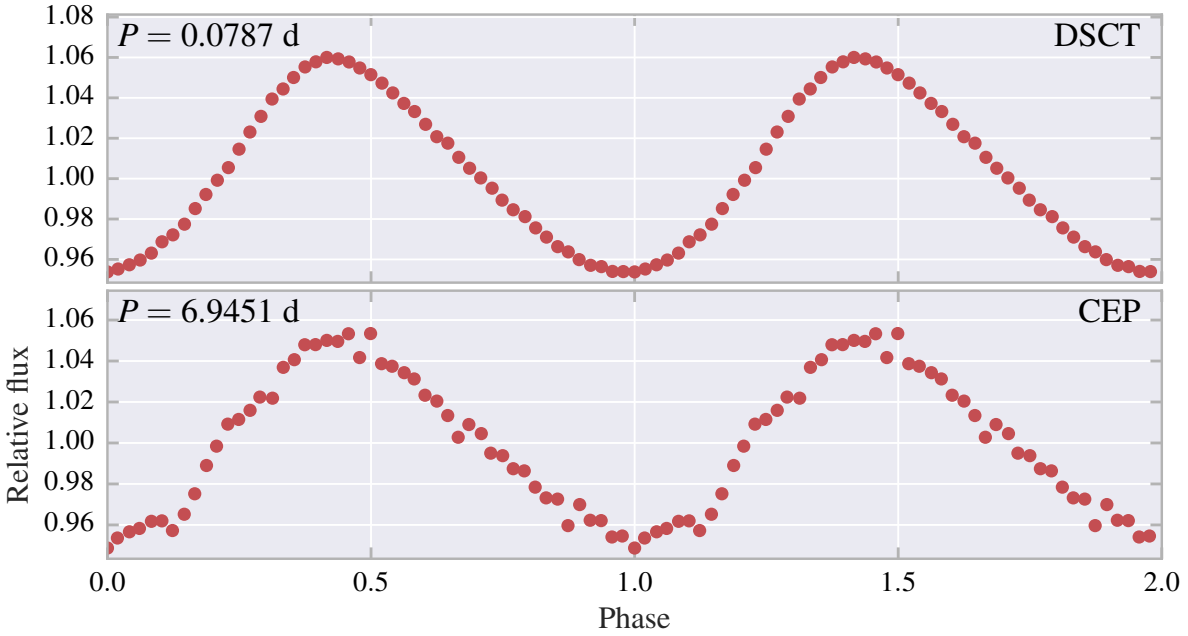


Figure 1.8: Top: δ Scuti (DSCT) variable from the AAVSO catalogue. Bottom: Likely Cepheid variable identified from NGTS lightcurves. δ Scuti variables as sometimes referred to as dwarf Cepheids due to their similar shape, but classical Cepheids have longer periods.

wasting telescope time on false-positives, it has become paramount to vet these systems. One way of doing this is to look at the centroid shift of the target star. The centroid shift is the difference between the centre of the aperture, and the centre of flux in the aperture. During an eclipse, the background system will undergo a loss of light, which will shift the centre of flux away from the system during the transit signal. If this shift is seen, it is a good indicator of a background system contaminating the aperture. This method requires very high precision on the photometry and aperture position, but has been done successfully for *Kepler* (N. M. Batalha et al., 2010), and recently also for NGTS for the first time from the ground (Günther et al., in prep.).

CHAPTER 2

Next Generation Transit Survey

The Next Generation Transit Survey (NGTS) is a wide-field photometric survey designed to discover transiting Neptune-sized and smaller exoplanets around bright stars ($V < 13$) (P. J. Wheatley et al., 2013; B. Chazelas et al., 2012). The facility is located at Paranal Observatory in Paranal, Chile, and is the first telescope project hosted, but not operated, by the European Southern Observatory (ESO) at the site. The excellent photometric conditions at the site, combined with state-of-the-art instruments, allows NGTS to reach 0.1% photometric precision over a wide field. This is the first telescope project to achieve such a precision from the ground. NGTS achieved its first light with one telescope in early 2015, and with all 12 telescopes in early 2016.

The NGTS project is a partnership between several UK universities (Warwick, Cambridge, Leicester, Belfast), Observatoire de Genève, DLR Berlin, and Universidad de Chile.

2.1 Design

Figure 2.1: The NGTS telescopes on their mounts withing their enclosure at Paranal Observatory. (Credit: ESO/R. West)

Table 2.1: Characteristics of NGTS.

Site	ESO Paranal Observatory
Number of telescopes	12
Diameter	20 cm
Aperture	$f/2.8$
CCD	2048×2048 deep-depleted red-sensitive
Mount	Equatorial
Bandpass	600–900 nm
FOV	88.8 square degrees (7.4 × 12)
Lifetime	~4 years
Cadence	10 s

2.1.1 Telescope and survey design

The design of NGTS builds on the experience from the Wide Angle Search for Planets (WASP; D. L. Pollacco et al. 2006), a successful wide-field survey for hot-Jupiters, with more than 132 such exoplanets confirmed to date[1]. NGTS is comprised of an array of twelve independent fully robotic 20-cm $f/2.8$ telescopes mounted on fork equatorial mounts. The optical design is a Newton reflector with a hyperbolic primary and a corrector. The cameras are 2048×2048 pixel red-sensitive deep-depletion CCDs designed for optimal sensitivity in the 600–900 nm range, thus targeting K and early-M spectral type stars. Each telescope has a field of view of 7.4 square degrees, giving a total field of view of 88.8 square degrees. The NGTS characteristics are summarised in Table 2.1.

The nominal survey lifetime is about four years. NGTS covers a new field every few months, and will at the end of its life have covered an area of about 8 times the size of the *Kepler* field. The photometric band is most sensitive to K and early-M stars, and together with sub-mmag red noise levels, NGTS can detect transit signals from Neptune-sized planets and smaller for stars of magnitude $V < 13$. NGTS thus aims to provide prime targets for radial velocity follow-up and characterisation of planetary interiors and atmospheric composition.

2.1.2 Science goals

The orientation and inclination of the orbital plane of a planet orbiting its host star follows a uniform distribution. The geometric probability of a Neptune-sized planet transiting a target star is about 5%. Occurence rates from Kepler shows the planet occurence of Neptunes with periods less than 10 days is also about 5%. Thus NGTS needs to observe a minimum sample of about 40 000 stars to observe about 100 transiting short-period Neptunes. Furthermore, the host stars should be bright enough to allow radial velocity follow-up, placing an upper limit around V=15. However, smaller host stars have larger reflex motions from the orbiting planet, thus increasing the upper limit for M stars.

Yield simulations of the planets and false positives for NGTS over its 4 year mission show that NGTS expects to find 4±3 super-Earths, 35±9 Neptunes, 55±8 Saturns, and 150±10 Jupiters, for a red noise level of 1 mmag (M. N. Günther et al., 2017). Eclipsing binaries (EBs) and background eclipsing binaries (BEBs) can produce signals that mimic that of a transiting planet, and are as such the most common astrophysical false positives in wide-field surveys for transiting planets. However, EBs can introduce other effects to the transit light curve that can be used to distinguish them from a planet. The transit depth is usually much deeper than that of a planet, and binaries orbiting close to each other have a distorted shape which introduces ellipsoidal variations in the light curves. One can also detect secondary transits

[1] http://www.exoplanets.eu (Oct 2016)

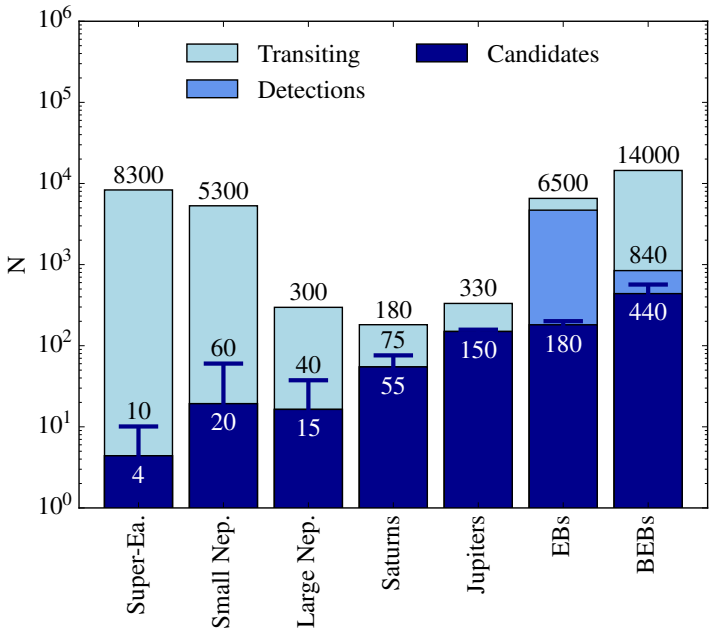


Figure 2.2: The expected yield for planets and false positives (EBs and BEBs) for NGTS from M. N. Günther et al. (2017). The light blue areas are the number of objects producing a transit event in the line of sight with $P < 20$ days. Blue is the number of objects that can be detected with a red noise level of 1 mmag. The dark blue areas are the number of planetary candidates after vetting against astrophysical false positives. The blue lines indicate the possible yield for a red noise level of 1 mmag.

(the occulation of the smaller star from the larger star), giving rise to different depths for odd and even eclipses. Transiting candidates can thus be vetted against these effects to identify the false positive nature of some of the transiting candidates. Before false positive vetting, NGTS expects to find 4688±45 EBs and about 843±75 BEBs as false positive planet candidates, compared to about 180 ± 12 and 439 ± 37 after vetting, for a red noise level of 1 mmag. A large number of EBs are expected to be diluted by foreground stars, given that the aperture radius is 3 pixels. The yield results are summarised in Fig. 2.2.

2.2 Pipeline

2.2.1 Data reduction and data flow

The NGTS observations are taken with 10 s cadence and are bias-subtracted, corrected for shutter travel time, and inter-pixel sensitivity variations and vignetting using flatfield exposures. The bias and flatfield correction frames used are master frames constructed from a couple of months of data. Dark frames are currently not used in the reduction as the dark current is negligible. The reduced images then go through the input catalogue, which defines where the apertures are placed. The input catalogue is constructed from a pre-survey of each field, where around 100 images are taken and stacked together and searched for sources using IMCORE from CASUTools[2]. The source list is then cross-matched with the 2MASS and APASS catalogues to cut sources with $I > 16$ are filtered out. A custom version of the WCSFIT tool computes an astrometric solution based on the input catalogue. The tool IMCORE_LIST then uses the astrometric solution computed by WCSFIT to compute the detector coordinates of the sources in the

[2] http://casu.ast.cam.ac.uk/surveys-projects/software-release

NGTS input catalogue for the field, and the apertures are placed at the detector coordinates. At this stage, aperture photometry is carried out on the output from IMCORE_LIST, measuring the fluxes within an aperture using an RCORE of 3.5 pixels. The resulting photometry files are next sent to a flagging tool that flags for e.g. saturation, CCD edge proximity, and apertures that are affected by the Very Large Telescope's (VLT) guide star lasers.

2.2.2 Systematic error removal

The reduced data from the pipeline inevitably contains systematic errors that can originate from a variety of sources, such as instrumental effects, pointing drift, PSF changes over the detector, environmental effects and atmospheric extinction. These low-level errors can be corrected for by noting that any systematic error should be present in a large sample of targets. The SYSREM algorithm of O. Tamuz, T. Mazeh, and S. Zucker (2005) was developed for the purpose of removing systematic errors in large sets of photometic light curves, and the NGTS pipeline uses an implementation of the SYSREM algorithm that is similar to that used in the WASP survey (A. Collier Cameron et al., 2006; D. L. Pollacco et al., 2006).

2.2.3 Sky background correction

The sky background correction algorithm is the standard correction tool used by IMCORE for the source extraction. The CCD is divided into cells of size 64×64 pixels. The raw background values are then derived in each cell using MAD iterative k-sigma clipping from the median to filter away stars. The cells undergo further filtering to produce a low-resolution background level, then a local background for each pixel is defined using bilinear interpolation. A global sky level and noise level are also estimated after the varying component is removed, and stored under the SKYBKG HDU name in the final data product.

2.2.4 Transit search and candidate vetting

BLS

Following the data reduction and correcting for systematic effects, all photometry is passed on to the ORION tool that performs a transit search on a grid of trial periods using an implementation of the Box-Least-Squares algorithm (BLS). The lightcurves are phase folded on different periods and a least-squares fit on a transit box model is carried out. The best 5 periods are extracted, as well as the corresponding transit widths, depths, and epochs. A range of parameters is further extracted from the lightcurve and nearby lightcurves to assess whether the identified period is likely to be a systematic period or astrophysical. Typically, ORION can identify up to 25% BLS candidates in a field. These are mostly from variable stars, such as eclipsing binaries, background eclipsing binaries, other variable stars, or from systematic periods.

Vetting tools

Candidates from BLS are further processed with CANVAS (**Can**didate **V**etting and **A**nalysi**s**), which generates a list of parameters and criteria that rejects most candidates. Some parameters that are evaluated are reduced proper motion of the star to exclude possible giants, reduced flux density points that may cause a false transit signal, isochrone for the star to assess the stellar radius, and the stellar density. Other tools were developed to check other aspects, such as difference between odd and even transits to exclude eclipsing binaries, proper transit modelling, likelihood of transit given the noise in the data, and nights with individual transits given the model. One of the most significant false-positives that are the most difficult to rule out are background eclipsing binary systems, see Section 1.4.3. A centroiding tool

Table 2.2: The structure of the NGTS data format in the FITS files with names and types as shown in the files. For the dimensions, $N_{\text{stars}} \times N_{\text{obs}}$ are the number of stars and number of observations, respectively. The CCD positions are the positions calculated from the astrometric solution from WCSFIT. Centroid is the position of the centre of flux of the star.

HDU Name	HDU Type	Dimensions	Description
PRIMARY	PrimaryHDU		
CATALOGUE	BinTableHDU	$N_{\text{stars}} \times 16$	Information per star (ID, mean flux, coordinates, ...)
IMAGELIST	BinTableHDU	$N_{\text{obs}} \times 197$	Information per observation (camera, PSF, seeing, ...)
HDJ	ImageHDU	$N_{\text{obs}} \times N_{\text{stars}}$	Heliocentric Julian Day, time of observation
FLUX	ImageHDU	$N_{\text{obs}} \times N_{\text{stars}}$	SYSREM-corrected flux (ADU per second)
FLUX_ERR	ImageHDU	$N_{\text{obs}} \times N_{\text{stars}}$	Error in flux
FLAGS	ImageHDU	$N_{\text{obs}} \times N_{\text{stars}}$	Photometric quality flags (bad weather, light contamination, ...)
CCDX	ImageHDU	$N_{\text{obs}} \times N_{\text{stars}}$	CCD position of star in x-direction
CCDY	ImageHDU	$N_{\text{obs}} \times N_{\text{stars}}$	CCD position of star in y-direction
CENTDX_ERR	ImageHDU	$N_{\text{obs}} \times N_{\text{stars}}$	Centroid error in x-direction
CENTDX	ImageHDU	$N_{\text{obs}} \times N_{\text{stars}}$	Centroid in x-direction
CENTDY_ERR	ImageHDU	$N_{\text{obs}} \times N_{\text{stars}}$	Centroid error in y-direction
CENTDY	ImageHDU	$N_{\text{obs}} \times N_{\text{stars}}$	Centroid in y-direction
SKYBKG	ImageHDU	$N_{\text{obs}} \times N_{\text{stars}}$	Sky background (ADU per second per pixel)

was developed to rule out these false-positives, being able to reach $1/1000$ pixel precision for our data (M. Günther et al. 2017, in prep.).

2.2.5 Data product

The end data products are stored in Flexible Image Transport System (FITS) format, with one file per field. FITS files are comprised of segments of Header/Data Units (HDUs) that contain a Header Unit with metadata in human-readable ASCII format, followed by a Data Unit. The first HDU is called the Primary HDU, and any number of additional HDUs may follow the primary. The additional HDUs can have three types of extensions:

- Image Extension: 0-999 dimensional array of pixels.
- ASCII Table Extension: Tabular information stored in ASCII format.
- Binary Table Extension: Tabular information stored in binary format.

The NGTS FITS files contain 14 HDUs of varying dimensions, as shown in Table 2.2, with the Primary HDU being empty.

Given a crowded field ($\sim$20 000 stars) observed over a period of four months, the resulting FITS files can have sizes of up to 90 GB. Thus reading some of the bigger files to memory should be done on a computer cluster with enough memory available. For the purpose of working with NGTS data, the python[3] package `ngtsio`[4,5] was developed for memory efficiency and ease of use.

[3] https://www.python.org
[4] https://pypi.python.org/pypi/ngtsio (available through `pip`)
[5] https://github.com/MNGuenther/ngtsio

2.3 Operations

2.3.1 Start of science observations

Science operations of NGTS started on 21 September 2015 with 4 cameras. Six weeks later, two more cameras were installed, followed by another two a fortnight later. These 8 telescopes observed ~400 hours per week with little downtime. The final 4 cameras were installed at the end of February 2016. since.

2.3.2 Downtime and issues

Shortly after the installation of the remaining cameras, the facility had to shut down for one month due to rats (!) having gnawed through cables that promptly had to be replaced. Upon resuming observations, the Chilean winter was upon Paranal with El Niño and bad weather. During a period of 12 weeks the facility averaged around 250 hours of observations per week. As the bad weather passed the facility had to shut down for two weeks due to an issue with the roof. After this, normal observations resumed with little downtime, averaging at 600 hours per week. One camera was not able to observe due to a shutter failure, which took some time to replace, resulting in some sustained downtime until the end of 2016. Shortly after all telescopes became operational, there were issues with stray light coming into the telescope, upon which longer baffles were installed to limit the amount of non-target light that entered the telescope.

CHAPTER 3

Data processing and analysis

This Chapter describes the general procedure that was followed when working on the data. In Section 3.1 we introduce the dataset that we used for this work, and how it has been processed in the NGTS standard pipeline. In Section 3.2 we match our targets to known catalogues. Section 3.3 describes the software that was written to analyse our photometry and noise in our fields, and finally Section **??** describes how it can be used to identify variable stars.

3.1 Observations and data reduction

The data used in this work is tagged as TEST16A, and was released in summer 2016. As of January 2017, new data under the tag TEST18 has been released, but for consistency we continued using TEST16A data for most of our analysis, with the exception of a few cases where we compare with TEST18 data. An overview of the observations are given in Table 3.1. The TEST16A data include observations from 30 fields, with coordinates (α, δ) in the range $03^{\mathrm{h}}04^{\mathrm{m}}$ to $23^{\mathrm{h}}46^{\mathrm{m}}$ and $-66°41'$ to $+08°26'$ and include ~280 000 stars. The nightly observations were taken between 21 Oct 2015 – 3 Jul 2016 and were reduced by the NGTS standard pipeline, which has been described in Section 2.2, but we will reiterate the most important steps here. The pipeline outputs 10 s cadence data that is bias-subtracted, flatfield-corrected, and corrected for shutter travel time. The data is further corrected for systematic effects with an implementation of the SYSREM algorithm (O. Tamuz, T. Mazeh, and S. Zucker, 2005), which searches for common flux variations in a subset of stars based on their rms level and brightness. The algorithm identifies four such features that are described by their basis functions, which are used to correct for the systematic effects.

3.2 Catalogue cross-matching

At the time of this work, the NGTS fields had not officially been cross-matched with existing catalogues (other than the input catalogues for the positions in the pipeline). Catalogue matching gives access to important stellar information useful for further work, so we include a quick summary of the process.

The TEST16A dataset was cross-matched with the catalogues below. For each NGTS star we searched catalogues for entries within a $10''$ radius of the NGTS coordinates, and chose the closest one as a match.

Table 3.1: Field names and characteristics used in the TEST16A dataset, sorted by right ascension coordinate. The name of each field starts with the letters "NG" followed by four digits that denote the right ascension at the centre of the field in units of hour angle, followed by a sign and four digits that denote the central declination in units of degrees.

Field	Camera	Date range	Span (days)	N_{nights}	N_{images}	$N_{\text{apertures}}$
NG0304 −1115	809	20151104–20160301	118	94	114 663	3 691
NG0313 −2230	810	20151104–20160225	113	80	118 880	3 838
NG0348 −3345	811	20151104–20160201	89	77	128 415	3 568
NG0409 −1941	812	20151118–20160317	120	90	122 125	4 860
NG0522 −2518	802	20150922–20160503	224	150	212 092	8 505
NG0531 −0826	806	20150924–20160421	210	134	183 210	11 194
NG0612 −2518	805	20150922–20160514	235	158	218 974	15 092
NG0618 −6641	801	20150922–20160515	239	158	233 945	11 362
NG1135 −2518	809	20151127–20160515	170	99	128 919	6 870
NG1213 −3633	810	20151129–20160525	178	89	106 302	10 181
NG1253 −1941	803	20160418–20160703	76	23	31 456	6 700
NG1315 −2807	812	20160106–20160515	170	99	96 291	8 371
NG1318 −4500	811	20151219–20160623	187	72	91 497	23 620
NG1340 −3345	808	20160418–20160703	76	33	50 976	11 596
NG1349 −1115	804	20160419–20160703	75	22	31 434	5 172
NG1416 −2518	806	20160106–20160515	130	68	84 404	12 032
NG1421 +0000	805	20160106–20160515	130	66	66 257	3 999
NG1428 −2518	802	20160106–20160518	133	66	75 433	9 217
NG1444 +0537	801	20160114–20160515	122	61	59 483	3 612
NG1947 −4200	806	20150922–20151201	70	44	36 999	14 551
NG2028 −2518	802	20150921–20151201	71	47	45 415	13 952
NG2047 −0248	810	20160421–20160622	62	17	8 701	20 313
NG2058 −0248	803	20160421–20160627	67	20	12 955	18 656
NG2126 −1652	804	20160421–20160627	67	23	22 748	6 357
NG2132 +0248	805	20160421–20160627	67	22	18 543	7 516
NG2142 +0826	801	20160421–20160627	67	21	13 835	8 069
NG2145 −3345	808	20160421–20160627	67	18	10 460	5 324
NG2150 −3922	812	20160421–20160623	63	20	15 399	5 210
NG2152 −1403	811	20160421–20160622	62	17	10 487	6 012
NG2346 −3633	802	20160504–20160627	54	18	14 057	3 498

Tycho-2

The Tycho-2[1] catalogue (E. Høg et al., 2000) contains proper motions and magnitudes (B and V) for 2.5 million of the brightest stars ($V < 12$), collected by the *Hipparcos* satellite of the European Space Agency. Proper motions are precise down to about 2.5 mas yr^{-1}, and components of double stars down to 0.8″ are included in the catalogue. We find ∼10 000 Tycho-2 stars in our fields, or about 4% of our current sample of stars.

Gaia

The first data from Gaia (DR1[2]) was released on 14 September 2016 and contains the full five-parameter astrometric solutions for 2.5 million stars from Tycho-2, as well as positions and single-band G magnitudes for 1 billion stars brighter than G=20.

[1] http://cdsarc.u-strasbg.fr/viz-bin/Cat?I/259
[2] http://www.cosmos.esa.int/web/gaia/dr1

Table 3.2: Results from catalogue cross-matching with our sample of stars.

Catalogue	Number of matches	% of our stars	Number of fields
Tycho-2	~ 10 000	3.7	30
Gaia	~260 000	94	30
GCVS	142	0.05	15
AAVSO	453	0.15	30
ACVS	226	0.08	30

GCVS

The General Catalogue of Variable Stars contains ~48 000 entries of variable stars with periods, amplitudes, magnitudes, and variability classification based on both photometry and spectroscopy. We find 142 matches with our sample of stars.

AAVSO

The American Association for Variable Star Observers maintain a variable star catalogue with ~400 000 entries of variable objects with period, magnitudes, and variability type. The catalogue is however mostly supported by amateur observations, therefore being prone to misclassifications. We find 453 matches for our sample of stars, most of them being classes that are easily recognised from photometry alone, e.g. various types of eclipsing binary stars and RR Lyraes. It is worth noting however that AAVSO contains variable stars from GCVS as well, and likely a number of these matches are from GCVS.

ACVS

The variable star catalogue from the All Sky Automated Survey (ASAS) holds ~50 000 variable stars, surveyed in the V and I bands. We find 226 matches with our stars, most of which are short-period high-amplitude variables such as RR Lyraes and eclipsing binaries.

The matching results are summarised in Table 3.2. Here we note that only about 4 per-cent of our objects are brighter than the *Hipparcos* faint limit ($V = 12$). The NGTS CCDs saturate at V~9, and is sensitive to V~15. Due to a non-negligible number of mis-matched NGTS objects with the Gaia, 2MASS, and APASS catalogues that provided wrong magnitudes for our targets, the author decided to solely work in electron units for the remainder of the work.

3.3 Data quality

3.3.1 Fractional rms

One common way of assessing the overall data quality of a photometric survey is to study so-called flux-rms diagrams. The stars are plotted by their brightness and root-mean-square (rms) level. A slightly modified version is the fractional rms diagram, which shows the inverse signal-to-noise ratio instead, as shown in the upper panel in Fig. 3.1. The curved shape of the fractional rms plot is determined from the total noise of the observations, which comes from various sources which will be introduced in Section 3.3.2.

The data is read on a per field basis. The data contains N_{frames} flux values at 10 s exposure from N_{LC} light curves, one per aperture on the CCD. The data spans the whole observing season for a given

field, which is up to ~4 months. In every field some apertures contain inconsistent data, which can be due to apertures being placed on stars that are close to the edge of the CCD, or because the star in question is close to the brightness limit, or other spurious effects. These frames are flagged in the pipeline by setting the flux values to zero. We therefore mask all flux data that are zero, and their error. Next, the data is averaged together (binned) to reduce white (Gaussian) noise. We find that binning the data to $t_{\rm bin} = 10$ min exposures provide a good balance between reducing most unwanted white noise, while preserving rapid astrophysical variability from some variable stars such as δ Scutis. We compute a weighted average of the masked flux from $N_{\rm bin} = t_{\rm bin}/t_{\rm exp}$ frames, weighted by the individual errors on the $t_{\rm exp}$=10 s data that are output from the pipeline. It is worth noting, however, that any errors introduced in the pipeline processing (e.g. by SYSREM) are not propagated forward. The binned data is then sigma-clipped at 5σ to remove strong outliers that are likely environmental in nature. The sigma-clipping is done only once, as an iterative approach could possibly clip strong astrophysical variability from variable stars. At this stage the median flux of each light curve is computed. We further calculate the median absolute deviation from the median (MAD) per light curve. We use the (theoretical) error of the MAD divided by the median flux value to calculate the fractional rms, $F_{\rm rms}$, of each light curve. Calculating $F_{\rm rms}$ in this way has the advantage that it is more robust to outlier effects, and can give a better idea of the typical flux variability of the light curves. The result is then a measure of the median noise level per star over the whole period it was observed with NGTS.

3.3.2 Noise model

In order to explain the flux-rms curve in Fig. 3.1 we need to build a noise model. Being able to characterise and understand the noise of our instruments and survey is fundamental for the NGTS experiment, as we are looking for transit signals at the mmag level.

White noise

White noise comes from the statistical uncertainty in the counting of photons, as previously described in Chapter 1. The various flux values in the data files are given in units of ADU $\rm s^{-1}$. The conversion back to physical units ($\rm e^-$) requires one to know the characteristics and settings of each camera at the time of processing. For the camera settings that NGTS uses, the characteristics are tabulated in Table 3.3. We multiply all flux values with the gain, $\mathcal{G}$, specific to the camera used, and our cadence, $t_{\rm exp}$. Our data is then the number of $\rm e^-$ after a $t_{\rm exp} = 10$ s exposure.

The sources are independent sources of noise and have to be added in quadrature to get the total white noise. We adopt a white noise model that includes noise from the source, sky background, readout, scintillation, and dark current. Adding these sources together, we get that the total white noise for our 10 min binned ($n_{\rm bin} = 60$ frames) data is

$$\sigma_{\rm white} = \frac{\sqrt{\sigma^2_{\rm read} + \sigma^2_{\rm star} + \sigma^2_{\rm sky} + \sigma^2_{\rm dark} + \sigma^2_{\rm scint}}}{\sqrt{n_{\rm bin}}}, \tag{3.1}$$

where the noise terms are described in more detail below.

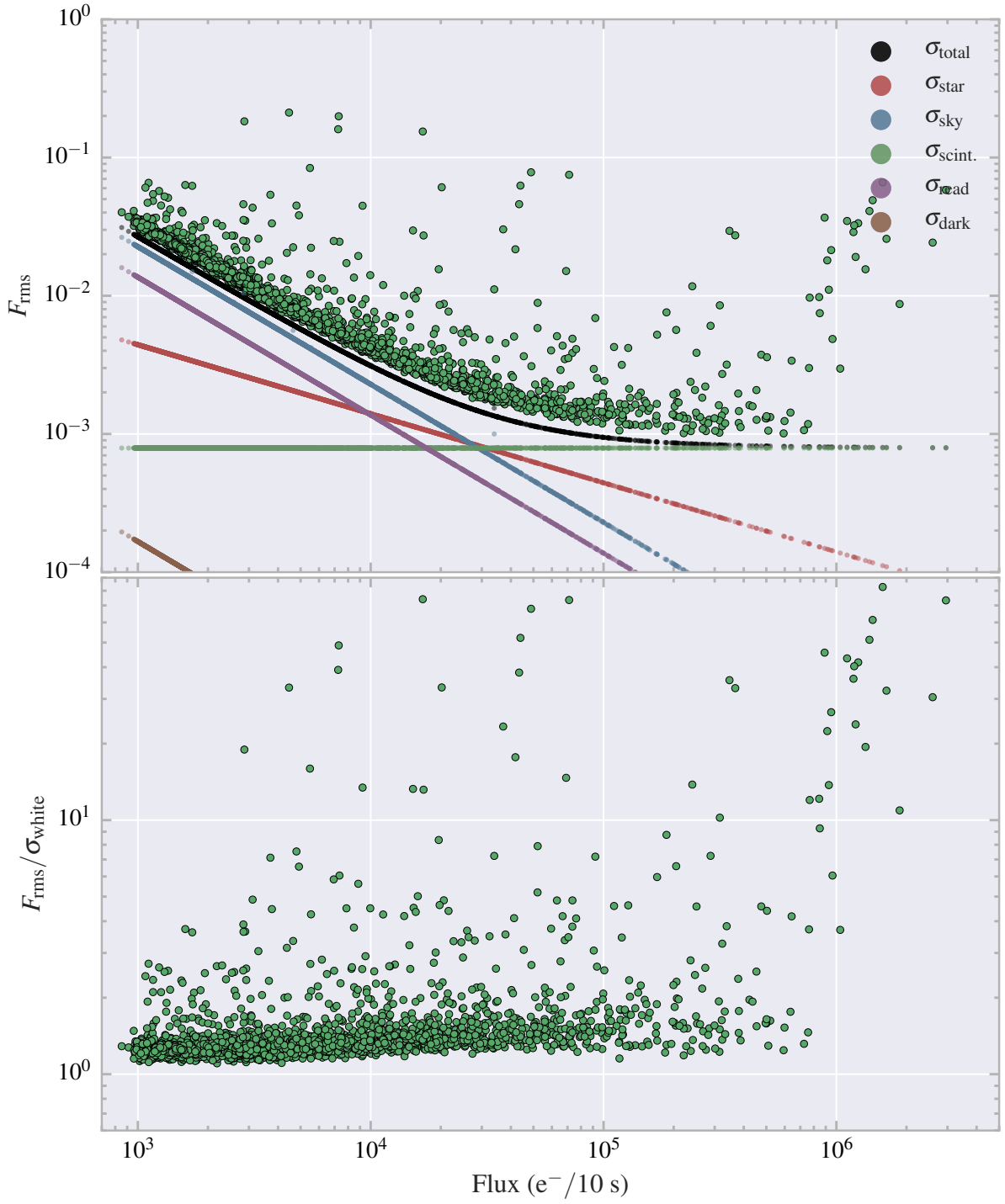


Figure 3.1: F_{rms} with white noise model for the field NG0304-1115.

Table 3.3: Camera characteristics for the NGTS CCDs, measured from lab tests at the University of Leicester.

Camera	Gain (e^-/ADU)	Readout rms (e^-/pixel)	Dark current rms (e^-/pixel/s)
801	2.05	12.9	0.012
802	2.93	12.8	0.028
803	2.79	12.7	0.028
804	1.89	13.1	0.012
805	3.04	14.9	0.040
806	2.08	14.1	0.004
807	1.94	13.0	0.005
808	2.07	13.0	0.004
809	2.05	18.4	0.005
810	2.07	13.2	0.007
811	1.99	13.8	0.005
812	2.04	14.2	0.005

Read noise

If we denote the number of readout electrons in a pixel as $R^2_{\rm read}$, and our aperture radius is 3 pixels, we have from Eq. 1.18 that the read noise in our model is

$$\sigma_{\rm read} = \sqrt{\pi r^2 \times R^2_{\rm read}} = \sqrt{9\pi R^2_{\rm read}}. \tag{3.2}$$

The readout electron rms for each camera is shown in Table 3.3

Dark current

If we denote the dark current values in Table 3.3 as $R_{\rm dark}$ with units e^- pixel^{-1} s^{-1}, Eq.1.19 then says

$$\sigma_{\rm dark} = \sqrt{\pi r^2 \times D \times t_{\rm exp}} = \sqrt{90\pi D}. \tag{3.3}$$

Shot (source) noise

If the source provides $R_{\rm star}$ ADU s^{-1}, then we have from Eq. 1.20 that the noise from the source is

$$\sigma_{\rm star} = \sqrt{R_{\rm star} \times t_{\rm exp} \times \mathcal{G}} = \sqrt{10 R_{\rm star} \times \mathcal{G}}. \tag{3.4}$$

Scintillation noise

We include atmospheric scintillation in our white noise model, following Eq. 1.21 in Chapter 1, using $t_{\rm exp} = 10$ s. We apply a correction factor of 1.56 for the Paranal site (J. Osborn et al., 2015b).

Sky background noise

For the sky background we use the SKYBKG HDU in the FITS files. Given the rate, $R_{\rm sky}$ ADU pixel^{-1} s^{-1}, from the sky background, the noise is

$$\sigma_{\rm sky} = \sqrt{R_{\rm sky} \times \pi r^2 \times t_{\rm exp} \times \mathcal{G}} = \sqrt{90\pi \times R_{\rm sky} \times \mathcal{G}}. \tag{3.5}$$

Red noise

From Fig. 3.1 we see that our white noise model can explain the noise for faint to intermediate stars, but can not explain the noise levels for the brighter stars. This is due to systematic noise that is unknown in origin, but can be due to e.g. instrumental effects or errors in the data processing in the pipeline. The systematic (red) noise does not bin down like statistical noise does, and the red noise can vary when looking at different timescales. We update Eq. 3.1 and have that our total noise model is

$$\sigma_{\text{total}} = \sqrt{\sigma^2_{\text{white}} + \sigma^2_{\text{red}}}. \tag{3.6}$$

Using F_{rms} diagrams we now describe how the red noise can be estimated.

