Transit shapes and self-organizing maps as a tool for ranking planetary candidates: application to Kepler and K2

D. J. Armstrong, D. Pollacco, A. Santerne

▶ **To cite this version:**

D. J. Armstrong, D. Pollacco, A. Santerne. Transit shapes and self-organizing maps as a tool for ranking planetary candidates: application to Kepler and K2. Monthly Notices of the Royal Astronomical Society, 2017, 465, pp.2634-2642. 10.1093/mnras/stw2881 . insu-03666207

HAL Id: insu-03666207

https://insu.hal.science/insu-03666207

Submitted on 12 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **465,** 2634–2642 (2017)
Advance Access publication 2016 November 8

doi:10.1093/mnras/stw2881

Transit shapes and self-organizing maps as a tool for ranking planetary candidates: application to *Kepler* and *K2*

D. J. Armstrong,[1,2]★ D. Pollacco[1] and A. Santerne[3,4]

[1]*Department of Physics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK*
[2]*ARC, School of Mathematics and Physics, Queen's University Belfast, University Road, Belfast BT7 1NN, UK*
[3]*Aix Marseille Université, CNRS, Laboratoire d'Astrophysique de Marseille UMR 7326, F-13388 Marseille, France*
[4]*Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, P-4150-762 Porto, Portugal*

Accepted 2016 November 4. Received 2016 November 4; in original form 2016 August 16

ABSTRACT

A crucial step in planet hunting surveys is to select the best candidates for follow-up observations, given limited telescope resources. This is often performed by human 'eyeballing', a time consuming and statistically awkward process. Here, we present a new, fast machine learning technique to separate true planet signals from astrophysical false positives. We use self-organizing maps (SOMs) to study the transit shapes of *Kepler* and *K2* known and candidate planets. We find that SOMs are capable of distinguishing known planets from known false positives with a success rate of 87.0 per cent, using the transit shape alone. Furthermore, they do not require any candidate to be dispositioned prior to use, meaning that they can be used early in a mission's lifetime. A method for classifying candidates using a SOM is developed, and applied to previously unclassified members of the *Kepler* Objects of Interest (KOI) list as well as candidates from the *K2* mission. The method is extremely fast, taking minutes to run the entire KOI list on a typical laptop. We make PYTHON code for performing classifications publicly available, using either new SOMs or those created in this work. The SOM technique represents a novel method for ranking planetary candidate lists, and can be used both alone or as part of a larger autovetting code.

Key words: methods: data analysis – methods: miscellaneous – methods: statistical – planets and satellites: detection – planets and satellites: general – binaries: eclipsing.

1 INTRODUCTION

Transit surveys both from the ground and space have been the most successful method of discovering planets to date. Instruments such as SuperWASP (Pollacco et al. 2006), HAT/HATnet (Bakos et al. 2004), KELT (Siverd et al. 2012), *Kepler* (Borucki et al. 2010), *K2* (Howell et al. 2014) and *CoRoT* (Auvergne et al. 2009) have found thousands of transiting exoplanets with a wide range of parameters. The light curves produced by these instruments are searched for planets using techniques such as the BLS algorithm (Kovacs, Zucker & Mazeh 2002). Lists of planetary candidates are produced, with some selected for further follow-up observations. While such lists contain many true planetary signals, they also contain instrumental signatures and astrophysical false positives such as contaminating eclipsing binaries (e.g. Almenara et al. 2009; Santerne et al. 2012, 2016).

The process of selecting the best and most likely real candidates to progress to further observations is a difficult one. Typically human inspection is used to select the best candidates (e.g. Pope, Parviainen & Aigrain 2016), a process which can be both time consuming and subject to biases. Some recent methods have been developed to address this problem (McCauliff et al. 2015; Coughlin et al. 2016), and we aim to present an enhancement to these here. We introduce a novel technique designed to separate planetary signals from false positives using the shape of the transit signal, utilizing self-organizing maps (SOMs; Kohonen 1982, 1990; Brett, West & Wheatley 2004; Armstrong et al. 2016). Both investigations of the transit shape (e.g. Thompson et al. 2015) and SOMs have been used in the astrophysical literature before, but not as yet in combination. Here, we apply SOMs to space-based data from the *Kepler* and *K2* missions, as these data are both public and provide a large number of known planets for testing. In the future, we aim to explore the applicability of the technique to ground-based surveys.

SOMs are a machine learning technique first introduced by Kohonen (1982). They have been used in astronomy for estimating galaxy photometric redshifts (Carrasco Kind & Brunner 2014), identifying variable stars (Brett et al. 2004; Armstrong et al. 2016) and investigating active galactic nuclei (Torniainen et al. 2008). Machine learning in general is a promising area only beginning to find application to the exoplanet field. Recent uses include the automatic selection of candidates using random forests (McCauliff et al. 2015), and the

★ E-mail: d.j.armstrong@warwick.ac.uk

© 2016 The Authors
Published by Oxford University Press on behalf of the Royal Astronomical Society

Downloaded from https://academic.oup.com/mnras/article/465/3/2634/2447827 by guest on 12 May 2022

automatic choice of planetary atmosphere components to include in models using deep neural networks (Waldmann 2016). Mislis et al. (2015) explored the use of random forests in planet detection, but did not test their method on light curves showing significant out of transit variability, and concentrated on white-noise-simulated light curves. In the current age of increasingly large surveys with previously unseen quantities of data, such automated techniques will prove necessary to fully exploit observations, for example in identifying planets, variable stars (Richards et al. 2012; Masci et al. 2014; Armstrong et al. 2016) or other interesting objects. The ability to automate parts of the planetary discovery process will allow the removal of biases introduced by human intervention, making future statistical studies easier to perform and more robust.

We present this technique both as a stand-alone ranking method for planetary candidates, and as a potential stage in the candidate selection process, suitable for combining with more complex methods. It is simple in concept and computationally cheap. The technique is described in Section 3, and methods of using it to classify candidates detailed in Section 4. We apply the SOM to planet candidates from the *Kepler* and *K2* missions, demonstrating its use and ranking those candidates, in Sections 5 and 6. Strengths and weaknesses of the technique are discussed in Section 8.

2 DATA

2.1 *Kepler*

Data from the *Kepler* satellite was used to provide a large set of already classified planets and false positives for testing. We are also able to classify currently unclassified candidates (see Table 2). This data spans approximately 4 yr, with a cadence of 1766 s. Additional data with a shorter cadence near 1 min is also available for some targets; we ignore this to maintain a uniform sample. We use the set of *Kepler* Objects of Interest (KOIs) available on the NASA Exoplanet Archive as of 2016 May 17. We download the Data Validation (DV; Wu et al. 2010) light curves directly from the archive.[1] These are the light curves used to detect the KOIs, and so we deem them the most relevant for testing this method. There were 6384 total dispositioned KOIs with DV light curves available. These include 2247 confirmed planets, 1785 candidates, and 2352 false positives. Transit parameters (period, epoch, and duration) are taken from the archive.

