Formation, Detection, and Characterization of Extrasolar Habitable Planets
Proceedings IAU Symposium No. 293, 2012
N. Haghighipour, ed.

doi:10.1017/S174392131301260X

The Kepler Completeness Study: A Pipeline Throughput Experiment

Jessie L. Christiansen, Bruce D. Clarke, Christopher J. Burke, Jon M. Jenkins and the Kepler Completeness Working Group

SETI Institute/NASA Ames Research Center,
Mail Stop 244-30, P.O. Box 1, Moffett Field, CA 94035-0001
email: jessie.l.christiansen@nasa.gov

Abstract. The Kepler Mission was designed to measure the frequency of Earth-like planets in the habitable zone of Sun-like stars. A requirement for determining the underlying planet population from a sample of detected planets is understanding the completeness of that sample—what fraction of the planets that could have been discovered in a given data set were actually detected. Here we describe an experiment designed to address a specific aspect of that question, which is the issue of signal throughput efficiency. We investigate the extent to which the Kepler pipeline preserves transit signals by injecting simulated transit signals into the pixel-level data, processing the modified pixels through the pipeline, and measuring their detection statistics. For the single channel that we examine initially, we inject simulated transit signal trains into the pixel time series of each of the 1801 targets for the 89 days that constitute Quarter 3. For the 1680 that behave as expected in the pipeline, on average we find the strength of the injected signal is recovered at 99.6% of the strength of the original signal. Finally we outline the further work required to characterise the completeness of the Kepler pipeline.

Keywords. (stars:) planetary systems, techniques: photometric, methods: statistical

1. Introduction

The *Kepler* Mission is a NASA Discovery mission designed to measure $\eta_{\oplus}$, the frequency of Earth-size planets in the habitable zone of Sun-like stars. It was launched in 2009, and since then has been nearly continuously monitoring the brightness ∼160,000 stars in 30-minute integrations, in a fixed field of view in the constellation Cygnus, looking for the periodic dimmings indicative of transiting planets. Thus far the project has released three catalogues of planet candidate events (Borucki *et al.* 2011a, Borucki *et al.* 2011b, Batalha *et al.* 2012).

In pursuit of *Kepler*'s primary goal of measuring $\eta_{\oplus}$, we are required to take the sample of planet candidate events and infer the parent population of planets. This process includes by necessity a set of assumptions which must be carefully chosen and justified in each analysis. Borucki *et al.* (2011b; referred to as B11 for the remainder of this proceeding), Youdin (2011), and Howard *et al.* (2012) describe initial analyses of the published *Kepler* planet candidate lists and preliminary attempts to constrain the underlying planet distribution. Two significant components of the analyses about which our knowledge is continuing to mature are the completeness of the planet sample (i.e. the false negative rate) and the reliability of the planet sample (i.e. the false positive rate).

Thus far there has been no concerted study into the false negative rate in the *Kepler* planet candidate lists. In B12 we showed that the number and distribution of additional planet candidates detected in that catalogue was significantly higher and systematically different from what would be expected from a simple signal-to-noise extrapolation of the distribution of planet candidates listed in the previous B11 catalogue.

This highlighted the incompleteness in the B11 catalogue, where planets with a strong enough signal in the original data were not detected. The B12 catalogue will also include some degree of incompleteness; indeed, several teams have identified planet candidates that were not included in that catalogue (Fischer *et al.* 2012, Huang *et al.* 2012, Ofir & Dreizler 2012).

This proceeding presents the first results of an experiment to measure the throughput efficiency of transit signals in the *Kepler* pipeline; i.e. the extent to which transit signals are preserved over the course of the data reduction. This is an essential ingredient in calculations of the planet candidate completeness, and in efforts to date it has been assumed to be 100%. We investigate the validity of this assumption by injecting fake transit signals into the *Kepler* pixels and processing the pixels through the pipeline in the same manner as the original data, and examine the detection statistics of the injected signals. In Section 2 we describe the experimental design and execution, in Section 3 we present the results, and in Section 4 we outline the future plans.

2. Experiment Design

We would like to assess the recoverability of a putative planet candidate around a given target star. Ideally, we would measure this by injecting the target light curve with a grid of fake transit signals, over a set of planet parameters of interest (e.g. size, orbital period), and for each signal, process the light curve through the *Kepler* science pipeline and directly measure the detection statistics. However, due to the number of *Kepler* targets (∼160,000), and the number of observations per target (40,000 and growing), this is computationally infeasible, and we need to restate the problem.

