


Planet Reliability Metrics: Astrophysical Positional Probabilities

KSCI-19092-001

Stephen T. Bryson and Timothy D. Morton

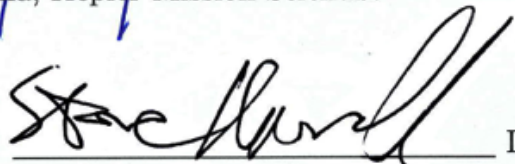
July 21, 2015

NASA Ames Research Center
Moffett Field, CA 94035

Prepared by:  Date: 7/21/15
Stephen T. Bryson, Kepler Science Office

Approved by:  Date: 7/21/15
Michael R. Haas, Kepler Science Office Director

Approved by:  Date: 7/21/15
Natalie Batalha, Kepler Mission Scientist

Approved by:  Date: 7/21/15
Steve B. Howell, Kepler Project Scientist

Document Control

Ownership

This document is part of the Kepler Project Documentation that is controlled by the Kepler Project Office, NASA/Ames Research Center, Moffett Field, California.

Control Level

This document will be controlled under KPO @ Ames Configuration Management system. Changes to this document **shall** be controlled.

Physical Location

The physical location of this document will be in the KPO @ Ames Data Center.

Distribution Requests

To be placed on the distribution list for additional revisions of this document, please address your request to the Kepler Science Office:

Michael R. Haas
Kepler Science Office Director
MS 244-30
NASA Ames Research Center
Moffett Field, CA 94035-1000
Michael.R.Haas@nasa.gov

The correct citation for this document is: S. T. Bryson and T. Morton, 2015, *Planet Reliability Metrics: Astrophysical Positional Probabilities*, KSCI-19092-001

Contents

1	Introduction	5
2	A Probabilistic Approach to Background False Positive Identification	8
3	From Likelihood to Probability	10
4	Implementation	11
4.1	Modeling the Transit Signal on a Known Star	11
4.1.1	Measuring the Location of a Transit Source in the <i>Kepler</i> Pipeline . . .	11
4.1.2	Modeling Transit Sources	12
4.1.3	Computing the Likelihood for Each Star	13
4.2	The Background Likelihood	15
5	Results	16
5.1	Host Star Relative Probability Quality	16
5.2	Host Star <i>a priori</i> Probability	17
5.3	Examples	19

1 Introduction

This document describes the *Kepler* astrophysical positional probabilities (APP) table hosted at the Exoplanet Archive¹. This table lists the stars with the highest probability of being co-located with the source of an observed transit, as well as the probability of the transit being on an unknown background source. The position of the transit signal source relative to the target star is found in the KOI tables at the Exoplanet Archive. These probabilities measure how well a star’s location matches the location of the transit signal – they do not measure the probability that the transit signal is consistent with a planet orbiting that star.

For each star known to be near a *Kepler* Object of Interest (KOI), we compute the relative probability that the star is co-located with the transit source on the sky. We also compute the probability that the transit source is due to an unknown background source, relative to the probability for the known stars. These probabilities are relative in the sense that if one star has twice the probability of another, then the first star is twice as likely to be co-located with the transit source. Or if the probability for a star is twice the probability of the background, then that star is twice as likely to be the source of the transit as an unknown background source.

For a particular KOI, the relative probability is computed for stars from the catalogs described in §5 that fall on the pixels associated with that KOI. The APP table reports those probabilities for the KOI host, for the two stars with the highest probability, one of which may be the host star, and the background probability.

The relative probability that the transit signal source is co-located with the KOI host star is of particular interest for exoplanet statistical studies. Derived planet properties depend critically on the details of the star that the planet orbits, and the planet properties reported in the *Kepler* planet candidate tables assume that the planet orbits the KOI host. So the probability that the transit signal is co-located with the host star, given in the APP table, provides a measure of the reliability that the derived planet properties are correct. For example, the probability that a KOI’s transit signal is a planet on the KOI host star can be computed as $P(\text{planet orbiting host}) = P(\text{planetary transit}) P(\text{transit at location of host})$, where $P(\text{planetary transit})$ is the probability that the transit signal is consistent with a planetary transit (such as the FPP table at the Exoplanet Archive) (Morton, 2012), and $P(\text{transit at location of host})$ is the relative probability of co-location with the host star given in the APP table.

The relative probabilities for some KOIs are more reliable than those for other KOIs, and some KOIs will not have computed probabilities. These probabilities are computed using the results of centroid analysis of *Kepler* data as described in §4.1.1. The quality of the computed probabilities depends on the quality of the centroid data, and when the data is of low quality, for example when the transit S/N is very low, the resulting probabilities may be unreliable. The APP table provides a metric measuring the quality of the host star probability computation, which can be used to exclude unreliable probability computations. For many KOIs these centroids are unavailable or are known to be invalid, such as when

¹<http://exoplanetarchive.ipac.caltech.edu>

the KOI host star is saturated or highly crowded. When the centroids are unavailable or invalid the probabilities are not computed, and the probability computation is declared to have failed.

The ability of the relative probabilities to distinguish between two known stars is determined by the accuracy of the underlying centroid data, which is in turn driven by the transit S/N. In some cases a star will have a relative probability near one while another star 1 arcsec away will have a probability near zero. In other cases, where the centroid measurements have lower spatial precision, the stars must be several arcsec apart to have different probabilities. The smallest distance that can be distinguished is 0.2 arcsec due to an observed centroid noise floor (see §4.1.3).

A simple alternative *a priori* probability of the transit signal being co-located with the target star is provided in the APP table for use in computing statistics when the probability computation is low quality or has failed completely. This *a priori* probability is simply the fraction of KOIs whose transit signals are known to be offset from the KOI host star, which is a strong function of Galactic latitude (see §5.2). This *a priori* probability should be used statistically, and should not be applied to the analysis of individual KOIs.

When the relative probability computation fails only the fields Kepler ID, KOI name and *a priori* probability are set.

The APP table has the following structure. Archive variable names are given in parentheses.

- **Identification parameters:**

- **Kepler ID** (*kepid*) of the KOI host star.
- **KOI name** (*kepoi_name*) of the transit being analyzed.
- **Period** (*pp_koi_period*) used in computing the relative probabilities. Not set when the probability computation fails.
- **Epoch** (*pp_koi_time0bk*) used in computing the relative probabilities. Not set when the probability computation fails.
- **Transit depth** (*pp_koi_depth*) in ppm used in computing the relative probabilities. Not set when the probability computation fails.

- **Host star probabilities:**

- **Relative probability that the transit is co-located with the KOI host star** (*pp_host_rel_prob*), or not set when the relative probability computation fails.
- ***A priori* probability that the transit is co-located with the KOI host star** (*pp_host_prior_prob*), to be used when the relative probability computation fails. This is computed for all KOIs.
- **Host star relative probability source flag** (*pp_host_prior_prov*). This flag takes one of the following values:
 - * **PROB:** The relative probability score is computed using the method described in this document.

