Algorithms and Data Structures for KIC Stellar Classification

Timothy M. Brown

Received ______________________; accepted ______________________

ABSTRACT

This document describes the algorithms and associated data sources used by the Stellar Classification Project (SCP) to classify stars and produce the outputs that appear in the KIC. I include descriptions of diagnostics that I applied to the KIC data to catch errors and control quality, including samples of most of the various kinds of plots made along the way. In Appendix 1 ("IDL Code"), I summarize the IDL routines that implement these algorithms, and the relationships between them. Appendix 4 ("Known Defects in KIC10") describes what is known to me as of 1 Oct 2008 concerning the failings of the KIC catalog, and ways in which it might tend to mislead the user.

1. Strategy

The processes and software described here represent an intermediate stage in the processing of Stellar Classification Project (SCP) data for the KIC. I received photometric catalogs from Mark Everett, each such file containing tabular information about an individual image taken at the telescope. I processed these data to obtain a list of detected stars in the Kepler field, and also to classify as many stars as possible in terms of their fundamental physical parameters. I then passed these reduced data to Dave Monet, who confederated them with information from other catalogs, and prepared the result for ingestion by the SOC.

Accordingly, the function of the software described here is fourfold. First, it ingests the raw photometric catalogs into a group of databases that allow convenient processing to obtain physical parameters of stars. Second, it estimates the atmospheric extinction suffered by each measurement, and corrects the instrumental stellar magnitudes to yield calibrated ones. Third, it combines the calibrated magnitudes with other information (stellar atmosphere models, for instance) to estimate the physical parameters of each observed star, including T_{eff}, $\log(g)$, $[Z]$, radius R, mass M, and interstellar extinction A_V. Fourth, it discards those stars (and putative stars) for which estimates are deemed particularly unreliable, and formats the remaining data for delivery to Dave Monet.

Data ingestion is a straightforward process, with its details defined almost entirely by the database structure that was defined at the outset, and that is described below. Complications in the database structure and related code arose almost entirely because of speed issues, which threatened to bog the system down as the volume of data became large.

Correction for atmospheric extinction is also simple in concept, depending mostly on the model adopted for extinction, and on the criteria for estimating the parameters in that model. Many of these decisions were guided by the choices made by the SDSS (Smith et al. 2002), since our filter set is similar to theirs.

Stellar parameter estimation was the most difficult piece of this part of the project. The underlying problem is that, for the purposes of the Kepler project, the most interesting parameter is the stellar radius R, which is to say $\log(g)$. But the (mostly) broadband filters we used provide data that are almost entirely insensitive to this parameter. Some gravity dependence is provided by the intermediate-band $D51$ filter, while for M-type stars, the $J-K$ color provides a gravity measure. Nevertheless, the photometric information that we could obtain on the timescale necessary for the project was barely sufficient for our needs. It therefore made sense to perform the parameter estimation in the context of astrophysical information other than our photometry. I did this by adopting various known distributions relating to stars in the Sun's neighborhood as Bayesian priors, and taking as each star's physical parameters the ones that maximized the posterior probability of obtaining the observed magnitudes and colors. As amplified below, this method has several drawbacks, but it does use the available data to far greater advantage than do methods that ignore the prior constraints. Much of the software described below is devoted to ways of quantifying the needed prior distributions.

Finally, filtering out bad data and formatting the result for the next stage in KIC assembly was also straightforward, consisting mostly of identifying spurious detections, failed fits, and so on. The criteria for passing results to the final KIC are described below.

2. Calibrated Photometry

The data received from Mark Everett consisted of one ascii file for each image taken at the telescope. Each file consisted of a header and two data tables. The header contained information relating to the image as a whole: Julian date, telescope coordinates, filter, exposure time, as well as various quantities such as PSF width, derived by Mark in the course of his analysis. The first data table contained one line for each star detected in the image. Values in each line included the star's coordinates obtained from a WCS fit, its instrumental magnitude, the local sky brightness, and various error estimates and goodness-of-fit parameters. The second table contained the coordinates, J, H, and K magnitudes, and unique integer identifier for all of the stars in the 2MASS catalog (Skrutskie et al. 2006) that fell within (or just outside of) the field covered by the image in question.

2.1. Database Organization

I organized 4 databases (or sets of databases) to hold these data and to facilitate later processing. These were as follows:

The *images* database contains one entry for each image processed by the system. The complete schema for this database is given in Appendix 2. In brief, however, it contains the header parameters from the input data file (JD, filter, exposure time, etc) as well as slots for image-level data quality parameters such as cloud flags.

The *blocks* database contains information about groups of sequential images (usually but not necessarily all the images taken in a single night) that are to be calibrated using one set of values for the atmospheric extinction parameters. Its full schema can be found in Appendix 2. Its most important contents are the starting and ending Julian date and image sequence number for the block, the extinction coefficients for each filter, and a quality-of-fit parameter describing the consistency of the extinction data.

The *stars* database contains one entry for each distinct star that is identified in the photometry. Its full schema can be found in the Appendix 2. It contains a unique identifier for each star, the 10-digit 2MASS identifier if it exists, the nominal coordinates (which are always the 2MASS coordinates, if the star exists in the 2MASS catalog), the difference between the nominal coordinates and the measured ones, the estimated extinction-corrected magnitudes in each filter and their errors, a mean Julian Date averaged over all observations for the star, and flags for variability, crowding, data source, and evidence of bad data.

To reduce search times, the stars database is actually subdivided into many *tiles*, with each tile covering one degree of right ascension by one degree of declination. Tiles are named according to the minimum values of RA and Dec that they contain, using names such as "t292+042". This tile would contain all stars with $292.0 \leq$ RA < 293.0 degrees, and $42.0 \leq$ Dec < 43.0. There is also a special tile "tstds" that contains all of the primary and secondary standard stars that are used for estimating atmospheric extinction parameters. This tile contains stars from all over the sky, although a large majority reside in one of 3 standard pointings within or near the Kepler field, which were used for frequent monitoring of extinction during normal observations.

The *observations* databases contain every observation obtained for each star that exists in the stars databases. The full observations schema can be found in the Appendix 2. This database contains data relating to single measurements of individual stars. These include the unique star identifier, the nominal and measured coordinates, the image sequence number, the filter name, the instrumental magnitude, its estimated error, and various fitting flags relating to it, the airmass of the star at the time of observation, and the chip or quadrant number and x, y coordinates of the star on the detector.

Separate groups of observations databases exist for each of the 7 filters in which we have observations, namely the SDSS-like filters u, g, r, i, z, and also the intermediate-band filters $Gred$ and $D51$. Early experience showed that observations with the u and $Gred$ filters were prohibitively expensive in observing time, so processing stages following the atmospheric extinction correction almost ignore these two filters. Besides being subdivided by filter, the observations databases also are subdivided into tiles in the sky, in a manner identical to the stars databases. In addition, a few tiles containing the standard pointings were observed so often that it was also necessary to subdivide them according to date. When necessary, observation tiles could be segmented in this fashion using routine tostd_segment, and segmented tiles could be reassembled into giant concatenations covering the entire duration of the observing campaign using the routine tostds_deseg.

2.2. Data Ingestion

Databases may be populated in two ways. For initialization, one must enter into the stars databases those stars that are to be used as primary photometric standards. I took these from the SDSS DR1 (Stoughton & Jester 2004) (the only SDSS catalog available at the time), choosing fields that could be observed by the Mt. Hopkins telescope in a single pointing, that contained a reasonable number of stars claiming good photometric precision, and that showed a wide range of color. Stars selected for inclusion as standards satisfied $14.5 \leq r \leq 16.5$, with estimated errors in all filters smaller than 0.03 magnitude. My aim was to choose stars bright enough to have accurate photometry, but faint enough to avoid saturation.

The primary standards that were entered in this way may be identified by searching the tstds stars database for objects with $source = 2$. There are 316 of these stars; ignoring 6 evident outliers, they span the range $-0.4 \leq (g - r) \leq 1.4$, and $-0.2 \leq (i - z) \leq 0.4$ (although the great majority of them lie within $0.3 \leq (g - r) \leq 0.8$).

Because of their source, primary standards do not have observed magnitudes in the Gred and D51 filters. To define the instrumental magnitude scale in these filters, I set the Gred and D51 magnitudes for these stars so as to force agreement with model $(g - Gred)$ and $(g - D51)$ colors, given the stars' observed (SDSS) $(g - r)$ colors. The model magnitudes used for this purpose were computed from the Basel models (Lejeune et al. 1997) with $[Z] = 0$ and $\log(g) = 4.0$, and from the estimated wavelength response for the 1.2m telescope 4-Shooter camera. These magnitudes were entered into the database using the routine narrow_set2.

Later in the catalog-building process, it was desirable to build fields of secondary standard stars. The stars in these fields have their magnitudes defined by repeated calibration against the primary standards. Two of the pointings that are used for secondary standards were chosen within the Kepler field itself. In the normal observing sequence, the telescope returned to one of these pointings on an hourly basis, so that we could obtain reliable measures of extinction in the part of the sky that was of most interest. Other secondary standard fields were defined near objects (for example, the cluster M67) that we wished to use in the astrophysical calibration of the photometry. In all cases, I used the routine new_std_field to identify suitable stars in the desired regions of the sky and copy their property into the tstds and tostds tiles in the relevant databases.