We can look at the ratio between the fractional rms and the white noise model, which we denote R, so $R \equiv F_{\text{rms}}/\sigma_{\text{total}}$. This is a measure of the rms of each star independent of brightness. The lower panel of Fig. 3.1 shows this ratio. From this, we look at two populations of stars: the faint population consists of stars that are virtually unaffected by red noise, while the intermediate to bright population are red noise dominated. We define the faint population as stars with a flux level, $F \leq 10^4\,\text{e}^-/10\,\text{s}$, and the bright stars with flux level $10^4 < F \leq 10^5\,\text{e}^-/10\,\text{s}$, as shown in Fig. 3.2. The overall distribution in R is expected

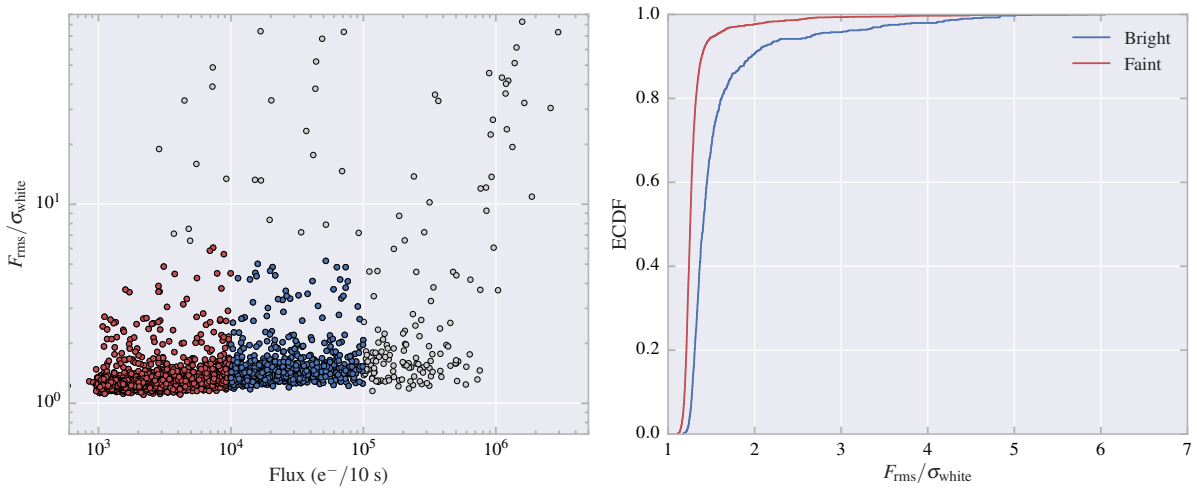


Figure 3.2: Left: Comparing R for two populations of stars. The faint and bright population of stars are dominated by different sources of noise and are used to estimate the red noise. Right: The empirical cumulative distribution function (ECDF) for the two populations.

to have a Gaussian component for the non-variable stars, with a separate variable star population that adds to the high-end Gaussian tail. To get to the underlying Gaussian population we therefore iteratively remove stars with high values of R until the change in the standard deviation is sufficiently small. We choose our cut-off at 5%, as shown in Fig. 3.3. To estimate the red noise floor, we add a base red noise level of 0.1 mmag, iteratively increment this value by 0.01 mmag and calculate the Kolmogorov-Smirnov (KS) test statistic comparing the two populations. The KS test statistic calculates the maximum distance between the two empirical cumulative distribution functions (ECDF) at a given R. When the test statistic reaches a minimum, we stop the iteration, and choose that value as our red noise level. The fractional rms plot with the red noise model included is shown in Fig. 3.4, and can now fit the noise at the bright end. We take note of the red noise level showing some brigthness dependence, as the noise on the bright end is overestimated.

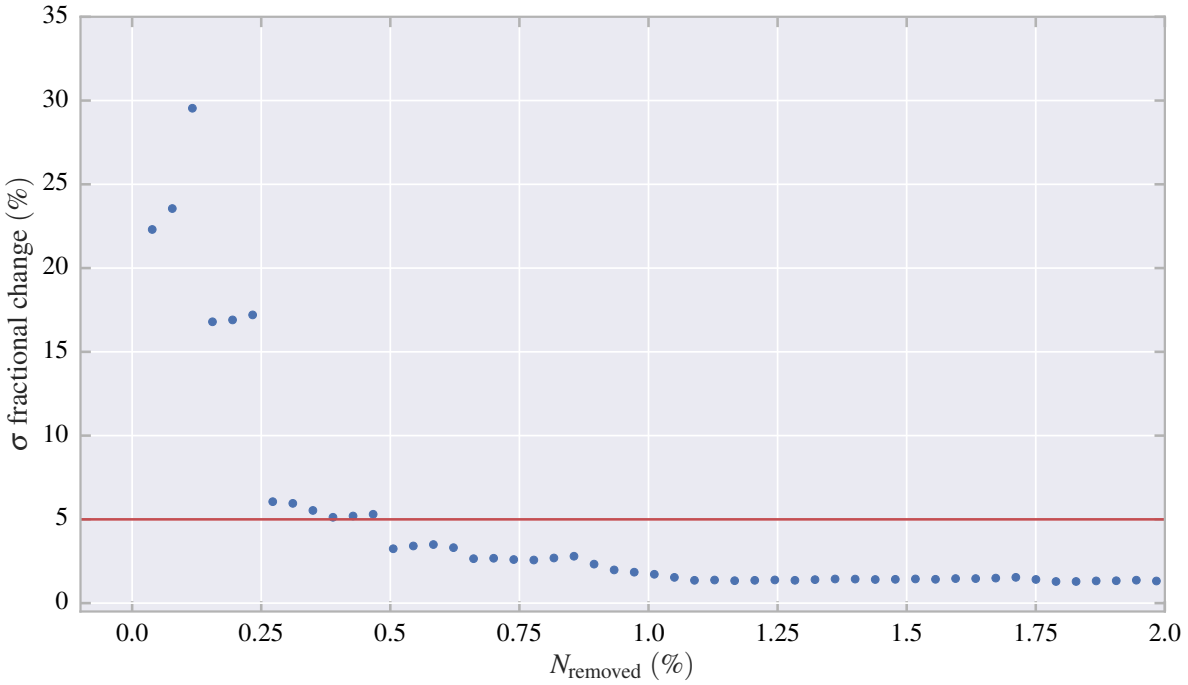


Figure 3.3: The standard deviation is found by iteratively removing the stars with highest R until the change in standard deviation is less than 5% (red horizontal line).

3.4 Identifying variable stars

Given an accurate noise model that properly accounts for both white and red noise, we can now in a global, statistical way identify stars that are of variable nature from the fractional rms plots. We show the probability density distribution of the ratio R in Fig. 3.5. We have a population that looks Gaussian, and the next step is to choose an appropriate cut-off for R, where we say that everything to the right of this value are variable stars. We assume that the distribution is Gaussian, and can choose to discard e.g. the 97th percentile and regard the remaining stars as variable. We look up the corresponding z-score in a normal distribution table and make our cut based on this. By this method can choose how strict we want our selection to be, and following our example of discarding the 97th percentile, we ensure that only approximately 3% of the non-variable stars are included in our variable population. Fig. 3.5 shows the probability distribution function (PDF) of R and indicates where we select our cut, and Fig. 3.6 shows our identified variable stars.

3.4.1 Lomb-Scargle periodogram

To find the period of the variable stars we use an implementation of the Lomb-Scargle (LS) periodogram with floating mean (M. Zechmeister and M. Kürster, 2009; N. R. Lomb, 1976; J. D. Scargle, 1982). The LS method is a least squares fit of sinusoids to the data, and is widely used for finding periodicity in irregularly sampled data, such as nightly observations with NGTS. While the LS periodogram is not the only choice of algorithm, it is known for ease of implementation and use. Other examples of algorithms are the autocorrelation function (ACF), that has been shown to be quite robust to measure rotation periods of stars for *Kepler* M-dwarfs (A. McQuillan, S. Aigrain, and T. Mazeh, 2013) with continuous data, and conditional entropy on sparse data (M. J. Graham et al., 2013).

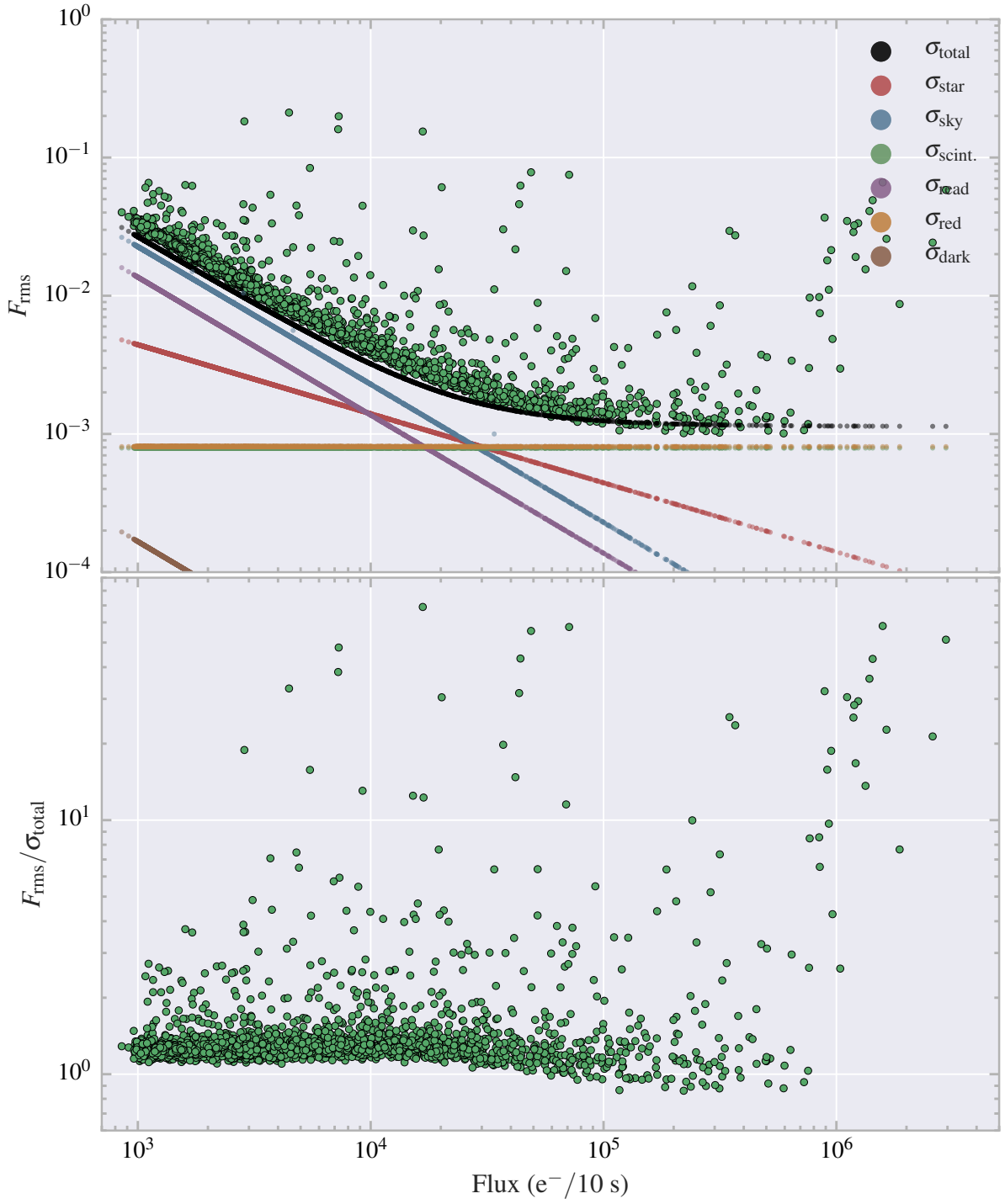


Figure 3.4: $F_{\rm rms}$ with R for the full noise model for the field NG0304−1115.

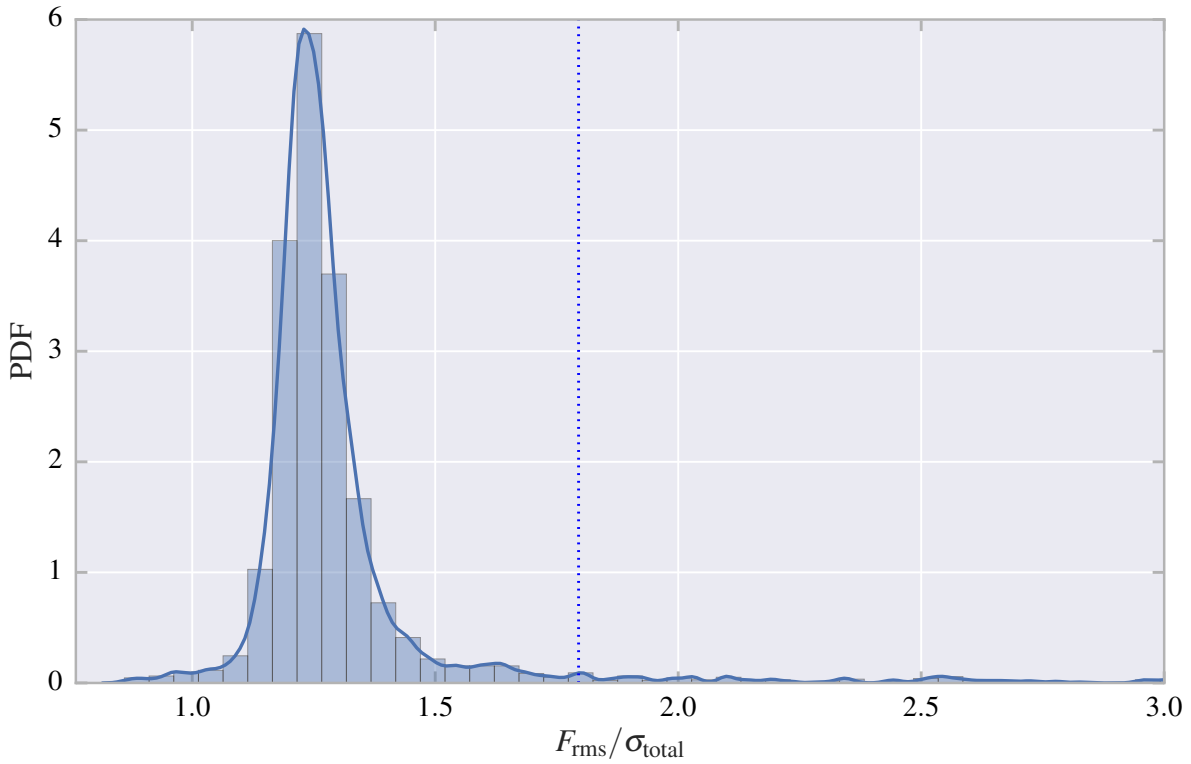


Figure 3.5: The probability density distribution of R for NG0304-1115, showing where we selected the variable threshold for this field.

We use a Python-based implementation[3] of the faster $O(N \log N)$ periodogram developed by W. H. Press and G. B. Rybicki (1989). We use an oversampling factor of $k = 20$ for the frequency spacing $\delta f = 1/(k\,T)$, where T is the total time period over which the observations were taken. Periods are searched in the range 1 hour up to half the time of the observing span, placing a limit at around 120 days for our longest observed fields in TEST16A. An example of the LS periodogram is shown in Fig. 3.7.

Mitigating systematic periods

The data has strong 1-day periodic signals, as well as its $^1/_5$, $^1/_4$, $^1/_3$, $^1/_2$ and $^2/_3$ aliases, as shown in Fig. 3.7. To avoid picking these periods, we remove the 1-day periodic signal and its harmonics by fitting the best-fitting sinusoid of the form

$$s = a \sin(2\pi f t) + b \cos(2\pi f t) + c,$$

where a, b, c are free parameters, $f = 1/P$ is the frequency removed, and t and s are time and flux. We extract the remaining dominant period following the systematic period removal. We proceed to remove the dominant period and find the next, selecting a total of three periods.

In many stars we also see power at period of the lunar cycle (28 d), and its half period. This is due to the sky background level likely not being treated in a sufficient manner, which will be discussed more in Section 5.1.1. We have not attempted to remove this signal since these targets show variability due to

[3] http://www.astroml.org/gatspy/

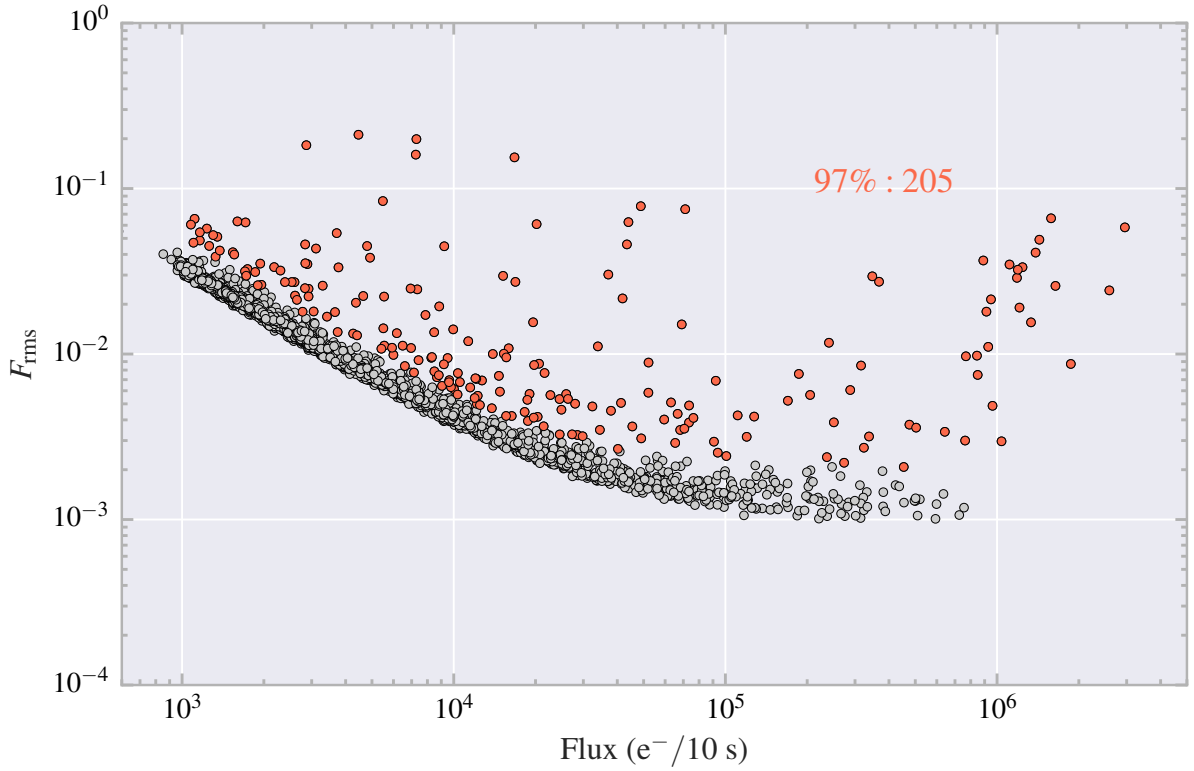


Figure 3.6: Field NG0304−1115, where the identified variable lightcurves are coloured in red, counting 205.

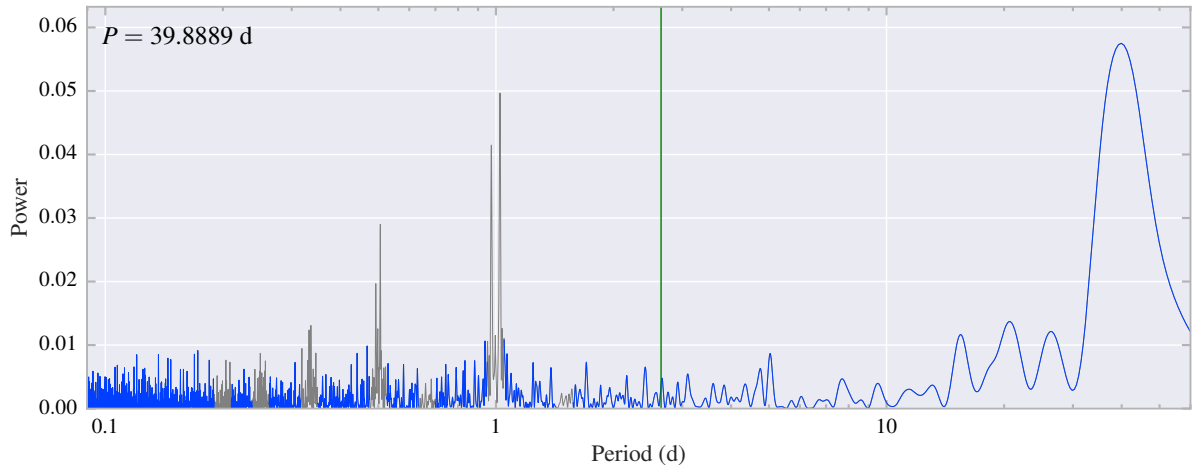


Figure 3.7: The Lomb-Scargle periodogram for an NGTS transiting planet candidate. The gray areas are the 1-day false period and its and aliases ±5% that show strong power in our lightcurves. The green line denotes the 2.6 day period found for the planet candidate. Here the M-dwarf in question shows a dominant periodic signal at ~40 days which could be due to rotation, or related to other systematic periods.

erroneous background subtraction, and it is not certain that they have other astrophysical variability, in which case they would just contribute to the population of noisy lightcurves.

Eclipsing binary stars

A significant fraction of the identified variable stars will be various types of eclipsing binary stars. The LS periodogram will often show the highest power at half the period. To tackle this we perform an EB test on each lightcurve. The lightcurve is binned to 10 min bins and phase folded on the dominant period. The phase folded lightcurve is further binned in 48 points between phases $\phi = [0, 1]$, and the minimum value is shifted to $\phi = 0$. We perform a test on each light curve by phase folding at double of the dominant period. If the object is an eclipsing binary, there should be an eclipse at $\phi = 0$ as well as $\phi = 0.5$. We perform two tests to check for eclipsing binary stars. We first check if the absolute value of the difference in the flux at these two points is greater than three times the standard deviation of the binned minimum flux between phases $\phi = 0.45$ and $\phi = 0.55$. Second, we check that the doubled period is less than $P = 20$ days. The second test is implemented because many long-period variables are identified as variable stars, where the difference in the flux between each period can be greater than the threshold from the first test due to amplitude modulation. Most eclipsing binary stars will not have periods above $P = 20$ days. If both these criteria are met, the dominant period is doubled. This is only done for the first extracted period. From the binned phase folded light curve we further calculate the semi-amplitude of the variation, $A = (f_{\mathrm{max}} - f_{\mathrm{min}})/2$. The maximum and minimum flux values, f_{min} and f_{max}, are chosen by taking the median of the three maximum and three minimum flux values to avoid picking outliers to estimate the period. Similar adjustments have been done in D. J. Armstrong et al. (2016).

To test the efficacy of our period-search algorithm we compare our periods with known variable stars in our fields from our cross-matches with the AAVSO, GCVS and ACVS catalogues, see Table 3.2. For the AAVSO and ACVS catalogues, their entries are mostly dominated by eclipsing binaries (EA, EB, EW), RR Lyraes (RRAB, RRC), DSCT, SR and Mira variables, most of which have been described in Section 1.4. The GCVS matches are mostly various types of eruptive variable stars, in particular Orion variables (IN, INA, INB). In Fig. 3.8, left plot, we show our periods compared to those in the catalogues. Here we have excluded our targets from fields with sparse data and red noise levels higher than 1.2 mmag (see Chapter 4). After the filtering we have 293 matching variable stars that we compare against. We find a matching rate of 40.0 per-cent on the same period, 71.7 per-cent when including double and half periods, and 77.5 per-cent when including all three periods that are extracted. The vertical dotted lines indicate the positions of the 0.33, 0.5, 0.66, 1 and 2 day periods, which show strong periodic signals in our lightcurves. Most of our mis-matched objects are found to lie at these periods. For the stars that have these systematic periods, we find that about half (~15) have catalogue periods higher than 100 days, belonging to Mira or semiregular variables. Our periodogram searches for periods up to half the observing span for a given field, which sets a maximum limit around 120 days, but most fields in our sample will have their maximum period lower than this. For these stars we find that our lightcurves phase folded on the catalogue period show no significant variability. Because of this the highest periodic signal is found to be at one of our systematic periods.

For the remaining lightcurves (~15) with catalogue periods less than 100 days, we looked at the phase folded lightcurves to investigate the reasons for finding systematic periods. In about half of the cases the lightcurves folded on the catalogue period shows a variable lightcurve. In removing the systematic periods with a sine fit, we only remove the exact period and its aliases. This removes most of the power around these periods, but there is still some power in the wings, which can be higher than other periodic signals (see Fig. 3.7). It can also be related to the sidereal day, which NGTS team members have shown to produce strong peaks in our data. For the remainder of the lightcurves, there are varying reasons for

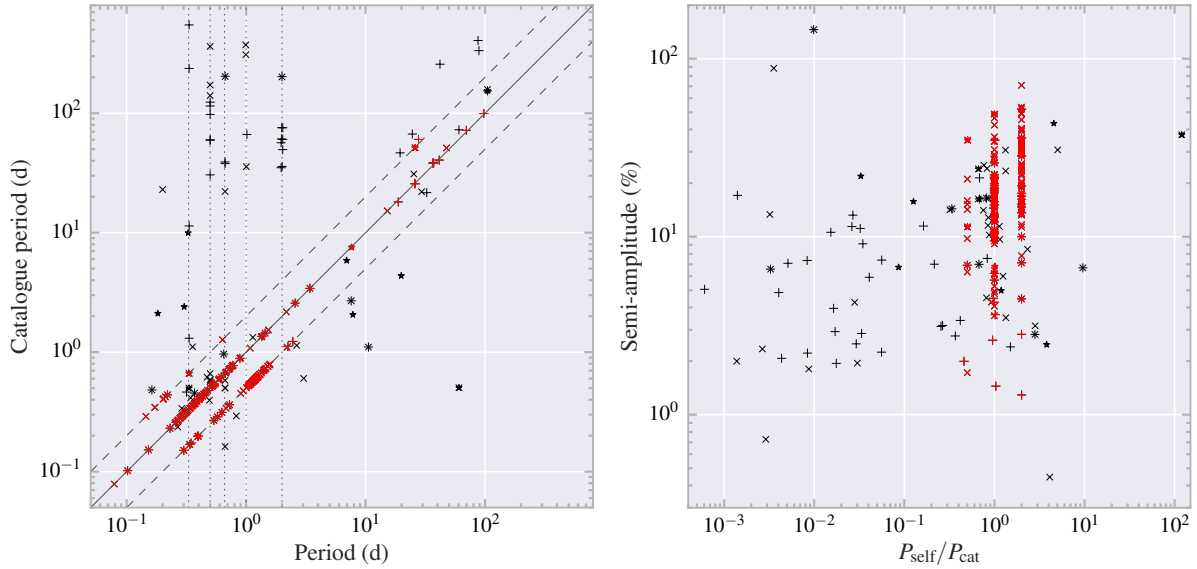


Figure 3.8: The performance of our Lomb-Scargle periodogram when comparing to catalogues with known variable stars in our field. In the left figure the x, plus and star markers are from AAVSO, ACVS and GCVS, respectively. The diagonal solid gray line indicates perfect agreement with the catalogue period, dashed lines are $^1/_2$ and double periods. We achieve an acceptance rate of 40.0 per-cent with the same period, 71.7 per-cent when including double and half periods, and 77.5 per-cent when including the second and third extracted periods. The vertical dotted lines mark the position of 0.33, 0.5, 0.66, 1 and 2 day periods, which are the strongest systematic periods. A large number of our mis-matched periods are due to picking up these systematics.

selecting the wrong periods. In some cases we find no indication of variability at the catalogue period. In others the lunar period was selected because it produced a higher peak, while in others the variable star is of a type that does not produce sinusoid-like signals, for which Lomb-Scargle finds higher power at our systematic periods which are more sinusoidal.

We further investigate the reasons for other non-matching periods that are not due to systematics. In the right plot of Fig. 3.8 we compare the ratios of our dominant period and the the from the catalogue against the amplitude we find. For the mis-matched targets in-between the 1:1, 1:2, and 2:1 period ratios, we find that for a number of them our period is a better estimate after inspecting some lightcurves. This may be due to period-modulation is some of these stars, or we pick up a lower amplitude signal since the catalogue observations do not have as good a precision as NGTS. Period changes in eclipsing binaries often indicates the presence of a third body in the system. In Fig. 3.9 we shows an example of a catalogued δ Scuti variable folded on its period. We show that our period is more likely, and that the variability is more reminiscent of a close-in binary star with strong ellipsoidal variation and some beaming. In Fig. 3.10 we show two more examples of detached eclipsing binaries that we have found more precise periods for. These cases are typical of variables from the AAVSO and ACVS catalogues. Most of our mis-matched periods from GCVS are due to the variables being eruptive in nature, thus being difficult to identify any periodicity. The same is true for semiregular variables from ACVS and AAVSO for which we have non-matching periods.

We also note that our period-check intended to double the periods for eclipsing binaries has the unwanted effect of doubling the period a large fraction of RR Lyrae variables. This can be mitigated in the future by e.g. also checking for symmetry in the phase region $\phi = [0.45, 0.55]$.

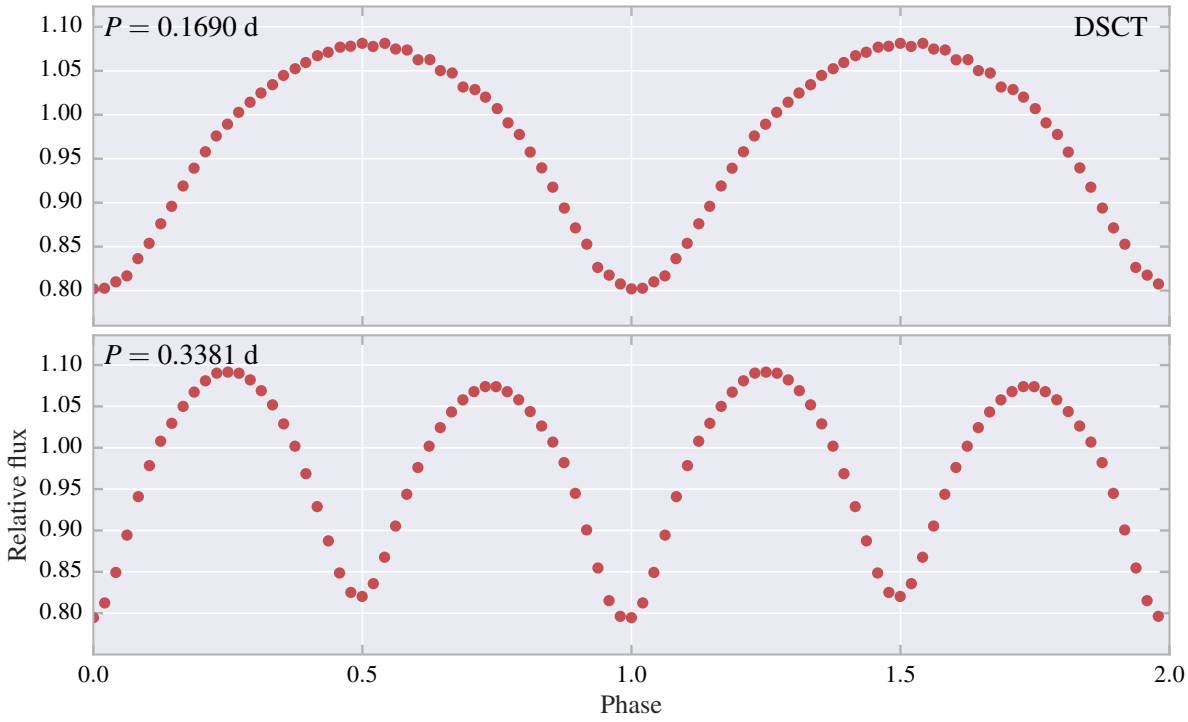


Figure 3.9: Upper panel: A catalogued DSCT from AAVSO phase folded on the catalogue period. Lower panel: Same object phase folded on our period, showing that it is likely an EW due to the different depths between primary and secondary eclipse. The different levels of out-of-eclipse flux also suggests a beaming effect.

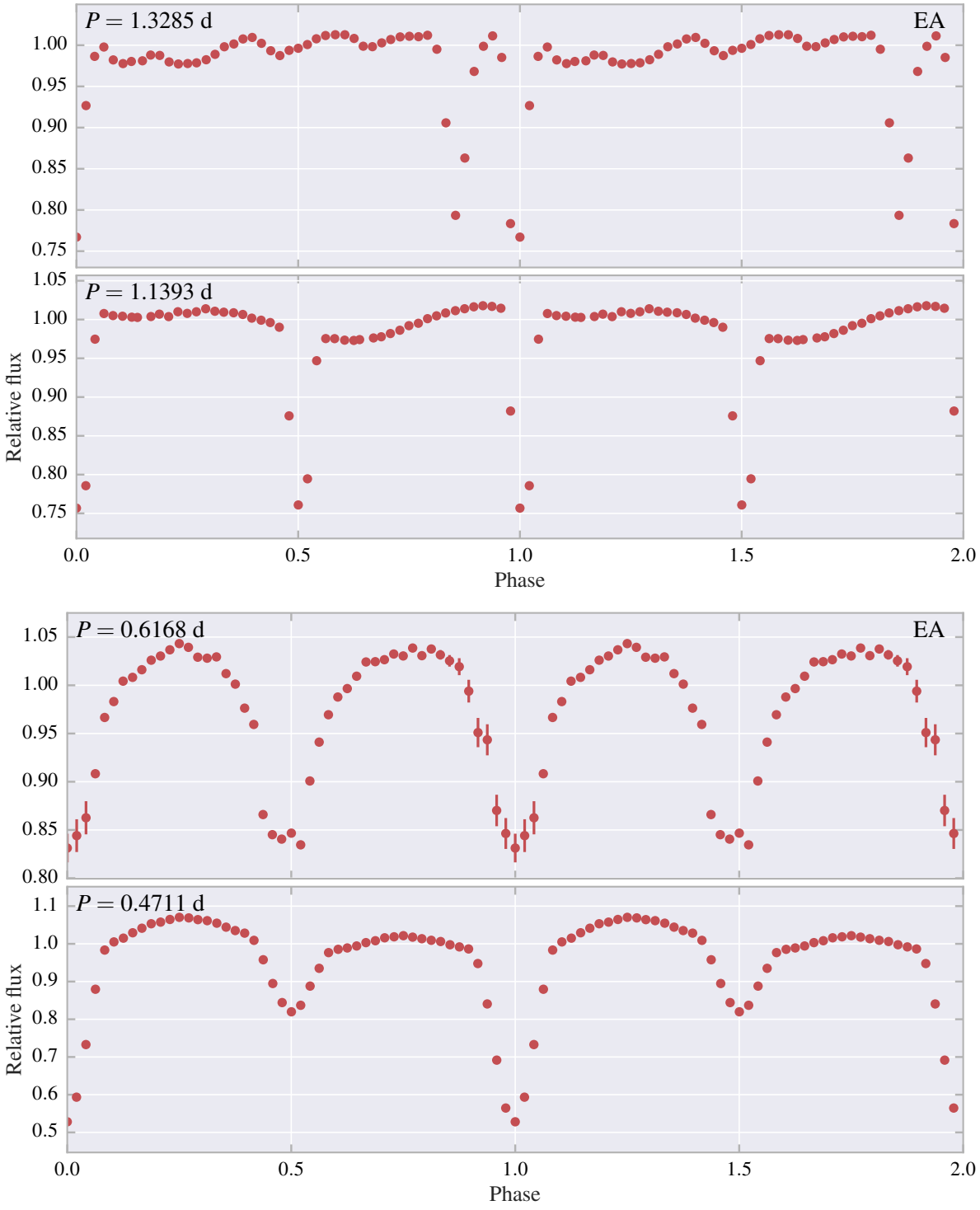


Figure 3.10: Two EA systems compared by their catalogue periods vs. our period. In both cases the variability type is correct from the catalogue, but the periods are wrong, likely due to period modulation since the observations were taken for the catalogue. The first system (upper two panels) shows beaming effects, while the second system (lower two panels) shows some ellipsoidal variation as well as beaming.