To prepare the transit shapes for entry into the SOM, each light curve is phase folded at the given ephemeris. We then cut to a region of phase within 1.5 transit durations of transit centre, making a 3 transit duration window. This region is binned down to 50 points, which we find gives a suitable resolution on the transit shape, and allows for uniform shape comparisons across KOIs with very different numbers of data points in transit. KOIs with few transits, such that any bin is left unpopulated even after phase folding, are ignored, leaving 6350 signals. Multiple planetary systems do not have other signals removed before phase folding and binning; we find that this has negligible effect on the SOM, and hence demonstrates its resilience to additional signatures both detected and undetected.

2.2 *K2*

We also apply the method to data from *Kepler*'s successor mission *K2*. In this instance, a substantially smaller number of candidates are presently available, due to differences between the missions and the relative youth of *K2*. *K2* observes single fields for ∼80 d campaigns, of which eight have been released to date. The cadence is the same as *Kepler*. We use the list of candidates presented in Crossfield et al. (2016), providing 184 objects including 108 planets and 21 false positives. We used ephemeris and transit durations as provided in that work. There are many options for detrending the raw *K2* data; we utilize the EVEREST pipeline (Luger et al. 2016) here, downloading data using the command line tool provided in that work.

K2 data were prepared similarly to *Kepler*. Due to the shorter baseline available with *K2*, fewer data points are often available for a given transit signal. The result is that 50 bins proved to be too many in several cases, leading to many targets with empty bins in the 3 transit duration window. We trialled smaller numbers of bins, finding that 20 bins were adequate for the SOM to perform. In cases where bins did not have any data points falling in their phase range, we linearly interpolate between nearby bin values.

2.3 PASTIS

We utilize simulated light curves for both testing and analysis of the SOM. These were generated with the PASTIS code (Díaz et al. 2014; Santerne et al. 2015), which produces light curves for various astrophysical scenarios while constraining the false-positive probability of planetary candidate signals. The ability to simulate light curves for different scenarios allows us to test degeneracies in the SOM method. Here, we create 1000 systems each for the six following scenarios: Planets (P), Eclipsing Binaries (EB), Eclipsing Triples (objects consisting of an eclipsing binary and companion, ET), Planets transiting the secondary star of a binary (PSB), Background Eclipsing Binaries (BEB), and Background Transiting Planets (BTP). Each system was drawn from the set of priors given in Table 1. 10 000 noise-free points were simulated within one orbital period, providing a phase curve for testing or injection into real data. The phase curve was smeared to account for the *Kepler* and *K2* long cadence exposure time.

3 SELF-ORGANIZING MAP

A SOM is an unsupervised machine learning algorithm. It finds clusters in the data given to it, without needing labels for that data to be pre-assigned. We will briefly describe the method here. For a more detailed overview, we refer the reader to Armstrong et al. (2016) or Brett et al. (2004).

A SOM consists of an N-dimensional array of 'pixels', in this case with periodic boundaries. The number of dimensions is unimportant here, but the number of pixels must be high enough to represent adequate variation in the input data for the task at hand. A good rule of thumb is to make sure you have a few times more pieces of input data than pixels. Each pixel is a template with the same form as the input data, and is initially randomized. At the end of training, each pixel will resemble a significant pattern in the input data (a typical planetary transit or binary eclipse for example), allowing such patterns to be investigated.

'Training' the SOM occurs over a set number of iterations. In each iteration, each piece of input data (here a single transit signal) is compared to the set of SOM pixels. The best matching pixel is determined, via the minimum Euclidean distance between the input data and pixels. That pixel and those near it are altered to become slightly closer to the piece of input data under consideration. The level of change allowed is determined by the learning rate, α. Pixels

[1] http://exoplanetarchive.ipac.caltech.edu/

Downloaded from https://academic.oup.com/mnras/article/465/3/2634/2447827 by guest on 12 May 2022

Table 1. Priors used to draw simulated systems for testing.

Parameter	Prior
Non-scenario specific	
P	Jeffreys 0.3–100 d
$e, P > 10$ d	Beta as Kipping (2013)
$e, P \leq 10$ d	0
ω	Uniform 0°–360°
i	Uniform in sin$i$, must transit
Target star	
Mass	Normal $1 \pm 0.15\,M_\odot$
[Fe/H]	Normal 0 ± 0.2 dex
Age	Normal 5 ± 2 Gyr[a]
Distance	100 pc
LD coefficients	Claret & Bloemen (2011)
Planet	
R_p	Power law in R_p^{-2}, $1\,R_\oplus$ to $2.2\,R_{jup}$
M_p	Normal, expected mass ±50 per cent
Albedo	0.1
Bound star	
Mass	IMF of Kroupa (2001)
Age	Fixed to target star
[Fe/H]	Fixed to target star
Background star	
Mass	IMF of Kroupa (2001)
[Fe/H]	Uniform −2.5–0.5 dex
Age	Uniform 0.1–13.7 Gyr[a]
Distance	Power law in D^2, 200 pc to 8 kpc
Interstellar extinction	0.7 mag kpc^{-1},[b]

[a]Old or massive stars excluded.
[b]Default in Besançon galactic model.

are altered based on their proximity to the best matching pixel, determined by the learning radius σ. Both α and σ decay during the course of the training, allowing finer levels of detail to emerge.

In our case, we are feeding phase-folded, binned transit light curves into the SOM, prepared as described in Section 2. As such, each SOM pixel will have 50 values (20 for *K2*), which form a template binned transit shape. The goal is to separate the input signals into groups of similar shape, and see if such groups have any power in distinguishing true planets from false positives. We utilize the SOM code provided in the PYMVPA PYTHON package (Hanke et al. 2009).[2] We use 500 training iterations. We set $\alpha = 0.1$ initially, with a linear decay to zero through the course of the iterations. We set $\sigma = 20$ initially (the radius of the SOM) and to decay exponentially as described in Armstrong et al. (2016). The values and decay forms of α and σ do not have a strong effect on the SOM's performance (Brett et al. 2004).

For the *Kepler* data, we use a 20 × 20 two-dimensional SOM consisting of 400 pixels, on 6350 KOIs. As only 184 candidates are available for *K2*, we reduce the size of the SOM to 8×8 and reduce σ accordingly. We choose two dimensions for the ease of visualization.