After obtaining a few dozen nights of data with all of the standard stars, some of them could be identified as unsuitable because they showed large temporal variability, or they were members of close binaries, or for other reasons. I used routine cull_refstar to weed these out of the secondary standard database tiles and replace them in the tiles from which they originally came.

Observations obtained with the 1.2m telescope were ingested using the routine obs2db4.pro, under control of a script called run_obs2db4. The functions of the script were twofold: to operate on many image files in sequence, and to assure that the size of the "observations" database tiles do not grow beyond the limits set for them when they are defined. The first function is a simple scripting loop. The second was performed by determining the number of entries in each observations database tile using routine obstilecount, and if this value became dangerously large, rewriting the tile contents with a larger size limit. This latter manipulation was done by routine rebuild_obs2db.

The ingestion process first checks to see that the image in hand is a new one. If so, it enters the image data into the images database. It then performs a test to see if the coordinate spread in the star coordinates is reasonable. This test was motivated by early failures in the WCS coordinate solutions, which put many spurious stars into the database at random positions. In reduction of the final data, no images failed this test. The detected stars were then inserted into the appropriate stars database tiles, using the routine obs2stardb3.pro. This routine creates new database tiles as needed, creates entries for stars that previously had none, and identifies cases that are new observations of already-detected stars. I match stars in the new set of observations with existing entries purely on the basis of position: if the coordinates of a new star match those of an existing one within a radius of 1.2 arcsec on the sky, they are considered to be the same, otherwise not. I first test for new 2MASS stars, adding any new ones to the database. Star nominal coordinates (the “RA” and “Dec” entries in the stars database) are taken to be those of the first instance of the star that is seen. Thus, the nominal coordinates are the 2MASS ones for any star that has them. Also, I test every group of stars against the tstds database before searching the database tile(s) corresponding to the sky position for the image, and I prevent stars that match tstds objects from being replicated elsewhere in the databases. I create new entries for any stars (either 2MASS or 1.2m) that do not match existing stars.

A new star entry can be created based on observations taken with any filter. Spurious detections therefore create star entries that may turn out to have apparently valid observations in only one filter. No attempt is made to test for this condition during the ingestion phase of the data processing.

After assuring that each new observed star has a corresponding entry in the stars databases, routine obs2obsdb.pro inserts the observed instrumental magnitudes and the differences between the nominal and measured positions into the appropriate observation database tiles. Also entered are each star’s airmass at the time of the observation, various quality-of-fit parameters, and the star’s pixel position on the detector(s).

In addition to the database entries, the routine obs2db4.pro writes several statistics and data quality diagnostics regarding every incoming image to a running logfile “obs2db.log”. A description of this file is found in Appendix 3, “Configuration and Input Data Files”.

2.3. Estimating Atmospheric Extinction Parameters

After ingesting all of the data from a given night, I run the routine block_hand to identify which images are to be grouped together for purposes of extinction correction. This interactive routine identifies "standard" and "target" pointings so that if one wishes to segment a night into several blocks, this can be done in a way that leaves adequate standard observations in each block. In practice, this facility was very seldom used, and in the final reduction of data, all blocks were defined as entire nights of data. The routine block_hand populates the blocks database, and sets default extinction parameters for new blocks. Some of these are normally replaced by actual fitted values later in the processing, as described below.

The interactive routine bad_hand elicits judgments from an operator about the quality of the data, storing flags among the entries for each image relating to clouds, and to errors in the astrometric and photometric fitting processes (blocks database entries "cldflag", "astflg", and "fitflg"). For this purpose it displays, for each block selected, several measured indices relating to the quality of the data, obtained from the obs2db.log file. The photometry display shows two plots: an estimate of the total cloud extinction for each image, and the value of the sky background in the filters g and i, for each image taken in those filters. I estimate cloud extinction by finding all 2MASS stars in the image that have $0.35 \leq J - K \leq 0.45$, and that are also detected in the 1.2m imagery. The code extrapolates estimated magnitudes for these stars from the 2MASS magnitudes using empirical relationships, approximately valid for our filters and for this limited $J - K$ color range. The median difference between the observed and extrapolated magnitudes for these stars is a measure of the total extinction, and is the plotted quantity.

The routine fitstds1 identifies all observations of standard stars (ie, those of primary or secondary standard stars, which are all stored in the 'tostds' databases) that are taken within a particular block of data. It then performs a robust minimum-chi-squared fit in order to estimate some parameters in the extinction model representing this block. These parameters are then saved to the blocks database. Optionally, the routine also generates postscript plots illustrating the input data and the quality of the fit. Other options allow changing the selection of extinction model coefficients that are fitted, and changing the behavior of the routine from the default mode (in which data are displayed to the screen and edited by hand) to a batch mode, in which no user intervention is required or allowed.

The extinction model is based on that used for the SDSS standard star grid (Smith et al. 2002), although in practice several of the coefficients in this model were found to be difficult or impossible to measure on a nightly basis, and were set to constant (sometimes zero) values. For filter i and an instance j (within a block) of observing a particular star k, this model represents the observed magnitude m_{ijk} as follows:

$$m_{ijk} \ = \ m_{0ik} \ + \ a_i \ + \ k_i(X_{kj} - X_0) \ + \ b_i(C_{ik} - C_{0i}) \ + \ c_i(C_{ik} - C_{0i})(X_{kj} - X_0) \ \ .$$

Here m_{0ik} is the true magnitude of star k in filter i, $X_{kj} - X_0$ is the difference between the airmass of star k at instance j and a standard airmass X_0, C_{ik} is the color of star k using a particular pair of filters appropriate to the filter i, and C_{0i} is the color of a 'typical' star using that same filter pair. The coefficients a_i, k_i, b_i, and c_i are parameters that may be chosen to give the best fit to the observations.

Note that by writing the extinction as above, the 'standard' magnitude of a star is what one measures when the star lies at airmass X_0, not when it is outside the atmosphere. One advantage of this approach is to reduce the effect of errors in estimation of the extinction per unit airmass (the k_i). For this analysis, I took the standard airmass X_0 to be 1.215. This value was chosen to coincide with the value 1.3 adopted by SDSS, with allowance for the higher altitude (hence lower air density) of the Sloan telescope relative to the 1.2m telescope at Mt. Hopkins.

In practice, I found that the observations of standard stars on a single night ordinarily did not suffice to give a reliable estimate of all of the coefficients in the extinction relation. Indeed, I obtained best results by fitting for the coefficients a_i each night, using values averaged over an entire observing season for the coefficients k_i, and taking the coefficients b_i from theoretical calculations based on model stellar fluxes and the estimated wavelength responses of the various filters. I took the coefficients c_i to be uniformly zero. Figure 1 shows the time variation of the coefficients a_i for all but 2 of the 205 nights (data blocks) for which we have data. To a large degree, these zero-point magnitudes reflect the arbitrary relation between stellar magnitudes and counts per second registered by the CCD through the various filters; there are, however, components of these zero points that have more physical interest. Signficant jumps in the zero points occurred with the inauguration of new cameras (changing from 4-shooter to MiniCam about JD 2453233, and from MiniCam to KeplerCam about JD 2453626). There also are trends and seasonal variations visible in these data, notably a loss of about 20% sensitivity in the KeplerCam data since the time it was installed.

After estimating the a_i coefficients with fitstds, I produce a second estimate of the extinction and the quality of the night's data using the routine plot_refvsx. This routine uses only selected stars observed at each reference pointing to estimate the a_i and k_i coefficients for each night. It chooses stars so that exactly the same subset of stars in each reference pointing is always used in the fitting process for a single night. This minimizes errors arising from uncertain standard magnitudes of the reference stars. The routine plot_refvsx thus produces nightly estimates of the k_i coefficients as well as the a_i (although the k_i are well-determined on something less than half of the nights). Figure 2 shows the variation of the k_i with time, measured on the nights when I judged the observations to be consistent enough to allow a measurement. The night-to-night variations tend to be large (up to 40% of the mean values), but they are also well-correlated among filters. Thus, there is good evidence for a variation of aerosol extinction, having very little wavelength dependence, and varying on a time scale of one to a few days. The SCP photometry could be improved by putting this information about the k_i back into the extinction model on a nightly basis. The corrections would not be large, however. Given the size of the errors in k and the typical difference between the airmasses of actual observations and the standard airmass ($X_0 = 1.215$), corrections in the observed magnitudes would be at most 0.09 mag, and would be less than .03 mag for 89% of the observations.

The routine plot_refvsx also produces a diagnostic plot for each data block, showing the variation of the apparent magnitudes of the chosen reference stars during the night. We used these plots, along with those produced by bad_hand, as a guide to which data should be marked as contaminated by clouds and hence excluded from the fits that yield true stellar magnitudes m_{i0}.

After estimating the extinction coefficients for a block of data, I used routine fitstars to estimate the m_{i0} magnitudes for every star that was observed in any filter within that block. The routine fitstars identifies every star that was observed at any time during a given block, and then gathers together every observation to be found in the databases for each of those stars. The routine uses a robust minimum-χ^2 process to make the magnitude estimates. In this context, "robust" means that the routine discards from the fit outliers among the observation list for each star. This process can only succeed if there are at least 3 observations, however, so for about 80% of the pointings in the Kepler field (those with 2 or fewer visits), no outlier rejection is performed. (Of the 1600 pointings that span the entire Kepler field, none have zero visits with acceptable photometry, 8% were visited only once, 71% twice, 17% three times, and 3% four or more times. Two pointings, corresponding to secondary standard fields, were visited more than 260 times each.)

