Astronomy in Focus, Volume 1, Focus Meeting 8
XXIXth IAU General Assembly, August 2015
Piero Benvenuti, ed.

doi:10.1017/S1743921316002842

Advances in the *Kepler* Transit Search Engine

Jon M. Jenkins

NASA Ames Research Center,
M/S 244-30, Moffett Field, CA 94035 U.S.A.
email: jon.jenkins@nasa.gov

Abstract. Twenty years ago, no planets were known outside our own solar system. Since then, the discoveries of ∼1500 exoplanets have radically altered our views of planets and planetary systems. This revolution is due in no small part to the *Kepler Mission*, which has discovered >1000 of these planets and >4000 planet candidates. While *Kepler* has shown that small rocky planets and planetary systems are quite common, the quest to find Earth's closest cousins and characterize their atmospheres presses forward with missions such as NASA Explorer Program's Transiting Exoplanet Survey Satellite (TESS) slated for launch in 2017 and ESA's PLATO mission scheduled for launch in 2024.

These future missions pose daunting data processing challenges in terms of the number of stars, the amount of data, and the difficulties in detecting weak signatures of transiting small planets against a roaring background. These complications include instrument noise and systematic effects as well as the intrinsic stellar variability of the subjects under scrutiny. In this paper we review recent developments in the *Kepler* transit search pipeline improving both the yield and reliability of detected transit signatures.

Many of the phenomena in light curves that represent noise can also trigger transit detection algorithms. The *Kepler Mission* has expended great effort in suppressing false positives from its planetary candidate catalogs. Over 18,000 transit-like signatures can be identified for a search across 4 years of data. Most of these signatures are artifacts, not planets. Vetting all such signatures historically takes several months' effort by many individuals. We describe the application of machine learning approaches for the automated vetting and production of planet candidate catalogs. These algorithms can improve the efficiency of the human vetting effort as well as quantifying the likelihood that each candidate is truly a planet. This information is crucial for obtaining valid planet occurrence rates. Machine learning approaches may prove to be critical to the success of future missions such as TESS and PLATO.

Keywords. statisics, machine learning, transit photometry, exoplanets

1. Transit Detection and Additional Statistical Tests

Jenkins (2002) introduced an adaptive, wavelet-based matched filter for detecting transiting planet signatures against time-varying, non-white observation noise. The transit detection is conducted in a joint-frequency domain through the wavelet transform, but can more simply be represented in the time domain as a simple matched filter where the whitened flux time series, $\tilde{x}$, is projected onto the whitened transit pulse waveform, $\tilde{s}$:

$$Z = \frac{\tilde{x} \cdot \tilde{s}}{\sqrt{\tilde{s} \cdot \tilde{s}}} = \frac{\sum_{i=1}^{p} \tilde{x}_i \cdot \tilde{s}_i}{\sqrt{\sum_{i=1}^{p} \tilde{s}_i \cdot \tilde{s}_i}}, \tag{1.1}$$

where Z is the detection statistic (thresholded at 7.1 σ for *Kepler* transit searches), and the rightmost term is written in terms of the individual transits, of which we assume there are p. Fig. 1 shows the light curve for one of *Kepler's* stars along with the whitened light curve. Fig. 2 shows a montage of the impulse response of the adaptive whitener,

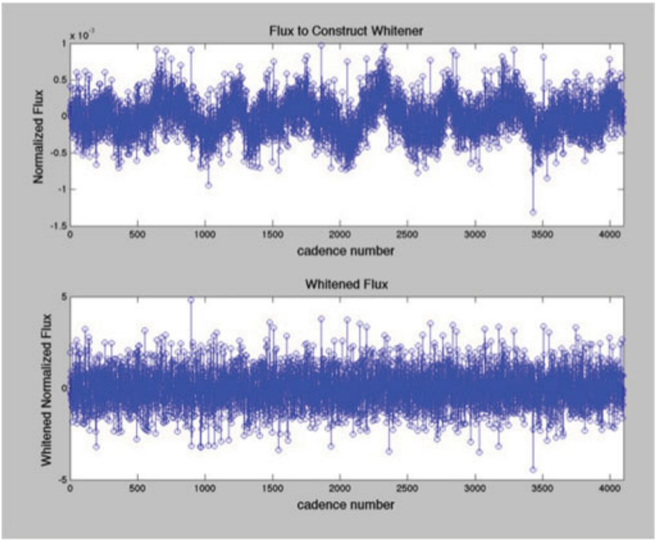

Figure 1. Scatter plot of a section of a flux time series for a *Kepler* star (top) and the whitened flux time series for this same star (bottom).