4 CLASSIFICATION

Once training is complete, a given candidate transit shape can be placed on the SOM by finding the best matching SOM pixel. The location of this pixel, (x, y), and Euclidean distance to it, can be extracted. The challenge at this point is to classify the pixel itself, and hence the candidate under consideration; a priori, we do not know if the pixel represents a planet transit or false positive. The pixel may also be unable to distinguish between the two. Here, we consider two key cases. First, when a large sample of classified objects is already available (i.e. *Kepler*, Case 1), and secondly earlier in a mission lifetime, when few or zero candidates have already been classified (i.e. *K2*, Case 2).

[2] http://www.pymvpa.org

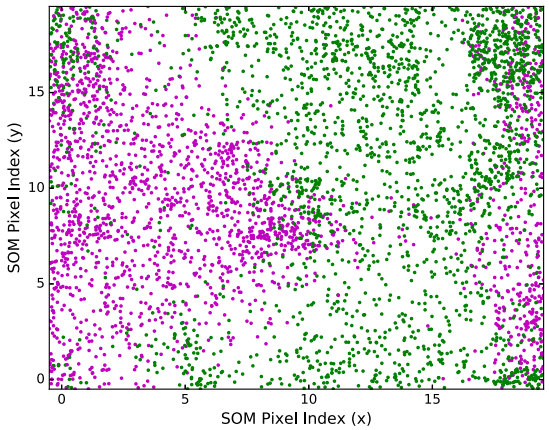

Figure 1. Trained SOM using higher SNR KOIs. KOI signals are plotted on their best matching SOM pixel, with a random offset between −0.5 and 0.5 added to each point for clarity. Planets are magenta and false positives are green. The boundaries are periodic.

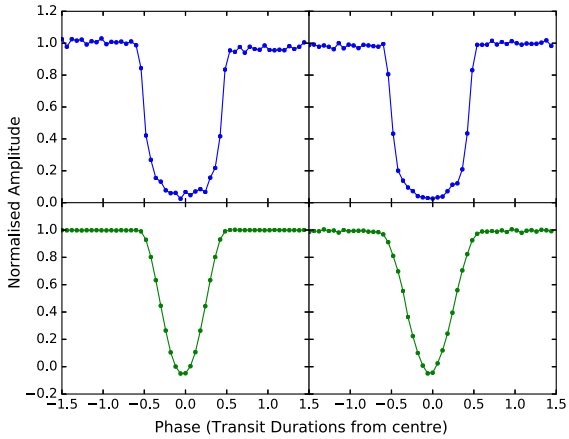

Figure 2. SOM pixel templates extracted from the SOM plotted in Fig. 1. Clockwise from top left the templates are from pixels [1,8], [5,10], [14,3], and [11,17]. The top two templates are from regions dominated by validated planets, and the bottom two templates from regions dominated by false positives.

4.1 Case 1: late in mission lifetime

In this case, a large sample of already dispositioned signals is available. We place the sample of KOIs on the trained *Kepler* SOM in Fig. 1, and show example trained SOM pixel templates in Fig. 2. It is clear that the SOM has power in separating confirmed planets from false positives. Note that the key distinction is between V-shaped and U-shaped transits, something that will be discussed in Section 8.

To classify a candidate signal into one of the two groups, we follow the following method. First, errors on the input signal must be considered. We account for these using a Monte Carlo procedure, whereby each input signal data bin is independently adjusted by a random offset drawn from the normal distribution with mean zero

Downloaded from https://academic.oup.com/mnras/article/465/3/2634/2447827 by guest on 12 May 2022

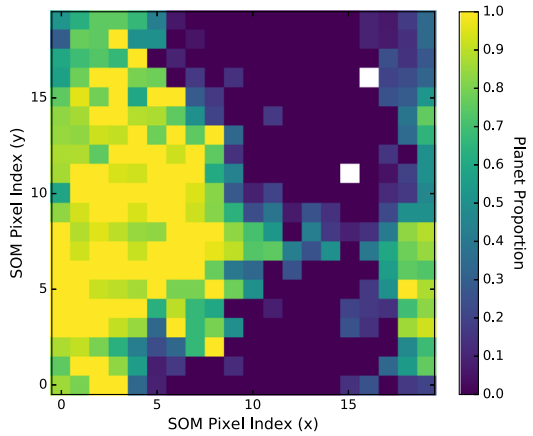

Figure 3. Proportion of dispositioned KOIs within each SOM pixel which are true planets, from the SOM in Fig. 1. Clear grouping is seen. White pixels are those where no dispositioned KOIs are found.

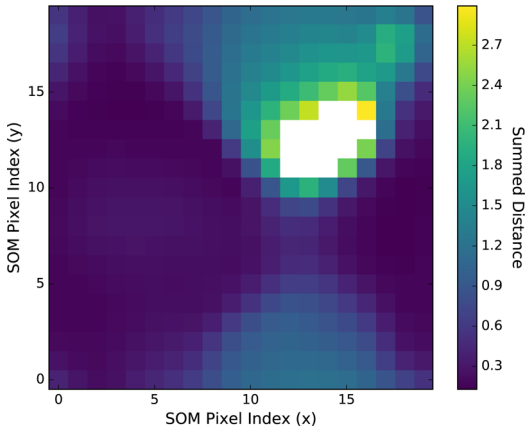

Figure 4. Summed distances between the Fig. 1 SOM pixel templates and simulated planetary signals $D_{\rm P}$, as defined in equation (5). Low distances represent a good match. Here, the grouping can be seen without relying on already dispositioned KOIs. The 14 largest distances have been masked for clarity.

and standard deviation the bin error. The signal is then repositioned on the SOM, and the process repeated for 1000 iterations. In this way, light curves with poorly defined transits can cover a range of pixels on the map, covering the map regions the light curve is compatible with according to its error. This returns a distribution of SOM pixel locations, (x_i, y_i), with i the Monte Carlo iteration index.

