Kepler: A Search for Terrestrial Planets

Kepler Data Processing Handbook

KSCI-19081-001
1 April 2011

NASA Ames Research Center
Moffett Field, CA. 94035

Prepared by: Michael N. Fanelli Date 30 March 2011

Michael N. Fanelli, Kepler Guest Observer Office

Approved by: Date 30 March 2011

Jon Jenkins, Co-Investigator for Data Analysis

Approved by: Michael R. Haas Date 3/30/11

Michael R. Haas, Science Office Director

Approved by: Date 20 March 2011

Thomas N. Gautier, Project Scientist

Note on Authorship

The Data Processing Handbook is the collective effort of the Data Analysis Working Group (DAWG) composed of members of the Science Office (SO), Science Operations Center (SOC), Guest Observer Office (GO) and the Science Team. Sections 4-10 were each written initially as a stand-alone description of the relevant Pipeline processing module. We have combined these contributions into a single document to provide a start-to-finish description of the data reduction, calibration, and transit analysis steps in the Pipeline. The authors are:

Fanelli, Michael N., KDPH Editor
Jenkins, Jon M., DAWG Chair (Section 8)
Bryson, Stephen T., (Section 4)
Quintana, Elisa, V., (Section 5)
Twicken, Joseph, D., (Sections 6 and 7)
Wu, Haley W., (Section 9)
Tenenbaum, Peter, (Section 10)

Allen, Christopher L.
Caldwell, Douglas, A. DAWG Co-Chair
Chandrasekaran, Hema
Christiansen, Jessie L.
Clarke, Bruce D.
Cote, Miles T.
Dotson, Jessie L.
Gilliland, Ronald (STSci)
Girouard, Forrest
Gunter, Jay P.
Hall, Jennifer
Haas, Michael, R.
Ibrahim, Khadeejah
Kinemuchi, Karen
Klaus, Todd
Kolodziejczak, Jeff (MSFC)
Li, Jie
Machalek, Pavel
McCauliff, Sean D.
Middour, Christopher K.
Morrris, Robert
Mullally, Fergal
Seader, Shawn
Smith, Jeffrey C.
Still, Martin
Thompson, Susan E.
Uddin, Akm Kamal
Van Cleve, Jeffrey
Wohler, Bill

The first authors of the various sections are noted above. We wish to thank our technical writers, Greg Orzech and Susan Blumenberg for their help in producing the original SPIE papers and this handbook.

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

TABLE OF CONTENTS

1 PHILOSOPHY AND SCOPE

1.1 Intended Audience

The *Kepler* Data Processing Handbook (KDPH) is written for a broad audience: the *Kepler* Science Team, Guest Observers (GOs), archival researchers, and peer reviewers of *Kepler* results submitted for publication. The KDPH first describes the selection of targets and specifications of their associated pixel sets. These target pixel data sets are collected by the Flight System as described in the *Kepler* Instrument Handbook (KIH). The KDPH then describes the transformation of these pixel sets into photometric time series, and the detection and validation of transits in those time series by the data analysis Pipeline as developed and operated by the *Kepler* Science Operation Center (SOC). The KDPH provides information about the algorithms, inputs, outputs, and performance of target management and Pipeline software components (called Computer Software Configuration Items or CSCIs). Together, the KIH and the KDPH supply the information necessary for understanding *Kepler* results, given the real properties of the hardware and the data analysis methods used.

1.2 Relationship to Other Documents

1.2.1 Kepler Instrument Handbook (KSCI-19033)

The KIH provides information about the design, performance, and operational constraints of the *Kepler* hardware, and an overview of the available pixel data sets. That document presents an overview of the *Kepler* instrument, and then tracks photons through the telescope, focal plane, and focal plane electronics. Details regarding targets, the pixels of interest around them, and operational details are specified, which will be helpful in both planning observations, and understanding the data reduction procedures described in this document.

1.2.2 Kepler Data Characteristics Handbook (KSCI-19040)

The Data Characteristics Handbook provides a description of a variety of phenomena identified within the *Kepler* data, and a discussion of how these phenomena are currently handled by the data reduction Pipeline.

1.2.3 Kepler Data Release Notes (KSCI-19041, etc)

With each quarterly release of data, a set of accompanying notes is created to give *Kepler* users information specific to the time period during which the data was obtained and processed. The notes provide a summary of flight system events that affect the quality of the data and the performance of the processing Pipeline. The Data Release Notes, along with other *Kepler* documentation, are located at the Multi-mission Archive (MAST) at Space Telescope. Once the user becomes familiar with the content of the Data Characteristics Handbook, they need only read the short Release Notes for details specific to that quarter.

1.2.4 Kepler Archive Manual (KDMC – 10008)

Data from the *Kepler* mission are archived at MAST, which serves as NASA's primary archive for ultraviolet and optical space-based data. The *Kepler* Input Catalog (KIC), processed light curves, and target pixel data are all accessed through MAST. The *Kepler* Archive Manual describes data products, file formats and the functionality of *Kepler* data access. The Archive Manual can be accessed from the MAST *Kepler* page: ://archive.stsci.edu/kepler, and is available in HTML, DOC, and PDF formats.

1.2.5 Kepler Guest Observer Documentation

The *Kepler* Guest Observer (GO) program web page, located at ://keplergo.arc.nasa.gov/ provides information and tools for observation planning, and data analysis.

1.2.6 The Kepler Mission Special Issue of Astrophysical Journal Letters

An initial description of the *Kepler* mission, data characteristics, and early science results is presented in Astrophysical Journal Letters, Volume 713. This compendium contains several papers providing background on mission definition (Koch et al. 2010), science operations (Haas et al. 2010), target selection (Batalha et al. 2010), instrument performance (Caldwell et al. 2010), the *Kepler* point spread function (Bryson et al. 2010c), the data processing Pipeline (Jenkins et. al 2010a), and the characteristics of the long (Jenkins et al. 2010b) and short cadence data (Gilliland et al. 2010).

1.3 Document Organization

This document presents an overview of Science Operations Center target management and data processing procedures, and then follows the data from target selection through aperture definition, pixel-level calibration, aperture photometry, and transit detection. For each operation applied to the data, the algorithms, inputs, and outputs are described, and example results shown.

2 INTRODUCTION TO KEPLER DATA PROCESSING

The *Kepler* mission was designed to determine the frequency of Earth-sized planets lying within the habitable zone of Sun-like stars (FGKM dwarfs). Extrasolar planets are detected using the transit method, in which the passage of a planet across the disk of a star produces small reductions in the light of that star. *Kepler* continuously monitors the brightness of stars using a space-based telescope and differential photometer within an ~115 square degree field of view (FOV). Approximately 160,000 main-sequence stars constitute the primary observing program. Several thousand additional sources are normally observed during each observing season; these sources comprise the asterosiesmology, guest observer, and other science targets. *Kepler's* mission profile defines many of the operating procedures and constraints imposed on the data collection and reduction process. The broad outline of the mission is described in several papers which the user may wish to read prior to working with *Kepler* data: Borucki et al (2010), Koch et al. (2010), and Haas et al. (2010). This section provides a brief overview of the data collection and reduction process.

2.1 Science Operations Architecture

Kepler was launched on 7 March 2009 into an Earth-trailing heliocentric orbit. The spacecraft points at one location on the sky for the entire duration of the mission. Continuous observing lasts for approximately one month, separated by a ~one day gap during which the spacecraft is oriented towards Earth for downlink of the stored data via NASA's Deep Space Network. Every quarter, the spacecraft rolls 90 degrees to maximize electric power production and thermal control. Science data contain gaps corresponding to both the quarterly rolls and monthly downlink times.

Data from the Deep Space Network are unpacked at the Mission Operations Center, located at the Laboratory for Atmospheric and Space Physics (LASP) in Boulder, CO. Ball Aerospace and Technology Company is responsible for *Kepler* Mission Operations, providing engineering support and directing the *Kepler* operations team at LASP. The data are then sent to the Data Management Center at the Space Telescope Science Institute, where they are collected into FITS-formatted files, and then transmitted to the Science Operations Center (SOC) at NASA Ames.

The SOC operates the data processing and calibration Pipeline that converts raw data numbers into calibrated pixels and light curves, and corrects for instrumental signatures in the calibrated light curves. The Pipeline also contains software to conduct a transit signal search within all observed light curves and attempts to characterize any detected transit signals.

Data products produced by the Pipeline are exported back to the Space Telescope Science Institute for archiving at the Multi-mission Archive at Space Telescope (MAST). These data include light curves, target pixels, and full-frame images. All users obtain their data through the search and retrieval web interfaces maintained by MAST.

The SOC is responsible for algorithm development, Pipeline operation, and code verification. Observation planning is conducted by the SOC in coordination with the Science Office prior to each observing season. New target definitions are generated, validated by the ground segment, and uploaded to the spacecraft during the quarterly roll operation. Newly downlinked data is processed by the most recent Pipeline code release and validated by the Data Analysis Working Group (DAWG). Once approved, data products are exported to MAST, and transit candidates are analyzed by the science team.

2.2 Science Concept to Sources: the Kepler FOV and Input Catalog

The search for terrestrial planets begins with an appropriate list of stars to be monitored for transits. After defining a specific field-of-view (FOV), The Project developed this list, a catalog of FGKM dwarf stars located within the selected FOV.

2.2.1 FOV

Kepler operates using the "stare" method, monitoring ~170,000 sources continuously. To provide long-term, continuous monitoring of the target list, a single sky field was selected. The FOV is located in the Cygnus-Lyra region, centered on galactic coordinates l = 76.32º, b = +13.5º (RA = 19h 22m 40s, DEC = +44º 30' 00"). This FOV was dictated by spacecraft operational considerations (power, sun angle) and the stellar content of the target field (Koch et al 2010).

2.2.2 Stellar Classification Program

The *Kepler* Project conducted a stellar classification program to provide a well-defined target list for the exoplanet survey. The basis for classification was an extensive optical observing program, which imaged the entire FOV. Calibrated photometry from this imaging program was used to classify stars by correlating against parameter grids generated from synthetic stellar spectra (Brown et al. 2011).

2.2.3 Kepler Input Catalog

The output of the Stellar Classification Program is the *Kepler Input Catalog*, or '"KIC", which is the primary source list for science programs executed by *Kepler*. The KIC contains an assigned ID number, known as the KepID (sometimes KIC#), coordinates, photometry and stellar classification parameters. Users "mine" the KIC to identify sources of interest to their science programs. Additional details regarding how KIC parameters are used within target management and Pipeline data processing are described in Section 3.

2.3 Target Management: Targets to Target Definitions

Almost all observing targets are selected utilizing the *Kepler Input Catalog*. The SO/SOC manage the process through which target lists are converted into target definitions. These tasks include:

- Calculating several source-specific parameters for each target: (a) the optimal photometric aperture that maximizes the signal-to-noise ratio of the derived light curve; (b) a crowding metric that quantifies the ratio of the flux from the target to the flux from the surrounding stars, and (c) the flux fraction which measures the fraction of total source flux in the optimal aperture.
- Defining individual pixel masks (shape, # of pixels) for all selected sources.
- Associating each required mask with the best-fitting mask drawn from a table of masks.
- Designing dedicated pixel masks, if needed, for bright or unusual sources.
- Defining the location and pixel masks for background flux collection.
- Creating target definitions, which associate specific locations on the detector (row, column, CCD module, CCD output), for each source.

The observing program is adjusted to meet operational constraints. The long cadence target definition table is constrained to 170K targets or 5.44 million pixels. The short cadence target definition table is limited to 512 targets or 43520 pixels. Section 4 details the algorithms and outputs used in target management.

2.4 Flight System: Pixel Tables to Cadence Data Sets

Every quarter target definitions tables are uploaded to the onboard memory: short cadence (~1 minute) target definitions can be swapped on a monthly basis, long cadence (~30 minute) target definitions are changed quarterly. Photometry is recorded by the solid-state recorders for only those pixels explicitly specified in the target definition tables, about 6% of all pixels. In addition to the target pixels, one complete *full-frame* image is also obtained each month. A full-frame image contains flux values for all 95 million visible pixels in the detector array. After downlink and pixel archiving, the data arrive at the SOC for calibration, processing, and analysis.

2.5 Photometry Pipeline: Cadence Data Sets to Flux Time Series

The *Kepler* Data Processing Pipeline converts raw data numbers recorded on the spacecraft into corrected and time-stamped photometric light curves. This process contains three major steps, each step defined by a software module. Each module is composed of a number of MATLAB routines largely directed at science data processing, and java software to manage I/O from the local database (Klaus et al. 2010a, 2010b; McCauliff et al. 2010; Middour et al. 2010). The principle software modules are: Calibration (CAL), Photometric Analysis (PA) and Pre-Search Data Conditioning (PDC).

Calibration converts raw data numbers for each observation (cadence) into calibrated pixels. The raw data include photometric (target and background) pixels, along with additional pixels termed “collateral data” that are used to eliminate bias, smear, and other instrumental artifacts. Section 5 details the algorithms and outputs of CAL.

Photometric Analysis assigns timestamps, removes background signals, computes photocenters, and performs optimal aperture photometry using the aperture definitions determined during target management. The result is a calibrated light curve with associated errors. Section 6 details the algorithms and outputs of PA.

Pre-Search Data Conditioning performs systematic error corrections to the light curves derived by PA. This module corrects instrumental signatures including pointing drift, velocity aberration, thermal effects, and cosmic rays. PDC is currently designed to condition the light curves for transit searches, not necessarily to robustly preserve astrophysical signals for all periods and amplitudes. The algorithmic basis for the data reduction procedures in PDC remain under development, as the Project attempts to better understand and remove systematic errors from the *Kepler* data. Section 7 details the algorithms and outputs of PDC.

2.6 Planet Search Pipeline: Flux Time Series to Valid Transits

All light curves produced by PA and PDC are exported to the *Kepler* archive as FITS-formatted files. These same data form the input for the Transiting Planet Search (TPS) and Data Validation (DV) Pipeline modules, which conduct the search for transit signals.

Transiting Planet Search identifies features in the light curves which match the expected signature of a transit with durations in the range of 1 to 16 hours (Jenkins 2002). Light curves whose maximum folded detection statistic exceeds a specified threshold are designated

Threshold Crossing Events (TCEs). Light curves containing a TCE are passed to DV. The TPS module computes *Kepler's* photometric precision to characterize the noise levels in the light curves, and assess the robustness of transit detection. The photometric precision metric is termed the Combined Differential Photometric Precision (CDPP). Section 8 details the algorithms and outputs of TPS.

Data Validation performs a series of tests to confirm the transit candidates identified by TPS. TCEs are subject to these tests in an automated way, and the results are recorded in documents termed DV Reports. Section 9 details the algorithms and outputs of DV.

Data Validation Fitter. DV module tests require a model of the planetary system which is consistent with the observed transit periods, durations, and depths. These models are produced by a planet-model fitter procedure. Section 10 details the algorithms and outputs of DV-Fitter.

2.7 SOC Software Releases

The *Kepler* Data Processing Pipeline is defined by its *release number*, which describes the body of code used to process a specific data set. Changes to the release number reflect major evolutionary changes or added functionality. Pipeline changes occur as our understanding of the instrument, data artifacts and processing features improves. *Kepler* data users should be aware that the output data products and formats may also change with time.

The current Pipeline version is SOC 6.2. An upgrade to SOC 7.0 is expected in July 2011.

TABLE 1
Kepler Data Processing Timeline

Quarter	SOC Version	Data Release Note	Release Date
Q0	6.1	5	04-Jun-2010
Q1	6.1	5	04-Jun-2010
Q2	6.1	7	16-Sep-2010
Q3	6.1	4	15-Apr-2010
Q4	6.1	6	22-Jul-2010
Q5	6.2	8	25-Oct-2010
Q6	6.2	9	23-Jan-2011
Q7	6.2	10	22-Apr-2011

2.8 Data Archive

The last step in the production of *Kepler* data is the export of the Pipeline products to the *Archive* maintained at MAST, from which users obtain their data. All data is archived, including raw and calibrated pixels, and calibrated and corrected light curves. *Kepler* data of interest includes:

- Target pixel images – the individual pixel masks collected for each observation during a quarter and calibrated by CAL
- Light curves – derived by PA and corrected by PDC
- Full-frame images – output from all 84 channels on the detector array

Delivery of each quarter's data initially occurs four months after the end of data collection. Re-processed data are delivered on a schedule developed by the Science Office. Re-processing occurs when the SOC Pipeline is updated to a new release, indicating an improved processing environment.

3 THE KEPLER INPUT CATALOG

The *Kepler Input Catalog* provides information on celestial objects within the *Kepler* field-of-view. About 4.4 million cataloged sources fall on the detector array for at least one *Kepler* observing season. The KIC serves two purposes. First, the KIC is the primary source for selection of targets for observation. Second, KIC parameters are used for observation planning, data reduction, and transit analysis within the Pipeline.

3.1 Catalog Content

The KIC was developed by the *Kepler* Science Team prior to launch and included an optical observing campaign as described by Brown et al (2011). Photometry was obtained using SDSS-like filters in the *g,r,i,z* bands, and an intermediate-band filter, labeled D51, used to separate cool dwarfs from giants. For stars with good photometry, estimates for stellar effective surface temperature, gravity, radii, and mean metal abundance were derived. The KIC also contains an estimate of source fluxes within the *Kepler* bandpass (Van Cleve & Caldwell 2009), quantified as the *Kepler* magnitude, *Kp.* This magnitude is roughly equivalent to a Johnson *R*-band magnitude for stars with $0 < (B-V) < 1.0$. The photometry and derived parameters were federated with data from other photometric survey catalogs (Tycho, 2MASS, Hipparcos, USNO-B) to provide a full census of sources in the field down to $g \sim 20$ corresponding to the completeness limit of the USNO-B catalog. Stellar parameters from the KIC form the basis for the selection of the exoplanet survey target set (Batalha et al. 2010).

3.2 KIC Parameters and the Pipeline

Kepler magnitudes derived from KIC photometry, coupled with the spatially-varying pixel response function (Bryson et al. 2010c) are used to define the required pixel masks (shape and number of pixels) for all sources, determined during target management. Assignment of target apertures and associated pixel masks are described in Section 4. The KIC is used to estimate (1) the degree to which each target aperture is contaminated by light from nearby sources, and (2) the fraction of source light contained in the photometric aperture compared to the total source flux. The first quantity is termed the *crowding metric*, and the second is termed the *flux fraction* in aperture. The crowding metric equals the flux of the target star in the photometric aperture expressed as a fraction of the total flux (star plus background flux). This metric can be used to select uncrowded targets and to correct flux time series for crowding during data calibration. It may be particularly useful for comparing light curves across quarters. Light curves are corrected for source crowding in the PDC module, but the output of PA is not corrected. Application of the crowding metric is detailed in Section 7.3.3. No correction for the flux fraction is applied in the current Pipeline.

4 TARGET AND APERTURE DEFINITION

4.1 Introduction

Communication bandwidth and data storage constraints limit the amount of data that may be stored and downlinked from the *Kepler* spacecraft. Pixels required for precision photometry on the ~170K target stars per quarter must be identified and selected for storage and downlink. Section 4 describes the method used to determine the required pixels for each target. The target definition process is repeated every quarter as the spacecraft rolls and some targets are replaced with alternates. One of the primary goals of pixel selection is to identify pixels that are optimal for aperture photometry, as described in 4.3.

4.1.1 Target and Aperture Definition Task Flow

The *Kepler* flight system software uses a table of 1024 *aperture masks* to collect all target data, so the pixels associated with each target must fit into one of these masks. Efficient design of these masks is required to gather the desired pixels. The pixel selection workflow proceeds as follows:

1. For each target an optimal aperture is defined (see § 4.3).
2. A series of 1 to 4 pixel halos are added to the optimal aperture based on the magnitude uncertainty of the target.
3. An undershoot column is added to the left of the outermost halo.
4. A mask is selected which most efficiently captures all these pixels (see § 4.4.3)

A single target, particularly a very bright stellar target, may be tiled with more than one mask.

The resulting mask assignments are collected into target definitions that include the aperture mask index and its location on the focal plane (module, output, row, column).

4.1.2 Kepler Pixel and Target Types

Pixel data are collected for several types of targets:

1. **Stellar** targets are point-like sources whose pixels are selected to maximize the signal-to-noise ratio (SNR) (see 4.3.2). Stellar targets are specified by a *Kepler* ID. These targets may be observed at either short or long cadence.
2. **Custom** targets are explicitly specified collections of pixels. Custom targets are defined by a reference pixel position and a set of offsets, one for each pixel, from that reference position. Custom targets are used for non-stellar sources and engineering purposes. These may be either short or long cadence.
3. **Background** targets are small (nominally 2 x 2) sets of pixels that sample the background in long cadence. These pixels are collected to support removal of the background signal (see § 4.3.3).
4. **Reference pixel** (RP) targets are special stellar targets which are downlinked bi-weekly via low-bandwidth X-band communications to monitor spacecraft health and performance (Chandrasekaran et al. 2010). Reference pixel stellar targets share a mask table with LC

and SC, and also appear on the LC target list. Custom masks are used to collect background, black and smear collateral data for the RP targets.

4.2 Pixel Selection Requirements

Memory, bandwidth, and flight software design impose several constraints on the final set of pixels selected for downlink:

Long cadence: There may be no more than 170,000 target definitions, with no more than 10,000 per output channel, and no more than 5.44 million pixels across the FOV. This implies an average of ≤ 32 pixels per target definition.

Short cadence: There may be no more than 512 target definitions and no more than 43,520 pixels. These values imply an average of ≤ 85 pixels per target.

Background: There may be no more than 1125 target definitions and no more than 4,500 pixels (for an average of 4 pixels per target) for each of 84 output channels.

Reference pixels: At the beginning of the *Kepler* mission, there may be no more than 96,000 reference pixels across the entire focal plane. As *Kepler* moves away from Earth, bandwidth degradation reduces the number of reference pixels that can be downlinked in a bi-weekly contact.

Aperture mask table: Up to 1024 aperture masks may be defined, using no more than 87,040 pixels. This implies an average of 85 pixels per mask for 1024 masks.

For reasons described in § 4.5, 252 entries in the aperture mask table are dedicated to supporting reference pixel observations, leaving 772 aperture masks for the LC and SC targets.

Basic *Kepler* mission requirements specify the capability to observe stars with magnitudes between 9 and 15. Simultaneously meeting the above constraints yet providing efficient mask assignment for a wide range of stellar magnitudes is a significant challenge. Observation of stars brighter than magnitude 9 is permitted, but pixel efficiency may suffer.

4.3 Pixel Selection

Pixels for stellar targets are selected to maximize the SNR, while background pixels are selected to estimate the background. Both pixel types are selected based on a model of the sky using the KIC and the *Kepler* photometer’s optical and electronic properties. The basic strategy for optimal pixel selection is to create two synthetic images: one with all stars in the region of the target star, and another with all stars except the target star itself. These images are then used to compare the signal from the target star with the noise from the target star, stellar background, and the instrument. The techniques in this section are based on work by Jenkins, Peters & Murphy (2004).

4.3.1 Synthetic Image Creation

A synthetic image is created using the following elements:

Kepler Input Catalog. The KIC provides J2000 right ascensions, declinations, and magnitudes in the Kepler bandpass (Brown et al. 2011).

Pixel Response Function (PRF) (Bryson et al. 2010c) is an observation-based super-resolution model of how light from a star falls on *Keple*r pixels at different locations in the focal plane. There is one PRF model for each output channel, and this model contains five PRFs that are linearly interpolated to capture intra-channel PRF variations. The observations used to generate the PRF model were taken at a 15-minute cadence, sufficient to capture the LC behavior of spacecraft pointing jitter. The PRF model includes intra-pixel variability.

Focal plane geometry (FPG) and pointing model (Bryson et al. 2010c) includes measurements of the relative CCD locations in the Kepler focal plane, plus models of the *Kepler* optics and of differential velocity aberration (DVA). These models are used to determine the pixel location of the central ray of each stellar target as a function of time. DVA can move a star as much as 0.6 pixels in a quarter.

Saturation model (Caldwell et al. 2010) includes information about the well depth of each output channel.

Zodiacal light model is a representation of the zodiacal light versus position on the focal plane as a function of time.

Read noise model (Bryson et al. 2010c) provides observed values of read noise for each output channel.

Charge transfer efficiency (CTE) model describes how much flux is lost with each charge transfer during readout.

The synthetic image used for pixel selection models the signal in calibrated pixels, so smear and other instrumental effects are not included. The pixels selected for optimal photometry (described in 4.3.2) are used for an entire quarter, so the pointing model is used to smear the PRF along the path for each star due to DVA.

The synthetic image for each output channel is generated star by star. For each star in the KIC that falls on the output channel:

1. The star's pixel position on the channel is computed from the star's RA and Dec in the KIC, using the FPG and pointing model (including sub-pixel position).
2. The PRF at the star's position is evaluated along the path taken by the star due to DVA, creating a smeared PRF covering the entire quarter.
3. The pixels resulting from the star's PRF are normalized using the *Kepler* magnitude in the KIC. The resulting target-only image is saved for each target.
4. The normalized pixels are added to the synthetic image at the appropriate pixel location.

At this point the synthetic image represents the collection of stars in the KIC as they would appear on the CCD without any saturation or zodiacal light background. A copy of the synthetic image, called the background image, is made for use in the optimal pixel selection. The zodiacal light is interpolated onto each pixel and added to the synthetic image. Saturation is then iteratively spilled along columns by moving flux which exceeds the well depth up and down the column. Finally the CTE model is applied. A comparison of the resulting synthetic image and the in-flight image of the same pixels is shown in Figure 4.1.

There are several sources of error in the generation of the synthetic image, which can compromise optimal apertures:

PRF errors: Each PRF model is computed as an average of several observed stars without consideration of color. The actual PRF of individual stars will differ because the optical point spread function underlying the PRF has minor color dependence. The PRF varies within a channel, with stronger variation near the edge of the FOV, and this variation is only approximately modeled by the linear interpolation included in the PRF model.

KIC errors: The KIC contains artifacts which do not correspond to real objects on the sky. Such artifacts can be see in Figure 4.1 around the bright star at (520, 580), where there is a faint diagonal line of KIC entries extending from the lower left to the upper right through the core of the star. This is likely a diffraction spike misidentified as faint stars and does not appear in the flight image. The KIC does not reflect stellar variability, which is unknown for many stars in the *Kepler* field (for stars whose variability is known the brightest magnitude is used). An example of such magnitude errors can be seen in the star at (540, 570) of Figure 4.1, which is brighter in the synthetic image than in the flight image. When a KIC underestimate of a target's magnitude has been identified, there is a mechanism to provide corrected magnitudes to the pixel selection process.

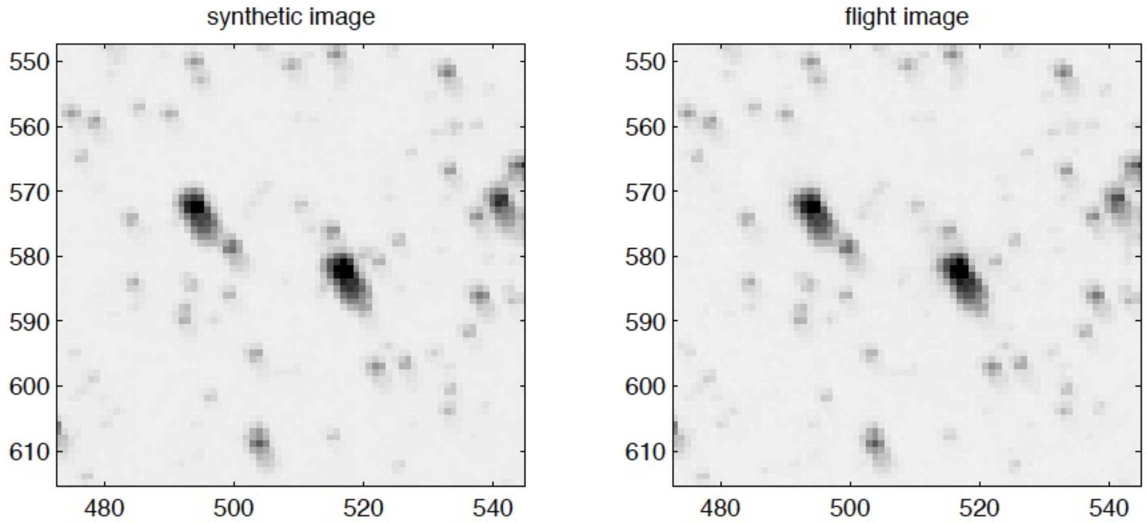


Figure 4.1. A comparison of a synthetic image (left) and a flight image of the same region of the sky, near the edge of the *Kepler* FOV where the PRF is large. For purposes of comparison with the sky image, the synthetic image is computed for a short time interval so it is nor smeared by DVA.

Saturation model: The current saturation model provides the well depth per output channel, assuming that the well depth does not vary within the channel and that the spill up and down a column is symmetric. Both of these assumptions are violated, with larger variations in both well depth and symmetry. These variations have a higher spatial frequency than can be measured with the bright stars in the *Kepler* FOV. To provide margin against this uncertainty in saturation spill asymmetry, the simulated saturation is extended by 50% in both directions.

Changing focus: The focus of the *Kepler* photometer undergoes seasonal changes as the Sun angle and Kepler orientation vary during the year. In addition, smaller high-frequency focus changes have been observed that are highly correlated with various heaters operating on the spacecraft. These focus changes cause PRF changes not captured in the current static synthetic image. Focus changes also induce plate scale variations, so the stars are not placed in the correct positions in the synthetic images. Though plate scale-induced position errors are small, their effects have been observed in photometry and corrective measures have been implemented in the Pipeline. As these other systematic effect are understood and corrected, the optimal aperture can be redefined to include halo pixels during subsequent processing runs to further improve the photometry.

4.3.2 Optimal Pixel Selection

The target-only image created in the previous step is subtracted from the background image, creating an image with all stars except the target. Saturation and CTE are then simulated as described above. Pixel values p_{target} in the target-only image define the signal in each pixel, while the pixel values p_{back} in the background image provide the background signal. The SNR of each pixel is estimated as

$$SNR_{\text{pixel}} = \frac{p_{\text{target}}}{\sqrt{p_{\text{target}} + p_{\text{back}} + \nu^2_{\text{read}} + \nu^2_{\text{quant}}}}, \qquad (4\text{-}1)$$

where ν_{read} is the read noise and ν_{quant} is the quantization noise.

Quantization noise is given by:

$$\nu_{\text{quant}} = \sqrt{\frac{n_C}{12}\left(\frac{w}{2^{n_b-1}}\right)^2}, \qquad (4\text{-}2)$$

where n_C is the number of cadences in a co-added observation (30 for LC), w is the well depth, and n_b is the number of bits in the analog-to-digital converter (= 14). This noise formula includes Poisson shot noise of both the target and background pixel values.

Given a collection of pixels, the SNR of the collection is given by the square root of the sum of the squares of the SNR of the component pixels. Optimal pixel selection begins by including the pixel with the highest SNR. The next pixel to be added is the pixel that results in the greatest increase in SNR of the collection. Initially the SNR will increase as pixels are added. After the bright pixels in the target have been included, dim pixels dominated by noise cause the aggregate SNR to decrease. The pixel collection with the highest SNR defines the optimal aperture. Figure 4.2 shows an example of the dependence of SNR on pixel inclusion.

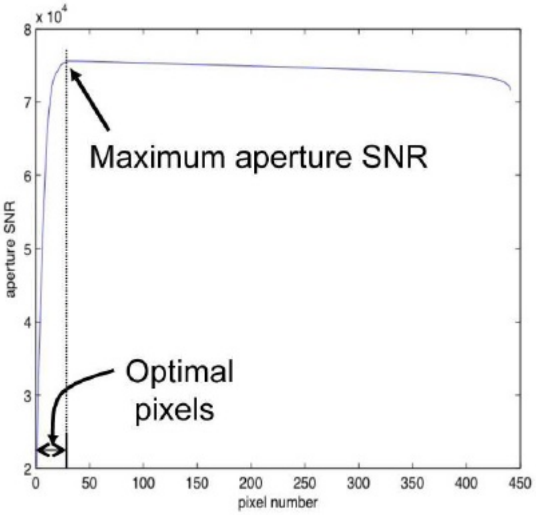


Figure 4.2. The aperture SNR curve built up as pixels are added in order of decreasing pixel SNR. The optimal set of pixels is the set at the maximum of this curve.

The background and target images are used to estimate crowding by calculating the fraction of flux in the optimal aperture due to the target star. The resulting crowding metric is useful for estimating the dilution of flux from the target in the optimal aperture, which has an impact on the detectability of transits (Batalha et al. 2010). The same measure on a 21 x 21 pixel aperture provides a sky crowding metric which may be useful for identifying uncrowded stars. The fraction of each target's flux that falls in its optimal aperture is also computed.

4.3.3 Background Pixel Selection

Background targets are selected to create a 2D polynomial representation of the background on each channel. This background polynomial is subtracted from the pixel values prior to aperture photometry. To lead to good polynomial fits, the background targets should be homogeneously and uniformly distributed on the output channel. Because the accuracy of background polynomials increases with an increasing density of data points but diminishes near the edge of the polynomial domain, more background targets are placed at the edges of the output channel. To achieve this distribution, the intersections of an irregular Cartesian mesh are used for the initial guess at background target positions. The two linear meshes provide 31 × 36 = 1116 intersections, close to the allowed maximum of 1125 background targets.

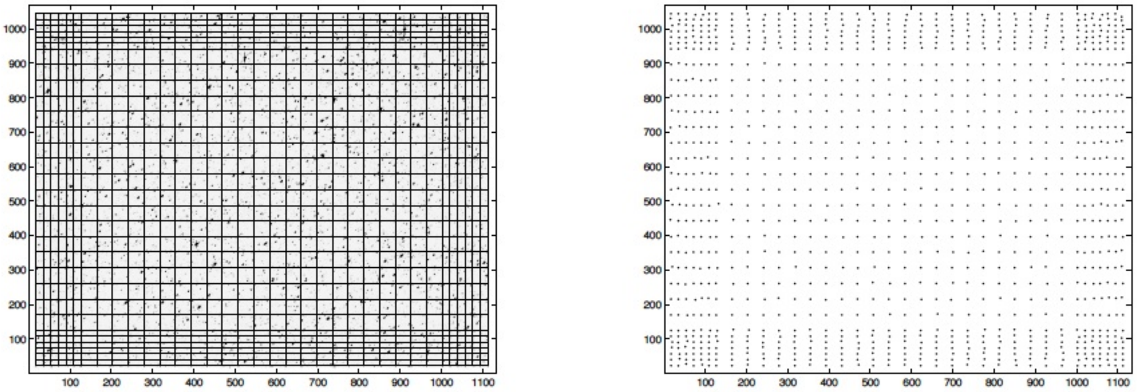


Figure 4.3. Left: The grid used to seed background target positions on an output channel. Right: Final locations of 2 × 2 pixel background targets for a typical module.

Because most pixels on an output channel do not contain stars, a pixel in the synthetic image is considered to be background if its value is less than the dominant mode of the pixel histogram. A 2 × 2 pixel background target is considered valid if all four pixels are below the background threshold. Initially, a background target is placed at each mesh intersection. If that target is not a valid background target, increasingly larger boxes centered on the mesh intersection are searched until either a valid background target location is found or the box exceeds a maximum size. In the latter case the target with the smallest summed pixel value is chosen. The performance of this search is enhanced by the creation of a 2 × 2-average binned image, so that the search in each box is done by taking the minimum of the binned image in that box. The initial mesh and resulting background locations are shown in Figure 4.3.

4.4 Mask Creation and Assignment

The pixel selection process described in section 4.3 combined with a typical assortment of custom apertures produces several thousand uniquely shaped pixel apertures across the *Kepler* field of view. These shapes must be mapped to 772 aperture masks with a minimum number of excess pixels collected. This presents a difficult combinatorial problem. A simple, near-optimal approach was developed that performs an iterative statistical estimate of which pixel apertures may be best used as entries in the mask table.

Our approach is tailored to produce aperture masks that are efficient for dimmer targets, where the vast majority of pixels are found, and does a relatively poor job for bright targets. This inefficiency for bright stars is addressed by creating several manually designed masks for high-priority targets. The mask creation and assignment process proceeds in three stages:

1. Required pixel apertures are defined by adding pixel margin to the requested pixels (see § 4.4.1).
2. The mask table is created from analysis of the final set of pixel apertures, supplemented by manual mask creation for bright targets (see § 4.4.2).
3. The required pixels for each target are mapped to one or more aperture masks in the mask table, creating the final target definitions (see § 4.4.3).

The mask table and target definitions are then delivered for quarterly upload to the spacecraft.

4.4.1 Required Pixels: Adding Pixel Margin

Once the pixels for a target have been specified, either by the optimal pixel selection process or by explicit specification for custom targets, the target may be assigned extra pixels. These extra pixels are of two types:

Halos are rings of pixels around the specified pixels for a target, typically applied as margin against uncertainty. A pixel is added to the halo if any of the eight adjacent pixels including corners are included in the target's specified pixels. Halos are added iteratively, for example, the third halo treats the previous two halos as specified pixels for the target.

Undershoot column is an extra column of pixels to the left ("upstream" in the pixel readout) of the specified pixels to provide data for the undershoot correction algorithm.

Halos and the undershoot column can be specified on a target-by-target basis, and when both are present the halos are applied first. The default for stellar targets is to add one halo and the undershoot column, and the default for custom targets is no halo or undershoot column. The final pixels, including the specified pixels and any halo or undershoot column pixels, are called *required pixels*.

4.4.2 Mask Table Creation

The table of 1024 aperture masks is divided into four components, each designed to accommodate specific target sets. Dim target masks are generated for the vast majority of stellar targets that fit required apertures with the smallest number of excess pixels. Dedicated masks are set to the required pixels for specific targets, and are used for oddly shaped diagnostic targets or very bright high-priority stars that are difficult to fit efficiently using other masks. Bright target masks are specially generated to fit long saturated columns or large cores of bright stars. Reference pixel target set masks are used to collect collateral data (black level and smear) for reference pixel targets. The dim target portion of the mask table, applicable to most stellar targets, is the largest component (736) and is filled as follows:

1. The aperture mask table is initialized to contain simple geometric shapes.
2. Mask assignment is performed using the full target set (as described in 4.4.3).
3. The unique required apertures are identified.
4. Masks that are perfect fits to some required aperture are identified and set aside.
5. Masks that were not perfect fits are sorted in descending order of the total number of excess pixels that are associated with that mask.
6. The masks in the last step are replaced with shapes from the unique set of required apertures for targets dimmer than a specified magnitude until the dim target region of the mask table is filled.

The bright target portion of the mask table (currently = 28) is filled in using two methods. First, the magnitude range between 7 and 11 is divided into several magnitude intervals (for the current mask table 18 intervals are used). Within each interval the required pixels of all targets with *Kepler* magnitude within that interval are combined to make a single aperture mask that fits all targets within that interval. For targets brighter than magnitude 7, masks are hand-specified to capture their saturation and large cores. The dedicated mask portion of the mask table is empty prior to final mask assignment, and is filled in as described in 4.4.3.

4.4.3 Mask Assignment

Once the mask table is complete these masks can be assigned to target definitions. Because SC targets are also on the LC target list, only the LC list needs to be considered. The majority of target definitions are assigned aperture masks by choosing the mask that contains the fewest excess pixels and all required pixels. Such a mask and the position of the target definition inside the mask is chosen by convolving the target's required pixels (rotated by 180 degrees as required by the definition of convolution) with the aperture mask pixels. If the convolution contains values that are equal to the number of required pixels, the target's pixels fit in the mask. The position of the mask relative to the target pixels is determined by the centroid of all entries in the convolution whose values are equal to the number of required pixels.

If no aperture mask contains the required pixels, the required pixel set is divided in two along the shorter axis and a new attempt is made to find an aperture mask for each piece. This division approach is applied iteratively until masks are found that cover all required pixels.

Dedicated mask targets are not algorithmically assigned masks by the above process. If there is not already an aperture mask that exactly fits the target's required pixels (several targets may use the same dedicated mask), a slot in the dedicated mask region of the aperture mask table (see 4.4.2) is filled in by the required pixels of this target. This process assumes that the dedicated mask portion of the mask table has been sized to accommodate all dedicated mask targets.

Examples of targets and their assigned masks are shown in Figure 4.4. In practice the number of unused pixels downlinked due to masks over-fitting targets is 4%, Mostly from bright stars. Figure 4.5 shows the number of mask pixels per target, the cumulative pixel count, and the number of excess pixels per target as functions of target magnitude. Most of the excess pixels are used by the bright targets while the dimmer targets, which make up the overwhelming bulk of targets, are well fit.