Results of the fitting process (including error estimates) are written back to the stars databases, replacing whatever values may have already been there.

2.4. Precision of Corrected Magnitudes

The precision (in the sense of repeatability) of stellar magnitude estimates may be assessed by collecting all the measurements for each star in the secondary standard fields (each having typically hundreds of individual observations), and estimating the scatter in these time series. I did this for as set of 5652 stars brighter than $r = 19.5$, each having at least 50 individual observations. For each star I estimated the time-series RMS from the interquartile dispersion q as $RMS = q/1.349$; this formula gives the right result for a gaussian distribution, but is insensitive to a small proportion of extreme outliers. I then computed for each filter the median over stars in 1-magnitude bins centered on integral values of the (time-series) median magnitude. The resulting RMS scatter as a function of stellar magnitude in each filter is plotted in Figure 3. Evidently the photometric errors are dominated by brightness-independent processes (probably errors in the atmospheric extinction correction) in all filters for stars brighter than about magnitude 14; for fainter magnitudes, photon statistics begin to have an effect, and become dominant by about magnitude 16. For the brighter stars, the repeatability of extinction-corrected measurements is about 2%, almost independent of filter.

The precision of color estimates (ie, differences between magnitudes measured in different filters) is usually somewhat better than for magnitudes (perhaps 1.5%), because much of the scatter in the extinction-corrected magnitudes comes from extinction processes that have long time scales, whereas the two measurements making up a color estimate were almost always taken within a few minutes of each other.

3. Model Stellar Atmospheres

The stellar classification process requires choosing among model stellar atmospheres (which have parameters $\{T_{eff}, \log(g), \log(Z)\}$) so as to best fit the photometric observations. For this purpose a grid of stellar atmospheres is needed. I used those by Castelli & Kurucz (2004); these cover $3500K \leq T_{eff} \leq 50000K$, $0 \leq \log(g) \leq 5.5$, and $-3.5 \leq \log(Z) \leq 0.5$, although not all gravities are represented at all temperatures.

The models provide fluxes as a function of wavelength. To convert these to stellar magnitudes (relative to an arbitrary zero point) the routine phot_models multiplies the stellar flux by an estimate of the CCD response and by the estimated filter transmission functions, and integrates these over wavelength to give the rate of photoelectron production as a function of filter and of each of the parameters that characterize the model atmospheres. The response curve that I used to describe the CCD is contained in the routine ccd_filtprofile, while the transmission curves for the filters are found in the on-line documentation directory /doc/filt_profiles. The filter bandpasses for which I computed model magnitudes were the Sloan-like filters u, g, r, i, z, the 2MASS filters J, H, K_s, and two intermediate-band filters $D51, G_{red}$ that were obtained specially for the SCP. These latter filters are characterized by center wavelengths of 510 nm and 432 nm respectively, and have fullwidths of about 15 nm. In practice, I found that the u and G_{red} filters required too much exposure time, so we stopped observing with these filters. Accordingly, the databases contain slots for these filters, but very few corresponding observations.

Zero points for the magnitudes computed by phot_models are in a sense arbitrary, since the model fluxes refer to surface areas on the radiating surface, not at the receiving one. Part of the input to phot_models is an ascii file containing values for these zero points, chosen so that model magnitudes and instrumental ones are in approximate agreement, given the assumption that the model i magnitudes are correct. A pointer to this input file as used for creation of KIC10 ("4s_survey_c2") is contained in Appendix 3.

After computing the 10 magnitudes described above the routine phot_models next forms differences to yield 9 independent colors, viz., $(u-g), (g-r), (r-i), (i-z), (z-J), (J-H), (H-K), (g-G_{red}), (g-D51)$. The arbitrary zero point cancels from these differences, so these numbers are a representation of the spectral energy distribution of the stallar models that does not depend upon stellar radii or distances. Most of the operations that later try to match stellar properties to observations rely on these 9 colors, sorted into this particular order and represented as an IDL vector. (Various strategems are used to avoid bad results when not all colors have valid observations.)

In practice, as amplified below, I found the model colors to have systematic differences with those observed for members of stellar clusters, for which the true stellar parameters are known with good accuracy. To remove these, all routines that use model colors apply to them a transformation of the form

$$c' \ = \ c \ + \ aa \ + \ bb(g-r) \ \ ,$$

where c' is the transformed model color (which may be any of $\{u-g, g-r, r-i\}$ etc.), c is the untransformed color, $(g-r)$ is the untransformed model $(g-r)$ color, and aa and bb are constants determined by fitting to observations. The aa coefficients are evidently completely degenerate with the zero-point constants contained in the file 4s_survey_c2 described above; the only difference is one of convenience, since the zero-point constants operate at the level of magnitudes, while the aa operate at the level of colors. The existence of two such similar sets of fitting coefficients operating at different places in the code is mostly a statement about the iterative nature of the algorithm development itself.

The output of phot_models is a table of colors as a function of sequential model number, with accompanying tables giving the parameters $\{T_{eff}, \log(g), \log(Z)\}$ as functions of the same sequential number. For subsequent processing it is convenient to rearrange these into a 3-dimensional array that gives the colors evaluated on a regular cartesian grid of T_{eff}, $\log(g)$, and $\log(Z)$. Moreover, to avoid unpredictable behavior in the process of optimally matching models to observations, it is desirable that the grid of model colors be complete (ie, that within the range of $\{T_{eff}, \log(g), \log(Z)\}$ covered by the "colors" array, all values should be present and artificial discontinuities should be avoided.) In other words, the search for an optimum fit to the observations may require evaluating colors for models that will never actually be chosen as best. Nevertheless, such models must return color values that do not cause the optimizing algorithm to crash, nor to diverge toward crazy values. The Castelli models do not meet these requirements, since they cover only combinations of $T_{eff}, \log(g)$ for which real stars are at least moderately common.

Thus, obtaining a usable colors array requires extrapolating the Castelli models into regimes where physically-based computations have not been done. This is accomplished by a small suite of routines: Routine basel_index1 computes index arrays that allow models with given values of $\{T_{eff}, \log(g), \log(Z)\}$ to be found quickly, if they exist. Routine color_fillin1 extrapolates each color from existing valid ones. For each color, it searches the desired 3-dimensional range in $\{T_{eff}, \log(g), \log(Z)\}$ for invalid entries. For each invalid entry found, it searches a 5x5x5-pixel cube centered on the cell of interest for valid values. If 30 or more of the possible 124 neighbors are found to be valid, it fits a 3-dimensional quadratic function to these valid values, assigns the result of evaluating this polynomial to the cell in question, and marks it as "valid". This process is then iterated, growing the region in which valid colors are found, until the entire 3-dimensional domain has been filled in. For historical reasons, the result of the extrapolation process is not the desired cartesian array of colors, but rather a list of color vectors, each tagged with appropriate values of the model atmosphere parameters. Routine color_block reformats this list into the 4-dimensional array "colblock", with dimensions spanning color (ie, $(u-g), (g-r)$, etc.), $T_{eff}, \log(g)$, and $\log(Z)$. The result of the extrapolation typically displays small-scale and small-amplitude oscillations. To suppress these, the routine colblock_smooth applies a boxcar smoothing to each $T_{eff} - \log(g)$ plane separately, with due attention to reasonable behavior at the boundaries of the colblock array.

After these extrapolations, the colors found in the colblock array are seen to be smooth and plausible extensions of the original Castelli model colors. It turns out, however, that even where the original Castelli colors are defined, these model colors fail to represent those we observe in a number of systematic ways. Notably, the Castelli models with near-solar metallicity appear never to produce $(g-r)$ colors redder than about 1.05, whereas there are many observed and likely unreddened stars with $(g-r)$ as red as 1.4. Successful operation of the classification algorithms requires that models be able to provide better fits to the observations than this, so it is necessary to edit the colblock array to agree better with the observed data. This is done using routines gmr_adjust and colors_adjust. The former adds a linear function of T_{eff} to the $(g-r)$ colors, for $T_{eff} \leq$ 4000K, with a maximum correction of +0.4 mag at 3500K. The resulting colors fit the observations much better, but there is no reason to believe the particular transformation connecting T_{eff} and $(g-r)$. Routine colors_adjust makes smaller adjustments (always less than 0.2 mag, and typically less than 0.05 mag) to several colors in various temperature ranges, with the aim of removing systematic differences between the observed colors and the best-fitting models. One effect of these adjustments is to weaken the tie with a physically-motivated T_{eff} scale, especially for temperatures below 4000K, where the models seem to work poorly.

For use within the model-fitting routines, the routine colors_blk_intrp uses tri-linear interpolation to provide estimated colors for arbitrary values of $T_{eff}, \log(g), \log(Z)$ within the range spanned by the colblock array.

4. Color Calibration

There are several places in the analysis code where the magnitudes and colors produced from the Castelli & Kurucz (2004) models are adjusted in order to fit the data better. I used an iterative procedure to determine these adjustments, starting by matching a small number of SDSS stars. By design, these stars spanned only a small range of $(g-r)$ color, and of course the Sloan data contained no information about the non-standard filters D51 and Gred. The remainder of the calibration process therefore consisted of a series of steps in which I used larger and larger groups of more poorly documented stars to refine the adjustments to the model magnitudes, and to expand their applicability to all filters and to a wide range of colors. This section describes that calibration process.