The disposition of each candidate signal is then calculated as follows. For each SOM pixel, we take the proportion of already disposition signals within it that are planets, and the proportion that are false positives. Each SOM pixel in the distribution (x_i, y_i) then moves a candidate signal towards either planetary or false-positive status based on the pixel's characterization (see Fig. 3). We further weight each SOM pixel on the number of known signals within it, meaning more well-characterized pixels are given increased classification power. The weights W and proportion of dispositioned signals in pixel (x, y) which are planets, $\alpha_{\rm planet}(x, y)$ are found by

$$W(x, y) = \sum_o (x_o = x, y_o = y) \tag{1}$$

and

$$\alpha_{\rm planet}(x, y) = \frac{\sum_{o=\rm planet} (x_o = x, y_o = y)}{W(x, y)}, \tag{2}$$

where o is an index representing each already dispositioned object. The output statistic is then calculated by

$$\theta_1 = \frac{\sum_i (\alpha_{\rm planet}(x_i, y_i) W(x_i, y_i))}{\sum_i (W(x_i, y_i))}. \tag{3}$$

Values of θ_1 above 0.5 represent planets, and those less than 0.5 false positives. The closer to unity θ_1 is, the more likely a candidate is to be a planet. We stress that this is not a posterior probability (and hence cannot be used for validation of planetary candidates), although it is related. Calibration may be possible in future to make θ_1 more closely resemble a probability, but the statistic would nevertheless be subject to various biases discussed in Section 8. The conversion of θ_1 into a true posterior probability is beyond the scope of this work, as it would need to consider factors such as galactic pointing (and hence crowding) as well as the myriad other inputs to common validation codes such as BLENDER (Torres et al. 2010), PASTIS (Díaz et al. 2014; Santerne et al. 2015), and VESPA (Morton 2012). An example of the usage of this case is given in Sections 5.1 and 6.1.

4.2 Case 2: early in mission lifetime

In this case, a candidate list may have been produced but large numbers of confirmed or validated planets are not yet available. This is the situation where ranking candidates may prove the most use; it is important to select the best candidates to observe for radial velocities for example, without wasting limited telescope resources.

We cannot use the above statistic θ_1. Here, we adapt θ_1 to make use of simulated transit signals, created using PASTIS in Section 2.3. The distance between each SOM pixel and each simulation is calculated, using the sum of the squared difference between each bin, considering only the bins specifically in transit. The average distance of a pixel to each scenario's (e.g. planet, eclipsing binary, and background eclipsing binary) simulation set is taken. The planet scenario distances for the *Kepler* SOM of Fig. 1 are shown in Fig. 4. As such, we calculate the average distance for each set of simulated light curves D_S, where S labels the type of scenario under consideration, by

$$D_{\rm S}(x, y) = \frac{1}{n_{\rm S}} \sum_s \sum_b (T_{\rm SOM}(x, y, b) - T_{\rm S}(s, b))^2, \tag{4}$$

where $T_{\rm SOM}(x, y, b)$ is the value of bin b in SOM pixel (x, y), $T_{\rm S}(s, b)$ is the value of bin b in the simulated light curve s of scenario S, and $n_{\rm S}$ is the number of simulated light curves in scenario S. Next, we calculate the average distance of each SOM pixel to the planet-like and false-positive-like scenarios as

$$D_{\rm P}(x, y) = \langle D_{S=\rm planet-like}(x, y) \rangle \tag{5}$$

and

$$D_{\rm FP}(x, y) = \langle D_{S=\rm false-positive-like}(x, y) \rangle, \tag{6}$$

where planet-like scenarios are the P and PSB scenarios from Section 2.3. We ignore the BTP scenario as the simulated transits are extremely shallow and hence generally uninformative. False-positive-like are the EB, ET, and BEB scenarios, also from Section 2.3. The output statistic is then calculated as

$$\theta_2 = \frac{1}{n_i} \sum_i \left(\frac{D_{\rm FP}(x_i, y_i)}{D_{\rm P}(x_i, y_i) + D_{\rm FP}(x_i, y_i)} \right), \tag{7}$$

Downloaded from https://academic.oup.com/mnras/article/465/3/2634/2447827 by guest on 12 May 2022

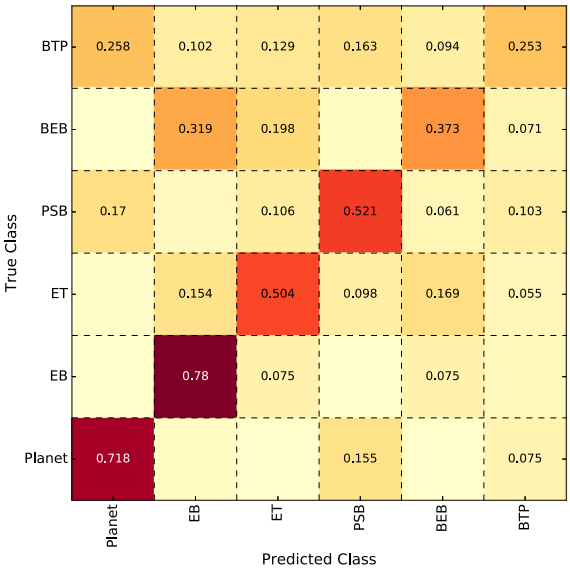

Figure 5. Confusion matrix showing the true class and predicted class for each simulated scenario described in Section 4.3. Light curves are classified using the method described in Section 4.1. Correct classification lies along the diagonal. The proportion of light curves in each scenario which lie in each box is shown. Only proportions greater than 5 per cent are shown for clarity. Confusion is typically between the three planetary scenarios (Planet, PSB, BTP) or between the three stellar scenarios (EB, ET, BEB). Typically shallower scenarios such as BTP are less well classified.

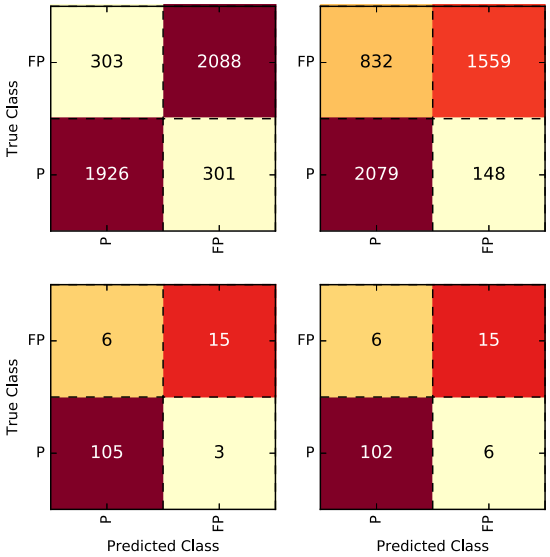

Figure 6. Confusion matrix showing the dispositions given to candidate signals, using a threshold in θ_1 or θ_2 of 0.5 to classify a signal. Matrices are shown for *Kepler* (top) and *K2* (bottom), and for Case 1, θ_1 (left), and Case 2, θ_2 (right). Each box shows the total number of signals classified. P represents planets, FP false positives.

where n_i is the number of Monte Carlo iterations performed. The values of θ_2 have the same properties as θ_1 above. Examples of this case are given in Sections 5.2 and 6.2.