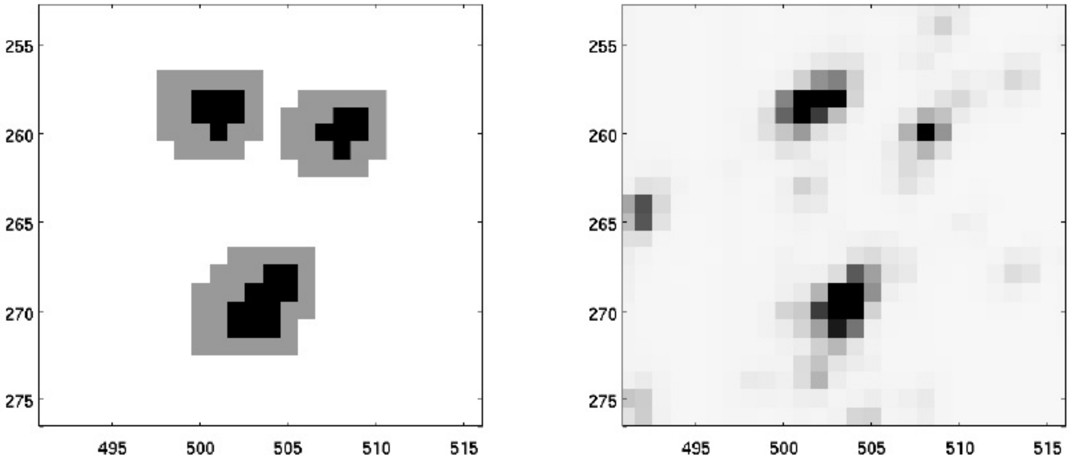


Figure 4.4. Left: Examples of the masks (grey) selected once the halo and undershoot column is added to the optimal aperture (black). Right: A flight image showing the targets captured by the masks on the left.

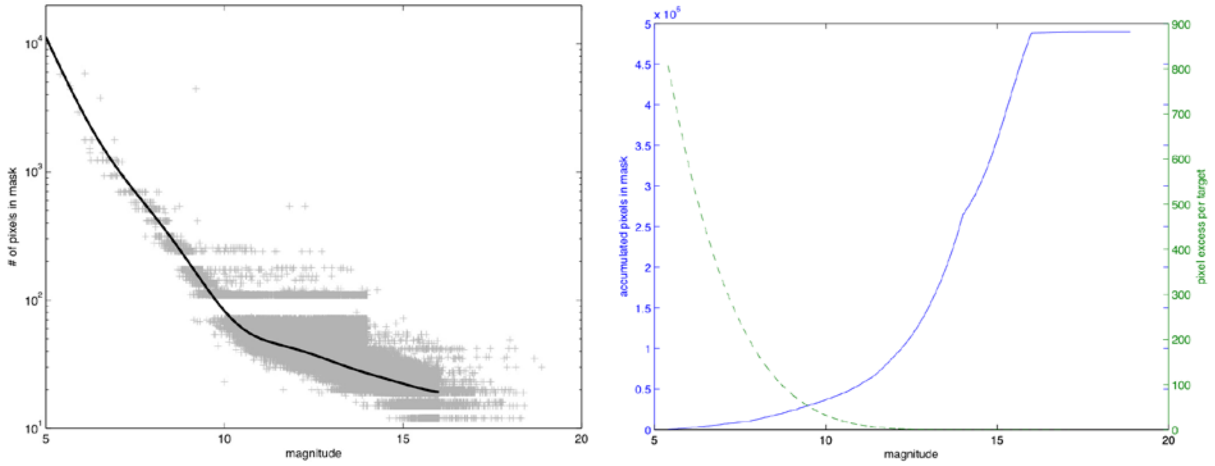
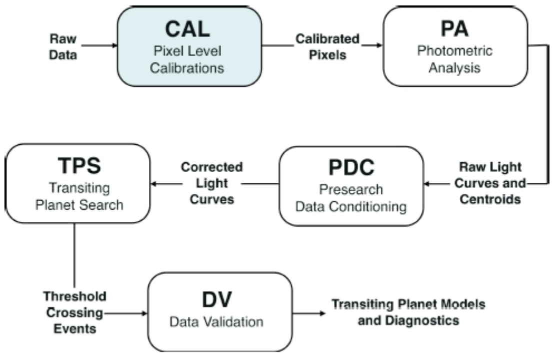


Figure 4.5. Left: The number of pixels per mask vs. target magnitude for stellar targets including pixel halo and undershoot column. The pixel number for each mask is shown by a grey cross, and a tenth order robust polynomial fit showing the mode of the pixel number distribution is shown by the black line. The smallest mask (for a one pixel optimal aperture) is 4 x 3 pixels. Banding for the brighter stars is due to the smaller number of masks available for these stars. Right: Cumulative number of pixels in a target's mask vs. magnitude (solid line, left axis), showing that most pixels are used by dim targets, and the number of excess pixels per target (dashed line, right axis), showing that bright targets account for most of the pixel excess.

5 PIXEL CALIBRATION

5.1 Introduction

In this section we describe the first component of the Pipeline, Calibration (CAL), which processes original flight pixel data provided by the Data Management Center at the Space Telescope Science Institute. Various data types are collected and differ by the number of integrations that compose each sampling time (or "cadence"), and also by the number and location of pixels that are collected. The raw data include photometric (target and background) pixels, along with a subset of the CCD termed "collateral data" which includes masked and virtual (over-clocked) rows and columns that are used primarily for calibration. Three types of data sets are processed within CAL: (1) select pixels from >150,000 long cadence targets that are collected every 29.4 minutes (with 270 exposures per cadence), (2) a smaller set of pixels from ≤ 512 short cadence targets that are sampled more frequently at 0.98-minute intervals (with nine exposures per cadence), and (3) full-frame image data which contain all pixels for a single long cadence. These data types are processed separately in the Pipeline, and the differences will be noted herein.



CAL	Calibrate pixels (collateral, background and target) for long cadence, short cadence, and full frame images
PA	Extract raw flux and compute photocenter (centroid) for each target and cadence from associated target pixels
PDC	Correct systematic and other errors in raw light curves, remove excess flux due to aperture crowding, and condition light curves for the transiting planet search
TPS	Perform transiting planet search and return Threshold Crossing Events (TCEs) for detections
DV	Fit transiting planet model to light curves with TCEs, search for additional transiting planets, and perform statistical tests to validate candidate planets

Figure 5.1. An overview of the science data flow in the SOC Pipeline is shown (left), along with a description of the primary functions of each component (right).

This section describes the data flow within CAL and the methodology and reasoning behind the individual corrections that are applied to the different pixel types. Many of the primary corrections use external models (Allen et al. 2010a) of each CCD that were developed from pre-flight hardware tests and FFI data taken during commissioning (Haas et al. 2010) prior to the dust cover ejection. We discuss how these models are applied within CAL to correct for 2D bias structure, gain, and nonlinearity of the conversion from analog-to-digital units (ADU) to photoelectrons; local detector electronics effects (undershoot and overshoot); and flat field (variations in pixel sensitivity). Other signals that are corrected include excess charge from saturated stars that leak into the masked and virtual regions, cosmic ray events, dark current and smear. Note that CAL does not include any time or motion corrections or coordinate transformations. We present an overview of the focal plane CCD components and the pixels that are calibrated in the § 5.2. In § 5.3 we describe the individual calibration steps, presented in the order that they are performed, along with the additional functionalities of CAL. In § 5.4, we present a summary of the CAL module and discuss future work that will help to improve the quality of the data.

5.2 Data Formats

The *Kepler* focal plane array is composed of 42 charge-coupled device (CCD) detectors (Figure 5.2). A CCD "module" refers to a pair of CCDs that share a field flattener and are read out simultaneously by the detector electronics. Each of the 21 modules is composed of four CCD "outputs" that are each read out by a separate analog signal chain. The CAL software component operates on a single CCD module/output, or "channel," at a time.

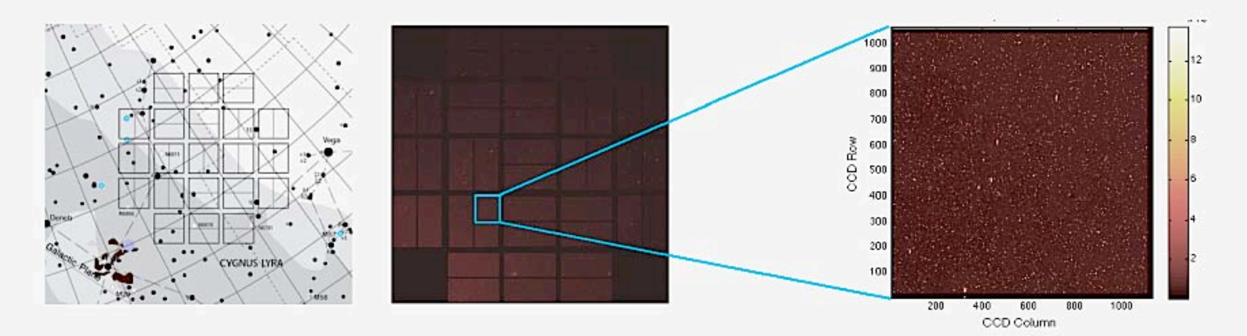


Figure 5.2. A celestial view of the *Kepler* focal plane (left), the first light image (middle), and a calibrated FFI from Q4 data (right, in units of 10^6 electrons/cadence) of module/output = 17/2 (channel 58) are shown.

5.2.1 Pixel Collection

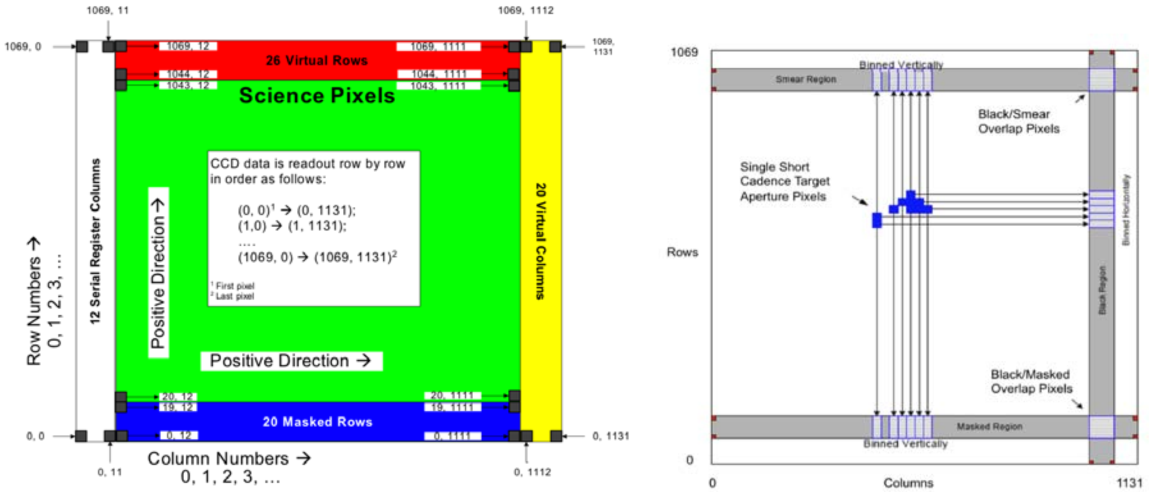
Each CCD channel consists of an array of pixels with 1070 rows and 1132 columns, of which only a subset (1024 x 1100) is photometric (Figure 5.3). The full (1070 x 1132) array of pixels is downlinked for FFI data, whereas only select target and background pixels are downlinked for LC and SC data due to limitations in memory, bandwidth, and the design of the flight software. For LC, an upper limit of 170,000 stellar targets and 1125 background targets are collected across the focal plane (with additional limitations on the number per channel and the total pixel count). In addition, LC collateral data (black, masked smear, and virtual smear pixels, which are described in the next section) are collected for calibration. For SC, a maximum of 512 stellar targets across the focal plane are collected, along with a subset of the collateral pixels: the black rows and smear columns that lie in the projection of the photometric pixels onto the collateral region (Figure 5.3), in addition to the black pixels that lie in the projection of the masked and virtual smear pixels (used to calibrate the smear).

5.2.2 Photometric and Collateral Data

The photometric pixels include all available target and background pixels on a CCD channel. In PA (the Pipeline module that follows CAL), the photometric pixels are packaged into individual stellar and background targets to create target flux time series. Within CAL, however, the target and background pixels are indistinguishable and the calibration steps are processed on the individual pixel flux time series.

The collateral data include the following:

- 12 "leading black" pixels that represent virtual (non-physical) pixels, which are read out before the photometric pixels in each row,



CCD data is readout row by row in order as follows:

(0, 0)[1] → (0, 1131);
(1,0) → (1, 1131);
....
(1069, 0) → (1069, 1131)[2]

[1] First pixel
[2] Last pixel

Figure 5.3. A schematic of the pixel regions in a single CCD channel (left) shows the location of photometric pixels, along with the collateral pixels on the perimeter of the CCD that are collected for calibration. Only a subset of collateral data (black columns and smear rows) is collected and co-added onboard the spacecraft (Van Cleve & Caldwell 2010). For LC data, all black rows (and a subset of columns) and all smear columns (and a subset of rows) are collected (gray region in right panel), whereas for SC only collateral data in the projection of pixels are collected.

- 20 “trailing black” pixels, which are read out after the photometric pixels in each row,
- 20 “masked smear” pixels, which are physical pixels closest to the serial register that are covered with an opaque aluminum mask, and
- 26 “virtual smear” pixels, which are read out after all of the photometric pixels are clocked out.

Only a subset of these pixels is downlinked for calibration, however, designated by the ground segment in a configuration map that is uplinked to the spacecraft. To estimate the black level correction, a subset of the trailing black region (columns 1119-1132) is collected, and the pixels in each row are co-added onboard the spacecraft prior to downlink. Each “black pixel” that CAL receives is therefore the sum of 14 black pixel values per row per cadence. The leading black pixels are not used for calibration due to the presence of image artifacts in those regions. Likewise, a subset of masked smear (rows 7-18) and virtual smear (rows 1047-1058) pixels is collected, resulting in 12 co-added masked smear and 12 co-added virtual smear pixels per column per cadence. For SC data, only a small number of targets are collected from each channel, and the rows/columns of each target determine which collateral pixels are collected.

Two additional pixel types are collected for SC: “masked black” pixels, which are the sum of the pixels in the cross-sections of trailing black columns and masked smear rows, and “virtual black” pixels, which are the sum of the overlapping virtual smear rows and trailing black columns. Each masked black or virtual black pixel is the sum of the number of black pixels (14) times the number of smear pixels (12), or 168 co-adds per cadence. For FFI data, all pixels are downlinked, but CAL uses the spacecraft configuration map to determine which collateral pixels should be used in calibration, and the data are processed as if it the pixels came from a single long cadence.

5.2.3 Processing Order

Data for each CCD channel are calibrated individually. Regardless of cadence type, the collateral pixels are always processed first to estimate the bias, smear, and dark levels. The photometric pixel calibration follows, using results calculated from the collateral data output. For FFIs, the collateral regions are also processed first to obtain the black, smear, and dark level estimates, and then the entire array (collateral plus photometric) is calibrated using these values.

5.2.4 Data Gaps

All pixel data are accompanied by logical arrays (the same size as the pixel arrays) of spatial and temporal gaps, and only the available pixels are calibrated. In some calibration steps, such as the black and dark level estimation, CAL interpolates across missing cadences. The only cases for which data may be gapped *within* CAL include (1) cadences that occur during a momentum dump, (2) cadences that occur when the spacecraft is not in fine point when the spacecraft attitude is not precisely controlled by the fine guidance sensors (Haas et al. 2010), and (3) masked or virtual columns that have excess flux due to saturated stars that bleeds into those columns.

5.3 Calibration

A schematic of the data flow in the CAL module is shown in Figure 5.4 and the primary calibration steps are described in this section. The boxes in Figure 5.4 with dashed lines show the steps that can be disabled in the Pipeline if desired. All cadence types (FFI, LC, and SC) and pixel types (collateral and photometric) are processed separately, but use the same MATLAB code base (with the exceptions noted in this section). For each cadence type and channel, the first invocation processes collateral (black and smear) pixels for all cadences. The outputs to this first pass include calibrated black and smear pixels, collateral and cosmic ray metrics, and more importantly the estimates for black, smear, and dark current for the specific channel that are needed to calibrate the photometric pixels. Due to the large volume of data, the photometric pixels are subdivided by rows for LC processing, since the undershoot correction operates on pixel rows. For SC data, there are typically only 2-5 SC targets on a given channel but far more (30x) cadences than LC data, so we divide the pixels into cadence chunks.

5.3.1 Focal Plane Models

Some of the calibration steps rely on external models that characterize each CCD. These focal plane characteristics (FC) models (Allen et al. 2010) were developed during extensive ground-based tests, were updated in flight while the spacecraft dust cover was still in place, and are continuously monitored by the FC Pipeline module. These time-dependent models include (1) a read noise model, which gives the read noise per channel; (2) a 2D black model, which provides a 2D map of the black/bias structure per channel; (3) a gain model, which gives the ADU-to-photoelectrons conversion factor per channel; (4) a linearity model, which provides a set of polynomial coefficients used to correct for any nonlinearity in the gain transfer function; (5) an undershoot model, which includes filter coefficients that are used to correct for undershoot/overshoot artifacts induced by the CCD local detector electronics (LDE); and (6) a flat field model consisting of a 2D map of values that are used to correct for pixel-to-pixel sensitivity.

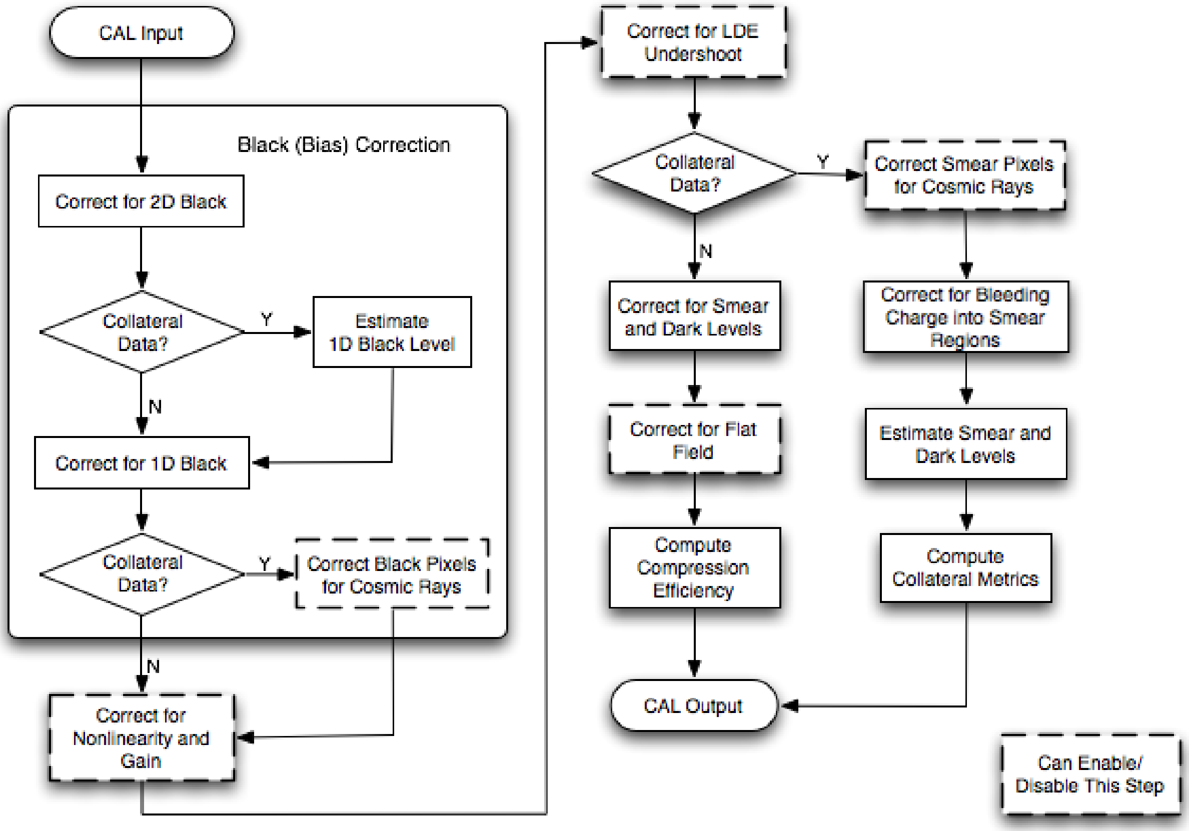

CAL Input
Black (Bias) Correction
Correct for 2D Black
Collateral Data?
Y
Estimate 1D Black Level
N
Correct for 1D Black
Collateral Data?
Y
Correct Black Pixels for Cosmic Rays
N
Correct for Nonlinearity and Gain
Correct for LDE Undershoot
Collateral Data?
Y
Correct Smear Pixels for Cosmic Rays
N
Correct for Smear and Dark Levels
Correct for Bleeding Charge into Smear Regions
Correct for Flat Field
Estimate Smear and Dark Levels
Compute Compression Efficiency
Compute Collateral Metrics
CAL Output
Can Enable/ Disable This Step

Figure 5.4. Data flow for calibrating collateral and photometric pixels. Dashed boxes indicate corrections that can be disabled in the Pipeline.

5.3.2 Fixed Offset, Mean Black, and Spatial Co-adds

Before the pixels are calibrated for black (bias) level, they are corrected for a “fixed offset” and “mean black” value. These values (which vary for SC and LC and across channels) were introduced to deal with spatial variations in bias and gain across the focal plane array, and to address issues with the pixel requantization scheme (Haas et al. 2010). Prior to downlink, all pixels are subject to requantization in which each pixel value is mapped to a discrete value in a pre-generated table in order to control the quantization noise (the round-off error resulting from digitizing the voltage signals) to within ¼ of the intrinsic measurement uncertainty (Caldwell et al. 2010). Because collateral pixels are spatially co-added and fall on a different part of the requantization table, the mean black and fixed offset work by adjusting all channels to a common zero point to ensure proper requantization. Given a pixel array ***P*** (in this case, for either collateral or photometric pixel data), the first correction performed within CAL for the available rows (*row*), columns (*col*), and cadences (*t*) is:

$$P_{all}(row,col,t) = P_{all}(row,col,t) - (fixed\ offset) + (mean\ black(t)). \qquad (5\text{-}1)$$

Note that MATLAB is used for all of the CAL science algorithms, so the operations are often performed on full or partial (n_{rows} x n_{cols} x $n_{cadences}$) pixel arrays rather than looping over any particular dimension. The (*row, col, t*) notation here is meant to help the reader understand which pixels are processed in each step along with the dimensions of the pixel arrays and/or corrections. CAL only operates on the available (non-gapped) rows, columns, and cadences.

The original photometric pixels are in units of ADU/cadence. The collateral pixels, however, need to be normalized by the number of spatial co-adds to convert to the same units:

$$\begin{aligned}
P_{black}\left(row,t\right) &= \frac{P_{black}\left(row,t\right)}{n_{black\,cols}} && \left(LC+SC\ data\right), \\
P_{masked\,smear}\left(col,t\right) &= \frac{P_{masked\,smear}\left(col,t\right)}{n_{masked\,smear\,rows}} && \left(LC+SC\ data\right), \\
P_{virtual\,smear}\left(col,t\right) &= \frac{P_{virtual\,smear}\left(col,t\right)}{n_{virtual\,smear\,rows}} && \left(LC+SC\ data\right), \\
P_{masked\,black}\left(t\right) &= \frac{P_{masked\,black}\left(t\right)}{n_{black\,cols}\cdot n_{virtual\,smear\,rows}} && \left(SC\ data\,only\right), \\
P_{virtual\,black}\left(t\right) &= \frac{P_{virtual\,black}\left(t\right)}{n_{black\,cols}\cdot n_{virtual\,smear\,rows}} && \left(SC\ data\,only\right).
\end{aligned} \tag{5-2}$$

5.3.3 Black Correction

The “black level,” or bias, in each CCD channel is an electronic offset that has been added to the CCD voltage to ensure that positive signals are input into the analog-to-digital converter (ADC). In addition, the black level has a 2D structure which includes various artifacts that were discovered during ground testing of the CCDs. These features are characterized in a 2D black model developed during ground testing and with reverse-clocked images (which omit starlight). Some causes of the image artifacts include heating of the readout electronics, start of line (SOL) transients, and FGS frame transfer and parallel transfer clocking crosstalk signals that are injected into the photometric region as the image is read out (Van Cleve & Caldwell 2010). Figure 5.5 shows an example of a 2D black model (left panel) which displays SOL features near the leading black region, and a close-up view (right panel) shows frame transfer (horizontal bands) and parallel transfer crosstalk (diagonal bands) signals. A 2D black model (in ADU/exposure) is extracted within CAL for each cadence and channel, scaled by the number of exposures, and is simply subtracted off all collateral and photometric pixels:

$$P_{all}\left(row,col,t\right)=P_{all}\left(row,col,t\right)-2Dblack\left(row,col,t\right)\ . \tag{5-3}$$

Once the 2D black level is removed, a fit to the residual bias is used to estimate a 1D black correction. The polynomial model order for the best fit is determined in an iterative fashion using the Modified Akaike Information Criterion (Akaike 1974). A robust fit is first performed to protect from outliers (neglecting charge injection rows in the virtual smear region), and a least squares with known covariance method is used with the computed best polynomial order to produce the fit, or “black correction.” For each cadence, the black correction in a given row is subtracted from all available pixels in that row:

$$P_{all}\left(row,\ t\right)=P_{all}\left(row,\ t\right)-1Dblack\left(row,\ t\right). \tag{5-4}$$

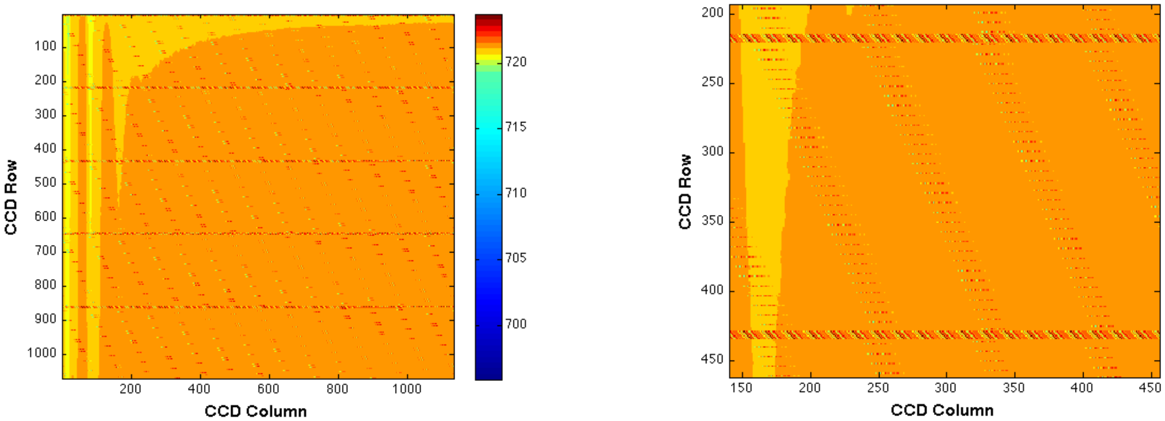


Figure 5.5. An example of a 2D black model (left, in units of ADU/exposure), and a close-up (right), that show the 2D bias structure that is subtracted from all pixels.

Following the 1D black correction, the black pixels are corrected for cosmic rays and saved for output to the Multi-mission Archive at Space Telescope, as they are no longer needed for calibration. Note that CAL only corrects LC and SC collateral pixels for cosmic rays, but uses the same methodology as the photometric pixel cosmic ray correction that is performed within PA .

5.3.4 Nonlinearity and Gain Correction

The gain and nonlinearity describe the transfer function from photoelectrons (e-) in the CCD to ADU coming out of the ADC. Gain is the average slope of the transfer function, and ranges from 94 to 120 e-/ADU across the focal plane (Van Cleve & Caldwell; Caldwell et al. 2010). Nonlinearity is a measure of the deviation from a linear transfer function at each ADU signal level. The nonlinearity model provides polynomial coefficients (Figure 5.6) for each exposure, and the correction can be estimated by evaluating the polynomial at the black-corrected pixel values. The range of this correction across the focal plane is within +/- 3%. The nonlinearity model is valid up until the full-well level, which is the maximum number of electrons a pixel can hold before saturation occurs (~10^6 e–). The gain model provides the gain value per channel and cadence in e-/ADU, and all pixels are simply multiplied by the gain following the nonlinearity correction. At this point, P_{all} represents either smear arrays (in which the rows can be ignored since these are essentially just functions of columns and cadences) or photometric pixel arrays:

$$\begin{aligned} P_{all}(row,col,t) &= P_{all}(row,col,t) - polynomial_{nonlin}(P_{all}(row,col,t)) \,, \\ P_{all}(row,col,t) &= P_{all}(row,col,t)\cdot gain(t) \,. \end{aligned} \tag{5-5}$$

5.3.5 LDE Overshoot / Undershoot Correction

Overshoot and undershoot are signal distortions that were discovered during ground testing of the CCDs, and result from operating a clamp circuit in the local detector electronics (LDE) with insufficient bandwidth (Philbrick 2009). The impulse response artifacts are most noticeable after light-to-dark (undershoot) and dark-to-light (overshoot) transitions, resulting in spikes in the pixel row time series of the affected targets (Figure 5.7). The undistorted image can be reconstructed by modeling these artifacts as a linear shift-invariant (LSI) system, which can be described by a set of difference equations that transforms an input signal *x(n)* into an output signal *y(n)*:

$$\begin{aligned} a(1)\cdot y(n) = b(1)\cdot x(n) + b(2)\cdot n(n-1) + \cdots + b(nb+1)\cdot x(n-nb) \\ - a(2)\cdot y(n-1) - \cdots - a(na+1)\cdot y(n-na)\ . \end{aligned} \tag{5-6}$$

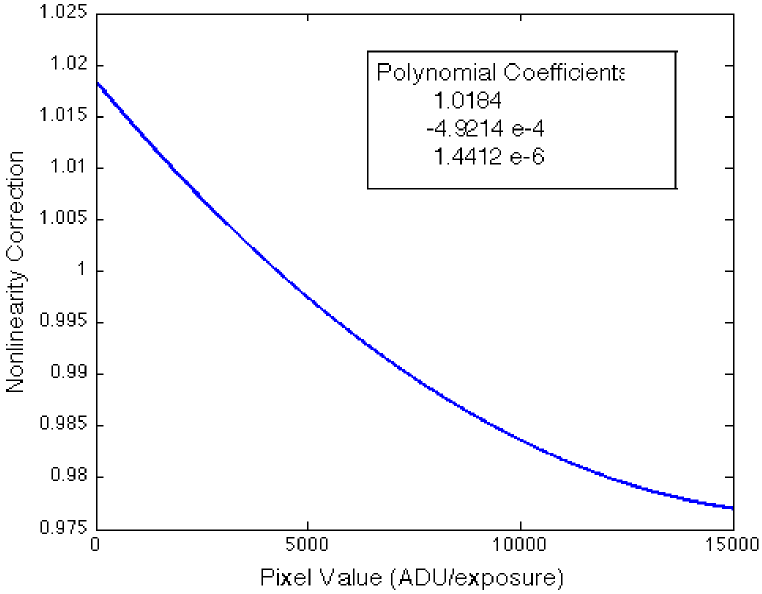


Figure 5.6. The nonlinearity correction, shown for a sample CCD channel, is the fractional deviation from the linear electrons-to-ADU transfer function at each pixel value.

Here ***n***-1 is the filter order, and ***a*** and ***b*** are the feedback and feedforward filter coefficients, respectively, that determine the z-transform system response ***H(z)***:

$$H(z) = \frac{b(1) + b(2)\cdot z^{-1} + \cdots + b(nb+1)\cdot z^{-nb}}{a(1) + a(2)\cdot z^{-1} + \cdots + a(na+1)\cdot z^{-na}}\ . \tag{5-7}$$

The undershoot model provides a set of 20 filter coefficients for ***a***, and an inverse filter is applied (with ***b*** = 1) to each row per cadence to correct for any undershoot/overshoot:

$$P_{all}(row,\ t) = filter\big(b,\ a,\ P_{all}(row,\ t)\big), \tag{5-8}$$

where *filter* is a built-in MATLAB function based on the above difference equations. Note that an extra column to the left (lower column number) of each target aperture is collected and downlinked to perform this correction.

Both collateral and photometric data are corrected for undershoot/overshoot; the median value across the focal plane array (Caldwell et al. 2010) is ~0.34%. In the collateral data invocation, the LC and SC masked and virtual smear pixels are next corrected for cosmic rays and saved for output. These pixels are still used to estimate the smear and dark current levels that are needed to correct the photometric pixels (as described in the next section), but the “calibrated smear” pixels that are output to the MAST consist of the calibrated pixels up to this point.

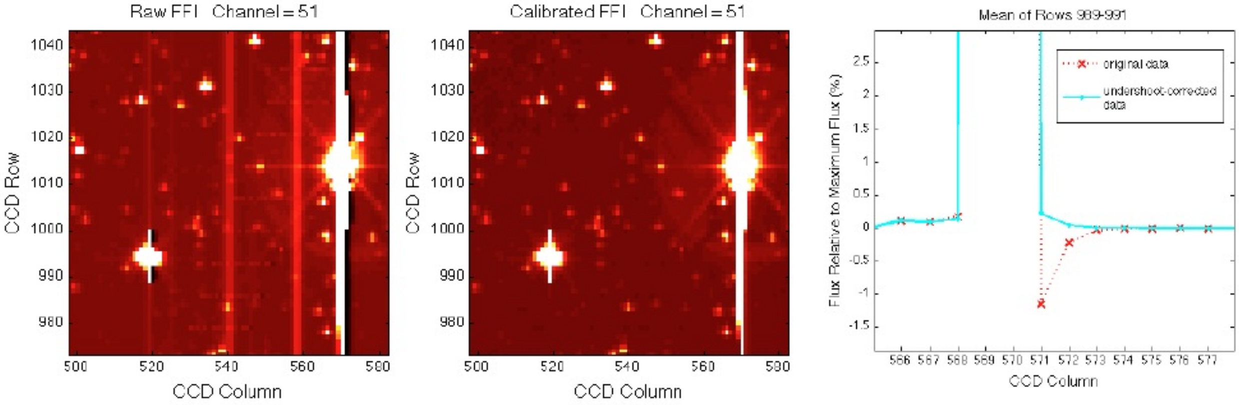


Figure 5.7. A close-up, stretched image of two saturated target stars that show pixel undershoot signatures resulting from bright-to-dark pixel transitions in the direction of the serial readout (left panel), and the calibrated image (middle). The mean of three pixel rows is shown for one target (right) with the undershoot response (the negative spike) along with the corrected pixel values.

5.3.6 Smear and Dark Correction

The target and background pixels are corrected for both smear and dark current levels. The *Kepler* photometer is operated without a shutter, so stars smear along columns as the CCD is read out and are clearly visible in FFI data (Figure 5.8). Dark current is a thermally induced signal in each physical pixel during an integration period, which includes the exposure time ($t_{exposure}$ ~ 6.02 s) and readout time (t_{read} ~ 0.52 s). Because the focal plane is maintained at such a low temperature (-85 degrees C), the dark current is very low with a median value of ~0.25 e-/pixel/sec across the focal plane (Caldwell et al. 2010). The smear and dark corrections are grouped together here because they can be estimated from linear combinations of the virtual and masked smear pixels.

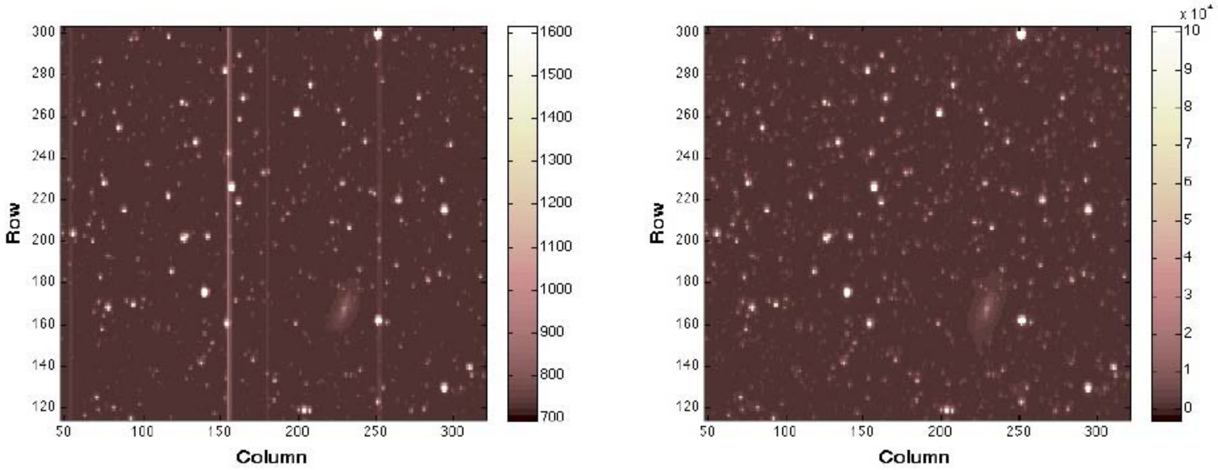


Figure 5.8. A portion of an uncalibrated FFI (in units of ADU/cadence, left) and the calibrated image (in photoelectrons per cadence, right) demonstrate the removal of smear from several columns.

The masked smear pixels, which are shielded from star flux, detect dark current during an integration ($t_{exposure}$ + t_{read}) and collect smear signal from the photometric and over-clocked virtual pixels during readout. The virtual pixels contain dark current that is accumulated during t_{read} only, but collect smear as they are clocked through the image. The dark level per cadence is computed by taking a robust mean of the masked and virtual smear differences from the common columns:

$$dark\ level(t) = mean\left(P_{masked smear}(col,t) - P_{virtual\ smear}(col,t) \cdot \left(\frac{t_{\mathrm{exposure}} + t_{\mathrm{read}}}{t_{\mathrm{exposure}}} \right) \right) . \tag{5-9}$$

We interpolate dark level values over missing cadences to ensure that a dark level is available for all cadences. To compute the smear level, the dark level is first removed from the masked and virtual pixels. Ideally, both masked and virtual pixels are available for each column and cadence, but either may be used if only one is available. If neither is available, however, the smear correction cannot be performed for the entire column. We use the (n_{cols} x $n_{cadences}$) logical gap indicator arrays ***G*** (where gaps = true) that are provided with the smear pixel arrays to estimate the smear levels:

$$\begin{aligned} P_{masked\ smear}(col,t) &= P_{masked\ smear}(col,t) - \left(dark level(t)\right) \cdot G_{masked\ smear}(col,t) \ , \\ P_{virtual\ smear}(col,t) &= P_{virtual\ smear}(col,t) - \left(dark level(t)\right) \cdot G_{virtual\ smear}(col,t) \cdot \left(\frac{t_{\mathrm{read}}}{t_{\mathrm{exposure}} + t_{\mathrm{read}}} \right) . \end{aligned} \tag{5-10}$$

The available smear pixels for each column are tracked using the following logic:

Available Masked	Available Virtual	$C_{masked\ smear}$	$C_{virtual\ smear}$
True	True	1/2	1/2
True	False	1	0
False	True	0	1
False	False	0	0

$C_{masked\ smear}$ and $C_{virtual\ smear}$ are coefficients in the linear combination of the dark-corrected masked and virtual smear pixels (where *G′* are logical arrays with gaps = false):

$$\begin{aligned} C_{masked smear}(col,t) &= \frac{1}{2} G_{masked\ smear}(col,t) \times \left(1 + G'_{virtual\ smear}(col,t)\right) \ , \\ C_{virtual\ smear}(col,t) &= \frac{1}{2} G_{virtual\ smear}(col,t) \times \left(1 + G'_{vmasked\ smear}(col,t)\right) \ , \end{aligned} \tag{5-11}$$

$$smear\,level(col,t) = P_{masked\,smear}(col,t) \times C_{masked\,smear}(col,t) + P_{virtual\,smear}(col,t) \times C_{virtual\,smear}(col,t). \quad (5\text{-}12)$$

The above smear and dark level estimates are computed during the collateral data calibration, resulting in a mean dark level value per channel and an array of smear levels per column per channel. These are later subtracted from the photometric pixels in each column:

$$P_{photometric}(col,t) = P_{photometric}(col,t) - \left(dark\,level(t)\right) - \left(smear\,level(col,t)\right). \quad (5\text{-}13)$$

An additional complication to the smear level estimate is bleeding charge from saturated targets into the masked or virtual smear regions that are clearly visible in FFI data. CAL currently detects and gaps columns that are corrupted by bleeding charge in LC masked or virtual smear data (there are typically only one or two bleeding columns per channel).