My first step was to establish approximate zero points for all filters by matching all of the observed colors of Sun-like stars. I first chose stars from the list of Sloan standard stars given by Smith et al. (2002) For 4 of these stars (0.715-218, SA 97-288, SA98-978, and SA 114-654) I was able to obtain $(B-V)$ colors from SIMBAD, and (J, H, K) magnitudes from 2MASS (Skrutskie et al. 2006). I then fit linear relations giving the observed Sloan and 2MASS colors for these stars $((g-r),(r-i),(i-z),(J-H),(H-K))$ in terms of $(B-V)$. I did not include $(z-J)$ in this list because I did not wish, at this stage, to mix results from the IR and visible-light surveys; I restricted attention to these 4 stars so that the linear fit against $(B-V)$ color would be meaningful. According to the $(B-V)-T_{eff}$ relationship from Sekiguchi & Fukugita (2000), for $T_{eff} = 5750K$, one expects $(B-V) = 0.632$. Evaluating the linear color-color relations just established at this value of $(B-V)$ gives empirical constraints on the various Sloan and 2MASS colors for $T_{eff} = 5750$K; these are $(g-r) = 0.427$; $(r-i) = 0.122$; $(i-z) = 0.026$; $(J-H) = 0.260$; $(H-K) = 0.076$. For the SCP catalog, I arbitrarily took the i magnitude to be correct; this allowed me to compute zero point corrections for all of the SCP filters $\{g, r, i, z\}$ I inserted these values into the input file for phot_models, and computed all of the model colors.

To extend the above results to include D51 and the 2MASS magnitudes, I needed to set zero points for the colors $(g - D51)$ and $(z - J)$. I did this by plotting scatter diagrams of observed $(g - D51)$ and $(z - J)$ against $(g - r)$, for a well-populated data tile t285+045, which lies at +17 degrees galactic latitude. Overplotting the corresponding model color-color relations allowed me to estimate the needed zero-point corrections needed to bring the model $(g - D51)$ and $(z - J)$ colors into agreement with the observations. I inserted these additional zero points (for $\{D51, J, H, K\}$) into the input file for phot_models, and recomputed all of the model colors. The final zero points estimated in this way are found in the file 4s_survey_c2, which is listed in the Appendix 3.

At this point it became clear that a large disparity, and an obstacle to further progress, was the difference between observed and model $(g - r)$ colors for cool stars. As mentioned above, Castelli model $(g - r)$ colors for solar-metallicity dwarfs never exceed about 1.2, while many cluster members and faint field stars (which are almost certainly dwarfs of near solar metallicity) have colors that approach 1.5. To allow more accurate modeling of the photometry for these stars, I applied ad hoc adjustments to the $(g - r)$ colors of all Castelli models with temperatures of 4000K or less. These changes were made using routine gmr_adjust, which adds to each model $(g - r)$ entry a value that depends only upon T_{eff}. The corrections applied were $\{4000K:\ +.179;\ 3750K:\ +.289;\ 3500K:\ +.402\}$. I arrived at these corrections by overplotting model colors onto color-color diagrams for field stars from tile t285+045, and adjusting only the $(g - r)$ color to obtain a plausible fit. As mentioned above, this ad hoc process renders the T_{eff} scale unphysical, at least so far as the $(g - r)$ color is concerned.

I next compared model colors to those measured for known single members of the open cluster M67, with the aim of estimating linear color transformations between the (adjusted) models and the observed colors. As described above, these transformations were of the form

$$c'_i \; = \; c_i \; + \; aa_i \; + \; bb_i(g-r) \;\; ,$$

where the coefficients aa_i, bb_i corresponding to the i^{th} color are determined from a robust fitting process in the routine m67_color_fit. The routine attempts fits for a sample of 116 stars for which reliable values of T_{eff} and $\log(g)$ are available from spectroscopic fits performed by Soeren Meibom. From these stellar atmosphere parameters, routine colors_blk_intrp computes model colors, and then m67_color_fit estimates transformation coefficients aa_i, bb_i that give the best match between the transformed model colors and observations of the corresponding observed stars. Between 63 and 77 stars in the "stars" database matched M67 stars from Meibom's list, depending on the color involved. Perhaps the most important failing of this sample of stars was its relatively strong concentration toward moderately red colors: it contained no stars bluer than $(g-r) = 0.21$, 10 stars redder than $g-r$=0.8, and only 1 star redder than 0.9.

Carrying out this process led to adequate fits to the data, but the coefficients bb_i were only marginally significant. Accordingly I performed a second fit, forcing the bb_i to zero. This also gave an adequate fit, with half as many free parameters. I thus adopted the result of the second fit for the color transformation; in effect, this process led to small adjustments in the zero point corrections already found in the file 4s_survey_c2. The color corrections were all smaller than 0.05 mag, coming out as follows: $aa\{(g-r),(r-i),(i-z),(z-J),(J-H),(H-K),(g-D51)\} \;\; = \;\; \{-.024,-.048,-.026,-.044,.030,-.023,-.005\}$. The corresponding (per star) RMS differences between the transformed model colors and the observed ones were $\{.021,.014,.011,.022,.029,.028,.017\}$ mag. The most obvious dependence in these fits is that the RMS scatter is about 50% larger for colors that involve 2MASS magnitudes than for those that do not.

It was now possible (using routines to be described below) to estimate the parameters

$\{T_{eff}, \log(g), \log(Z)\}$ for a large number of field stars, including effects of interstellar reddening. I did this for a sample of about 2000 stars in 4 tiles covering the cluster M67, and for each star computed the fitting error, ie, the difference between the observed colors for the star and the colors predicted by the best-fit model. In the ideal case, one would hope these fitting errors to be normally distributed with zero mean, uncorrelated with one another, and moreover to show no systematic dependence on any of the fitted parameters. Of course, there were variances from these ideals. The most prominent errors appeared as nonzero mean values for the color differences. I eliminated these differences by adjusting (yet again) the color transformations for the various colors. The new aa_i values were $\{-0.022, -0.054, -0.005, -0.075, 0.025, -0.036, 0.0\}$. These values, corrected for reddening and based on a much larger sample of stars, are all within 0.03 mag of the previous ones.

A worry concerning the color-temperature relation is that the median observed $(g-r)$ is 0.416 for the 35 M67 stars with T_{eff} lying between 5700K and 5800K (with median T_{eff} = 5763K). Since these suffer reddening of about 0.04, the intrinsic $(g-r)$ colors must be about 0.376, which is too blue by almost 0.05 mag. This corresponds to T_{eff} values that are in error by about 150K.

A large component of the remaining differences between models and observations is well-correlated among colors, and all are correlated with the model values of T_{eff}. This suggests small systematic errors in the model flux distributions and hence colors. To reduce these, I applied small ad hoc adjustments to the colors $\{(r-i), (i-z), (J-H), (H-K)\}$. These were linear in temperature and equal to zero at T_{eff} = 5000K, with slopes per 1000K equal to $\{.006, -.021, -.036, .011\}$. All of these adjustments were independent of $\log(g)$ and $\log(Z)$.

The difference between observed and model $(g-D51)$ color showed a more complicated behavior than the others. This difference was essentially zero for $T_{eff} \geq 4500$K, but turned sharply positive for cooler temperatures. On the other hand, $(g - D51)$ never grows larger than 0.4, which is roughly the value it attains in models of cool low-Z stars. To match the observed behavior (for mostly high-Z stars) and to retain the expected behavior at low Z, I adjusted the $(g - D51)$ colors in the following way: for $Teff \leq 4300$K,

$$(g - D51)' = (g - D51) - 0.23 f_g (T_{eff} - 4300)/1000.$$

where f_g is unity for $\log(g) \geq 3.5$, and decreases linearly with $\log(g)$ to zero at $\log(g) = 0$.

All of the adjustments to $(r - i), (i - z), (J - H), (H - K), (g - D51)$ are carried out by routine colors_adjust.

5. Reddening

Interstellar reddening is significant for most of the stars in the Kepler field, so it is necessary to include it in the models of stellar colors. I assumed that the wavelength dependence of reddening is described by the formalism of Cardelli et al. (1989). In all cases I used $R_V = A_V/E_{B-V} = 3.1$, which Cardelli et al. found to be the typical value for diffuse interstellar dust clouds.

Strictly, the reddening suffered by starlight depends on the spectrum of the star, and hence on its parameters $\{T_{eff}, \log(g), \log(Z)\}$. The magnitude of this effect is fairly small, however, so for speed of computation, I precomputed reddenings for our filters and a range of stellar parameters. I then adopted the reddening vectors for typical stars in the Kepler field ($T_{eff} = 5000$K, $\log(g) = 4.0$, $\log(Z) = 0$ as applying to all stars. Examination of results for different stellar parameters shows that these reddening corrections are in error by typically 0.02 mag per magnitude of extinction per 1000K in T_{eff}.

Computation of Cardelli et al. reddening vectors occurs in routine redden, while application of the parameterized version of the reddening law is done by the routine red_vector.