Instead, we decide to assess the recoverability of a putative planet candidate around an *average* target star, and thus reduce the question to ensemble statistics. This decreases the number of tests required by several orders of magnitude to the point that it becomes tractable. To achieve this, instead of running multiple simulations on each target, we inject each target light curve with a different transit signal arising from a randomly generated single planet candidate, described below, and then measure the average recovered detection statistics.

Another way to reduce the computational burden is to reduce the observation baseline, since the number of searches performed by the pipeline increases as N^2, for N observations. Therefore we perform this initial experiment using a single quarter of data—Quarter 3, which spanned 89 days from 2009 September 18 to 2009 December 16. Unfortunately, this limit on the duration of the observations means that for planets with periods longer than 90 days, we would only be able to inject one transit event per light curve, and on average we would sample very few of the systematics and features in those light curves. However, since we are largely concerned with the average signal distortion introduced by the processes in the pipeline, we can treat each separate transit event in the light curve as an independent statistical test of the distortion. Therefore, as long as the transit events are separated well enough in time so as to not mutually influence each other's detection statistics, we can place them arbitrarily closely together in the light curve. The variance window over which the Transiting Planet Search (TPS) pipeline module calculates the noise, in order to determine the significance of a detection, is 30 times the duration of the box pulse being tested; therefore, nothing outside of this window can have any effect on the measured SNR of a given transit. For this investigation, we separate the injected transit events by 50 times their duration.

https://doi.org/10.1017/S174392131301260X Published online by Cambridge University Press

2.1. *Transit model construction*

For each target, we generate a transit model to be injected. The model is initially constructed from three observable parameters: (1) The signal strength, which is randomly drawn from a normal distribution between 2σ and 20σ. This range is to allow us to examine any dependence of the distortion on the initial signal strength. (2) The signal duration, which is randomly drawn from a normal distribution between 1 and 16 hours. In the pipeline, we search for box pulses with durations from 1.5–15 hours; we choose the larger range in the injected pulses to examine the recoverability of the signals when they are outside our nominal search range. (3) The phase at which the first transit in injected, which is randomly drawn from a normal distribution from 0–1; in a following step this is used with the separation of the transit events to calculate the initial epoch.

From a couple of starting assumptions and knowledge of the target star, we can then use those observable parameters to reverse engineer the planet parameters required to generate a model with those features. For this test, we assume circular orbits (eccentricity of zero) and central-crossing transits (impact parameter of zero). Eccentricity has only a very slight impact on the transit shape and therefore recoverability, and assuming circular orbits allows us to easily calculate the orbital period of the injected planet from the selected signal duration. The impact parameter has a large effect on the shape of the transit, however we are primarily concerning ourselves with the question of distortion in the measured signal strength: for an initial signal with 10σ significance, what is the typical final signal strength measured by the pipeline? By allowing the impact parameter to be an additional input parameter, when generating the model we would have to adjust the injected planet size to recover the input signal strength, i.e. for a given signal strength, the impact parameter and the planet size are degenerate parameters.

To generate the model, we use the stellar parameters (surface gravity, effective temperature, stellar radius and metallicity) from the Kepler Input Catalog (KIC; Brown *et al.* 2011) and default to solar values for unclassified stars. We find the rms CDPP (Combined Differential Photometric Precision; Christiansen *et al.* 2012) previously calculated by the pipeline for the target light curve, for the duration which is closest to our model signal duration. We calculate the model transit depth, δ, from the product of the rms CDPP (which is the average depth of a 1-sigma signal in the target light curve) and the model signal strength. Note that when we inject the model into the real data, local noise will result in individual transit events having a range of measured depths, independent of distortions caused by the pipeline. We control for this in our final analysis by comparing each individual transit, processed through the pipeline, to the same model transit injected at the same place in the light curve, without having been processed by the pipeline. From the transit depth, we can calculate the planetary radius, R_p, from $\delta = (R_p/R_\star)^2$, where $R_\star$ is the stellar radius.

We estimate the orbital period from the scaling relation:

$$P = \left[\frac{t_{\rm dur}}{(1.4 * M_\star^{-1/3} * R_\star)} \right]^3 \tag{2.1}$$

where $M_\star$ and $R_\star$ are the target star mass and radius in units of kg and m respectively. This then allows us to calculate the semi-major axis, a, from Newton's modification of Kepler's third law, from which we can calculate the geometric ratio $a/R_\star$. We can then use the Mandel & Agol (2002) analytic transit model formalism to generate a model, centred at the calculated starting epoch, and repeated along the light curve at our predefined transit separation, which is 50 times the model transit duration, as described earlier. The model is generated at a higher sampling rate than the light curve by a factor

https://doi.org/10.1017/S174392131301260X Published online by Cambridge University Press

of 30, and then re-sampled onto the light curve time stamps. We use the KIC magnitude for each target star to convert relative depth as calculated by the model into the total number of photoelectrons that need to be subtracted from the light curve.