- * **MATCH:** The host star relative probability is set to zero because this KOI has been determined to be a period-epoch match and the parent is not among the known stars used in the relative probability computation (Coughlin et al., 2014).
 - * **FPWG:** The relative probability computation failed but the host star relative probability is set to zero because the *Kepler* False Positive Working Group has examined this object and determined that the transit source is not co-located with the KOI host star.
 - * **FAILED:** The relative probability computation failed and there is no alternative source of relative probabilities.
- **Relative probability quality** ($pp_host_prior_score$). This quality value ranges from 0 to 1, and a value below about 0.3 indicates that the relative probability computation is likely to be untrustworthy.
- **The Two Highest Relative Probability Stars:** The two known stars with the highest probability of being co-located with the transit signal source. There may be only one known star considered in the probability computation. Several parameters are given for each star:
 - **Identifier of the star** (pp_1hi_starid , pp_2hi_starid). This may be a catalog number or a reference flag
 - **Star right ascension** (pp_1hi_ra , pp_2hi_ra) in degrees
 - **Star declination** (pp_1hi_dec , pp_2hi_dec) in degrees
 - **Star *Kepler* magnitude** (pp_1hi_kepmag , pp_2hi_kepmag). Depending on the source of information for this star, the *Kepler* magnitude uncertainty may be as large as two magnitudes.
 - **Relative probability that the transit is co-located with this star** ($pp_1hi_rel_prob$, $pp_2hi_rel_prob$).
 - **The modeled transit depth** ($pp_1hi_mod_depth$, $pp_2hi_mod_depth$) in parts per million that best reproduces the observed transit depth used in the relative probability computation described in §4.1.2. We do not expect this modeled depth to be accurate, but we provide it because it can indicate possible planetary-size transiting objects on stars other than the host star. This modeled depth depends critically on the accuracy of the catalog used to model the flux around the KOI host star and may be significantly in error. Therefore depths as large as 3 million ppm are reported, so the modeled star may have to contribute negative flux in order to reproduce the observed depth. Stars whose modeled depths are greater than 3 million ppm are rejected from consideration by the relative probability computation. The probability computation is itself relatively insensitive to such large errors.
 - **Star provenance flag** ($pp_1hi_prob_prov$, $pp_2hi_prob_prov$) indicates the source of the position and magnitude information for this star.

- **Background Relative Probability** (*pp_unk_rel_prob*): The relative probability that the transit source is co-located with an unknown background object rather than a known star, described in §4.2.
- **Background Density** (*pp_bkgd_density*): The modeled density of background sources from Morton and Johnson (2011) used in the computation of the background relative probability.
- **DV Run Identifier** (*pp_dv_run_id*): The identifier of the *Kepler* pipeline data validation (DV) run that produced the centroid analysis used in the relative probability computation.

The stars considered in the relative probability computation are from several sources. In the initial APP table release these sources are the Kepler input catalog (Brown et al., 2011) and the UKIRT catalog (Lawrence et al., 2007). The source for a particular star is denoted by “KIC” or “UKIRT” in their ID, followed by the catalog identifier (called “source ID” in the UKIRT catalog)².

Every KOI in the APP table that has a successful relative probability computation (Host star relative probability source flag = PROB) has an APP report giving a table of all stars that are considered in the relative probability computation. Each report also has a figure per star, showing its position relative to the target star and, when possible, contours showing the observed and modeled position distributions described in §4.1.3. These reports are linked from the KOI’s entry in the APP table.

The rest of this document describes the computation of the astrophysical position probabilities. §2 introduces and motivates the basic approach. In §3 we describe how probability is derived from likelihood via Bayesian hypothesis testing, and §4 describes the implementation. Specifically, §4.1.1 summarizes how transit locations are computed from the PRF-fit technique. §4.1.2 describes how transits are modeled on each known star. §4.1.3 derives the likelihoods from the modeled and observed data via smooth bootstrap techniques, with mathematical details given in appendix A. The likelihood of the background is treated in §4.2. Results are given in §5, starting with a discussion of where the probability computation is unreliable. §5.3 illustrates the current method with a few examples.

2 A Probabilistic Approach to Background False Positive Identification

The *Kepler* Mission detects transiting exoplanets as well as background false positives that are observationally separated from the target star (Koch et al., 2010). The most common method for identifying background false positives is deriving the transit source location from various centroid techniques, and flagging a KOI as a false positive if its transit source location is more than 3σ from the target star (Bryson et al., 2013). Using the 3σ threshold makes

²For instructions accessing the UKIRT catalog, see <http://keplergo.arc.nasa.gov/ToolsUKIRT.shtml>.

it very unlikely that a transit signal on the target star will be misidentified as being on a background source. This 3σ threshold has been used to identify offset false positives in the KOI tables at the Exoplanet Archive (Burke et al., 2014; Mullally et al., 2015; Coughlin et al., 2015). This threshold is, however, somewhat crude and has the following weaknesses:

- Methods that measure centroids are subject to unknown systematic biases.
- When there are one or more known field stars within 3σ of the target star, this threshold provides no information about whether these field stars are consistent with the data.
- The reliability of the claim that the transit signal is on the target star is interestingly different when the measured signal source position is, for example, 1σ vs. 2.8σ from the target star, but both cases pass the threshold.
- The rate of background binaries depends strongly on Galactic latitude (Bryson et al., 2013), but this is not reflected in the threshold.

Generally speaking, there is more information available about the location of a transit signal than a single position and a 3σ circle. This paper presents an analysis of the position of transit signals measured by *Kepler* that uses this additional information to compute the probability that the transit source is on a known star. These probabilities often provide more insight than the 3σ approach.

Several methods are used by the *Kepler* Mission to identify background false positives by determining that the observed position of the transit signal is not consistent with the target star position, as described in Bryson et al. (2013). In this paper we concentrate on the PRF-fit difference image technique (summarized in §4.1), which is the most robust and provides the highest precision. This method measures the position of a transit signal relative to the target star for each *Kepler* observational quarter via averaging pixel flux values across all transits in that quarter. These measurements are subject to various systematic errors, in addition to photometric shot noise, which results in quarter-to-quarter variations of the measured centroid position. Averaging individual transits within a quarter is possible because *Kepler's* exceptional pointing stability means that each transit's flux variations stay on the same pixels. Across quarters, however, the stars fall on different pixels, preventing averaging at the pixel level.