In stellar parameter estimation, the estimated reddening is derived using a model of reddening in the Milky Way. This model assumes that dust is distributed in a smooth disk aligned with the plane of the Milky Way, having an exponential decay of density with height above the plane. For all computations described here (except for certain test cases, such as parameter estimations for M67), I took the e-folding height for disk density to be 150 pc, and the density in the plane to be such as to cause 1 magnitude of extinction in the V band in a path length of 1000 pc.

6. Stellar Magnitude Error Estimation

Routine fitstars computes a formal error for the magnitude estimates for each star in each filter. This is based on the formal errors reported by Mark Everett's photometry software, and represents a lower limit on the true errors. I also compute a variability index based on the scatter (max minus min) of all the values for each star and each filter. The variability index is rather coarse: a value of 0 indicates total scatter less than 0.07 mag; 1 indicates between 0.07 and 0.3 mag, and 2 indicates more than 0.3 mag. Both error measures are stored in the stars database.

7. Fitting Stars to Observations: Linked Parameters

A star and the light we see from it may be characterized by many parameters, not all of them independent. These include not only its photometric colors, but also its apparent magnitude, distance, reddening, and intrinsic properties $\{T_{eff}, \log(g), \log(Z)\}$, mass, radius, and luminosity. Different photometric measurements speak to different combinations of these parameters; to obtain sensible estimates of the intrinsic parameters, it is important to respect the relations that connect them. Routine stellar_parms1 does the parameter estimation for each star in a list of database tiles, writing its results to the "properties" database. It does this in a way that (with the help of some approximations) takes account of relations among the various stellar parameters.

For a given composition, the stellar atmospheres models are defined on a grid in $\{T_{eff}, \log(g)\}$ When comparing observed colors c_i to model ones, it is therefore sensible to work in this space. But in color space, the effects of reddening E_{B-V} are largely degenerate with those of T_{eff}, and (in the context of my extinction model), reddening depends on the distance d to the star, and on the star's height above the galactic plane (or, given d, its galactic latitude b).

The apparent r magnitude of a star depends on its luminosity L, and hence on its radius R, on T_{eff}, and on the bolometric correction BC, but also on d and on the extinction A_r affecting its light. Finally, R is related to the stellar mass M and to $\log(g)$. Thus, the 9 observables are the photometric colors (in the typical case there are 7 of these), the apparent magnitude in any one filter (I use r), and the galactic latitude b. The eleven parameters we wish to estimate are $T_{eff}, \log(g), \log(Z), M, R, L, BC, d, A_r, A_V, E_{B-V}$. But connecting the parameters are the relations:

$$
\begin{aligned}
A_r &= \kappa_r \int_0^d \exp(-s \sin b) ds \\
A_V &= A_r/0.88 \\
E_{B-V} &= A_V/3.1 \\
L &= R^2 (T_{eff}/T_{sun})^4 \\
r &= r_{sun} - 2.5 \log L + BC_r + A_r + 5 \log d - 5 \\
\log(g) &= \log(g)_{sun} + \log M - 2 \log R \ ,
\end{aligned}
$$

where κ_r is the assumed r-band interstellar extinction coefficient in magnitudes per pc, R, M, and L are in solar units, d is in pc, and A_r, A_V, BC, and E_{B-V} are in magnitudes. The simple numerical relations between A_r, A_V, and E_{B-V} result from integrating the Cardelli et al. model of interstellar reddening against typical (5000K dwarf) stellar flux distributions, as described above.

Applying the above constraints reduces the number of unknowns from 11 to 5, but this is still an uncomfortably high-dimensional space to search for best goodness-of-fit. Worse, given realistic errors in the data, an unrestricted search in this space tends to lead to unphysical solutions in which various parameters are driven to extreme values in an attempt to balance misfits relative to the observations. To reduce these problems I have adopted two further ad hoc relations, namely

$$BC \ = \ BC(T_{eff}, \log(g))$$

and

$$M \ = \ M(T_{eff}, \log(g)) \quad .$$

The BC relation is valid for fixed composition, and for computational purposes I have simply adopted the results of the Castelli models for $\log(Z) = 0$. The inaccuracies resulting from ignoring the $\log(Z)$ dependence are justified by the relative scarcity of low-metallicity stars in the solar neighborhood.

The $M(T_{eff}, \log(g))$ relationship is true only in a statistical sense, because age and (to a lesser extent) metallicity effects cause the evolutionary tracks of stars of different mass to cross in the HR diagram. This causes inevitable ambiguity, especially for giants and subgiants. However, for the cool main-sequence stars that are of the most interest to Kepler, the approximation is fairly good. To estimate this relation, I took stellar evolution isochrones by Girardi et al. (2000), and weighted each by its age (thus, in effect, assuming a constant star formation rate), with 5 isochrones covering ages between 63 My and 4.5 Gy

I then smoothed the resulting distribution of masses in $\log(T_{eff}) - \log(g)$ space to yield typical masses at each point. This averaging was performed by routine masstable, which saved the desired values in a 2-dimensional numerical table for use by other routines.

An implication of adopting the $BC(T_{eff}, \log(g))$ and $M(T_{eff}, \log(g))$ relationships is that, at the same level of consistency, several other useful parameters of stars may also be treated as being functions only of T_{eff} and $\log(g)$. These include the luminosity L, the $V - r$ color (used in relating extinction to distance), and absolute magnitudes in V and r bands. I precomputed tables for all of these quantities, and used the routine parmblkintrp to interpolate into the tables to give values for any of these quantities as functions of T_{eff} and $\log(g)$.

The process of computing a goodness-of-fit statistic for an individual star then proceeds as follows. In routine stellar_parms1, the observed colors, the r magnitude, and the galactic latitude b are fetched for each star. If some but not all of the colors $\{(g - r), (r - i), (i - z), (z - J), (J - H), (H - K), (g - D51)\}$ from the stars database are valid, then routine guessval fills in the remaining colors with plausible values, setting the corresponding photometric errors to very large values so that they will not enter into later χ^2 optimizations. If at least 3 visible-light magnitudes or the 2MASS magnitudes are valid, then stellar_parms1 calls routine find_prop1 to estimate the intrinsic properties of the star.

Routine find_prop1 performs a straightforward (ie, not particularly efficient) search in $\{T_{eff}$, $\log(g)$, $\log(Z)\}$, seeking the minimum value of a merit function that is the sum of a χ^2 statistic based on the photometry and the negative logarithm of a prior probability. This section describes the χ^2 computation; the justification for this general Bayesian procedure and the computation of the prior probability distrubution will be described below. The search for the minimum involves evaluating the merit function on successive 2-dimensional subsets of the search space; for practical purposes this means evaluating model colors for a grid of models, each with specified T_{eff}, $\log(g)$, and $\log(Z)$, and each relating to a star with a specified r magnitude and galactic latitude. Routine grid_likeli does this. It first uses routine linked_props2 to apply some of the constraints listed above to provide values for the star's luminosity, distance, and extinction, given specified values for $\{T_{eff}, \log(g), \log(Z), r, b\}$. This is done by using an interpolated table look-up (via routine parmblkintrp) to give L, BC, and $V - r$ from T_{eff} and $\log(g)$, and then to compute the distance d and extinction from r, b, and the galactic extinction model. Grid_likeli then uses routine colors_blk_intrp to generate model colors for each T_{eff}, $\log(g)$, $\log(Z)$ in the grid that is of current interest to it, and knowing the extinction for each of these, it applies a corresponding reddening correction to the model colors. Last, it differences the observed and model colors, and (using the estimated observational errors) computes a χ^2 value for each grid point.

8. Bayesian Maximum-Likelihood Estimation

When fitting stellar parameters to our photometric data, pure χ^2 minimization often leads to unreasonable results. This is because outlandish combinations of the stellar parameters, corresponding to stars that are seldom or never seen in nature, may yield marginally better fits to the observations than do more plausible parameter combinations. (This tendency seems to be aggravated since, quite reasonably, nobody has bothered to compute real models for outlandish stellar atmospheres.)

A way to control this behavior is to apply Bayesian methods. To apply Bayes' Theorem to the stellar classification problem, I denote the intrinsic stellar parameters $\{T_{eff}, \log(g), \log(Z)\}$ by the vector $\mathbf{x}$, and the various photometric observations of a particular star by the vector $\mathbf{q}$. I assume that we know something about the statistical distribution of stars, so that it is meaningful to talk about the prior probability distribution

$P_0(\mathbf{x})$, so that the probability that a star lies within a small volume of parameter space is given by $P_0(\mathbf{x})d\mathbf{x}$. Then Bayes' Theorem says that if I add to this a priori information about stars a set of observations relating to a particular star, then the updated probabilities are

$$P(\mathbf{x}|\mathbf{q}) \;=\; [P_0(\mathbf{x})P(\mathbf{q}|\mathbf{x})]/P_0(\mathbf{q}) \quad .$$

The maximum-likelihood choice of stellar parameters is then the choice of $\mathbf{x}$ that maximizes this posterior probability. The denominator in this expression, the a priori probability of observing photometric indices given by $\mathbf{q}$, is a normalization that does not depend upon the parameters $\mathbf{x}$ of interest; for purposes of maximizing the posterior probability, it may be ignored. In the ideal case of independent Gaussian errors in the photometric observations, one can write

$$P(\mathbf{q}|\mathbf{x}) \;=\; \exp(-\chi^2) \quad ,$$

where χ^2 (not the reduced χ^2) is the usual goodness-of-fit statistic. Taking the logarithm of the posterior probability, it follows that maximizing this probability is equivalent to maximizing

$$\ln P(\mathbf{x}|\mathbf{q}) \;=\; \ln P_0(\mathbf{x}) \;-\; \chi^2 \quad .$$

This is what the stellar parameter-estimation procedures try to do.