5.3.7 Flat Field Correction

The flat field is the final major calibration step, and operates on photometric pixels to correct for spatial and temporal variations in pixel sensitivity to a uniform light source. Differences in pixel response can be due to variations in quantum efficiency or throughput changes in the field flattener lenses or anti-reflection coating of the CCD. The flat field model includes a geometric large-scale map combined with a small-scale (pixel-to-pixel) flat field map that is computed using a 9x9-pixel high-pass filter. The values represent the percent deviation from the local mean with a median value across the focal plane (Caldwell et al. 2010) of ~0.96%, and the 2D flat field model is divided from the appropriate photometric pixels for each cadence:

$$P_{photometric}(row,\ col,\ t) = \frac{P_{photometric}(row,\ col,\ t)}{flat\,field(row,\ col,\ t)}. \quad (5\text{-}14)$$

5.3.8 Additional Functionality in CAL

At the end of the last invocation of CAL for each CCD channel, the theoretical and achieved compression efficiency of the data is computed. These metrics, along with time series of black, smear, and dark level metrics also computed within CAL, are used by the Photometer Performance Assessment (PPA) module (Li et al. 2010) to track and trend data. The uncertainties can be computed within CAL by the propagation of uncertainties (POU) module (Clarke et al. 2010). The primary noise sources for *Kepler* include read noise, which is an additive noise source due to the readout process, quantization noise that is stochastic and results from quantization in the ADC and pixel requantization, and Poisson-like shot noise. The uncertainties in the raw pixel data are computed at the start of CAL, and (if enabled) POU runs in parallel with CAL and the uncertainties are propagated at each transformation step. If POU is disabled, the outputs to CAL are the raw uncertainties corrected only for gain.

5.4 Summary

We have described the pixel-level corrections that are performed in the CAL Pipeline for LC, SC, and FFI flight data. The data corrections include 2D and 1D black, gain, nonlinearity, undershoot and overshoot distortions from the LDE electronics, cosmic rays, bleeding charge, dark current, smear, and flat field variations. The algorithms were validated using simulated flight data from the End-To-End-Model (Bryson et al. 2010b) that was developed in the SOC, which simulates every layer of the data – from CCD and instrument artifacts to transit light curves – and has proven to be a powerful tool in the development and testing of the Pipeline modules. Output from CAL that is exported to the MAST includes raw and calibrated black, smear, and photometric pixels, along with the associated gap indicators and uncertainties. Additional metrics for cosmic ray detection, black level, smear level, and dark current level estimates are also provided.

Some issues that may be addressed in future versions of CAL include incorporating a more detailed time-varying 2D black model, refining the cosmic ray algorithms, and addressing bleeding charge in SC data. Improving the performance of the algorithms is also a high priority because of the large volume of data that is processed each quarter (Q1, for example, yielded ~2TB for LC and SC CAL output alone), and the fact that the SOC will be reprocessing data throughout the mission.

6 PHOTOMETRIC ANALYSIS

6.1 Introduction

We describe the Photometric Analysis (PA) software component and its context within the *Kepler* science processing Pipeline. The primary tasks of this module are to compute the photometric flux and photocenters (centroids) for over 160,000 long cadence (~thirty minute) and 512 short cadence (~one minute) stellar targets from the calibrated pixels in their respective apertures. We discuss science algorithms for long and short cadence photometry: cosmic ray cleaning; background estimation and removal; aperture photometry; and flux-weighted centroiding. We discuss the end-to-end propagation of uncertainties for the science algorithms. Finally, we present examples of photometric apertures, raw flux light curves, and centroid time series from *Kepler* flight data. Introduction

Science data acquired in-flight are processed in the *Kepler* science processing Pipeline (Jenkins et al. 2010a; Middour et al. 2010). Pixel values are calibrated for each cadence in the Calibration software component, described in section 5. Raw flux light curves are extracted and target photocenters (centroids) are computed in the PA component. The first primary task of PA is to extract raw flux light curves from calibrated pixels in the apertures of the respective long and short cadence targets. Prior to computation of the photometric flux for each target, so-called Argabrightening events are (optionally) mitigated, cosmic rays are (optionally) removed, and a background estimate is subtracted from the pixels in the target apertures. The second primary task of PA is to estimate CCD coordinates of the aperture photocenter (centroid) for each target and cadence from background-removed pixels associated with the respective targets. Other tasks of PA are to compute barycentric timestamp corrections per target and cadence based on a spacecraft trajectory reconstruction, and to compute metrics (brightness and encircled energy) for monitoring instrument performance and, potentially, to support systematic error correction in the PDC component.

An overview of PA and the flow of data through the Pipeline module are presented in § 6.2. PA science algorithms are described in § 6.3; cosmic ray cleaning is discussed in § 6.3.1; background estimation and removal in § 6.3.1; aperture photometry in § 6.3.1; and flux-weighted centroiding in § 6.3.1.

6.2 Overview and Data Flow

The primary tasks of the PA Pipeline module are to compute raw flux light curves and centroids of target stars from calibrated pixels in their respective apertures. The standard PA unit of work (Klaus et al. 2010a, 2010b) for LC and SC science data processing is a single module output for a duration of one quarter or one month. PA science data includes calibrated background and target pixels. Due to the large volume of calibrated pixel data, PA is typically executed in multiple invocations (Klaus et al. 2010b). For LC units of work, all of the background pixel data are processed in the first PA invocation. Background pixels are utilized to estimate the background level to be subtracted from each target pixel on each cadence. Target pixels are then processed in as many subsequent invocations as required. Calibrated pixel data for targets with PPA_STELLAR labels are processed first in PA (although they may span multiple PA invocations). After all PPA_STELLAR pixels have been processed, then calibrated pixels for targets without PPA_STELLAR labels are processed. Targets with the PPA_STELLAR label represent gold standards for motion polynomial fitting, centroid seeding, and computation of PA metrics. Approximately 200 of these are pre-selected per module output. There are no background pixels in SC units of work, so the processing of SC science data is restricted to target pixels only.

Data flow in the PA Pipeline module is shown in Figure 6.1. Data generally flow from left to right and top to bottom in the figure. Functional blocks in the diagram are referenced with bold type where they are discussed in the text, and section numbers are included for the algorithms that are described in detail later in this section.

Mitigate Argabrightening Events. It has been observed in flight data that there are cadences for which the flux values of all background and target pixels are elevated (Jenkins et al. 2010b). These have been dubbed *Argabrightening* events after Vic Argabright, the Ball Aerospace engineer who first observed and reported them. Examination of the affected images suggests that these brightenings can be attributed to nearby particles passing across the telescope FOV, possibly induced by micro-meteoroid impacts (Witteborn et al. 2011). If Argabrightening mitigation is enabled by the PDC module parameter' then Argabrightening detection statistics are formulated for each cadence and a threshold is applied to identify the affected cadences. All background and target pixel time series are subsequently gapped for Argabrightening cadences in each of the PA invocations in the unit of work. Argabrightening statistics are formulated from the background pixels for LC units of work and from target pixels outside of the optimal apertures of the respective targets in the first invocation for SC units of work.

Clean Cosmic Rays (6.3.1). If cosmic ray cleaning is enabled, cosmic rays are identified and removed from the calibrated background (LC only) and target pixels. Cosmic ray metrics are computed separately from the background and target cosmic ray event lists. Background cosmic ray events and metrics are written to the Pipeline Data Store after the background pixels have been processed in the first invocation of LC units of work. Target cosmic ray events and metrics are written to the Data Store after the last target invocation in all PA units of work.

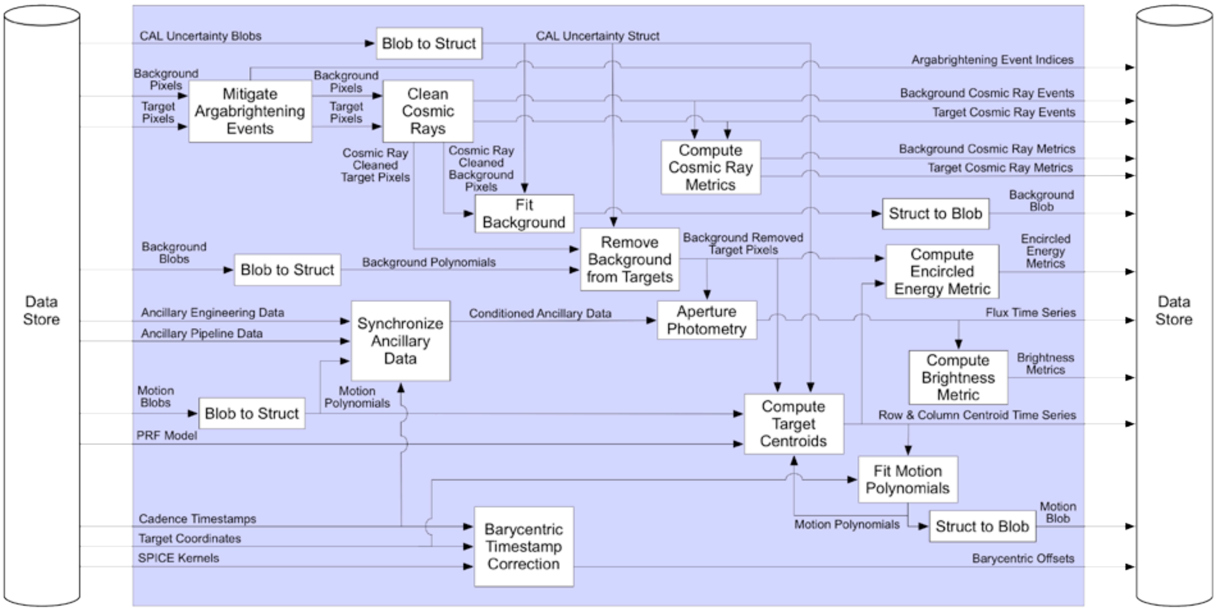


Figure 6.1. Data flow diagram for the Photometric Analysis Pipeline module. Inputs are shown at the left and outputs are shown at the right. Inputs are obtained from the Data Store and outputs are written to the Data Store.

Fit Background and Remove Background from Targets (6.3.2). For LC units of work, a two-dimensional (background) polynomial is fitted to the cosmic ray-cleaned background pixels on a cadence-by-cadence basis. The background is removed from cosmic-ray-cleaned target pixels by evaluating the background polynomial at the coordinates of the given target pixels and subtracting the resulting background estimates from the target pixel values on a cadence-by-cadence basis. For SC units of work, background polynomials must be normalized and interpolated at the short cadence timestamps before they can be evaluated to estimate the background.

Aperture Photometry (6.3.3). Flux time series are obtained by performing aperture photometry on the background-removed target pixels. In Simple Aperture Photometry (SAP), pixel values are summed for all pixels in the optimal aperture of a given target. SAP is the only photometry method currently supported by PA.

Compute Target Centroids (6.3.4). Row and column centroid time series are computed from the background-removed target pixels on a cadence by cadence basis. Flux-weighted centroids are computed for all PA targets. Centroids based on fitting a predetermined Pixel Response Function (PRF) (Bryson et al. 2010c) to calibrated target pixel values in the respective target apertures may also be computed for PPA_STELLAR or other targets. PRF-centroid light curves are not exported to MAST.

Fit Motion Polynomials. Two-dimensional (motion) polynomials are separately fitted to the collection of row and column centroids of the PPA_STELLAR targets (if a sufficient number of such targets are available) as a function of target right ascension and declination. The motion polynomials are then evaluated at the celestial coordinates of the non-PPA_STELLAR targets in order to seed the PRF-based centroid computation if PRF-based centroiding is enabled for such targets. The motion polynomials are written to the Data Store. In short cadence units of work, motion polynomials are obtained for centroid seeding by interpolation of long cadence motion polynomials provided to PA from the Data Store. The motion polynomials are computed cadence by cadence. In addition to providing seeds for centroids, motion polynomials are utilized elsewhere in the Pipeline for focal plane geometry fitting (Tenenbaum and Jenkins 2010), systematic error correction, attitude determination, and computation of instrument performance metrics (Li et al. 2010).

Compute Encircled Energy Metric and Compute Brightness Metric. The encircled energy metric is computed for each cadence from the background-removed target pixels and centroids of the PPA_STELLAR targets. This metric is a robust measure of the average radius required to capture a specified fraction of flux from the respective targets. The brightness metric is computed for each cadence from the raw flux light curves for the PPA_STELLAR targets. This metric is a robust measure of the average ratio of observed-to-estimated flux for the given targets. The raw flux light curves, row and column centroids, and encircled energy and brightness metrics are all written to the Data Store.

Uncertainties in calibrated background and target pixels are propagated by standard methods to uncertainties in the raw flux, centroids, metrics, and two-dimensional polynomials generated in PA. Uncertainties propagated in PA are also written to the Data Store. If rigorous Propagation of Uncertainties (POU) is enabled, uncertainties are propagated from covariance matrices for calibrated background and target pixels (Clarke et al. 2010). Otherwise, uncertainties in PA products are propagated only from uncertainties in calibrated background and target pixels under the assumption of statistical independence.

Barycentric Timestamp Correction. Barycentric timestamp corrections are computed for each target and cadence given the celestial coordinates of the target and a reconstructed spacecraft trajectory. The offset for each mid-cadence timestamp is computed such that the sum of timestamp plus offset yields the time that the flux would have been captured at the solar system barycenter. The barycentric offsets allow investigators to account for modulation of the timing of observed transits due to the heliocentric orbit of the photometer. Barycentric corrections produced in PA also compensate for small time slice offsets introduced in the multiplexed readout of the focal plane array (KIH chapter 5, pg. 64).

6.3 Science Algorithms

6.3.1 Cosmic Ray Removal

Cosmic rays have been problematic for all space flight missions with CCD detectors. The expected cosmic ray flux rate (Jenkins et al. 2010b) for the *Kepler* CCD array is 5 $\text{cm}^{-2}\ \text{sec}^{-1}$. Each 27-$\mu$m pixel is therefore expected to receive a direct hit approximately three times per day. Energy may be deposited over a range of pixels for each hit, however, depending on the angle of incidence.

In the PA Pipeline module, cosmic rays are cleaned (identified and subtracted) in both background and target pixels after Argabrightening events have been mitigated. Cosmic ray identification in the relatively low flux background pixels is more effective than in target pixels where deposited energy is often a small fraction of the total flux. Most cosmic rays are not expected to be detected, but the desire is to detect the cosmic rays with significant charge with a low false detection rate.

A flow chart describing the cosmic ray cleaning algorithm is shown in Figure 6.2. The cosmic ray identification process for PA background and target pixels is a general pixel time series outlier identification algorithm. To that end, both positive and negative outliers are identified. The algorithm therefore goes beyond identification of cosmic rays only, as these are restricted to deposits of positive charge. A parameter has been included in the module interface (Klaus et al. 2010a, 2010b), however, to define a threshold multiplier for negative events. This multiplier may be specified to be arbitrarily large; in that event, the algorithm is insensitive to negative flux outliers.

A two-dimensional array of all background or target pixel time series in a given PA invocation is first detrended along the time dimension with a low order polynomial. The detrended time series for the respective pixels are smoothed with a short median filter (relative to typical astrophysical events and stellar variations). The residuals (cosmic ray deltas) are then computed between the detrended time series and the time series obtained by median filtering each (detrended) pixel time series. A sliding Median Absolute Deviation (MAD) is computed on the residuals for each pixel time series to obtain a time-varying estimate of the residual scatter, and detection statistics are formulated by computing the ratio of the residuals to the MAD for each pixel and cadence. Outliers are identified by applying a threshold to the absolute value of the detection statistics.

Once the residual values that exceed the detection threshold have been identified, the number of outlier counts per pixel time series is determined. The pixels for which there are an excessive number of counts are identified and reprocessed as before with a longer duration median filter. The motivation for doing so is to prevent out-of-family cosmic ray identification based on the nature and variability of specific targets. Implicit is the assumption that cosmic rays or other outliers should be distributed uniformly among all background or target pixels.

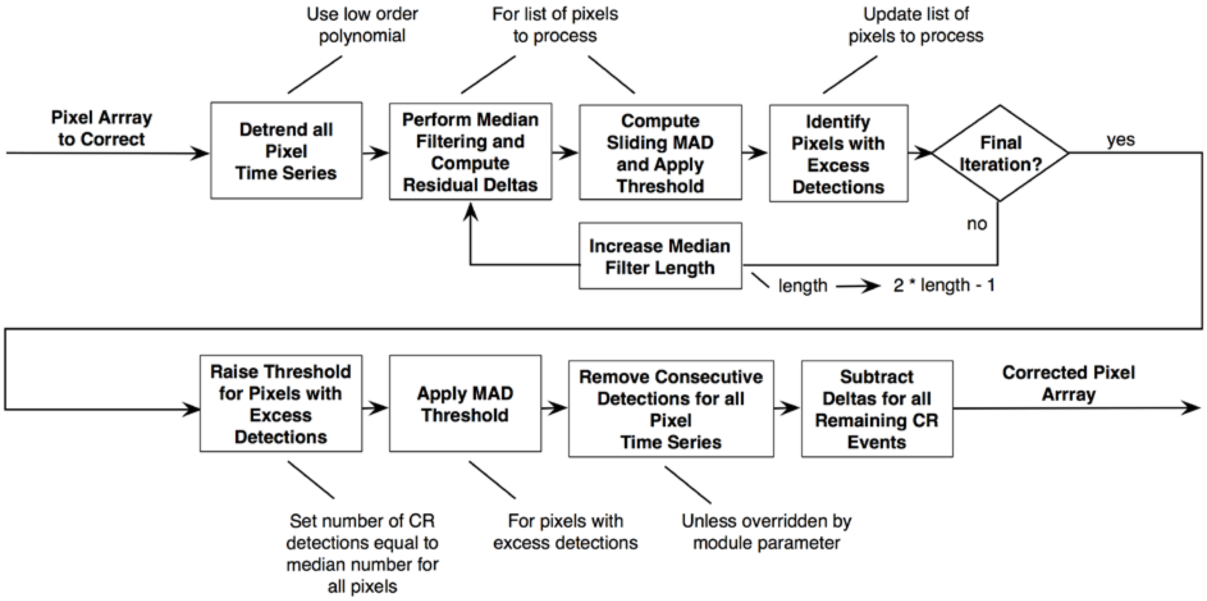

Pixel Arrray to Correct
Use low order polynomial
Detrend all Pixel Time Series
For list of pixels to process
Perform Median Filtering and Compute Residual Deltas
Compute Sliding MAD and Apply Threshold
Update list of pixels to process
Identify Pixels with Excess Detections
Final Iteration?
yes
no
Increase Median Filter Length
length → 2 * length - 1
Raise Threshold for Pixels with Excess Detections
Apply MAD Threshold
Remove Consecutive Detections for all Pixel Time Series
Subtract Deltas for all Remaining CR Events
Corrected Pixel Arrray
Set number of CR detections equal to median number for all pixels
For pixels with excess detections
Unless overridden by module parameter

Figure 6.2. Flow chart for the cosmic ray cleaning algorithm. The loop in the upper path is executed twice in the current PA release.

After the process has repeated, the detection threshold is raised for those pixels that continue to have an excessive number of hits (in relation to the ensemble of background or target pixels in the given invocation). The number of cosmic rays identified for each of these pixels is set equal to the median number across the ensemble. Once outliers have been identified for all pixels and cadences, temporally consecutive events are invalidated as cosmic ray detections. Consecutive cosmic ray hits in a single pixel time series are extremely unlikely and almost certainly indicative of false detections. Finally, the residuals (cosmic ray deltas) after median filtering are subtracted from the original pixel array for the valid remaining cosmic ray events to obtain an array of cosmic ray-cleaned pixel values.

Cosmic ray detection statistics for a Q3 target pixel are shown in Figure 6.3 along with the pixel time series before and after cosmic ray cleaning. Cosmic ray events above the specified 12 MAD detection threshold are circled in the original time series. The difference between the original and corrected pixel value for each cadence with an identified cosmic ray is equal to the residual (or delta) after median filtering. For each cosmic ray event, the mid-cadence timestamp is written to the Data Store along with the row and column coordinates of the pixel and the cosmic ray delta. Duplicate detections are first removed from the cosmic ray events lists. These may result if a pixel falls in multiple target apertures.

6.3.2 Background Estimation and Removal

Background pixels are acquired at the long cadence rate in a grid pattern on each of the focal plane module outputs. 4,464 background pixels are defined on each channel (see Figure 4.3). The background pixel time series represent spatial and temporal samples of the global background level and an attempt is made to prevent them from being corrupted by flux from neighboring stars or from saturated targets on the same columns (section 4). Nevertheless, there are background pixels in the flight data that exhibit high flux levels due to nearby or saturated targets. These pixels necessitate the development of a robust (against outliers) method for background estimation and subtraction.

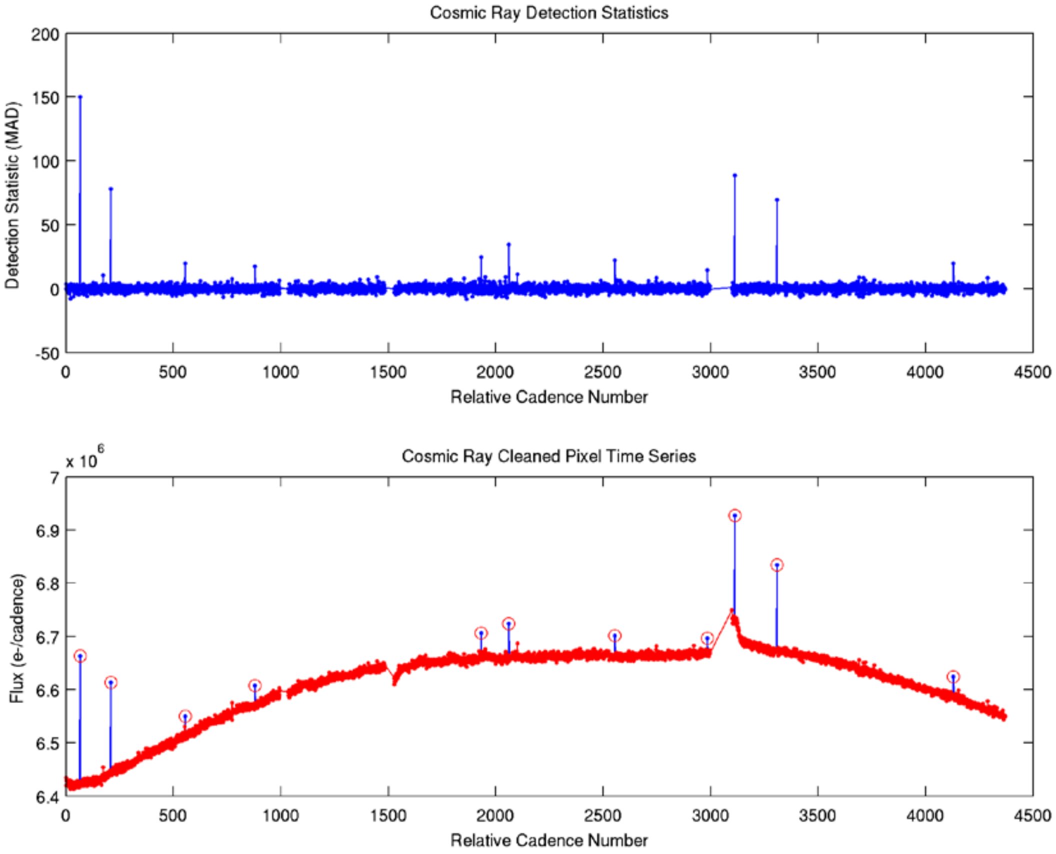


Figure 6.3. Cosmic ray detection statistics for a single Q3 pixel time series are shown in the upper panel. The target pixel time series before and after cleaning are shown in the lower panel. Cosmic ray events in the original pixel time series are circled. Flux transients following the monthly downlinks near cadences 1500 and 3000 are thermally induced.

Background estimation is performed in two steps. The process begins by fitting a two-dimensional (background) polynomial to the calibrated, cosmic ray-corrected background pixel values as a function of the CCD row and column coordinates for each cadence. This occurs on the first PA invocation in each long cadence unit of work. In all subsequent target invocations, the background is estimated by evaluating the background polynomial for each cadence at the CCD coordinates of the respective target pixels. This produces a spatially smooth estimate of the local background for each target without dedicating any background pixels to specific targets. In fact, the ratio of background pixels to targets is only on the order of 2:1 across the entire focal plane.

Once the background is estimated for each target pixel and cadence, the background is removed by subtracting the estimated value from the calibrated, cosmic ray-corrected target pixel value. Uncertainties in the background-removed target pixels are propagated from uncertainties in the background and target pixels by standard methods. The covariance matrix for the background polynomial is determined from uncertainties in the background pixels for each long cadence. The background polynomial covariance matrix is then propagated through the background estimation and subtraction processes. Uncertainties may be propagated in an approximate (but computationally inexpensive) fashion, or in a rigorous treatment (that utilizes, as a starting point, full covariance matrices (Clarke et al. 2010) for the background pixels and target pixels in each stellar aperture) depending on the logical value of a PA module parameter.

Background removal for short cadence targets is complicated by the fact that there are no short cadence background pixels. In short cadence units of work, the long cadence background polynomials are provided as input to PA on the first target invocation. The background polynomials are then normalized and interpolated in time at the midpoints of the short cadence intervals. The background estimation and removal process proceeds with the interpolated background polynomials as before. It should be noted that changes in the background that occur on time scales shorter than the long cadence rate cannot be captured in the short cadence background estimation process.

The order for the two-dimensional polynomial that is fit to the background pixels for each long cadence is determined by default in the Pipeline with the Akaike Information Criterion (AIC) (Akaike 1974). This criterion includes a penalty that increases with fit order and seeks to optimize the trade-off between fit order and goodness of fit. In PA, the order that is used for the background fit for all cadences is determined by the order (up to a specified maximum) that minimizes average AIC over all cadences in the unit of work. For a background fit of order *K*, the number of background polynomial coefficients for each cadence is *(K+1) * (K + 2) / 2*.

For the set of calibrated, cosmic ray-corrected background pixels ***B***, let p_B and σ_B designate the pixel values and uncertainties for any given cadence such that $p_{B,n}$ and $\sigma_{B,n}$ represent the value and uncertainty of the n^{th} background pixel. Furthermore, let the (one-based) CCD row and column coordinates of the background pixels be designated by i_B and j_B respectively. The background polynomial x_{BP} is then determined for order K_B by minimizing the weighted χ^2 defined in equations (6-1) and (6-2) for the given cadence:

$$\chi_B^2 = \sum_{n \in B} \left[\left(p_{B,n} - \sum_{k=0}^{K_B} \sum_{l=0}^{k} x_{BP,m} \; i_{B,n}^{k-l} \, j_{B,n}^{l} \right) / \; \sigma_{B,n} \right]^2 , \qquad (6\text{-}1)$$

where the background polynomial coefficient index is given by

$$m = k(k+1)/2 + l \, . \qquad (6\text{-}2)$$

For the set of pixels ***P*** in a specified target aperture, the background level may then be estimated for the given cadence from the background polynomial x_{BP} and subtracted from the calibrated, cosmic ray-corrected target pixels p_T to obtain the background-removed target pixel values p_S as follows:

$$p_S = p_T - p_{\hat{B}} , \qquad (6\text{-}3)$$

where the CCD row and column coordinates of the target pixels are designated i_T and j_T respectively, and the background estimate for the n^{th} target pixel in ***P*** is given by

$$p_{\hat{B},n} = \sum_{k=0}^{K_B} \sum_{l=0}^{k} x_{BP,m} \; i_{T,n}^{k-l} \, j_{T,n}^{l} \, . \qquad (6\text{-}4)$$

The background polynomial coefficient index *m* is defined as in equation (6-2).

The weighted least squares problem defined in equation (6-1) is solved in an iterative, robust manner in order to deemphasize outlier background pixel values that would otherwise perturb the two-dimensional fit. The problem may still be posed for each cadence as one of *Ax = b.* The least squares solution is the familiar *x = Tb*, where $T = (A'A)^{-1}A'$. If the rigorous propagation of uncertainties is enabled in PA, the covariance matrix C_{BP} for the background polynomial x_{BP} is determined from the covariance matrix C_B of the background pixels (Clarke et al. 2010) and the background polynomial transformation matrix T_{BP} for the given cadence by

$$C_{BP} = T_{BP} C_B T_{BP}'. \tag{6-5}$$

Uncertainties in the background polynomial coefficients are then propagated to the background-removed pixels in the target aperture for the given cadence by

$$C_S = C_T + \hat{A}_T C_{BP} \hat{A}_T', \tag{6-6}$$

where C_T is the covariance matrix for the cosmic ray-corrected target pixels (Clarke et al. 2010), and the linear transformation to evaluate the background polynomial in the target aperture is given by

$$p_{\hat{B}} = \hat{A}_T x_{BP}. \tag{6-7}$$

6.3.3 Aperture Photometry

It is not possible to save and later downlink all of the pixel values acquired by the focal plane array. The storage and bandwidth required to support that are excessive. Rather, an aperture is defined for each target that specifies the pixels necessary to support pixel-level calibrations, light curve extraction, and computation of the target photocenter. Only pixels in the specified target apertures plus the background and collateral pixels for black level and smear corrections in CAL are written to the Solid State Recorder aboard the spacecraft and downlinked for Pipeline processing. Great care is taken in generation of the aperture definitions to ensure that all of the pixels required to support the *Kepler* science mission are captured. Pipeline software may be enhanced over time, but the science pixels may only be acquired once.

Within each target aperture, the subset of pixels required for photometric extraction of light curves is referred to as the *optimal aperture*. The size of the optimal aperture for any given target depends on a number of factors, including target magnitude, PRF, noise level, local crowding and Differential Velocity Aberration (DVA). The apertures and associated optimal apertures are redefined for each observing season as the targets move from one CCD to another with each quarterly roll of the photometer. Images illustrating the mean flux level for pixels (after background subtraction) in two sample target apertures are shown in Figure 6.4. The pixels in the optimal aperture for each target are marked with circles. The *Kepler* magnitude (*Kp*) is 13 for the PPA_STELLAR target on the left, and 10 for the saturated target on the right. Charge bleeds along the column where the flux concentration is highest for the saturated target. The optimal aperture is stretched vertically to permit flux to be captured in one saturated column or another as the target moves due to DVA.

The optimal aperture does not necessarily contain all of the stellar flux for a given target. The *flux fraction* in the aperture refers to the ratio of target flux contained in the optimal aperture to the total flux of the target. Furthermore, not all flux in the optimal aperture is due to the primary target. The *crowding metric* refers to the fraction of flux in the optimal aperture that is due to the target. The flux fractions and crowding metrics are computed for each target table when the apertures are defined (see Section 4) and are provided to the Pipeline modules that require them. Excess flux due to crowding in the optimal apertures is removed when the light curves are corrected in PDC.

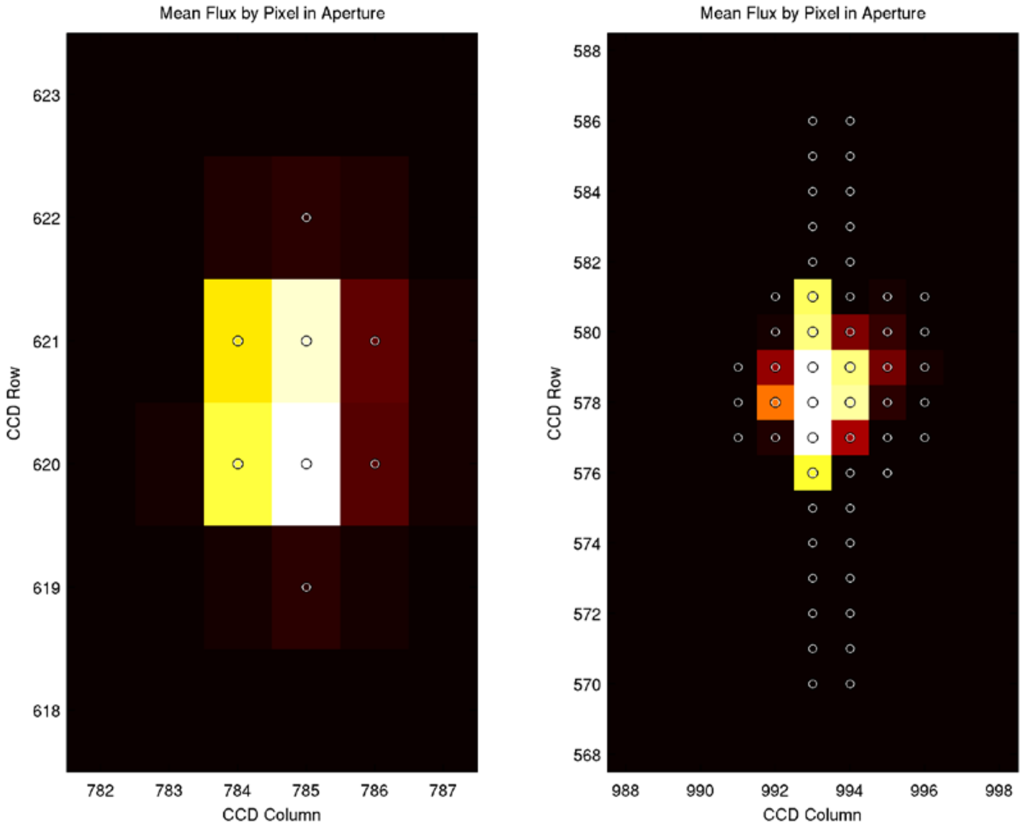


Figure 6.4. Mean flux levels for pixels in two target apertures. The PPA_STELLAR target on the left is 13th magnitude. The saturated target on the right is 10th magnitude. Pixels in the optimal aperture for each target are indicated with circles. The color map range for the saturated target is nearly ten times larger than that of the PPA_STELLAR target.

Light curves are computed in PA by a method referred to as Simple Aperture Photometry (SAP). It is possible that additional photometric methods will be supported in future releases. Once the calibrated target pixels have had cosmic rays corrected and background removed, the raw flux is obtained per target and cadence by the unweighted summation of pixels in the associated optimal aperture. For the set of target pixels ***O*** in the optimal aperture of a specified target, the raw flux f_R is computed by SAP from the background removed pixels p_S in the target aperture for a given cadence by

$$f_R = \sum_{n \in O} p_{S,n} = t_{SAP} p_S, \tag{6-8}$$

where the n^{th} element of the row vector t_{SAP} is equal to one if the n^{th} pixel is in the optimal aperture ***O***, or zero otherwise.

It should be noted that if a data gap exists for any pixel and cadence in the optimal aperture of a given target, then a data gap is set for the raw flux value (and associated uncertainty) for that target and cadence. Discontinuities cannot be artificially introduced into the light curves by extracting the raw flux from only a subset of the pixels in the optimal aperture. As this is written, it has not been observed in-flight that data availability is pixel dependent. Rather, it has always been true that either all or none of the pixels are valid for a given cadence.

Some representative light curves for long cadence targets are shown in Figure 6.5. Raw flux is plotted versus cadence number for all cadences in the unit of work (Q3 for module output 7.3). A quiet target is shown in the upper panel, a variable target in the middle panel, and an eclipsing binary is shown at the bottom. Multiple-cadence data gaps are due to loss of fine point and two monthly downlinks. Single-cadence gaps due to momentum desaturations and mitigated Argabrightening events are also present (but not visible) in the light curves.

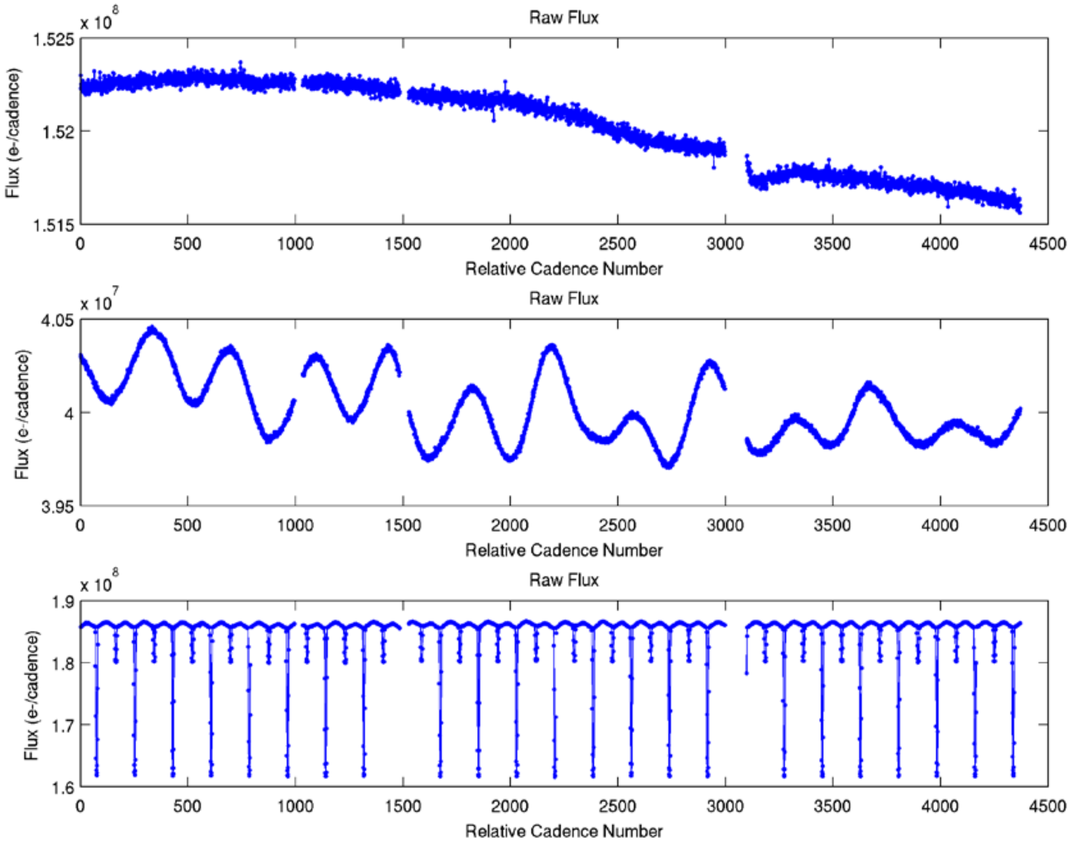


Figure 6.5. Raw flux light curves for representative targets computed by SAP in Q3 for module output 7.3. A quiet PPA_STELLAR target is shown in the upper panel, a variable target in the middle panel, and an eclipsing binary in the lower panel.

If the rigorous propagation of uncertainties is enabled, the variance v_F for the raw flux value computed by SAP in equation (6-8) is determined from the covariance matrix C_S for all pixels in the target aperture on the given cadence by

$$v_F = t_{SAP} C_S t'_{SAP}. \tag{6-9}$$

6.3.4 Centroiding

There is substantial motion of the target positions on the focal plane in the flight science data. The dominant source of long-term target motion is DVA, which causes the targets to trace small but significant elliptical paths across the respective CCD detectors over the period of the heliocentric orbit of the photometer. The maximum motion due to DVA is 0.6 pixels per observing quarter (Jenkins et al. 2010b). There is also target movement due to pointing jitter, pointing drift, focus changes (that appear to result from temperature changes in the photometer), and commanded attitude adjustments to compensate for pointing drift. It should be noted that the photocenters of variable targets also move in crowded apertures.

Flux levels vary with target motion at even the sub-pixel level. Systematic effects that result in target motion therefore produce correlated signatures in the associated light curves. Precise computation of the target locations on each cadence is critical for correcting systematic errors later in the Pipeline. It is also required for precise a posteriori definition of optimal target apertures, for precision reconstruction of spacecraft attitude, and for monitoring instrument performance. The photocenter of each target in its aperture is referred to as the target *centroid*. It should also be noted that analysis of centroid motion (section 9) is a critical tool for distinguishing between legitimate transits due to orbiting planets and apparent transits due to background eclipsing binaries.

Two centroiding methods are employed in PA. Flux-weighted centroids are computed for every target and cadence and are currently exported to the MAST. These centroids are computationally inexpensive and essentially determine the photometric center of mass in the target aperture. Flux-weighted centroids are computed in an aperture that includes the optimal aperture plus a single halo ring for each target. PRF-based centroids are computed in an aperture that includes all of the available pixels for each target. In both cases, uncertainties are propagated to the computed centroid row and column coordinates based on the Jacobian for each centroid computation.