4.3 Testing with PASTIS

Given the set of simulated transit signals, it is possible to test this method for degeneracies. We perform this test by injecting the simulated signals into the *Kepler* DV light curves, and creating binned phase folded transits as described in Section 2. We increase the depth of simulated transits such that each signal is at least marginally detectable. This ensures each simulation contributes information to the SOM; boosting the depth is possible as the purpose here is to test degeneracies between different scenarios rather than to find specific recovery rates. We then train a SOM with these simulations as the input data. We follow the Case 1, proportions based method of classifying to attempt to separate the injected groups. The results are shown in Fig. 5. Some success is found for all scenarios, but two clear groups are formed within which scenarios are degenerate. One is of scenarios where the transiting object is a planet (P, PSB, BTP), the other where the transiting object is a star (EB, ET, BEB). Some mixing between these two groups is seen towards the top right of Fig. 5; in this region, the scenarios used have typically shallower and hence less well-defined transit shapes. We conclude that the method distinguishes stellar eclipses from planetary transits successfully, but cannot exclude false-positive scenarios involving planets (background transiting planets for example). This behaviour is expected, as such false positive are among the hardest to identify, and motivates the planet-like and false-positive-like groups used in the Case 2 method. Furthermore, transits with low signal to noise (SNR) can be confused; this is expected to be a problem for candidates where even the binned transit has a poorly defined shape, and is allowed for in Section 4 using the bin errors. The exact success rates seen in Fig. 5 have no relevance to the success of the method on real data, as they are inherently functions of the distributions of simulated light curves used. The results here are only informative in so much as they highlight degeneracies between true planets and some false-positive classes found when using this method. Testing the method's results on real data is performed in Section 5.

5 APPLICATION TO *KEPLER*

5.1 Case 1

We take the *Kepler* transit signals as prepared in Section 2. We found best results from training the SOM only on the higher SNR transits, and use a cut at 30 in the SNR parameter given by the NASA Exoplanet Archive, leaving 3078 KOIs. The SOM is then trained as described in Section 3. Note that we can obtain classifications for all signals, despite only training on a subset. The locations of all the *Kepler* signals on this SOM are shown in Fig. 1, with examples of the SOM pixel templates underlying the map in Fig. 2.

We will apply both classification methods to the *Kepler* data. The proportions of planets and false positives in each SOM pixel are shown in Fig. 3. The SOM was unaware of the disposition of each signal before training. Hence, we can use every input transit to test the method. We apply equation (3) to the *Kepler* signals and show a histogram of the resulting statistic in Fig. 7. Values of θ_1 greater than 0.5 represent planets, with the strength of the result increasing as θ_1 rises. 1923 of the 2227 dispositioned planets are classified correctly, and 2093 of the 2391 false positives, making an overall success rate of 87.0 per cent. The success rate rises to 91.7 per cent when considering only the higher SNR KOIs used to train the SOM. Taking only objects with 'well-determined' classification (4188 of the 4618 dispositioned KOIs, defined as θ_1 greater than 0.6 or less than 0.4 for planets and false positives, respectively) improves the results for all KOIs to 89.8 per cent, and the results for higher SNR KOIs to 92.8 per cent. Confusion matrices for this case and the others tested are shown in Fig. 6. We note however that the aim of

Downloaded from https://academic.oup.com/mnras/article/465/3/2634/2447827 by guest on 12 May 2022

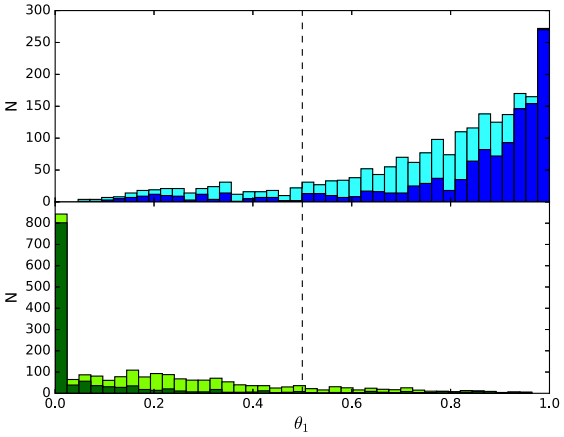

Figure 7. θ_1 statistic for the KOIs. Top: true planets. Bottom: false positives. The lighter shaded histograms show the whole sample, while darker shades represent the higher SNR KOIs.

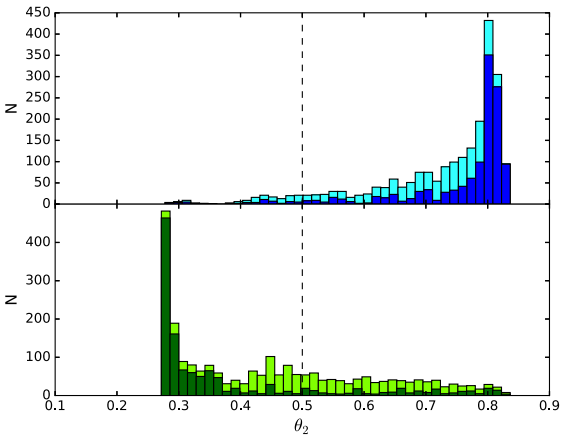

Figure 8. As Fig. 7 for θ_2.

this method is to rank candidates rather than classify them; Fig. 7 is more useful for gaining an understanding of the method's success for this purpose, through the distribution of θ_1 for planets and false positives.

5.2 Case 2

Although we expect Case 1 to work better where possible, it is interesting to compare to Case 2 (using simulated light curves for classification, although not for training). We use the same SOM as Section 5.1 to classify the KOIs using equation (7). The resulting histogram is shown in Fig. 8. Note that θ_2 is generally unable to fully use the parameter space between 0 and 1, because even SOM pixels heavily dominated by planets still have a finite distance to false-positive-simulated light curves and vice versa. 3638 KOIs are classified correctly, making a success rate of 78.8 per cent, considering only higher SNR KOIs as above increases this to 88.9 per cent, nearly as effective as Case 1, while retaining 3572 of the 4618 dispositioned KOIs. Considering only 'well-determined' classifications increases these results to 84.6 per cent and 93.9 per cent for all KOIs and higher SNR KOIs, respectively, equivalent to Case 1. A side effect of θ_2 not using the full 0–1 space is that the distribution of θ_2 is unbalanced; more planets are classified correctly than false positives, as can be seen in Fig. 6 (top-right). This may be a desired outcome, if for example it is more important to maintain planets than remove false positives. The balance could be adjusted by using different thresholds of θ_2, as necessary. We note that these success rates will likely change from mission to mission, but do represent the effectiveness of the method, and a good test of comparison between classification cases. It is possible that future developments may improve them, such as using different and physically motivated numbers of each simulated scenario.