As is well known, Bayesian methods have both advantages and disadvantages. In the current case one advantage is that wildly implausible stellar classifications are ruled out a priori; the emerging values of $\{T_{eff}, \log(g), \log(Z)\}$ are guaranteed to resemble those of known kinds of stars. The corresponding disadvantage is that stars with rare properties are almost certain to be misclassified, since such stars are represented insignificantly or not at all in the statistical samples that one uses to estimate $P_0(\mathbf{x})$. The Bayesian maximization places stars in the most plausible of the already- existing pigeonholes, irrespective of the quality of the fit.

Thus, for the principal Kepler purpose of distinguishing GKM dwarfs from giants, the Bayesian approach can be expected to work well. For other purposes (identifying possible brown dwarfs, say, or distant highly-reddened blue supergiants), it is better to rely directly on the photometry, and ignore the classifications.

Another problem with Bayesian estimation is that commonly there is no good basis for estimating prior probabilities. Fortunately, in the case of stellar classificaton, one can do rather well. For purposes of the KIC classification, I expressed $P_0(\mathbf{x})$ as the product of 3 terms:

$$P_0(\mathbf{x}) \; = \; P_{HR}[T_{eff}, \log(g)] P_Z[\log(Z)] P_z(z) \quad ,$$

where z is height above the galactic disk. Expressing P_0 as a product distribution implicitly assumes that there are no correlations between the various dependencies. Of course, this is not completely true. Because of the relations between age and Z, and between age and z, the product distribution is not strictly valid. The relatively low frequency of old, low-metallicity stars makes this distribution a reasonable approximation, however. It is nevertheless worth remembering that as I have implemented it, the classification scheme knows nothing of the existence of low-metallicity halo stars, nor of the absence of young, hot, low-metallicity stars, etc.

To estimate P_{HR}, routine mk_hrblock uses the Hipparcos (ESA (1997),Høg et al. (2000)) catalog to create a histogram of the nearby star distribution, sampled in $(B-V)_{Tycho}$, V_{abs} space. For this purpose it takes all the Hipparcos stars with parallaxes that are known to better than 10%. This sample contains 9590 stars. Routine prob_hr reads the selected Hipparcos data from file hipparcos.txt using routine rd_hip, and estimates distances and hence absolute V magnitudes from the stars' parallaxes. An unexplained oddity of constructing the HR diagram in $(B-V)$, V space is that the Tycho passbands are rather different from Bessell bands, so that by rights, the Tycho magnitudes should be transformed (according to ESA SP-1200) before being used to create color-magnitude diagrams. If one does this, however, the resulting diagram fits very badly with standard (eg Girardi) isochrones, in particular the one for M67, which probably approximates the typical star population in the solar neighborhood. The untransformed magnitudes fit theoretical isochrones much better; accordingly I have not transformed the Tycho magnitudes before using them.

The routine prob_hr then uses routine vbmv2tg to map the $B-V$ color and absolute V magnitude into log(T_{eff} and log(g). This is done by inverting the relations described above giving $B-V$ and BC as functions of $\{T_{eff}, \log(g)\}$, which in turn is done by iterative interpolation into precomputed tables. Once each star is tagged with its calculated T_{eff} and log(g), it is a simple matter to construct a histogram giving the fraction of stars found within each cell in that space.

After the histogram of star frequencies in $T_{eff} - \log(g)$ space was created, I performed several edits and additions to make it more realistic, complete, and useful. This work was done by routine hrb_adjust. I first set star frequencies to zero for regions where only 1 or 2 stars are found in isolated cells. I next added stars to the area of the histogram occupied by bright giants, since the volume covered with good parallax precision by Hipparcos is not large enough to populate this part of the diagram. I then scaled the frequencies for intrinsically faint stars to account for the search volume that is implied by the stellar absolute magnitudes, since this volume is determined by the astrometric precision only for relatively bright stars. After doing this, the very faintest M stars were still unrepresented, so I added stars to populate the faint tail of the main sequence. Finally, I smoothed the histogram to yield one that represents the broad features of the local stellar population, but that has little small-scale structure. Routine mk_hrblock returns the natural logarithm of this histogram, which is taken to be the logarithm of P_{HR}.

To estimate $P(Z)$, I used the compilation by Nordström et al. (2004) of metallicities for about 14000 bright nearby stars. Routine prob_z evaluates a polynomial fit to the natural log of the histogram of Nördstron's metallicity values for any input value of $\log(Z)$, and returns the logarithm of the corresponding probability.

Routine prob_dist accepts the log of the estimated distance to a target star (which has already been computed from its known apparent magnitude and galactic latitude, and from its assumed values of T_{eff} and $\log(g)$). It returns the ln of the probability of finding a star per unit solid angle in the sky and per unit log(distance). This routine assumes the galactic disk's number density decreases exponentially, with a vertical scale height of 300 pc. It also allows for the cubic increase in sample volume with increasing distance.

Once all of these probabilities and the photometric χ^2 statistic have been estimated, the routine grid_likeli (remember grid_likeli?) returns the ln of the prior likelihood as $\ln(P_0(\mathbf{x}))$, and also $\chi^2(\mathbf{x})$.

9. Likelihood Maximization

Routine find_prop1 invokes grid_likeli to compute the prior probability and the photometric χ^2 statistic for every point in a 2-dimensional subgrid of the interesting space of $\{T_{eff}, \log(g), \log(Z)\}$. It does this several times, each time searching for the maximum of the likelihood function on the subgrid in question, and each time refining the subgrid, either by shrinking the size of the region examined or by changing the choice of variables (T_{eff} and one or the other of $\log(g)$ and $\log(Z)$) on which to compute the grid. At the end of this process, it looks to see if the minimum on the last grid (which lies in $T_{eff}, \log(g)$-space) lies on a boundary of the searched grid. If it does, the box is shifted by a standard distance towards the maximum value, and the maximum of the likelihood function is found again. The boundary test is then repeated; if more than 3 repeats are necessary, then the star's fit convergence flag is set to "failed" (but the last values of the stellar parameters are nonetheless returned to the calling routine). If a boundary failure does not occur, then interpolated best values of $\{T_{eff}, \log(g), \log(Z)\}$ are returned, along with error estimates. These estimates are formal errors. They are based on the curvature of the likelihood function along the axes $\{T_{eff}, \log(g), \log(Z)\}$, which is to say, they take no account of correlations among errors on the different parameters.

The stellar parameter estimation procedure takes place independently for each star in a tile in *RA*, *Dec*. When all of the stars in a tile have been done, all of the results for that tile are written to the "properties" database. This calculation is trivially parallelizable by assigning different tiles to different processors. After data ingestion, it is the most time-consuming part the data analysis. Running on 5 processors, however, it is possible to recompute all of the stellar parameters in about 3 days with current (2008) computing facilities.

10. Deriving Secondary Quantities

Once the likelihood maximization has been done for a star, several derived quantities are computed and stored in the properties database. Routine find_prop1 uses linked_props2 to estimate the extinction for the target star, based on the final values of $T_{eff}, \log(g)$. The method for doing this amounts to computing the absolute magnitude of the star from the table lookup implemented in routine parmblkintrp, computing the distance to the star from its observed r magnitude and the model of galactic dust distribution, and then evaluating the dust model extinction, given the star's distance and galactic latitude. Note that this process does not involve an explicit fit to the observed colors. Such fitting happens only implicitly, in that the fitted $\{T_{eff}, \log(g), \log(Z)\}$ are consistent with the dust model and the implied distance to the star.

For given $T_{eff}, \log(g)$, the routine parmblkintrp returns (among other things) an estimate of the stellar luminosity. Routine stellar_parms1 combines this estimate with T_{eff} to obtain an estimate of the stellar radius, which is stored in the properties database.

Routine stellar_parms1 also estimates the "Kepler" stellar magnitude kepmag (defined in AB magnitudes) that the star would appear to have as seen by the Kepler photometer. This is done by routine kepler_mag. The method used is to compute a linear transformation of the observed g, r, i magnitudes; alternate transformations are used if any one of these three magnitudes is invalid. I derived the transformations by fitting linear relations between spectral systhesis model magnitudes $\{g, r, i, k\}$, where the k magnitude was derived from Kepler photometer spectral response data provided by Dave Koch. The fitting formulas are accurate to about ± 0.2 mag (relative to the simulations) for stars hotter than 3500 K except if only g and r are valid, in which case errors for M stars may exceed 0.6 mag, in the sense that kepler_mag returns magnitudes that are too faint.

11. Star Selection, Formatting, and Output

After the stellar parameter estimation has been done, it is necessary to select stars that have information of high enough quality to be included in the final catalog, to extract the data for these stars from the database and format it in a way suitable for ingestion into Dave Monet's KIC database, and to write it to output files in a form that can be readily transmitted. These functions are performed by routine dbdump.