Every centroid computed in PA is validated against the bounding box of the associated centroid aperture, and each is gapped in the event that it does not fall within the bounding box for any reason. Gaps are also set for PRF-based centroids that cannot be successfully computed for any target and cadence due to failure of the iterative fitting algorithm.

Let the subset of pixels in the aperture of a given target that are utilized for computation of the flux-weighted centroids on each cadence be denoted by ***W***. If the row and column centroid coordinates are denoted by *r* and *c* respectively, then the flux-weighted centroid is computed for any given cadence by

$$r = \frac{\sum_{n \in W} i_{T,n} p_{S,n}}{\sum_{n \in W} p_{S,n}} \quad \text{and} \quad c = \frac{\sum_{n \in W} j_{T,n} p_{S,n}}{\sum_{n \in W} p_{S,n}}. \tag{6-10}$$

As in equation (6-4), the CCD row and column coordinates of the pixels in the target aperture are designated i_T and j_T respectively.

Flux-weighted centroid row and column time series for a long cadence PPA_STELLAR target on module output 7.3 in Q3 are shown in Figure 6.6. The centroid row time series is shown in the upper panel and the centroid column time series in the lower panel. The dominant factor driving long term centroid drift is DVA. Transients in the respective centroid time series at the beginning of the quarter and following the monthly downlinks to Earth (near cadences 1500 and 3000) are due to focus changes which result from thermal changes in the photometer. It is interesting to note that the transients are more pronounced in the centroids than they are in the light curve for this target shown in the upper panel of 6.5.

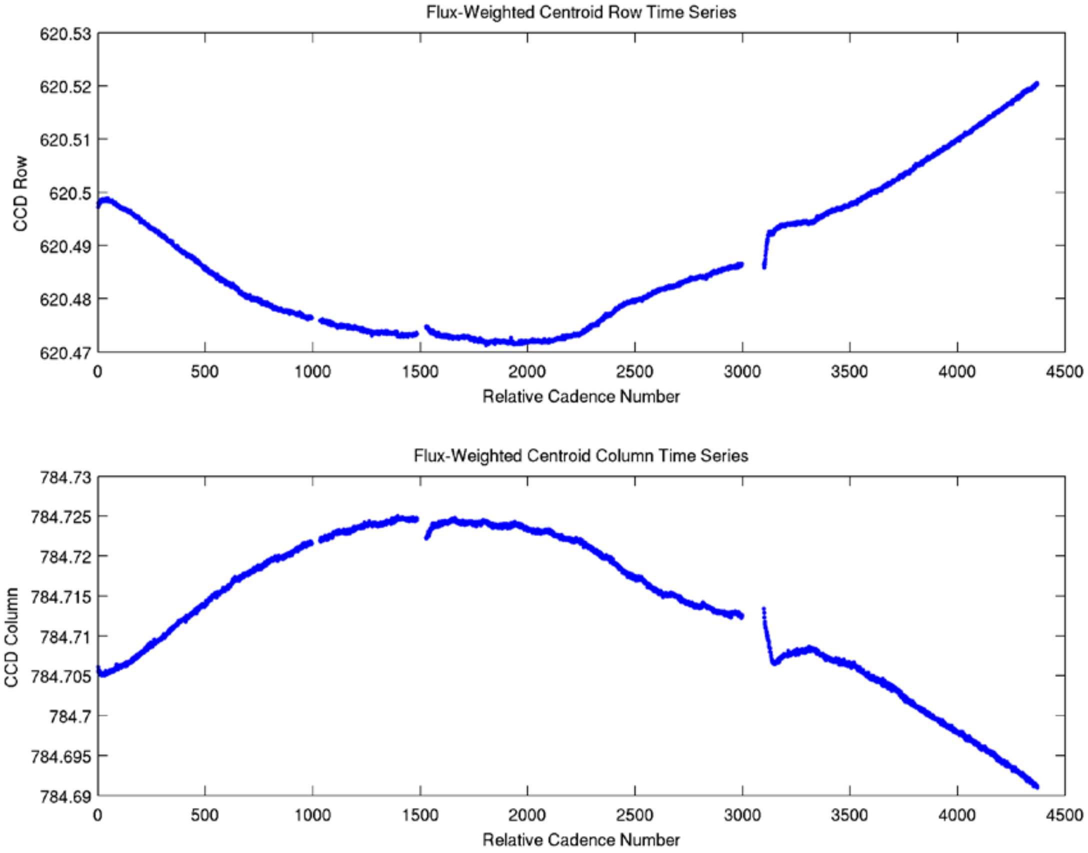


Figure 6.6. Flux-weighted centroid row and column time series for the PPA_STELLAR target also shown in the left panel of Figure 6.4 and the upper panel of Figure 6.5. The data gaps are due to a loss of fine point anomaly and monthly data downlinks. Transients are due to focus changes that result from thermal changes in the photometer.

If the rigorous propagation of uncertainties treatment is enabled by the Pipeline operator, the variances v_R and v_C for the respective row and column centroids on the given cadence are computed by

$$v_R = t_{FWR} C_S t'_{FWR} \quad \text{and} \quad v_C = t_{FWC} C_S t'_{FWC} \,. \tag{6-11}$$

The covariance matrix C_S of the background-removed pixels in the target aperture was defined in equation (6-6). The Jacobian transformations t_{FWR} and t_{FWC} are row vectors where the n^{th} elements of the respective transformations are given by

$$t_{FWR,n} = \frac{i_{T,n} - r}{\sum\limits_{v \in W} p_{S,v}} \text{ and } t_{FWC,n} = \frac{j_{T,n} - c}{\sum\limits_{v \in W} p_{S,v}} ,$$

if the n^{th} pixel is in the flux-weighted centroid aperture ***W***, or zero otherwise.

6.4 Summary

The primary tasks of the Photometric Analysis module of the *Kepler* SOC Science Processing Pipeline are to compute the photometric flux and photocenters (centroids) for all long and short cadence targets from the calibrated pixels in their respective apertures. In this section, we have shown how cosmic rays are cleaned from the calibrated background and target pixel data, then discussed the process for estimation and removal of background flux from the pixels in the apertures of the respective science targets. We then described the extraction of raw flux light curves via simple aperture photometry from the background-subtracted pixels in the optimal apertures of the science targets. Finally, we discussed how flux-weighted centroids are computed per target and cadence and also described how uncertainties are propagated to the various Photometric Analysis data products from covariance matrices for the calibrated background pixels and the calibrated pixels in the respective target apertures. Raw flux light curves, centroid time series, and barycentric timestamp corrections produced by PA are exported to the archive and made available to the general public in accordance with the mission data release policies.

7 PRE-SEARCH DATA CONDITIONING

7.1 Introduction

In this section, we describe the correction of light curves generated by the PA module for systematic and other errors found in the data. The first task of Pre-Search Data Conditioning (PDC) is to identify and correct flux discontinuities that cannot be attributed to known spacecraft or data anomalies. Systematic error correction is then performed by removing flux signatures in the respective light curves that are correlated with ancillary engineering or Pipeline data. Systematics in the flight data are attributable to a variety of sources, and are present on multiple time scales over a wide dynamic range. The second task of PDC is to remove excess flux from the light curves due to background sources within the respective target apertures (crowding). The final task of PDC is to condition the light curves for the transiting planet search. This involves identification and removal of flux outliers and filling of data gaps.

Corrected flux light curves are exported to MAST in the light curve files. Flux outliers identified and removed in PDC for the purpose of conditioning light curves for the transiting planet search in the SOC Pipeline are restored in the corrected flux light curves exported to the MAST. Filled data gaps are removed from the exported light curves.

An overview of PDC and the flow of data through the Pipeline module are presented in § 7.2. PDC science algorithms are described in § 7.3; identification and correction of random flux discontinuities is discussed in § 7.3.1, systematic error correction in § 7.3.2, and removal of excess flux due to aperture crowding in § 7.3.3. A summary and conclusions are presented in § 7.4.

7.2 Overview and Data Flow

The primary tasks of PDC are to correct systematic and other errors in the raw flux light curves produced in the PA module, to remove excess flux in light curves due to crowding in the respective target apertures, and to condition light curves for the transiting planet search performed in the TPS module (section 8) by identifying and removing flux outliers and filling data gaps.

Systematic errors are present in the flight science data on a range of time scales and may be traced to a variety of sources (Jenkins et al. 2010b). Targets also exhibit native variability over a range of time scales and with a wide variety of astrophysical signatures (Jenkins et al. 2010b). The goal of PDC is to remove systematic errors from raw flux light curves while leaving the native target variation and astrophysical signatures intact. It is difficult, if not impossible, to do this for all targets. There is an attempt in the current code release (SOC 6.2) to identify those targets for which systematic error correction performs badly; in these cases, the raw flux is passed through to the back end of PDC uncorrected.

The standard PDC unit of work (Klaus et al. 2010a, 2010b) for LC and SC science data processing is a single module output for a duration of one quarter or one month. Raw flux light curves for all targets on a module output are provided as input through the module interface and corrected flux light curves are passed back through the module interface both with and without fitted harmonic content. Indices of flux outliers are produced as an output of PDC along with the associated outlier values and uncertainties. Indices of filled data gaps are also returned by PDC. Unless a raw flux light curve is pathological, all data gaps should be filled by PDC. Corrected flux light curves (with harmonic content) for LC and SC targets are exported periodically to the

MAST. Outlier values and uncertainties are restored in the export files and the filled data gaps are removed.

Data flow in the PDC Pipeline module is shown in Figure 7.1. Data generally flow from left to right and top to bottom in the figure. Functional blocks in the diagram are referenced with bold type where they are discussed in the text.

Synchronize Ancillary Data. Ancillary engineering data, ancillary Pipeline data (i.e., ancillary data produced in another Pipeline module), and temporal motion polynomial sequences produced for the given module output in PA are synchronized to mid-cadence timestamps in the unit of work to support systematic error correction. Signatures that are correlated with synchronized ancillary data are removed from raw flux light curves to correct the systematic errors. It is not necessary that ancillary engineering data, ancillary Pipeline data, and motion polynomials all be present in the PDC input structure. Systematic error correction is performed given whatever ancillary data are available. The synchronization process involves binning to cadence, sub-cadence, or super-cadence intervals followed by digital resampling (decimation or interpolation) where necessary. Gaps may be filled in the synchronization process.

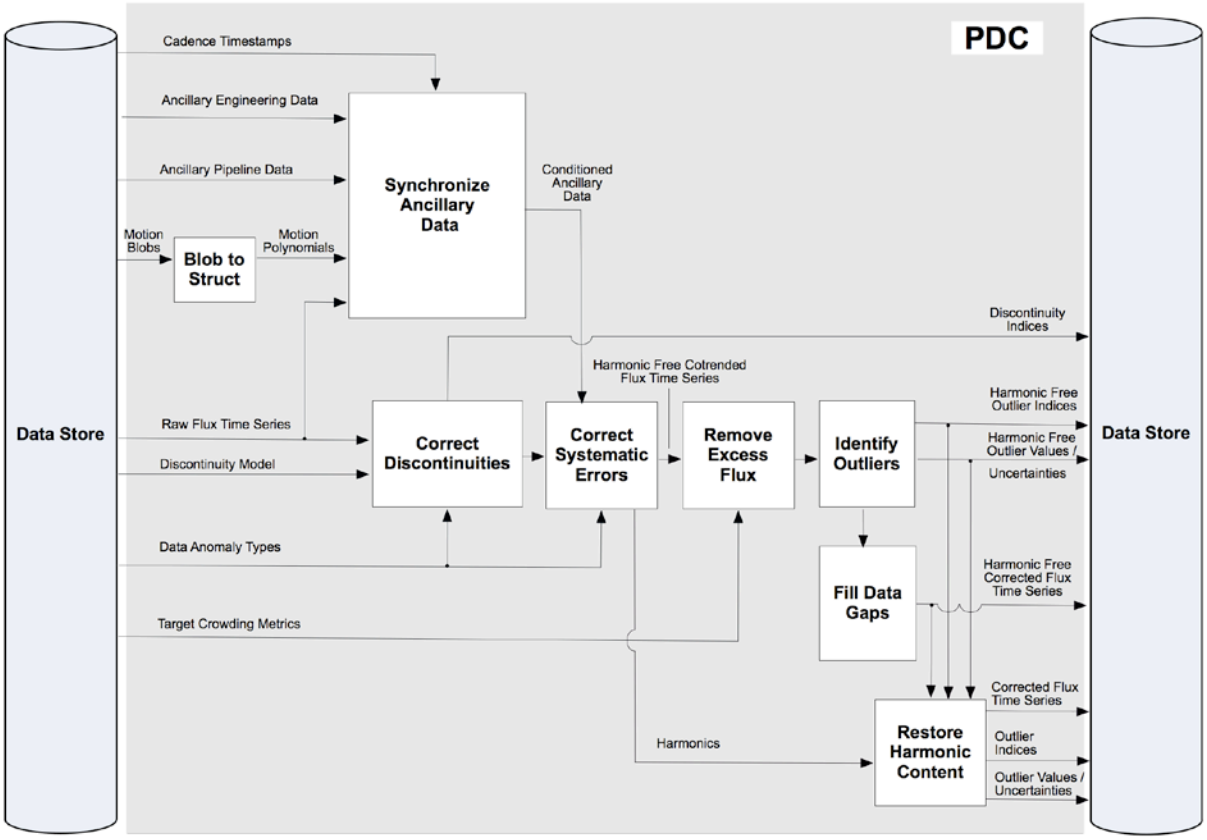


Figure 7.1. Data flow diagram for the Presearch Data Conditioning Pipeline module. Inputs are shown at the left and outputs are shown at the right. Inputs are obtained from the Data Store and outputs are written to the Data Store

Correct Discontinuities (7.3.1). Random flux discontinuities have been observed since the first flight data were acquired (Jenkins et al. 2010b). The random discontinuities are differentiated from discontinuities introduced into many target light curves by spacecraft activities and anomalies (monthly downlinks, safe modes) and commanded attitude adjustments. It is likely that the random flux discontinuities are caused by impacts of cosmic rays or other energetic particles on CCD pixels. Prior to correction of systematic errors in PDC, an algorithm is employed to identify and correct random discontinuities in the raw flux light curves. A discontinuity template is correlated against the numerical second derivative of raw flux for each target and a threshold is applied to identify significant events. In addition to cadence indices of detected discontinuities, step sizes are also estimated. Discontinuities are subsequently corrected for each target by adjusting the flux values following each identified discontinuity by the associated step size. The process of discontinuity identification and correction is iterated for each target to allow for correction of multiple cadence discontinuities.

Correct Systematic Errors (7.3.2). Systematic errors are corrected by a process referred to as *cotrending*. A design matrix is created by separately filtering each of the synchronized ancillary data time series into selectable lowpass, midpass, and highpass components. Each raw flux time series is then projected into the column space of the design matrix in a least squares sense, and the residual (with mean level restored) between the raw flux and least squares fit determines the systematic error corrected flux for each target. This process essentially removes flux signatures that are correlated with the ancillary data on the specified time scales. A Singular Value Decomposition (SVD) is utilized to perform the projection in a computationally efficient and numerically stable manner. Uncertainties in corrected flux values are propagated from uncertainties in the raw flux values in accordance with standard methods. Memory limitations prevent the creation of full covariance matrices for each corrected flux time series, however.

An attempt is made to identify variable targets and to fit the light curve for each such target with a superposition of phase shifting harmonics. Reliable identification of variable targets is difficult, however, in the presence of large data anomalies. Harmonic content is subsequently removed from light curves of the variable targets for which harmonic fitting was successful. All of the harmonic free light curves are again subjected to the cotrending process. A decision is made whether to use the standard or harmonic free cotrending result for each of the targets initially identified as variable. Systematic error correction performance is finally evaluated for all targets in the unit of work and error-corrected flux is replaced with raw flux for each of the targets determined to have been badly corrected.

Remove Excess Flux (7.3.3). Following the correction of systematic errors, excess flux due to crowding in the optimal aperture is removed from the light curve for each target. The amount of excess flux is determined from the crowding metric that is provided to PDC for each target. The crowding metric is defined as the fraction of flux in the optimal aperture due to the target itself. The metric is computed in the Target and Aperture Definitions component (Section 4) of the Pipeline. A single value is provided for each target and target table even though crowding may vary with long-term motion of targets and background sources primarily due to differential velocity aberration.

Identify Outliers. Outliers are identified in each flux time series based on robust estimates of the mean and standard deviation in a sliding scan window. The window size and outlier detection threshold are PDC module parameters and are separately tuned for long and short cadence units of work. Flux values marked as outliers are gapped and later filled along with other data gaps. The purpose of outlier identification and removal is to prevent the triggering of false threshold crossing events in TPS. Outlier values and uncertainties are restored in the corrected flux light curves exported to the MAST.

Fill Data Gaps. An attempt is made to fill all data gaps in PDC. The transiting planet search requires that samples be available for all cadences. Gap filling for each target proceeds in two steps: first, “short” data gaps are filled, and then any remaining “long” data gaps are filled. Short and long refer not to the type of cadence data being processed, but to the length of the gaps to be filled. The boundary between short and long data gaps is determined by the gap filling module parameter set.

Short data gaps are sequentially filled with available flux samples to the left and right of the respective gaps. An autoregressive algorithm is employed to estimate sample values in the gaps with a linear prediction based on the flux correlation in the neighborhood of the gap. Uncertainties in short gap-filled samples are produced from uncertainties in the samples used to fill them. Long data gaps are filled in a process that involves folding and tapering blocks of available samples from the left and right of the respective gaps. Wavelet domain coefficients are then adjusted to ensure statistical continuity across the filled gaps. There is no attempt to estimate uncertainties for long data gap-filled samples. Gap-filled values are not included in the corrected flux light curves exported to the MAST.

Restore Harmonic Content. Harmonic content identified and removed for harmonically variable targets is restored to the corrected flux light curves before PDC runs to completion. Harmonic content is also restored to the outlier values identified earlier. PDC therefore produces two corrected light curves and two sets of outliers for each target, one based on the standard flux time series for the given target, and one based on the harmonic-free flux time series. For targets without fitted harmonic content, the standard and harmonic-free results are identical.

Kepler is first and foremost a transit photometry mission. Every effort is made to preserve transits in PDC and to prevent them from compromising performance of the science algorithms. When discontinuities are identified and corrected, an attempt is made to ensure that large transits (and other astrophysical events such as binary eclipses and flares) do not trigger the detector. In the correction of systematic errors, an attempt is made to prevent large transits and astrophysical events from corrupting the least squares fitting process. These events are restored after fitting is performed.

Large transits and astrophysical events are also masked prior to performing outlier identification. There are two reasons for doing so. First, transits are masked to prevent them from corrupting the estimates of mean and standard deviation that are utilized in setting the robust outlier detection threshold. Second, transits are masked in order to prevent them from being identified as outliers. An attempt is also made in the gap-filling process to prevent transits and other astrophysical events in available science data samples from being used to fill both short and long data gaps.

7.3 Science Algorithms

7.3.1 Discontinuity Identification and Correction

Random flux discontinuities have been observed for some targets since the first flight data were acquired (Jenkins et al. 2010b). The random discontinuities are differentiated from discontinuities introduced into many of the target light curves as a result of spacecraft activities and anomalies (monthly downlinks, safe modes) and commanded attitude adjustments. Random discontinuities are most often attributed to abrupt decreases in sensitivity, perhaps due to impacts of cosmic rays or other energetic particles on CCD pixels. They are sometimes followed by a partial exponential rebound. There is no attempt in the current PDC release (6.2) to model and correct exponential rebounds.

A flow chart describing the discontinuity identification algorithm is shown in Figure 7.2. The process is performed independently on all raw flux light curves in a given unit of work. An attempt is first made to replace giant transits and other astrophysical events. Savitzky-Golay filtering is performed on the raw flux time series to compute the numerical derivatives for orders zero, one, and two. A sliding discontinuity template (which is provided as a parameter through the module interface) is then correlated against the filtered second derivative of the time series. Statistics are computed for the correlation time series and a threshold is applied to identify discontinuity candidates.

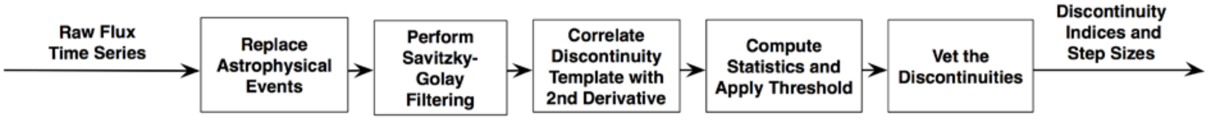

Raw Flux Time Series
Replace Astrophysical Events
Perform Savitzky-Golay Filtering
Correlate Discontinuity Template with 2nd Derivative
Compute Statistics and Apply Threshold
Vet the Discontinuities
Discontinuity Indices and Step Sizes

Figure 7.2. Flow chart for the discontinuity identification algorithm. The process is performed independently for each target in the unit of work.

Candidates are vetted before results are returned. Discontinuities that coincide with known spacecraft and data anomalies are excluded, as these are addressed as part of the systematic error correction process. Discontinuities that may have been identified as artifacts of interpolated data gaps or masked astrophysical events are also discarded. Discontinuities that fail a gradient test and are apparently due to single cadence outliers are discarded as well. Cadence indices and discontinuity step sizes are returned for each target for all discontinuity candidates that survive the vetting process.

Correction of discontinuities is straightforward. For any given target, the portion of the flux time series following each identified discontinuity is adjusted by the estimated discontinuity step size until all discontinuities have been corrected. The process of discontinuity identification and correction is repeated until no additional discontinuities are found or an iteration limit is reached. The iterative process allows multiple cadence discontinuities to be identified and corrected. If discontinuities are still identified for a given target after the iteration limit has been reached, the process is deemed to have failed for that target and the initial flux values are restored without correction of any discontinuities.

A discontinuity detection example for a long cadence Q3 target on module output 7.3 is shown in Figure 7.3. The raw flux time series is plotted versus cadence in the upper panel. There are two flux discontinuities present that are not the result of known spacecraft activities or anomalies. The first of these occurs near cadence number 1450 and the second occurs around cadence number 2700. The associated discontinuity detection statistics are shown in the lower figure. Both events clearly exceed the specified 5σ detection threshold. Detection statistics for cadence indices returned by the detector are circled in the lower figure. Multiple random discontinuities are not common for targets in a single unit of work. The discontinuity following the second downlink of the quarter (near cadence 3000) and the associated thermal transient are addressed later in PDC when systematic error correction is performed.

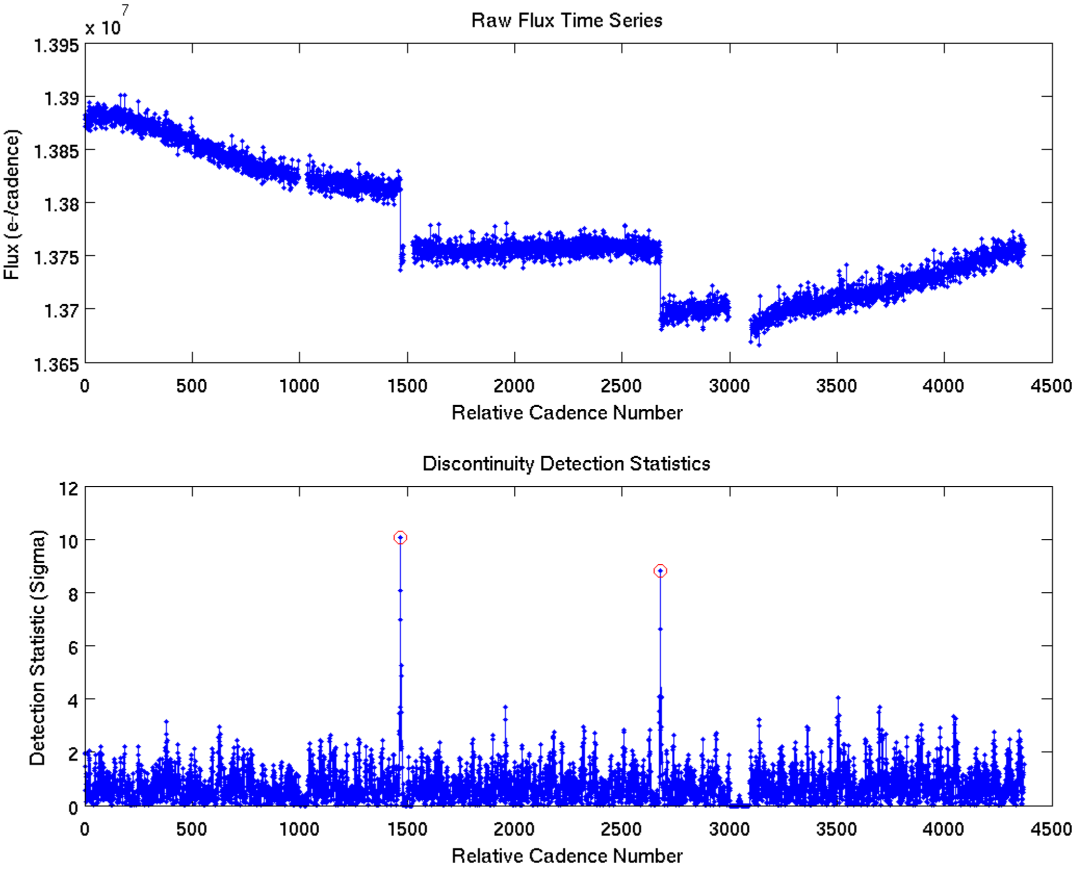


Figure 7.3. A raw flux time series with two random discontinuities is shown in the upper panel. The detection statistics formulated by correlating the second derivative of the time series with the discontinuity template are plotted in the lower panel. The detection statistics are circled for the cadence indices returned by the detector.

7.3.2 Systematic Error Correction

Systematic errors are introduced into long and short cadence light curves by a variety of sources over a broad spectrum of dynamic ranges and time scales. In the first year of science data collection it has become apparent that the systematic errors are caused primarily by target motion at the pixel or sub-pixel level. Target motion in turn produces changes in target flux levels. The motion polynomials produced in PA by fitting the centroids of selected targets for each cadence as a function of the celestial target coordinates are therefore well suited for removing systematic effects on a module output basis.

The dominant long-term systematic effect is DVA, which causes targets to trace small ellipses on the respective CCD detectors over the period of the heliocentric orbit of the photometer (Jenkins et al. 2010b). The maximum motion due to DVA is 0.6 pixels per observing quarter. Other significant systematic errors in flight science data have resulted from variable (eclipsing binary) Fine Guidance Sensor reference targets, short-period (~3 hours) reaction wheel heater cycling, long duration (~4-5 days) thermal transients following safe modes and monthly downlinks, and commanded photometer attitude adjustments. Early in the mission it was necessary to perform multiple attitude adjustments (see DRN7 for Q2 details) in a single quarter (Q2) to accommodate drift in photometer pointing (Haas et al. 2010).

The ability to correct systematic errors in PDC directly impacts performance of the transiting planet search in TPS and hence, the detectability of the very planets that the *Kepler Mission* was designed to discover. Systematic errors must be corrected so that they do not trigger massive numbers of TCEs in the transiting planet search. The scale of the systematic errors in the light curves may be multiple orders of magnitude larger than the transit signatures of Earth-size planets.

Systematic error correction is performed in PDC by identifying signatures in the raw flux light curves that are correlated with ancillary engineering and Pipeline data and temporal motion polynomial sequences. Ancillary data are first synchronized to mid-cadence timestamps of the science data. A least squares fitting algorithm is employed, utilizing the SVD of the ancillary design matrix. The projection of raw flux for each target into the column space of the design matrix is based on the rank of the design matrix and is performed in a computationally efficient and numerically stable manner. Uncertainties are propagated for each target given the linear transformation of raw to cotrended flux, although the full covariance matrix for the cotrended flux cannot be computed due to memory constraints.

A flow chart describing the systematic error correction algorithm is shown in Figure 7.4. The synchronized time series (to the cadence timestamps) for the respective ancillary engineering and Pipeline data mnemonics and temporal motion polynomial coefficient sequences are packed into the columns of a design matrix. The length of each time series (and hence the number of rows in the design matrix) is equal to the number of cadences in the unit of work. The mean value is subtracted from each synchronized time series, and each is then divided by its maximum absolute value for the purpose of numerical conditioning. A constant column (containing all ones) is included in the matrix. Gapped values are interpolated so that the columns may be subsequently filtered into bandpass components. There cannot be any gaps in the synchronized data, however, on cadences with valid science data.

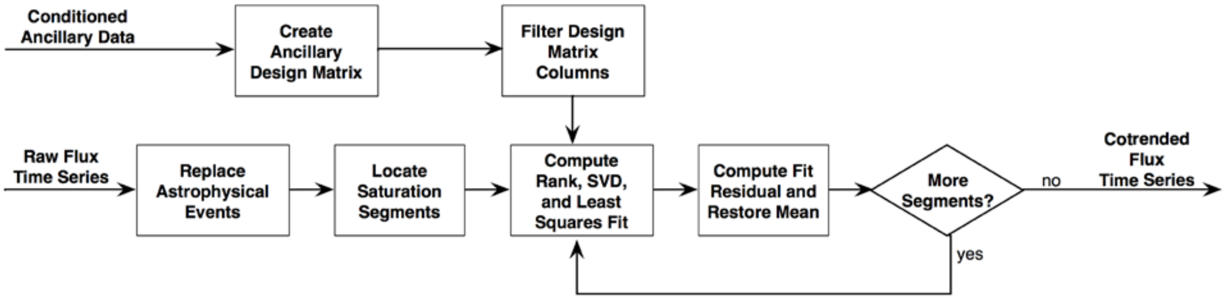

Conditioned Ancillary Data
Create Ancillary Design Matrix
Filter Design Matrix Columns
Raw Flux Time Series
Replace Astrophysical Events
Locate Saturation Segments
Compute Rank, SVD, and Least Squares Fit
Compute Fit Residual and Restore Mean
More Segments?
no
yes
Cotrended Flux Time Series

Figure 7.4. Flow chart for the systematic error correction algorithm. The design matrix is generated and filtered once for all targets in the unit of work. The Singular Value Decomposition (SVD) and least squares projection are also performed once for all targets (as long as they have matching data gaps). Targets with saturation segments, however, must be corrected segment by segment. Saturated segments are demarcated by abrupt changes in curvature as pixels in the associated optimal aperture enter or exit saturation.

The columns of the design matrix (with the exception of the constant column) are then filtered into selectable bandpass (lowpass, midpass, and highpass) components. This action permits correction of systematic errors by separately identifying signatures in the target light curves that are correlated with the ancillary data on multiple time scales. The bandpass components are obtained in a cascade of Savitzky-Golay filters. Flux for each target is first filtered into lowpass and highpass components, and the initial highpass component is optionally filtered again into midpass and highpass components. The filter orders and durations (which determine the frequency cutoffs) are tuned separately for long and short cadence science data processing. Generation of the design matrix and filtering of the columns is performed once per PDC unit of work.

Astrophysical events (such as large transits, eclipses, flares, and microlensing events) are identified in the raw flux light curves prior to performing the cotrend fit, and are replaced temporarily with values interpolated across the cadences of the respective events. Random noise is added to the interpolated values based on the statistics of the light curves excluding the astrophysical events. The purpose of this is to prevent astrophysical events from perturbing the fit of the synchronized ancillary data to the raw flux light curves. The intent of PDC is to fit (and remove) systematic error signatures in the science data; it is not the intent to fit the astrophysical events.

Large systematic effects in the light curves due to thermal transients (following safe modes and monthly downlinks) and photometer attitude adjustments may be inadvertently misidentified as astrophysical events. If they are subsequently masked from the least squares fitting, then they cannot be corrected. To resolve this problem, the astrophysical events are vetted against the known spacecraft and data anomalies that are provided as input to all Pipeline modules. If an identified event occurs on or near a known anomaly (monthly downlink, safe mode, or commanded attitude adjustment), the event is not replaced prior to cotrending. This unfortunately represents an engineering trade-off. True astrophysical events that occur on or near major spacecraft and data anomalies are compromised so that those same anomalies may be corrected in the flux for most targets.

After astrophysical events have been identified and replaced temporarily for all targets, any saturated time series segments are located for each target that is sufficiently bright to saturate the CCD detectors. Saturated segments are demarcated by changes in the curvature of the respective flux time series (with astrophysical events removed). A Savitzky-Golay filter is utilized to compute the numerical second derivatives and a threshold is applied for the bright targets based on the statistics of the second derivative time series. If breakpoints are identified in the flux for saturated targets, those targets are separately cotrended segment by segment after processing has completed for all of the targets without saturation breakpoints.

Cotrending is performed with a linear least squares fit of the filtered ancillary design matrix columns to the raw flux time series for each target. Gaps are first squeezed from the raw flux light curves and the associated rows of the design matrix. The data gaps must match for targets that are cotrended at once. If that is not true for all targets in the unit of work, then cotrending must be performed separately on subsets of the targets that do have matching data gaps. Without any loss of generality, given the design matrix A and a raw flux time series f_{RAW}, we seek to find the least squares solution to the set of linear equations:

$$Ax = f_{RAW} . \tag{7-1}$$

Let the dimension of A be $m \times n$ and the reduced SVD of A be denoted by

$$A = USV' , \tag{7-2}$$

where *U* has dimension *m* × *n*, *S* is a diagonal matrix of singular values with dimension *n* × *n*, and *V* also has dimension *n* × *n* if *m* > *n*, as is generally the case in PDC. It may then be shown with matrix algebraic manipulation that the least squares solution to (1) is given by

$$x = VS^{-1}U' f_{RAW} . \tag{7-3}$$

In PDC, we are more interested in the fit to the raw flux, however, than we are in the actual fit coefficients *x*. Given equations (7-2) and (7-3), we may compute the least squares fit of the filtered ancillary data to the raw flux light curve by

$$f_{FIT} = Ax = USV'\left(VS^{-1}U' f_{RAW}\right) = UU' f_{RAW} . \tag{7-4}$$

The projection of the raw flux into the column space of the design matrix depends only on the unitary matrix *U*. If the design matrix A is not full rank (i.e., the columns of the design matrix are not independent), then we seek to limit the dimension of the least squares fit. If the rank of the design matrix *A* is denoted by *r*, and U_r denotes the first *r* columns of *U*, then in PDC the least squares fit is performed as follows:

$$f_{FIT} = U_r U'_r f_{RAW} . \tag{7-5}$$

The residual between the raw flux (with astrophysical events restored) and the fitted flux determines the cotrended flux f_{COT}. The mean raw flux level f_{RAW} is also included as follows:

$$f_{COT} = f_{RAW} - f_{FIT} + \mu_{RAW} = \left(I - U_r U_r{}'\right) f_{RAW} + \mu_{RAW} . \tag{7-6}$$

Propagation of uncertainties from raw to cotrended flux is straightforward in principle. Memory constraints prevent computation of the full covariance matrix for the cotrended flux, which has dimension *m* × *m* where *m* is the number of cadences in the unit of work. If C_{RAW} and C_{COT} denote the covariance matrices for temporal samples of the raw and cotrended flux time series for a given target, the uncertainties may be propagated (disregarding the uncertainty in the mean level which may be considered to be negligible) by:

$$C_{COT} = T_{COT} C_{RAW} T'_{COT} , \tag{7-7}$$

where the transformation T_{COT} is defined by:

$$T_{COT} = I - U_r U'_r . \tag{7-8}$$

Due to the aforementioned memory constraint, only the diagonal elements of the covariance matrix C_{COT} are computed in PDC from the diagonal covariance matrix C_{RAW}. The uncertainties in the cotrended flux are given by the square root of the respective diagonal elements of the covariance matrix C_{COT}. Diagonal elements of C_{RAW} are squares of the uncertainties in the raw flux time series produced by PA.

It has been stated earlier that the intent of cotrending is to fit and remove systematic signatures in the data and not to fit the astrophysical events (such as transits, eclipses, and flares). To that end, an attempt is made to mask such events from the least squares fitting process. The situation becomes complicated, however, when native variability of the targets is considered. The least squares combination of filtered ancillary data may corrupt variability that is inherent in the stellar targets, and in some cases it has been observed to remove the variability completely.

After all targets have been cotrended as described above, an attempt is made to identify variable targets in the unit of work. Those targets for which the center-to-peak flux variation is observed to exceed a specified threshold are flagged. The typical variability threshold in the Pipeline is set to 0.5% of the median target flux level. Coarse detrending (accounting for known spacecraft and data anomalies, thermal transient characteristics, and DVA) is performed on the raw flux light curves for the variable targets and an attempt is made to fit the detrended flux for each with a superposition of phase shifting harmonics (Jenkins et al. 2010a; 2010c). The phase shifting harmonics differ from a conventional Fourier representation in that the frequencies are permitted to vary linearly with time. Fitted harmonic content is removed from the raw flux for each of the variable targets and saved for restoration later in PDC. The residual flux light curves with harmonic content removed are then subjected to the cotrending process as before to remove systematic effects. The harmonic content is identically zero in cases where phase-shifting harmonics cannot be fitted to the variable light curves.

Reliable identification of variable targets is difficult in the presence of large data anomalies. Once the apparently variable targets have been corrected, a decision is made for each regarding whether to utilize the standard or harmonic-free cotrending result going forward. Performance is assessed in each case based on a robust estimate of the ratio of the power at short time scales (defined by module parameter) in the cotrended result to the power at the same time scales in the raw flux. If a target does not appear to be variable (excluding large transits or other astrophysical events) after cotrending without removal of harmonic content and if the standard systematic error correction performance appears to be good, then the standard result is retained. Otherwise, the cotrending result is retained for the residual flux after removal of harmonics. The harmonic content is restored later in PDC.

The final step in the systematic error correction process is to identify targets for which cotrending has not performed to an acceptable degree. Raw flux (after correction of random flux discontinuities) is substituted for cotrended flux for such targets based on a comparison of the performance metric discussed above with a specified performance limit. For these targets, systematic effects are not addressed in PDC because it is not possible to do so without corrupting the inherent character of the light curves. Targets in this category are typically variable stars for which the light curves are not harmonic, or for which the phase-shifting harmonics cannot adequately represent the stellar variation.

Figure 7.5 illustrates systematic error correction performance for a quiet target on module output 7.3 in Q2. This quarter featured a series of significant spacecraft and data anomalies (DRN7). The raw flux light curve is shown in the upper panel. The least squares fit of the filtered ancillary data to the raw flux is shown in the middle panel. The difference between raw and fitted flux is virtually indistinguishable on the scale shown in the figure. All of the information regarding anomaly-induced flux discontinuities and recovery transients is captured by the ancillary data and motion polynomial sequences utilized for correction of the systematic errors. The fit residual is shown in the lower panel with the mean flux level restored. The scale of the raw flux artifacts for this target due to the various spacecraft and data anomalies is multiple orders of magnitude larger than the transit depth of an Earth-size planet orbiting a Sun-like star.

Systematic error correction performance for a Q1 target located on module output 2.1 is shown in Figure 7.6. This module output has been observed to be particularly sensitive to focus changes in the photometer. A 200-cadence segment of the raw flux light curve is shown in the upper panel. The least squares fit of the filtered ancillary data to the raw flux is shown in the middle panel. The oscillations in the raw flux are caused by the cycling of a spacecraft reaction wheel heater outside the photometer. The mechanism by which heat generated outside the photometer causes the focus to change is not yet understood. The cotrending process nevertheless produces a fit that essentially tracks the flux oscillations, and the oscillations are significantly reduced in the residual flux shown in the lower panel.