CHAPTER 4

Results

In this Chapter we report the results from our analysis, as described in Chapter 3. We ingested the full TEST16A dataset (Table 3.1) into our data quality pipeline, which generated empirical and expected noise levels for each star, flagged variable stars, and calculated their period and amplitude from our period-search algorithm. In Appendix A we provide a summary figure per field of the various output from our codes, which will be referred to throughout this Chapter.

In Section 4.1 we report our findings on the limiting noise levels and data quality, while in Section 4.2 we report the numbers on the identified variable stars, their parameters, and show some example lightcurves.

4.1 Noise

4.1.1 Fractional rms plots and noise models

Overall, our noise model performs well in explaining the flux-rms curve for the data in most of our fields. In some cases the model underestimates the noise throughout, leaving a small gap between the total noise curve and the F_{rms} population, as shown in Fig. 4.1. This could be due to various reasons. One possibility is the sky background being underestimated, as will be discussed in Section 5.1.1. It could also be due to the read noise values being slightly wrong. The readout rms values in Table 3.3 are taken from lab tests done by Andor, and could possibly be different by some non-negligible amount compared to what NGTS team members at Leicester University found in their camera testing. However, the latter results proved to be difficult to find, and so the Andor values were used. A third reason for the model discrepancy could be due to something that was discovered at the time of writing. For the TEST16A dataset, the gain setting for cameras 808, 809, 810, 811, and 813, changed from PAG2[1] to PAG1, but this change was not reflected in the FITS data files, so some of the gain values that were used could be for the wrong CCD gain setting. Some of these cameras were reverted back to PAG2 at a later stage. Field NG1318−4500 is an example where the camera gain setting was changed mid-season that the code did not account for, which produces the erratic rms curve that is shown in Fig. 4.2. However, the difference between the PAG1 and PAG2 gain values is a factor of ∼2, which would produce a larger model discrepancy than we observe. Therefore it is unlikely that this is a problem for all other fields except NG1318−4500.

[1] The PAG setting is used to determine the conversion factor between measured electrons and ADU.

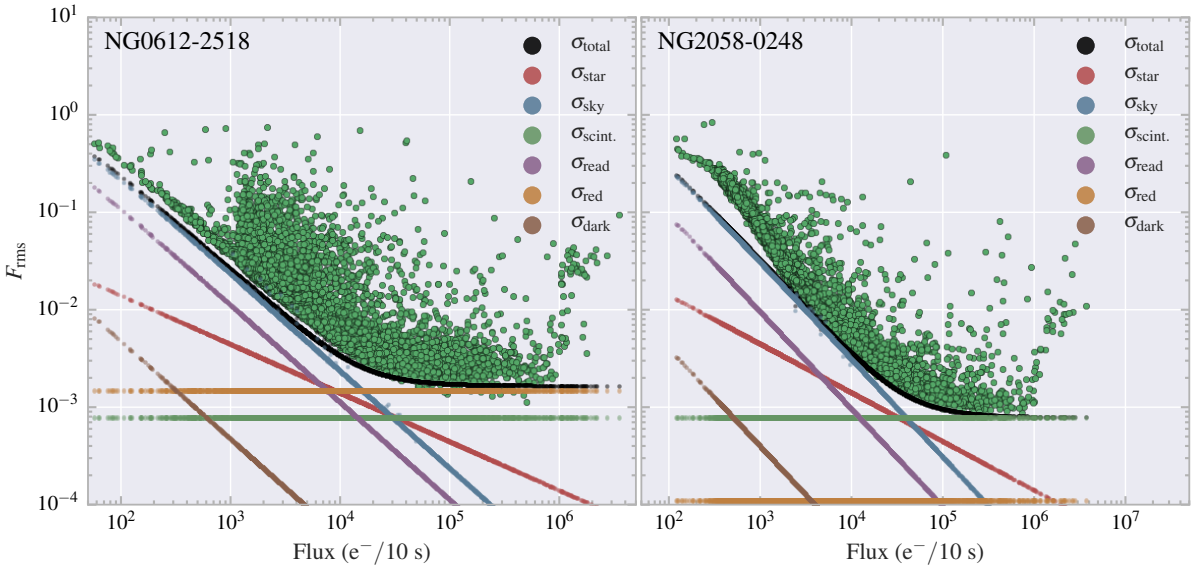


Figure 4.1: F_{rms} curves for NG0612−2518 (left) and NG2058−0248 (right).

Extra faint end noise

In some fields we notice a sharp noise increase for the faint end that our model cannot reproduce, as seen to the right in Fig. 4.1 for field NG2058−0248. These effects are present in fields with less than a month of observations. It could be due to the increased sky background from the full moon, as our model takes the mean sky background as an estimate. For observations less than a cycle of the moon the mean will underestimate the increased background level during full moon, but will average out as the field is observed for longer, as seen in other fields with several months of observations.

Diluted stars

For fields with more than ∼15 000 stars having their flux extracted by aperture photometry we should start to see effects of dilution, where flux from a nearby star is entering the aperture of another star. This introduces extra noise that is not taken into account by our model, possibly producing the larger spread that is seen in some crowded fields. Due to this we expect to pick up a larger fraction of "false" variables, i.e. noisy stars, in our variable star filtering. A solution to this would be to filter away stars that are diluted. At the time of writing, dilution parameters per star were not implemented properly in the data files, and so was not taken into account.

4.1.2 Scaling with bin width

We have also looked at how the noise per field scales with the bin width of the data. For this we selected a sample of 10 stars from three regions of brightness corresponding to V∼14, V∼12, and V∼10[2], for a total of 30 stars. The stars in each bin were chosen according to the lowest R value in those bins. We computed their F_{rms} for bin widths up to t_{bin}∼3.3 h, and choose the median F_{rms} value of the 10 stars in each magnitude bin as the noise level at each iteration. We compare it to our white noise model at each

[2] Formally, this corresponds to flux values $3\times10^3 < f < 4\times10^3$, $2\times10^4 < f < 3\times10^4$, and $1\times10^5 < f < 2\times10^5$, respectively

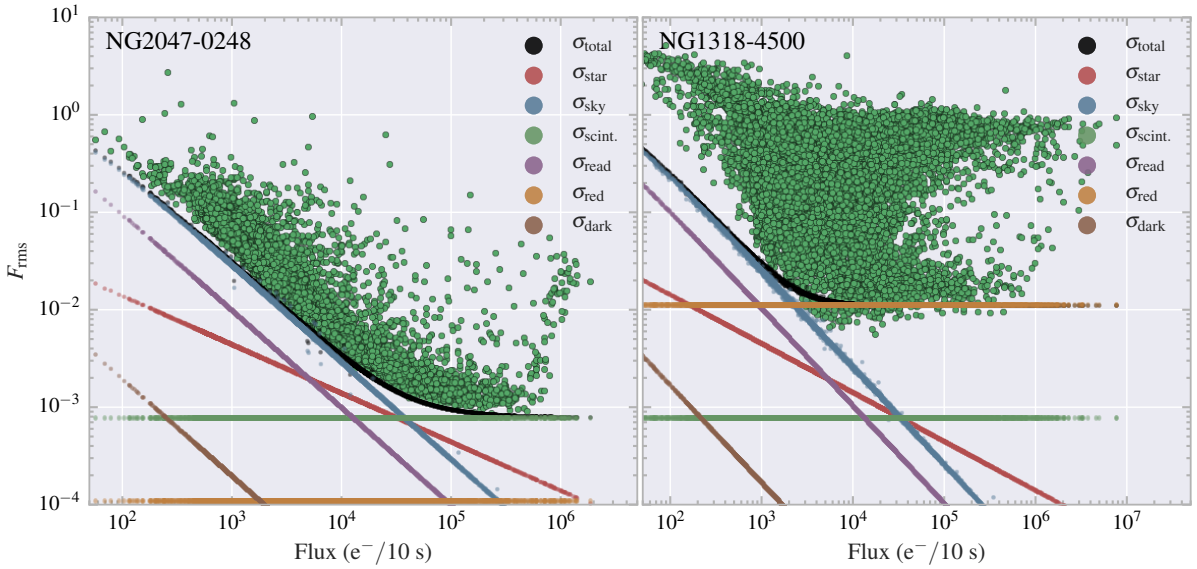


Figure 4.2: F_{rms} curves for NG2047−0248 (left) and NG1318−4500 (right), both with ~20 000 apertures each. The two fields have red noise levels of 0.11 and 11.2 mmag, respectively. For NG1318−-4500, the camera PAG setting was changed mid season.

bin, which follows $1/\sqrt{n}$ behaviour as shown in Fig. 4.3 for three magnitude bins. The circle points are the calculated F_{rms} at each win width, and the dotted lines are the calculated white noise from our model, in each magnitude bin. Our noise bins down with increased bin width, but not as one would expect for complete white noise behaviour. This can point to the nights binning down differently, possibly due to varying flux levels around full moon.

4.1.3 Red noise levels

The red noise levels found for each field by our algorithm is tabulated in Table 4.1. For the TEST16A data we detect red noise levels in the range 0.1–1.5 mmag, with an additional outlier at 11.2 mmag for NG1318−4500. Seventy per-cent of our fields reach sub-mmag red noise levels on the data with t_{bin} = 10 min. We reach a median red noise level of 0.82±0.31 for our fields. There does not seem to be any cameras that are consistently performing poorly. Fields NG2047−0248 and NG2058−0248 are stated to reach a 0.11 mmag red noise levels. In reality the red noise is zero (< 0.1) for these fields. In the red noise estimation, the initial red noise value is set to 0.1, and incremented by 0.05 mmag until a minimum of the Kolmogorov-Smirnov test statistic is reached, which happens already at the first iteration. It is also clear from Figs. A.22 and A.23 that the white noise model provides a good fit to the data. Both of these fields have less than 20 days of data, which demonstrates the full potential of NGTS when not limited by long-term effects that add systematic noise to the photometry.

We have also processed TEST18 data to compare with TEST16A levels. TEST18 data contains 6 months of more data for our TEST16A fields with short observating duration. find that the red noise levels are in the range 0.48–1.28 mmag, with a median 0.82±0.10 across fields. This level is consistent with the TEST16A dataset, and the smaller error indicates that the red noise level is more consistent for each field because the number of observations in each field are comparable. For the TEST18 dataset we were unable to process data from some fields due to software issues with the data extraction that could not be resolved in time. The fields in question are represented by havng missing values in Table 4.1.

Table 4.1: Red noise levels and variable numbers per field from the TEST16A and TEST18 datasets. TEST18 data contains 6 months of more data for our short fields. At the bottom of the table, the uncertainty on the mean is the standard deviation, while the uncertainty on the median is the MAD error.
†Fields tagged as "bad", due to high (>1.2 mmag) red noise levels in TEST16A, or sparse data without enough observations (downtime). NG1349−1115 is tagged as a bad field because a significant fraction of the identified variables are due to a saturated star contaminating a the whole pixel column (Fig. 4.7, middle panel). Its red noise is still included in the calculations.
*Field suffers from processing error and is not included in any final numbers at the bottom of the table.

Field	Camera	TEST16A				TEST18			
		$\sigma_{\rm red}$ (mmag)	$N_{\rm var}$	$N_{\rm var}^{\rm real}$	$f_{\rm var}^{\rm real}$ (%)	$\sigma_{\rm red}$ (mmag)	$N_{\rm var}$	$N_{\rm var}^{\rm real}$	$f_{\rm var}^{\rm real}$ (%)
NG0304 −1115	809	0.81	205	85	2.43	0.77	208	72	2.06
NG0313 −2230	810	0.95	176	54	1.47	0.92	202	49	1.33
NG0348 −3345†	811	1.28	169	56	1.59	1.08	179	66	1.88
NG0409 −1941	812	0.81	203	74	1.56	0.65	231	84	1.77
NG0522 −2518	802	0.84	293	99	1.19	0.75	301	90	1.08
NG0531 −0826†	806	1.2	466	200	1.82	–	–	–	
NG0612 −2518†	805	1.46	515	195	1.45	–	–	–	
NG0618 −6441†	801	1.25	290	136	1.22	1.28	273	121	1.08
NG1135 −2518	809	0.82	254	107	1.59	–	–	–	
NG1213 −3633	810	1.22	354	78	0.79	–	–	–	
NG1253 −1941	803	0.49	237	120	1.84	0.48	227	97	1.49
NG1315 −2807	812	1.06	348	88	1.11	1.04	303	75	0.95
NG1318 −4500†,*	811	11.2	866	4	0.02	8.0	820	24	0.11
NG1340 −3345	808	0.72	395	111	0.97	0.79	290	84	0.74
NG1349 −1115†	804	1.08	234	62	1.23	0.9	164	69	1.37
NG1416 −2518	806	0.71	342	118	1.00	0.71	391	142	1.21
NG1421 +0000	805	0.76	156	55	1.41	0.84	128	–	–
NG1428 −2518	802	0.92	315	148	1.63	1.07	302	99	1.09
NG1444 +0537	801	0.99	215	83	2.35	0.94	206	93	2.64
NG1947 −4200†	806	1.26	505	106	0.74	–	–	–	
NG2028 −2518	802	0.75	280	103	0.76	0.61	337	93	0.68
NG2047 −0248†	810	0.11	577	135	0.69	0.9	581	109	0.56
NG2058 −0248†	803	0.11	492	95	0.52	0.77	375	126	0.69
NG2126 −1652†	804	0.87	203	52	0.83	0.96	214	82	1.32
NG2132 +0248†	805	0.61	294	64	0.88	0.89	241	83	1.14
NG2142 +0826†	801	0.78	233	84	1.11	0.85	260	120	1.58
NG2145 −3345†	808	0.94	209	68	1.32	0.76	237	93	1.81
NG2150 −3922†	812	0.4	192	38	0.76	1.14	173	47	0.94
NG2152 −1403†	811	0.38	190	34	0.71	0.73	180	68	1.43
NG2346 −3633†	802	0.3	198	50	1.49	0.77	248	130	3.88
Total			7976	2698	1.12		5792	2116	1.04
Mean		0.82 ± 0.34			1.26 ± 0.47	0.85 ± 0.18			1.37 ± 0.75
Median		0.82 ± 0.31			1.22 ± 0.37	0.82 ± 0.10			1.26 ± 0.32
Only good fields									
Total			3174	1245	1.31				
Mean					1.49 ± 0.49				
Median					1.47 ± 0.36				

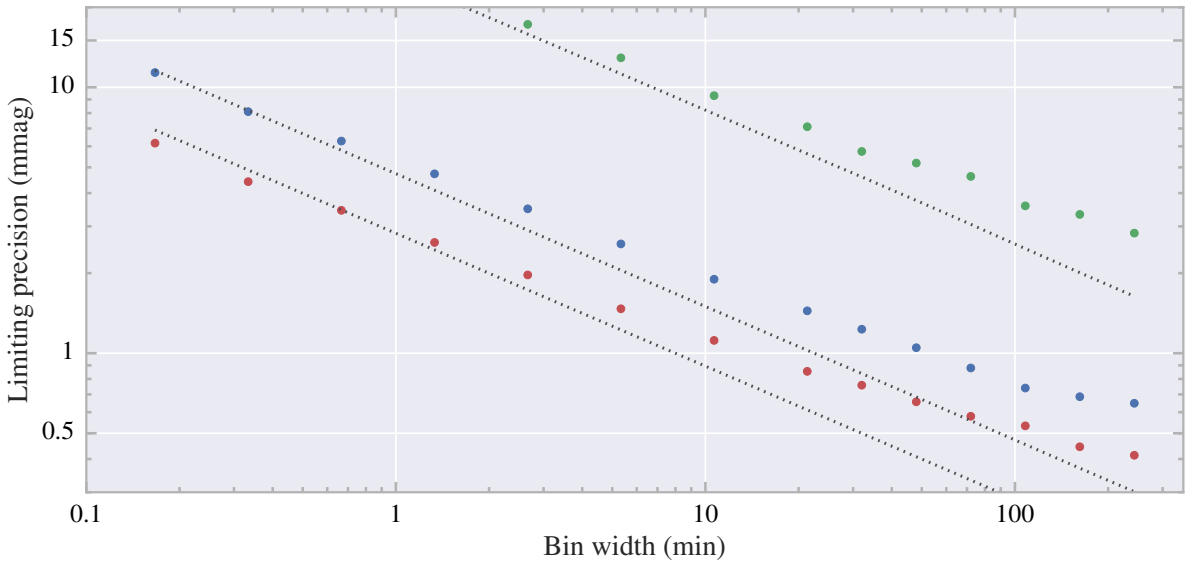


Figure 4.3: Noise scaling with bin width for a sample of non-variable stars of mag $V{\sim}10$ (red), $V{\sim}12$ (blue), and $V{\sim}14$ (green). The triangles shows the white noise level from our model.

4.2 Variable stars

4.2.1 Detection

Across our 30 fields our algorithm detects 7 976 variable lightcurves for our threshold of 97%. A large fraction of these, around ⅔, are found to be false detections due to systematic noise. These often have a dominant period at one of the systematic periods related to 1-day effects. We perform a vetting of the fields to estimate the true fraction of astrophysical variable stars. We exclude fields with sparse observations, as stars from these fields often present with systematic periods. We also exclude fields with a red noise level higher than 1.2 mmag. Finally, we also exclude NG1349−1115, as most of its identified variables are due to contamination from a saturated star (Fig. 4.7). After making our cut we further mask objects with a dominant period at one of the systematic periods. We are then left with 1245 variable stars at high confidence. The fraction of true variable stars is then found to be 1.31%. The fraction of variables in each field is in the range 0.52–2.43 %, with median 1.47±0.36. The numbers are summarised in Table 4.1.

It is clear that the result of our variable star filtering heavily depends on the general scatter in the rms curves due to systematics. In figures where there is an increased scatter for the faint end stars, our algorithm will mostly pick up these lightcurves only, without ever probing the intermediate to bright targets where the noise is lower. These effects are illustrated in Fig. 4.4, where we compare two fields where the algorithm picked up mostly noise with systematic periods (left), and variables stars (right). Our red noise estimates have a tendency to raise the total noise level enough to cut through the stars at the low F_{rms} end. This leads to an overestimation of the noise at the low end compared to the rest of the rms curve, which means the algorithm does not pick up as many bright variable light curves. The reason for the overestimation is that the red noise is found from comparing faint stars with stars up until flux levels of 10^5 $\mathrm{e}^-/10$ s. This is to avoid having saturation effects from the brightest stars affect the red noise estimate.

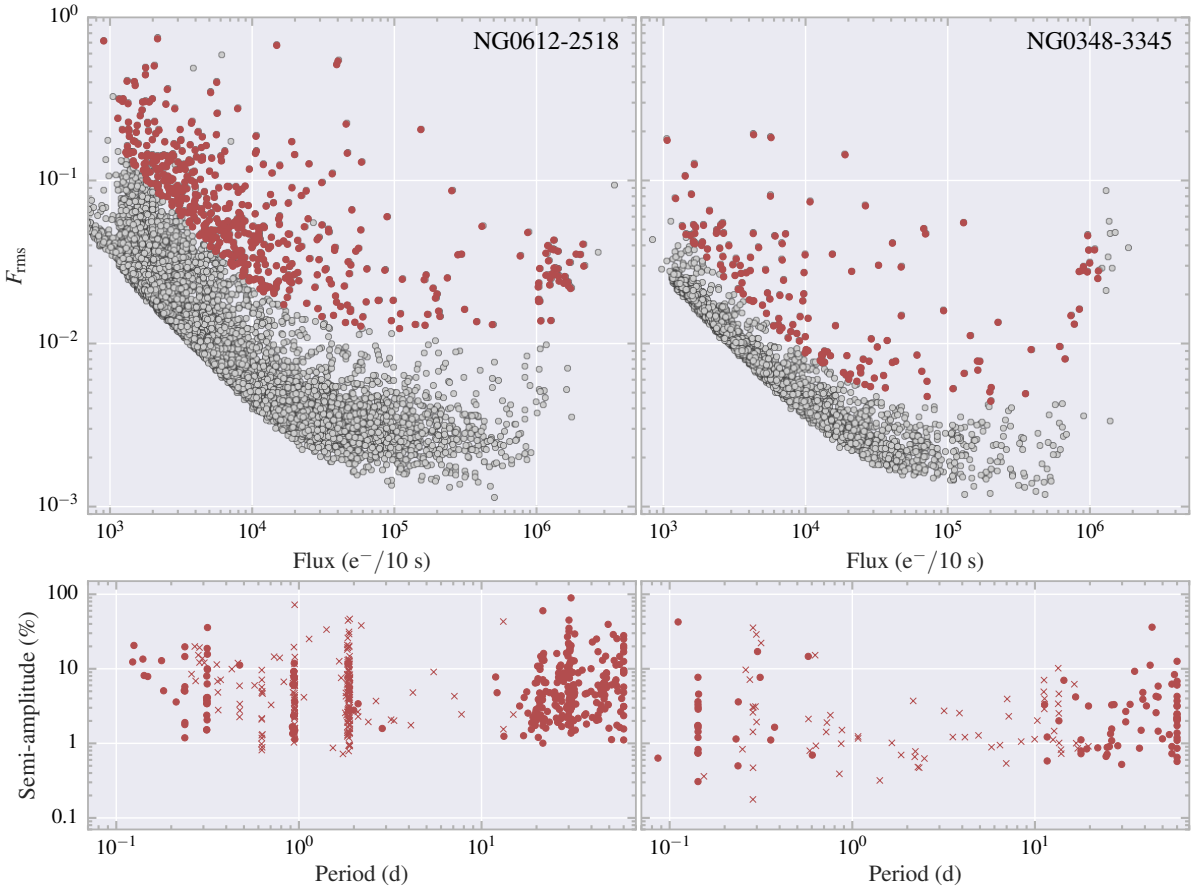


Figure 4.4: The variable star identification performance between a field where most of the identified variables have systematic periods related to 1-day effects and moon correlations (left), and a field where there are fewer false variables (right).

4.2.2 Parameters

Our code flags stars with variable lightcurves, and a period is searched with an implementation of the generalised Lomb-Scargle periodogram. Once the dominant period is identified, following some criteria described in Section 3.4.1, the lightcurve is folded on its period. The semi-amplitude of the signal is found from the phase-folded lightcurve. After taking a look at a large sample of variable lightcurves from all our fields, it was noted that some of our fields, mainly with $\alpha \geq$ 20h have poor data coverage due to issues explained in Section 2.3.2, which is also reflected in Table 3.1 in the observing span versus actual nightly observations. The poor time coverage leads our LS periodogram to pick up a large fraction of systematic periods. Short period variables $\leq$1 day, such as RR Lyrae variables, are sometimes correctly picked up even with the sparse observations, as seen in Fig. 4.8, which demonstrates the robustness of the Lomb-Scargle periodogram for sparsely sampled data. By filtering away fields with that have obvious poor time coverage with red noise levels $\sigma_{\rm red} \geq 1.2$ mmag, we improve the matching rate of our periodogram. We show all our identified variables in period-amplitude phase in Fig. 4.5 for all fields (top), and fields with a red noise level of less than 1.2 mmag and excluding sparse fields (bottom). The points are colour-coded by field, and in the upper panel we notice various streaks of systematic periods. Most of these are due to harmonics of the 1-day signal. We see strong line clustering at 2, 1, $^2/_3$, $^1/_2$, $^1/_3$,

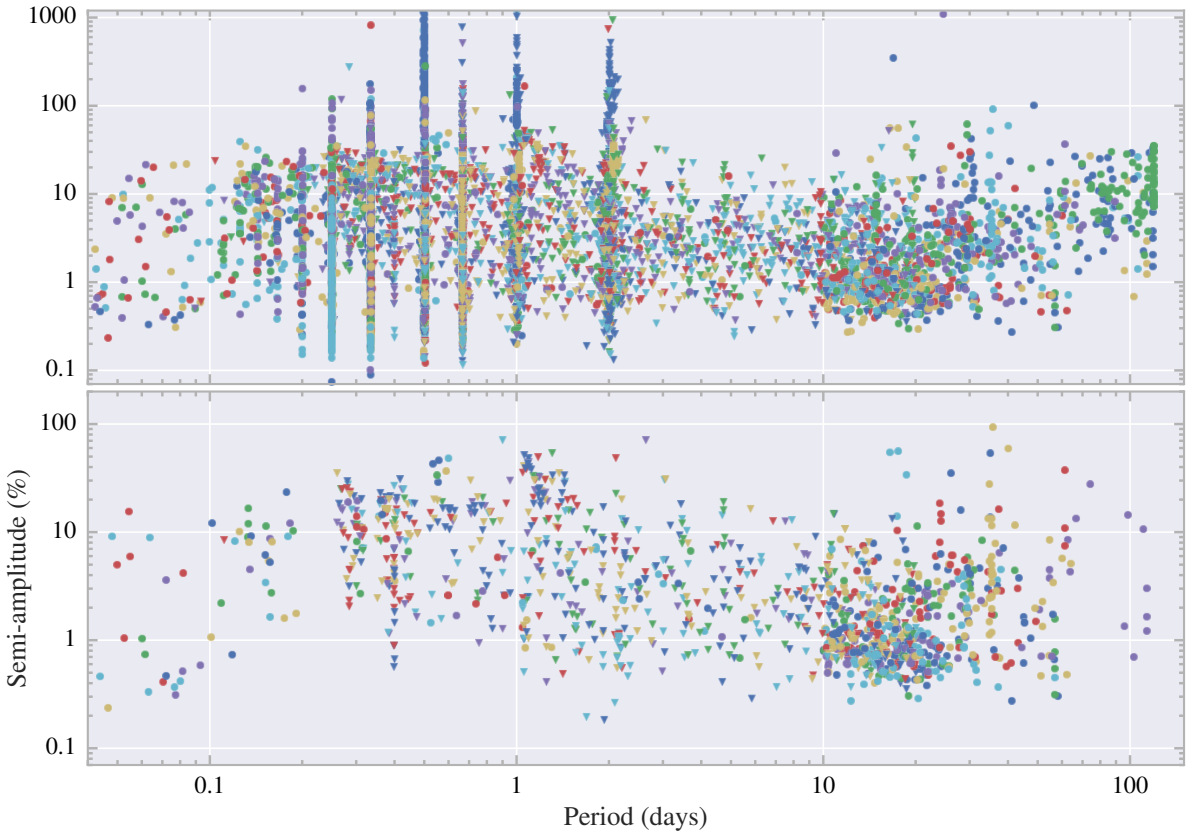


Figure 4.5: All our identified variables in period-amplitude phase coloured by field. Triangles are objects that had their dominant period doubled from the period check for eclipsing binaries. Top: All fields. Bottom: Fields with $\sigma_{\text{red}} \leq 1.2$, without sparse data, and masked common systematic periods ±4%.

1/4, 1/5, 1/6, and 1/7 days. There are also systematics specific for some fields, such as the cluster of green points at max period around 110 days.

In the lower panel we have excluded fields with sparse data and high red noise levels, and have masked systematic periods at 1 day and factors $1/n$, for integers n up to 7, as well as 2/3. After masking we notice a streak of periods just below 0.4 days, which is likely connected to our window function. We see another streak of yellow points at 35 days, which looks like a systematic period specific for that field.

We notice several populations of variable stars here. To the left, for periods less than 0.04–0.2 days, there are roughly two populations of stars. One has higher periods and amplitudes around 10%, while the other has low periods and amplitudes ranging from a few mmag to 10%. It is likely that both of these populations are mostly δ Scuti variables, as the amplitudes are consistent with this type. In the period regime 0.2–1 day we have a large population of RR Lyraes and close eclipsing binaries. The cluster of high-amplitude stars with doubled periods at ~1.2 days is likely to have a number of RR Lyrae stars that erroneously had their period doubled by our algorithm, since most of the points are triangles. The same can be true for other variables with lower amplitudes in this regime. Variable stars in the range 1–10 days are very often affected by the period doubling, most of it likely erroneously. Variables in this regime can be Cepheids or similar pulsators. For periods >30 days we expect to see semiregular variables, and some Mira variables at >90 days.

The most notable cluster in this plot is for 10–30 days at low amplitudes, most around 1%. This

shows a clear peak in the period histogram in Fig. 4.6. The main stellar targets of NGTS are G, K, and early M stars, which are expected to have rotation periods in this regime. A. McQuillan, T. Mazeh, and S. Aigrain (2014) showed for the full sample of *Kepler* main sequence stars that the expected amplitudes from brightness variations due to rotation of spot-covered stars is 0.2–30 mmag. We find that this is consistent with our data, which can imply that we are indeed picking up the upper envelope of these stars. We investigated their lightcurves in more detail to rule out systematic effects, and found that they have astrophysical variability. A large fraction of these stars do in fact show brightness variations that are consistent with spot-covered rotating stars. Some number of them do have periods at half the period of the moon phase, but no clear indication of systematic effects was present.

We show the period and amplitude distrbutions in Fig. 4.6. For the period we clearly see the peak in the 10–30 day regime. We also notice a peak around 2 days. In our period masking we mask periods ±4%, which would be too narrow for this period as it still has an overdensity. We found that for the other period, increasing past ±4% had a negative effect. The structures leftwards of 1 day are a result of our frequent masking of periods less than one day, which will produce artifical overdensities and underdensities. In the amplitude diagram, to the right, we note an apparent bimodal distribution. We do not expect any such feature for the amplitude, and find that it is not of importance. The peak around ∼1% is due to the abundance of low-amplitude signals that can be from rotating spot-covered stars.

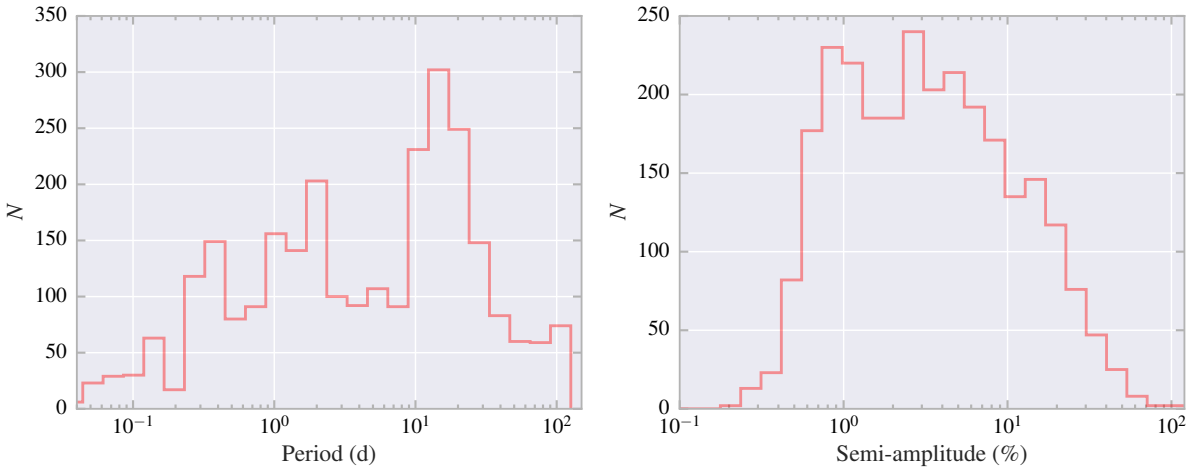


Figure 4.6: Histograms of period and amplitude for our final sample of high confidence variable stars.

Notwithstanding the few remaining systematic periods in the bottom panel of Fig. 4.5, we notice that a significant fraction of the variable stars have sub per-cent amplitudes. It is indeed reassuring that we are able to pick up (likely) non-systematic periodic signals down to a few mmag. This highlights the potential for NGTS to detect low-mass transiting planets.

4.2.3 Variables on the CCD

In Fig. 4.7 we show the positions on the CCD for identified variable lightcurves with colour-coded periods for three fields. For a well-behaved field not dominated by systematic noise we expect a random distribution across the CCD. To the left are the variables for NG2047−0248, following non-systematic behaviour, as is also shown in Fig. 4.2. Just below the middle region we note an apparent underdensity, but find that it is likely not significant. In the middle plot we show the CCD positions for NG1349−1115 where most of the identified variable lightcurves are due to contamination from a saturated star in the

lower right corner. The effect of the saturation is noticable in Fig. A.15 where there are small clusters of stars with similar brightness sharing the same noise level. The figure to the right shows the variables from NG1318−4500 with gain-related systematics.

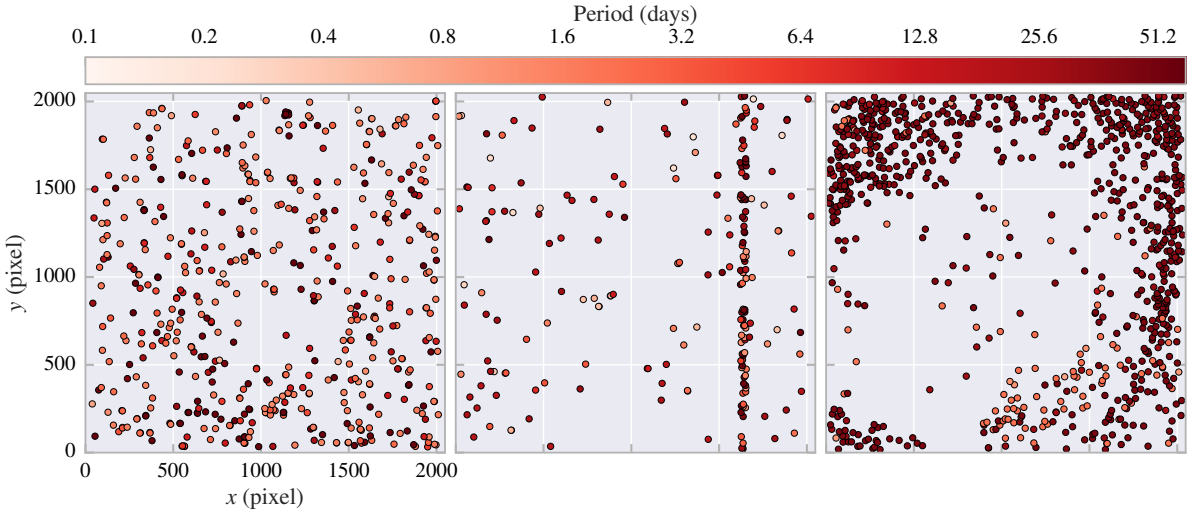


Figure 4.7: Variable objects on their CCD for fields NG2047-0248, NG1349-1115, and NG1318-4500, in order from left to right.