Table 2. Output statistics for KOIs, sorted by θ_1. Full table online.

Kepler ID	θ_1	θ_2
005297298	1.000	0.800
004275191	1.000	0.699
003351888	1.000	0.799
003935914	1.000	0.699
008219268	1.000	0.811
010723750	1.000	0.800
009818381	1.000	0.814
008552719	1.000	0.797
005780885	1.000	0.782
007869917	1.000	0.795
010418224	1.000	0.800
007515679	1.000	0.799
⋮	⋮	⋮

5.3 Results

The results for the KOIs are given in Table 2. This table provides a ranked list of the undispositioned *Kepler* candidates, which we hope to be useful to anyone considering selecting targets for follow-up. It can be combined with other diagnostics, such as those provided on the NASA Exoplanet Archive, or codes. It is important to be aware of the biases involved in this selection. First, as demonstrated in Section 4.3, false positives involving planetary transits are not distinguished. Secondly, as the SOM separates false-positive signals primarily on V shape, grazing planets will likely be classified as false positives. While this is regrettable, grazing planets are difficult to follow-up, and are relatively few in number.

Several KOIs dispositioned as planets are given low θ values, in both θ_1 and θ_2. We examined by eye the cases where θ_1, the most successful method, was unable to classify planets successfully. 80 per cent of the 301 failure cases either showed a very low SNR signal (∼40 per cent), or no signal at all (∼40 per cent), implying that either the planet is too small to be detectable using 50 bins, or that the ephemeris provided by the archive was erroneous. The remaining 20 per cent were clear transits with a V shape, and hence are likely grazing planets as discussed above. We investigate the effects of SNR on performance by considering the ratio of planets classified correctly as a function of SNR. We estimate SNR by taking the difference between the average out-of-transit and average in-transit bins, divided by the standard deviation of the out-of-transit bins. The dependence of performance on SNR is shown in Fig. 9, and shows a clear decrease at low SNR. We also tested against planetary radius as given by the NASA Exoplanet Archive, and found that for low planetary radius a decrease in performance was seen (as expected given the SNR decrease which accompanies low radius). We found no other dependence on planetary radius.

Downloaded from https://academic.oup.com/mnras/article/465/3/2634/2447827 by guest on 12 May 2022

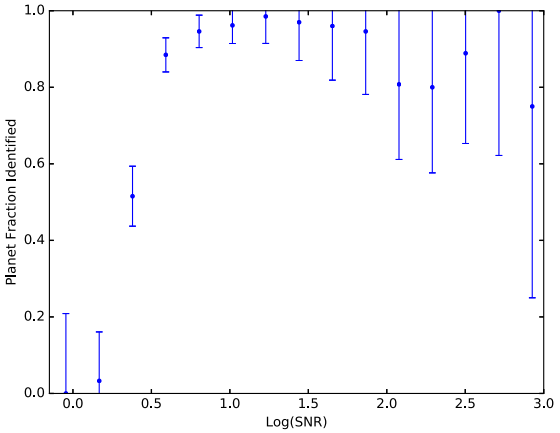

Figure 9. Fraction of successfully classified planets as a function of SNR, as defined in Section 5.3. A clear drop off at low SNR is seen. Bins are spaced evenly in log (SNR) space, and contain between 4 and 555 KOIs each. The error bars represent the Poisson counting error on the number of samples in each bin. 12 KOIs with the lowest SNR were not detected, are not shown for clarity.

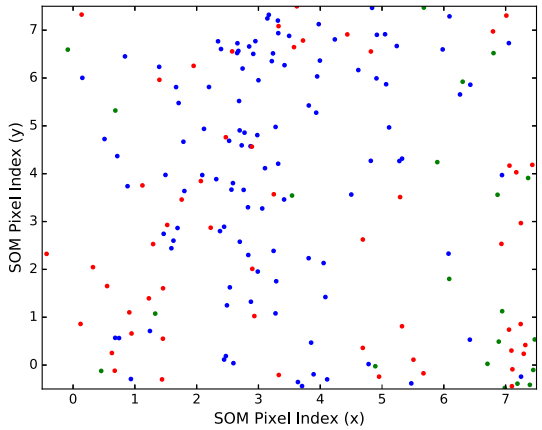

Figure 10. As Fig. 1 for the *K2 sample*. Blue points represent validated planets, green validated false positives, and red undispositioned candidates.

6 APPLICATION TO *K2*

6.1 Case 1

We apply the SOM to *K2* similarly to *Kepler*. *K2* transits are binned down to 20 bins rather than 50, but we find encouragingly that the method is still effective. We do not make any SNR cuts for training the *K2* SOM, due to the lower number of signals.

The location of *K2* signals on the trained SOM is shown in Fig. 10. The grouping is less clear due to the lower numbers, although it is apparent when considering the distances to the simulated light curves as described in Section 4.2, and shown in Fig. 11. The results of attempting to use equation (3) are shown in Fig. 12. For the planets θ_1 performs well, but due to the low numbers of confirmed false positives the histogram for these is poorly populated. The success rate is reasonable, with 104 of the 108 planets classified correctly and 16 of the 21 false positives, giving a 93 per cent success rate overall. Given the low numbers of false positives however this is potentially spurious. Furthermore, in the case of upcoming missions such as *PLATO* (Rauer et al. 2014) or *TESS* (Ricker et al. 2014), even this low number of dispositioned candidates will not be initially available.

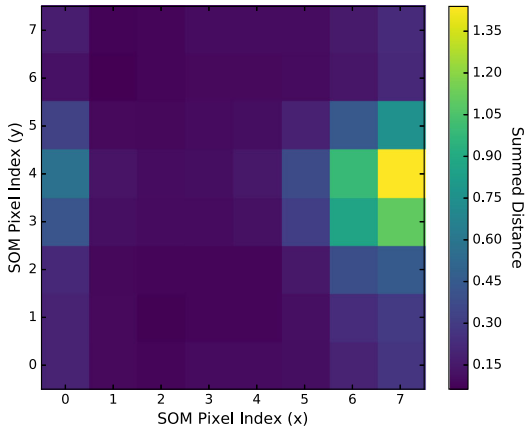

Figure 11. Summed distances between the Fig. 10 SOM pixel templates and simulated planetary signals. Low distances represent a good match, and highlight the grouping seen.