The offset of the transit signal from the target star is estimated via an average across quarters of each quarter's offset measurement. Bryson et al. (2013) describes how this average is computed as a χ^2 minimizing fit. This average, however, may not represent the true location of the transit source. Quarter-to-quarter systematics produce scatter in the quarterly position measurements, which can be thought of as a sampling of an unknown distribution of positions. The best estimate of the transit source location is given by the average of this unknown underlying distribution. The statistical bootstrap is an effective method of producing a *distribution of averages* of the measured quarterly transit location relative to the target star. The traditional bootstrap method provides the distribution of averages as a set of discrete points. We convert these points to a continuous distribution with the *smooth bootstrap* technique, which uses kernel density estimation techniques. We produce a

continuous distribution of average *observed* positions $D_o(x, y)$ for the transit location and, via modeling of the transit on each star s , continuous *modeled* distributions $D_s(x, y)$. The models are based on stellar catalogs, observed transit parameters, the effective PSF of the Kepler instrument, and known noise sources as described in §4.1. We consider the degree of overlap of these distributions as a measure of the likelihood that a transit on star s is consistent with the observed transit location. We define the likelihood that the transit is on star s as the integral over the product of the distributions: $L_s = \int D_o(x, y) D_s(x, y) dx dy$. Because D_o and D_s are densities L_s has units of per square arcsecond.

We compare the likelihood L_s that the transit is on star s with the likelihood L_t that the transit is on star t by computing the ratio $H_{st} = \frac{L_s}{L_t}$. All stars known to fall on the pixels collected for the target star are considered, as well as an unknown background. Similar to the treatment in Gregory (2010), we convert the hypothesis ratios H_{st} into probabilities in §3.

This modeling approach addresses the above described weaknesses of the 3σ threshold approach in several ways:

- Systematic crowding bias is accounted for, so long as that crowding is due to known stars.
- A continuous probability estimate more clearly describes borderline cases such as when there are stars within 3σ of the target star.
- The background binary density is accounted for so, for example, KOIs at low Galactic latitude are more likely to be due to background objects.

3 From Likelihood to Probability

Hypothesis testing considers the ratio $H_{st} = \frac{L_s}{L_t}$, where L_s was defined in §2. Hypothesis s is considered more likely than hypothesis t if $H_{st} > 1$.

The hypotheses ratios satisfy $H_{st} = H_{ts}^{-1}$, so there is a large amount of redundancy among the various H_{st} . In particular, thinking of H_{st} as a matrix for bookkeeping purposes, any element can be expressed in terms of the elements of a specified column. For example, we can express any of the H_{st} in terms of the first column H_{s1} :

$$H_{st} = \frac{L_s}{L_t} = \frac{L_s L_1}{L_1 L_t} = H_{s1} H_{1t} = \frac{H_{s1}}{H_{t1}}.$$

We eliminate this redundancy and convert each column into a set of relative probabilities by normalizing each column by its sum: for each column t ,

$$H_{st} \rightarrow \hat{H}_{st} \equiv \frac{H_{st}}{\sum_w H_{wt}} = \frac{L_s}{L_t} \frac{1}{\sum_w \frac{L_w}{L_t}} = \frac{L_s}{\sum_w L_w}. \quad (1)$$

These normalized hypothesis ratios \hat{H}_{st} are independent of column: $\hat{H}_{st} = \hat{H}_{sw}$ for any t and w , and $\sum_i \hat{H}_{st} = 1$. So we can define the *probability of hypothesis s relative to the other*

hypotheses as $R_s = \frac{L_s}{\sum_w L_w}$. R_w can be interpreted as a probability because $0 \leq R_s \leq 1$ and $\sum_s R_s = 1$.

4 Implementation

4.1 Modeling the Transit Signal on a Known Star

4.1.1 Measuring the Location of a Transit Source in the *Kepler* Pipeline

To set the context for how the transit signals are modeled and the resulting positions are measured, we briefly summarize how observed transit signal positions are measured relative to the target star using the PRF-fit difference image method. For details see Bryson et al. (2013).

An observed transit signal associated with a target star is identified and characterized from the flux light curve obtained by summing the pixels in an optimal photometric aperture around that target star (Jenkins et al., 2010). This photometric aperture is a subset of a larger pixel mask collected for each target star (Bryson et al., 2010b). For each quarter that contains transits, the in-transit cadences are identified. All pixels associated with this target star are then averaged over the in-transit cadences, creating the average *in-transit image*. Cadences on either side of each transit in a quarter are similarly used to create an average *out-of-transit image*. Subtracting the in-transit image from the out-of-transit image creates the *difference image* for each quarter. Assuming that the transit signal is the only source of flux variation between the in- and out-of-transit images, the difference image provides a direct image of the transit source. The location of the transit source is measured by fitting a Pixel Response Function (PRF) model (Bryson et al., 2010a) to the difference image. Given a star’s position and magnitude, the PRF provides the flux in each pixel due to that star, so a PRF fit is the determination of the star position for which the PRF-modeled flux distribution best matches the pixel values. The PRF fit provides a formal propagated uncertainty based on the input pixel value uncertainties. These quarterly uncertainties do not include the systematics described below.

The quarterly offset of the transit source from the target star is obtained by subtracting the position of the target star from the position of the transit source obtained from the PRF-fit position to the difference image. The uncertainty of this offset is computed via standard propagation of errors. The position of the target star is obtained from either the catalog position of the target star or from a similar PRF fit to the out-of-transit image. Using the PRF fit to the out-of-transit image is preferred because systematic PRF fit error due to inaccuracies in the PRF model are largely common to both the difference image and out-of-transit fits and therefore cancel out. Using the PRF fit to the out-of-transit image assumes, however, that the target star is well-isolated so that this fit position gives the actual position of the target star. Crowding due to background stars will introduce a bias into the out-of-transit PRF fit position that is typically not present in the difference image. This introduces a crowding bias into the measurement of the transit source offset relative to the target star. One of the motivations for the work in this paper is to estimate this crowding

bias for each target star. In extreme cases when there is another star of comparable or greater brightness than the target star, the PRF fit will often give the position of the nearby star. In this case the out-of-transit PRF fit will be invalid. We do not include target stars in the analysis described in this paper when the out-of-transit PRF fit gives a target star location that is more than two arc seconds from the target star’s catalog location.

The PRF fit to the difference image is subject to various other systematics (Christiansen et al., 2012), particularly due to flux variations other than the transit source, which introduce noise into the difference image. The result is that the quarterly offsets of the transit signal location relative to the target star will have some scattered distribution. While this scatter is statistically near-Gaussian (in particular it is zero-mean) when averaged over all targets, it may be far from Gaussian for specific target stars.

4.1.2 Modeling Transit Sources

For each target star, in each quarter in which transits occur, we create a synthetic out-of-transit image for the pixels in that target star’s pixel mask using techniques similar to those described in Bryson et al. (2010b). Specifically, stellar catalogs and the *Kepler* PRF model (Bryson et al., 2010a) are used to add the flux from each star in or near the mask to the pixels, scaled by that star’s catalog flux. The specific catalog used depends on the star and is identified in each target star’s APP report.