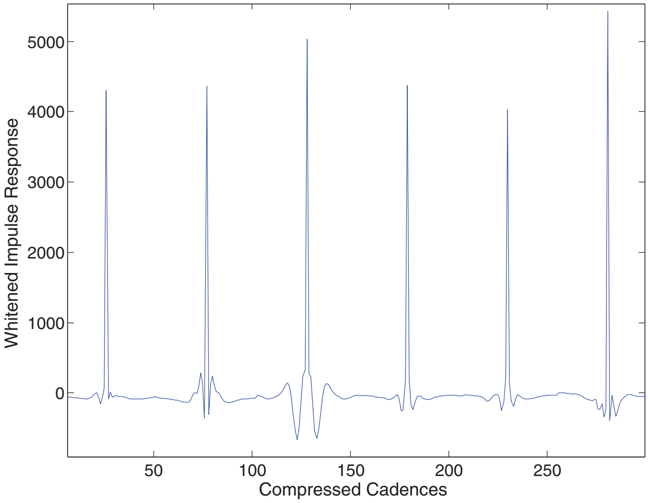

Figure 2. Montage of time-varying impulse response of the whitener applied in Fig. 1.

illustrating the fact that the whitener adapts to non-stationary noise. Eq. 1.1 explicitly takes the non-stationary noise into account when formulating the detection statistic.

To help control the false alarm rate due to structured noise in the flux light curves, Seader *et al.*(2013) added a robust version of the detection statistic, Z_{robust} based on an iterative kernel-based weighting of the individual in-transit data points as per Holland & Welsch (1977). Z_{robust} is thresholded typically at 6.8 σ. Seader *et al.*(2013) also formulated two different χ^2 metrics that examine the degree to which the scatter of the residual in-transits points matches expectations after the transit model has been subtracted from the data. $\chi^2_{(2)}$ pursues this question at the level of the individual transits and has $p-1$ degrees of freedom, while χ^2_{GOF} handles this at the level of the in-transit points, and has $n_{transit}$ degrees of freedom, where $n_{transit}$ is the number of points in transit.

https://doi.org/10.1017/S1743921316002842 Published online by Cambridge University Press

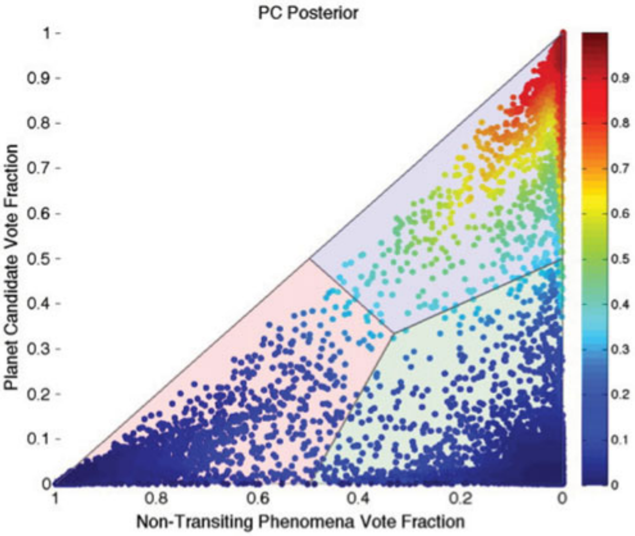

Figure 3. Ternary plot of the posterior probability of the classification of ‘planet candidate’ for ~16,000 TCEs based on the vote fractions of a random forest. The color bar indicates the posterior probability for each TCE.

If a transit-like feature exceeds the thresholds for all four tests, it is registered as a Threshold Crossing Event (TCE) and subjected to a suite of diagnostic tests and physical model fitting, and passed on the the Threshold Crossing Event Review Team (TCERT) for dispositioning as a planet candidate, an astrophysical false positive, or an artifact.

2. Random Forests for Automated Classification of Transit-Like Features

We are exploring supervised machine learning approaches to dispositioning TCEs that leverage the extensive history of the TCERT in vetting previous sets of TCEs. Figure 3 shows the result of training a random forest on the TCEs resulting from a search through the first 16 quarters of *Kepler* data and overlaying a Bayesian classification scheme on top of the feature space defined by the vote fractions of the random forest for each of three classes: planet candidate, astrophysical false positive, and non-transiting phenomena. This allows us to construct posterior probabilities for each class based on the behavior of the training sets. These posteriors can be used in occurrence rate calculations to reduce the sensitivity to which objects are or are not included.

References

Breiman, L. 2001 *Machine Learning*, 45, 1

Jenkins, J. M. 2002, *ApJ* 575, 493

Holland, P. W.,& R. E. Welsch, Communications in Statistics: Theory and Methods, A6 , 1977, 813–827.

Seader, S., Tenenbaum, P., Jenkins, J. M., & Burke, C. J. 2013, *ApJS*, 206, 25

https://doi.org/10.1017/S1743921316002842 Published online by Cambridge University Press