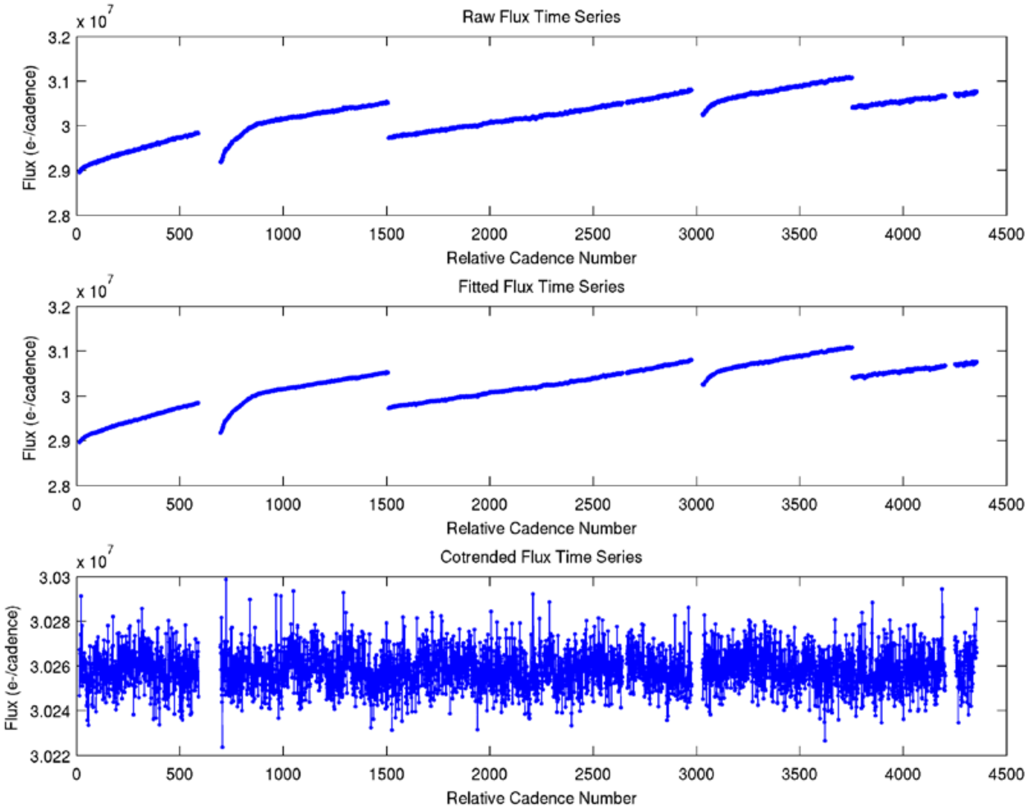


Figure 7.5. The raw flux time series for a quiet long cadence Q2 target is shown in the upper panel. A number of significant spacecraft and data anomalies are clearly visible in the light curve. These include thermal transient at the start of the quarter, safe mode recovery transient near cadence 700, attitude adjustment near cadence 1500, Earth point recovery transient near cadence 3000, and attitude adjustment near cadence 3750. There is also a loss of fine point data gap near the end of the quarter. The least squares fit of the synchronized ancillary data to the raw flux is shown in the middle panel. The fit residual with mean level restored is shown in the lower panel.

It should be noted that the peak-to-peak variation in flux oscillations for this target is on the order of 0.1% (before removal of excess flux due to aperture crowding), which is approximately equivalent to the transit depth of a Neptune-size planet orbiting a Sun-like star, and ten times the transit depth of an Earth-size planet orbiting a Sun-like star. Hence, even small-scale systematic effects in the flight data dwarf the transit signatures that the *Kepler Mission* has been designed to detect.

7.3.3 Excess Flux Removal

Optimal apertures may include flux from sources other than the targets with which they are associated. It is necessary to estimate and remove excess flux in order that the relative transit depths in the corrected flux light curves produced by PDC faithfully represent the true transit depths of the target system. Otherwise, transits will be systematically diluted and planet radii will be systematically underestimated in the Pipeline and by consumers of Pipeline data.

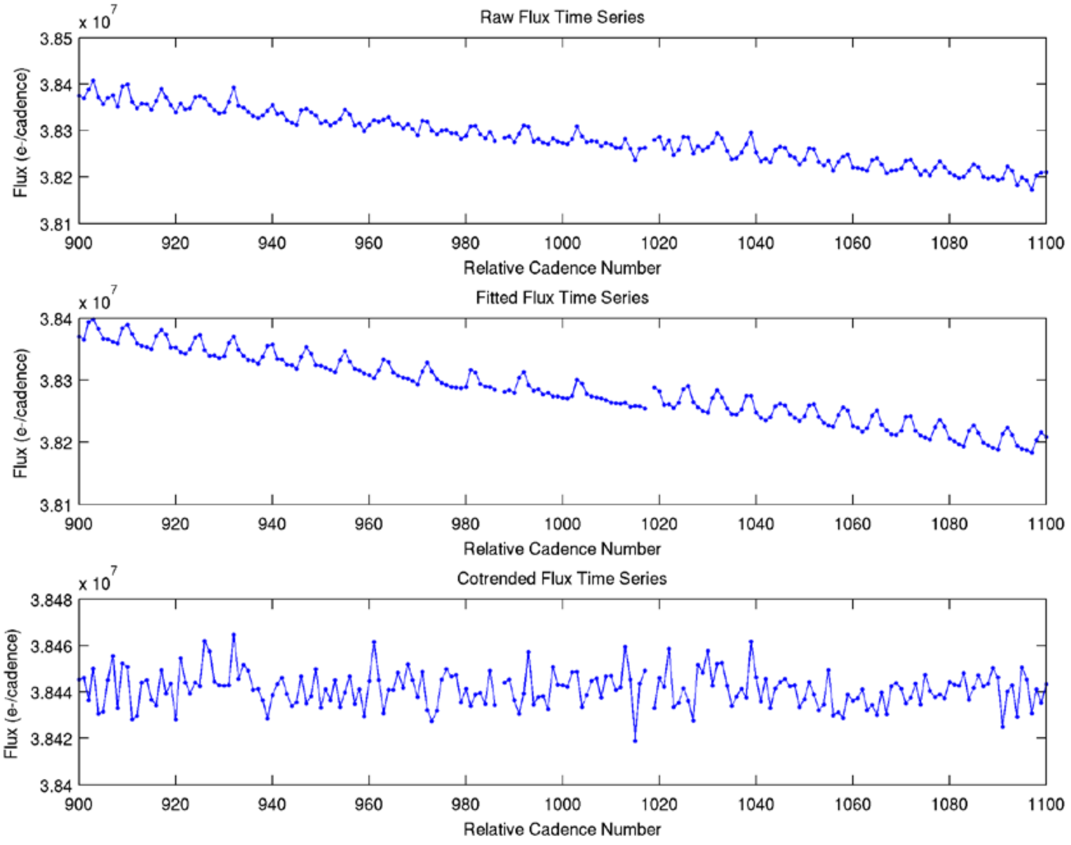


Figure 7.6. A 200-cadence segment of a raw flux time for a quiet long cadence target in Q1 is shown in the upper panel. Oscillation in the light curve is due to cycling of a reaction wheel heater outside the photometer. The least squares fit of the synchronized ancillary data to the raw flux is shown in the middle panel. The corresponding segment of the fit residual with mean level restored is shown in the lower panel.

The so-called *crowding metric* is computed during TAD module (Section 4) for each target and is defined to be the fraction of flux in the optimal aperture due to the target itself. The crowding metric is a scalar value representing the average aperture crowding over the effective date range of a given target table. Crowding may be dynamic, however, changing as background sources of light enter and exit the optimal aperture due to DVA. To that extent, computation of the crowding metric and removal of excess flux in PDC are only approximate and may need to be revisited in the future.

For each target, a constant excess flux level is estimated over the duration of the unit of work based on the crowding metric for the given target and the median value of the cotrended flux for that target. The excess flux level is then subtracted from the cotrended flux value for every cadence with valid science data. Let the crowding metric for a given target be denoted by α and the median value of the cotrended flux time series be denoted by m_{COT}. The constant excess flux level f_{XS} due to crowding in the optimal aperture is then estimated by

$$f_{XS} = (1-\alpha)\, m_{\mathrm{COT}} . \tag{7-9}$$

The crowding-corrected flux time series f_{COR} is finally determined by subtracting the excess flux level from the cotrended flux time series as follows:

$$f_{COR} = f_{\mathrm{COT}} - f_{XS} = f_{\mathrm{COT}} - (1-\alpha)\, m_{\mathrm{COT}} . \tag{7-10}$$

Uncertainties are not propagated for the excess flux correction. Uncertainties for the crowding-corrected flux values are set equal to those of the cotrended flux. The uncertainty in the median flux estimate over all cadences is assumed to be negligible in comparison with the uncertainty in the systematically error-corrected flux value for any given cadence.

7.4 Examples of PDC Output

In this section we provide additional examples of PDC output. PDC gives satisfactory results on most stars which are either intrinsically quiet (Figure 7-7), or have well-defined harmonic light curves above the detection threshold (Figure 7-8). In most of these cases, the standard deviation of the corrected flux is within a factor of 2 of the noise expected from read and shot noise in the calibrated pixels summed to form the uncorrected light curve. PDC also performs well in many cases where the star is variable, but without a dominant harmonic term (Figure 7-9). However, PDC will sometimes not identify a target-specific discontinuity (Figure 7-10), and will sometimes introduce noise into complex light curves (Figure 7-11). Conversely, PDC sometimes identifies eclipses as discontinuities and introduces a discontinuity in an attempt to correct the false discontinuity (Figure 7-12).

PDC can remove astrophysical signatures if they are:

1. Harmonic signals above the threshold for which the fit is good, but PDC incorrectly determines that target was cotrended well when treated as non-variable (Figure 7-13)
2. Harmonic, but have periods > 5 days and fall below PDC's detection threshold for stellar variability. For Q5, the center-peak threshold is 0.5%, which for otherwise quiet stars allows a harmonic with a peak-to-peak amplitude < 1.0% to go undetected.
3. Spikes a few cadences wide such as flares, or other steep gradients in flux (Figure 7-14).
4. More or less linear ramps over the processing interval.
5. Harmonic signals above the threshold, but the harmonic fit does not produce a good fit, and current algorithm fails to recognize that cotrending has performed badly.
6. Non-harmonic signals for which current algorithm fails to recognize that cotrending has performed badly.

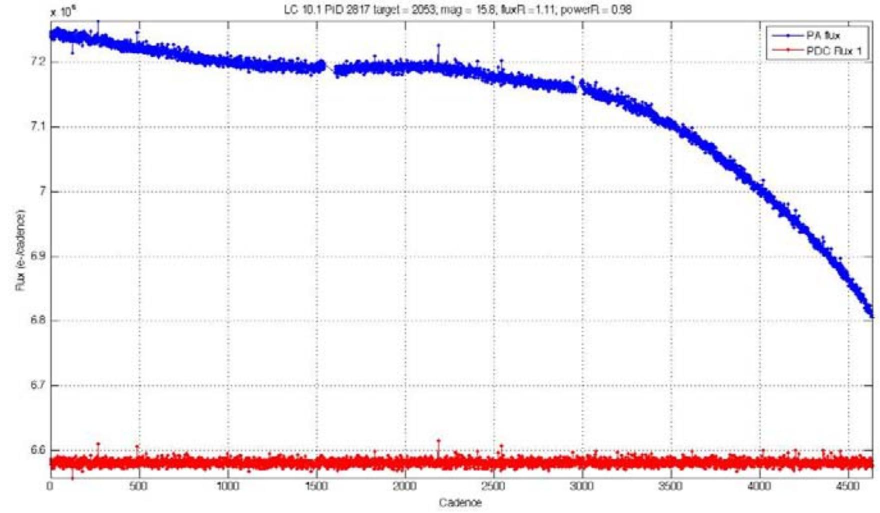


Figure 7-7. Q5 example of PDC removal of trends and discontinuities from the light curve of a quiet star with *Kp* = 15.8. The noise in the corrected light curve is only 11% greater than the noise expected from the calibrated pixels, a considerable improvement over the uncorrected light curve.

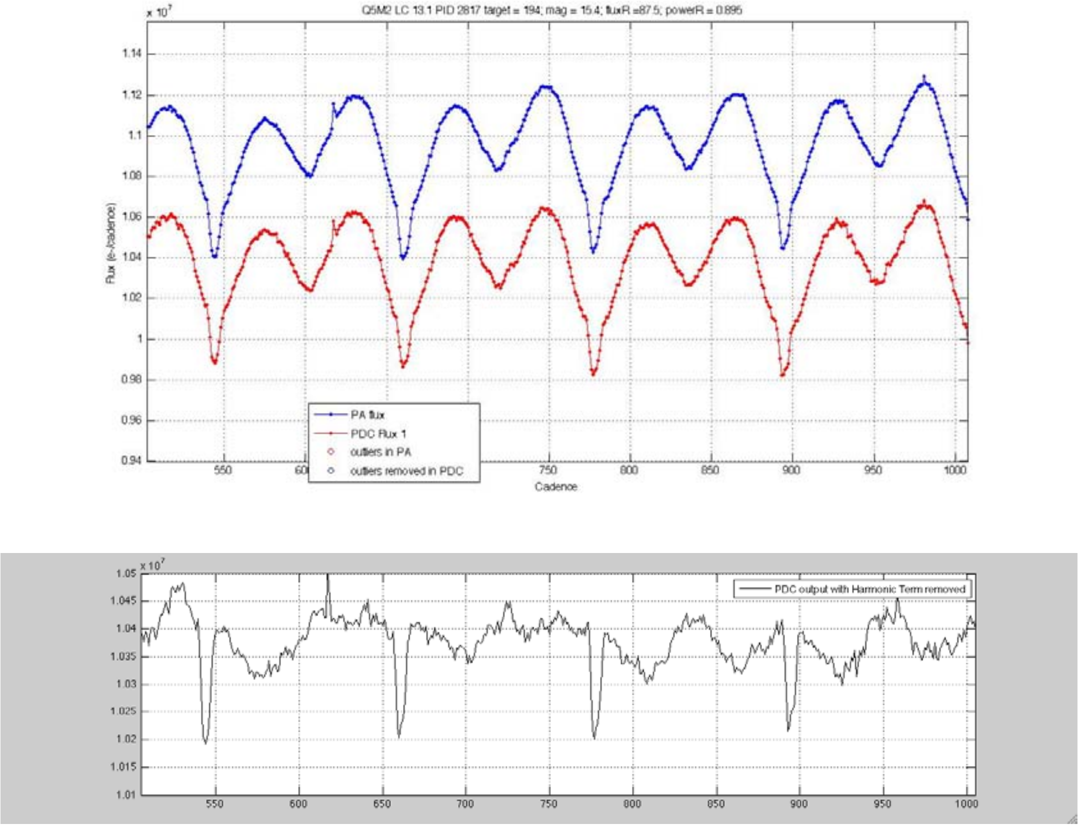


Figure 7-8. Q5 example of PDC correction of a harmonically variable star with eclipses. MAST users receive the light curve corrected for systematic errors, with the harmonic variability restored and gaps in the data represented by –Infs. The corrected light curve delivered to MAST is shown in red in the upper panel of this Figure. The lower panel shows the light curve with harmonics removed illuminating the efficacy of PDC harmonic removal, even with light curves with other features such as eclipses.

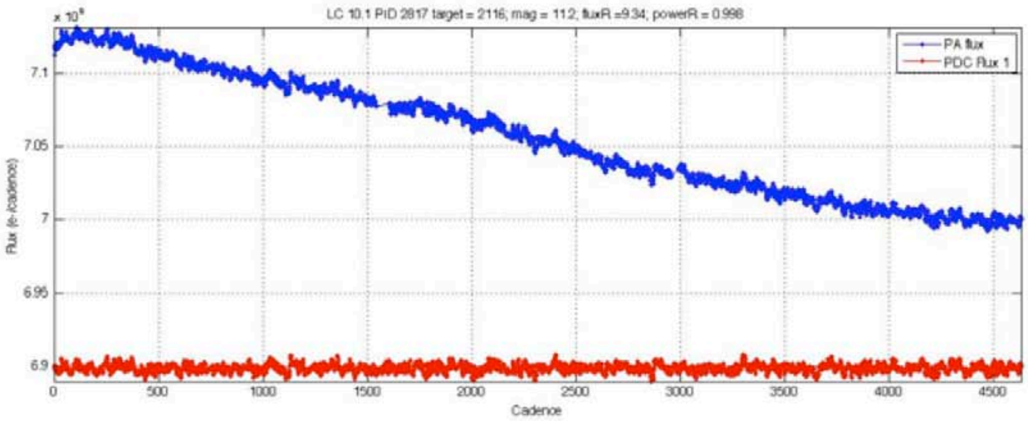


Figure 7-9. Q5 example of PDC correction of a non-harmonically variable star. The RMS of the corrected (red) curve is about 9x the noise calculated from the read and shot noise in the calibrated pixels, and is believed to be almost entirely due to intrinsic stellar variability.

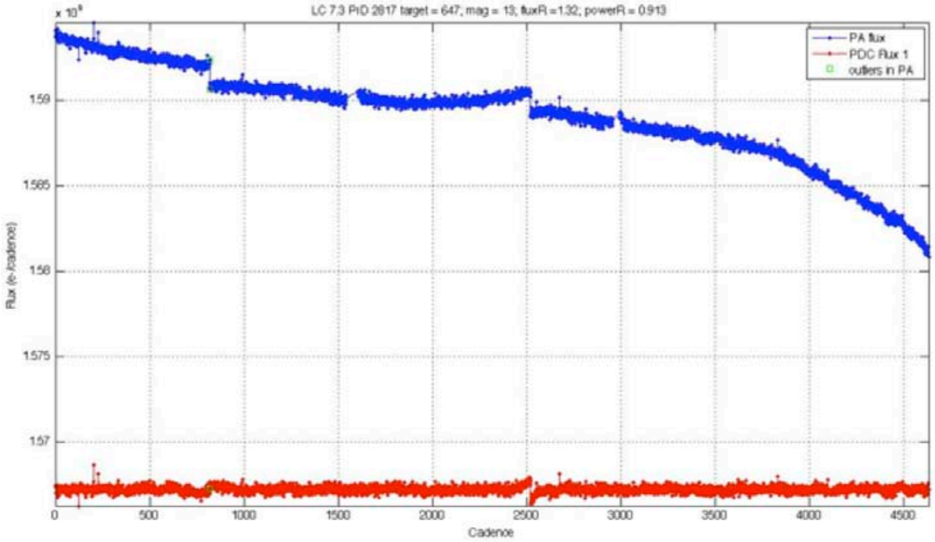


Figure 7-10. Q5 example of an unidentified and hence uncorrected target-specific discontinuity, located at cadences 800 and 2500.

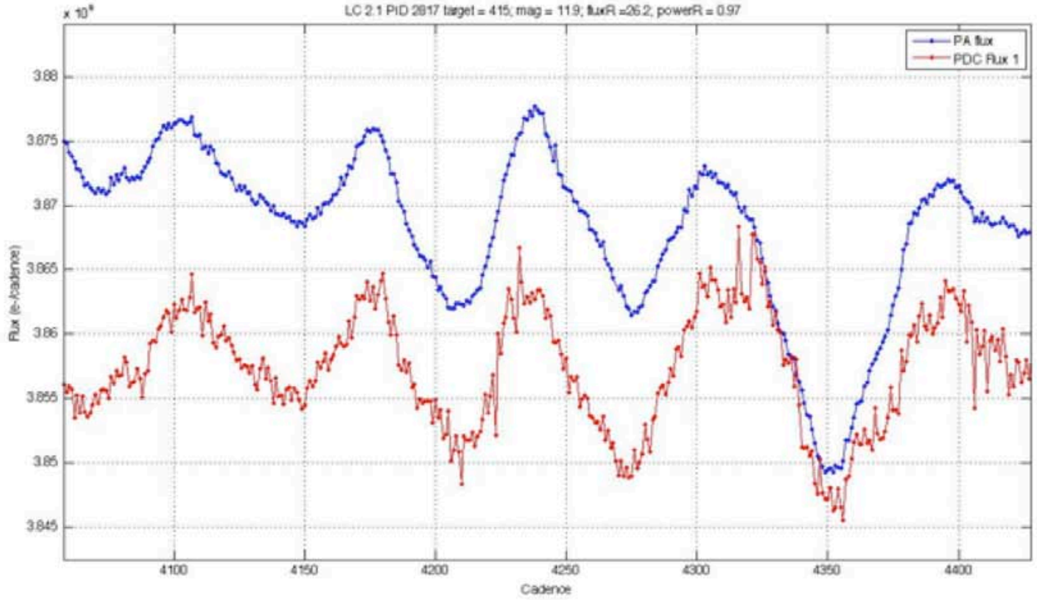


Figure 7-11. Q5 example of PDC adding short-period noise to a light curve, which it did not identify as a bad cotrend.

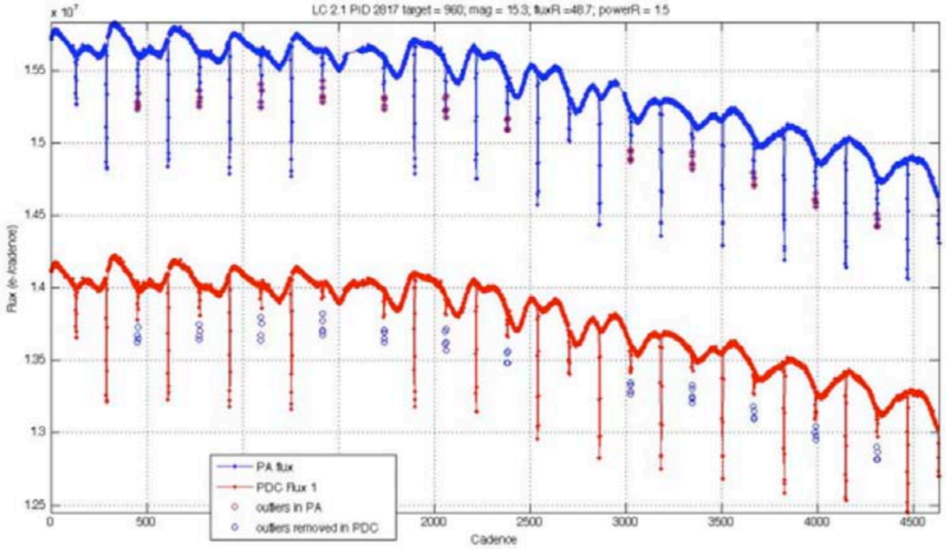


Figure 7-12. Q5 example of PDC mis-identifying an eclipse as an outlier.

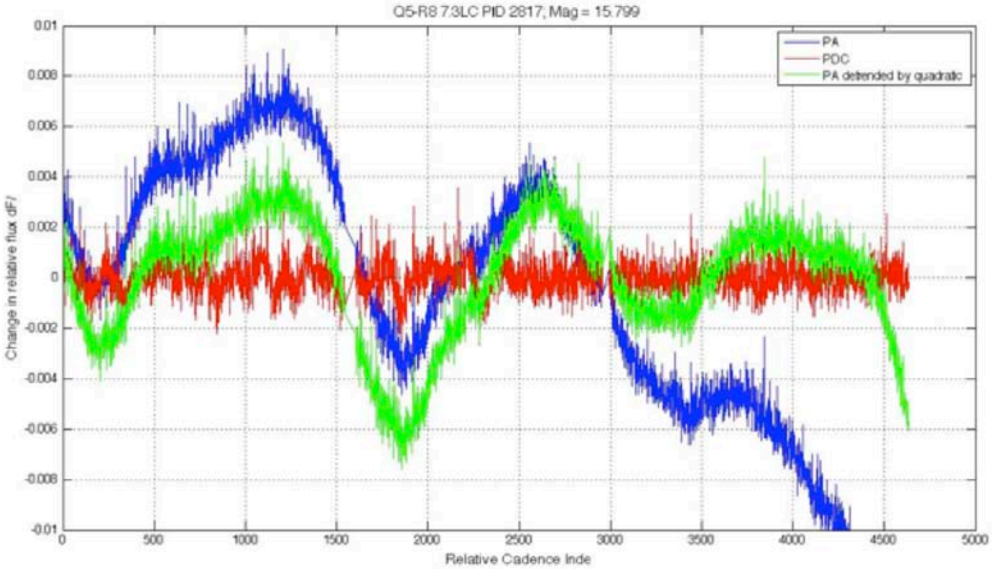


Figure 7-13. Q5 example of PDC removal of harmonic stellar variability. The amplitude of the variability with respect to a quadratic trend is ± 0.54%, just over the 0.50% threshold, before initial cotrending. Standard cotrending eliminated the variability but increased the noise, which PDC did not detect in this case. If PDC had detected the increase in noise, it would have retained the result of the harmonic-removed cotrending, and restored the harmonic content before export to MAST.

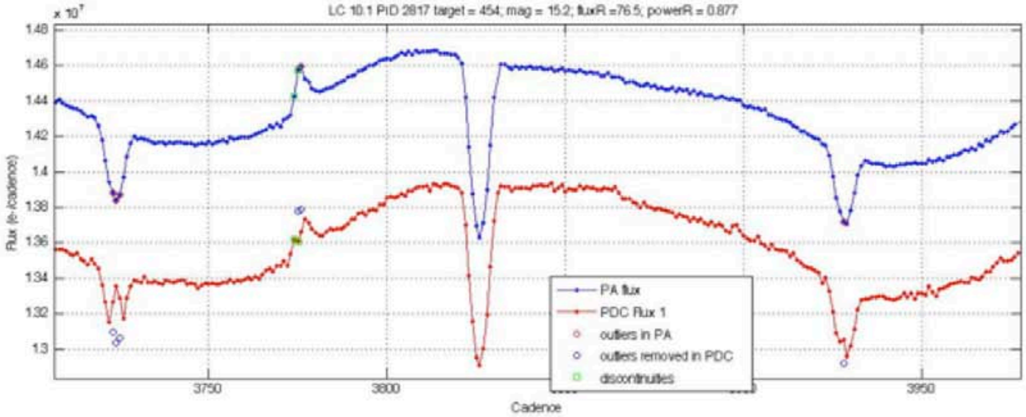


Figure 7-14. Q5 example of astrophysical events, possibly flares, identified by PDC and partially removed from the corrected light curve at cadence 3775. Open green squares show astrophysical events identified as discontinuity anomalies and "corrected." Unlike outliers, discontinuities are not restored to the light curves before delivery to MAST.

7.5 Summary

The primary tasks of the PDC module of the *Kepler* science processing Pipeline are to correct systematic and other errors in the raw flux light curves, remove excess flux in the light curves due to crowding of the respective target apertures, and condition light curves for the transiting planet search by identifying and removing flux outliers and filling data gaps. We first presented an overview of the PDC module. We have then shown how random flux discontinuities are identified and corrected. We have discussed the correction of systematic errors and shown examples of the correction of both large and small-scale systematic effects in the flight data. We have described removal of excess flux from light curves due to background sources in the respective target apertures. We have also described the propagation of uncertainties from raw to corrected flux light curves in PDC. Proper correction of systematic errors for all targets, while preserving their native variability over large periods and amplitudes, has proved to be a major challenge. The shortcomings of the current PDC release are recognized, and it remains a work in progress.

8 TRANSITING PLANET SEARCH

8.1 Introduction

The science Pipeline processes photometric data collected from the *Kepler* photometer and furnishes calibrated pixels, raw and systematic error corrected light curves, and centroid time series (Jenkins et al. 2010a; Middour et al. 2010; Haas et al. 2010). These data are passed to the Transiting Planet Search (TPS) module, which applies a wavelet-based, adaptive matched filter to identify transit-like features with durations in the range of 1 to 16 hours (Jenkins 2002). Light curves whose maximum folded detection statistic exceeds 7.1 are designated Threshold Crossing Events and are subjected to a suite of diagnostic tests in the Data Validation module to fit a planetary model to the data and to establish or break confidence in the planetary nature of the transit-like events (further described sections 9 and 10).

Monitoring the photometric precision obtained by *Kepler* is a high priority and is done on a monthly basis as a byproduct of the noise characterization performed by TPS. The photometric precision metric is called Combined Differential Photometric Precision and is defined as the root mean square (RMS) photometric noise on transit timescales. Each month CDPP, along with a suite of performance metrics developed during processing as the data proceed through the Pipeline, is monitored and reported by the Photometric Performance Assessment component (Li et al. 2010). The SOC monitors CDPP for transit durations of 3, 6, and 12 hours. The typical transit duration varies from a few hours for planets close to their stars to 16 hours for a Mars-size orbit (Koch et al. 2010). Thus, TPS contributes in two primary ways: (1) it produces 3-, 6-, and 12-hour CDPP estimates for each star each month, and (2) it searches for periodic transit pulse sequences.

The *Kepler* spacecraft rotates by 90º approximately every 93 days to keep its solar arrays directed towards the Sun (Haas et al. 2010). The first 10 days of science data obtained as the last activity during commissioning is referred to as Q0. There were 34 days of observations during Q1 following Q0 at the same orientation. Subsequent quarters are referred to as Q2, Q3, etc., and these each contain 93 days of observations. Transit searches are performed nominally every three months after each quarterly data set has been downlinked to the ground and processed from CAL through PDC.

As illustrated in Figure 8.1 there are three major subcomponents in TPS needed to facilitate the full transit search. Since each target star falls on a different CCD in each quarter, TPS needs to combine quarterly segments in such a way as to minimize the edge effects and maximize the uniformity of the apparent depths of planetary transit signatures across the entire data set. The first component of TPS “stitches” the quarterly segments of each flux time series together before presenting the light curve to the transit detection component. The second component characterizes the observation noise as a function of time from a transit’s point of view and correlates a transit pulse with the time series to estimate the likelihood that a transit is occurring at each point in time. These tasks are accomplished by a wavelet-based adaptive matched filter (Jenkins 2002). The third and final component of TPS uses the noise characterization and correlation time series to search for periodic transit pulse sequences by folding the data over trial orbital periods spanning the range from one day to the length of the current data set.

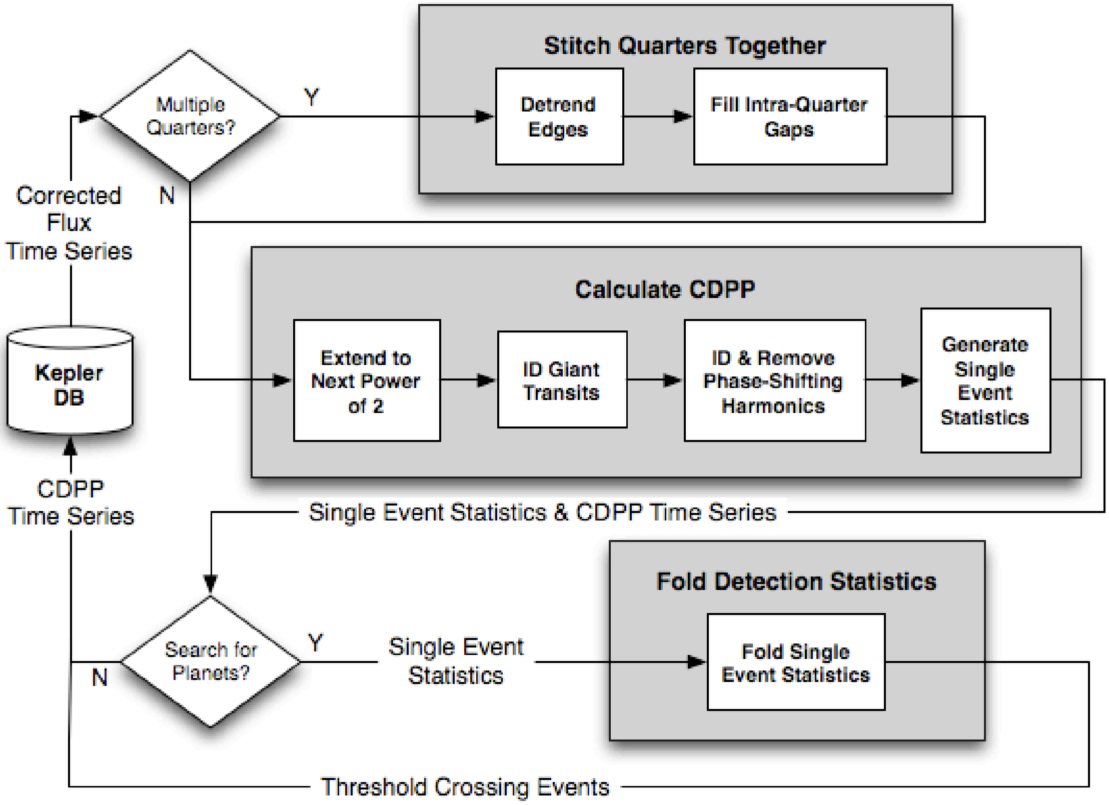

Stitch Quarters Together
Multiple Quarters?
Y
N
Detrend Edges
Fill Intra-Quarter Gaps
Corrected Flux Time Series
Kepler DB
Calculate CDPP
Extend to Next Power of 2
ID Giant Transits
ID & Remove Phase-Shifting Harmonics
Generate Single Event Statistics
CDPP Time Series
Single Event Statistics & CDPP Time Series
Fold Detection Statistics
Search for Planets?
Y
N
Single Event Statistics
Fold Single Event Statistics
Threshold Crossing Events

Figure 8.1. Block diagram for TPS. When TPS is run in transit search mode, the data from the beginning of the mission to the most recent data are first stitched together at the boundaries between quarterly segments. Single-event statistics are generated over a range of transit durations from 1 to 12 hours, which are then folded at trial periods from one day to the length of the observations. A by-product of the generation of single-event statistics is a measure of the photometric precision for each star, called CDPP. Stars for which multiple-event statistics exceed 7.1 are designated Threshold Crossing Events and are written to the *Kepler* Data Base, along with the CDPP time series and other information, such as the epoch and period of the most likely transit pulse train. For monthly data sets, TPS measures the photometric precision achieved for as many as 170,000 target stars, in which case the first and third subcomponents are skipped.

A number of issues identified since science operations commenced on May 12, 2009 have required significant modifications to the Science Processing Pipeline and to TPS. Many target stars exhibit coherent or quasi-coherent oscillations. The wavelet-based detector was designed to deal with solar-like stellar variability for which any such oscillations occur on timescales much shorter than the LC observation interval of 29.4 minutes, and while it works well for broad-band, non-white noise processes, it is not optimal for coherent background noise that is concentrated in the frequency domain. To mitigate this phenomenon, TPS has been modified to include the ability to identify and remove phase-shifting harmonic components. This code is also used by PDC to condition the flux time series prior to identifying and removing instrumental signatures. The algorithm is based on that of Jenkins and Doyle (2003). This step is performed after the quarterly segments have been conditioned, just prior to noise characterization.

This section is organized as follows. Stitching multiple quarters of data together is presented in § 8.2. § 8.3 discusses the identification and removal of harmonic content, the identification of deep transits and eclipses, and summarize the wavelet-based, adaptive-matched filter. § 8.4

describes the software used to fold single-event statistics to search for periodic transit signatures. The conclusion and summary of future work is given in § 8.5.

8.2 Stitching Multiple Quarters

This subcomponent of TPS is under development, as up to now we have only exercised TPS on individual quarterly light curves. This feature will be completed as part of the next SOC development cycle and is scheduled as part of the SOC 7.0 release in mid-2011. Therefore, the details of implementation for this subcomponent are subject to change as we conduct an implementation and test cycle.

The first step in this process is to detrend the edges of each quarterly segment in order to "stitch" together the segments into one continuous time series. The median of each segment is first subtracted from the light curve and divided into it to obtain a time series that represents the time evolution of fractional changes in the brightness of the target star. Next, a line is robustly fitted to the first and last day of each quarter. The slopes and values of the fitted lines at the ends of the segment then completely determine the coefficients of a cubic polynomial that is then subtracted from the segment. Depending on the details of the stellar variability exhibited in the light curve, the cubic polynomial may introduce large excursions from the mean flux level (now zero). To identify if this is the case, the residual is tested against two statistical criteria used to determine if the residual is well-modeled as a zero-mean stochastic process. The number of positive data points is compared to the number of negative data points, and then the area under the positive points is compared to the negative area under the negative points (essentially, the sum of the positive points compared to the sum of the absolute value of the negative points). If the negative and positive metrics of these tests are within a specified tolerance of each other (typically taken to be 20%), then TPS proceeds to the next step. If these criteria are not satisfied, then TPS robustly fits constrained polynomials to the residual whose value and slope are zero at the end points of the segment starting with a quartic polynomial, and retests the residual against these criteria. The order of the polynomial is increased until either the criteria are met or a specified maximum polynomial order (10) is reached.

All polynomials p(x) whose values and slopes are zero at the endpoints 0 and 1 have the form:

$$p(x) = x^2(x-1)^2 q(x), \tag{8-1}$$

where x is the time t normalized as $x = (t - t_1) / (t_N - t_1)$ by the first and last time tags t_1 and t_N, and $q(x) = c_0 + c_1 x^2 + ... + c_{M-1} x^{M-1}$ is a normal M^th^-order polynomial. The design matrix for solving for such a constrained polynomial is the standard one whose rows are multiplied by the constraint polynomial terms evaluated at each time tag, $x^2(x-1)^2$. That is, the elements of design matrix A are defined by:

$$A_{i,j} = x_i^2(x_i - 1)^2 x_i^j, \tag{8-2}$$

where i = 1,...,N and j = 0,...,M−1.

The next step in stitching the quarterly segments together is to fill the gaps between them. Gaps within quarters are filled by PDC prior to TPS. Short data gaps of a few days or less are filled by an autocorrelation approach using auto-regressive stochastic modeling (Jenkins 2002). Filling longer data gaps, as is necessary for target stars that are not observed each and every quarter, will require other methods. Currently, the long data gap fill method employed by the Pipeline reflects and tapers data on either side of a gap across the gap and then performs a wavelet analysis in order to adjust the fill data to make the amplitude of the stochastic variations consistent with those of data adjacent to the gap. For light curves with entire quarterly segments missing we will likely modify the reflection approach to deal with cases where the gap may be longer than the available data on one or the other side of the gap, as will happen for targets observed in Q1 but not Q2, and then observed in Q3. The gap filling is necessary to allow the next subcomponent of TPS to operate: applying a wavelet-based, matched filter to either calculate CDPP on a monthly or quarterly basis, or to furnish single-event statistics for the full transit search.

8.3 Calculating CDPP

This component of TPS performs the noise characterization central to the task of detecting transiting extrasolar planets. Prior to applying the wavelet-based, adaptive matched filter it is necessary to extend the flux time series to the next power of 2, as this detection scheme invokes Fast Fourier Transforms (FFTs). It is also necessary to screen out coherent and quasi-coherent harmonic signals in the flux time series, as well as deep transits and eclipses, for which the original wavelet-based approach is not well suited. The first step is to identify and remove strong transit signatures and then extend the time series to the next power of 2 via Jenkins's (2002) methods.

8.3.1 Identifying Giant Transiting Planets and Eclipses

The transit detection scheme baselined for the Pipeline is designed to search for weak transit signatures buried in solar-like variability and observation noise. The noise characterization will sense the presence of strong transit signatures from giant planets or eclipsing binaries and tends to "annihilate" them as part of the pre-whitening step in the detection process. To identify such signatures, TPS applies Akaike's Information Criterion (Akaike 1974) to fit a polynomial to each light curve and identifies clusters of negative residuals that are many median absolute deviations from the fit. Such points are removed and the process is repeated until no additional sets of consecutive, highly negative points are identified. The residuals are subjected to a search for harmonic signatures.

8.3.2 Identifying and Removing Phase-Shifting Harmonics

Once the deep transits and eclipses have been identified, the cadences containing such events are temporarily filled using the autocorrelation-based short data gap fill algorithm. The time series is extended to the next power of 2 using Jenkins's approach (Jenkins 2002). A Hanning window-weighted periodogram is formed. The background Power Spectral Density (PSD) of any broadband, non-white noise process in the data is estimated in a two-step process. First, a median filter is applied to the periodogram and the result is smoothed with a moving average window. The median filter ignores isolated peaks in the periodogram. Next this background PSD estimate is divided pointwise into the periodogram. The whitened PSD is then examined for statistically significant peaks, and the frequency bins of such peaks are fed to a nonlinear least squares fitter as the seed values for a fit in the time domain to phase-shifting harmonic signals. These are sinusoids in time that allow for the center frequency to shift linearly in time. For complex harmonic signals, this process can take a while. Figure 8.2 shows two examples of light curves with strong, coherent harmonic features as they are fitted and removed with this approach. The resulting harmonic-cleaned flux time series is then ready for the wavelet-based matched filter.