The uncertainty σ_i^{OOT} of each out-of-transit pixel i is taken to be the observed uncertainty σ_i^{obs} computed by the *Kepler* pipeline. To compute the uncertainty of the difference image pixels, we estimate the non-photometric component of the σ_i^{OOT} by subtracting in quadrature the out-of-transit image’s photometric uncertainty $\sigma_i^{\text{phot,OOT}}$ from the observed pixel uncertainty σ_i^{obs} : $\sigma_i^{\text{pix}} = \sqrt{(\sigma_i^{\text{obs}})^2 - (\sigma_i^{\text{phot,OOT}})^2}$. Here $\sigma_i^{\text{phot,OOT}} = \sqrt{f_i^{\text{OOT}}}/\sqrt{N}$ is the Poisson photon noise for each pixel with flux f_i , scaled by the square root of the number of out-of-transit cadences N .

For each star s in the target star’s pixel mask, an in-transit image is created by the same method as for the out-of-transit image, but with the flux of star s scaled by $(1 - d_s)$ where d_s is the fractional depth of the simulated transit/eclipse on that star. The simulated depth d_s is set separately in each quarter to match the average observed depth d^{obs} in the target star’s optimal aperture for that quarter. This observed depth is not corrected for dilution by other flux in the aperture. The simulated depth d_s is seeded with the estimate $d^{\text{obs}} f_{\text{target}}/\sum_s f_s$ where f_{target} is the flux of the target star and f_s is the flux of star s (the target star is included in the sum over s). When the seed estimate of d_s is less than one, the d_s that matches d^{obs} is computed via a nonlinear Levenberg-Marquardt fit (Levenberg, 1944; Marquardt, 1963). When the seed estimate of d_s is greater than one, d_s is set equal to the seed value. This simulated depth is strongly dependent on the accuracy of the *Kepler* magnitudes of the stars used to model the flux on the pixels and can be inaccurate. We conservatively allow the depth to be as great as 3 (implying a reduction in the flux of the star by 300%). Stars whose modeled depth is greater than 3 are not considered in the probability computation and should be considered to have probability zero. We do not simulate transits on stars

outside the pixel mask because centroid measurements for such stars are unreliable and can be very misleading.

The uncertainty of the modeled difference image pixels is estimated as $\sqrt{(\sigma_i^{\text{OOT}})^2 + (\sigma_i^{\text{IT}})^2}$, where $\sigma_i^{\text{IT}} = \sqrt{(\sigma_i^{\text{phot,IT}})^2 + (\sigma_i^{\text{pix}})^2}$ is the estimated uncertainty of the in-transit image. Here $\sigma_i^{\text{phot,IT}} = \sqrt{f_i^{\text{IT}}}/\sqrt{N}$ is the Poisson photon noise for each in-transit image pixel with flux f_i^{diff} , scaled by the square root of the number of out-of-transit cadences N , and σ_i^{pix} is that pixel's non-photometric noise estimated from the out-of-transit image as described above.

The result is that for each quarter we have a simulated out-of-transit image and a collection of in-transit images, with each difference image modeling the transit on a different known star in the target star's pixel mask. Each of these images have associated estimated pixel uncertainties. The reader may ask why we did not build the simulated in-transit image from the observed out-of-transit image by injecting transit signals in the out-of-transit pixels. Simulated transits are placed on known stars at their catalog positions, and these catalog positions often disagree with the actual star positions due to, *e. g.*, catalog error and proper motion. Constructing the out-of- and in-transit images using the same catalog positions guarantees that the difference image is consistent with the out-of-transit image, eliminating the possibility of introducing biases in the modeling.

Similar to the observational data described in §4.1.1, a PRF fit is performed on the modeled out-of-transit image in each quarter q in which there is a transit, and, for each star s , on the modeled in-transit images. Taking the quarterly difference between the PRF fits of the in- and out-of-transit PRF fits gives us observed offsets $\Delta_{o,q} = (\Delta_{\text{RA}}, \Delta_{\text{DEC}})_{o,q}$ with covariance matrix $\tilde{\Sigma}_o$, and modeled offsets $\Delta_{s,q} = (\Delta_{\text{RA}}, \Delta_{\text{DEC}})_{s,q}$ with covariance matrix $\tilde{\Sigma}_s$ for each star s in the target star's pixel mask. In the next section we use these quarterly offsets to estimate distributions of the average observed and modeled offsets.

4.1.3 Computing the Likelihood for Each Star

The observed offsets of a transit signal location from the target star can be thought of as the sampling of an unknown distribution of offsets. The average of these observed offsets gives an estimate of the transit source location relative to the target star. If the transits were observed at a different time, we would get a different sampling of the unknown underlying distribution, with a different estimate of the transit source location. In this section we construct a distribution of these average offsets for both the observed and modeled transits. The overlap of the observed modeled distributions, defined as the integral of the product of the distributions, is the likelihood that the modeled star is the source of the transit.

For each collection of observed quarterly offsets $(\Delta_{\text{RA}}, \Delta_{\text{DEC}})_{o,q}$ and modeled quarterly offsets $(\Delta_{\text{RA}}, \Delta_{\text{DEC}})_{s,q}$ with their associated uncertainties, we construct a continuous distribution of mean average offsets using the *smooth bootstrap* technique. The smooth bootstrap starts with a conventional ensemble of bootstrap averages, and replaces each average value with a Gaussian distribution, similar to kernel density estimation (Silverman, 1986). Gaus-

sian distributions are very convenient because products of Gaussians are Gaussians, and Gaussians lend themselves to explicit integration.

Given Q quarterly offsets $\Delta_{o,q}$ or $\Delta_{s,q}$, where Q is the number of quarters with a transit for the target, the bootstrap method generates an ensemble of N resampled offset sets, each of length Q , via resampling with replacement. We set $N = 500$ unless $Q < 5$, in which case we include every permutation of the data including repetitions (so $N = Q^Q$). Because $Q \leq 17$ this provides a sufficient sampling for the bootstrap estimate. The set of averages \mathbf{b}_k , $k = 1 \dots N$, of each resampled set returned by the bootstrap method provides N average offsets. Each \mathbf{b}_k is a two-dimensional vector with components giving the average RA and Dec offsets for each resampling. We denote the ensemble based on the observed offsets as $\mathbf{b}_{o,k}$ and those based on the modeled offsets for each star s as $\mathbf{b}_{s,k}$.