4.2.4 Variable lightcurves

Here we show some example lightcurves of newly discovered variable stars identified by our algorithm. By eyeballing several hundred lightcurves, we note that we very often identify the correct period for eclipsing binary stars, which suggests the modification we made to the period search is working as intended. An unfortunate consequence is that the period of most RR Lyrae stars are doubled as well.

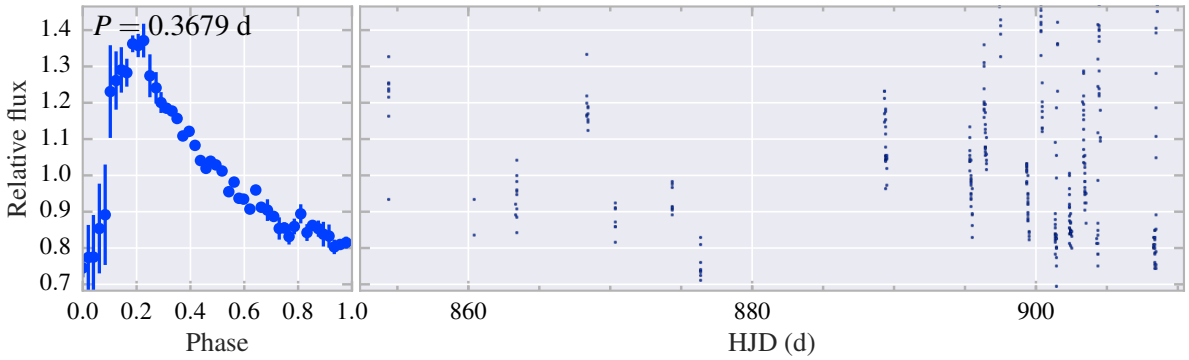


Figure 4.8: RR Lyrae (RRAB) variable star with the correct period identified from only 17 nights of data spanning 54 days. To the right is the 10 min binned light curve with individual nights. To the left the lightcurve folded on its period, showing the phase variation.

In Fig. 4.10 we show several variable stars with intermediate periods. These are not catalogued variable stars, thus it is difficult to be sure what types they are. Based on lightcurve characteristics alone, the top panel can be a γ Doradus variable, showing oscillation at two frequencies. The upper middle and bottom

panel could variable due to spots on the stellar surface as the star rotates. We take note of the photometry showing clear variations at less than a per-cent. The lower middle panel is likely a classical Cepheid variable, showing a flare at 750 days.

In Fig. 4.11 we show a few long-period variables. The top could be a semiregular variable (SR). The upper middle panel shows a periodic variable star with very low amplitudes, maximum difference being around 1%. This example shows the excellent photometry that NGTS can achieve for long-period observations. The two bottom lightcurves look very much opposite of each other, and have the same amplitudes, which is suspicious. Both of these stars are from the same field, but some quick investigation found that they are not in proximity of each other on the CCD. Due to time restrictions, we did not pursue this further and are still unsure about the nature of these stars.

Systematic periods

In Fig. 4.9 we show two examples of systematic effects in our lightcurves. The upper panel shows a typical example of a lightcurve affected by moon correlations. The dominant period coincides with the lunar cycle at ~28 days. In this example the effect induces a dip in the lightcurve at full moon from oversubtraction. In other lightcurves the correlations manifest as brightness peaks at the phase of full moon. In the lower panel we show an effect specific to the field NG0313−2230. A number of lightcurves show a trend where the brightness increases towards the middle of the season, then falls down towards the end of the observing season, with a total amplitude of about 1%. It is not clear why this is happening, and given time restrictions, no further investigation was done. Some possible scenarios involve bad weather at Paranal, or SYSREM is not detrending the shorter nights towards the end of the season as well as full nights earlier in the season.

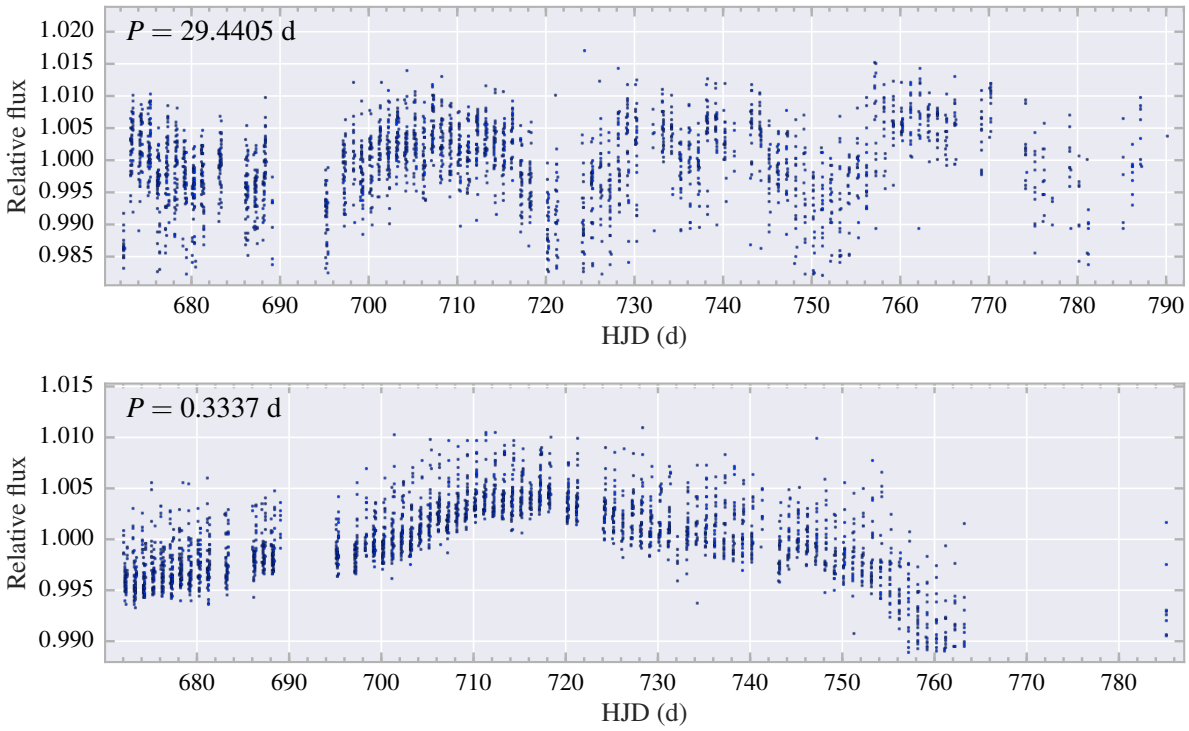


Figure 4.9: Two examples of systematic effects seen in lightcurves. Top: Photometry from NG0304−1115 affected by the Moon's phase. This is seen in many lightcurves and is thought to be the limiting systematic noise NGTS currently has. Bottom: Lightcurve from NG0313−2230 showing a ramp up, then down at the end of the observing season. This is seen in other lightcurves from this field as well.

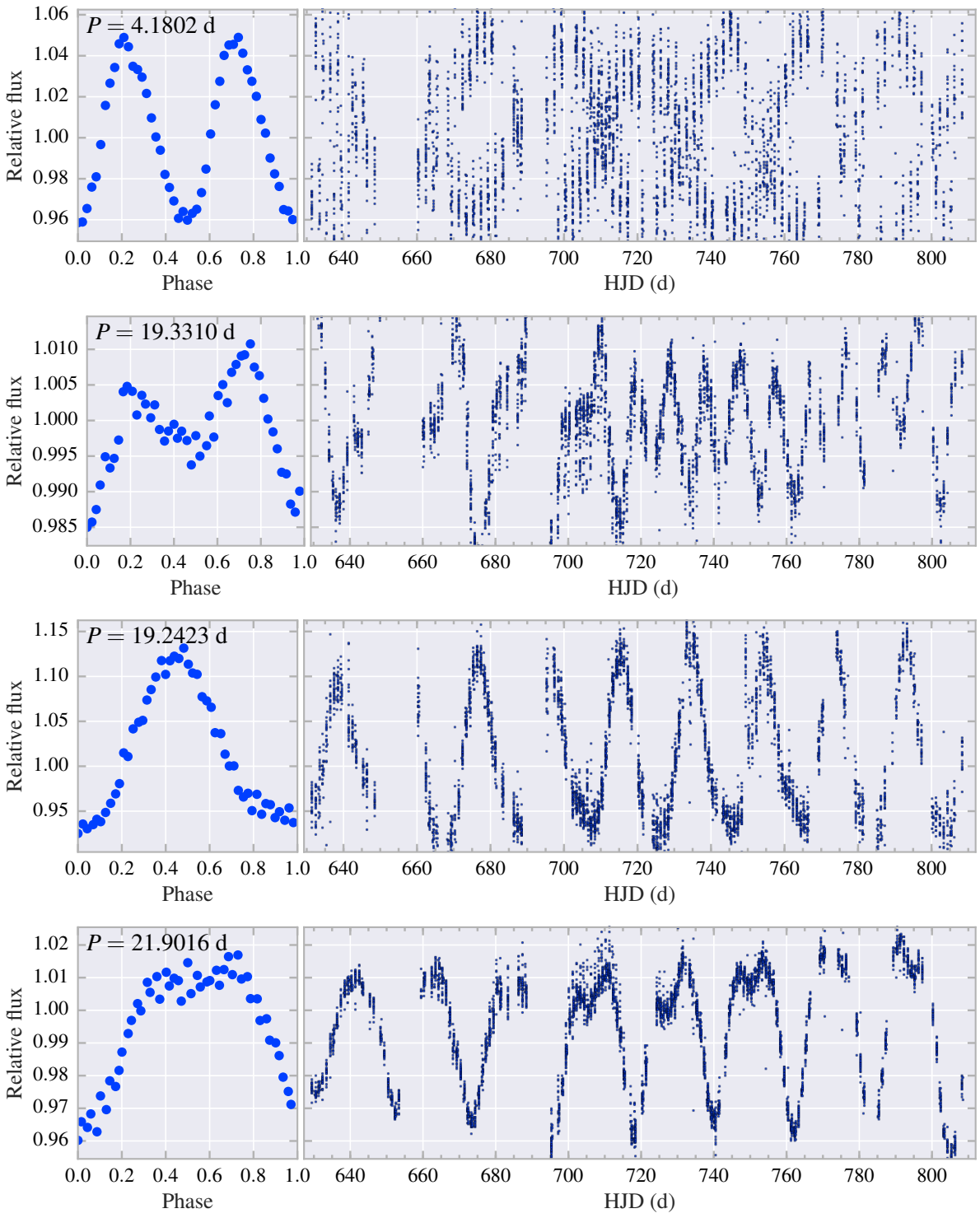


Figure 4.10: Variable stars from fields NG0531−0826 and NG0618−6441. Top: Possible γ Doradus (GDOR) variable oscillating at two periods. Upper middle: Possibly star with spots. Lower middle: Likely Cepheid variable, displaying a flare at HJD = 750. Bottom: Not clear. Shows variations reminiscent of spots, but amplitudes are very high.

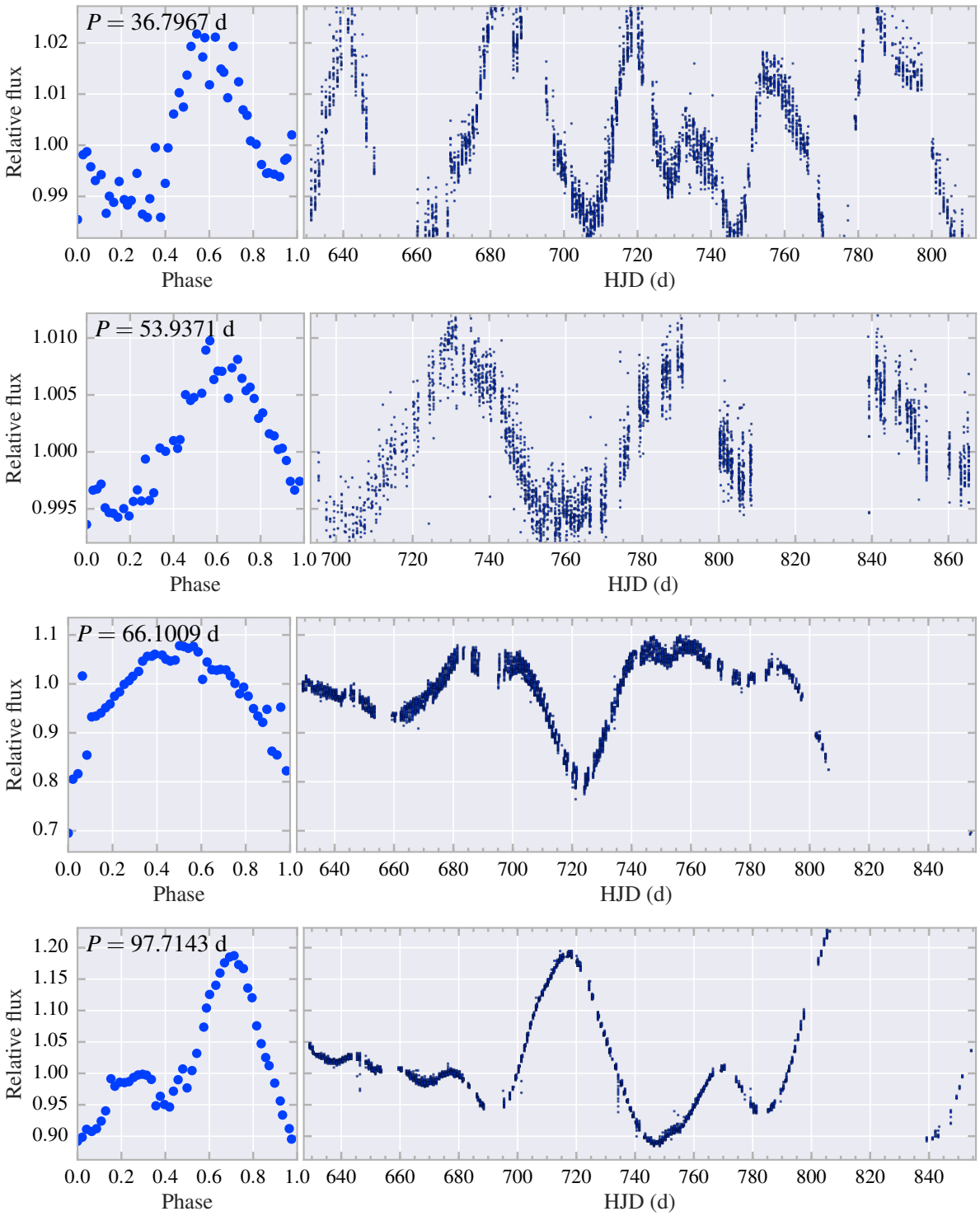


Figure 4.11: A series of long-period variables from fields NG0531−0826, NG1135−2518, and NG0522−2518. In the upper middle panel we notice the slow variation at sub per-cent amplitudes that NGTS is able to pick up, with a combination of long observations and high precision. These variations truly show the excellent photometry that NGTS is capable of.

CHAPTER 5

Discussion

In this Chapter we discuss our results further, and well as some broader implications of what we have found in Chapter 4. We mention problems with the data, suggest solutions, and discuss some future avenues of work.

5.1 Systematic noise

5.1.1 Moon phase correlations

In early stages of this work it became evident that some light curves have periodic signals of ~28–31 days correlated with the phase of the Moon due to an improper sky background estimation. Affected light curves can show a brightness increase (moon peaks^{TM}) or brightness dimming up to a ~2% level. These effects were first suspected to only affect faint stars. In their case, the effect would manifest as a brightness decrease over some days, mimicking a very long duration transit. However, later it was discovered that that the $F_{\rm rms}$ curve per field will vary between nights at the period of the moon phase, becoming a global problem that affects our long-term photometry and is ultimately limiting our precision and possibly inhibiting planet detection. In Table 4.1, the red noise level for all our fields with less than ~30 days of observations are generally 0.5 mmag lower than for other fields with several months of data. In some cases the red noise level for these fields is negligible, and the data can be fully explained by the white noise model. This goes to show that adding more data adds red noise to the long-baseline photometry, which is likely driven by the moon correlations. By applying simple corrections to the sky background algorithm, it may be possible to reduce the mean global red noise level of NGTS to 0.5 mmag, or even beyond.

The sky background correction algorithm is described in Section 2.2.3. The background levels are estimated from cells of 64×64 pixels. In crowded regions this could be a too small of an area to get a good background level. Further, the pixel cells undergo iterative k-sigma clipping to remove the stars. It is possible that this process stops prematurely, or one uses a too high number, k, that does not completely remove the stars, which will affect the background estimation. Possible improvements to the algorithm could be to increase the pixel grid sizes in which the background level is estimated, or change the k factor in the sigma clipping.

The pixel cells are treated completely independent of each other in the algorithm above. Each cell will have different estimates of the background level, which is attempted to be smoothed by interpolating between pixels between neighbouring cells. However, this will not pick up other variations in the background level in a different region of the CCD. Other algorithms take this into account,

such as NEBULISER[1]. This algorithm was developed to estimate the background level in regions with nebulosity. It will estimate the background by iterative sliding median and mean filters applied to both axes simultaneously, thus sampling a larger region of the CCD in the shape of a "plus" sign, reaching a more global estimate of the sky background that is less affected by locality.

5.1.2 Crowded fields

Dilution from overlapping apertures in crowded fields is a limiting factor in the number of true variable stars that we identify. This mainly affects fainter stars because they are more numerous, manifesting in an increased rms scatter at the faint end. Diluted stars can be identified by a catalogue cross-matching process, taking into account the magnitude limits of NGTS and looking for sources close enough together for their apertures to overlap. Given their distance from each other, a dilution factor can be computed. This has been done by NGTS team members and is in the process of being included in the final data files. It is expected that a re-analysis of the variable star identification process, excluding diluted stars, will have a substantial effect in identifying real astrophysical variability.

5.1.3 Brightness-dependent noise

In Fig. 3.1, lower panel, we showed the ratio R as function of flux after including the red noise level found for NG0304−1115. For a perfect noise model we would expect the floor of the population to follow a straight line around $R = 1$, where the rms scatter for the most quiet stars is described by the total noise model. Instead, we see that the red noise overestimates the total noise at the bright end, which curves the population of bright stars downwards. This indicates that the true red noise level is dependent on brightness, and not at a constant level. The source of this brightness dependence is difficult to identify, but a possible reason could be non-linearity in the CCD.

5.1.4 Airmass effects

Each NGTS telescope will observe a given field for a duration of 3–4 months. Throughout this period, the number of hours per night that the field is visible in the night sky will vary. At the start of an observing season, as the field rises above the horizon, it may be observable for a couple of hours, reaching about 8 hours mid-season, before the field starts setting again and becomes observable only for a few hours per night. Adding to this, as Earth rotates the altitude of the field will also vary throughout the night. At the beginning of the night the field will be close to the horizon where the light has to pass through more atmosphere than when the field is observered at a higher altitude on the sky. The change in airmass throughout the night adds a varying level of scintillation noise in the lightcurve, and manifests as "ramps" at the beginning and ends of nights in some lightcurves. These ramps are a common source of false transit detections by the BLS algorithm, and contributes to the overall systematic noise in the photometry. Some investigation was carried out to quantify the effect of these ramps, but nothing conclusive was found due to the dominating systematic noise level being from moon correlations.

5.2 Implications for planet detections

Red noise is the limiting factor in the search for small planets around bright stars. The planet yield estimates for NGTS M. N. Günther et al. (2017) was shown to be heavily dependent on the red noise

[1] http://casu.ast.cam.ac.uk/surveys-projects/software-release/background-filtering

level that NGTS ultimately achieves. The simulated yield for NGTS over its full mission (4 years) for 1 mmag red noise is 4±3 super-Earths, 19±5 small Neptunes, 16±4 large Neptunes, 55±8 Saturns, and 150±10 Jupiters. Detection of large planets is not very affected by red noise at the mmag level. WASP (D. L. Pollacco et al., 2006) had a limiting noise level around 1%, which was enough for the science case of detecting Jupiter-sized planets. Achieving a 0.5 mmag red noise level was therefore shown to mainly increase the planet yield for smaller planets: 10±3 super-Earths, 60±10 small Neptunes, 38±4 large Neptunes, 76±10 Saturns, and 158±11 Jupiters.

In Table 4.1 we show that the mean red noise level for the TEST18 dataset is 0.87±0.17 mmag, which places the expected planet yield somewhere between the two results above. The level of red noise that the moon correlations induce are thought to be of the order 0.5 mmag, based on the red noise levels of the short fields. It is believed that a correction to the sky background subtraction will at the very least bring us closer to the 0.5 mmag red noise level, which will greatly enhance the detection capabilities of Neptune-sized planets. Decreasing the red noise also has the effect of increasing sensitivity to detect small planets at longer orbital periods.

5.3 Variable stars

5.3.1 Detection fraction

The number of detected variables in a survey depends on the stellar populations surveyed, magnitude limit, length and sampling of observations, instrument methods, and analysis tools. Excluding fields with sparse data, and fields that suffer from high red noise, we find that the fraction of stars we are confident are astrophysically variable across all our fields is 1.4±0.65 per-cent. We can compare these numbers to what has been found from other surveys, as reviewed in L. Eyer and N. Mowlavi (2008). From the *Hipparcos* mission it was found that 2.26% of stars were periodic variable stars, i.e. pulsating variables and eclipsing binaries. Another 4.62% were non-periodic/not classified, while 2.79% were not investigated. The *Hipparcos* magnitude limit was 12 in the V band, while NGTS is sensitive to 15. The number of stars increases sharply for fainter limits, which makes it unsurprising that we find more variable stars. The third data release from ASAS (G. Pojmanski, 1997) contains results from a survey of 7 300 000 stars at $V < 15$, of which 0.34% of stars were found to be variable. Our precision with NGTS is about an order of magnitude better, which means that we are sensitive to low-amplitude variations from e.g. spotty stars with periods ~10–30 days, which are numerous in our sample (Fig. 4.5). OGLE-I and II (A. Udalski et al., 1992) surveyed the Galactic bulge at $I < 18$ and found a variable fraction of 0.67%. NGTS observes strictly away from the Galactic bulge due to higher stellar densities that increases the number of false positive detections due to blended eclipsing binaries. It is not clear whether, or how, the variable star population changes as a function of Galactic latitude, which also makes comparisons difficult.

5.3.2 Detection potential

Due to the nature of the variable lightcurve filtering, we will only find lightcurves with consistent variation – enough to affect the median flux level of the full observing duration. With a small modification to the code, it is possible to do variable lightcurve filtering on a nightly basis. This has the benefit of picking up transient astrophysical effects, such as stellar flares, which was shown in early stages of this work, but has not been explored further because the motivation behind the work was to assess the long-term photometric quality of NGTS. However, it is possible to build further on this by identifying stars with nightly variations and cross-match with those filtered by the full observing duration. Stars that are found

in common can be discarded, as short-period variables such as δ Scuti, RR Lyrae, and EBs will also be picked up by this method. A population of the remaining lightcurves should be transients/eruptive variables that can be studied further.

Flare stars are particularly interesting in the context of exoplanet science for various reasons. The prime targets of NGTS are K and M-dwarf stars. These, in particular M-dwarfs and even more loss mass stars, are known to have high surface activity with turbulent atmospheres, high rate of flares and XUV emission, which has sparked some discussion on the habitability potential near these stars. Planets around low-mass stars will undergo strong irradiation which could destroy a potential atmosphere (P. J. Wheatley et al., 2017; E. Bolmont et al., 2017), but it is not clear what effect high-energy irradiation would have on habitability. The activity on these stars also make exoplanet hunting more difficult with radial velocity measurements. Stars with high rates of flares typically have high chromospheric activity and turbulent atmospheres, which can have limiting effects on the detection of planet-induced reflex motions of the star. Studying flares may help us understand these stars and their activity, thus aiding in exoplanet detection.

5.3.3 Systematic periods

Finding a way to correctly deal with systematic periods is important for any future work on NGTS variable stars. Throughout this study we experimented with various ways of doing this. A first approach consisted of only masking commonly known systematic periods, as well as periods ±5% within these. It was later found out that this had a profound effect on the efficacy of our period-search algorithm when comparing with known variable stars from catalogues. Upon investigating the reasons for our low number of matching periods, it was found that in most cases a harmonic of the period was selected due to the dominant period being in viscinity of the masked periods. A large number of our identified variable stars are indeed RR Lyraes, short-orbit eclipsing binaries and δ Scutis, which are quite often found close to these periods. To mitigate this, we tried to model the periodic variations with a simple sum of a sine and cosine term without masking, which is the approach used in the current study. The catalogue period matching exercise yielded better results, but from Fig. 4.5 we see the systematic periods persist. A solution, which was tried in early stages of this work, would be to also remove ±5 of frequencies close to the systematic periods. Similar adjustments have been done in D. J. Armstrong et al. (2016). However, it was found that this was computationally demanding due to the fine frequency spacing we use. While the study progressed we found that at this stage we do not require a highly accurate period-finding algorithm, but note that future studies on e.g. automated classification would benefit greatly, and demand, an improved algorithm.

However, the systematic periods can be used to our advantage. Most lightcurves that have a dominant systematic period are likely to be non-variable, or variable at a very long period, as any quasi-variability on the timescale we use will probably show a dominant LS signal. One could then simply remove these lightcurves and take the remaining stars as true variable stars to continue with the analysis. This may work as a simple solution if the goal is to do quick variable star science, but for this work it was more important to understand what causes the systematic noise in the first place, and perhaps identify possible solutions.

5.3.4 Choice of periodogram

The autocorrelation function (ACF) was also explored as a means to identify periods in our lightcurves. ACF works by calculating the correlation of the lightcurve with itself while shifting the signal in time. Given a periodic lightcurve, the ACF will show a peak every integer multiple of the period. The period is then typically found from an averaged distance between the first three peaks in the periodogram. This

periodogram is more robust for detecting correct periods of eclipsing binaries, as a correlation at half the binary period will always be smaller than for the full period for binaries of unequal sizes. Additionally, ACF is less sensitive to long-term trends in the data, which can sometimes show strong peaks in LS periodograms. In the literature, ACF has also proven to be very robust for measuring rotation periods of *Kepler* M-dwarf stars (A. McQuillan, S. Aigrain, and T. Mazeh, 2013; A. McQuillan, T. Mazeh, and S. Aigrain, 2014). However, ACF generally requires better treatment of the data, and is more affected by un-even sampling than LS periodograms.

Due to general noise and systematics in data, the ACF periodograms will often show spurious peaks, making it difficult to compare the peaks alone. A smoothing of the periodogram, typically with a Gaussian kernel, is required. The width of the smoothing kernel will depend on the periodic signal that one searches for, and is typically found from trial-and-error. For this work we were interested in searching a range of periods and variable stars, most of which resemble sinusoid signals, which LS is more sensitive to. The LS periodogram, with its versatility and ease of implementation was found to be better suited for our purposes.

5.4 Future steps

5.4.1 Rotation periods of NGTS stars

The concentration of stars with periods 10–30 days in Figs. 4.5 and 4.6 are consistent with rotation periods of G, K and M stars, assuming they are main sequence, as shown for the full sample of *Kepler* main sequence stars by A. McQuillan, T. Mazeh, and S. Aigrain (2014). Furthermore, their study showed that the expected amplitude of brightness variations from their rotation with spot-covered surfaces are in the range of 0.5–30 mmag, with a downwards trend towards low-mass stars. Our amplitudes (5–15 mmag) are consistent with the upper envelope of their sample. Further study needs to be done to rule out systematic effects related to moon correlations, but if confirmed that they are rotation periods of main sequence stars, it would open up new avenues for studies of stellar evolution from the ground with NGTS.

5.4.2 Pipeline variable filtering

Assuming that dilution effects in crowded fields are properly accounted for and the sky background correction is changed to behave more adequately – as is being worked on by other team members – the software should perform remarkably better in identifying bona fide variable stars. At this stage it could be ingested into the NGTS pipeline in filtering variable lightcurves. Variable star filtering in the pipeline is useful for several reasons. Before transit searches are performed on the lightcurves, the data undergo systematic error corrections, as described in Section 2.2.2. Although the SYSREM algorithm uses an rms cut-off for its reference star selection, low-amplitude stars may still pass the filtering and affect the global detrending of lightcurves. A reliable variable star filter would result in better detrending, which in turn may help uncover planetary signals.

For NGTS data and most transit surveys, transits are searched for with the highly efficient BLS algorithm (G. Kovács, S. Zucker, and T. Mazeh, 2002), which attempts to fit a transit box model to the data phase folded on trial periods. This has the unwanted effect of catching a range of variable stars, most notably eclipsing binaries (Section 1.4.3). Time spent on follow-up vetting of these false positives can be minimised if the data undergo an accurate variable filtering prior to the transit search.

5.4.3 Automated classification

The onset of large volumes of data from various astronomical surveys (WASP, LSST, TESS, HAT, find more) has pushed astronomers to use automated tools to analyse data that can no longer be done manually. In the context of automated variable star classification, machine learning techniques have gained traction and have proven to be efficient and accurate for both evenly sampled space data as well as sparse ground-based data. The idea is to extract a number of "features" from the lightcurve based on various noise characteristics, periods, amplitudes, and more, that is used as input to a machine learning classifier that will put the lightcurve in the most probable variable class based on its inputs. The classifier is first trained on a sample of stars with known classes, but based on inputs from new lightcurves, the classifier can automatically evolve and learn. J. W. Richards et al. (2011) applied a Random Forest (RF) classifier to OGLE and *Hipparcos* data and showed a 22.8% error rate on 25-class dataset. For pulsational variables alone, the discovery efficiency was at 99.1%, and 98.2% for eclipsing systems. For the classification, they used 32 periodic features extracted from the generalised Lomb-Scargle periodogram, as well as 20 other non-periodic features extracted from the lightcurve. The selection of a high number of features introduces an ambiguity, as a number of these features may or may not be descriptive of all variable classes. In an attempt to minimise the selection of features, D. J. Armstrong et al. (2016) applied a Self-Organising Map (SOM) on *K2* data prior to an RF classifier. The SOM is a type of clustering algorithm that – based on the shape of the lightcurve – will group similar variables together in a two-dimensional space. This method proved to be very efficient in separating close binaries from detached binary stars, as well as other pulsators such as RR Lyraes and δ Scutis. The position of each variable (SOM Index) is then used as a feature in the RF classifier. When comparing the importance of each feature in classifying a lightcurve, it was shown that the SOM index was the second most important feature in the classifier, next to the dominant period, which potentially holds more information than a number of other features would combined.

The original intent of this work was to identify variable stars through F_{rms} diagrams and building a noise model, then perform automated machine learning classification of the identified periodic variable lightcurves using an RF classifier with the SOM as an input. However, upon completing the variable star identification process, subsequent period search unveiled that a large fraction of the variable lightcurves were noise and/or dominated by systematic periods due to issues previously discussed in this Chapter, namely moon correlations, dilution, window function effects, sidereal day correlation, airmass effects, and likely other hitherto undentified effects. To perform an adequate classification, most of these issues need to be addressed and improved on so that astrophysical variability will dominate the signal. Most of the issues are currently being worked on in the NGTS team, and future data quality will significantly benefit from these efforts. For these reasons it was decided that automated classification at this stage was premature, and more effort was put into investigating the sources of noise that is limiting the experiment, which constitutes the bulk of this work.

5.5 Conclusion

In this study we have carried out a global assessment of the noise level for NGTS' long time-series photometry from its first year of observations. We have built a white noise model for NGTS taking into account survey and camera characteristics, and from this estimated the level of systematic noise in the data. We found that the median red noise level across the 30 fields we analysed is 0.82±0.10 mmag, demonstrating that NGTS indeed is able to reach sub-mmag noise levels over the 3–4 months that each camera is observing a given field. Further investigation was carried out in an attempt to identify possible sources of systematic noise, and we found that an incorrect sky background subtraction leads to a

global ~28 day periodic variation in the noise level for all our fields. This is currently in the process of being corrected by NGTS team members. Addressing this problem may bring the red noise level below 0.5 mmag, which will have a great effect on the number of super-Earth size exoplanets NGTS can detect.

We used our noise model to develop a robust, automated way of identifying periodic/semi-periodic variable stars using a brightness-independent rms threshold. Assuming a Gaussian distribution in brightness-independent rms (R) of non-variable stars, we find 7976 variable lightcurves outside the 97th percentile. We find that a high fraction (~66%) of these are lightcurves with high noise and dominant systematic periods. After applying vetting critera on systematic periods and noise levels for, we find 1245 stars that we believe are astrophysically variable at high confidence. This constitutes 1.31% from our reduced sample of stars. About 900 of these are new variable stars, including δ Scutis, RR Lyraes, various types of eclipsing binaries, Cepheids, and semiregular variables. There is also indication that we are picking up rotation periods of spotted main sequence stars, which would be the first time for a ground-based wide-field survey. Efforts are currently being made to address issues with systematic noise by NGTS team members, and we expect that this will lower the number of non-astrophysical variables, in turn greatly increasing the efficiency of our algorithm in detecting bona fide variable stars. For future work, we expect that this can be ingested into the data pipeline to optimise systematic detrending, and the sample of identified variable stars can be used for automated classification using e.g. machine learning techniques.

Bibliography

Armstrong, D. J. et al. (2016). "K2 variable catalogue - II. Machine learning classification of variable stars and eclipsing binaries in K2 fields 0-4". In: *MNRAS* 456, pp. 2260–2272. DOI: 10.1093/mnras/stv2836. arXiv: 1512.01246 [astro-ph.SR] (cit. on pp. 36, 58, 60).

Bakos, G. et al. (2004). "Wide-Field Millimagnitude Photometry with the HAT: A Tool for Extrasolar Planet Detection". In: *PASP* 116, pp. 266–277. DOI: 10.1086/382735. eprint: astro-ph/0401219 (cit. on p. 2).

Batalha, N. M. et al. (2010). "Pre-spectroscopic False-positive Elimination of Kepler Planet Candidates". In: *ApJL* 713, pp. L103–L108. DOI: 10.1088/2041-8205/713/2/L103. arXiv: 1001.0392 [astro-ph.EP] (cit. on p. 18).

Bolmont, E. et al. (2017). "Water loss from terrestrial planets orbiting ultracool dwarfs: implications for the planets of TRAPPIST-1". In: *MNRAS* 464, pp. 3728–3741. DOI: 10.1093/mnras/stw2578. arXiv: 1605.00616 [astro-ph.EP] (cit. on p. 58).

Borucki, W. J. et al. (2010). "Kepler Planet-Detection Mission: Introduction and First Results". In: *Science* 327, p. 977. DOI: 10.1126/science.1185402 (cit. on p. 2).

Burke, C. J. et al. (2015). "Terrestrial Planet Occurrence Rates for the Kepler GK Dwarf Sample". In: *ApJ* 809, 8, p. 8. DOI: 10.1088/0004-637X/809/1/8. arXiv: 1506.04175 [astro-ph.EP] (cit. on p. 10).

Charbonneau, D. et al. (2000). "Detection of Planetary Transits Across a Sun-like Star". In: *ApJL* 529, pp. L45–L48. DOI: 10.1086/312457. eprint: astro-ph/9911436 (cit. on p. 4).

Chazelas, B. et al. (2012). "NGTS: a robotic transit survey to detect Neptune and super-Earth mass planets". In: *Ground-based and Airborne Telescopes IV*. Vol. 8444. Proc. SPIE, 84440E. DOI: 10.1117/12.925755 (cit. on p. 19).