The routine dbdump accepts a list of tilenames and the name of an output directory. For each tilename in the input list, the routine opens the corresponding database tiles in the stars and properties databases. It reads the required data for all stars from each of these two database tiles, and scales some variables (RA, for instance) into the form desired for output. It then counts for each star the number of valid visible-light magnitudes and the number of valid NIR magnitudes, and it discards all stars that do not have at least 2 valid optical magnitudes or 3 NIR magnitudes. This rejection step commonly removes something like 2/3 of the stars from the catalog. Next, it sets the astrophysical parameters $\{\log(g), \log(Z), A_V, E_{B-V}\}$ (but not T_{eff}, which is assumed to be reasonably well determined so long as one optical color can be performed) to be invalid for all stars where the number of valid visible-light nagnitudes is less than 4, or where the stellar parameter estimation code did not converge. This step typically causes only a small fraction of the remaining stars to be tagged with invalid physical parameters. Finally, dbdump writes the desired data to an ascii file in the desired directory, with a name that is derived from the input tilename. The data are formatted using the "|" character as a separator, to allow easy ingestion into Dave Monet's SQL database.

The gzipped tar file containing all of the output data typically was less than 150 MB in size. I transferred this to Dave Monet via anonymous FTP.

12. Testing and Diagnostics

A number of diagnostic plots are generated as a matter of course in the process of ingesting and extinction-correcting the photometric data, and as checks on the outcome of estimating stellar parameters. Moreover, I have tested the output of these routines against several well-studied sets of stars, with the aim of understanding the accuracy of the results and of diagnosing the errors that remain.

12.1. Plots from bad_hand

The routine bad_hand makes 3 plots, each containing 2 panels and each plotting diagnostics of photometric quality against the fractional part of the JD for a single night. The diagnostics plotted by bad_hand are actually computed during the ingestion process by the routine obsdiagnos, and stored in the file obs2db.log. Plots from bad_hand for every night of KIC observations are available at www.cfa.harvard.edu/kepler/photometry/plots1/plots_ps/. A representative set of bad_hand plots is shown in Figures 4, 5, and 6.

The top panel of the first plot (Figure 4) is labeled "diff SDSS", but this is an error; it should be called "diff 2MASS". This plot is intended to help identify periods of cloud. For each image in a night, bad_hand locates all of the stars contained within the image that have 2MASS $(J - K)$ colors in the range 0.35 to 0.45. These stars are matched to stars observed in the image, and for each matching star, the routine computes an expected magnitude, given the filter in which the image was taken and the J and K magnitudes of the star. The plotted quantity is the median (or average, if there are too few stars to make a meaningful median) of the difference between the observed and expected magnitudes. The extrapolated 2MASS magnitudes have systematic errors that cause a repeated pattern in the differences corresponding to the observing program cycling through the filters $g, r, i, z, D51$ (with several filters having both long and short exposures). Aside from this pattern, values that exceed about 0.7, trends, or values that are more than about 0.5 mag from the mean for the night are all indicators of suspect data.

The bottom panel of the first plot shows the sky brightness in the filters g and i. For both filters, the units are scaled so that a near-zenith observation on a clear dark night gives values of about unity. This plot is particularly useful for identifying intervals of cirrus when the Moon is up. If flags are set indicating likely clouds, or astrometric errors, or fitting errors for a given image, symbols indicating this fact are overplotted on the sky brightness plot.

The top panel of Figure 5 shows the number of "unmatched" stars found on the image. For this purpose, an unmatched star is one that has a measured magnitude less than 14.5 in the current filter, and that is not matched in position to a 2MASS star. A large number (greater than 10, typically) of such unmatched stars often indicates an error in the astrometric solution for the image.

The bottom panel of Figure 5 shows the RMS differences between star positions as fitted for each image and the matching 2MASS stars that are brighter than magnitude 14.5 in the current filter. These values are typically 0.1 to 0.2 arcsec; values approaching 1 arcsec are evidence for a failed astrometric solution.

The top panel of Figure 6 shows the DAOphot fitting parameter *sharp* for each image. This parameter is an index of the sharpness of stellar images; values with magnitudes more than 1 typically indicate some serious problem with the imaging.

The bottom panel of Figure 6 shows the DAOphot fitting parameter *chi* for each image. This parameter is an index of the quality of the fits to image PSFs; the median values over all stars brighter than 12.5 (which is the plotted quantity) generally clusters near +1.0. Occasional values near zero or +2 may occur in images that have few bright stars. More extreme values than these indicate errors in the star image fitting.

Figure 7 shows the first plot from bad_hand on a night with sporadic cloud. The diff SDSS plot is irregular both in its zero point and in the sizes of the excursions during the normal filter cycle. The Sky Brightness plot shows moonrise at about JD 0.67, followed by a period of smoothly increasing brightness, and then large and irregular fluctuations between JD of 0.8 and 0.95, presumably caused by clouds. The string of diamond symbols near JD = 0.83 indicates images that have been flagged as bad.

12.2. Plots from fitstds

The routine fitstds uses routine resid_plot to produce 5 plots, one for each of the filters $\{g, r, i, z, D51\}$. Each plot contains 3 panels, in which residuals about the best-fit extinction curve are plotted against airmass (top panel), $(g - r)$ color (middle panel), and time in hours after local noon (bottom panel). Each residual corresponds to a single observation of a single standard star (ie, one in the tstds database tile), after correction using the atmospheric extinction parameters for that block (night) of data. Data points that contributed to the extinction fit are plotted with "+" symbols; those that were excluded from the fit are plotted as points. A representative set of plots from fitstds and filter g is shown in Fig 8. The other filters look similar.

In the fitstds plots, data that are affected by clouds or that are otherwise bad show up as increased scatter in the residuals, often with one or more groups in time or airmass showing systematic offsets away from residual values of zero. In these cases, the plot against $(g - r)$ color often appears bloated or even bimodal. Figure 9 shows a typical bad night of g-filter data, when passing cirrus caused a few tenths of a magnitude of time-varying extinction.

12.3. Plots from plot_refvsx

The routine plot_refvsx also generates 5 plots for each block (night) of data, one plot for each filter. Figure 10 shows a representative (not particularly good) plot for the g filter. Each plot has 2 panels. The plot titles indicate which filter was used, and also a block number, which is related in an obscure way to the date on which the data were taken. The mapping between dates and block numbers is given in the file "processing.txt", which is described in Appendix 3, and found in www.cfa.harvard.edu/kepler/kic/ancillary_files.

Both panels are based on the magnitudes of a sample of typically 20 bright stars that are selected because they appear on every one of the standard-field images in which they lie. By working on this unvarying sample of stars, I avoid errors resulting from stars (with standard magnitudes that may be systematically erroneous) occasionally dropping out of the sample. The upper panel of each plot shows the median of the (instrumental - standard) magnitude, taken over the sample of good stars, as a function of the mean airmass (averaged over stars) for each observation. On these plots, cross symbols overlain with diamonds indicate points observed east of the meridian; crosses overlain with triangles indicate points observed west of the meridian. Vertical error bars show the $\pm$ 1 quartile points for the residual measurements at each time. The dashed line shows the result of a robust straight-line fit of observed magnitude against airmass; the fit coefficients and their formal errors are displayed in the plot's upper lefthand corner.

The lower panel of each plot shows the residuals about the extinction fit just described, plotted as a function of the FWHM of the stellar images measured in binned (2x2) pixels. For the KeplerCam, a binned pixel projects to 0.672 arcsec on the sky.

12.4. Quality of fits to the photometry (kic_qc3)

After the stellar parameter estimation code has been run on all tiles, three routines kic_qc1, kic_qc2, and kic_qc3 are run to provide quality-control information regarding the fitting process and the status of the catalog as a whole. The numbering of these routines is historical; for explanatory purposes it makes sense to discuss them in the order kic_qc3, kic_qc1, kic_qc2. Routines kic_qc1 and kic_qc3 produce one set of plots for each database tile; kic_qc2 makes a set of plots that relate to the entire Kepler field. All of these plots are available at www.cfa.harvard.edu/kepler/kic/kic_qc/.

Routine kic_qc3 helps to determine whether the model photometry implied by the derived stellar parameters is an accurate representation of the observed photometry. One passes it a list of tiles to be processed. For each tile, it retrieves the extinction-corrected stellar magnitudes for all stars from the stars database tile, and the stellar parameters $\{T_{eff}, \log(g), \log(Z), E_{B-V}\}$ from the corresponding properties database tile. It then selects stars for which the properties fit converged (as indicated by a small value for the "err_logg" entry in the properties database), and for which all of the magnitudes $\{g, r, i, z, D51, J, H, K\}$ have valid values. For each star in this set, colors_blk_intrp and red_vector are called to generate model colors. The routine then forms the difference between the observed colors and the model ones. It then generates three plots, each with two panels; examples of these are shown in Figs 11-13. The title of each panel gives the name (hence coordinates) of the tile being plotted, and the galactic latitude of the tile. Each of the 6 panels shows the differences plotted as points on a different color-color diagram (eg $(r - i)$ vs $(g - r)$, or $(J - K)$ vs $(g - i)$). The colors are chosen so that every filter is represented at least once. Of interest are the centroids of the plotted clouds of points, their scatter in each color, and the presence or absence of correlations between the residuals in different colors.