2.2. *Pixel-level transit injection*

For each observation, we use a set of bright target stars on each channel to derive the conversion from celestial coordinates—Right Ascension (RA) and Declination (Dec)—to pixel coordinates, fitting a polynomial to the KIC RA and Dec of the bright stars and their measured pixel location in that observation Twicken *et al.* 2010. By injecting transit signals into every target, and increasing the number of events injected per target above expectations, we risk introducing noise into these derived polynomials that would not be present in a normal pipeline run. Therefore we use the polynomials derived from an identical 'clean' pipeline run, without the transits injected.

Using these polynomials, we calculate the location of each target for each observation from the KIC RA and Dec. The PRF is a function of position, so for each target we generate a local PRF by interpolating between five PRF models to the derived location (Bryson *et al.* 2010). We then render the modelled PRF onto the pixels that comprise the target star, which tells us the fractional contribution of each pixel to the total flux from that star. For each pixel we then subtract that fraction of the total number of photoelectrons required to be subtracted for that observation. We subtract from the flux instead of scaling the flux because individual pixels can have flux contributions from multiple targets, and we only want to reduce the flux from the target of interest. This is particularly important for the false positive tests based on the change in the location of the centre of the flux in and out of transit—scaling the flux would not preserve the spatial information in the distribution of flux in the local scene. In this initial experiment, we inject the transits on the location of the target star, so do not expect flux centroid offsets.

2.3. *Pipeline processing*

For an overview of the *Kepler* science pipeline, see Jenkins *et al.* 2010. The modified pixels are processed through the pipeline as normal. The only departure from standard operations is that, like the position polynomials, the cotrending basis vectors (CBVs) used in the Presearch Data Conditioning (PDC) module for systematic correction are generated from a 'clean' pipeline run. Again, this is to avoid corruption of the CBVs from the presence of many transits in every light curve, since the CBVs are generated from the data themselves. Of course, in reality there are some number of real transit events in a fraction of the light curves, however they should be stochastically scattered throughout and significantly outnumbered by light curves without transit events. In general, CBVs should not be affected by the presence of transits in a fraction of the light curves, although those transits can be affected by the CBVs themselves during correction.

In summary, the final order of processing is that we run the calibrated pixels (the output of CAL) of Quarter 3 through the Photometry Analysis (PA), Presearch Data Conditioning (PDC), and Transiting Planet Search (TPS) pipeline modules, without any modification, to generate the position polynomials, the cotrending basis vectors, and the rms CDPP for each target. We then inject the transit signals into the calibrated pixels, one planet for every target in the full focal plane (80 channels), and re-run the modified pixels through PA, PDC and TPS, utilising the previously generated information as described.

https://doi.org/10.1017/S174392131301260X Published online by Cambridge University Press

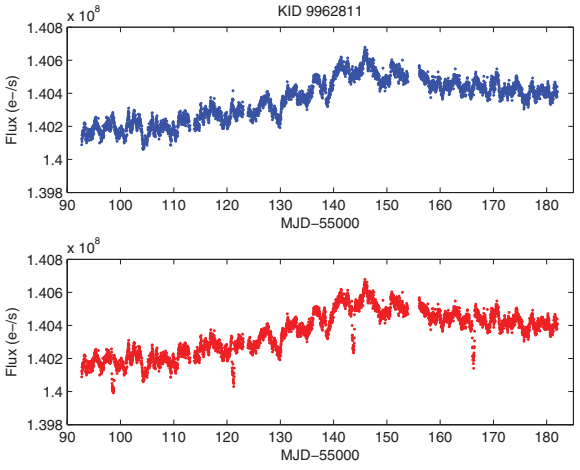

Figure 1. The simple aperture photometry flux time series created by summing the calibrated pixels in the optimal aperture of KID 9962811 before (upper panel) and after (lower panel) injecting a simulated planet transit signal.