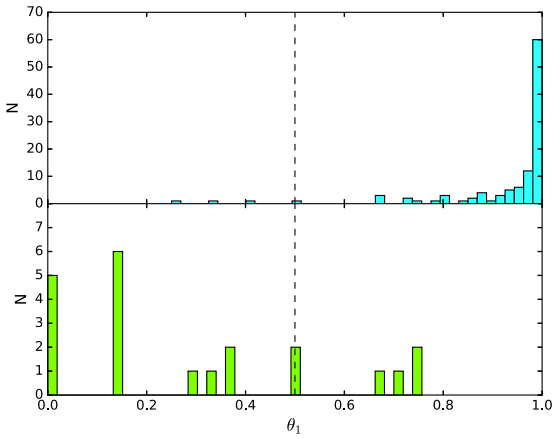

Figure 12. θ_1 statistic for *K2*. Top: true planets. Bottom: false positives.

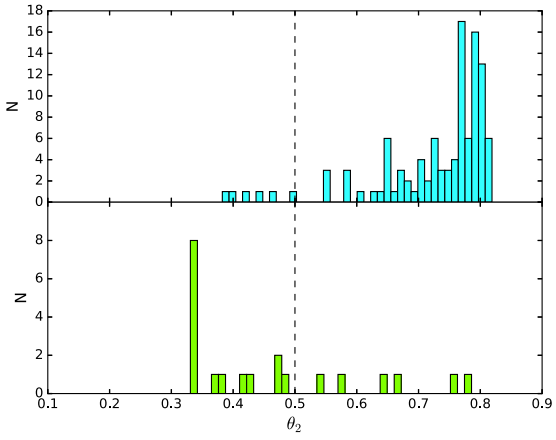

Figure 13. As Fig. 12 for θ_2.

6.2 Case 2

As such we turn to Case 2. The results of applying equation (7) to the *K2* signals are shown in Fig. 13. These have a similar success rate to Case 1, but require no already known candidates and are more robust. While still poorly populated and hence hard to test, the histogram of θ_2 values for the false positives has a clearer distribution towards the low end, as desired. 102 of the 108 planets

Downloaded from https://academic.oup.com/mnras/article/465/3/2634/2447827 by guest on 12 May 2022

Table 3. Output statistics for *K2* objects, sorted by θ_1. Full table online.

EPIC ID	θ_1	θ_2
201445392	1.000	0.774
201295312	1.000	0.770
201324549	1.000	0.773
201713348	1.000	0.755
201247497	1.000	0.803
201549860	1.000	0.781
201565013	1.000	0.811
201565013	1.000	0.794
201596316	1.000	0.801
201596316	1.000	0.754
201345483	1.000	0.743
201702477	1.000	0.797
201613023	1.000	0.754
⋮	⋮	⋮

are classified correctly, and 15 of the 21 false positives, representing similar performance to Case 1. As the calculation of θ_2 does not depend at all on the low number of known false positives, we believe it to be more reliable in this instance.

6.3 Results

The results for the *K2* candidates are shown in Table 3. Users should be aware of the same caveats as highlighted in Section 5.3.

7 CODE AVAILABILITY

The code used to create and train the SOM is already part of a public package, PYMVPA.[3] To make using this method easier for readers, we make available code to classify a *Kepler* or *K2* light curve using our pre-trained SOM, along with convenience functions for users to create their own SOMs. This is available on GITHUB,[4] along with documentation describing its use.

8 DISCUSSION

SOMs have proven to be effective at separating true planetary signals from false positives, using only the shape of the candidate signal. The source of this discriminatory power is made clear by Fig. 2. In short, transiting planets produce U-shaped transits, whereas most stellar eclipses are V shaped, primarily due to the different radius ratios involved in each case. This difference is well known; it forms a key part of planetary candidate selection in most current surveys. To date, however, the choice of whether a candidate is too V shaped to continue observing has typically been made by humans, with the subjective biases and inconsistent thresholds that implies. 'How V shaped is too V shaped?' is a question often answered on a case by case basis. The SOM provides an opportunity to standardize, quantify, and speed up this process. We expect the SOM to be useful either as a fast pre-screen of a given candidate list, or as input to a more comprehensive autovetting code which incorporates other inputs such as secondary eclipse detections.

[3] http://www.pymvpa.org
[4] https://github.com/DJArmstrong/TransitSOM

It is possible for true planets to give V-shaped transits. This occurs for grazing transits, where only part of the planet occults the stellar disc, as well as for near-grazing planets and short-period planets observed at a very long cadence. As such, in removing V-shaped signals we are removing some true planets. This problem is common to both human selection and the SOM. While it would be preferable to maintain all planets, losing some at the expense of the majority of false positives is generally considered worthwhile given limited telescope time. Furthermore, grazing transits are difficult to model, as they present a degeneracy between radius ratio, impact parameter and inclination which is easier to separate in the full transit case.

A potential issue in our development of the SOM arises from the KOI sample. The majority of this sample has been dispositioned using validation (e.g. Morton et al. 2016), without separate observations of the planetary mass. This process relies on finding the false-positive probability of a candidate, using its galactic pointing, local crowding, transit shape and host star parameters. As we are relying exclusively on the transit shape, one of the key validation inputs, there is a danger of bias in using the validated sample to confirm the method. We have mitigated for this by marking 'confirmed' planets (those with detected masses) separately in the SOM. Confirmed planets follow the same groupings, supporting our conclusions. Furthermore, testing with PASTIS (Section 4.3) successfully and independently checked the effectiveness of our method. We note that the success of the SOM demonstrates the power of the transit shape alone in the validation process, at least to the point of separating stellar eclipses from planetary transits.

In a reversal of the main goal of this work, it is possible to use the SOM to identify eclipsing binaries and triple stars. Catalogues of eclipsing binaries (Armstrong et al. 2015, 2016; LaCourse et al. 2015) are useful science products of planet surveys, with the most recent catalogue for *K2* using SOMs to identify the binary stars in the sample. Eclipsing binaries provide one of the only direct tests of stellar evolution models, and can even host planetary systems themselves (e.g. Doyle et al. 2011).

We have developed this method with the aim of separating astrophysical false positives from true planetary signals. In doing this, we ignore candidates produced due to instrumental noise and apparently periodic noise patterns. These can be removed with other techniques; looking for clusters of candidates in epoch space for example. Here, the SOM can also contribute. First, noise-candidates do not typically show a transit-like shape. As such, they will have a large distance from even their best matching pixel on the SOM. This can be used to separate candidates. If noise-candidates are included in training the SOM, they will develop their own region on the map, one which does not resemble any simulated astrophysical light curve; again, this can be utilized. If all such non-matching candidates are designated false positives, the planet sample will be preserved.