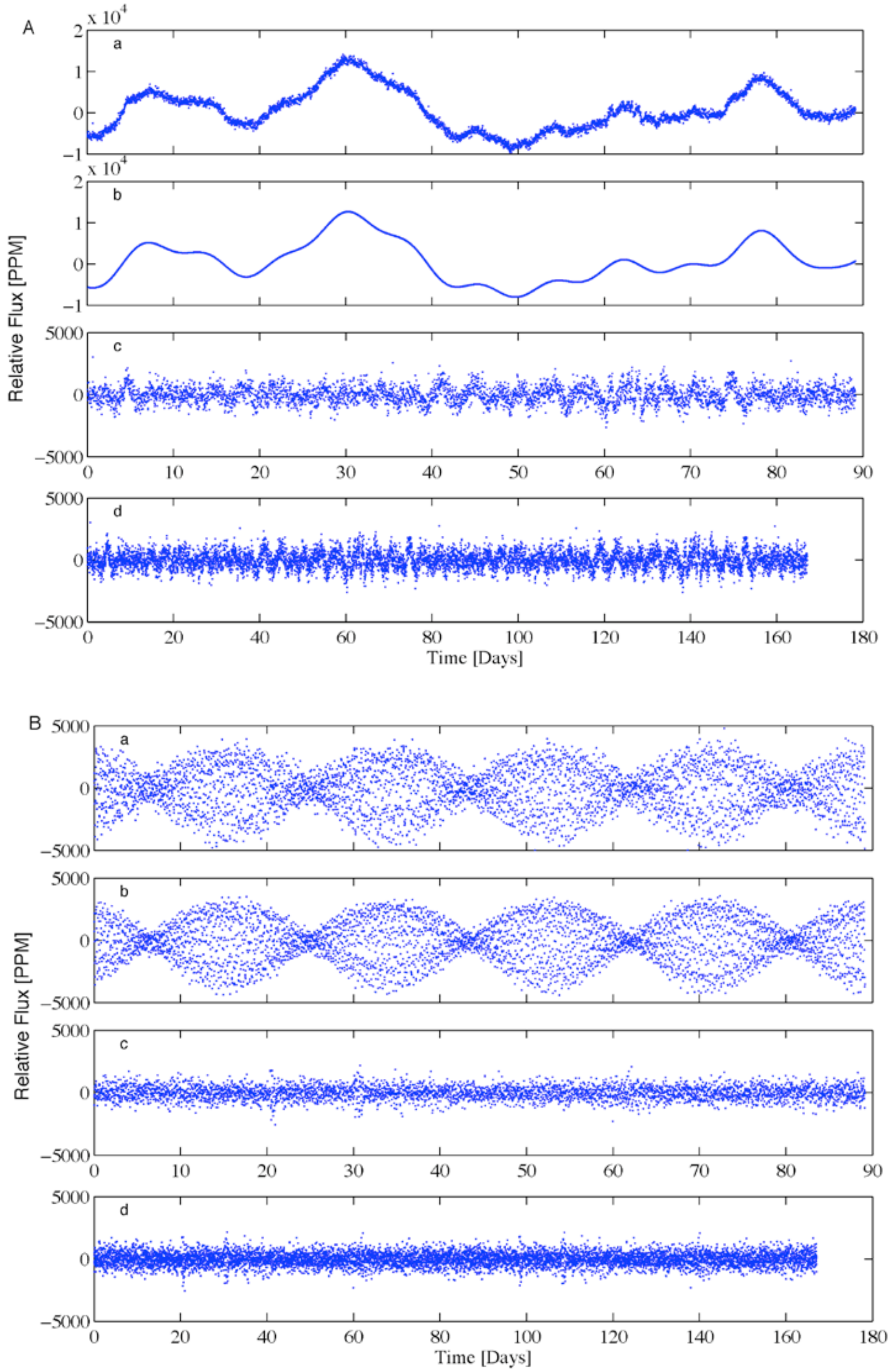


Figure 8.2. Harmonic removal and extension of two flux time series. A: a star with low-frequency oscillations; B: a star with high-frequency oscillations and amplitude modulation. a: original flux time series; b: detected harmonic signature; c: flux time series with harmonics subtracted; d: harmonic-free time series extended to 8192 samples.

8.3.3 A Wavelet-Based Matched Filter

The optimal detector for a deterministic signal in colored Gaussian noise is a pre-whitening filter followed by a simple matched filter (Kay 1999). In TPS we implement the wavelet-based matched filter (Jenkins 2002) using Debauchies's (1988) 12-tap wavelets. The wavelet-based matched filter uses an octave-band filter bank to separate the input flux time series into different band passes to estimate the PSD of the background noise process as a function of time. This scheme is analogous to a graphic equalizer for an audio system. TPS constantly measures the "loudness" of the signal in each bandpass and then dials the gain for that channel so that the resulting noise power is flat across the entire spectrum. Flattening the power spectrum transforms the detection problem for colored noise into a simple one for white Gaussian noise (WGN), but also distorts transit waveforms in the flux time series. TPS correlates the trial transit pulse with the input flux time series in the whitened domain, accounting for the distortion resulting from the pre-whitening process. This is analogous to visiting a funhouse "hall of mirrors" with a friend of yours and seeking to identify your friend's face by looking in the mirrors. By examining the way that your own face is distorted in each mirror, you can predict what your friend's face will look like in each particular mirror, given that you know what your friend's face looks like without distortion. Let's briefly review the wavelet-based matched filter.

Let *x(n)* be a flux time series. Then we define the over-complete wavelet transform (OWT) of *x(n)* as:

$$W\{x(n)\} = \{x_1(n), x_2(n), \ldots x_M(n)\}\ , \tag{8-3}$$

where:

$$x_i(n) = h_i(n) \ast x(n),\ i = 1, 2, \ldots , M, \tag{8-4}$$

and '$\ast$' denotes convolution, and $h_i(n)$ for $i = 1,\ldots,M$ are the impulse responses of the filters in the filter bank implementation of the wavelet expansion with corresponding frequency responses $H_i(\omega)$ for $i = 1,\ldots,M$.

In Figure 8.3 we present a signal flow graph illustrating the process. The filter, H_1, is a high-pass filter that passes frequency content from half the Nyquist frequency, $f_{Nyquist}$, to the Nyquist frequency ($[f_{Nyquist} / 2, f_{Nyquist}]$). The next filter, H_2, passes frequency content in the interval $[f_{Nyquist} / 4; f_{Nyquist} / 2]$, as illustrated in Figure 8.4. Each successive filter passes frequency content in a lower bandpass until it reaches the final filter, HM, the lowest bandpass, which passes DC content as well. The number of filters is dictated by the number of observations and the length of the mother wavelet filter chosen to implement the filter bank. In this wavelet filter bank there is no decimation of the outputs, so there are M times as many points in the wavelet expansion of a flux time series, $\{x_i(n)\}$, $i = 1,\ldots,M$, as there were in the original flux time series *x(n).* This representation has the advantage of being shift invariant, so that we need only compute the wavelet expansion of a trial transit pulse, *s(n),* once. The noise in each channel of the filter bank is assumed to be white and Gaussian and its power is estimated as a function of time by a moving variance estimator (essentially a moving average of the squares of the data points) with an analysis window chosen to be significantly longer than the duration of the trial transit pulse.

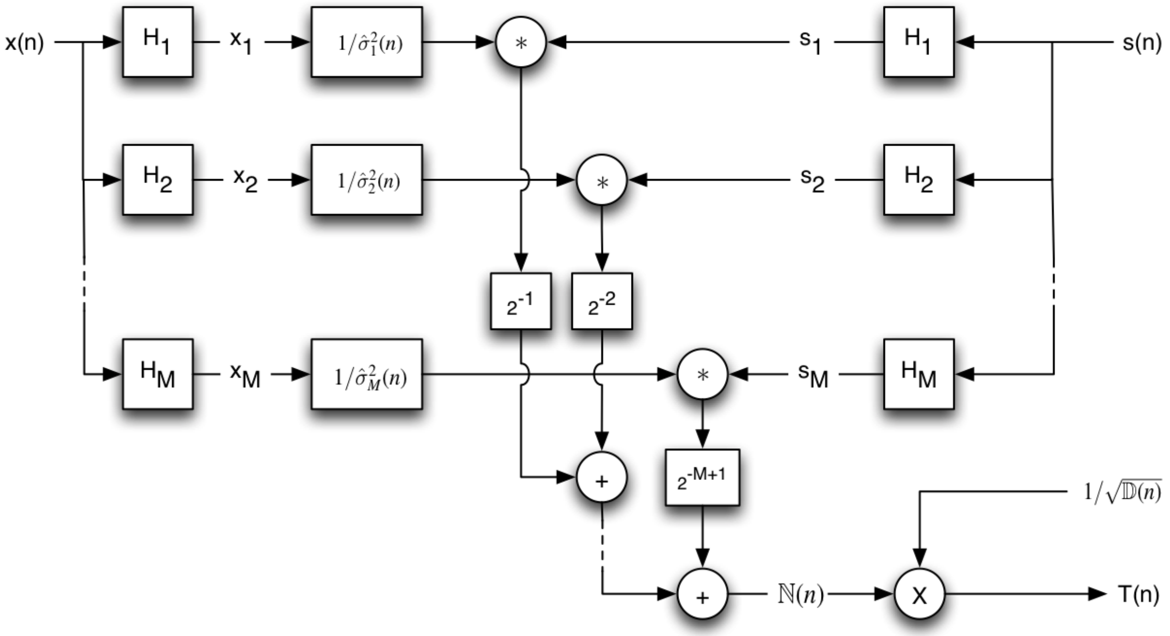


Figure 8.3. Signal flow diagram for TPS. The wavelet-based matched filter is implemented as a filter bank with bandpass filters $H_1,...,H_M$ progressing from high frequencies to low frequencies. The flux time series, *x(n),* is expanded into *M* time series *x_i(n),* for *i* = 1,...,*M*. Noise power, $\sigma_i^2(n)$, *i* = 1,...,*M* is estimated for each bandpass and then divided into the channel time series, *x_i(n),* in order to whiten the flux time series in the wavelet domain. The trial transit pulse is processed through a copy of the filter bank and convolved with the doubly pre-whitened flux time series in each bandpass. Parseval's theorem for undecimated, wavelet representations allows us to combine the results for each bandpass together to form the numerator term, N*(n),* of Eq. 8-7. A similar filter bank arrangement is used to furnish D*(n)* from Eq. 8-7 by replacing the flux time series *x(n)* in this flow diagram with the trial transit pulse *s(n),* and by using the same bandpass noise estimates to inform pre-whitening. The single-event detection statistic, *T(n)*, is obtained by dividing the correlation term, N*(n)* by the square root of the denominator term, D*(n).*

The detection statistic is computed by multiplying the whitened wavelet coefficients of the data by the whitened wavelet coefficients of the transit pulse:

$$T = \frac{\tilde{x}\cdot\tilde{s}}{\sqrt{\tilde{s}\cdot\tilde{s}}} = \frac{\sum_{i=1}^{M} 2^{-\min(i,M-1)} \sum_{n=1}^{N} \left[x_i(n)/\hat{\sigma}_i(n)\right]\left[s_i(n)/\hat{\sigma}_i(n)\right]}{\sqrt{\sum_{i=1}^{M} 2^{-\min(i,M-1)} \sum_{n=1}^{N} s_i^2(n)/\hat{\sigma}_i^2(n)}}, \tag{8-5}$$

where the time-varying channel variance estimates are given by:

$$\hat{\sigma}_i^2(n) \; = \; \frac{1}{2^i K+1} \sum_{k=n-2^{i-1}K}^{n+2^{i-1}K} x_i^2(k), \; i=1,\square\;, M, \tag{8-6}$$

where each component *x_i(n)* is periodically extended in the usual fashion and 2*K*+1 is the length of the variance estimation window for the shortest time scale. In TPS, *K* is a parameter set to typically 50 times the trial transit duration.

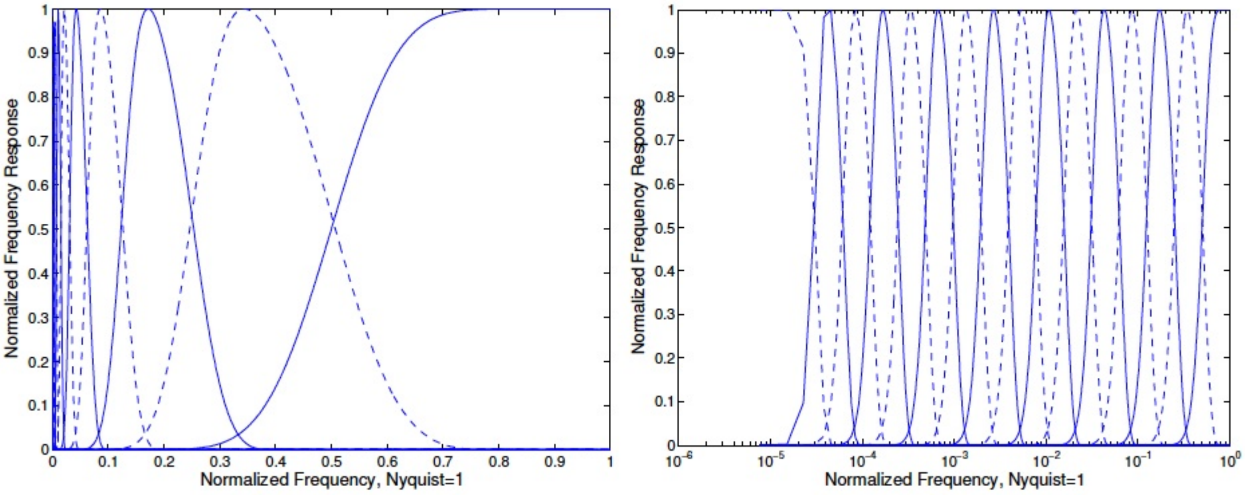


Figure 8.4. Frequency responses of the filters in the octave-band filter bank for a wavelet expansion corresponding to the signal flow graph in Figure 8.3 using Debauchies's 12-tap filter. Left: frequency responses on a linear frequency scale. Right: frequency response on a logarithmic frequency scale, illustrating the "constant-Q" property of an octave-band wavelet analysis.

To compute the detection statistic, *T(n),* for a given transit pulse centered at all possible time steps, we simply "doubly whiten" W {*x(n)*} (i. e., divide *x_i(n)* point-wise by $\hat{\sigma}_i^2(n),$ for *i* = 1,...,M), correlate the results with W {*s(n)*}, and apply the dot product relation, performing the analogous operations for the denominator, noting that $\hat{\sigma}_i^{-2}(n)$ is itself a time series:

$$T(n) = \frac{N(n)}{\sqrt{D(n)}} = \frac{\sum_{i=1}^{M} 2^{-\min(i,\ M-1)} \left[x_i(n) / \hat{\sigma}(n) \right] * s_i(-n)}{\sqrt{\sum_{i=1}^{M} 2^{-\min(i,\ M-1)}\ \hat{\sigma}_i^{-2}(n) * s_i^2(-n)}}. \qquad (8\text{-}7)$$

Note that the "−" in *s_i(−n)* indicates time reversal. The numerator term, N*(n),* is essentially the correlation of the reference transit pulse with the data. If the data were WGN, the result could be obtained by simply convolving the transit pulse with the flux time series. The expected value of Equation 8-7 under that alternative hypothesis for which *x_i(n)* = *s_i(n)* is

$$\sqrt{\sum_{i=1}^{M} 2^{-\min(i,M-1)}}\ \hat{\sigma}_i^{-2}(n) * s_i^2(-n).$$

Thus, $\sqrt{D(n)}$ is the expected signal-to-noise ratio (SNR) of the reference transit in the data as a function of time. The CDPP estimate is obtained as:

$$CDPP(n) = 1 \times 10^6 / \sqrt{\mathrm{D}(n)}\ , \qquad (8\text{-}8)$$

in units of parts per million.

For stars with identified giant planet transits or eclipses, an alternate route is taken to estimate the correlation and expected SNR. The data located in transit are removed and filled by a simple linear interpolation. The resulting time series is then high-pass filtered to remove trends on timescales >3 days and then a simple matched filter is convolved with the resulting time series. A moving variance supplies the information necessary to inform the expected SNR. Figure 8.5 illustrates the process of estimating CDPP for a star exhibiting strong transit-like features. Once the time-varying power spectral analysis performed by TPS, we can search for periodic transit pulses.

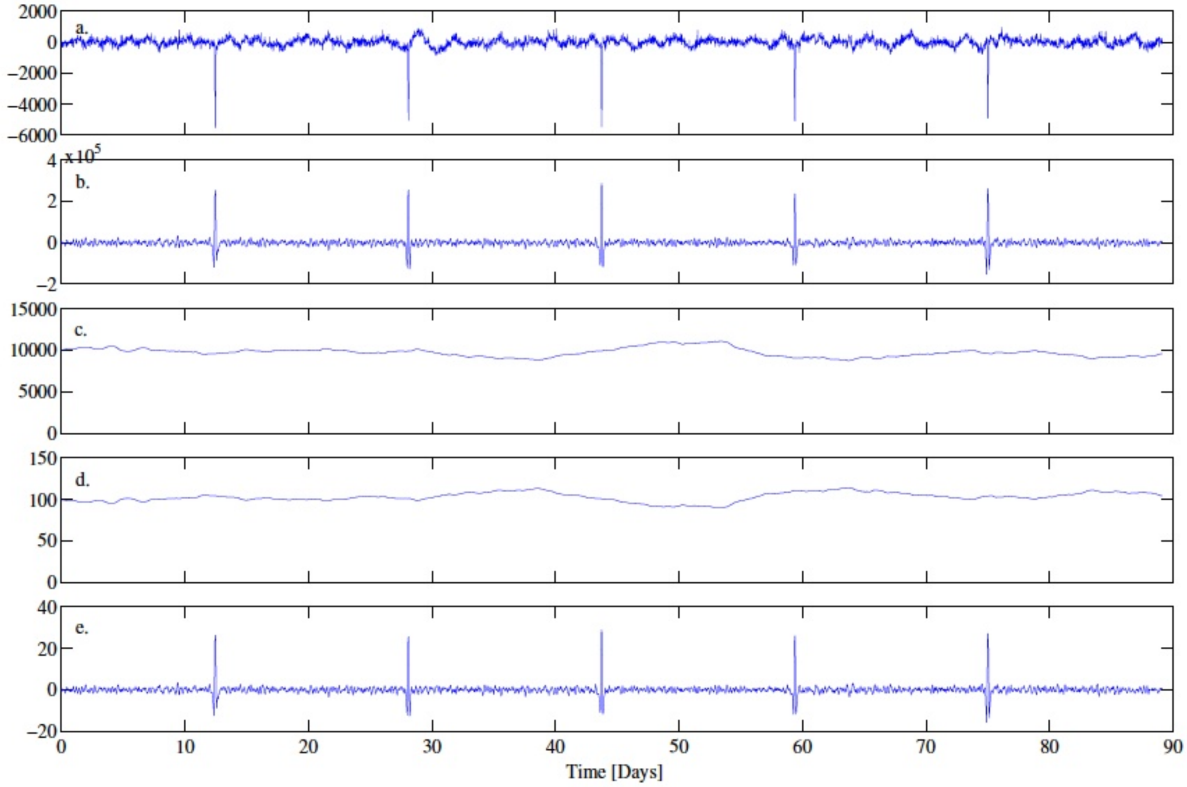


Figure 8.5. Calculation of CDPP for one target star. (a) Normalized target flux in parts per million (ppm). (b) Correlation time series N*(n)* from Eq. 8-7. (c) Normalization time series D*(n)* from Eq. 8-7. (d) Three-hour CDPP time series. (e) Single-event statistic time series, *T(n).* In all cases, the trial transit pulse, *s(n),* is a square pulse of unit depth and three-hour duration.

8.4 Folding Detection Statistics

The third and final stage of TPS is to fold the single-event detection statistics developed in 8.3 over the range of potential orbital periods. Applying a matched filter for a deterministic signal with unknown parameters is equivalent to performing a linear least-squares fit at each trial point in parameter space, which for transit sequences is the triple composed of the epoch (or time to first transit), orbital period, and transit duration, $\{t_0, T_p, D\}$. Clearly, we can't test for all possible points so we must lay down a grid in parameter space that balances the need to preserve sensitivity with the need for speed.

As given by Jenkins and Doyle (1996), one measure of sensitivity is the correlation coefficient between the model planetary signatures of neighboring points in parameter space. If we specify the minimum correlation coefficient, ρ, required between neighboring models, then we can derive the step sizes in period, epoch, and duration. For the case of simple rectangular pulse trains, a real transit will have a correlation coefficient with the best-matched model of no worse than $\rho + (1 - \rho) / 2$. The correlation coefficient as a function of the change in epoch, Δt_0, is given by $c(\Delta t_0) = (D - \Delta t_0) / D = 1 - \Delta t_0 / D$, where D is the trial transit duration.

Similarly, for a change in transit duration we have $c(\Delta D) = (D - \Delta D) / D = 1 - \Delta D / D$, so that $\Delta D = (1 - \rho)D$. So for a given minimum correlation coefficient, ρ, we have $\Delta t_0 = (1 - \rho)D$. The step size in orbital period, ΔT_p, is strongly influenced by the number of transits expected in the data set at the trial period itself. In this case, $c \approx 1 - N\Delta T_p / 4D$, where N is the number of expected transits, or the ratio of the length of the data set to the trial period, so that:

$$\Delta T_p = 4(1-\rho)D / N = 4\Delta t_0 / N. \tag{8-9}$$

The default choice for TPS is $\rho = 0.9$ for orbital period and epoch. Trial transit duration is specified by a discrete list furnished to TPS, and we have accepted $\rho = 0{:}5$ for the transit duration minimum correlation coefficient, although this will be tightened up as we proceed to search for transiting planets over multiple quarters. Starting with the minimum trial orbital period (usually one day), TPS applies equation 8-9 to determine the next trial orbital period, continuing until the maximum trial orbital period, the length of the time series, is reached. To form a multiple-event statistic for given point $\{t_0; T_p; D\}$, TPS computes the correlation and SNR time series, $\mathrm{N}(n)$ and $\mathrm{D}(n)$, and then loops over the trial orbital periods, folding these time series at each orbital period (rounded to the nearest number of samples) and summing the numerator and denominator terms falling in the each epoch bin. TPS identifies the maximum multiple-event statistic and its corresponding epoch. TPS also identifies and returns the maximum single-event statistic for each trial transit duration (McCauliff et al. 2010). Figure 8.7 illustrates this process for the flux time series appearing in Figure 8.5.

To preserve sensitivity to short-duration transits and small orbital periods, TPS supports a super-resolution search with respect to epoch and orbital period. This is accomplished by shifting the trial transit pulse by a fraction of a transit duration, generating the single-event statistic time series components for this shifted transit, then interleaving the results with the original transit pulse's single-event statistics. For example, a three-hour transit pulse lasts six LC samples: [0,–1,–1,–1,–1,–1,0]. Shifting this transit by 10 minutes or one-third of an LC earlier, we obtain the sequence [–1/3,–1,–1,–1,–1,–1,–2/3,0] with corresponding single-event detection statistics $\mathrm{N}_{+1/3}(n)$ and $\mathrm{D}_{+1/3}(n)$. Shifting the original transit pulse by 10 minutes later, we obtain the sequence [0,–2/3,–1,–1,–1,–1,–1,–1/3] with corresponding single-event detection statistics $\mathrm{N}_{-1/3}(n)$ and $\mathrm{D}_{-1/3}(n)$. We combine the results from all three analyses schematically as:

$$\mathrm{N}(n) = \{\cdots, \mathrm{N}_{+1/3}(k), \mathrm{N}_0(k), \mathrm{N}_{-1/3}(k), \mathrm{N}_{+1/3}(k+1), \mathrm{N}_0(k+1), \mathrm{N}_{-1/3}(k+1), \cdots\}, \tag{8-10}$$

where we've denoted the original time series as $\mathrm{N}_0(n)$. A similar expression applies for the super-resolution denominator term, $\mathrm{D}(n)$. The folding proceeds exactly as before, except that now a sample is 9.8 minutes rather than 29.4 minutes.

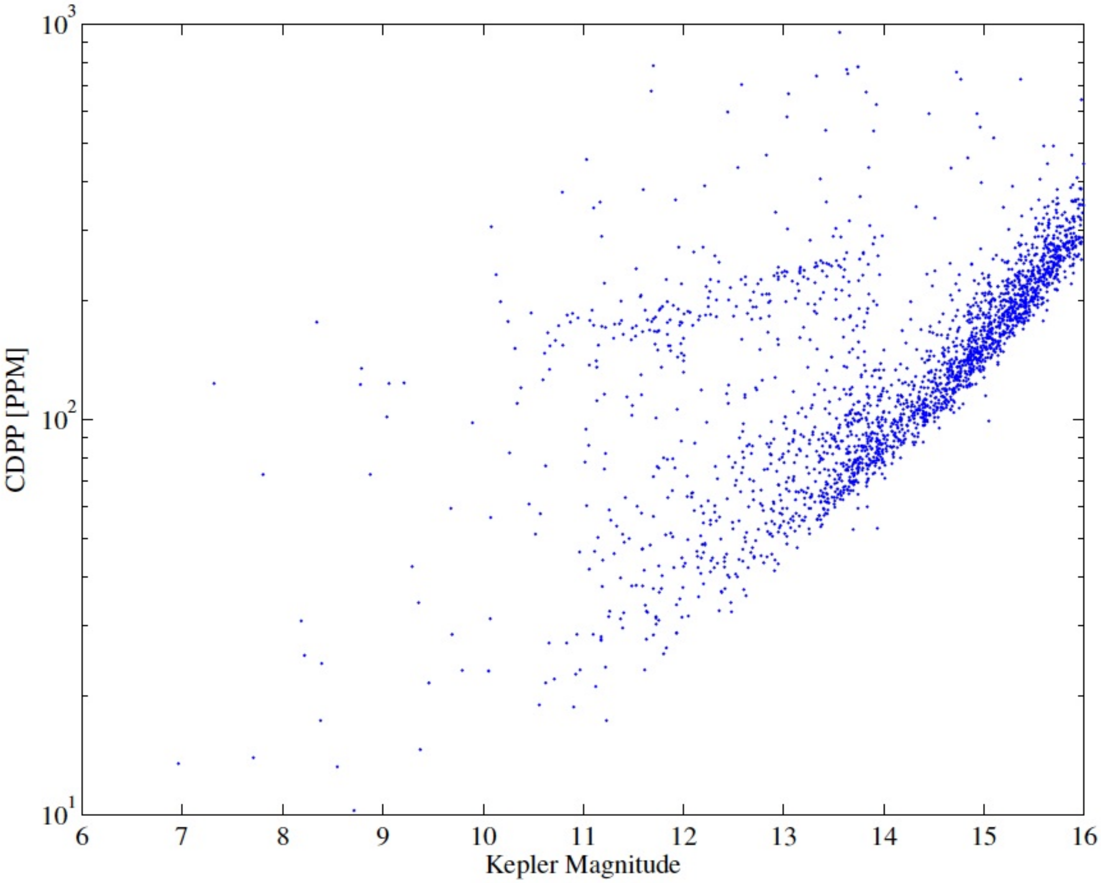


Figure 8.6. Three-hour CDPP as a function of *Kepler* magnitude for 2,286 stars on module 7, output 3, for one representative quarter.

8.5 Summary

Fifteen planets have been announced by the *Kepler* team as of February 2011 (Borucki et al. 2010, 2011b) including the remarkable six-planet system, Kepler-11 (Lissauer et al. 2011). Several hundred potential planets are currently being vetted by the Kepler Science Team. TPS has been quite productive in identifying Threshold Crossing Events in individual quarters and soon will be capable of detecting planetary signatures across the complete data set. TPS does trigger TCEs for a significant number of non-transit or eclipse events due to pixel sensitivity dropouts, flare events, and other isolated and cluster outliers. Near-term development includes better identification of step discontinuities due to pixel sensitivity dropouts in systematic error corrections made by PDC and also increasing the robustness of TPS to such events. These steps should reduce the number of TCEs that are analyzed by the DV component while preserving TPS's sensitivity to transit signatures.

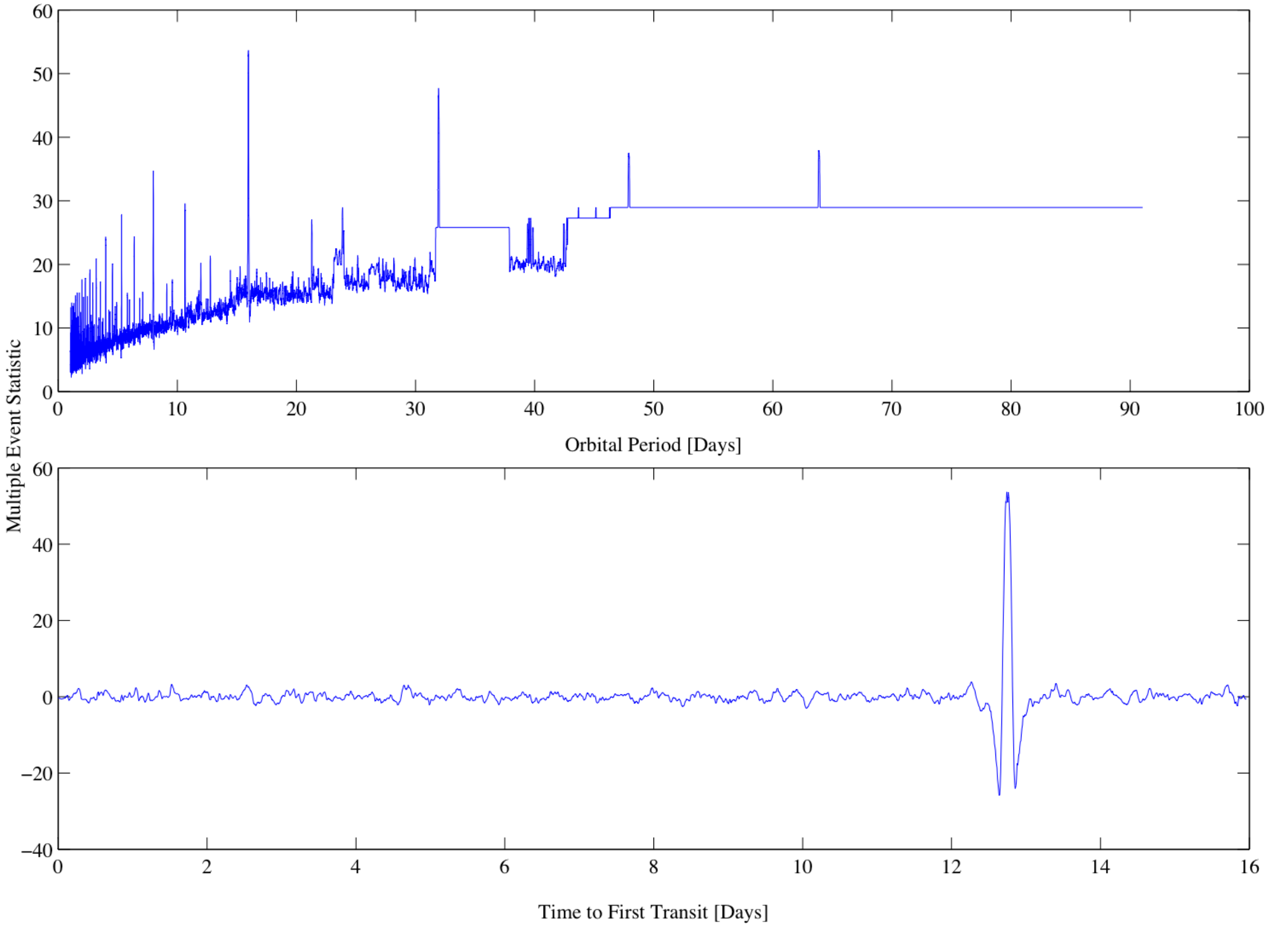


Figure 8.7. Multiple-event statistics determined by folding the single-event statistics distribution. Top: maximum multiple-event statistic as a function of fold interval (orbital period), showing a peak at 15.97 days, corresponding to the orbital period of the transiting object in the data of Figure 8.4. Bottom: multiple-event statistic for 15.97 day period as a function of lag time, showing a peak at 12.74 days, corresponding to the mid-time of the first transit shown in Figure 8.6.

9 DATA VALIDATION

9.1 Introduction

The primary task of the Data Validation (DV) software component is to perform an automated validation of the many threshold crossing events produced by the TPS module. DV is provided with the TCE for each threshold-crossing target corresponding to the maximum multiple-event detection statistic over the set of trial transit pulse durations. A transiting planet model is fitted to the light curve for the given target to obtain model parameters for the initial planet candidate. The model fit is subtracted from the light curve and a search for additional transiting planets is performed on the residual light curve. If an additional detection occurs, the transiting planet model is fitted to the residual flux based on the new TCE. A suite of automated validation tests is performed when no additional planet candidates can be identified through the multiple planet search (or when the operator-configurable iteration limit is reached). The main purpose of the automated validation tests is to facilitate the identification of true planet candidates from the large number of false positive transiting planet detections, astrophysical and otherwise.

The automated tests performed in DV are by no means the final validation of new planet discoveries by the *Kepler Mission*. In fact, DV is only the beginning of the vetting process for *Kepler* planet discoveries. Pipeline results from TPS and DV are exported to the *Kepler* Science Analysis System (KSAS). There, they are federated with prior results, and planet candidates are scored and ranked in accordance with a list of science criteria. Promising planet candidates are screened by the Threshold Crossing Event Review Team (TCERT), which is composed of the *Kepler* Science Principal Investigator and selected members of the *Kepler* Science Office and Science Team. Very promising candidates suited to vetting from the ground are further investigated from ground-based observatories through the Follow-up Observing Program (FOP) (Gautier et al. 2010). Planet discoveries are announced only after extensive review and follow-up observation where applicable.

In this section, we describe the nature of the automated validation tests. § 9.2 presents an overview of DV and data flow through this software module. § 9.3 describes the transiting planet signal generator and limb-darkening model. § 9.4 describes the automated validation tests for centroid motion, eclipsing binary discrimination, and detection significance. Conclusions are discussed in § 9.5.

9.2 Overview and Data Flow

DV addresses only LC targets for which the transiting planet detection threshold is exceeded in TPS. The DV unit of work may include one or more targets to support load balancing on worker machines in the Pipeline cluster (Klaus et al. 2010a). The duration of DV's standard unit of work is a single science data acquisition quarter (~93 days). A future version of DV will accommodate light curves spanning multiple LC target tables and quarterly spacecraft rolls. At that point, DV will likely be invoked quarterly with all data acquired since the beginning of Quarter 1 (12 May 2009) for targets with TCEs.

Figure 9.1 illustrates data flow through DV within the *Kepler* SOC Pipeline, including major fields in DV input and output structures and DV's primary components. DV also automatically generates an extensive report in PDF (not shown in Figure 9.1) for each target processed and saves it to the *Kepler* Database (McCauliff et al. 2010) with other DV outputs when the Pipeline module is executed.

Data timestamps in the Pipeline are specified in Modified Julian Days (MJD), representing the start, middle, and end of each LC observation aboard the spacecraft. Timestamps are adjusted for each target to the solar system barycenter to prevent modulation of the transit timing by the heliocentric orbit of the photometer. Sky coordinates of individual targets are obtained from the *Kepler Input Catalog* and NAIF SPICE kernels. The latter contain the reconstructed spacecraft trajectory and solar system ephemeris and are produced by the JPL Navigation organization. Timestamp corrections also include small offsets introduced by the multiplexed readout of the CCD array (KIH, section 5.1).

Ancillary engineering data (sensor data such as temperature and state of reaction wheels) from the spacecraft, ancillary Pipeline data from other SOC software modules, and motion polynomials from PA are utilized to detrend the target light curves for planet model fitting and for performing the DV Centroid Test. Ancillary data and motion polynomials are first synchronized to the LC timestamps.

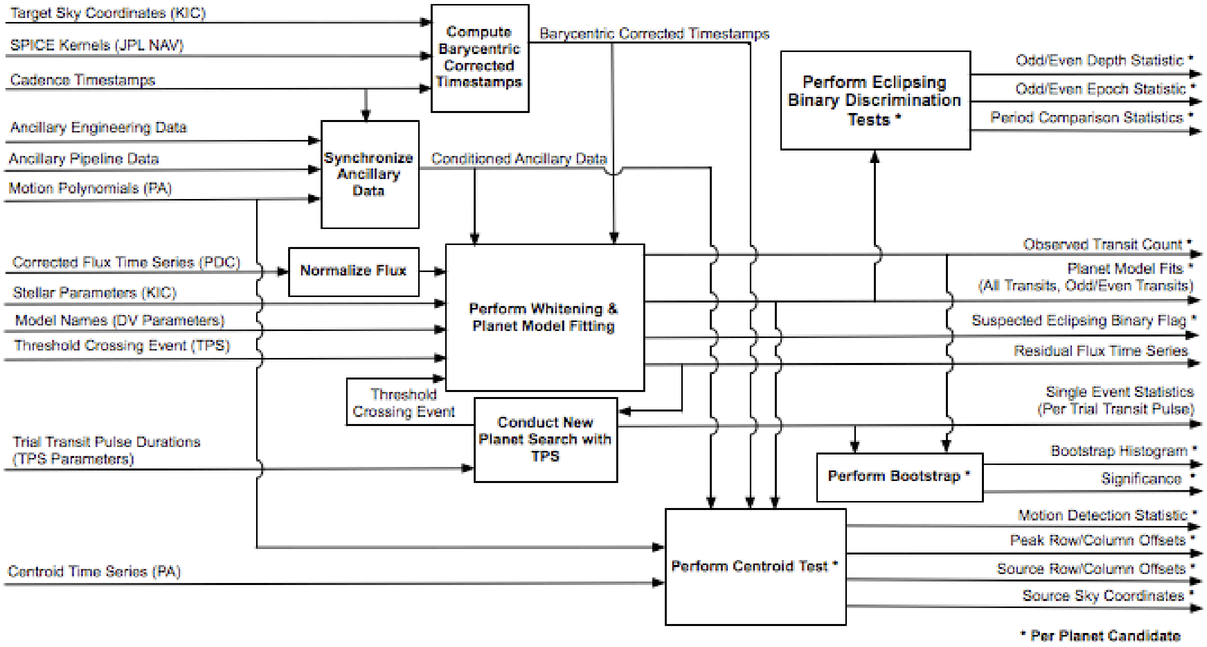

Target Sky Coordinates (KIC)
SPICE Kernels (JPL NAV)
Cadence Timestamps
Compute Barycentric Corrected Timestamps
Barycentric Corrected Timestamps
Perform Eclipsing Binary Discrimination Tests *
Odd/Even Depth Statistic *
Odd/Even Epoch Statistic *
Period Comparison Statistics *
Ancillary Engineering Data
Ancillary Pipeline Data
Motion Polynomials (PA)
Synchronize Ancillary Data
Conditioned Ancillary Data
Corrected Flux Time Series (PDC)
Normalize Flux
Stellar Parameters (KIC)
Model Names (DV Parameters)
Threshold Crossing Event (TPS)
Perform Whitening & Planet Model Fitting
Observed Transit Count *
Planet Model Fits *
(All Transits, Odd/Even Transits)
Suspected Eclipsing Binary Flag *
Residual Flux Time Series
Threshold Crossing Event
Conduct New Planet Search with TPS
Single Event Statistics
(Per Trial Transit Pulse)
Trial Transit Pulse Durations
(TPS Parameters)
Perform Bootstrap *
Bootstrap Histogram *
Significance *
Perform Centroid Test *
Centroid Time Series (PA)
Motion Detection Statistic *
Peak Row/Column Offsets *
Source Row/Column Offsets *
Source Sky Coordinates *
* Per Planet Candidate

Figure 9.1. Data flow diagram for the Data Validation (DV) component of the *Kepler* SOC Pipeline. Processes performed for all planet candidates associated with each DV target and outputs produced for all planet candidates are shown with asterisks. Inputs are obtained from the *Kepler* DB and outputs are written to the *Kepler* DB.

The corrected flux light curve generated in PDC for each DV target is initially normalized such that the values inside the transit indicate the fractional transit depth and those out-of-transit are zero-valued. The normalized flux is whitened to account for stellar variability and fit with a transiting planet model in an iterative process. The transit model is also separately fit to sets of odd and even transits for further analysis of the planet candidate. The DV fit algorithms are sufficiently complex that they are described in detail in section 10, but an overview of the DV transit model generator is described in § 9.3.

After the fitting process is complete, the fitted transits are removed and TPS subjects the residual flux to a search for additional planets. The whitening and fitting process is repeated for a new planet candidate if an additional TCE is generated. The search for additional planets concludes when no additional TCEs are produced or an iteration limit is reached. After all planet candidates have been identified, the final residual flux time series, single-event statistics for all trial transit pulses, parameter values, and associated covariance matrices for all model fits are saved to the *Kepler* DB, as is a flag for each planet candidate that the fitter suspects to be an eclipsing binary.

A centroid motion test is performed on the centroid time series for each target to ascertain whether there is statistically significant motion of the aperture photocenter during transits of the respective planet candidates. Centroid motion can be a strong indicator that observed events may be due to a background eclipsing binary present in the stellar aperture. Centroid motion alone cannot be used to rule out true planet candidates, however, as there will be motion of the centroid for a target with a legitimate transiting planet if there is any significant crowding in the aperture. The peak centroid row and column offsets during transit are determined, and the change in brightness during transit is utilized to determine the actual row and column offsets of the transit (or eclipse) source from the nominal out-of-transit centroid coordinates. The test is also intended to produce the celestial coordinates of the source. The barycentric corrected timestamps are also utilized when the centroid test is performed.