Collier Cameron, A. et al. (2006). "A fast hybrid algorithm for exoplanetary transit searches". In: *MNRAS* 373, pp. 799–810. DOI: 10.1111/j.1365-2966.2006.11074.x. eprint: astro-ph/0609418 (cit. on p. 22).

de Wit, J. et al. (2016). "A combined transmission spectrum of the Earth-sized exoplanets TRAPPIST-1 b and c". In: *Nature* 537, pp. 69–72. DOI: 10.1038/nature18641. arXiv: 1606.01103 [astro-ph.EP] (cit. on p. 4).

Dravins, D. et al. (1998). "Atmospheric Intensity Scintillation of Stars. III. Effects for Different Telescope Apertures". In: *PASP* 110, pp. 610–633. DOI: 10.1086/316161 (cit. on p. 13).

Eyer, L. and N. Mowlavi (2008). "Variable stars across the observational HR diagram". In: *Journal of Physics Conference Series*. Vol. 118. Journal of Physics Conference Series, p. 012010. DOI: 10.1088/1742-6596/118/1/012010. arXiv: 0712.3797 (cit. on p. 57).

Fressin, F. et al. (2013). "The False Positive Rate of Kepler and the Occurrence of Planets". In: *ApJ* 766, 81, p. 81. DOI: 10.1088/0004-637X/766/2/81. arXiv: 1301.0842 [astro-ph.EP] (cit. on pp. 4, 10).

Gillon, M. et al. (2013). "SPECULOOS: Search for habitable Planets EClipsing ULtra-cOOl Stars". In: *Protostars and Planets VI Posters* (cit. on p. 4).

Gillon, M. et al. (2016). "Temperate Earth-sized planets transiting a nearby ultracool dwarf star". In: *Nature* 533, pp. 221–224. DOI: 10.1038/nature17448. arXiv: 1605.07211 [astro-ph.EP] (cit. on p. 4).

Graham, M. J. et al. (2013). "Using conditional entropy to identify periodicity". In: *MNRAS* 434, pp. 2629–2635. DOI: 10.1093/mnras/stt1206. arXiv: 1306.6664 [astro-ph.IM] (cit. on p. 32).

Günther, M. N. et al. (2017). "A new yield simulator for transiting planets and false positives: application to the Next Generation Transit Survey". In: *MNRAS* 465, pp. 3379–3389. DOI: 10.1093/mnras/stw2908. arXiv: 1611.02526 [astro-ph.EP] (cit. on pp. 10, 20, 21, 56).

Haswell, C. A. (2010). *Transiting Exoplanets* (cit. on pp. 3, 6).

Henry, G. W. et al. (2000). "A Transiting "51 Peg-like" Planet". In: *ApJL* 529, pp. L41–L44. DOI: 10.1086/312458 (cit. on p. 4).

Høg, E. et al. (2000). "The Tycho-2 catalogue of the 2.5 million brightest stars". In: *A&A* 355, pp. L27–L30 (cit. on p. 26).

Kovács, G., S. Zucker, and T. Mazeh (2002). "A box-fitting algorithm in the search for periodic transits". In: *A&A* 391, pp. 369–377. DOI: 10.1051/0004-6361:20020802. eprint: astro-ph/0206099 (cit. on p. 59).

Léger, A. et al. (2009). "Transiting exoplanets from the CoRoT space mission. VIII. CoRoT-7b: the first super-Earth with measured radius". In: *A&A* 506, pp. 287–302. DOI: 10.1051/0004-6361/200911933. arXiv: 0908.0241 [astro-ph.EP] (cit. on p. 2).

Lomb, N. R. (1976). "Least-squares frequency analysis of unequally spaced data". In: *Astrophysics and Space Science* 39, pp. 447–462. DOI: 10.1007/BF00648343 (cit. on p. 32).

Mayor, M. and D. Queloz (1995). "A Jupiter-mass companion to a solar-type star". In: *Nature* 378, pp. 355–359. DOI: 10.1038/378355a0 (cit. on p. 2).

McQuillan, A., S. Aigrain, and T. Mazeh (2013). "Measuring the rotation period distribution of field M dwarfs with Kepler". In: *MNRAS* 432, pp. 1203–1216. DOI: 10.1093/mnras/stt536. arXiv: 1303.6787 [astro-ph.SR] (cit. on pp. 32, 59).

McQuillan, A., T. Mazeh, and S. Aigrain (2014). "Rotation Periods of 34,030 Kepler Main-sequence Stars: The Full Autocorrelation Sample". In: *ApJS* 211, 24, p. 24. DOI: 10.1088/0067-0049/211/2/24. arXiv: 1402.5694 [astro-ph.SR] (cit. on pp. 48, 59).

Osborn, J. et al. (2015a). "Atmospheric scintillation in astronomical photometry". In: *MNRAS* 452, pp. 1707–1716. DOI: 10.1093/mnras/stv1400. arXiv: 1506.06921 [astro-ph.IM] (cit. on p. 13).

– (2015b). "Atmospheric scintillation in astronomical photometry". In: *MNRAS* 452, pp. 1707–1716. DOI: 10.1093/mnras/stv1400. arXiv: 1506.06921 [astro-ph.IM] (cit. on p. 30).

Pojmanski, G. (1997). "The All Sky Automated Survey". In: *AcA* 47, pp. 467–481. eprint: astro-ph/9712146 (cit. on p. 57).

Pollacco, D. L. et al. (2006). "The WASP Project and the SuperWASP Cameras". In: *PASP* 118, pp. 1407–1418. DOI: 10.1086/508556. eprint: astro-ph/0608454 (cit. on pp. 2, 9, 20, 22, 57).

Press, W. H. and G. B. Rybicki (1989). "Fast algorithm for spectral analysis of unevenly sampled data". In: *ApJ* 338, pp. 277–280. DOI: 10.1086/167197 (cit. on p. 34).

Richards, J. W. et al. (2011). "On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data". In: *ApJ* 733, 10, p. 10. DOI: 10.1088/0004-637X/733/1/10. arXiv: 1101.1959 [astro-ph.IM] (cit. on p. 60).

Ricker, G. R. et al. (2014). "Transiting Exoplanet Survey Satellite (TESS)". In: *Space Telescopes and Instrumentation 2014: Optical, Infrared, and Millimeter Wave*. Vol. 9143. Proc. SPIE, p. 914320. DOI: 10.1117/12.2063489. arXiv: 1406.0151 [astro-ph.EP] (cit. on p. 2).

Scargle, J. D. (1982). "Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data". In: *ApJ* 263, pp. 835–853. DOI: 10.1086/160554 (cit. on p. 32).

Tamuz, O., T. Mazeh, and S. Zucker (2005). "Correcting systematic effects in a large set of photometric light curves". In: *MNRAS* 356, pp. 1466–1470. DOI: 10.1111/j.1365-2966.2004.08585.x. eprint: astro-ph/0502056 (cit. on pp. 22, 25).

Udalski, A. et al. (1992). "The Optical Gravitational Lensing Experiment". In: *AcA* 42, pp. 253–284 (cit. on p. 57).

Wheatley, P. J. et al. (2013). "The Next Generation Transit Survey (NGTS)". In: *European Physical Journal Web of Conferences*. Vol. 47. European Physical Journal Web of Conferences, p. 13002. DOI: 10.1051/epjconf/20134713002. arXiv: 1302.6592 [astro-ph.EP] (cit. on p. 19).

Wheatley, P. J. et al. (2017). "Strong XUV irradiation of the Earth-sized exoplanets orbiting the ultracool dwarf TRAPPIST-1". In: *MNRAS* 465, pp. L74–L78. DOI: 10.1093/mnrasl/slw192. arXiv: 1605.01564 [astro-ph.EP] (cit. on p. 58).

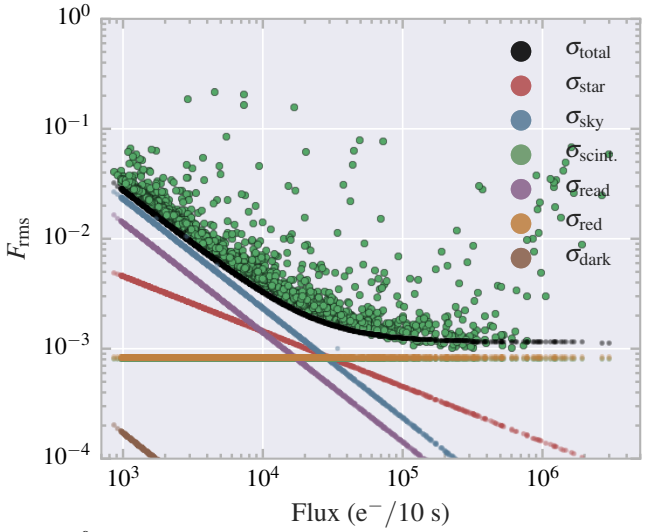
Zechmeister, M. and M. Kürster (2009). "The generalised Lomb-Scargle periodogram. A new formalism for the floating-mean and Keplerian periodograms". In: *A&A* 496, pp. 577–584. DOI: 10.1051/0004-6361:200811296. arXiv: 0901.2573 [astro-ph.IM] (cit. on p. 32).

APPENDIX A

Field summary plots

Here follows summary plots for each field that we have analysed. The figure contents are the same for all subsequent figures.

Top left: F_{rms} plot for the field including data from the whole observing season, with noise model. Middle left: The selected variable stars in that field (red). Grey points in coloured regions are stars with less than 20% data than the median number of observations in the field and are excluded. Middle right: Noise scaling with bin width for a typical V=10 star from the field. The circular points are the binned data, and the dotted line is the $1/\sqrt{n}$ white noise behavour, using the unbinned data as reference. Lower left: Identified variables plotted by their positions on the CCD, colour-coded by their dominant period. Lower right: Identified variables plotted in period-amplitude space. The dotted lines indicate positions of common aliases of the 1-day systematic period, suggesting those targets do not have astrophysical variability.



Field	NG0304-1115	-
Camera	809	-
Date range	20151104–20160301	-
Span	118	days
Nights	94	days
Apertures	3691	-
Red noise	0.81	mmag
Variables	190	-
Real variables	85	-
Variable frac.	2.43	%
Median period	1.21	days

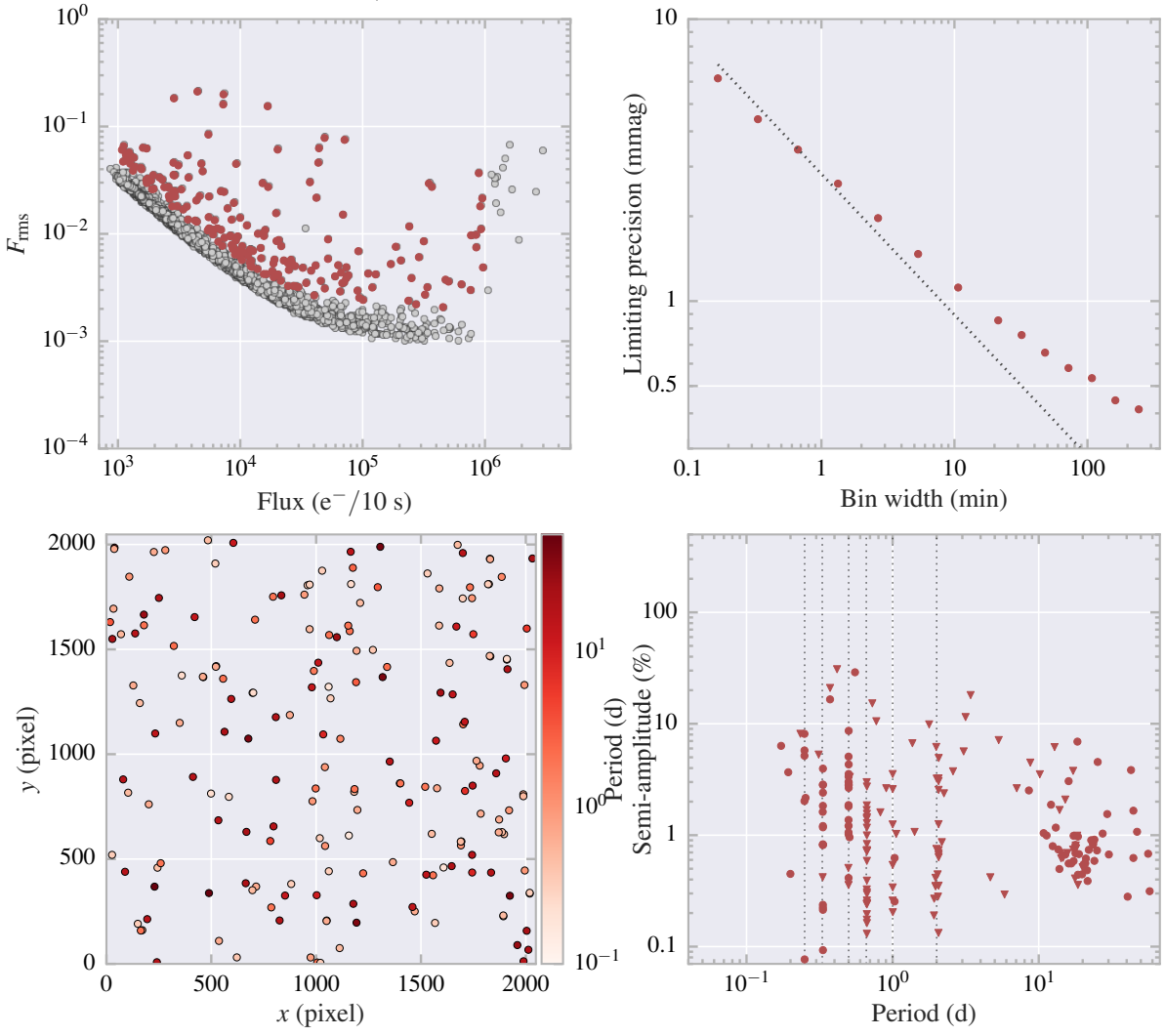
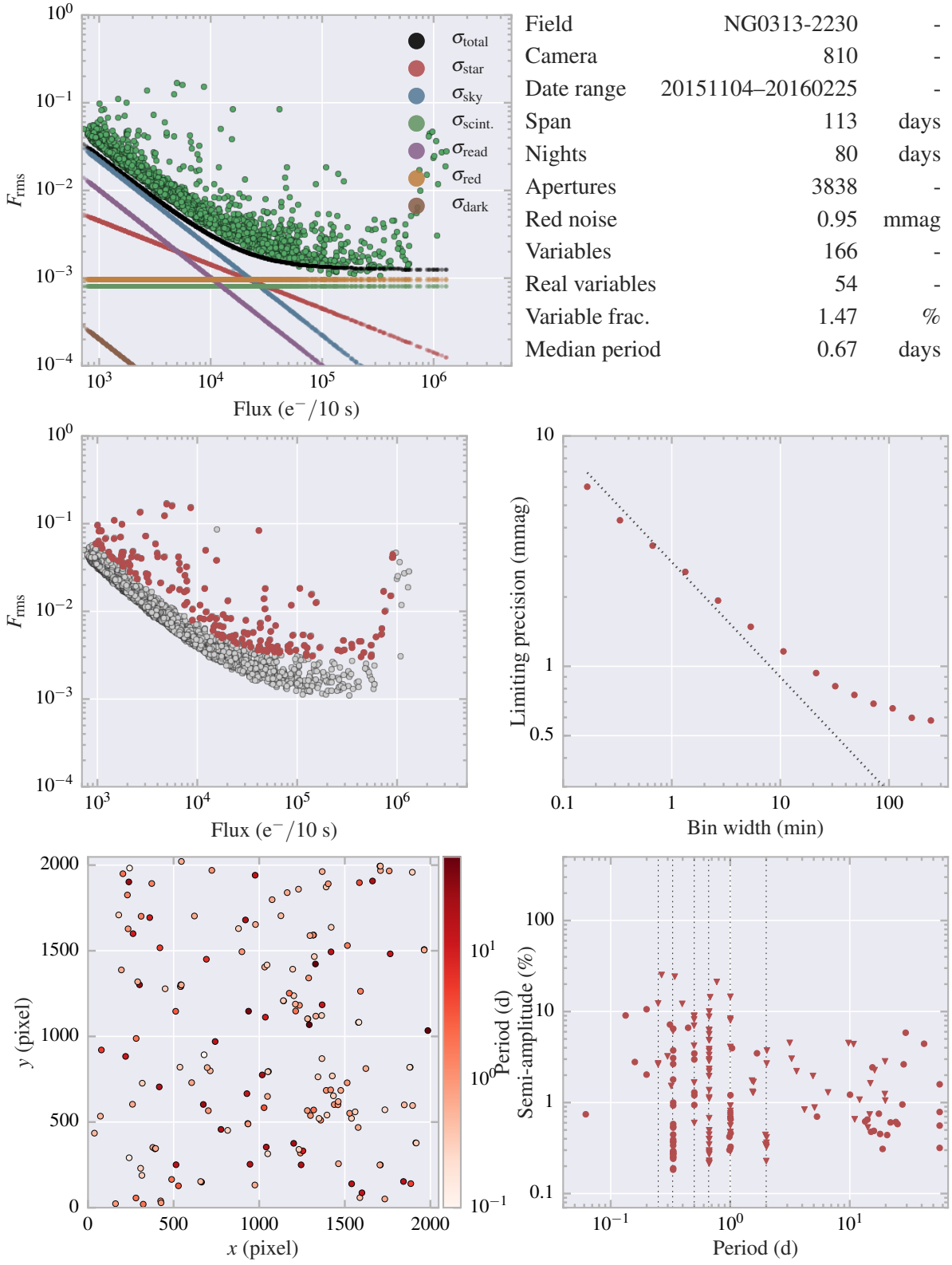
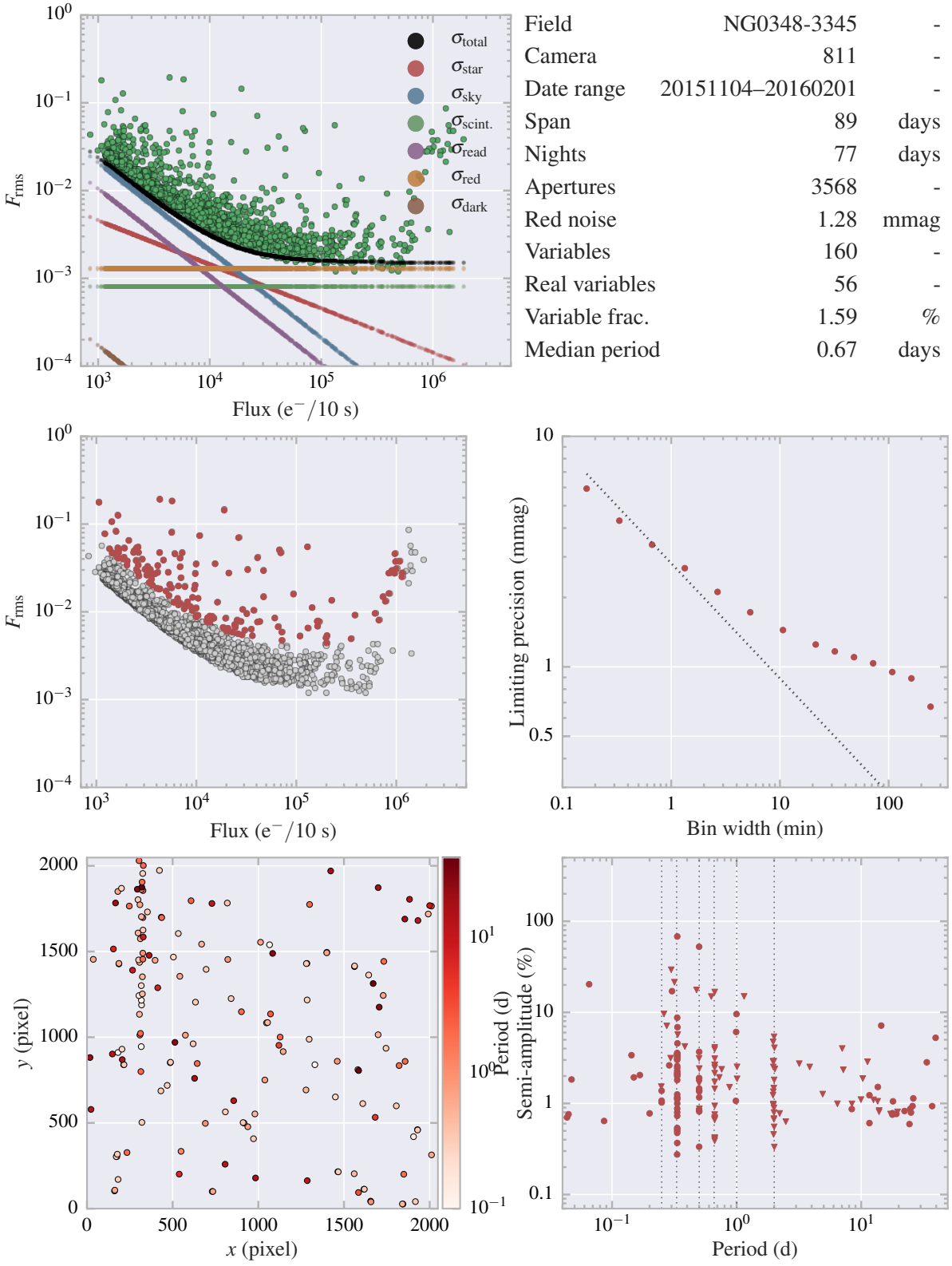


Figure A.1: Field summary for NG0304−1115.



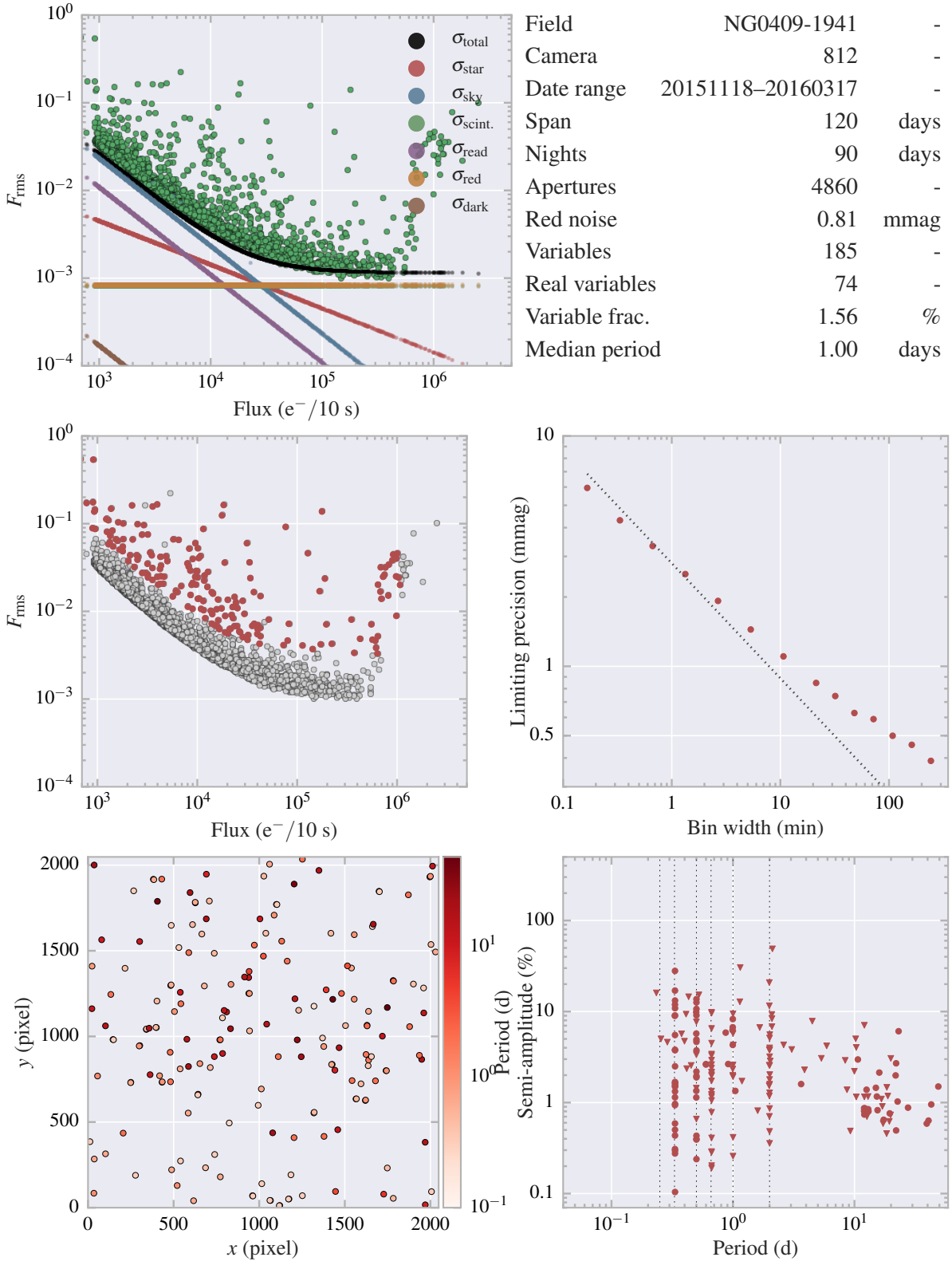
Field	NG0313-2230	-
Camera	810	-
Date range	20151104–20160225	-
Span	113	days
Nights	80	days
Apertures	3838	-
Red noise	0.95	mmag
Variables	166	-
Real variables	54	-
Variable frac.	1.47	%
Median period	0.67	days

Figure A.2: Field summary for NG0313−2230.



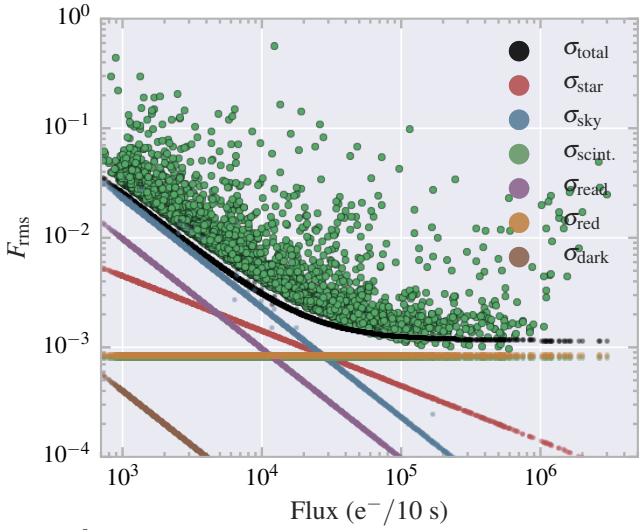
Field	NG0348-3345	-
Camera	811	-
Date range	20151104–20160201	-
Span	89	days
Nights	77	days
Apertures	3568	-
Red noise	1.28	mmag
Variables	160	-
Real variables	56	-
Variable frac.	1.59	%
Median period	0.67	days

Figure A.3: Field summary for NG0348−3345.

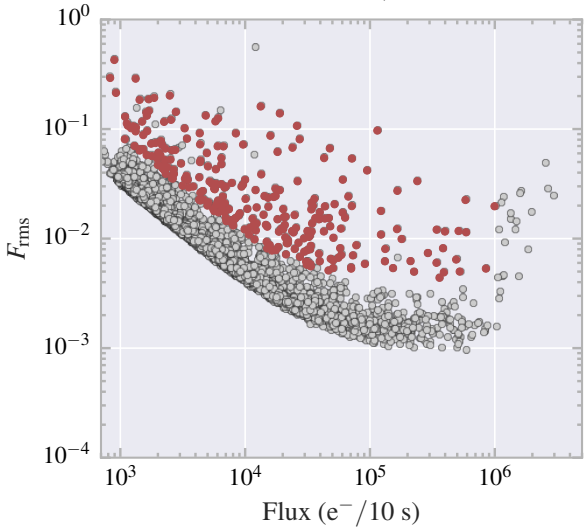


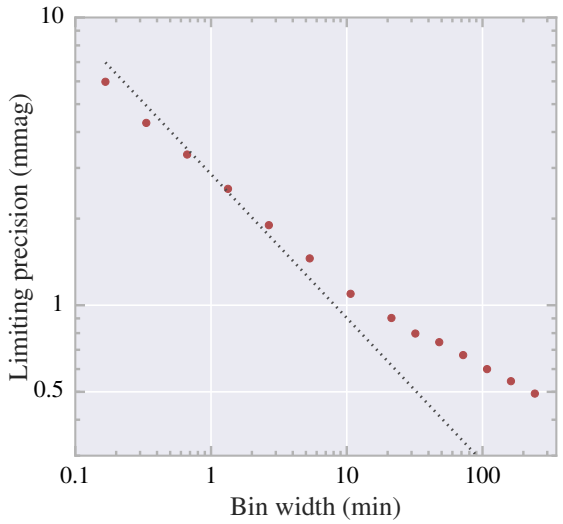
Field	NG0409-1941	-
Camera	812	-
Date range	20151118–20160317	-
Span	120	days
Nights	90	days
Apertures	4860	-
Red noise	0.81	mmag
Variables	185	-
Real variables	74	-
Variable frac.	1.56	%
Median period	1.00	days

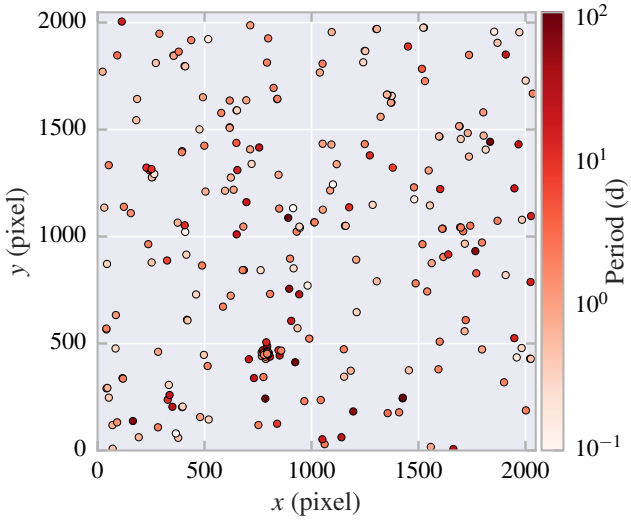
Figure A.4: Field summary for NG0409−1941.



Field	NG0522-2518	-
Camera	802	-
Date range	20150922–20160503	-
Span	224	days
Nights	150	days
Apertures	8505	-
Red noise	0.84	mmag
Variables	264	-
Real variables	99	-
Variable frac.	1.19	%
Median period	1.98	days







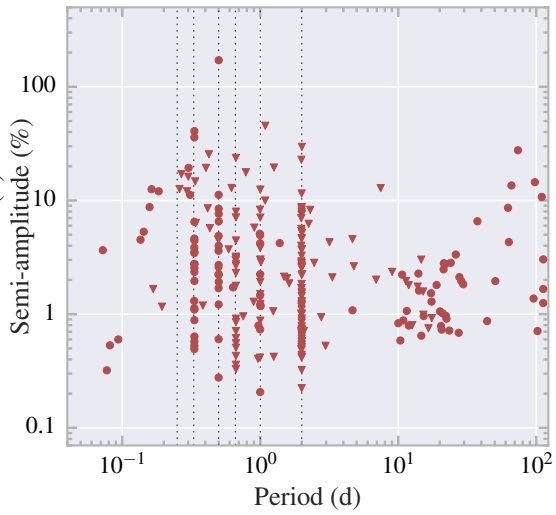
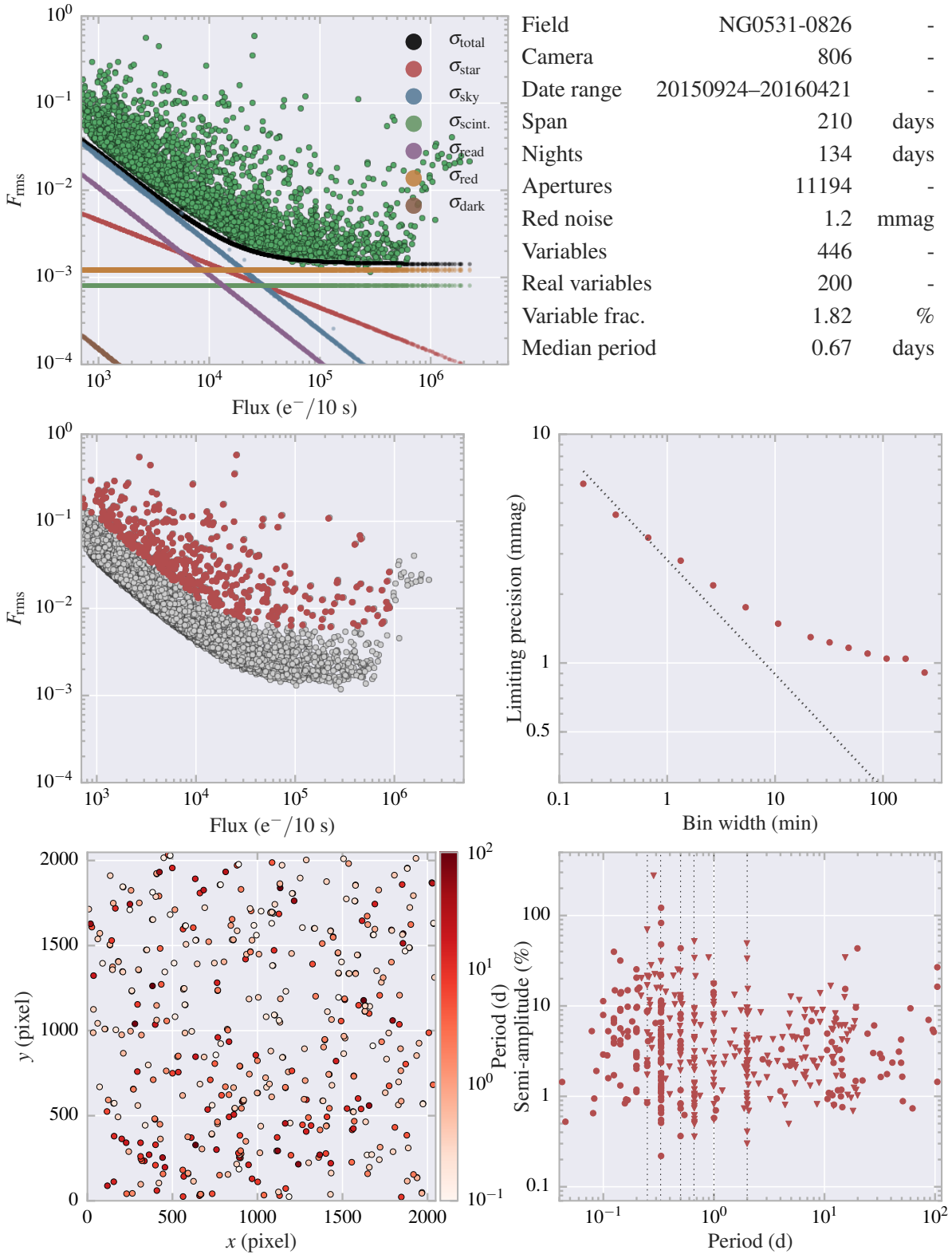
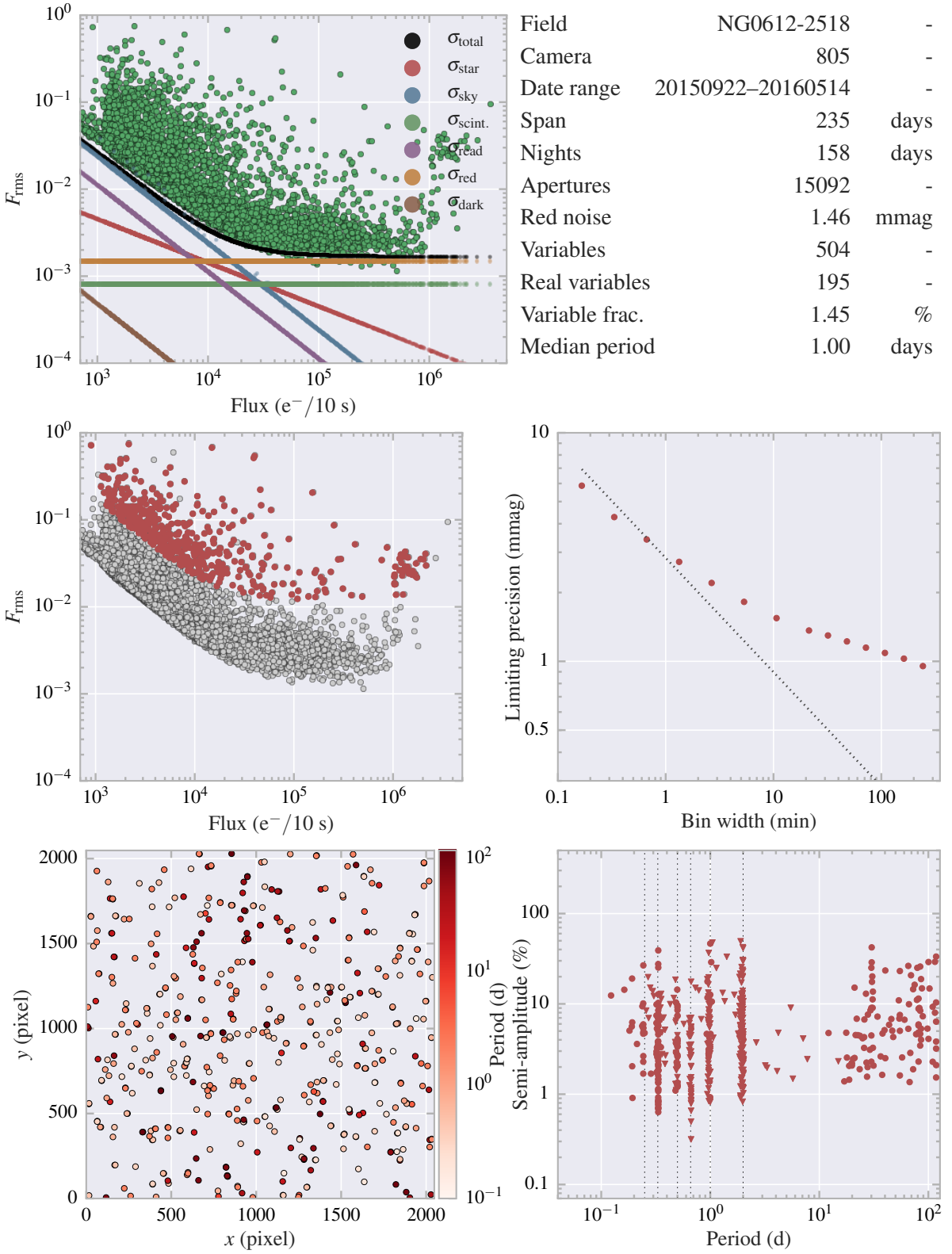


Figure A.5: Field summary for NG0522−2518.