We smooth the bootstrap ensemble using a normalized two-dimensional Gaussian for each bootstrap average \mathbf{b}_k

$$G(\mathbf{x}, \mathbf{b}_k, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{b}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{b}_k)\right] \quad (2)$$

where \mathbf{x} and \mathbf{b}_k are two-dimensional vectors (RA and Dec offsets in our case), and $\boldsymbol{\Sigma}$ is the covariance matrix that determines the smoothing, based on the covariance matrix of the original offsets $\tilde{\boldsymbol{\Sigma}}_o$ or $\tilde{\boldsymbol{\Sigma}}_s$ depending on whether we are smoothing the observed or modeled distribution. We choose a two-dimensional generalization of Scott's rule-of-thumb (Scott, 1992):

$$\boldsymbol{\Sigma}_o = N^{-\frac{2}{d+4}} \tilde{\boldsymbol{\Sigma}}_o, \quad \boldsymbol{\Sigma}_s = N^{-\frac{2}{d+4}} \tilde{\boldsymbol{\Sigma}}_s \quad (3)$$

where N is the number of averages in the bootstrap ensemble \mathbf{b}_k and, in our case, $d = 2$. To account for an observed small residual bias in centroid offsets described in Bryson et al. (2013), a noise floor term of $(0.2/3 \text{ arcsec})^2$ is added to the diagonal terms of $\boldsymbol{\Sigma}_o$ and $\boldsymbol{\Sigma}_s$. This imposes a minimum size on the bootstrap distributions.

We define our smooth bootstrap distributions of the observed and modeled average offsets as

$$D_o(\mathbf{x}) = \frac{1}{N} \sum_{k=1}^N G(\mathbf{x}, \mathbf{b}_{o,k}, \boldsymbol{\Sigma}_o), \quad D_s(\mathbf{x}) = \frac{1}{N} \sum_{k=1}^N G(\mathbf{x}, \mathbf{b}_{s,k}, \boldsymbol{\Sigma}_s). \quad (4)$$

With this definition $\int D_o(\mathbf{x}) d\mathbf{x} = \int D_s(\mathbf{x}) d\mathbf{x} = 1$, where $d\mathbf{x}$ is the two-dimensional RA and Dec area element.

We show in appendix A that our desired likelihood for the star s based on the smooth bootstrap technique is given by

$$L_s = \int D_o(\mathbf{x}) D_s(\mathbf{x}) d\mathbf{x} = \frac{1}{N^2} \sum_{k=1}^N \sum_{j=1}^N G(\mathbf{b}_{o,j}, \mathbf{b}_{s,k}, \boldsymbol{\Sigma}_o + \boldsymbol{\Sigma}_s) \quad (5)$$

From these likelihoods we compute the probability that the star is in the same sky position as the transit source relative to other known stars as $R_s = \frac{L_s}{\sum_w L_w}$, where the sum includes the likelihood of the background described in the next section. When $R_s < 10^{-30}$, we set $R_s = 0$ in the table.

4.2 The Background Likelihood

We model the hypothesis that the transit signal is due to an eclipse on an unknown background binary star using the background model of Morton and Johnson (2011). This model varies with target star *Kepler* magnitude and Galactic latitude. The *Kepler* magnitude dependence is due to this model's requirement that the background binaries produce a detectable transit-like signal when diluted by the target star. For a specific target star's pixel aperture this model is sufficiently slowly varying that we can take it as locally constant. We define b as the model background binary density per square arcsecond, evaluated for the target star.

To estimate b we use Equation (14) and Table 1 from Morton and Johnson (2011). We note that this table as published has the column values reversed: the absolute values of c_0 should be the largest and those of c_4 should be the smallest. For the targets modeled in this paper, b ranges from 1.6×10^{-6} to 5.1×10^{-5} background binaries per square arcsecond, with a median of 1.1×10^{-5} .

Given b , in order to satisfy the requirement that our distributions are normalized we define the background distribution as

$$D_{\text{bgd}}(\mathbf{x}) = \begin{cases} b, & r \leq \frac{1}{\sqrt{\pi b}} \\ 0, & r > \frac{1}{\sqrt{\pi b}} \end{cases} \quad (6)$$

where $r = \sqrt{x^2 + y^2}$. So $D_{\text{bgd}}(\mathbf{x}) = b$ in a circle of radius $R_0 = 1/\sqrt{\pi b}$, is zero outside this circle, and $\int D_{\text{bgd}}(\mathbf{x}) dx dy = 1$. The smallest radius for this circle occurs when b is largest, where $R_0 = 79$ arcseconds or about 20 *Kepler* pixels. A 20 pixel radius is larger than the largest pixel mask for a non-saturated target star, so this normalization is appropriate for comparison with the normalized Gaussians we use to compute our likelihoods.

The background likelihood is $L_{\text{bgd}} = \int D_{\text{bgd}} D_o d\mathbf{x}$, and because b vanishes outside the circle of radius R_0 , the product in the integrand vanishes as well. This is plausible because for any reasonable measurement of the centroid position, D_o should essentially vanish outside a circle that is considerably smaller than 79 arcseconds: the pixel mask for a 12th magnitude star typically has a radius of about 25 arcseconds. Therefore no information is lost by imposing a background model that vanishes outside a circle of radius ≥ 79 arcseconds. We demonstrate this by assuming that D_o is a Gaussian with a diagonal covariance matrix with equal entries, so D_o becomes a function of r only. Then the likelihood of the background hypothesis is

$$L_{\text{bgd}} = \int D_{\text{bgd}} D_o d\mathbf{x} = \int_0^\infty D_{\text{bgd}} D_o dr = b \int_0^{R_0} D_o dr. \quad (7)$$

Now $\int_0^\infty D_o dr = 1$ and $\int_0^{R_0} D_o dr = \int_0^\infty D_o dr - \int_{R_0}^\infty D_o dr = 1 - \int_{R_0}^\infty D_o dr$, so

$$L_{\text{bgd}} = b \left(1 - \int_{R_0}^\infty D_o dr \right). \quad (8)$$

If we make the conservative assumption that the uncertainty of D_o is 5 arcseconds (see Fig 33 of Bryson et al. (2013)), then $\int_{R_0}^\infty D_o dr \approx 10^{-110}$, which can be neglected. Because

this example is computed based on the highest background density and largest reasonable measurement uncertainty, we can generally take $L_{\text{bgd}} = b$. While this analysis made various simplifying assumptions, a more realistic analysis is not expected to significantly change the results.

5 Results

The purpose of the methods described in this paper is to compute the relative probability that

- the transit signal source is in the same location as the target star
- the transit source is in the same location as a known star other than the target star
- the transit source is in the background population of unknown stars.

The relative probabilities described in this paper have been computed for identified KOIs (Coughlin et al., 2015) that have the following properties:

- The *Kepler* magnitude is dimmer than 10, because PRF fitting breaks down for brighter targets, which are highly saturated.
- The PRF fit to the observed out-of-transit image, which measures the position of the target star (see §4.1.1), is within 2 arcsec of the catalog position of the target star. When this condition is violated, either there is sufficient crowding to invalidate the centroid data or the target star catalog position is incorrect, invalidating the modeling behind the relative probability computation.

KOIs not satisfying these criteria are marked “FAILED” in the APP table and no relative probabilities are available.