A series of eclipsing binary discrimination tests is conducted on key model-fit parameters to determine whether the planet candidate is statistically likely to be a true transiting planet or an eclipsing binary. The depths of the odd and even transit sequences for each planet candidate are compared statistically for equality. The timing of the first transit in the odd and even transit sequences are compared statistically for consistency with the period for all observed transits. In the cases of both depth and timing, equality is consistent with a true planet. Finally, the period for each of the planet candidates associated with a given target is compared statistically with the next shorter and next longer period of all planet candidates for the given target. Equality here is indicative that the candidate is *not* a true planet.

A statistical bootstrap test is performed for each planet candidate to determine the likelihood that the detection statistic reported in the transiting planet search would have been produced in the absence of any transits by noise alone. A histogram of multiple-event statistics is populated based on single-event statistics computed from the final residual time series for the target and the number of observed transits for the given planet candidate. The probability of false detection (i.e. bootstrap significance) is given by the probability that multiple-event statistics represented in the histogram exceed the value of the maximum multiple-event statistic for the TCE associated with the given planet candidate. The bootstrap results, in addition to the results of the other automated DV tests, are saved to the *Kepler* DB.

9.3 Transit Model

The DV fitter performs iterative fits of a planet model to potential candidate light curves (see § 10). The planet model uses TCE parameters and stellar parameters obtained from the KIC (stellar radius, effective temperature, and surface gravity) to estimate limb-darkening coefficients and compute light curves at barycentric-corrected cadence timestamps.

TCE parameters (duration of trial transit pulse, phase of first transit, orbital period, and maximum event detection statistic) are combined with KIC parameters to generate the following set of parameters to seed the fit: the transit epoch (time to first mid-transit), orbital eccentricity, longitude of periastron, minimum impact parameter, star radius, transit depth, and orbital period. Note that we assume central transits to seed the fit (minimum impact parameter = 0) and circular orbits throughout (eccentricity = 0 and longitude of periastron = 0). Once the initial planet model is generated and fitted, we compute and output the planet semi-major axis, planet radius, transit duration, and transit ingress time in order to obtain a complete set of model parameters. After the initial fit is performed, we use only physical parameters for subsequent fits, which include planet radius and semi-major axis instead of transit depth and orbital period.

Tables developed by Claret (2000) supply nonlinear limb darkening coefficients for a range of stellar parameters. To estimate the coefficients for each TCE, we interpolate across the Claret tables using stellar surface gravity and effective temperature, turbulent velocity and stellar metallicity equal to zero, and values for the *Kepler* R-band. To generate the light curve, we first compute the orbit at the exposure times within each transit (using the *Kepler* CCD exposure time, readout time, and number of exposures per cadence). The orbit is then rotated to obtain the desired minimum impact parameter projected onto the plane of the sky. The Mandel-Agol (2002) methodology is used to compute the light curve at the exposure times from the time-dependent impact parameters, limb-darkening coefficients, and ratio of the planet/star radii. If the normalized radius of the eclipsing body is less than ~0.01, a small-body approximation is implemented which speeds up the algorithm. In this approximation, it is assumed that the surface brightness of a star is constant under the disk of the eclipsing object, and the semi-major axis is large compared to the size of the star so that the orbit is essentially a straight line. The time series that is output from the transit model represents the integral of the transit signal over each LC and is normalized to zero during out-of-transit periods for use by the fitter. The DV inputs are designed to accommodate multiple-planet and limb-darkening models, but we currently only support the Mandel-Agol (2002) analytic models and the Claret (2000) limb-darkening tables.

9.4 Validation Tests

9.4.1 Centroid Test

In this section we describe the implementation of the DV Centroid Test.

9.4.1.1 Overview

The purpose of the DV Centroid Test is to assess correlations between variations in the centroid (photocenter) time series and fitted transit signatures in the corrected flux time series. If the centroid variations are uncorrelated with a transit signature, the transit signature is likely due to variations in flux from the target star and not from a background source within the target aperture. One possible source of such variations is a planetary transit of the target star.

If the centroid variations are highly correlated with a transit signature, the transit signature may be due to a background source such as a faint eclipsing binary. A high correlation does not necessarily rule out a planetary transit of the target star; particularly if the target aperture is crowded, the centroid shift may, in fact, be due to a planetary transit. It is therefore necessary to follow up detected correlations with an estimate of the location of the centroid-perturbing source.

An estimate of the source location in row and column coordinates on the focal plane may be obtained from the fractional depth of the transit feature in the flux time series, the absolute offset of the corresponding feature in the centroid time series, and the nominal out-of-transit centroid value. These row and column coordinates may then be converted to celestial coordinates and compared to the known location of the target star and other nearby background stars.

A measure of the correlation—the detection statistic—is obtained by applying a matched filter to the whitened row and column centroid time series and adding in quadrature. The relevance of the measured correlation—the significance—is developed assuming the detection statistic is a Chi-squared variable with two degrees of freedom. The significance has a value between zero and one. It is likely that a detection statistic at least as large as the one calculated would be obtained from uncorrelated data containing only random statistical fluctuations. For the DV Centroid Test, a reported significance of zero indicates high confidence that transit features in the flux time series are correlated with features in the centroid time series and implies that transit features in the flux time series may be due to a background source. A reported significance of one indicates low confidence of correlation and implies that transit features in the flux time series may be due to a planetary transit of the target star.

9.4.1.2 Implementation

The DV Centroid Test processes one target at a time. The planned implementation is shown on the data flow diagram in Figure 9.2a. Inputs are the corrected flux time series (from PDC), residual flux time series, fitted transit models, barycentric timestamps, row and column centroid time series, and motion polynomials (the latter two are from PA). First, the residual flux time series is used to construct a whitening filter for the corrected flux time series as shown in Figure 9.2b (note that the whitener in the centroid test is performed independently of the whitener in the fitter). Then the corrected flux time series is median-filtered and whitened. The length of the median filter is selected to preserve features with time scales on the order of the shortest fitted transit identified previously in DV by the fitter. Next, the row and column centroid time series are detrended against ancillary data to remove systematic variations correlated with known sources such as differential velocity aberration and temperature variations of the CCD readout electronics. These are then passed through the same median filter as the corrected flux time series. The row and column centroid time series and the fitted transit models are passed to an iterative whitener which produces the row-and-column-whitened centroid time series, the median row and column out-of-transit centroid value, the background source offset from the median centroid (in row and column) and the row and column centroid detection statistic.

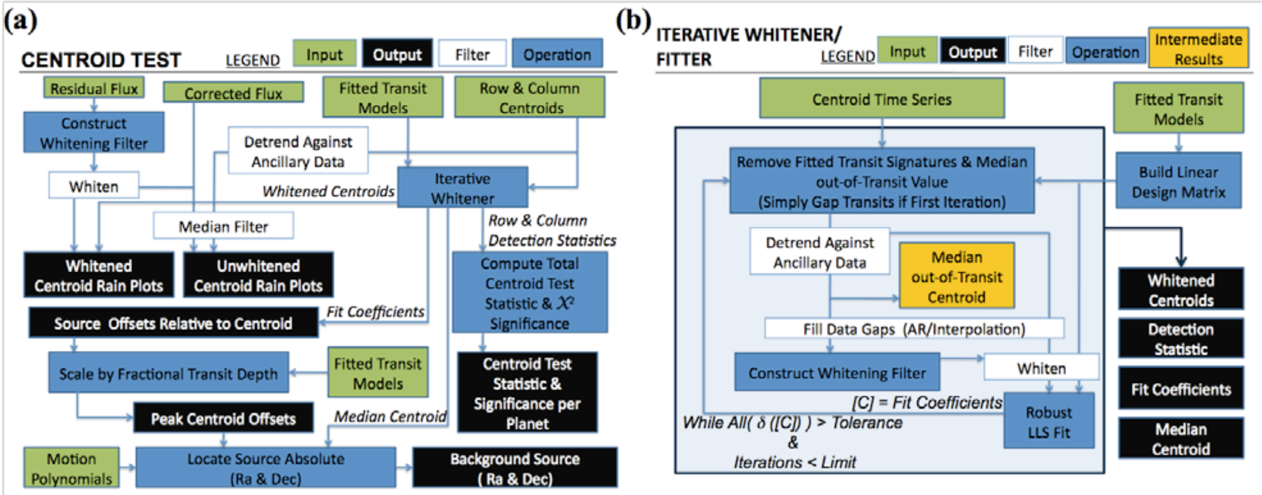

(a)
CENTROID TEST
LEGEND
Input
Output
Filter
Operation
Residual Flux
Corrected Flux
Fitted Transit Models
Row & Column Centroids
Construct Whitening Filter
Detrend Against Ancillary Data
Whiten
Whitened Centroids
Iterative Whitener
Median Filter
Row & Column Detection Statistics
Whitened Centroid Rain Plots
Unwhitened Centroid Rain Plots
Compute Total Centroid Test Statistic & X^2 Significance
Fit Coefficients
Source Offsets Relative to Centroid
Scale by Fractional Transit Depth
Fitted Transit Models
Centroid Test Statistic & Significance per Planet
Peak Centroid Offsets
Median Centroid
Motion Polynomials
Locate Source Absolute (Ra & Dec)
Background Source (Ra & Dec)
(b)
ITERATIVE WHITENER/ FITTER
LEGEND
Input
Output
Filter
Operation
Intermediate Results
Centroid Time Series
Fitted Transit Models
Remove Fitted Transit Signatures & Median out-of-Transit Value (Simply Gap Transits if First Iteration)
Build Linear Design Matrix
Detrend Against Ancillary Data
Median out-of-Transit Centroid
Whitened Centroids
Detection Statistic
Fill Data Gaps (AR/Interpolation)
Construct Whitening Filter
Whiten
Fit Coefficients
[C] = Fit Coefficients
While All(δ ([C])) > Tolerance
&
Iterations < Limit
Robust LLS Fit
Median Centroid

Figure 9.2. (a) Data flow diagram for the DV Centroid Test. (b) Data flow diagram for the iterative whitener used by the DV Centroid Test.

The median-filtered corrected flux is plotted as a function of the detrended median-filtered centroids. This “cloud” or “rain” plot is used as a qualitative diagnostic to check for correlations between the flux and centroid time series. Figure 9.3 shows such plots in the unwhitened domain. The strength and direction of any “wind” observed in these plots indicate the magnitude and sign of any correlations. This plot is generated in the whitened domain as well.

The outputs of the centroid test are the following: centroid detection statistic, significance of the centroid statistic, maximum centroid offset in row and column, location of the background source relative to the nominal target centroid in row and column, and celestial location of the background source in right ascension and declination.

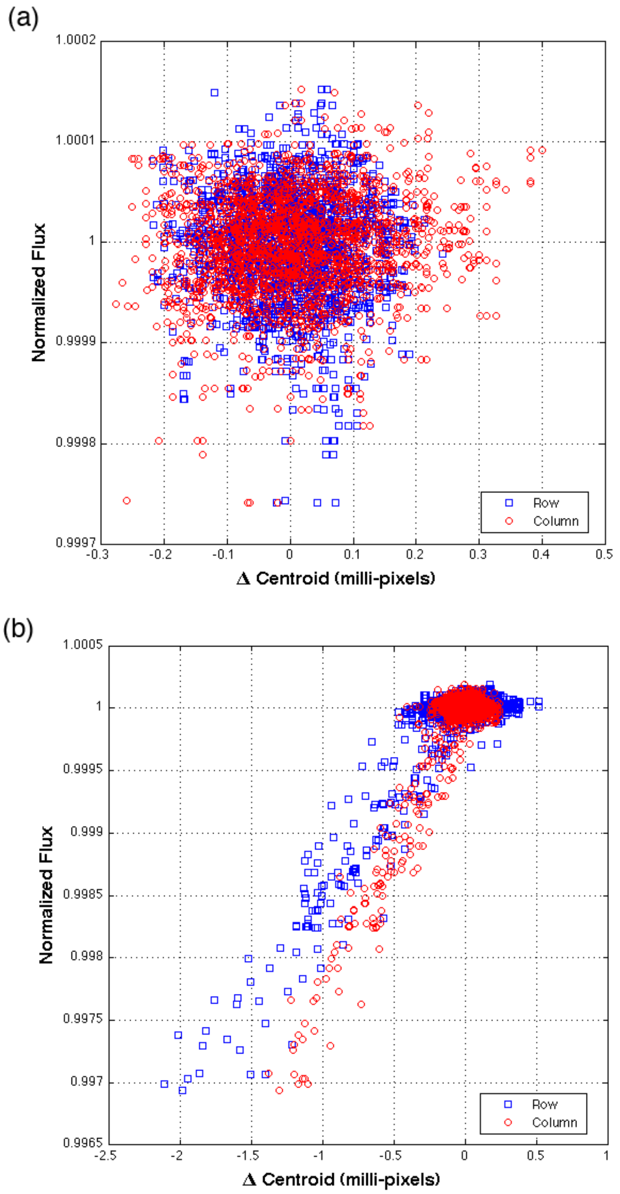


Figure 9.3. Flux-weighted centroids “rain” or “cloud” plot in the unwhitened domain for two synthetic targets. (a) The transit signatures of a planet candidate consist of noise and there is little or no correlation between transit signatures and centroid shifts. (b) A background eclipsing binary within the aperture of a target of interest; transit signature is above noise and correlated with up to 2 milli-pixel centroid shift, hence the “rain” with “wind.”

9.4.1.3 Estimating the Location of the Background Source

Flux in the target aperture is the sum of the flux from the target star plus the flux from all other sources in the aperture, i.e., the background. The change in the centroid due to a background eclipse event is the ratio of the actual distance between the background binary and the target on the CCD to the change in brightness over the target aperture. The centroid shift in terms of the brightness contributed by the target star (*B*) and the brightness contributed by the background source (*b*) is:

$$\delta x = \frac{b\Delta x}{B+b} - \frac{(b-\delta b)\ \Delta x}{B+b-\delta b}, \qquad \text{(9-1)}$$

where *b* = brightness of the background binary, *B* = brightness of the target star, *δb* = change in brightness of the background binary during eclipse, Δ*x* = spatial offset of the background binary from the target centroid, and *δx* = change in centroid during the transit feature.

If the fractional transit depth is small compared to unity, a background binary eclipse can mimic a planetary transit of the target. For all planetary transit candidates identified by the fitter, it is the case that the apparent fractional transit depth is small compared to unity, e.g., *δb/(B + b)* << 1. With this approximation, the offset of the background source relative to the nominal out-of-transit centroid is:

$$\Delta x = \frac{\delta x}{\delta b / (B+b)}. \qquad \text{(9-2)}$$

Standard propagation of errors (assuming independence) gives the variance of Δ*x* as:

$$\sigma_{\Delta x}^2 = \left(\delta b / (B+b)\right)^{-2} \left[\sigma_{\delta x}^2 + (\Delta x)^2 \sigma_{\delta b/(B+b)}^2\right]. \qquad \text{(9-3)}$$

As implemented in the DV Centroid Test, the background source offset from the target centroid (Δx) and its variance ($\sigma_{\delta x}^2$) are determined directly within the iterative whitener from a fit of the whitened centroid data to a linear combination of whitened transit models. The resulting fit coefficients (e.g., model scale factors) are in fact the background source offsets in the unwhitened domain. The corresponding centroid shift (δx) and its variance ($\sigma_{\delta x}^2$) in the unwhitened domain are then calculated using Equations (9-2) and (9-3). Adding Δx to the median out-of-transit centroid gives the background source location in row and column on the CCD. Inverting the motion polynomial for the cadence associated with the median out-of-transit centroid gives the celestial source location in right ascension and declination.

9.4.1.4 Generating the Detection Statistic

The detection statistic provides a measure of the relevance of the Linear Least Squares (LLS) fit results by comparing the size of the fitted signal to the nominal noise level in the data. It is calculated in the whitened domain as the inner product of the data and the candidate signal, normalized by the nominal standard deviation of the data and by the norm of the candidate signal:

$$l = \frac{b \cdot s}{\sigma \sqrt{(s \cdot s)}}, \tag{9-4}$$

where *l* = detection statistic, *b* = raw data, *s* = signal to detect, and σ = nominal standard deviation.

A separate detection statistic is calculated for the row and column centroid time series for each transit signature modeled by the fitter. The sum of the squares of the row and column detection statistics form the total centroid detection statistic. One total detection statistic is produced for each fitted planet candidate. The row and column detection statistics are assumed to be independent Chi-square variables making the total detection statistic a Chi-square variable as well, but with two degrees of freedom. Evaluating the Chi-square cumulative distribution function (CDF) for the total detection statistic value and two degrees of freedom yields the probability of producing a statistic less than or equal to the one observed given uncorrelated data containing only random statistical fluctuations (the null hypothesis). In the DV Centroid Test, the row and column centroid detection statistics are easily determined within the iterative whitener from the output of the last iteration. For each planet candidate, the detection statistic is the square root of the Chi-squared of the corresponding scaled whitened transit model.

According to the DV convention, a significance of one shall be consistent with the detection of a planet and a significance of zero shall be consistent with no planet detected. We therefore report the complement of the Chi-square CDF result as the significance of the detection. The reported statistical significance is a value between zero and one where zero indicates high correlation (the transit feature in the flux time series may be due to a background source), and one indicates no correlation (the transit feature in the flux time series may be due to a transit of the target star).

9.4.2 Eclipsing Binary Discrimination Tests

The eclipses of an eclipsing binary system and the transits of a planet around a star may appear similar in a flux time series. To discriminate between them, we have designed and developed several tests based on their different characteristics. This section describes the tests, which collectively are called the *Eclipsing Binary Discrimination (EBD) Tests.* The EBD tests are statistical hypothesis tests on the consistency of key transit parameters. The fitter provides the parameters for each TCE. The EBD tests consist of the following: Odd/Even Transit Depth Test, Odd/Even Transit Epoch Test, and Orbital Period Test.

The depths of multiple transits of a planet are ideally the same, and the transits of a planet are evenly spaced in time. In contrast, the depths of primary and secondary eclipses of an eclipsing binary system are generally different due to the difference in size and brightness of the two stars. The difference in the epoch times of the primary and secondary eclipses is usually not equal to half of the orbital period of the eclipsing binary system, since the orbit of two stars moving around their gravitational center is often not circular. The *Odd/Even Transit Depth Test* and the *Odd/Even Transit Epoch Test* are designed to distinguish the flux time series of an eclipsing binary system whose primary and secondary eclipses are identified as one TCE. For each TCE, the transits are divided into odd and even sets, and the depths and epoch times of odd and even transit sequences are estimated separately in the fitter. The null hypothesis of the Odd/Even Transit Depth Test is that the estimated transit depths of odd/even transit fits of the TCE are consistent, and the null hypothesis of the Odd/Even Transit Epoch Test is that the difference of the epoch times of the odd/even transits is consistent with the average period of the transits. A small significance level leads to the rejection of the null hypothesis—i.e., the TCE is unlikely to be due to a planet.

Two planets of nearly equal size cannot be in stable orbits at the same period around a star (de Pater & Lissauer 2001). However, in an eclipsing binary system, two stars move in one orbit around a common center of gravity with a single orbital period. If the primary and secondary eclipses are identified as two TCEs, the observed periods will be the same. The *Orbital Period Test* is designed to distinguish between the flux time series of a star with two transiting planets and that of an eclipsing binary system with primary and secondary eclipses, reported as two separate TCEs. The null hypothesis is that the orbital periods of the two TCEs are consistent. A small significance level leads to the rejection of the null hypothesis—i.e., the two TCEs are unlikely to be primary and secondary eclipses of an eclipsing binary system.

In a general case, the consistency check of *N* independent measurements of a parameter, denoted as {x_i}, *i* =1, …, *N*, associated with uncertainties {σ(x_i)}, can be modeled as a statistical test with the null hypothesis: {x_i} are drawn from *N* independent Gaussian distributions with the same mean and standard deviations equal to {σ(x_i)}. The statistic and significance level is determined as

$$s = \left(x_1-\bar{x}\right)^2/\sigma^2\left(x_1\right)+\cdots+\left(x_N-\bar{x}\right)^2/\sigma^2\left(x_N\right), \tag{9-5}$$

and

$$p = \Pr\left\{\chi^2\left(N-1\right)>s\right\}, \tag{9-6}$$

where $\bar{x}$ is the weighted mean of the measurements, and the weight of x_i is inversely proportional to $\sigma^2(x_i)$. $\chi^2(N–1)$ denotes a Chi-squared distribution with *N–1* degrees of freedom and *Pr{·}* denotes “probability of.” A small significance level (typically less than 0.05) leads to the rejection of the null hypothesis, i.e., the measurements are inconsistent.

Figure 9.4 shows the normalized flux time series of two target stars for second quarter flight data. In Figure 9.4a, one TCE is reported by TPS, and transits of the TCE are labeled with dash-dot lines. The difference between the odd/even transit depths are much larger than the uncertainties, resulting in a large statistic (~1800) and a small significance level (~0) of the odd/even transit depth test. Therefore, the null hypothesis is rejected with high confidence—i.e., the light curve shown in Figure 9.4a is unlikely to be due to a planet. In Figure 9.4b, two TCEs are reported by TPS, and the transits of the first and second TCEs are labeled with dash-dot lines and dash lines, respectively. The calculated statistic (~0) and significance level (~1) of the orbital period test can be verified using the following observation: the estimated orbital periods of the two TCEs are almost equal, suggesting that the two TCEs are due to primary and secondary eclipses of an eclipsing binary system.

9.4.3 Bootstrap test

9.4.3.1 Overview

In the search for transiting planets, the multiple-event statistics of a planet candidate with SNR=8 (typical for four transits of an Earth-size planet orbiting a 12th magnitude Sun-like star) are represented by the alternative hypothesis, H1, as depicted in Figure 9.5a. If all transit-like features are removed from the flux time series, a subsequent search for transits will generate the null multiple-event statistics as depicted by H0. The DV Bootstrap Test seeks to evaluate the likelihood that a TCE produced under H1 could alternatively be generated from the null-event statistics by chance alone, i.e., it seeks to evaluate the cumulative sum of the probabilities from the detection statistic that triggered the event to the end of the tail in H0. DV Bootstrap first constructs a histogram of the tail end of the null multiple-event statistics starting from the search transit threshold, η. From this histogram, it obtains the probabilities at each detection statistic, then computes the cumulative sum of the probabilities to obtain the complementary cumulative distribution function (CCDF). The false alarm of a planet candidate is evaluated from the CCDF at the TCE detection statistic by either interpolating or extrapolating.

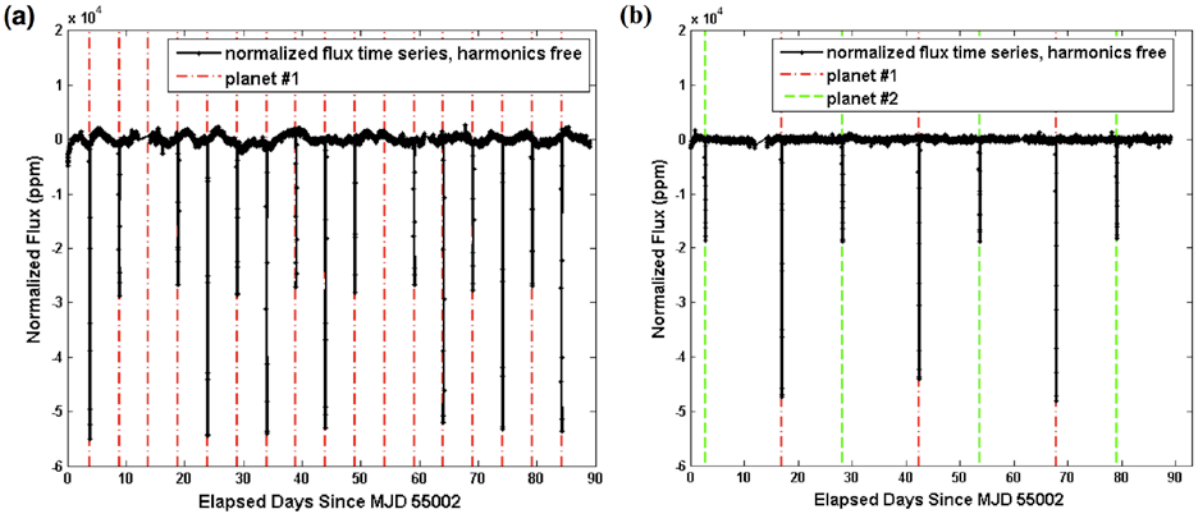


Figure 9.4. Normalized flux time series of target stars for second quarter flight data. (a) one TCE, (b) two TCEs.

To construct H0 in the traditional way for a search consisting of T transits, we need to form the multiple-event statistics for all combinations of N single-event statistics, i.e., H0 consists of N^T multiple-event statistics with all transits removed. For example, in a flux time series of 4500 cadences (~1 quarter) of data with five transits, there are $4500^5 = 2 \times 10^{18}$ null multiple-event statistics that can be formed; for 4500 cadences of data with six transits, there are $4500^6 = 8 \times 10^{21}$ null multiple-event statistics. The computational burden scales by the number of single event statistics to the power of number of transits, and generating the null distribution in this manner is computationally prohibitive. The solution to this problem lies in realizing that to compute the false alarm probability of a planet candidate, only the tail of the distribution of H0 above η is of interest.

In statistics, *bootstrapping* (Good 2005) is a method for estimating the sampling distribution of an estimator by repeatedly sampling, with replacement, from the original sample. We take a “modified” bootstrap approach by realizing that we are only interested in the upper tail portion of H0, and that the number of ways that a multiple-event statistic can be formed from the single-event statistic is known (Jenkins 2002). In this “modified” bootstrap approach, a counter, representing the indices that form the multiple event statistics, is used to index, obtain combinations, and update a histogram for the construction of H0.

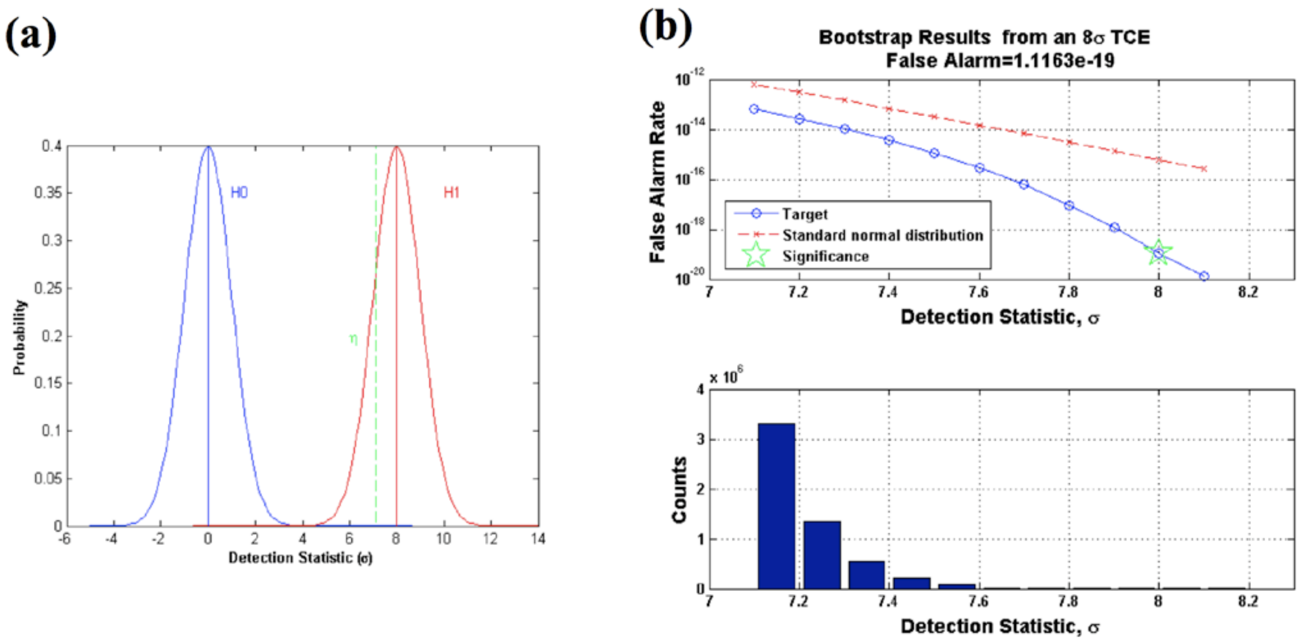


Figure 9.5. (a) Probability distribution for null detection statistics (H0), and for detection statistics of a transiting planet with SNR = 8 (H1). Search transit threshold is designated by η and is set at 7.1σ. (b) Bottom: tail end of the histogram formed following the bootstrap algorithm. Top: cumulative sum of the probabilities (derived from the histogram below) from upper tail to η; false alarm probability is indicated by the star.

9.4.3.2 Algorithm

In the case of a flux time series containing four transits for a duration of four years (~72000 cadences) with all transits removed, the algorithm is described as follows. First, sort the transit-free single-event statistics in descending order, preserving the "numerator" and "denominator" (Jenkins 2002) so that the multiple-event statistics can be formed. Set up a histogram by choosing a bin width of 0.1σ in which the minimum is at or below η and the maximum bin is the highest null multiple event statistic, rounded up to the nearest bin. Begin with a counter set at [1, 1, 1, 1], form the multiple event statistic from the associated single event statistics and compute the number of combinations for these digits (1). Add 1 count (number of combinations for [1, 1, 1, 1]) to the histogram bin corresponding to this detection statistic. Increment the counter to [1, 1, 1, 2], form the multiple event statistic, compute the number of combinations (4) and update the histogram with 4 counts in the bin with the corresponding detection statistic. This process is repeated many times and the digits in the counter are increased monotonically. If the formed multiple event statistic falls below η, the adjacent digit is incremented to 2 so that the counter reads [1, 1, 2, 2]. The procedure stops when successive multiple event statistics formed fall below η or when the counter reads [72000, 72000, 72000, 72000]. If there are multiple searches with different trial transit pulse widths, as is the case in DV (3, 6, 12 hour searches), then a histogram is formed for each. To obtain the probabilities, the histogram counts for each trial transit pulse are summed and divided by the total number of events and the false alarm rate is calculated as the cumulative sum of the probabilities from left to right. Finally, a logarithmic robust linear fit is performed on the curve, and the false alarm probability for the TCE is evaluated.

The procedure described above is still computationally intensive. To ameliorate this, we have implemented a skip count feature for targets with many transits. If the detection statistic formed is above η, the counter is incremented by a fixed deviate. The minimum histogram bin is chosen to be conservatively below η to account for inaccuracies when skip counts are implemented. After histograms have been generated, their counts are scaled by 1 + skip count.

9.4.3.3 Example

We apply the procedure above using a series of 72000 normally distributed random numbers to simulate a transit-free single event statistic time series over the course of a four-year mission. We assume that the TCE was triggered from an 8σ event. We then evaluate the likelihood that the TCE was caused by chance alone via our bootstrap method. Figure 9.5b illustrates the bootstrap results: tail end of the null distribution is represented by the histogram in the lower panel, its cumulative sum is shown by the circles on the top panel. We interpolate and compute a false alarm probability of 1.1×10^{-19} as indicated by the star, or ~1 in 9×10^{18} that this observation was produced by chance alone.

9.4.3.4 Limitations

In certain cases, the bootstrap algorithm cannot be used, e.g., hot Jupiters with periods on the order of a few days that generate 20 or more transits per quarter; these cases cannot be bootstrapped because calculating the combinations of the digits in the counter depends on calculating and/or representing factorial of the number of transits. For most computers, 20 is the maximum factorial that can be calculated accurately. To prevent halting downstream processes in DV, we have implemented a limit to the number of iterations that bootstrap will run before aborting. Perhaps the biggest limitation to the bootstrap test is the assumption that all transit signatures have been removed and the null single event statistic is composed purely of noise. The DV Bootstrap Test is most useful with low numbers of transits that result in small changes in transit depths that trigger TCEs in the vicinity of η —i.e., TCEs that suggest Earth analogues. Hot Jupiter-like planets that exhibit deep transits, with periods on the order of days that trigger TCEs much greater than η (e.g., >1,000σ), will yield false alarm values of nearly zero if bootstrapped. In this respect, the bootstrap test is especially effective at flagging TCEs triggered by Earth-like planets where analyses have not accounted for all transit-like features, which results in non-Gaussian statistics and higher false alarm probabilities.

9.5 Summary

We have presented a suite of statistical validation tests performed in DV, which consists of a test to assess correlations between centroid shifts and transit signatures, eclipsing binary discrimination tests, and a false alarm bootstrap test. DV test performance was evaluated using a set of 70 simulated (Bryson et al. 2010b) targets in which the ground truth was known. These targets consisted of synthetic Earths, Jupiters, eclipsing binary systems, background eclipsing binary systems, and a combination of these.

10 DATA VALIDATION FITTER

10.1 Introduction

The Data Validation (DV) module tests described in section 9 require a model of the planetary system that is consistent with the observed transit periods, durations, and depths. Such models are produced by a *planet-model fitter,* which is incorporated into DV. The fitter algorithm is described here.

10.2 Pre-fitting Analysis

The starting point for the DV fitter is the output from the TPS module. For each target star in the data set, TPS produces a Threshold Crossing Event data structure. Each TCE specifies the period, center time of first transit (also known as the "transit epoch"), and approximate duration of the set of features in the flux time series that most strongly resemble a transiting planet. While the depth of the transit is not included directly in the TCE, the strength of the transit signal is included, in units of multiples of the noise limit for the detection of a signal with the specified period and duration. This allows the DV fitter to produce an initial estimate of the transit model that seeds the fit, as described in § 10.3.1.

Prior to generation of the initial transit model estimate, a number of pre-fitting steps are taken to improve the accuracy of the parameters provided in the initial TCE. These steps also allow some poorly formed transiting planet candidates to be identified and rejected prior to fitting.

10.2.1 Transiting Planet Search

The TPS metric for the strength of a transit signature is Multiple Event Statistic (MES): this is the ratio of the detected signature's strength to the noise limit for a transit signature of the selected period and duration, or equivalently the signal-to-noise ratio for the detection of the series of transits. For a given light curve, TPS analyzes a set of user-specified transit durations, and reports TCEs for the period and epoch that result in the maximum MES for each selected transit duration. All light curves that have a TCE with MES over a specified threshold are then passed to DV for additional analysis. In the current configuration, TPS searches for 3-, 6-, and 12-hour transit durations, and TCEs with a MES of 7.1σ or greater are analyzed in DV.

The nature of the TPS algorithm is such that it does not attempt to ensure that all transits associated with a given MES are of comparable depth; as a result, a light curve with one or two very deep transit-like features is likely to be flagged as having a MES that is larger than the threshold. In Figure 10.1, we see examples of two classes of TCEs. In the top plot is the flux time series for a true transiting planet candidate, with transit-like features of comparable depth. In the bottom plot is a flux time series containing a single very large feature and a bump that is slightly above the noise floor; the TPS algorithm identifies this flux time series as containing a transit signature with a MES of 14.7σ, despite the fact that it is actually just the artifact that contains all the meaningful contributions to the MES.

The DV fitter performs an additional screening on the TCE obtained from TPS to identify cases resembling the bottom plot in Figure 10.1. This is accomplished by examining the Single Event Statistic (SES) that contribute to the MES, where each transit's SES is the SNR for detection of that transit (similar to the way that the MES is the SNR for detection of the series of transits). DV computes the ratio between the MES and the largest SES that contributes to the MES; if this ratio is smaller than a specified ratio, the event is considered too poor a candidate to fit, and DV moves on to its next target star.

In the limit of Gaussian noise and an infinite number of samples per single transit-like event, the relationship between the MES and the SES for a true transit signature is MES = $\sqrt{N_{\text{event}}}$ SES. Since there are finite samples per transit and the noise in the light curves is non-Gaussian and non-stationary, the SES in a given MES have a distribution of values, and the largest SES value can be significantly larger than the typical value; however, it is necessary to consider the largest SES value rather than, for example, the median SES value in order to identify cases such as the one shown in the lower half of Figure 10.1. Given all these issues, the current cutoff for further analysis is a MES value of at least 1.25 times the largest SES value. With the cutoff thus configured, about 50% of all TCEs passed to DV from TPS are rejected. The ensemble of rejected targets is overwhelmingly dominated by flux time series that contain clearly visible artifacts that are driving the detection process, and contains relatively few flux time series that appear to the eye to contain potential transiting planet signatures; in short, the applied MES/SES cut mainly eliminates uninteresting targets while also eliminating a tolerably small number of interesting targets.

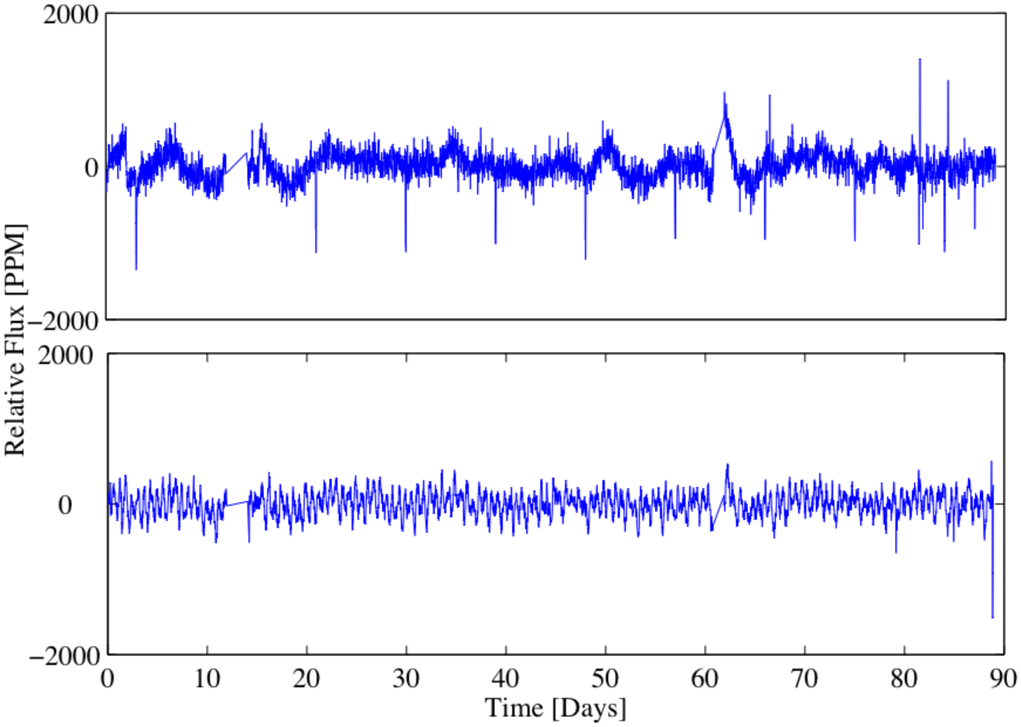


Figure 10.1. Two examples of light curves which produce threshold crossing events. Top: a planet candidate, with multiple transit-like features of comparable depth. Bottom: a spurious candidate, with a single large feature at the end of the light curve and a small bump at 79 days that are combined and interpreted as a threshold crossing event by TPS.

10.2.2 Transit Timing Estimate Improvement

A problem similar to the spurious TCE in 10.2.1 can cause the estimated period from TPS to be a harmonic of the actual period. Consider a flux time series as shown in the top portion of Figure 10.2: a fluctuation that occurs close to the midpoint between two actual transits can cause the MES of the TCE with half the actual period to be slightly larger than the MES of the TCE that has the correct period. In this case, the TCE with half the correct period will be sent to DV, resulting in an initial estimate of the period which is far from correct. This in turn yields an extremely poor seed from which the fitting process is unlikely to be able to recover.

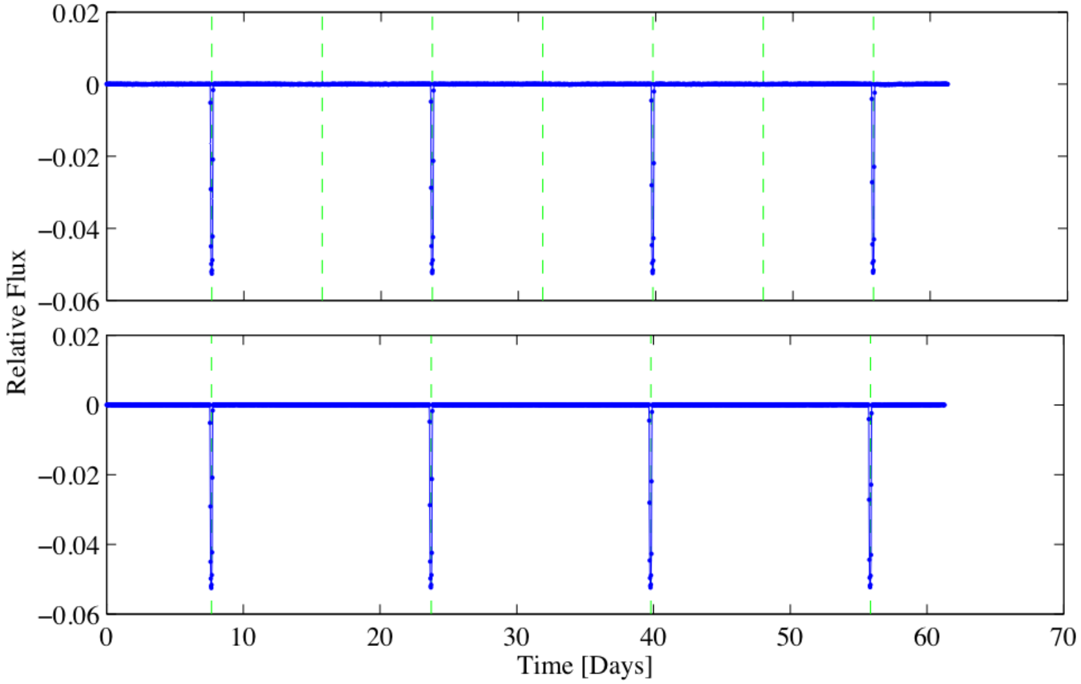


Figure 10.2. Example of a flux time series that requires improvement of its timing estimate. Top: flux time series with an incorrect period estimate. Bottom: the same flux time series after timing estimate improvement. In both cases, the predicted transit timings are indicated by vertical green dashed lines.