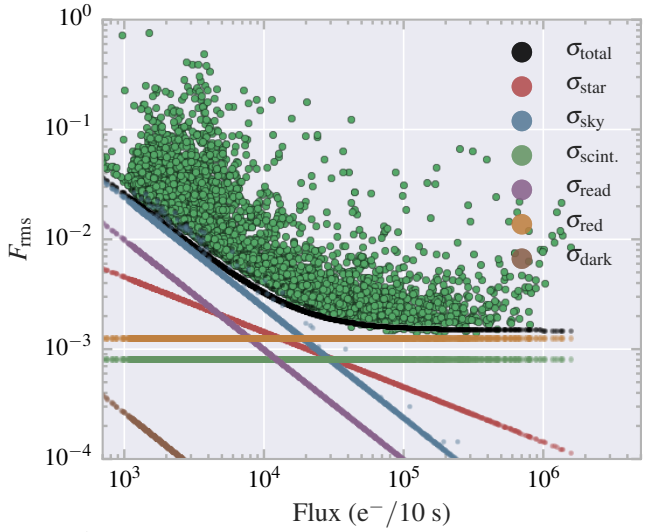
Field	NG0531-0826	-
Camera	806	-
Date range	20150924–20160421	-
Span	210	days
Nights	134	days
Apertures	11194	-
Red noise	1.2	mmag
Variables	446	-
Real variables	200	-
Variable frac.	1.82	%
Median period	0.67	days

Figure A.6: Field summary for NG0531−0826.

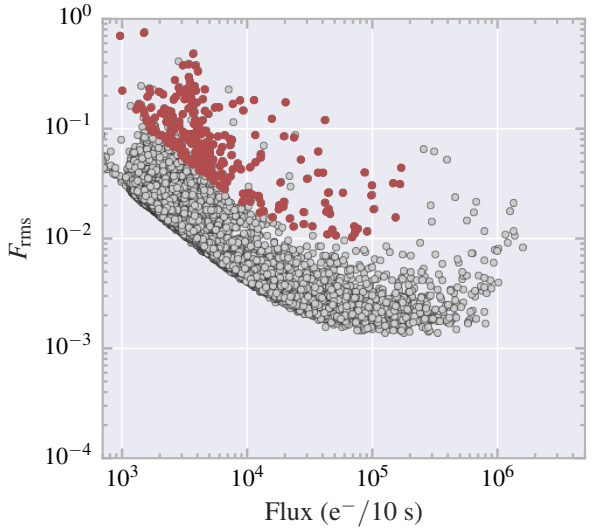


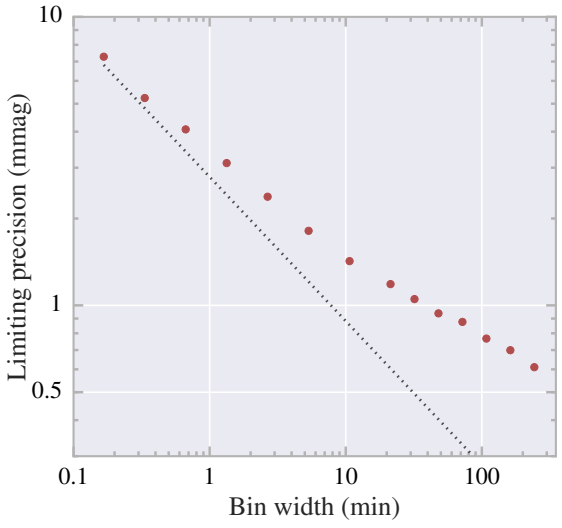
Field	NG0612-2518	-
Camera	805	-
Date range	20150922–20160514	-
Span	235	days
Nights	158	days
Apertures	15092	-
Red noise	1.46	mmag
Variables	504	-
Real variables	195	-
Variable frac.	1.45	%
Median period	1.00	days

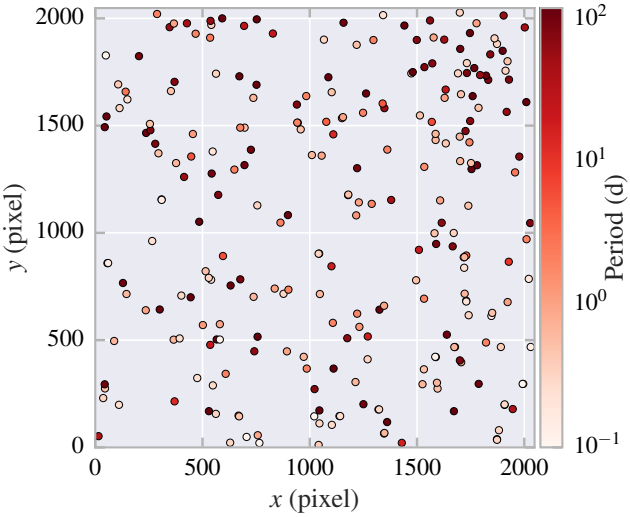
Figure A.7: Field summary for NG0612−2518.



Field	NG0618-6441	-
Camera	801	-
Date range	20150922–20160518	-
Span	239	days
Nights	158	days
Apertures	11362	-
Red noise	1.25	mmag
Variables	258	-
Real variables	136	-
Variable frac.	1.22	%
Median period	1.24	days







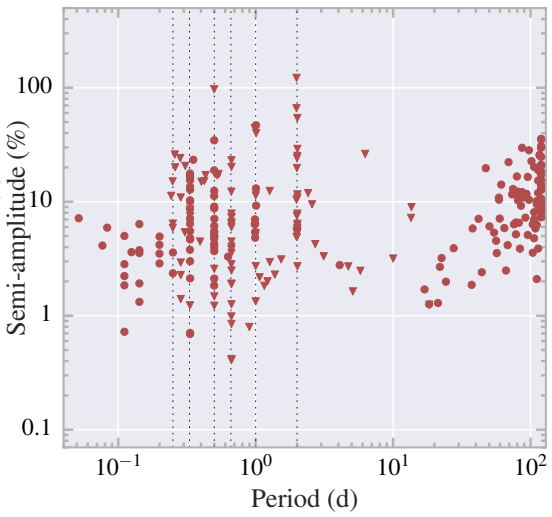
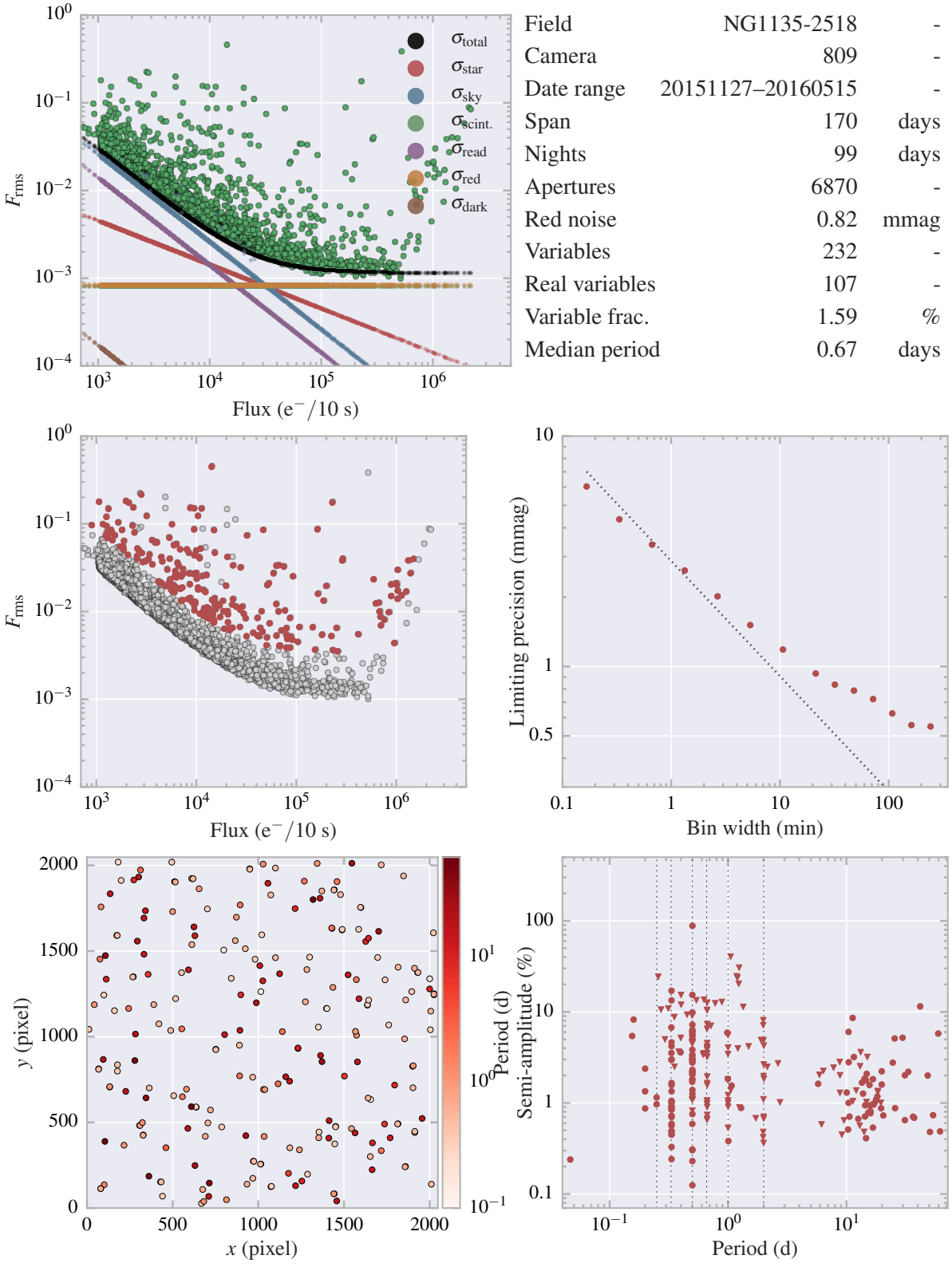
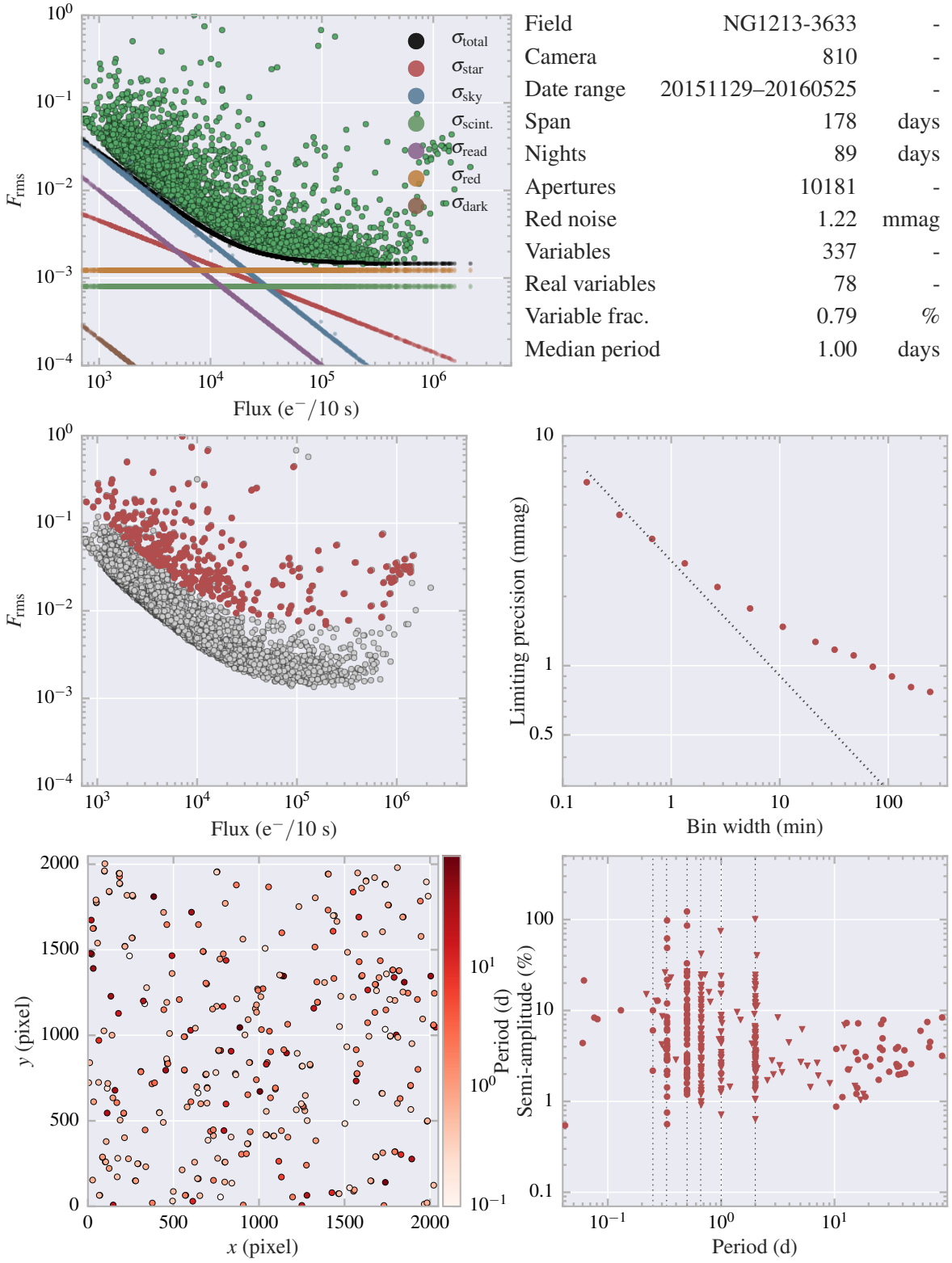


Figure A.8: Field summary for NG0618−6441.



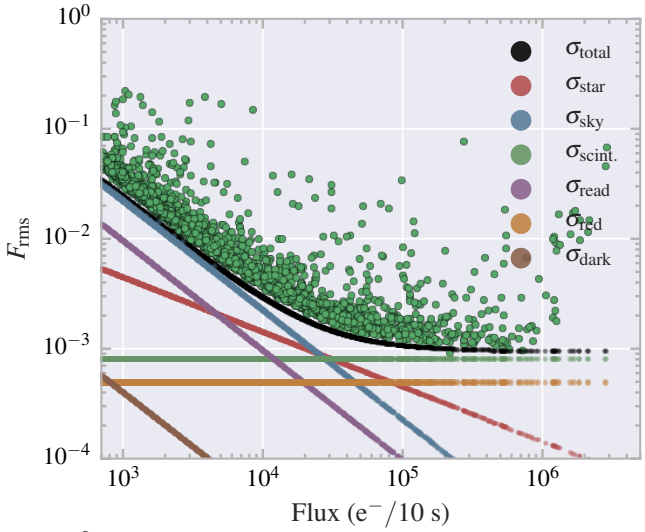
Field	NG1135-2518	-
Camera	809	-
Date range	20151127–20160515	-
Span	170	days
Nights	99	days
Apertures	6870	-
Red noise	0.82	mmag
Variables	232	-
Real variables	107	-
Variable frac.	1.59	%
Median period	0.67	days

Figure A.9: Field summary for NG1135−2518.

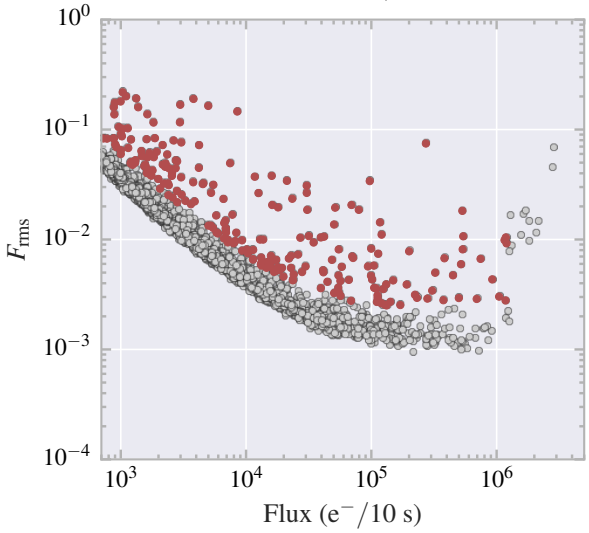


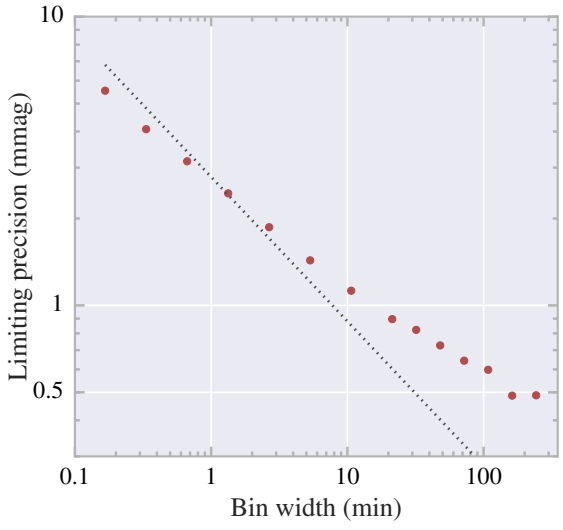
Field	NG1213-3633	-
Camera	810	-
Date range	20151129–20160525	-
Span	178	days
Nights	89	days
Apertures	10181	-
Red noise	1.22	mmag
Variables	337	-
Real variables	78	-
Variable frac.	0.79	%
Median period	1.00	days

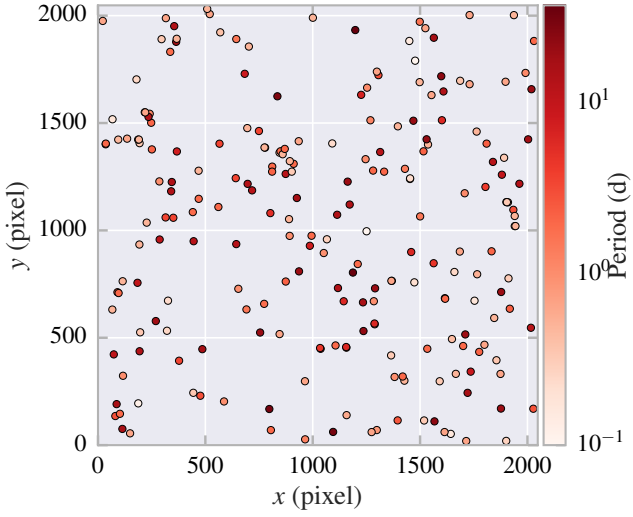
Figure A.10: Field summary for NG1213−3633.



Field	NG1253-1941	-
Camera	803	-
Date range	20160418–20160703	-
Span	76	days
Nights	23	days
Apertures	6700	-
Red noise	0.49	mmag
Variables	225	-
Real variables	120	-
Variable frac.	1.84	%
Median period	1.80	days







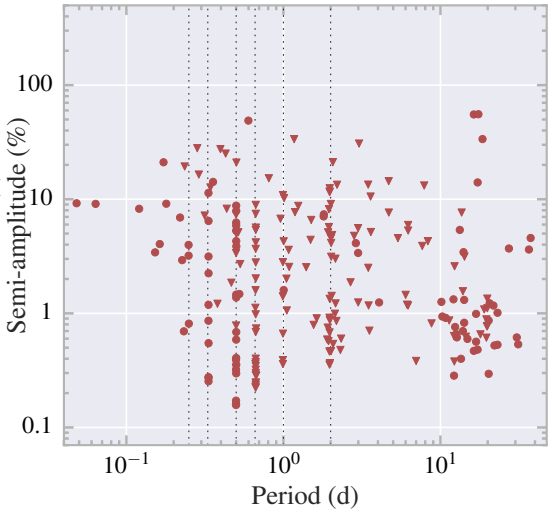
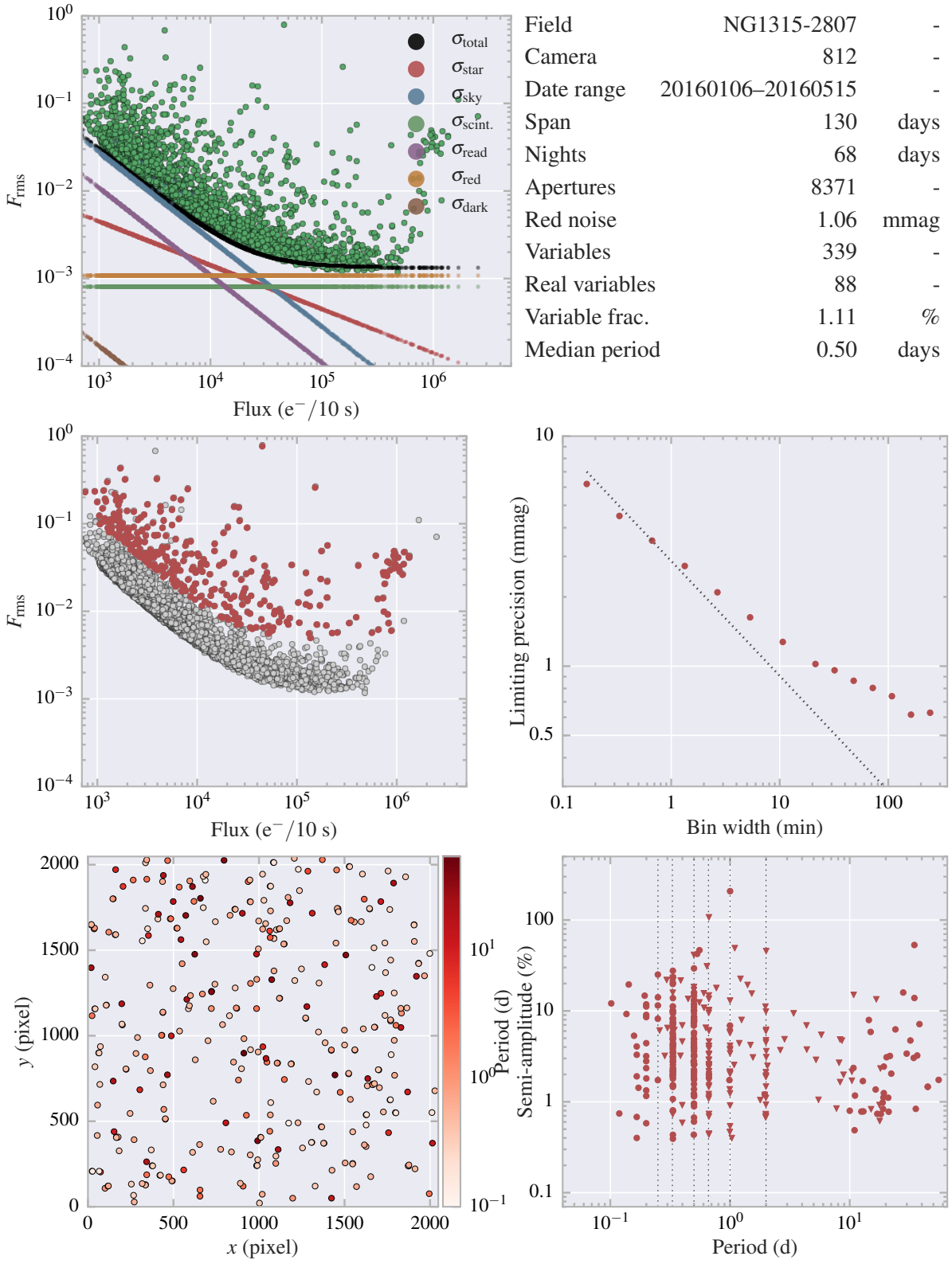
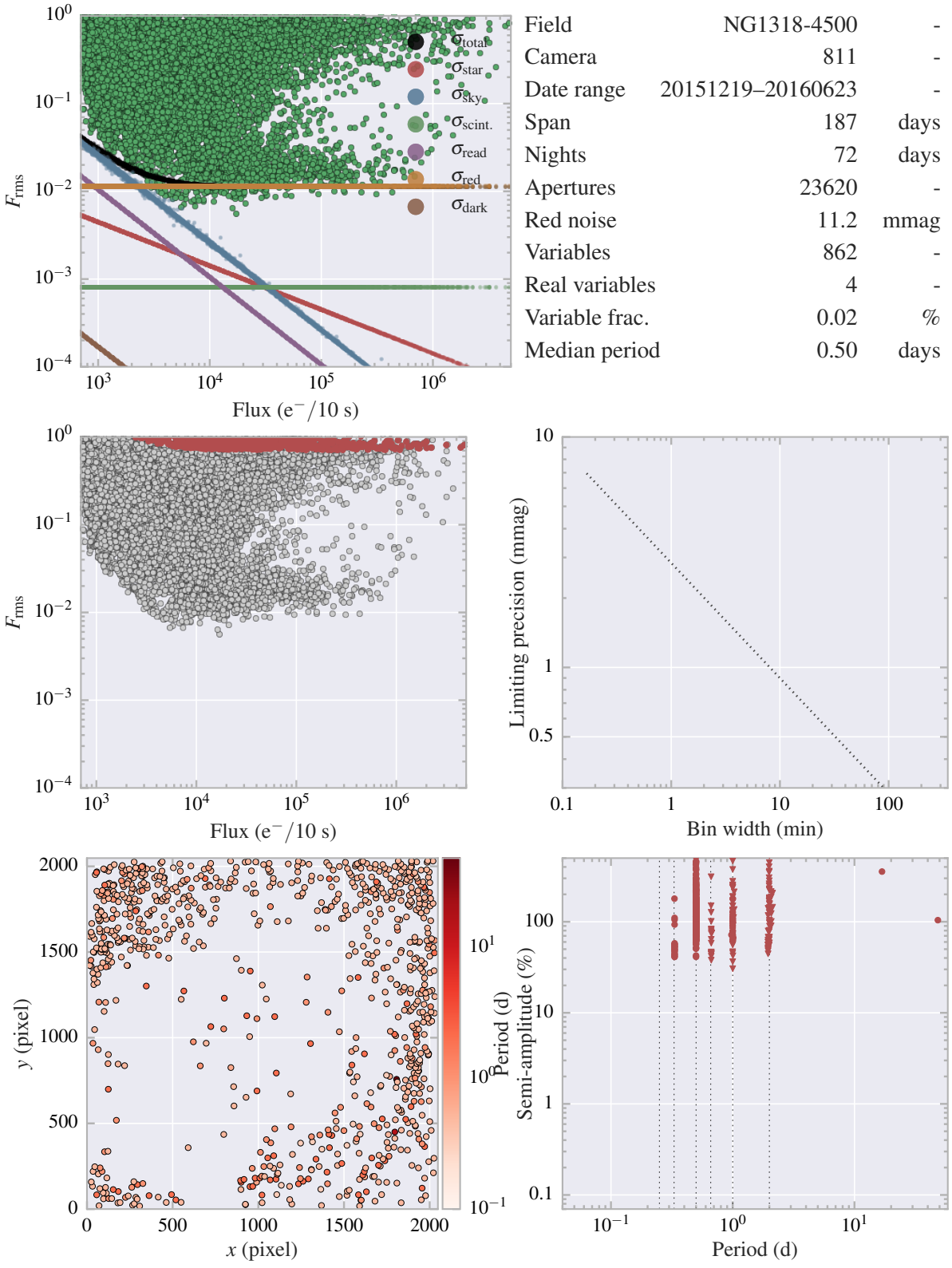


Figure A.11: Field summary for NG1253−1941.



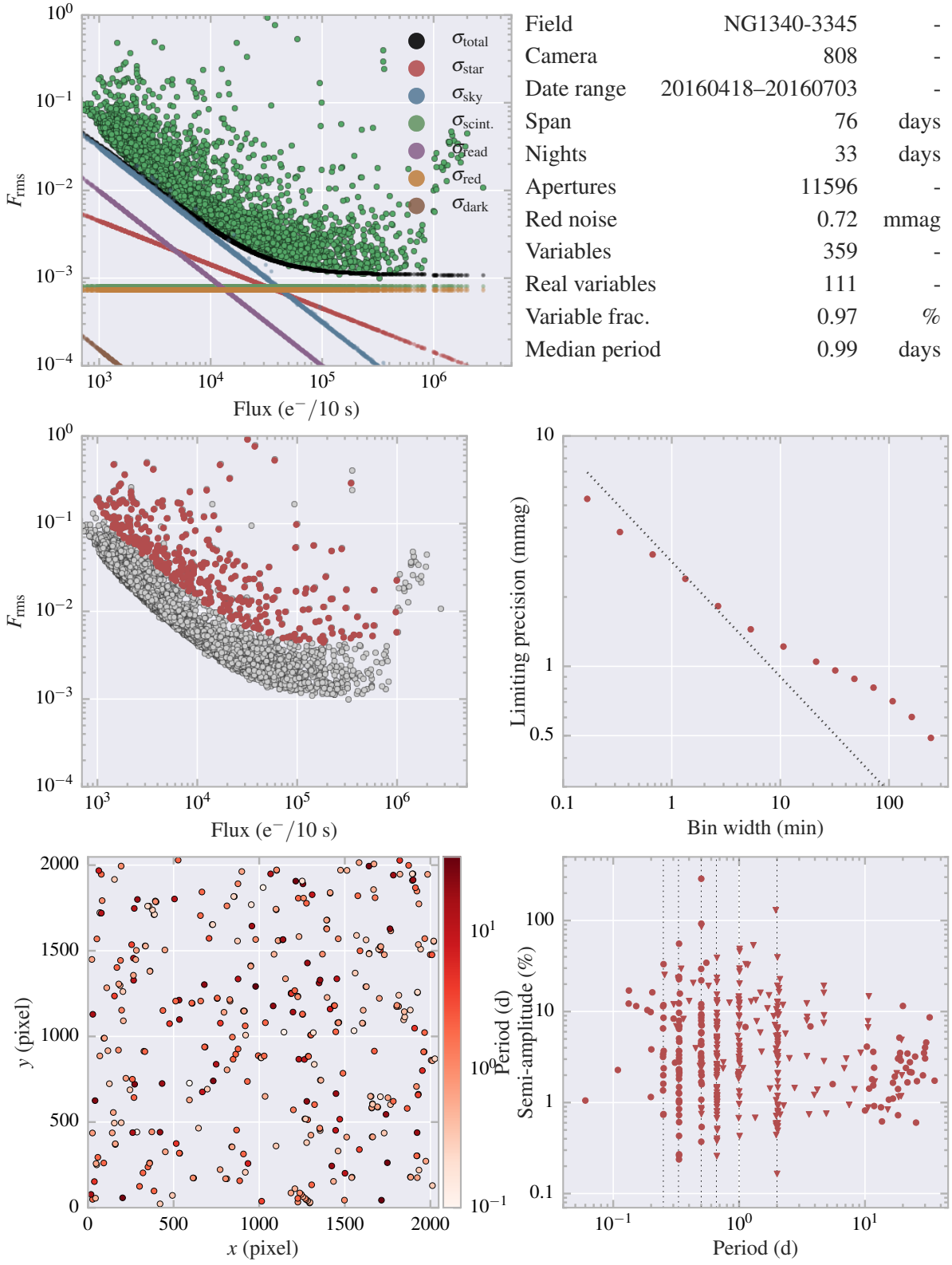
Field	NG1315-2807	-
Camera	812	-
Date range	20160106–20160515	-
Span	130	days
Nights	68	days
Apertures	8371	-
Red noise	1.06	mmag
Variables	339	-
Real variables	88	-
Variable frac.	1.11	%
Median period	0.50	days

Figure A.12: Field summary for NG1315−2807.