For creation of the synthetic scene we use the *Kepler* Input Catalog (KIC) (Brown et al., 2011) supplemented by the UKIRT catalog (Lawrence et al., 2007). The UKIRT catalog was federated by removing stars already in the KIC and estimating *Kepler* magnitudes from the UKIRT J magnitudes assuming all stars are on the main sequence. The errors on the UKIRT *Kepler* magnitudes can be as large as 2 magnitudes.

Future releases of the FPP table will include stars detected via high-resolution imaging such as those found by the Kepler Community Follow-up Observing Program (CFOP)³.

5.1 Host Star Relative Probability Quality

There are several ways in which the computed probabilities can be misleading. We concentrate on the two most common problem cases:

³<https://cfop.ipac.caltech.edu/home/>

- Transits with $S/N < 10$ often do not have enough signal in each pixel of the difference image to produce a reliable PRF fit.
- Crowding by bright field stars can invalidate the probability analysis. Removing target stars that are fit to more than 2 arcsec from their catalog positions removes most of these cases, but bias may remain in some cases.

We indicate the host star relative probability quality using a numerical score that measures the likelihood of both of these problems. This score ranges from 0 to 1, and we recommend trusting the probabilities reported for the target star when this score is about 0.3 or above. This score is the product of two metrics, each of which is normalized to range from 0 to 1:

- **Difference Image Quality** which measures how well the transit signal difference image resembles a star. In each quarter we compute the correlation of the fitted PRF model with the pixel data (Bryson et al., 2013). The difference image quality metric is the number of quarters where the correlation is > 0.7 divided by the total number of quarters in which a transit was observed.
- **Local Crowding** which compares the flux in the target star’s optimal aperture with the flux outside the optimal aperture in the star’s pixel mask. When the flux outside the optimal aperture exceeds that in the optimal aperture by about a factor of two, then the PRF fit to the out-of-transit image is considered unreliable. The local crowding metric uses the inverse of this ratio with a non-linear sigmoid function to produce a crowding score between 0 and 1, with 1 indicating that most of the flux is from the optimal aperture, and 0.5 indicating that about the same amount of flux is inside the optimal aperture as outside. This metric is provisional, and we expect improvements after applying the probability analysis to the results of the *Kepler* transit injection study.

5.2 Host Star *a priori* Probability

When the host star relative probability quality score described in §5.1 is below threshold, there are several alternatives. When possible, the user should examine the transit data to make a determination of the quality of the centroid measurements using the concepts in Bryson et al. (2013). When this is not possible, such as when many targets are being analyzed for a statistical study, or when examination shows that there is essentially no information on the transit signal source location, we recommend using the *a priori* target star probability values supplied in the table. These are based on the observation that the probability that a *Kepler* object of interest is not on the target star depends sensitively on Galactic Latitude (Bryson et al., 2013). Both the observed offset false positive fraction shown in Figure 1 and modeling (Morton and Johnson, 2011) indicate such a dependence on Galactic latitude. This fraction’s dependence on Galactic latitude is relatively insensitive to target star and transit properties. The *a priori* target star probability is a fit to the fraction of KOIs that are on the target star to the total number of KOIs at a given Galactic latitude.

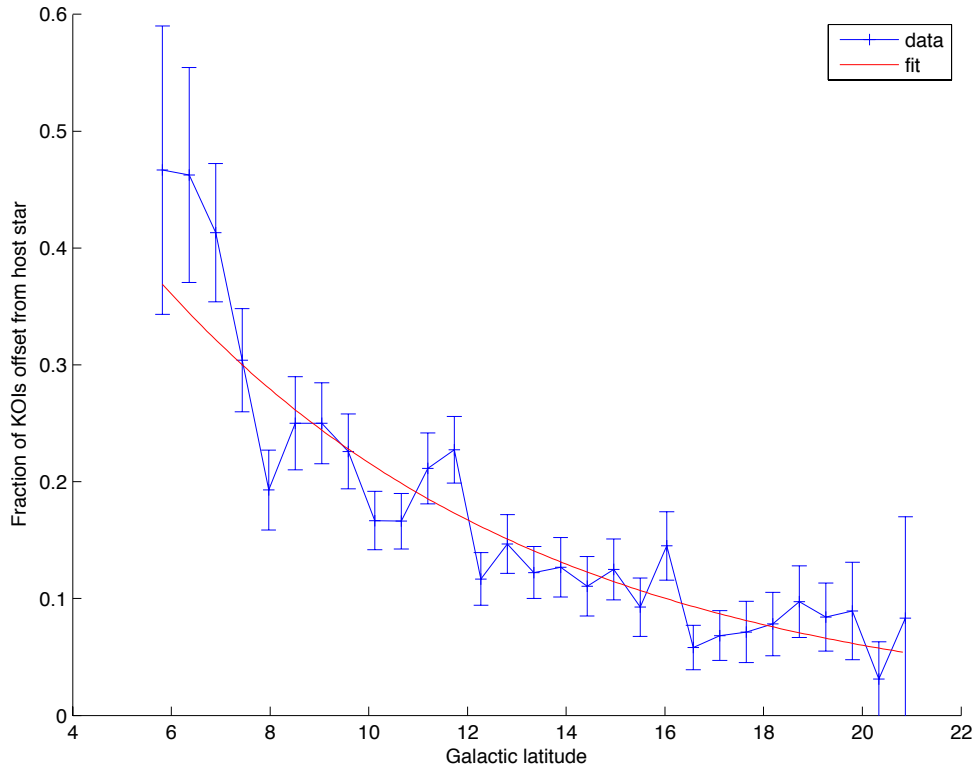


Figure 1: Blue: The distribution of the fraction of KOIs that have been identified as background false positives via offsets from the KOI host star as a function of Galactic Latitude. The error bars show the 1σ Poisson uncertainty. Red: the fit used to compute the KOI host star a priori probability. The fit is $y = 0.775 \times 10^{(-0.0556x)}$ where x is the Galactic latitude in degrees.

5.3 Examples

The confirmed planet Kepler-11c is shown in Figure 2. In this example the observed centroids are clustered around the KOI host star KIC 6541920, so the target star is near the center of the bootstrap distribution of observed averages, shown by the green contours. Modeling the transit on the target star produces a distribution of averages that is also nearly centered on the target star. The overlap of the observed and modeled distributions leads to a large likelihood and a 100% probability that the transit signal is co-located with the target star.

A background false positive, KOI 109.01, associated with a known background star is shown in Figure 3. In this case the green contours, showing the distribution of observed averages, is very near the background star KIC 4752452 and very far from the KOI host star KIC 4752451, indicating that KOI host is unlikely to be the source of the transit signal. The KOI host star has a probability of zero, while there is a 98.9% relative probability that the transit source is at the location of KIC 4752452. Because the green observed contours have only a small overlap with the magenta contours obtained by modeling the transit on KIC 4752452, the background has a probability of 1.1%.