Upon deeper examination, the incorrect period selection is almost always caused by an unusually shaped distribution of MES versus orbital period: the distribution shows a peak at both the incorrect and the correct period, but while the central value of the peak at the correct period is higher than the central value of the peak at the incorrect one, the peak associated with the incorrect period will have an off-center “bump” with an even higher value; it is this “bump” that is detected by TPS and reported as the MES and orbital period for the system. A TCE with the correct period is obtained by performing an additional set of searches in TPS. Instead of performing a single search across the full range of possible periods, each additional search covers a much-reduced range of periods; the searches cover the period in the original TCE plus the first few subharmonics of that period. In each of these searches, the MES reported is the value at the center of the detected peak, rather than the absolute maximum value; this allows the refined search to ignore an outlier MES and correctly determine the orbital period. Since the original development of TPS and DV, TPS has been revised to always use the peak-center value rather than the absolute maximum value of MES. This has dramatically improved the ability of TPS to return a correct orbital period without any additional search. At this time, the first four subharmonics are searched, each across a window of ±2 days. Thus, an initial TCE with a period of 10 days would trigger additional searches that examine periods of 8-12 days, 18-22 days, 28-32 days, and 38-42 days. This yields a total of four new TCEs, one from each subharmonic. The new TCE with the maximum MES is then accepted as the correct one and used to seed the fit. The bottom half of Figure 10.2 shows the results of such a search: in this case, the second subharmonic has been correctly identified as the optimal one for the search.

In the event of an eclipsing binary in which the intervals between eclipses are close to rational (for example, 2:1 or 3:1), the procedure outlined above will still not converge upon the correct period. For example, in a 2:1 case, the third harmonic of the correct period will have the maximum MES because it overlaps all of the primary eclipses of the system, plus all of the secondary eclipses of the system, and also a set of times in which there is no eclipse at all; the correct period will overlap only the primary eclipses. In order to address this issue, the TCEs from the subharmonic searches are examined for evidence of extremely large transits at the locations predicted by the TCEs. For example, in the case of an eclipsing binary with 2:1 intervals, the transits predicted by the first and second subharmonic TCEs do not all line up with large transit events, while the transits predicted by the third subharmonic TCE will all line up with large transit events. This difference allows the correct period to be deduced even in such cases.

10.2.3 Eclipsing Binary Removal

In general, the DV fitter is capable of fitting a light curve in which the transit-like features are due to an eclipsing binary. In the specific case of eclipsing binaries with extremely deep transits, the fitter is generally unable to converge correctly. This issue is addressed by logic that prevents the fitter from operating in any case in which all of the transits predicted by the TCE are deeper than a threshold, which is currently set to 15%. In the event of such an eclipsing binary system, the transit parameters, epoch, period, and depth, are recorded for later use in DV's binary discrimination tests (§ 9.4.2), the transits themselves are marked as gapped, and the remaining data in the flux time series are sent back to TPS to be searched for additional planet candidates.

10.3 Iterative Whitening and Fitting

Once the pre-fitting processes outlined in above are carried out, the flux time series is fitted using a robust Levenberg-Marquardt fitter (Levenberg 1944; Marquardt 1963), which is a modified version of the MATLAB function *nlinfit* (Mathworks 2007). In order to do this, the contribution of slow stellar variability must be removed, since it can be quite large compared to the transit features: for example, transits of the Earth across the sun result in a peak flux reduction (transit depth) of about 100 parts per million; solar variability over a one-day period is approximately 10 parts per million, but over a one-year period is closer to 1000 parts per million. A wavelet-based whitening filter is used to remove the variations in the flux time series occurring over timescales that are long compared to a transit. A consequence of this process is that the shape of the transit is distorted by the filter; it is therefore necessary to apply the same filter to the model flux time series used in the fit, such that the model transits and the data are properly matched to one another in shape.

The wavelet-based whitening filter is determined individually for each fit and optimized to the noise spectrum of the given target star. Given that the purpose of the whitening filter is to remove slow variations in the star's flux time series, it is beneficial to generate the whitening filter using a time series that contains only the slow variations and has been stripped of the transits that are to be fitted. This is accomplished by first subtracting the current best-estimate transit light curve from the flux time series and whitening the residual; a whitened version of the transit light curve is then added back to the whitened residual flux time series, and this whitened total flux time series is used as the data that constrains the fit. In this case the whitening process depends on the current estimated transit model, but the fitted transit model depends on the whitening process. As a result, it is necessary to iteratively perform the whitening and Levenberg-Marquardt fitting until a self-consistent combination of whitener and transit model is obtained.

10.3.1 Initial Estimate of Transit Parameters

As described in § 10.2, the initial model of the transit is derived from the TCE, which in turn is furnished by the TPS software module. The TCE contains the epoch and period of the maximum multiple event statistic, as well as the values of the multiple event statistic (MES) and the maximum single event statistic (SES). It also returns the transit duration that was used in the search: the current configuration uses 3-, 6-, and 12-hour transits to produce TCEs and returns the TCE with the largest MES. Prior to the start of fitting, the parameters in the TCE must be converted to physical parameters for a transiting-planet solar system; these parameters are then used to seed the fit.

The conversion is simplified by assuming a circular orbit and a central transit, and by using the star radius parameter for the target star given in the *Kepler Input Catalog* (KIC). The remaining parameters are the epoch, semi-major axis, and planet radius. The epoch is obtained directly from the TCE, and the semi-major axis can be obtained from the orbital period *T*, the star radius *R*, and the surface gravity *g* of the star from Kepler's Third Law:

$$a = \left(\frac{T^2 g R^2}{4\pi^2} \right)^{1/3} . \tag{10-1}$$

The surface gravity of the star, like the radius, is available from the KIC.

The ratio of the planet radius to the star radius is given by the product of the square root of the transit depth and a correction for limb-darkening, and therefore the planet radius can in principle be deduced from the transit depth, limb-darkening parameters, and star radius. Unfortunately the transit depth is not returned as part of the TCE, but the single-event statistic can be converted to an estimate of the transit depth:

$$D \approx \mathrm{SES} \cdot \sigma_1 \cdot \sqrt{N_{\mathrm{meas}}} , \tag{10-2}$$

where σ_1 is the typical relative noise in a single 30-minute measurement of the flux of the target star (typically in PPM) and N_{meas} is the number of 30-minute measurements in the transit duration used in the search (i.e., $N_{\mathrm{meas}} = 12$ for a 6-hour transit duration).

The depth estimate that is thus obtained from the SES is only approximately correct, but it is sufficiently accurate to use as the starting point for the fit. A minor improvement in accuracy is obtained by requiring the depth to be the minimum of *D*, as estimated above, and the full range of variation of the flux time series.

The top panel of Figure 10.3 shows a flux time series containing a series of transits and stellar variability. The second panel shows a model series of transits that matches the actual transit depths in the flux time series. It is clear that in order to make the functional form of these two curves match well enough to use the former as constraints for fitting the latter, it will be necessary to remove the slow component of the stellar variability while leaving the transit-like features intact. Furthermore, given that stellar variability is a non-stationary process, the frequency content of the stellar variability is itself varying with time, and any attempt to filter the flux time series must take this into account.

Given the nature of this problem, the removal of a non-stationary, non-white noise component from a time series, a joint time-frequency representation of the noise and of the filter is indicated; for this reason, a wavelet-based whitening filter is used for removing the stellar variability of target stars (Jenkins 2002). In order to prevent the whitener from removing or degrading the transit-like features of the flux time series, the following procedure is followed:

- A *residual flux time series* is formed by subtraction of the model transit from the flux time series.
- The whitening filter for the fit is constructed from the residual flux time series.
- The whitening filter is applied separately to the residual flux time series and the model transit, and the sum of the whitened residual flux time series and the whitened transit model is used to constrain the fit.

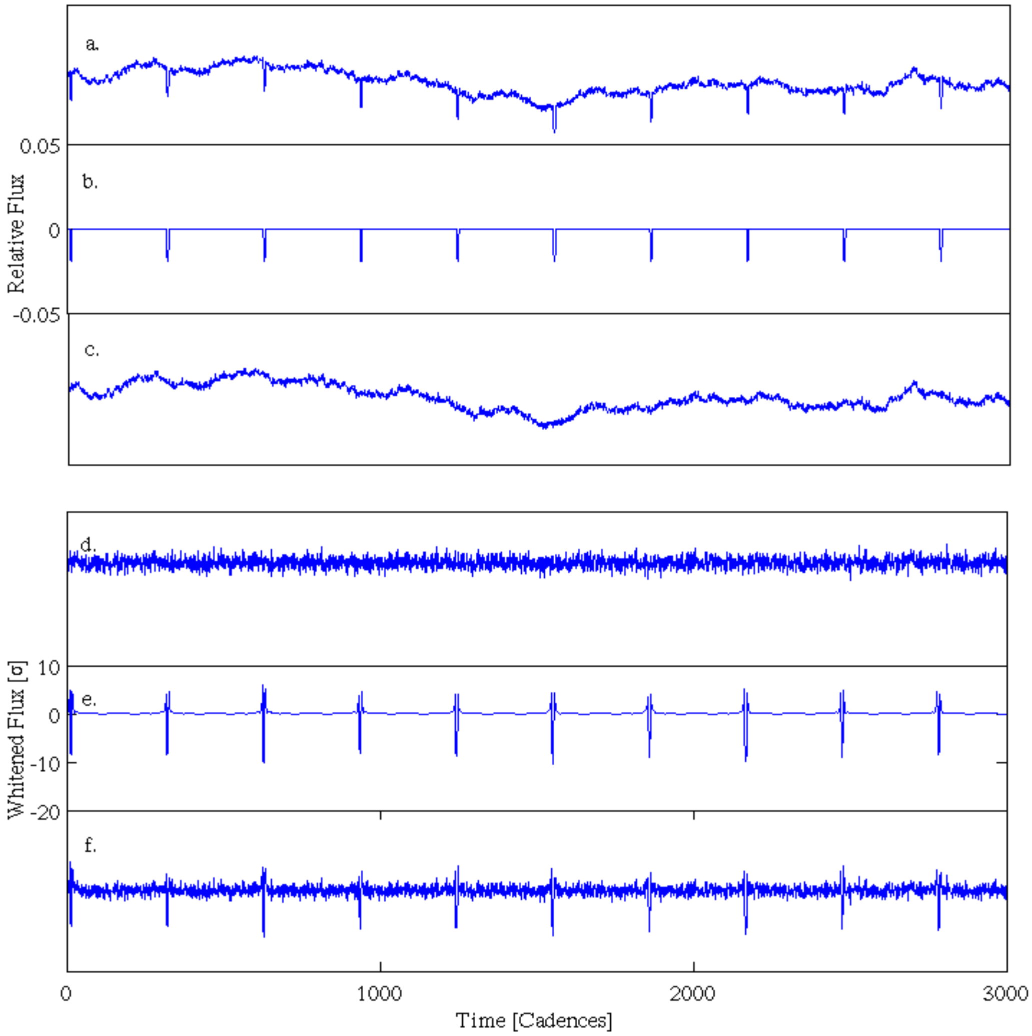


Figure 10.3. Whitening process for a flux time series: (a) Time series with transits and stellar variation; (b) Model transit flux time series; (c) Residual flux time series; (d) Whitened residual flux time series; (e) Whitened transit model; (f) Whitened flux time series, sum of (d) and (e). Vertical scale for (a-c) dimensionless relative flux, (d-f) is multiples of the standard deviation of (d).

The whitening filter that is generated from the residual flux time series is then used in the fit, as described in the next section. In Figure 10.3, the fourth panel shows the results of applying the whitening filter to the residual flux time series. As expected, the residual flux time series is converted to white noise with unit variance. The fifth panel shows the results when the whitening filter is applied to the model transit: the shape of the transit is distorted, such that the original transit (second panel), which was purely negative, now has positive "wings" on either side of the transit. The bottom panel shows the sum of the whitened residual flux and the whitened transit model. Note that the whitener also performs a scale transformation: in the whitened domain, the dimensions are multiples of the RMS of the white noise.

10.3.2 Levenberg-Marquardt Fit of Transit Model

The Levenberg-Marquardt fit of the transit model uses a modified version of the MATLAB *nlinfit* function. The *nlinfit* function supports robust fitting, in which an initial fit is performed and the data points are reweighted based on the magnitude of their residuals to the fit; the fit is repeated, new weights are applied to the data, and this process of reweighting and refitting is iterated until the change in fit parameters from one iteration to the next falls below a predetermined threshold. The key modification to *nlinfit* for use in the DV fitter is a less strict convergence criterion: while the unmodified *nlinfit* continues to iterate until all parameters are stable to within 1.5 x 10^{-8} of their values, the modified *nlinfit* continues to iterate until all parameters are stable to within 0.5 of their estimated uncertainties. The nominal converge criterion requires a much larger number of iterations than the modified convergence criterion, and represents a case of severely diminishing returns given that the changes in parameters for the nominal convergence criterion are usually extremely small compared to the estimated parameter uncertainties.

The fit is constrained by the whitened flux time series described in the previous section. For each new set of fit parameters, an unwhitened transit model is generated and then passed through the current whitening filter; the difference between the whitened transit model and the whitened flux time series is the quantity that is minimized by the fit. The initial fit is weighted by the estimated uncertainties in the flux time series, rescaled according to the scale factor between the unwhitened and whitened domains. In the robust fitting stages, weights are a product of the rescaled uncertainties and the robust weights.

10.3.2.1 Parameterization of the Transit Model

For a transiting planet, there are seven physical parameters required to describe the system: transit epoch, star radius, planet radius, semi-major axis, impact parameter, eccentricity, and longitude of periastron. The *Kepler* data lack the time resolution required to directly determine the eccentricity, and thus all orbits are modeled as circular. This leaves five free physical parameters that can be freely converted to equivalent observable parameters: epoch, orbital period, transit depth, transit duration, and transit ingress time. The ensemble of physical parameters has the advantage that any combination of valid physical parameters is itself a valid description of a transiting planet system; by contrast, there are combinations of observable parameters which are not valid as an ensemble (a trivial example of this is a system in which the transit duration exceeds the orbital period). The physical parameters are therefore the optimal ones to use for Levenberg-Marquardt fitting.

During testing and development of the DV fitter, two degeneracies in the parameters were uncovered. The first degeneracy is between the semi-major axis and the star radius: to lowest order, changing either the semi-major axis or the star radius primarily changes the observed orbital period of the system, so the fit is unable to distinguish between the two parameters. This was solved by replacing the star radius fit parameter with an orbital period fit parameter; for each new combination of parameters requested by the Levenberg-Marquardt process, the model function would use Kepler's Third Law to compute the implied star radius from the period, the semi-major axis, and the catalog value of the surface gravity, via a rearrangement of Equation (10-1); in this way, the transit model still uses a purely physical set of parameters even though the Levenberg-Marquardt process uses one observable parameter.

The second degeneracy is related to the impact parameter. Consider a central transit with a given period *T*, depth *D*, and duration *t*: one can formulate a transit with non-zero impact parameter *b* that has the same values of *T, D,* and *t*, but which is distinguishable from the central transit by the shape of its ingress and egress regions (essentially, the non-central transit has a longer ingress and egress time). *Kepler's* time resolution is 30 minutes and typical ingress/egress times for transiting planets in the HZ are under an hour, thus the ingress and egress times are only a weak constraint on the fit. The impact parameter is therefore only weakly constrained, and since the impact parameter is strongly covariant with the other physical parameters, inclusion of the impact parameter in the fit causes poor performance overall. To combat this, the fit is initially performed with the impact parameter held constant at zero. Once the fit has converged, the star radius in the fitted model is compared to the star radius in the KIC. A radius that is smaller than the KIC radius indicates that the data can be fitted with a model that holds the star radius constant at the KIC value and fits the impact parameter. In such cases the fit is repeated with the epoch, planet radius, semi-major axis and impact parameter used as fit parameters and the star radius held constant at the KIC value.

An additional subtlety to the parameterization is that the impact parameter is constrained to lie in the range [0, 1], but the Levenberg-Marquardt algorithm implicitly requires all fit parameters to be valid over all real values. To address this mismatch, a nonlinear transformation is performed between the "internal parameter" used by Levenberg-Marquardt and the "external parameter" used in the transit model (James 1994). This transformation maps the range [–∞,+∞] used by Levenberg-Marquardt to the range–[1, 1] in the transit model; the transit model then treats negative impact parameters as identical to their absolute values.

The units of the fitted parameters are as follows: transit epoch in barycentric-corrected Modified Julian Date, planet radius in Earth radii, semi-major axis in Astronomical Units, and period in days. The impact parameter is dimensionless.

10.3.2.2 Parameter Step Sizes in Levenberg-Marquardt Fitter

The *nlinfit* function performs a finite-difference calculation on each of the fit parameters to determine its Jacobian. For the purposes of fitting the light curve of a transiting planet system, the main constraint on the finite-difference calculation is that the step size should be small enough that the model transits do not move by a significant fraction of their duration. This is important because the transits occupy only a small fraction of the total light curve: therefore, if the Jabobian calculation is allowed to "jump" a model transit by an interval that is comparable to the transit duration, the model can easily get into a state in which the transits in the model line up with the inter-transit intervals in the data; at this point, small changes in the transit timing have no impact on the goodness of fit, and the fitter becomes irretrievably lost.

The *nlinfit* function uses a default minimum step size of 5.05 x 10^{-6} of each parameter to compute its Jacobian matrix. For most of the parameters this is acceptable, but for epoch it is not. The typical epoch MJD values are around 55,000, so *nlinfit* will change the epoch by 0.33 days when computing the Jacobian. This is addressed by forcing *nlinfit* to use a minimum step size for the epoch that is 0.1 times the data sample duration, or about three minutes for standard *Kepler* long cadence data.

10.3.3 Convergence Criteria

As described above, the DV fitter must iterate the process of deriving a whitening filter and performing Levenberg- Marquardt fitting in the whitened domain; the iteration is necessary because the fit results depend upon the whitening filter, but the design of the whitening filter depends upon the subtraction of the transit signature from the flux time series (i.e., the whitener depends upon the fit results). Iterations of whitening filter design and model fitting cease when one of the following conditions occurs:

- The number of iterations of whitening filter design and model fitting reaches a user-selected limit (currently set to 100)
- The total time spent performing fits on the current target star reaches a user-selected limit (currently set to nine hours)
- The change in parameter values from the previous iteration to the current one is smaller than some user-selected fraction of the estimated parameter uncertainty (currently set to 0.01).

If the user has requested robust fitting, the iterative whitening-fitting process is first run to convergence without the application of robust weights. Once the non-robust fit has converged, the fitter begins a new series of whitening-fitting iterations that includes robust weights. Note that this fitting process can be extremely time-consuming: the fitter internally iterates the Levenberg-Marquardt fit with varying weights, and the whitening-fitting loop iterates the robust fit until full internal consistency is reached. For this reason, the robust fitting process is not performed until a non-robust version of the fit has converged. Finally, in the case in which the fitted star radius is smaller than the KIC value of the star radius, an additional set of iterations is performed in which the impact parameter is fitted and the star radius is held fixed at the KIC value. Again, the fit is allowed to converge in a non-robust manner, after which the robust fit is performed. Once the fit is complete, the final fit with fixed impact parameter is compared to the final fit with fitted impact parameter, and the fit with the lowest reduced χ^2 is returned as the best fit to the data.

For each ensemble of whitening-fitting processes (non-robust with fixed impact parameter, robust with fixed impact parameter, non-robust with fitted impact parameter, robust with fitted impact parameter), the fitter is permitted to execute up to 100 iterations of whitening and fitting. In an extreme case, the total number of iterations could reach 400. The amount of clock time allowed per target star in the fitter remains fixed at nine hours, regardless of whether robust fitting is selected or fitting with variable impact parameter is required.

Figure 10.4 shows the results of a transit model fit. The upper plot shows the whitened flux time series, the lower plot shows the same data folded at the fitted period and averaged into 30-minute-wide bins.

10.3.4 Fitting of Odd-Numbered and Even-Numbered Transits

Once the main fit has completed, the DV fitter performs separate fits of the odd-numbered and even-numbered transits in the flux time series. This additional step is executed to provide information for the eclipsing binary discrimination tests, which are described briefly below and in detail in section 9.4.2. The all-transits fit is used to seed the odd- and even-transits fits, and the undesired transits (even-numbered transits in the odd-transits fit and vice-versa) are removed by marking their entries in the flux time series as missing.

The odd- and even-transits fits generally proceed in the same manner as the all-transits fits, with two exceptions. First, the fit parameterization of the all-transits fit (fixed or fitted impact parameter) is used for the odd- and even-transits fit. Second, depending on the total number of transits in the light curve, there may be only one transit in either the even-transits fit or in both the odd- and even-transits fits. If this is the case, then the number of parameters in the fit must be reduced by one, since the orbital period is no longer available as a fit constraint; this is accomplished by fitting only the epoch, planet radius, and semi-major axis, while holding the orbital period and the impact parameter fixed at their values determined by the all-transits fit.

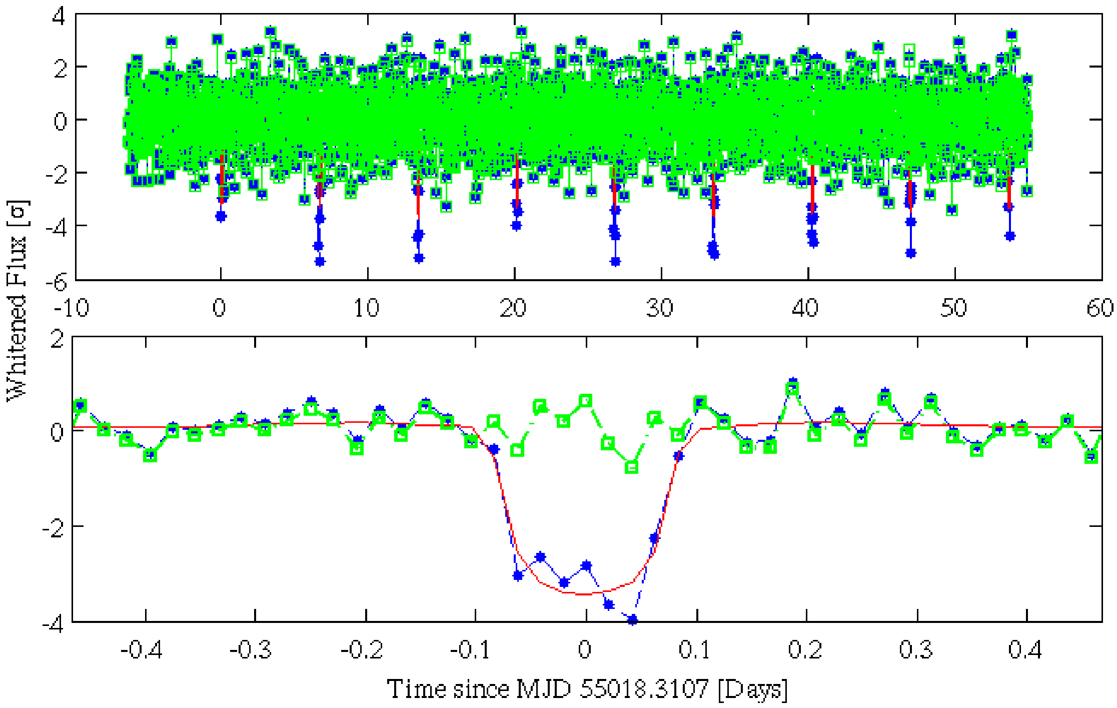


Figure 10.4. Sample results of the DV fitter. Top plot: whitened flux time series over the full duration of the data set. Bottom plot: whitened flux folded at the fitted period, binned and averaged to 30-minute intervals. In both plots, the original data (blue circles, dashed line), fitted light curve (red solid line), and residual time series (green squares, dot-dashed line data minus fit) are shown.

10.4 Multiple Planet Search

The TPS algorithm can only identify one planet candidate at a time, and that planet candidate will always be the one with the largest MES; this is generally the candidate with the deepest transits. In order to locate additional planet candidates in the flux time series of a given target star, it is first necessary to remove the transit signatures of the earlier, larger candidates. This is accomplished by using the fitted transit model from the previous planet candidate to identify the timestamps that correspond to transits in the flux time series; these timestamps are then marked as missing and the gapped flux time series is sent back to TPS to search for additional planet candidates. If additional TCEs that are above the detection threshold are found, the gapped flux time series and the new TCEs are sent back into the DV fitter.

For planet candidates with extremely deep transits, it is possible for the actual transits in the data to be of longer duration than the model transits in the fit. This is especially true if the quality of the fit is not extremely good; for example, when a planet model is fitted to an eclipsing binary. In such cases the transit model's identification of which data points are in transit is of limited accuracy, and data points on the outskirts of a model transit can be sufficiently darkened as to trigger a transiting planet detection. This undesirable outcome is prevented by marking a number of data points adjacent to each model transit as missing data. At this time, the fitter is configured to remove data points over a time span of three transit times at each transit location; in other words, for model transits with a 10-hour duration, the multiple planet search removes the 10 hours of each model transit, and also the preceding and following 10 hours of data.

There are two additional protections against repeated detections of the same transit signature in a light curve. First, the maximum number of planet fits for each target star is currently limited to four; once four planet candidates have been fitted, DV will proceed to the next target even if the multiple planet search detects a fifth planet candidate. Second, the maximum clock time that may be used in processing any given target star is limited to nine hours.

10.5 Applications of the Fitted Planet Models

The transit model fits from the DV fitter are used as inputs to a number of additional tests performed in DV. These tests are described in greater detail elsewhere and are only summarized here.

10.5.1 Centroid Motion Test

If the target star is actually a blend, then the photocenter will move during transits; this indicates that the transits might actually be caused by a background eclipsing binary. The centroid motion test uses the fitted transit model to determine the data timestamps that correspond to the maximum reduction in flux, and searches for a change in the photocenter location that occurs at these times. The statistical significance of the photocenter motion is then assessed.

10.5.2 Eclipsing Binary Tests

An eclipsing binary or background eclipsing binary can imperfectly mimic a transiting planet signature. The DV module performs a number of tests that can be used to discriminate between a planet and an eclipsing binary:

- For an eclipsing binary star with a circular orbit, the primary and secondary eclipses will be flagged by TPS as a single planet candidate. In this case, the odd-transits fit and even-transits fit will converge to different values of the transit depth. The depth test uses the fitted depths to determine the probability that the planet candidate is a circularized eclipsing binary.
- For an eclipsing binary star with a near-circular orbit, the interval from primary to secondary eclipse will be slightly different than the interval from secondary to primary. In this case, the fitted transit epoch of the odd-transit fit will not agree with the fitted epoch of the even-transit fit. The epoch test uses the fitted epochs to assess the probability that the planet candidate is actually a nearly-circularized eclipsing binary.
- For an eclipsing binary with an elliptical orbit, the primary and secondary eclipses will be identified as two distinct planets, but their fitted orbital periods will be identical. The period test compares the fitted periods of the all-transits fits of the planet candidates on a given target star to determine the probability that two planet candidates are actually the two

eclipses of an eclipsing binary. This test also uses the estimated periods of TCEs that have been rejected from fitting due to their depth, as described in section 10.2.3.

10.5.3 Bootstrap Analysis

At the conclusion of model fitting, it is possible to identify and remove all data points that occurred in or near a model transit. The remaining data points contain only the stellar variation and instrument noise contributions to the flux time series. These data points are therefore ideal for performing an after-the-fact bootstrap analysis of the fitted transits, which allows a more accurate estimate of the probability that each TCE was a result of a statistical fluctuation rather than a true astrophysical signature.

10.6 Performance of the Planet Fitter

The DV planet fitter was validated in an exercise that used simulated data with known ground-truth parameters. The exercise included 70 targets with an assortment of single and multiple planet systems, eclipsing binaries, and background eclipsing binaries. Out of 72 simulated true planets in the ensemble, 53 were correctly identified as planets and fitted, while 19 planets were not identified (*false negatives*). Of these, nine planets were missed because the simulated planet was too small to produce a TCE above the detection threshold; the remaining 10 false negatives were due to a number of issues in the science processing Pipeline. The same exercise produced 12 false positives, in which non-planet signatures were mistakenly identified as planets. The vast majority of the false positives were caused by eclipsing binaries, which is an expected outcome of the DV fitter. These cases can be identified by the eclipsing binary discrimination tests in DV.

The DV planet fitter was also successfully exercised against a 90-day sample of flight data. Performance was generally good, although a few percent of the TCEs in the flight data could not be fitted successfully for reasons that are still under study.

The main fitter issue exposed by both tests was the execution time of the DV fitter. In order to perform all the fits required for the 90-day flight data set, a total of 98 DV processes running in parallel required over four days of clock time. At the time of this writing, this is the most time-consuming process in the *Kepler* Science Data Processing Pipeline (Middour et al. 2010). A number of worthwhile optimizations have been identified, and will be implemented in the near future.

10.7 Summary

The DV fitter is a tool that performs automated fitting of transiting planet models to flux time series for the *Kepler Mission.* It has been successfully tested against simulated and real *Kepler* flight data with generally good results. A number of areas of potential improvement have been exposed, most significantly in the realm of execution time. In the near future, we expect to integrate DV into the *Kepler* Science Data Processing Pipeline and to use its results to guide selection and prioritization of targets for follow-up observation.

11 REFERENCES

Reference Documents

Document Name	Release Date
Kepler Instrument Handbook	July 2009
Kepler Archive Manual	Version 1.0 - 31 Aug 2009
Data Release Notes	At each data delivery to MAST
Kepler Data Characteristics Handbook	February 2011

Publications

Akaike, H. 1974, “A New Look at the Statistical Model Identification”, *IEEE Transactions on Automatic Control*, AC-19 (6), 716

Allen, C., Klaus, T., & Jenkins, J. 2010, “Kepler Mission’s Focal Plane Characterization Models Implementation”, *Proceedings of the SPIE,* 7740, 77401E

Batalha, N.B., et al. 2010, “Selection, Prioritization, and Characteristics of Kepler Target Stars”, *Astrophysical Journal Letters,* 713, L109

Borucki, W., et al. 2010, “Kepler Planet-detection Mission: Introduction and First Results”, *Science* 327, 977

Borucki, W., et al. 2011a, “Characteristics of Kepler Planetary Candidates Based on the First Data Set”, *Astrophysical Journal*, 728, 117

Borucki, W., et al. 2011b, “Characteristics of Planetary Candidates Observed by Kepler, II: Analysis of the First Four Months of Data”, submitted to ApJ, arXiv:1102.0541

Brown, T.M., Latham, D.W., Everett, M., & Esquerdo, G.A. 2011, “Kepler Input Catalog: Photometric Calibration and Stellar Classification”, *Astronomical Journal*, submitted

Bryson, S.T., et al. 2010a, “Selecting Pixels for Kepler Downlink”, *Proceedings of the SPIE*, 7740, 77401D

Bryson, S.T., et al. 2010b, “The Kepler End-to-end Model: Creating High-fidelity Simulations to Test Kepler Ground Processing”, *Proceedings of the SPIE,* 7738, 773808

Bryson, S.T., et al. 2010c, “The Kepler Pixel Response Function”, *Astrophysical Journal Letters,* 713, L9

Caldwell, D.A., et al. 2010, “Instrument Performance in Kepler’s First Months”, *Astrophysical Journal Letters*, 713, L92

Chandrasekaran, H., et al. 2010, “Semi-weekly Monitoring of the Performance and Attitude of Kepler Using a Sparse Set of Targets”, *Proceedings of the SPIE*, 7740, 77401B

Claret, A. 2000, “Non-linear Limb-darkening Law for LTE Models”, *Astronomy & Astrophysics,* 363, 1081

Clarke, B.D., et al. 2010, “A Framework for Propagation of Uncertainties in the Kepler Data Analysis Pipeline”, *Proceedings of the SPIE*, 7740, 774020

Debauchies, I. 1988, “Orthonormal Bases of Compactly Supported Wavelets”, *Communications on Pure & Appied Math,* 41, 909

De Pater, I., & Lissauer, J.J. 2001, in *Planetary Sciences*, Cambridge University Press, (Cambridge, UK), p. 24

Gautier, T.N., et al. 2010, “The Kepler Follow-up Observation Program”, *arXiv:1001.0352*

Gilliland, R.L., et al. 2010, “Initial Characteristics of Kepler Short Cadence Data.” *Astrophysical Journal Letters,* 713, L160

Good, P.I. 2005, in *Resampling Methods: a Practical Guide to Data Analysis*, Birkhauser, (Boston, USA), p. 8

Haas, M.R., et al. 2010, “Kepler Science Operations”, *Astrophysical Journal Letters* 713, L115

Hall, J.R., et al. 2010, “Kepler Science Operations Processes, Procedures, and Tools”, *Proceedings of the SPIE,* 7737, 77370H

James, F. 1994, *MINUIT Function Minimization and Error Analysis Reference Manual*, CERN, (Geneva, Switzerland)

Jenkins, J.M. 2002, “The Impact of Solar-like Variability on the Detectability of Transiting Terrestrial Planets”, *Astrophysical Journal Letters,* 575, L493

Jenkins, J.M., & Doyle, L.R. 2003, “Detecting Reflected Light From Close-in Extrasolar Giant Planets with the Kepler Photometer”, *Astrophysical Journal Letters,* 595, 429

Jenkins, J.M., Doyle, L. R., & Cullers, K. 1996, “A Matched-filter Method for Ground-based Sub-noise Detection of Extrasolar Planets in Eclipsing Binaries: Application to CM Draconis”, *Icarus,* 119, 244

Jenkins, J.M., Peters, D.J., & Murphy, D.W. 2004, “An Efficient End-to-end Model for the Kepler Photometer”, *Proceedings of the SPIE*, 5497, 202

Jenkins, J.M., et al. 2010a, “Overview of the Kepler Science Processing Pipeline”, *Astrophysical Journal Letters,* 713, L87

Jenkins, J.M., et al. 2010b, “Initial Characteristics of Kepler Long Cadence Data for Detecting Transiting Planets”, *Astrophysical Journal Letters,* 713, L120

Jenkins, J. M., et al. 2010c, “Transiting Planet Search in the Kepler Pipeline”, *Proceedings of the SPIE,* 7740, 77400D

Kay, S. 1999, “Adaptive Detection for Unknown Noise Power Spectral Densities”, *IEEE Transactions on Signal Processing,* 47, 10

Klaus, T.C., et al. 2010a, “The Kepler Science Operations Center Pipeline Framework”, *Proceedings of the SPIE,* 7740, 774017

Klaus, T.C., et al. 2010b, “The Kepler Science Operations Center Pipeline Framework Extensions”, *Proceedings of the SPIE*, 7740, 774018

Koch, D.G., et al. 2010, “Kepler Mission Design, Realized Photometric Performance, and Early Science”, *Astrophysical Journal Letters,* 713, L79

Latham, D.W., Brown, T.M., Monet, D.G., Everett, M., Esquerdo, G.A., & Hergenrother, C.W. 2005, “The Kepler Input Catalog”, *Bulletin of the American Astronomical Society*, 37, 1340

Levenberg, K. 1944, “A Method for the Solution of Certain Non-linear Problems in Least Squares”, *Q Applied Math* 2, 164

Li, J., et al. 2010, “Photometer Performance Assessment in Kepler Science Data Processing”, *Proceedings of the SPIE*, 7740, 77041T

Lissauer, J.J., et al., 2011, “A Closely-packed System of Low-mass, Low-Density Planets Transiting Kepler-11”, *Nature,* 470, 53

Mandel, K., & Agol, E. 2002, “Analytic Lightcurves for Planetary Transit Searches”, *Astrophysical Journal,* 580, L171

Marquardt, D. 1963, “An Algorithm for Least-squares Estimation of Nonlinear Parameters”, *SIAM Journal on Applied Math,* 11, 431

Mathworks. 2007, MATLAB r2007a release.

McCauliff, S., et al. 2010, “The Kepler DB: A Database Management System for Arrays, Sparse Arrays, and Binary Objects”, *Proceedings of the SPIE,* 7740, 77400M

Middour, C.K., et al. 2010, “Kepler Science Operations Center Architecture”, *Proceedings of the SPIE,* 7740, 77401A

Philbrick, R.H. 2009, “Correction of Artifacts in Correlated Double-sampled CCD Video Resulting From Insufficient Bandwidth”, *Proceedings of the SPIE,* 7244, 72440M

Quintana, E.V., et al. 2010, “Pixel-level Calibration in the Kepler Science Operations Center Pipeline”, *Proceedings of the SPIE*, 7740, 77401X

Tenenbaum, P. and Jenkins, J.M. 2010a, “Focal Plane Geometry Characterization of the Kepler Mission”, *Proceedings of the SPIE*, 7740, 77401C

Tenenbaum, P., et al. 2010b, “An Algorithm for Fitting of Planet Models to Kepler Light Curves.” *Proceedings of the SPIE*, 7740, 77400J

Twicken, J. D., et al. 2010a, “Photometric Analysis in the Kepler Science Operations Center Pipeline”, *Proceedings of the SPIE*, 7740, 774023

Twicken, J. D., et al. 2010b, “Presearch Data Conditioning in the Kepler Science Operations Center Pipeline”, *Proceedings of the SPIE,* 7740, 77401U

Van Cleve, J. & Caldwell, D. A. 2009, *The Kepler Instrument Handbook,* KSCI 19033-001 (Moffett Field, CA: NASA Ames Research Center)

Witteborn, F.C., Van Cleve, J., Borucki, W., Argabright, V., & Hascall, P. 2011, in preparation

Wu, H. et al. 2010, “Data Validation in the Kepler Science Operations Center Pipeline”, *Proceedings of the SPIE*, 7740, 774019

12 ACRONYMS AND ABBREVIATION LIST

ADC	Analog-to-Digital Converter
ADU	Analog-to-Digital Unit
CAL	Calibration Pipeline Module
CCD	Charge Coupled Device
CDPP	Combined Differential Photometric Precision
CfA	Center for Astrophysics
CSCI	Computer Software Configuration Items
CTE	Charge Transfer Efficiency
DAWG	Data Analysis Working Group
DV	Data Validation Pipeline Module
DVA	Differential Velocity Aberration
FFI	Full Frame Image
FGS	Fine Guidance Sensor
FOP	Follow-up Observation Program
FOV	Field of View
GO	Guest Observer
KADN	Kepler Ames Design Notes
KASC	Kepler AsteroSeismology Consortium
KDPH	Kepler Data Processing Handbook
KIC	Kepler Input Catalog
KIH	Kepler Instrument Handbook
LASP	Laboratory for Atmospheric and Space Physics
LC	Long Cadence
LCO	Las Cumbres Observatory
LDE	Local Detector Electronics
MAST	Multi-Mission Archive at Space Telescope
MJD	Modified Julian Date
PA	Photometric Analysis Pipeline Module
PDC	Pre-Search Data Conditioning Pipeline Module
PPA	Photometer Performance Assessment Pipeline Module
ppm	Parts-per-million
PRF	Pixel Response Function
PSD	Power Spectral Density
RF	Reference Pixel
SC	Short Cadence
SDSS	Sloan Digital Sky Survey
SES	Single Event Statistic
SNR	Signal-to-Noise Ratio
SO	Science Office
SOC	Science Operations Center
SOL	Start of Line
STScI	Space Telescope Science Institute
2MASS	2-Micron All Sky Survey
TCE	Threshold Crossing Events
TPS	Transiting Planet Search Pipeline Module
USNO	United States Naval Observatory