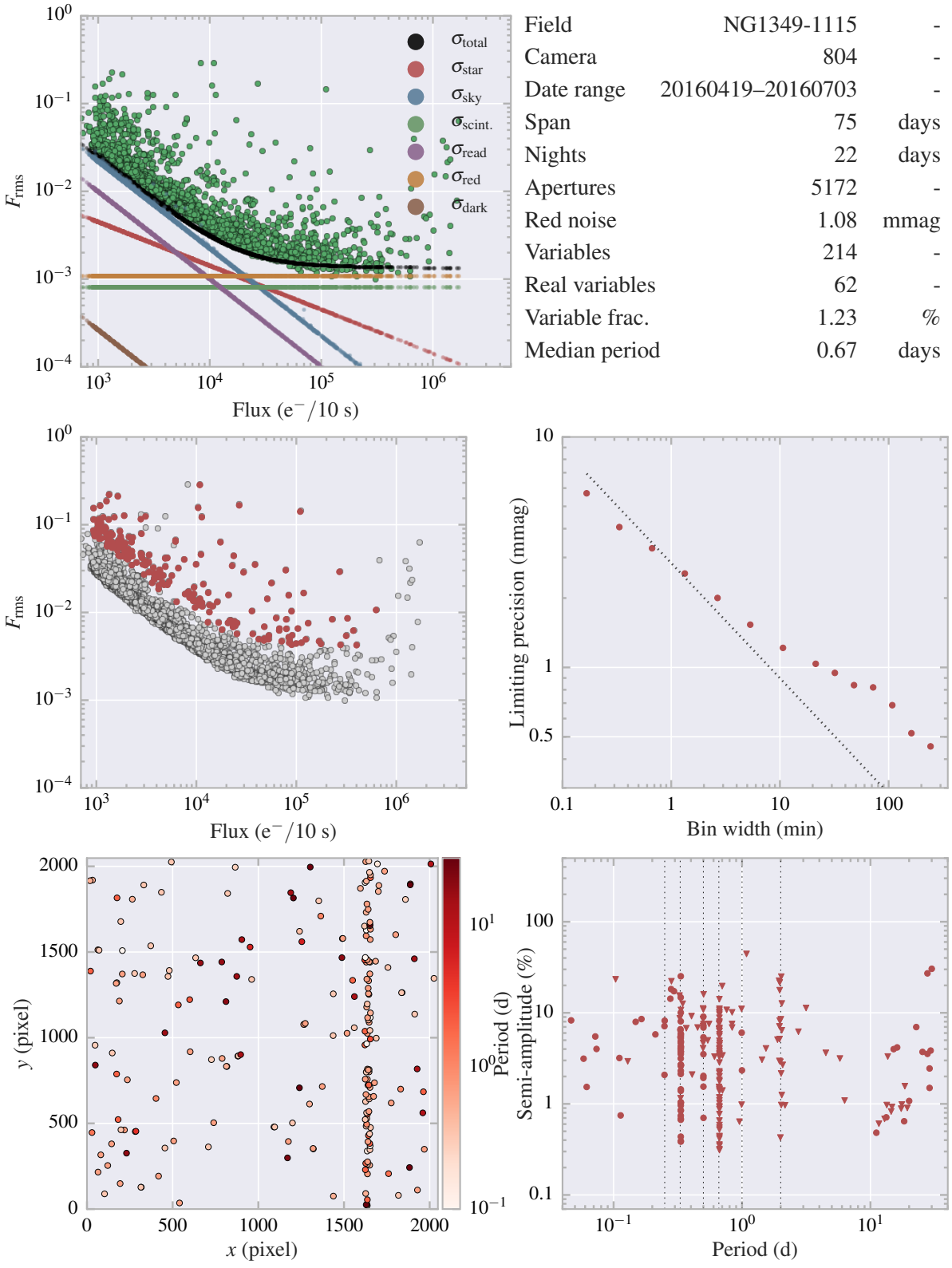
Field	NG1318-4500	-
Camera	811	-
Date range	20151219–20160623	-
Span	187	days
Nights	72	days
Apertures	23620	-
Red noise	11.2	mmag
Variables	862	-
Real variables	4	-
Variable frac.	0.02	%
Median period	0.50	days

Figure A.13: Field summary for NG1318−4500.



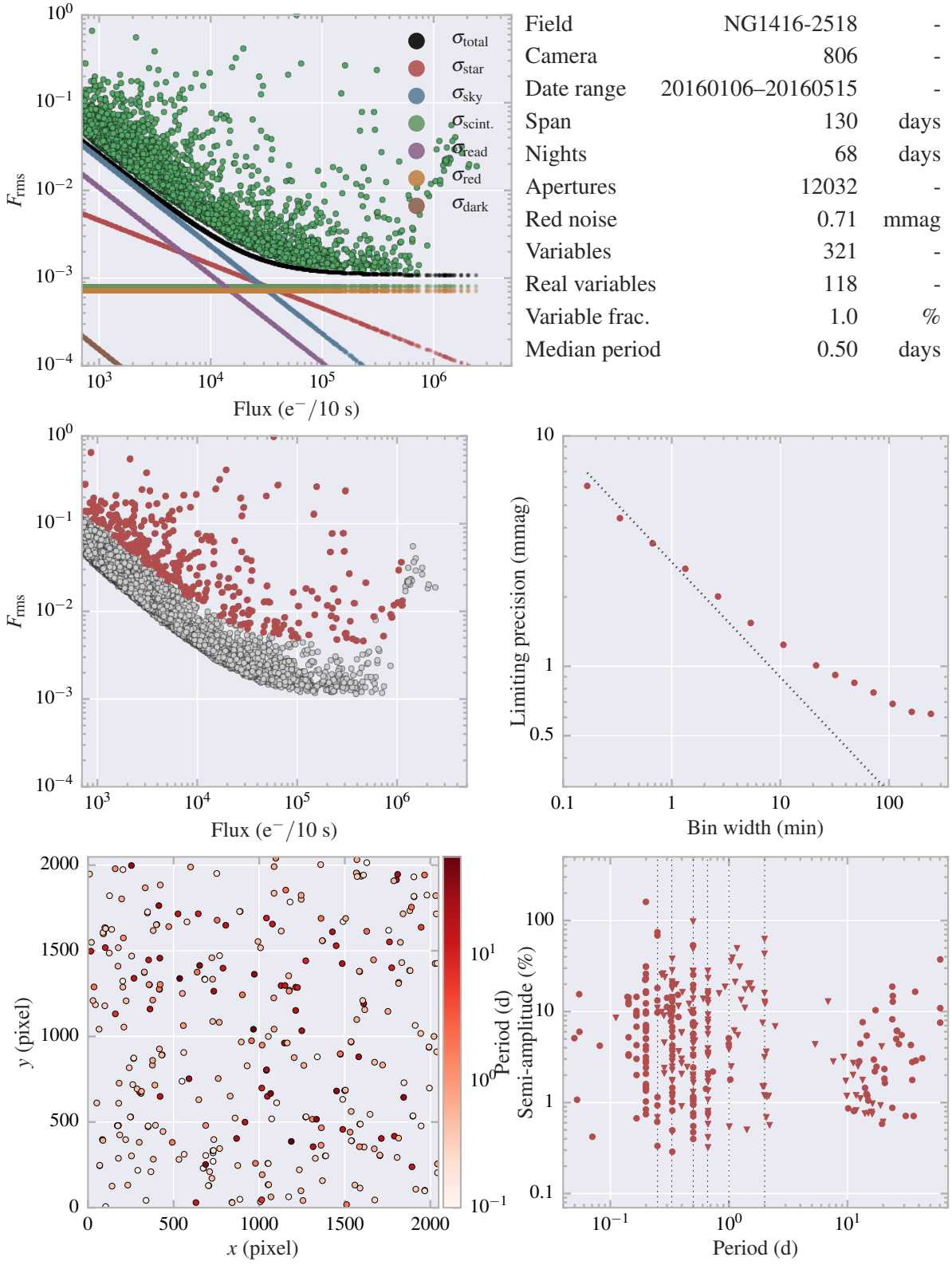
Field	NG1340-3345	-
Camera	808	-
Date range	20160418–20160703	-
Span	76	days
Nights	33	days
Apertures	11596	-
Red noise	0.72	mmag
Variables	359	-
Real variables	111	-
Variable frac.	0.97	%
Median period	0.99	days

Figure A.14: Field summary for NG1340−3345.



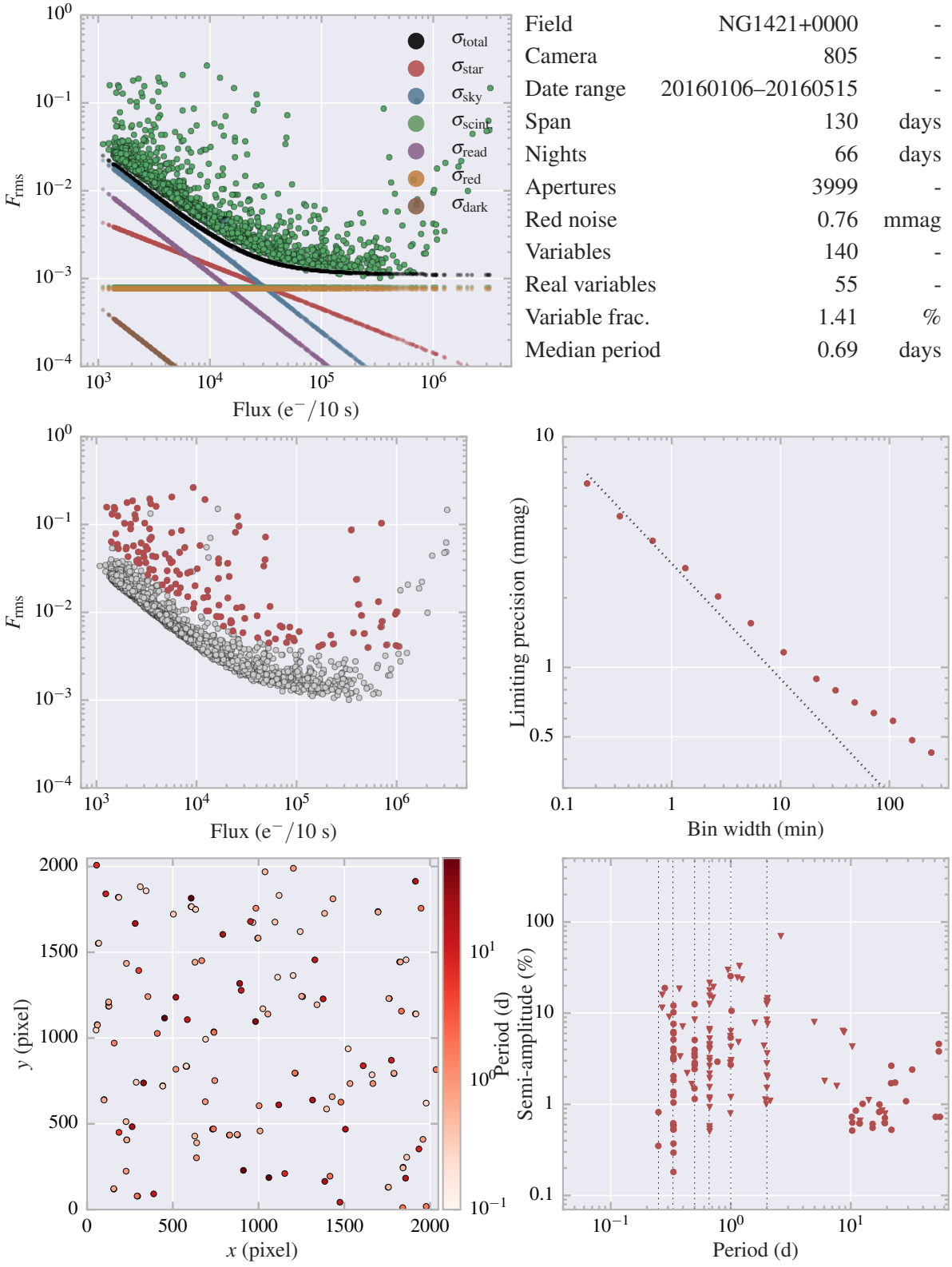
Field	NG1349-1115	-
Camera	804	-
Date range	20160419–20160703	-
Span	75	days
Nights	22	days
Apertures	5172	-
Red noise	1.08	mmag
Variables	214	-
Real variables	62	-
Variable frac.	1.23	%
Median period	0.67	days

Figure A.15: Field summary for NG1349−1115.



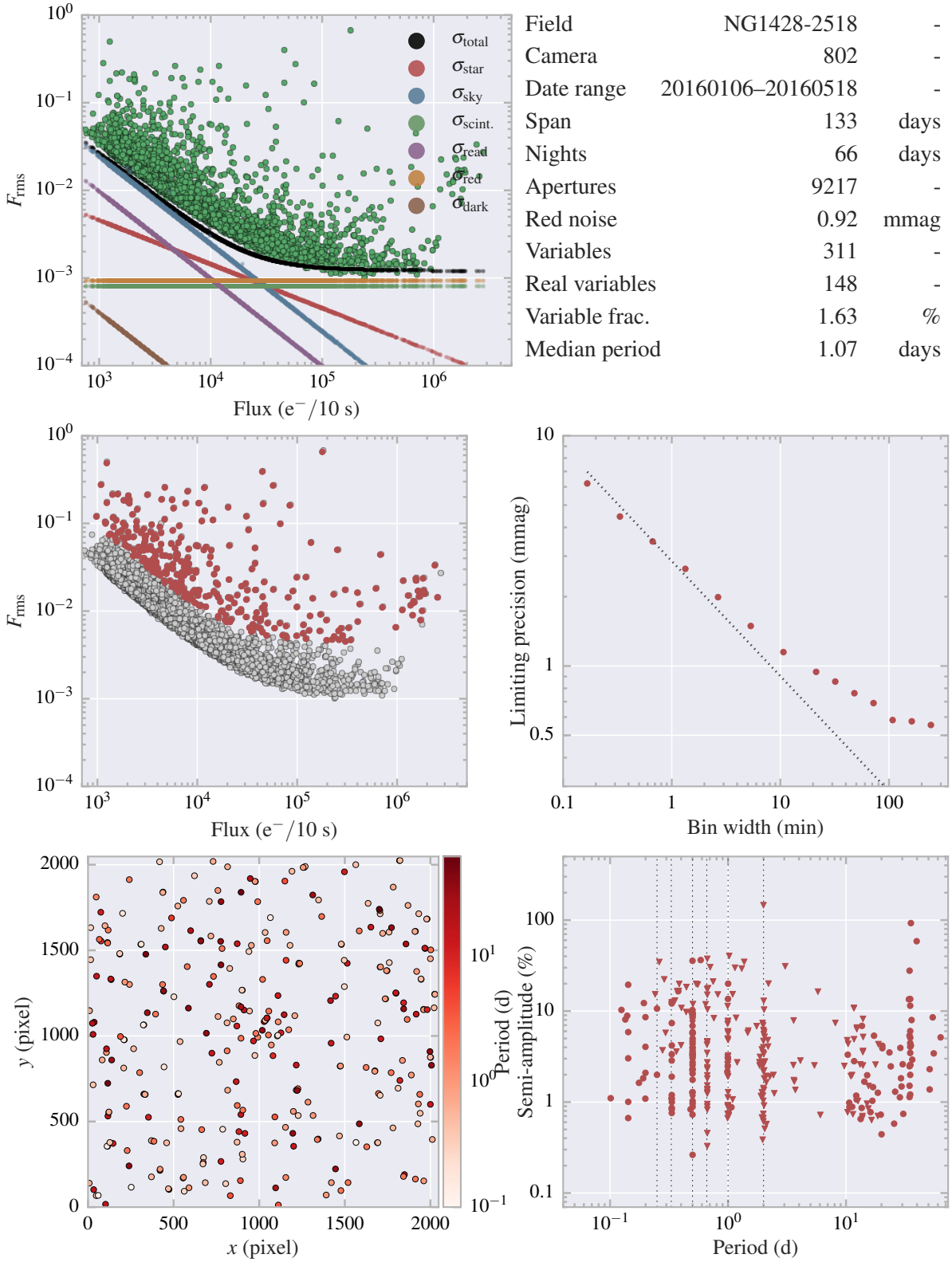
Field	NG1416-2518	-
Camera	806	-
Date range	20160106–20160515	-
Span	130	days
Nights	68	days
Apertures	12032	-
Red noise	0.71	mmag
Variables	321	-
Real variables	118	-
Variable frac.	1.0	%
Median period	0.50	days

Figure A.16: Field summary for NG1416−2518.



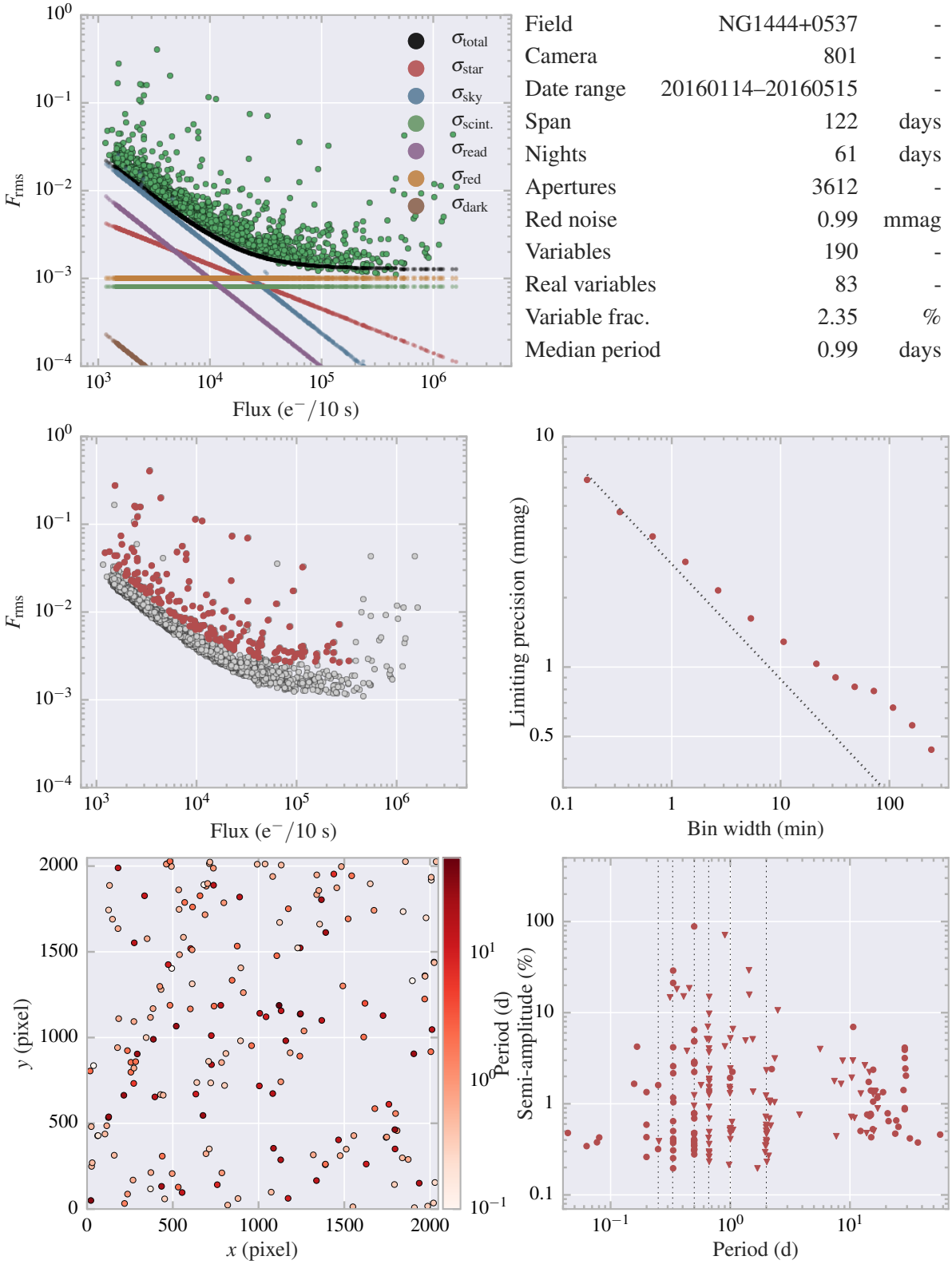
Field	NG1421+0000	-
Camera	805	-
Date range	20160106–20160515	-
Span	130	days
Nights	66	days
Apertures	3999	-
Red noise	0.76	mmag
Variables	140	-
Real variables	55	-
Variable frac.	1.41	%
Median period	0.69	days

Figure A.17: Field summary for NG1421+0000.



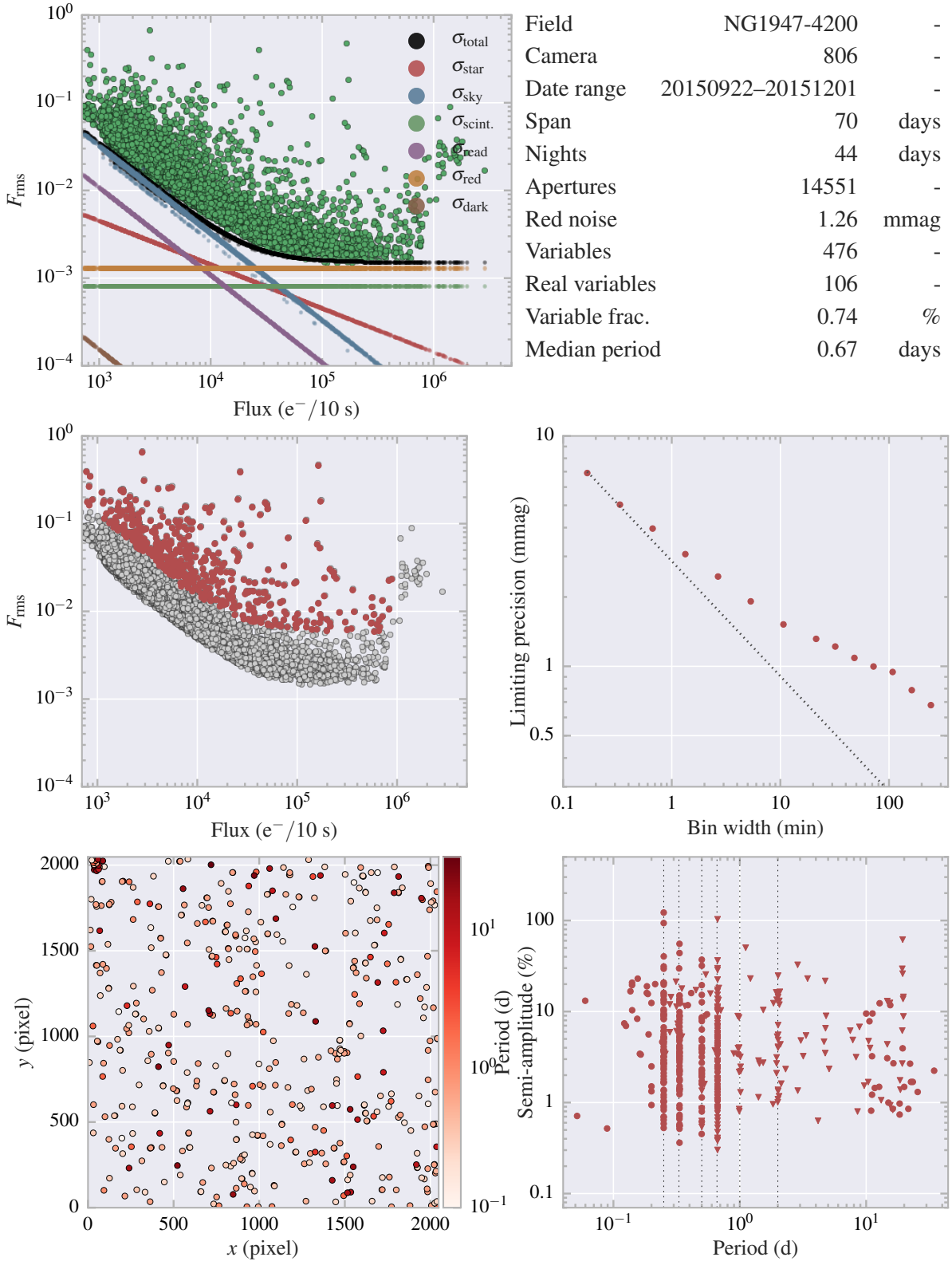
Field	NG1428-2518	-
Camera	802	-
Date range	20160106–20160518	-
Span	133	days
Nights	66	days
Apertures	9217	-
Red noise	0.92	mmag
Variables	311	-
Real variables	148	-
Variable frac.	1.63	%
Median period	1.07	days

Figure A.18: Field summary for NG1428−2518.



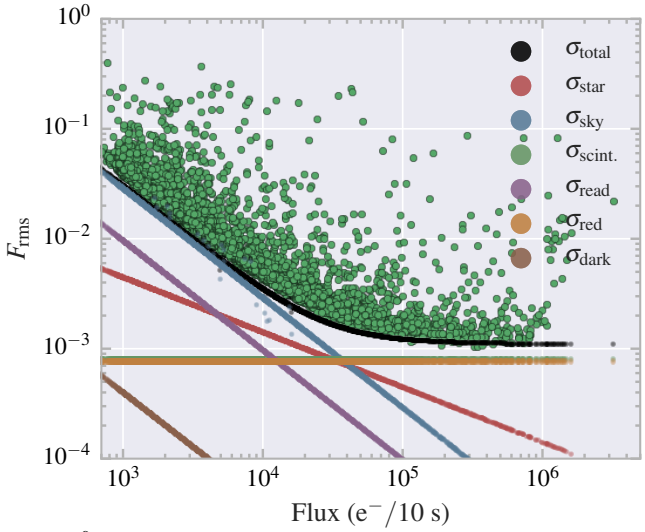
Field	NG1444+0537	-
Camera	801	-
Date range	20160114–20160515	-
Span	122	days
Nights	61	days
Apertures	3612	-
Red noise	0.99	mmag
Variables	190	-
Real variables	83	-
Variable frac.	2.35	%
Median period	0.99	days

Figure A.19: Field summary for NG1444+0537.

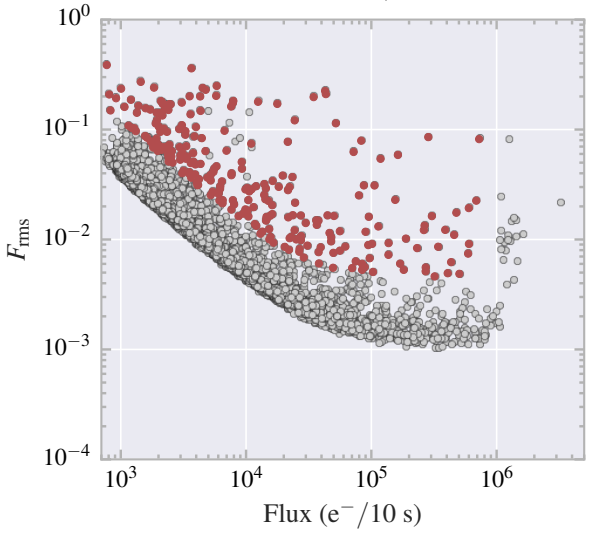


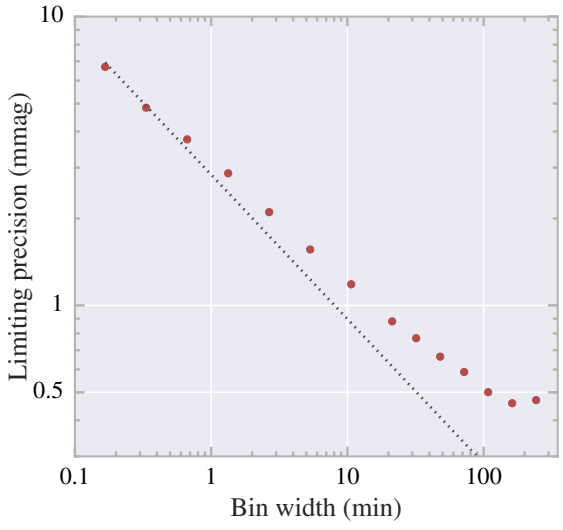
Field	NG1947-4200	-
Camera	806	-
Date range	20150922–20151201	-
Span	70	days
Nights	44	days
Apertures	14551	-
Red noise	1.26	mmag
Variables	476	-
Real variables	106	-
Variable frac.	0.74	%
Median period	0.67	days

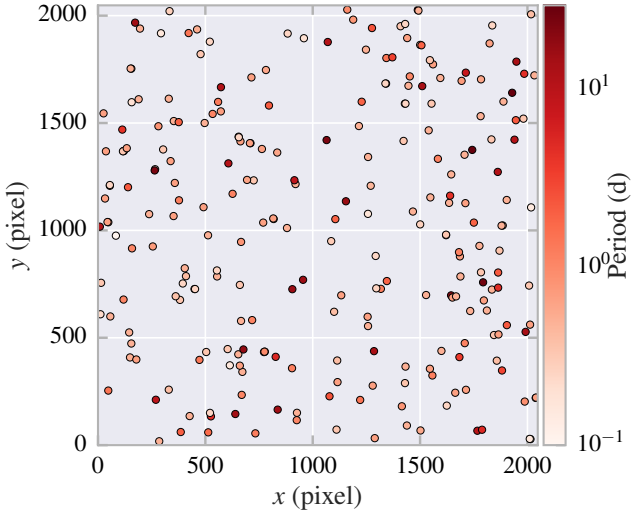
Figure A.20: Field summary for NG1947−4200.



Field	NG2028-2518	-
Camera	802	-
Date range	20150921–20151201	-
Span	71	days
Nights	47	days
Apertures	13952	-
Red noise	0.75	mmag
Variables	252	-
Real variables	103	-
Variable frac.	0.76	%
Median period	0.64	days







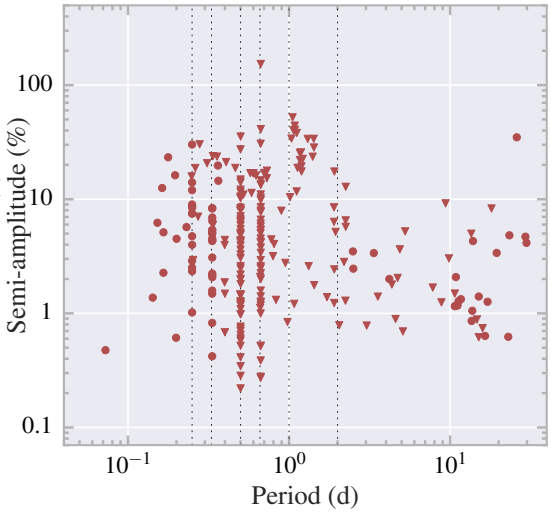
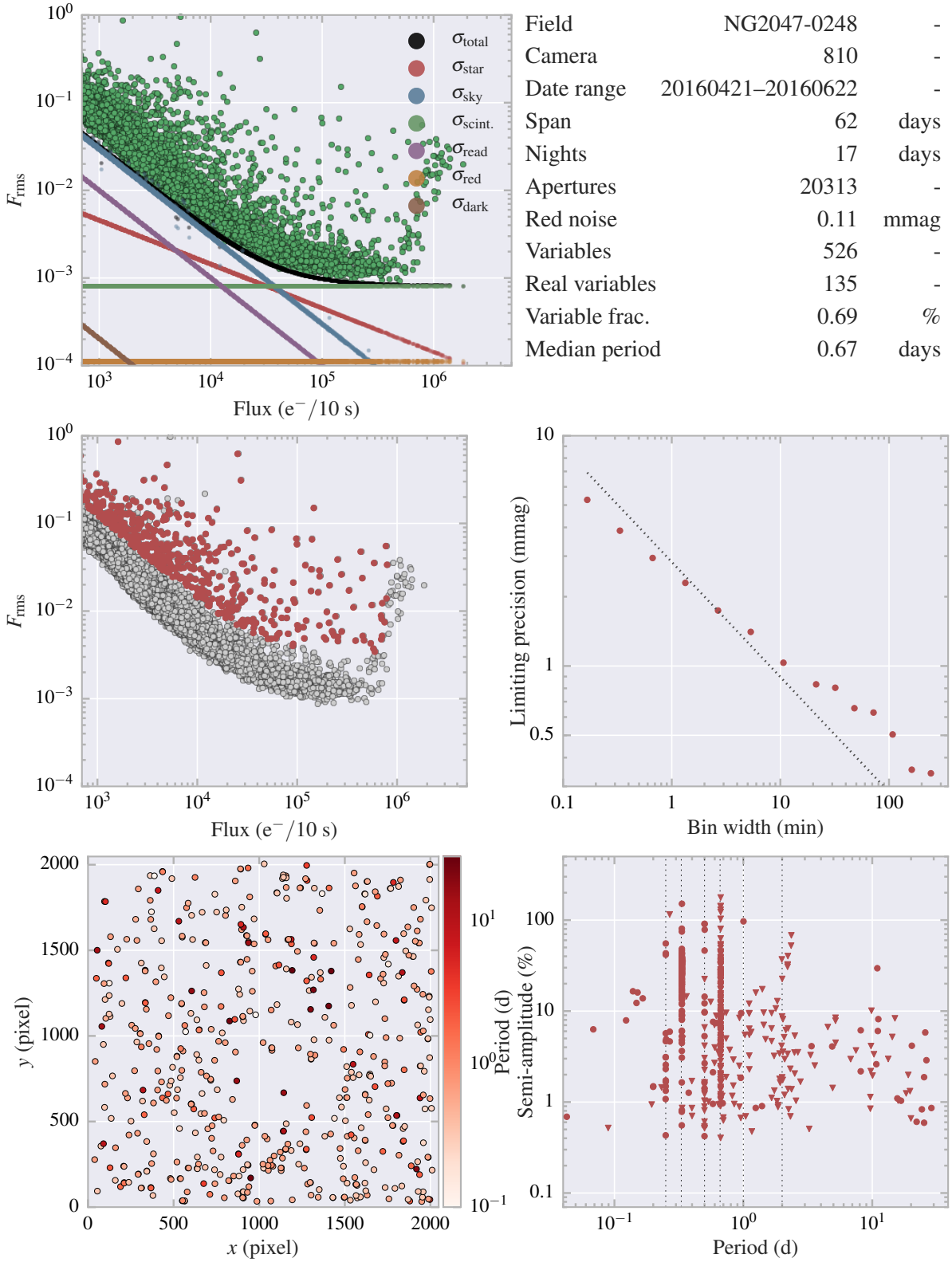
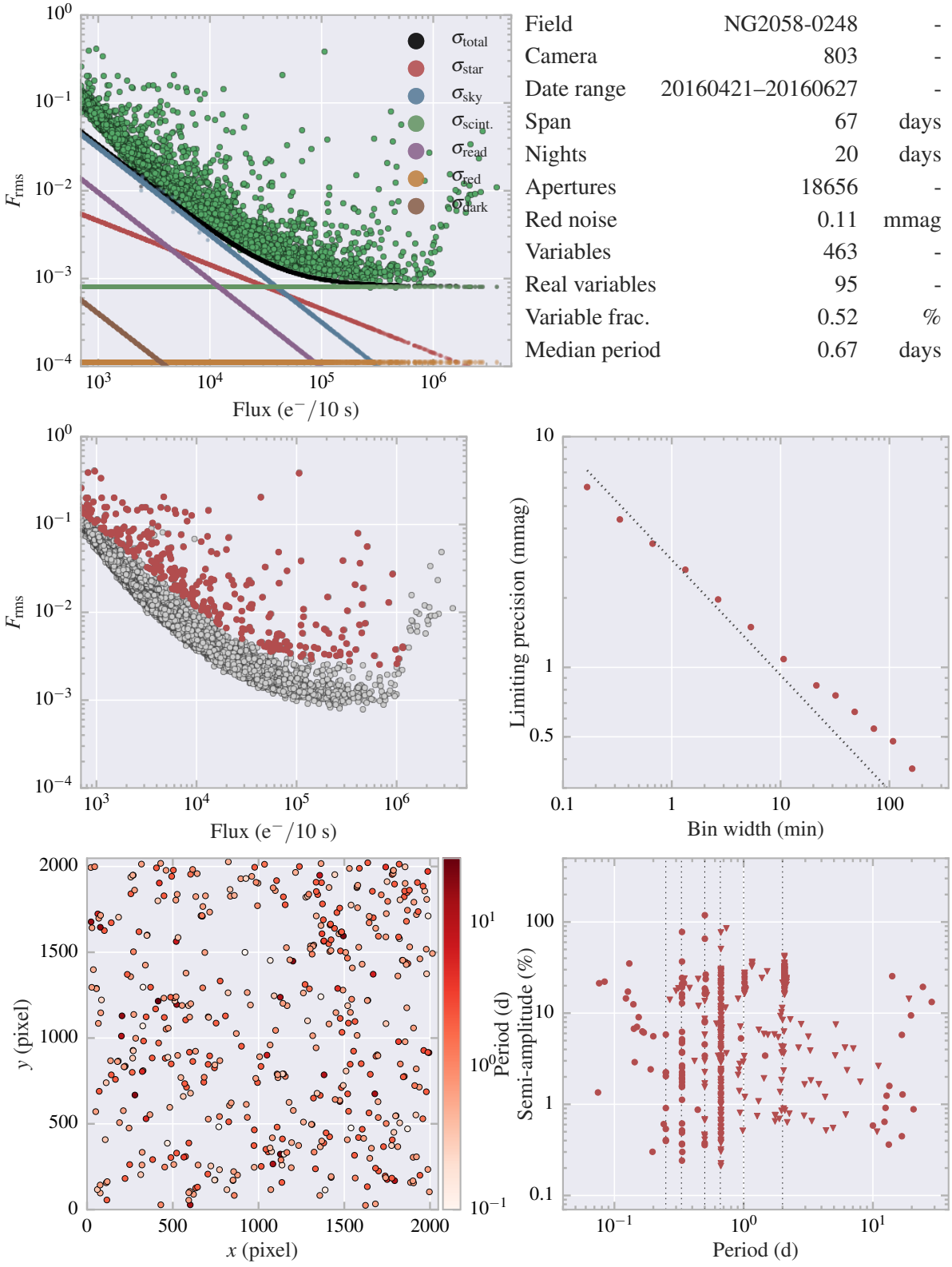


Figure A.21: Field summary for NG2028−2518.