3. Results

Here we show the results for a single channel, Channel 30, which had 1801 exoplanet targets in Quarter 3. Figure 1 shows an example generated light curve: the upper panel is the original simple aperture photometry flux time series of target KID 9962811, and the lower panel is the flux time series after injection of the simulated transiting planet. When examining the results after the pipeline processing, we noticed two aspects of the pipeline correction that were quite sensitive to the artificially close spacing of the transits: the first is that PDC treats targets that it identifies as variable in a different manner to those that do not pass the variability threshold, and the injection of many transits into a given light curve can push the measured variability of that light curve over the threshold. This results in a different PDC treatment and a significant systematic change in the recovered signal strength. The second is that the Sudden Pixel Sensitivity Dropout (SPSD) detector in PDC does not attempt to correct SPSDs if many are detected in a given light curve, which can occasionally be the case for a large number of injected transits. Since in this experiment we are concerned with the signal throughput efficiency of single events, we filter out these light curves where large scale changes in the pipeline treatment of the light curve distorted the results; we are planning another experiment with realistic spacing where we can assess the impact of the two effects listed on real signals. This leaves 1680 of the 1801 targets for analysing the signal preservation.

In Figure 2, we plot the SNR of the original input transit signal against the SNR measured by the pipeline after being processed through PA, PDC and TPS for those 1610 targets. Each cadence that was impacted by the injected transits is plotted here; all cadences associated with a given target are the same colour. A robust fit to the data points gives a slope of 99.6±0.2%, which indicates a very high fidelity between the original and processed signals. As noted earlier, although we inject the simulated transit signals with SNR between 2 and 20σ, they are injected into the local noise and artifacts (and potentially astrophysical signals) already in the light curves, and hence the actual injected SNR varies over a wider range.

https://doi.org/10.1017/S174392131301260X Published online by Cambridge University Press

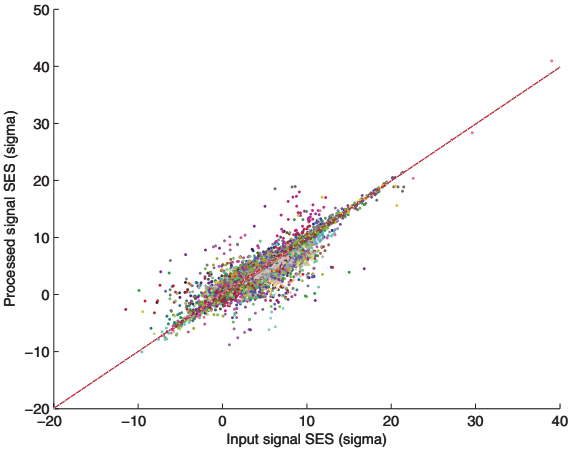

Figure 2. Comparing the initial input transit SNR to the final measured transit SNR for 1610 targets a single channel. A robust linear fit (the red dashed line) to the data gives a signal preservation rate of 99.6±0.2% of the original signal strength.

4. Discussion and Future Plans

This is an encouraging result for the validity of planet population models which have thus far been based on the theoretical SNR of putative planets in the *Kepler* field. We are currently analysing the full set of 84 channels, and anticipate similar levels of signal preservation; the full set of results will be presented by Christiansen *et al.* (in prep). We plan to run a series of tests, including realistic spacing of transits, multiple quarters of data, and centroid offsets to mimic background false positives, in order to examine further aspects of the pipeline throughput efficiency. Another important aspect of signal recovery, that of real signals being masked by stronger artifacts in the data, is something we are addressing in the latest version of the pipeline. For each light curve, we now search iteratively down to the detection threshold (7.1σ), rejecting systematics that do not pass the validation tests and then re-searching the light curve, in order to reliably detect any valid transit signals that pass the threshold.

References

Batalha, N. B., *et al.* 2013, *ApJS*, 204, article id. 24
Borucki, W. J., *et al.* 2011b, *ApJ*, 728, 117
Borucki, W. J., *et al.* 2011a, *ApJ*, 736, 19
Brown, T. M., *et al.* 2011, *AJ*, 142, 112
Bryson, S. T., *et al.* 2010, *ApJ*, 713, 97
Christiansen, J. L., *et al.* 2012, *PASP*, 124, 1279
Fischer, D. A., *et al.* 2012, *MNRAS*, 419, 2900
Howard, A. W., *et al.* 2012, *ApJS*, 201, 15
Jenkins, J. M., *et al.* 2010, *ApJL*, 713, L87
Huang, X., Bakos, G. A., & Hartman, J. D. 2013, *MNRAS*, 429, 2001
Mandel, K. & Agol, E. 2002, *ApJ*, 580, 171
Ofir, A. & Dreizler, S. 2013, *A&A*, 555, article id.A58
Twicken, J. D., Clarke, B. D., Bryson, S. T., Tenenbaum, P., Wu, H., Jenkins, J. M., Girouard, F., & Klaus, T. C. 2010, *ProcSPIE*, 7740, 774023
Youdin, A. N. 2011, *ApJ*, 742, 38

https://doi.org/10.1017/S174392131301260X Published online by Cambridge University Press