Figure 4 shows the interesting case of KOI-582.01, where an apparent offset in the centroids from the KOI host star turns out to be spurious, caused by centroid bias due to crowding. This bias is revealed by modeling the transit on the KOI host star KIC 9020160. The modeled magenta contours show that the expected distribution of averages is offset from KIC 9020160 in the same direction and distance as the green observed contours. Therefore the original disposition of KOI-582.01 as an offset false positive because it is more than 3σ from the target star (shown by the cyan circle) was incorrect.

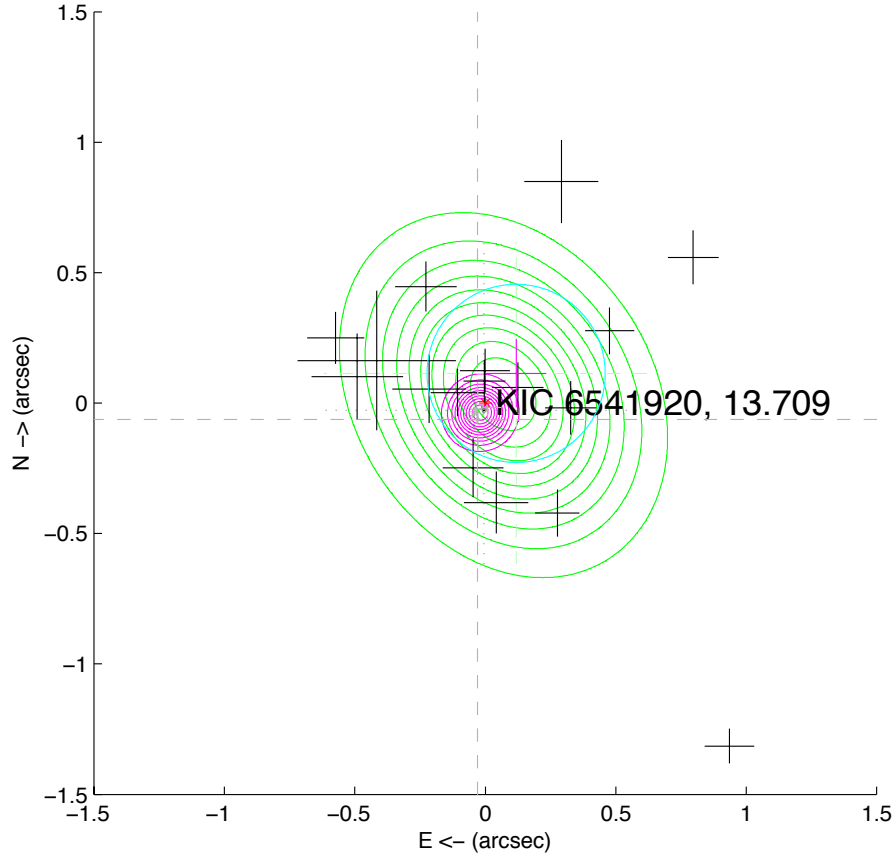


Figure 2: Results of the probability analysis for KOI-157.01 (confirmed planet Kepler-11c) with target star KIC 6541920, *Kepler* magnitude 13.709. In this example the transit signal location is strongly consistent with the target star. The distribution of averages of the observed transit positions D_o is rendered as green contours, while the magenta contours show the distribution D_{target} with the transit modeled on the target star. The target star is shown as a red asterisk, the black crosshairs are the observed quarterly transit locations, and the light grey dashed crosshairs are the modeled transit locations. In this example the modeled transit locations are very tightly clustered so the magenta contours appear as a dot near the target star. The magenta crosshair and cyan circle show the χ^2 average position and 3σ radius used in conventional planet candidate vetting.

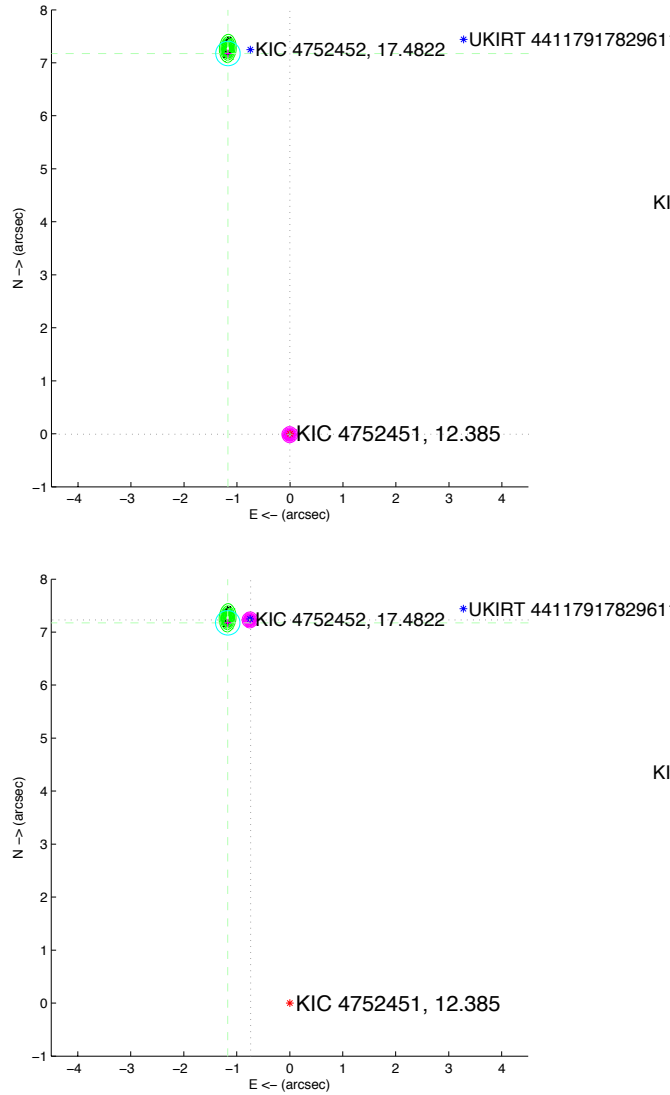


Figure 3: Results of the probability analysis for KOI-109.01 with target star KIC 4752451, *Kepler* magnitude 12.385. In this example the transit signal location is strongly inconsistent with the target star, and is consistent with the 17th magnitude field star KIC 4752452, shown by the circled blue asterisk. Top: the transit modeled on the target star KIC 4752451, so the modeled magenta contours and the observed green contours are very far apart. Bottom: the transit modeled on KIC 4752452, so the modeled magenta contours are close enough to overlap with the green observed contours. Though the overlap is small, it is large enough for a 98.9% probability that the transit is co-located with KIC 4752452, while the background probability is 1.1%. See Figure 2 for a description of elements of the figure.