9 CONCLUSION

A new method for identifying the best planetary candidates for follow-up has been developed, tested, and applied to the *Kepler* and *K2* data sets. The SOM replies only on the transit shape, and can achieve accuracies of nearly 90 per cent in distinguishing known *Kepler* planets from false positives. We apply the technique to the unclassified *Kepler* and *K2* candidates, and hope the resulting rankings will be useful to the community.

This method adds to the developing body of techniques for automatic vetting of planetary candidates. SOMs can contribute both as a quick initial screening step and as a part of larger autovetting codes.

Downloaded from https://academic.oup.com/mnras/article/465/3/2634/2447827 by guest on 12 May 2022

Such codes are beginning to become available for *K2* (Coughlin et al. 2016), and we intend to apply this method in combination with similar techniques in the future. Autovetting is a growing field, and will become increasingly important as new missions such as *TESS* and *PLATO* begin to produce data. The unprecedented large data volume of these missions will require automatic techniques to maximize their effectiveness. In addition to follow-up efficiency, automatic techniques allow faster and more detailed studies of completion rates in planetary surveys, allowing statistical studies to be made more easily and more robustly. We expect developments in this field to progress rapidly from now on.

ACKNOWLEDGEMENTS

We would like to thank the anonymous referee for providing useful comments on the manuscript. DJA acknowledges funding from the European Union Seventh Framework programme (FP7/2007-2013) under grant agreement no. 313014 (ETAEARTH). This publication was aided by the international team led by J. Cabrera on 'Researching the Diversity of Planetary Systems' at ISSI (International Space Science Institute) in Bern. Part of this work was supported by Fundação para a Ciência e a Tecnologia, FCT, (ref. UID/FIS/04434/2013 and PTDC/FIS-AST/1526/2014) through national funds and by FEDER through COMPETE2020 (ref. POCI-01-0145-FEDER-007672 and POCI-01-0145-FEDER-016886). AS is supported by the European Union under a Marie Curie Intra-European Fellowship for Career Development with reference FP7-PEOPLE-2013-IEF, number 627202. This paper includes data collected by the *Kepler* mission. Funding for the *Kepler* mission is provided by the NASA Science Mission directorate. The data presented in this paper were obtained from the Mikulski Archive for Space Telescopes (MAST). STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. Support for MAST for non-*HST* data is provided by the NASA Office of Space Science via grant NNX13AC07G and by other grants and contracts.

REFERENCES

Almenara J. M. et al., 2009, A&A, 506, 337
Armstrong D. J. et al., 2015, A&A, 579, A19
Armstrong D. J. et al., 2016, MNRAS, 456, 2260
Auvergne M. et al., 2009, A&A, 506, 411
Bakos G., Noyes R. W., Kovacs G., Stanek K. Z., Sasselov D. D., Domsa I., 2004, PASP, 116, 266
Borucki W. J. et al., 2010, Science, 327, 977
Brett D. R., West R. G., Wheatley P. J., 2004, MNRAS, 353, 369
Carrasco Kind M., Brunner R. J., 2014, MNRAS, 438, 3409
Claret A., Bloemen S., 2011, A&A, 529, A75
Coughlin J., Mullally F., Mullally S., Colón K. D., Barentsen G., Quintana E. V., Burke C. J., Barclay T., 2016, Am. Astron. Soc. Meeting Abstr. 228, 102.04
Crossfield I. J. M. et al., 2016, ApJS, 226, 7
Díaz R. F., Almenara J. M., Santerne A., Moutou C., Lethuillier A., Deleuil M., 2014, MNRAS, 441, 983
Doyle L. R. et al., 2011, Science, 333, 1602
Hanke M., Halchenko Y. O., Sederberg P. B., Hanson S. J., Haxby J. V., Pollmann S., 2009, Neuroinformatics, 7, 37
Howell S. B. et al., 2014, PASP, 126, 398
Kipping D. M., 2013, MNRAS, 434, L51
Kohonen T., 1982, Biol. Cybern., 43, 59
Kohonen T., 1990, Proc. IEEE, 78, 1464
Kovacs G., Zucker S., Mazeh T., 2002, A&A, 391, 369
Kroupa P., 2001, MNRAS, 322, 231
LaCourse D. M. et al., 2015, MNRAS, 452, 3561
Luger R., Agol E., Kruse E., Barnes R., Becker A., Foreman-Mackey D., Deming D., 2016, ApJ, 152, 100
McCauliff S. D. et al., 2015, ApJ, 806, 6
Masci F. J., Hoffman D. I., Grillmair C. J., Cutri R. M., 2014, AJ, 148, 21
Mislis D., Bachelet E., Alsubai K. A., Bramich D. M., Parley N., 2015, MNRAS, 455, 626
Morton T. D., 2012, ApJ, 761, 6
Morton T. D., Bryson S. T., Coughlin J. L., Rowe J. F., Ravichandran G., Petigura E. A., Haas M. R., Batalha N. M., 2016, ApJ, 822, 86
Pollacco D. et al., 2006, Ap&SS, 304, 253
Pope B. J. S., Parviainen H., Aigrain S., 2016, MNRAS, 461, 3399
Rauer H. et al., 2014, Exp. Astron., 38, 249
Richards J. W., Starr D. L., Miller A. A., Bloom J. S., Butler N. R., Brink H., Crellin-Quick A., 2012, ApJS, 203, 32
Ricker G. R. et al. 2014 in Oschmann J. M., Jr, Clampin M., Fazio G. G., MacEwen H. A., eds Proc. SPIE Conf. Ser. Vol. 9143, Space Telescopes and Instrumentation 2014: Optical, Infrared, and Millimeter Wave. SPIE, Bellingham, p. 914320
Santerne A. et al., 2012, A&A, 545, A76
Santerne A. et al., 2015, MNRAS, 451, 2337
Santerne A. et al., 2016, A&A, 587, A64
Siverd R. J. et al., 2012, ApJ, 761, 123
Thompson S. E., Mullally F., Coughlin J., Christiansen J. L., Henze C. E., Haas M. R., Burke C. J., 2015, ApJ, 812, 46
Torniainen I. et al., 2008, A&A, 482, 483
Torres G. et al., 2010, ApJ, 727, 24
Waldmann I. P., 2016, ApJ, 820, 107
Wu H. et al., 2010, in Radziwill N. M., Bridger A., eds, Proc. SPIE Conf. Ser. Vol. 7740, Software and Cyberinfrastructure for Astronomy. SPIE, Bellingham, p. 774019

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Table 2. Output statistics for KOIs, sorted by θ_1.
Table 3. Output statistics for K2 objects, sorted by θ_1.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TEX/LATEX file prepared by the author.

Downloaded from https://academic.oup.com/mnras/article/465/3/2634/2447827 by guest on 12 May 2022