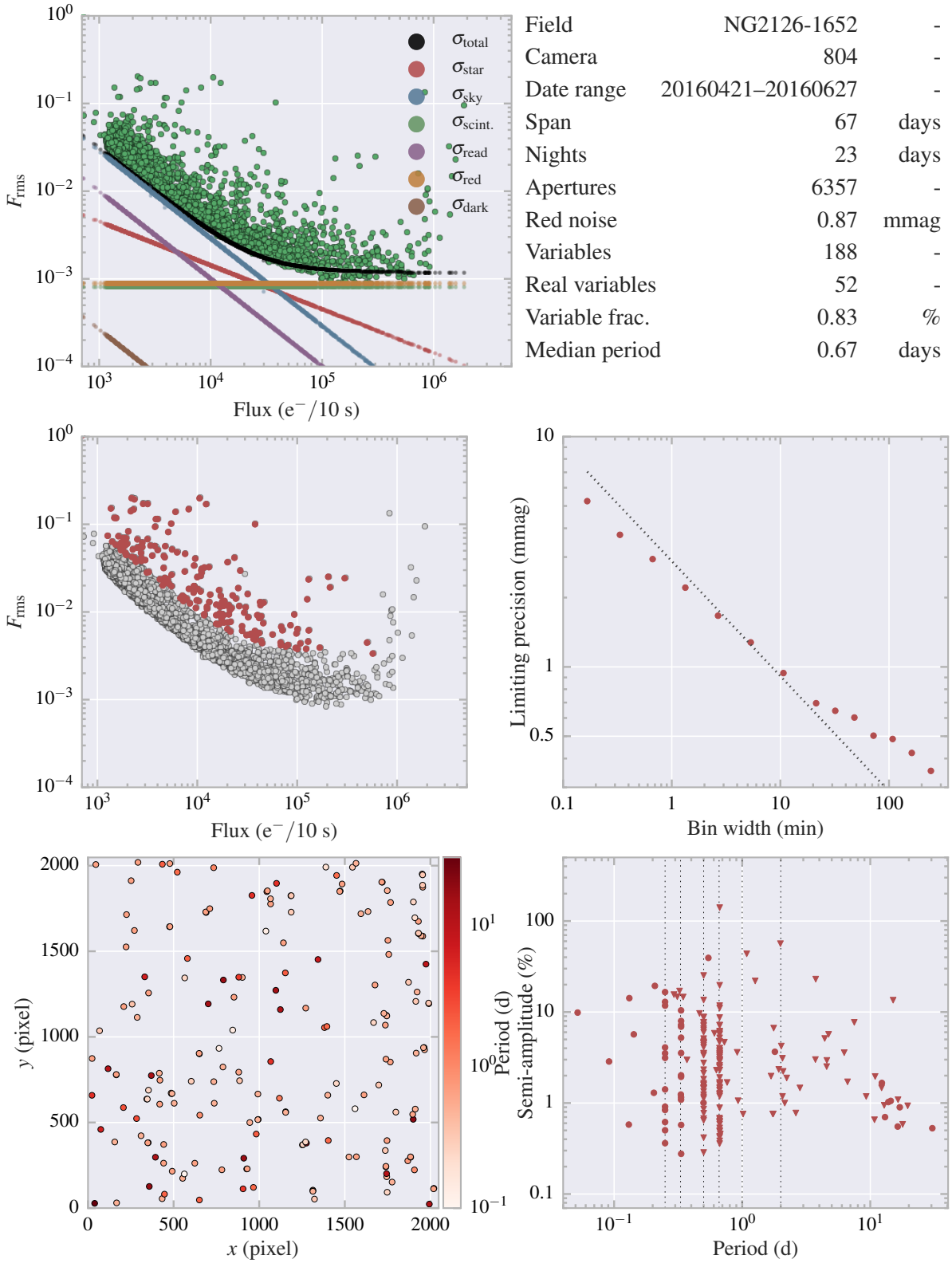
Field	NG2047-0248	-
Camera	810	-
Date range	20160421–20160622	-
Span	62	days
Nights	17	days
Apertures	20313	-
Red noise	0.11	mmag
Variables	526	-
Real variables	135	-
Variable frac.	0.69	%
Median period	0.67	days

Figure A.22: Field summary for NG2047−0248.



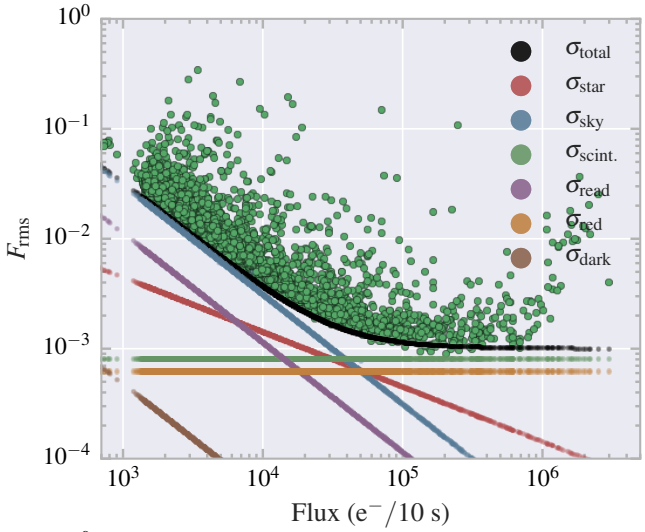
Field	NG2058-0248	-
Camera	803	-
Date range	20160421–20160627	-
Span	67	days
Nights	20	days
Apertures	18656	-
Red noise	0.11	mmag
Variables	463	-
Real variables	95	-
Variable frac.	0.52	%
Median period	0.67	days

Figure A.23: Field summary for NG2058−0248.

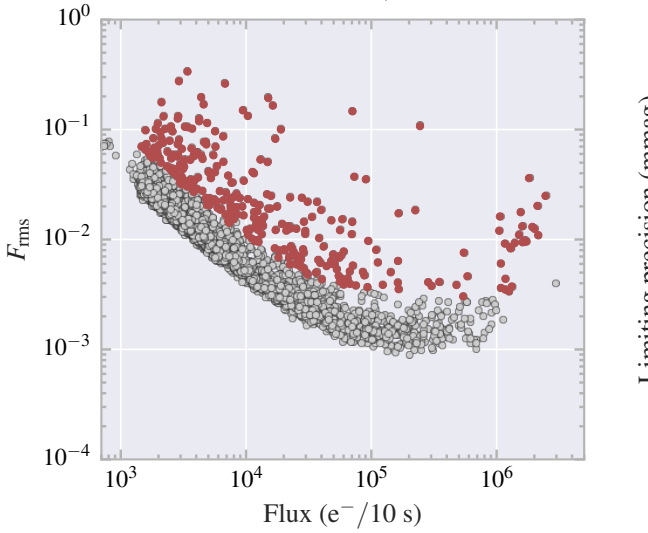


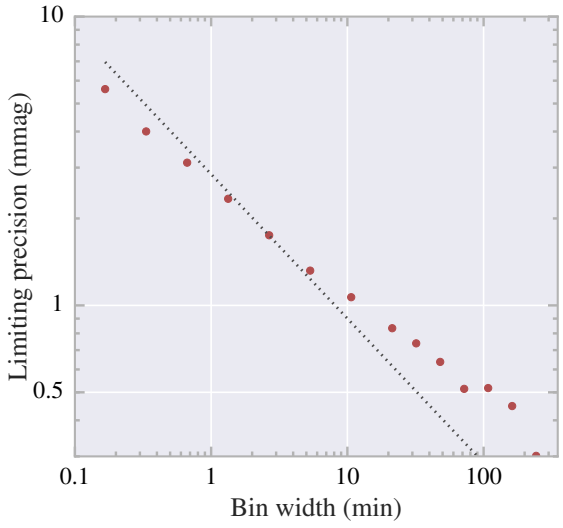
Field	NG2126-1652	-
Camera	804	-
Date range	20160421–20160627	-
Span	67	days
Nights	23	days
Apertures	6357	-
Red noise	0.87	mmag
Variables	188	-
Real variables	52	-
Variable frac.	0.83	%
Median period	0.67	days

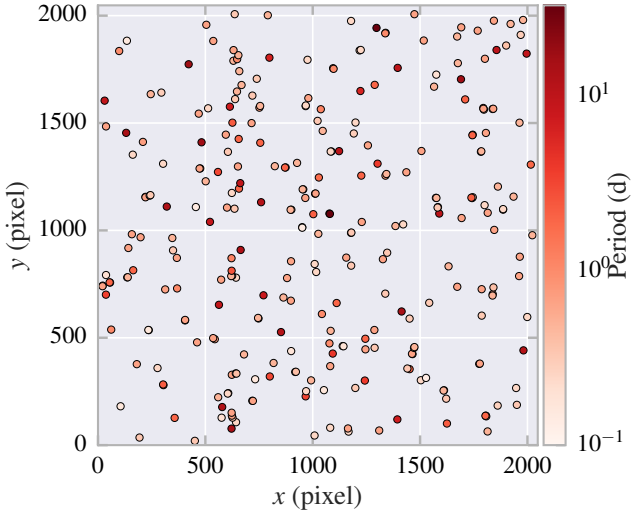
Figure A.24: Field summary for NG2126−1652.



Field	NG2132+0248	-
Camera	805	-
Date range	20160421–20160627	-
Span	67	days
Nights	22	days
Apertures	7516	-
Red noise	0.61	mmag
Variables	293	-
Real variables	64	-
Variable frac.	0.88	%
Median period	0.66	days







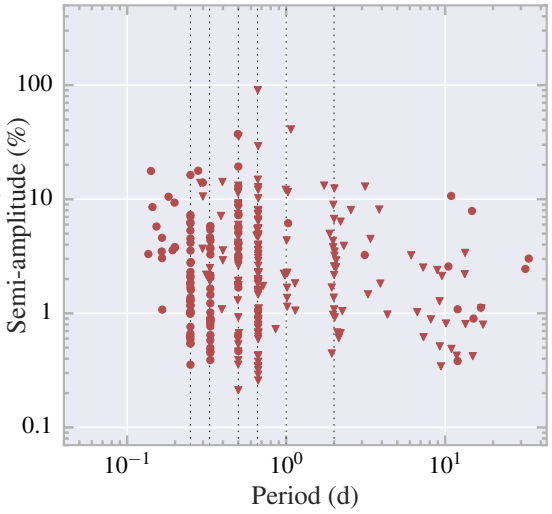
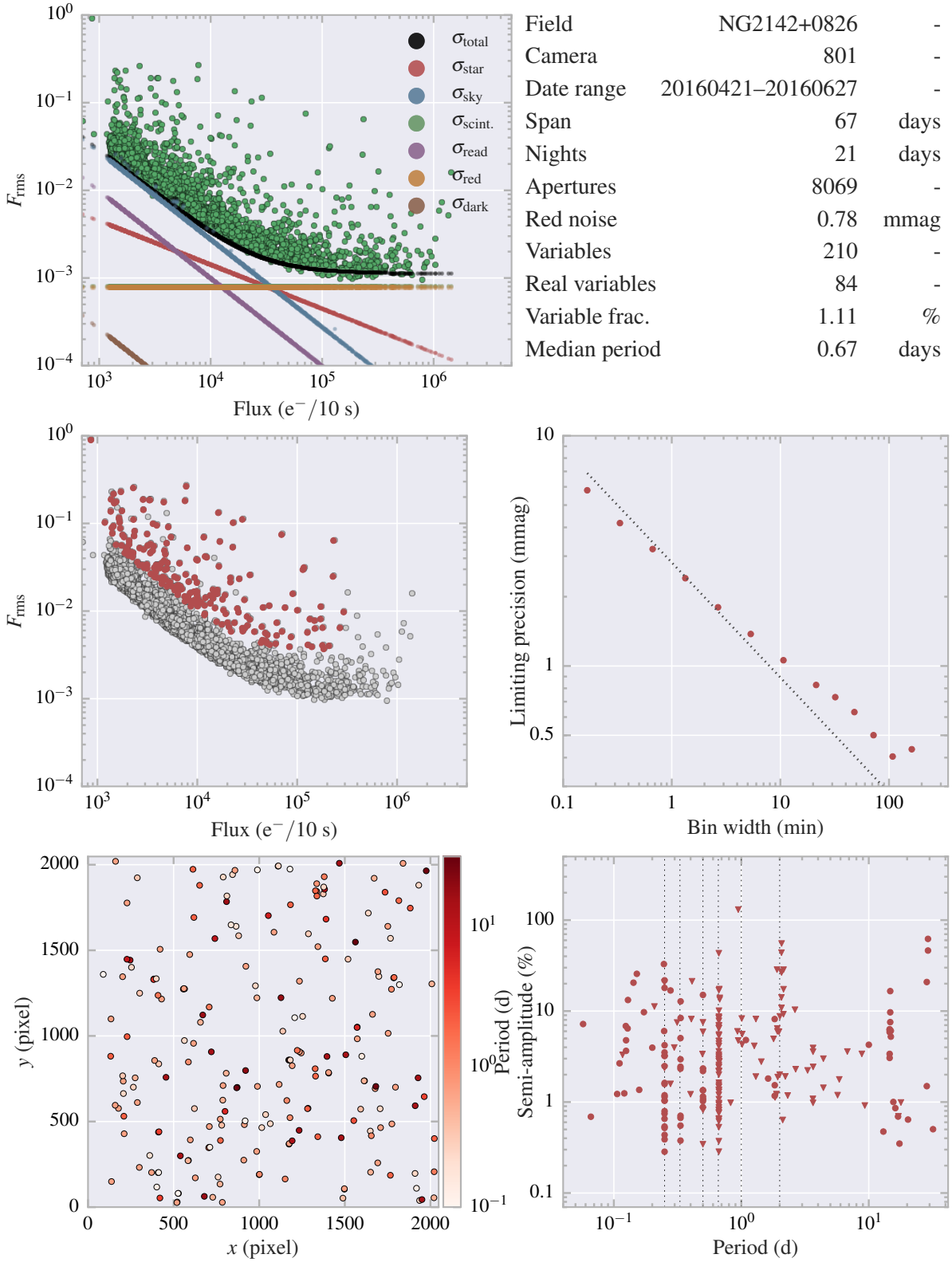
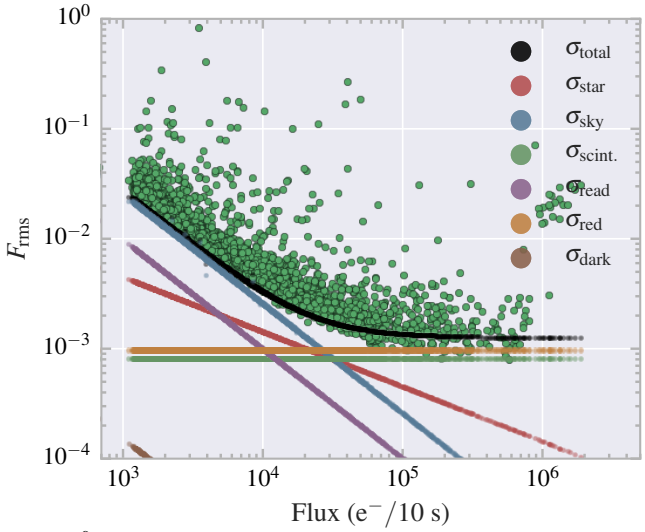


Figure A.25: Field summary for NG2132+0248.

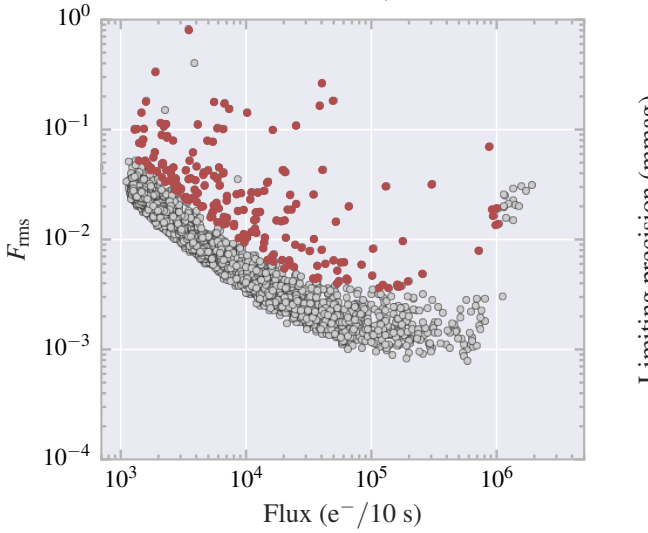


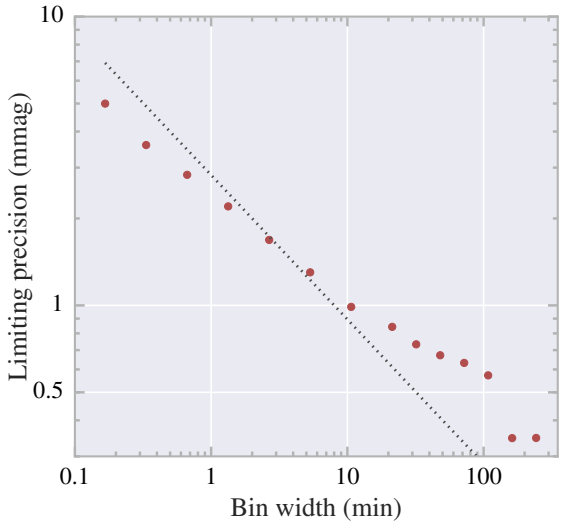
Field	NG2142+0826	-
Camera	801	-
Date range	20160421–20160627	-
Span	67	days
Nights	21	days
Apertures	8069	-
Red noise	0.78	mmag
Variables	210	-
Real variables	84	-
Variable frac.	1.11	%
Median period	0.67	days

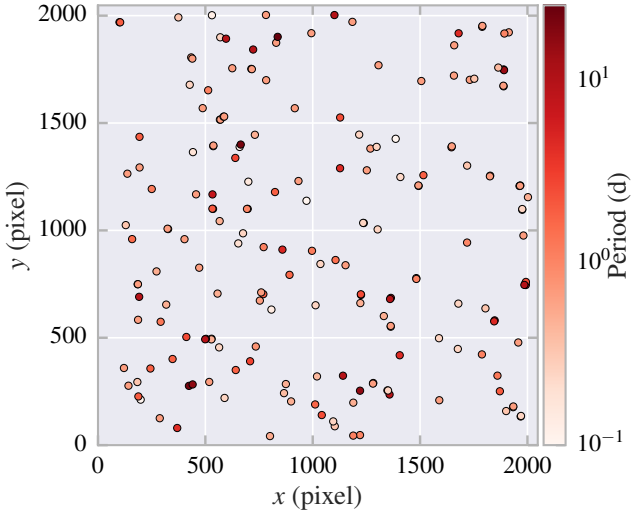
Figure A.26: Field summary for NG2142+0826.



Field	NG2145-3345	-
Camera	808	-
Date range	20160421–20160627	-
Span	67	days
Nights	18	days
Apertures	5324	-
Red noise	0.94	mmag
Variables	192	-
Real variables	68	-
Variable frac.	1.32	%
Median period	0.67	days







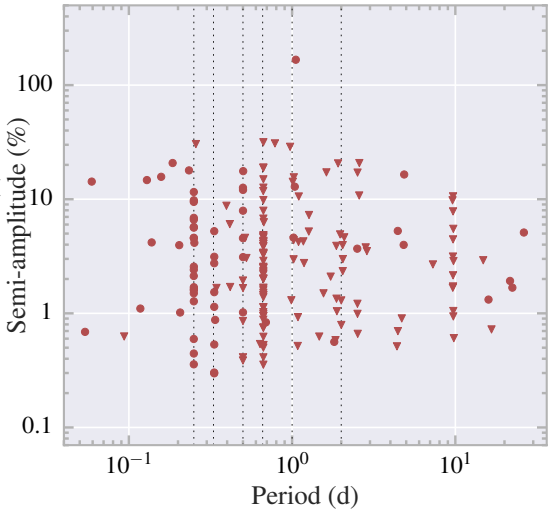
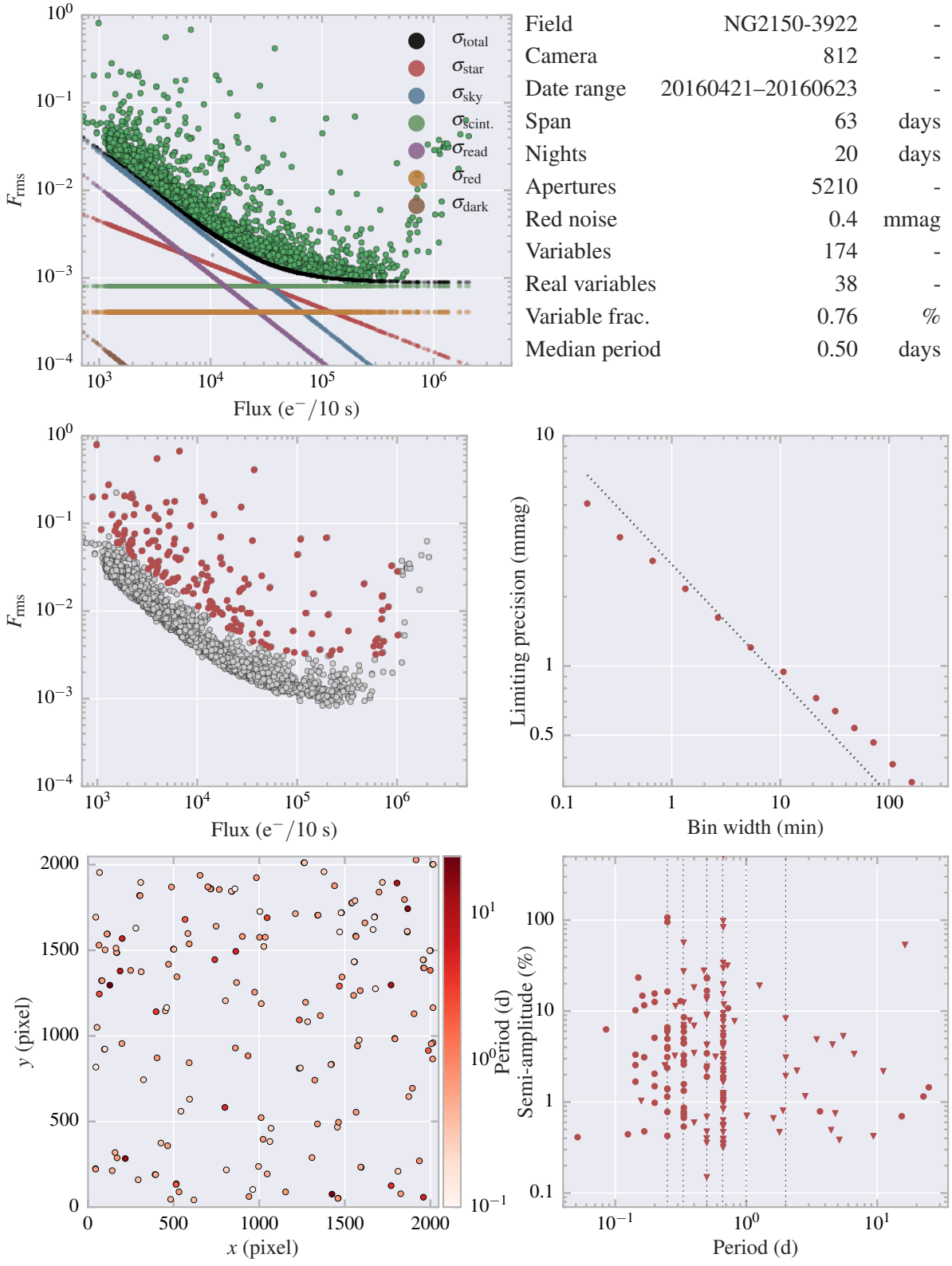
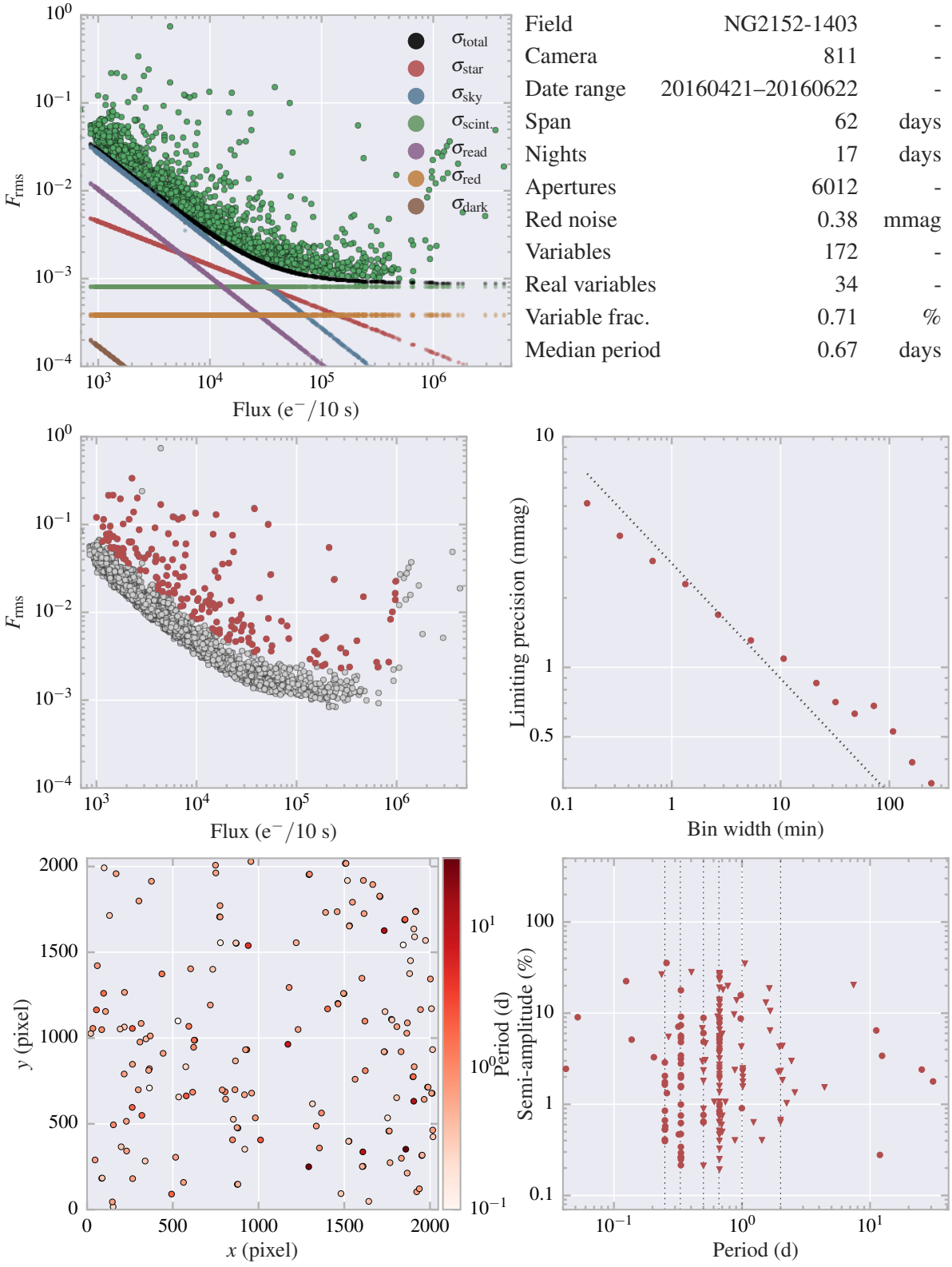


Figure A.27: Field summary for NG2145−3345.



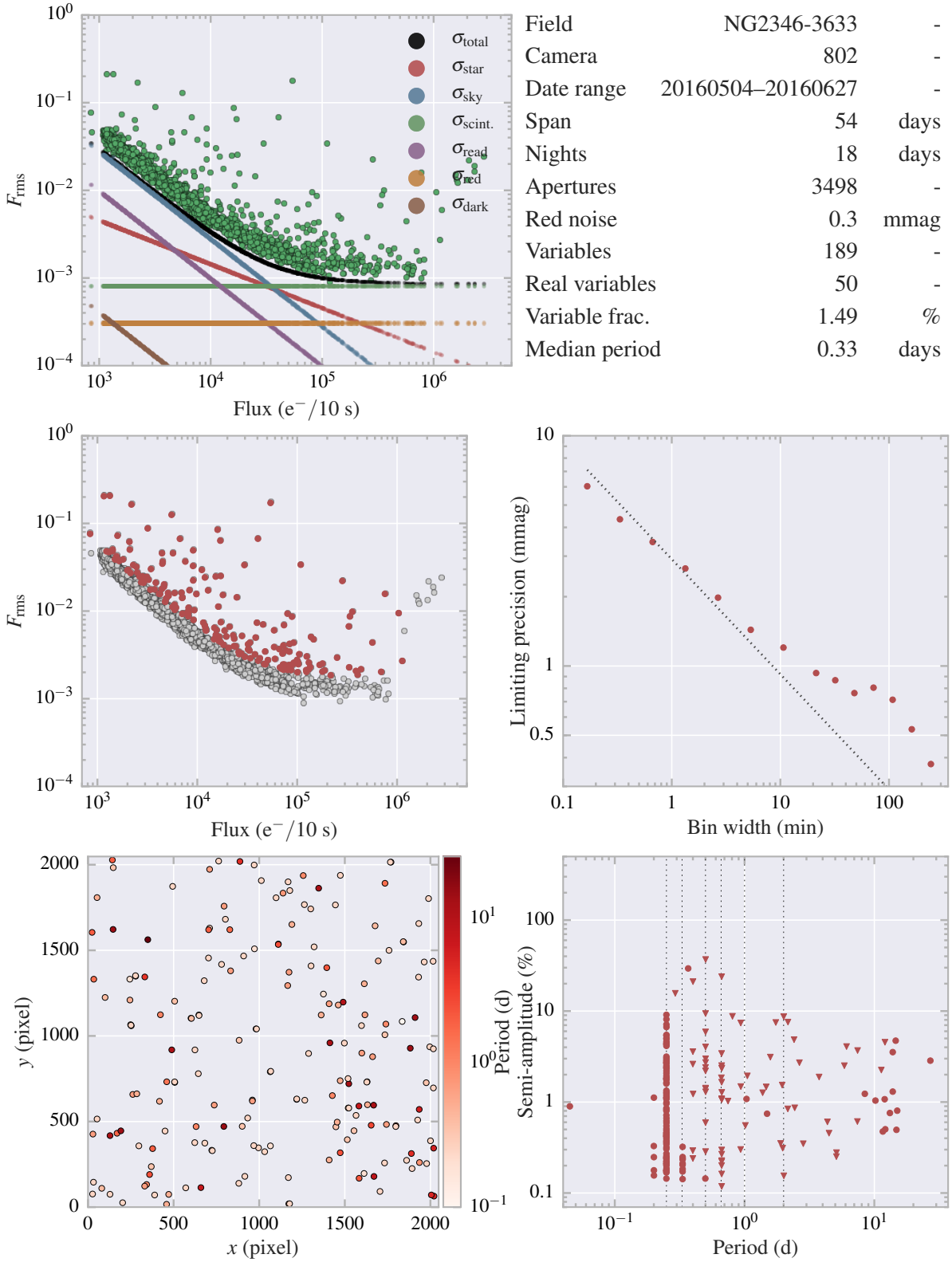
Field	NG2150-3922	-
Camera	812	-
Date range	20160421–20160623	-
Span	63	days
Nights	20	days
Apertures	5210	-
Red noise	0.4	mmag
Variables	174	-
Real variables	38	-
Variable frac.	0.76	%
Median period	0.50	days

Figure A.28: Field summary for NG2150−3922.



Field	NG2152-1403	-
Camera	811	-
Date range	20160421–20160622	-
Span	62	days
Nights	17	days
Apertures	6012	-
Red noise	0.38	mmag
Variables	172	-
Real variables	34	-
Variable frac.	0.71	%
Median period	0.67	days

Figure A.29: Field summary for NG2152−1403.



Field	NG2346-3633	-
Camera	802	-
Date range	20160504–20160627	-
Span	54	days
Nights	18	days
Apertures	3498	-
Red noise	0.3	mmag
Variables	189	-
Real variables	50	-
Variable frac.	1.49	%
Median period	0.33	days

Figure A.30: Field summary for NG2346−3633.