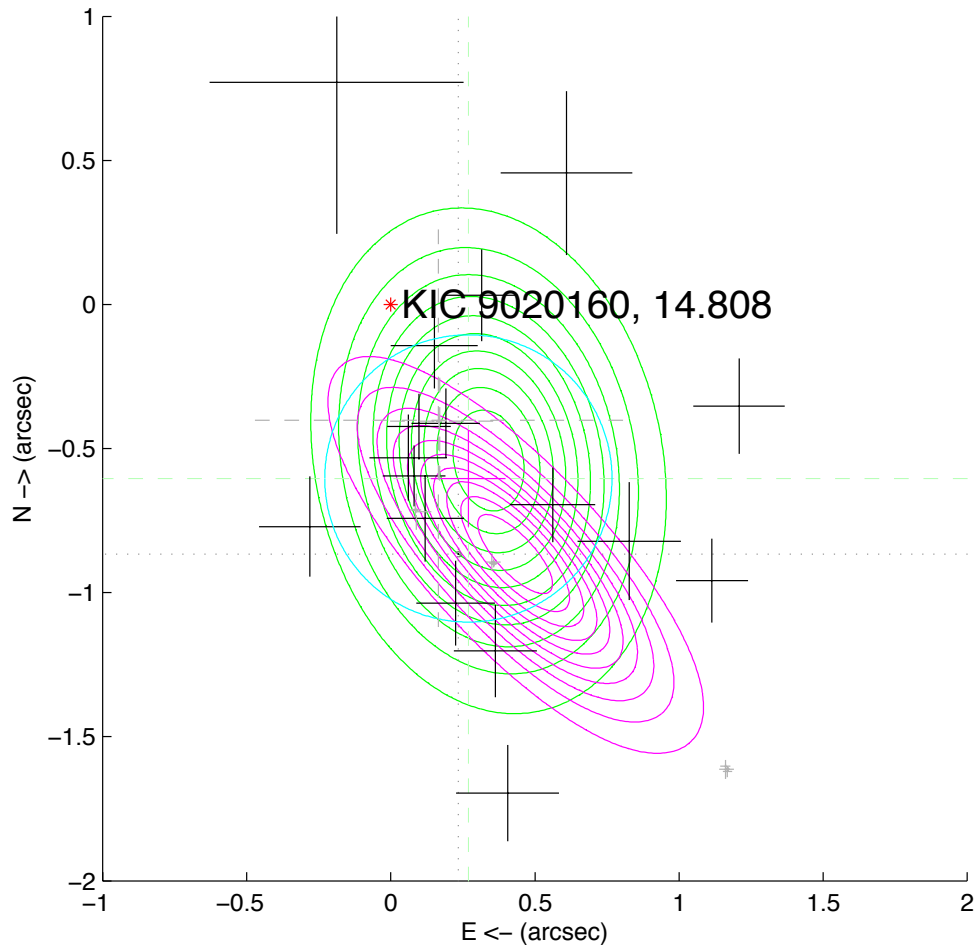


Figure 4: Results of the probability analysis for KOI-582.01 with target star KIC 9020160, *Kepler* magnitude 14.808. In this example the transit signal location is more than 3σ away from the target star, as indicated by the cyan circle offset to the SW. The conclusion that the transit source is offset from the target star is reinforced by the location of the black crosshairs showing the quarterly transit offsets. The observed distribution shown by the green contours is also offset, consistent with the 3σ circle. But the modeled distribution shown by the magenta contours shows that the distribution of offsets is expected to be offset in this way from the target star due to crowding bias. Therefore it would be incorrect to declare this KOI to be a background false positive. See Figure 2 for a description of elements of the figure.

Appendix A

We use the notation of §4.1.3. Our desired likelihood for each star s is the integral of the product

$$L_s = \int D_o(\mathbf{x}) D_s(\mathbf{x}) d\mathbf{x}. \quad (9)$$

The product of two Gaussians is the Gaussian

$$G(\mathbf{x}, \mathbf{b}_{o,j}, \Sigma_o) G(\mathbf{x}, \mathbf{b}_{s,k}, \Sigma_s) = c_{j,k} G(\mathbf{x}, \mathbf{m}_{j,k}, \Phi) \quad (10)$$

where

$$\begin{aligned} c_{j,k} &= G(\mathbf{b}_{o,j}, \mathbf{b}_{s,k}, \Sigma_o + \Sigma_s) \\ \mathbf{m}_{j,k} &= (\Sigma_o^{-1} + \Sigma_s^{-1})^{-1} (\Sigma_o^{-1} \mathbf{b}_{o,j} + \Sigma_s^{-1} \mathbf{b}_{s,k}) \\ \Phi &= (\Sigma_o^{-1} + \Sigma_s^{-1})^{-1}. \end{aligned} \quad (11)$$

so

$$\begin{aligned} L_s &= \int D_o(\mathbf{x}) D_s(\mathbf{x}) d\mathbf{x} \\ &= \frac{1}{N^2} \int \sum_{k=1}^N \sum_{j=1}^N G(\mathbf{x}, \mathbf{b}_{o,k}, \Sigma_o) G(\mathbf{x}, \mathbf{b}_{s,j}, \Sigma_s) d\mathbf{x} \\ &= \frac{1}{N^2} \int \sum_{k=1}^N \sum_{j=1}^N c_{j,k} G(\mathbf{x}, \mathbf{m}_{j,k}, \Phi) d\mathbf{x} \\ &= \frac{1}{N^2} \sum_{k=1}^N \sum_{j=1}^N c_{j,k} \end{aligned} \quad (12)$$

because $\int G(\mathbf{x}, \mathbf{m}_{j,k}, \Phi) d\mathbf{x} = 1$.

References

- Brown, T. M., et al. 2011, ApJ 142, 112
- Bryson, S. T., et al. 2010a, ApJ 713, L97
- Bryson, S. T., et al. 2010b, Proc. SPIE 7740, 77401D
- Bryson, S. T., et al. 2013, PASP 125, 889
- Burke, C. J., et al. 2014, ApJS 210, 19
- Christiansen, J., et al. 2013, *Kepler Data Characteristics Handbook*, KSCI-19040
- Coughlin, J., et al. 2014, ApJ 147, 119
- Coughlin, J., et al. 2015, *in preparation*
- Gregory, P. 2010, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press
- Jenkins, J., et al. 2010, Proc. SPIE 7740, 77400D
- Koch, D., et al. 2010, ApJ 713, L79
- Lawrence, A., et al. 2007, MNRAS 379, 1599
- Levenberg, K. 1944, Quarterly of Applied Mathematics 2, 164
- Marquardt, D. W. 1963, Journal of the Society for Industrial and Applied Mathematics, 11, 431
- Morton, T. D., and Johnson, J. A. 2011, ApJ 738, 170
- Morton, T. D. 2012, ApJ 761, 6
- Mullally, F., et al. 2015, ApJS 217, 31
- Scott, D. W. 1992, *Density Estimation: Theory, Practice and Visualization*, John Wiley & Sons, New York, Chichester
- Silverman, B. W. 1986, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London