Inspection of the plots suggests several general conclusions. In the mean, the fitting process appears to be working well: for tiles at relatively high galactic latitude (7 degrees and above), the residuals have nearly zero mean in all colors, with distributions that are more or less Gaussian. Some colors have significantly more scatter than others – this is particularly true of any color involving a 2MASS filter, and z as well appears to be unusually noisy. Also, there are clear correlations (usually anti-correlations, in fact) between several pairs of colors. Notably, $(r - i)$ and $(i - z)$ are anti-correlated, as are $(i - z)$ and $(z - J)$, and also $(J - H)$ and $(H - K)$. Further analysis indicates that these correlations arise from two separate processes, one of which affects exclusively the 2MASS magnitudes, while the other is exclusively connected with the visible-light data. Beyond this, the causes of these correlations are as yet unknown. Tiles at very low galactic latitudes (less than 7 degrees) typically show much larger scatter than higher-latitude tiles, and also significant displacement of the mean residuals from zero. My working hypothesis is that these effects arise from large and spatially nonuniform interstellar extinction. Proof for this idea is lacking, however.

12.5. Comparison of stellar parameters with stars in M67

An important test of the analysis procedure is to compare fitted parameters for a group of stars with those that can be reliably estimated using other means. One group of stars for which this comparison can be done with confidence is the cluster M67. For this cluster D. Latham and S. Meibom provided me a list of 116 stars that are thought to be single members of the cluster, along with spectroscopically- based estimates of their T_{eff} and $\log(g)$, assuming a near-solar metallicity. Figure 14 shows the T_{eff}-$\log(g)$ diagram based on Meibom's spectroscopic results. The main sequence, turnoff, subgiant and giant domains are clear delineated. Figure 15 shows the comparison between Meibom's T_{eff} values for M67 and those provided by the KIC analysis for the same stars. (Note that M67 lies far outside the Kepler field of view, so these stars do not appear in the KIC10 proper.) With the exception of one extreme outlying star, the agreement is generally good. The RMS difference between the measurements is about 150 K, and the systematic differences appear to be small. Figure 16 shows the comparison between Meibom's $\log(g)$ measurements with those using the KIC analysis. Again the general agreement is good, with an rms difference between the meaturements of about 0.4 dex. Systematic differences are discernable in $\log(g)$, however. The most significant of these is a tendency for subgiants in Meibom's analysis to appear as main-sequence stars in the KIC analysis. This is not surprising, since in the T_{eff} range where the turnoff to the subgiant branch occurs in M67, none of the filter combinations that were observed for the KIC are sensitive to gravity. In particular, $(g - D51)$ shows a useful gravity dependence only for $(g - r)$ colors greater than 0.65, whereas the main-sequence turnoff in M67 lies at roughly $(g - r) = 0.38$. Thus, I believe the performance of the KIC analysis on M67 is about as good as can be expected, and for the purpose of separating GKM dwarfs from giants at (say) the 95% level, it is satisfactory.

12.6. Giant-Dwarf separation in color-color space (kic_qc1)

The routine kic_qc1 create plots that help diagnose the performance of the star classifier in separating giants from dwarfs, and that also allow simple sanity checks on the behavior of several parts of the code. This routine produces five plots, each with two panels. Examples of these appear in Figures 17-21. Each plot represents one database tile, covering a region on the sky spanning 1 degree in RA by 1 degree in Dec. The title of each plot identifies the southwest corner of the tile, and also the galactic latitude of this point, in degrees. The first plot shows the $(g - D51)$ vs $(g - r)$ color-color diagram for one tile. In the top panel, stars classified as giants (with $\log g$) < 3.6) are indicated by heavy sympbols, and dwarfs (with larger $\log(g)$ are plotted as light symbols. In the lower panel, this convention is reversed. In this plot, main-sequence stars occupy a locus that trends lower left to upper right as T_{eff} falls, until $(g - r) \approx 0.65$ is reached. At this point the $(g - D51)$ colors turn sharply blueward as $(g - r)$ continues to get redder. At $(g - r) \approx 1.0$ this trend reverses and the $(g - D51)$ color reddens rapidly. The result is that main-sequence stars form a sickle shape: hot stars form the handle of the sickle, and cooler ones form the blade. The area inside the blade of the sickle (and particularly the upward extension of the handle) is occupied by stars with lower surface gravity, or with low metallicity. In these stars the Mg b lines are weak compared to what one sees on the main sequence, resulting in more flux in the $D51$

band, and a redder $(g - D51)$ color. Most often, the kic_qc1 plots show a clear demarcation between the dwarf- and giant-star regions, with few stars having contrary classifications appearing in either region.

Figure 18 shows the second plot created by kic_qc1. This is similar to the first, except that it plots $(J - K)$ against $(g - i)$. As is well known (eg Bessell & Brett (1988)), color-color diagrams involving $(J - K)$ bifurcate for M-type stars, with main-sequence stars limited to $(J - K)$ colors smaller than about 0.9, while low-gravity (and also low metallicity) stars continue to grow redder with lower T_{eff}. Again, most plots show a clean separation between dwarfs and giants on this diagram. Note that both plots 1 and 2 use dereddened colors, where the reddening $E_{B-V} = A_V/3.1$, and the reddening in the colors we actually compute is taken to be a constant multiplier times A_V, given in routine red_vector.

Figure 19 shows a plot of $\log(g)$ against T_{eff} in the top panel, and extinction vs g magnitude in the bottom. These are mainly useful to assure that the code did not go crazy and produce completely unreasonable results. The top panel should look a bit like an HR diagram (remembering that $\log(g)$ is almost constant on the main sequence). The lower panel typically shows a hard upper limit on A_V, which is the maximum extinction allowed by the dust model at the galactic latitude of the plotted tile. Apparently faint hot stars appear in the upper right corner of the populated area, since their distance must be large. Faint cool stars may still be nearby, and the sloping right-hand edge of the populated area reflects the relation between temperature and distance at the limiting magnitude of the plot (which is set in r band, not g. The weak positive correlation between A_V and g simply says that faint stars are likely to be farther away than bright ones, hence more extincted.

The top panel of Figure 20 shows observed (not dereddened) $(g - r)$ as a function of the estimated T_{eff}. This relation is normally fairly tight, with $(g - r)$ varying by 0.2 mag or less for a given T_{eff}, and T_{eff} varying by perhaps 1000 K peak-to-peak at the hottest temperatures, and 300 K near $T_{eff} = 4000$K. Often there is an indication of bimodality, especially at the hotter temperatures. The bottom panel shows a histogram of the derived metallicity for the stars in the tile. This normally is strongly peaked near $\log(Z) = 0.2$.

The top panel of Figure 21 shows log(Radius) plotted against log(mass), both quantities being in solar units. This plot was originally intended as a trap for a bad build of the KIC analysis code, which gave all stars a mass of 1.0. The very thin giant branch is an artifact of the assumption that mass (and many other parameters) is a function of T_{eff} and $\log(g)$ only. The way this functional dependence is calculated has the effect of assigning pretty much all giants log(Mass) of about 0.2, as shown in the plot. The bottom panel shows the RA, Dec positions of stars in the tile. For most tiles this is relatively boring, but if there are missing pointings or a star cluster in the tile, it can become more interesting. This plot makes no distinctions among stars of different magnitudes; each one gets a point.

12.7. Catalog coverage and kic_qc2

Routine kic_qc2 makes four plots that display the density of stars in the Kepler field of view, and the mean (averaged over stars) colors as a function of position within the field, shown in three different visible-light and infrared colors. The titles describe what is being plotted. Spatial resolution in these plots is about 10 arcmin. these are displayed in Figures 22-25.

REFERENCES

Bessell, M. S., & Brett, J. M. 1988, PASP, 100, 1134

Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, ApJ, 345, 245

Castelli, F., & Kurucz, R. L. 2004, arXiv:astro-ph/0405087

ESA, "The Hipparcos and Tycho Catalogues", ESA SP-1200

Girardi, L., Bressan, A., Bertelli, G., & Chiosi, C. 2000, A&AS, 141, 371

Høg, E., et al. 2000, A&A, 355, L27

Lejeune, T., Cuisinier, F., & Buser, R. 1997, A&AS, 125, 229

Nordström, B., et al. 2004, A&A, 418, 989

Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525

Sekiguchi, M., & Fukugita, M. 2000, AJ, 120, 1072

Skrutskie, M. F., et al. 2006, AJ, 131, 1163

Smith, J. A., et al. 2002, AJ, 123, 2121

Stetson, P. B., Bruntt, H., & Grundahl, F. 2003, PASP, 115, 413

Stoughton, C., & Jester, S. 2004, AGN Physics with the Sloan Digital Sky Survey, 311, 475

This manuscript was prepared with the AAS LaTeX macros v5.0.

Figure Captions

Fig. 1.— The atmospheric extinction (combined with instrumental) zero points for the g and i filters, as estimated by the routine fitstds for 203 of the 205 nights on which KIC data were obtained. The remaining 2 nights gave values that are extreme outliers, falling outside the range of these plots. Note the strong color dependence of the temporal variations.

Fig. 2.— Variation in the k_i coefficient (extinction per unit airmass in the i^{th} filter), shown as a function of time, for nights for which reliable linear fits to the extinction could be obtained, during a part of one observing season. The time span chosen here was one of the most variable that we encountered in 5 seasons of observing, but is nonetheless typical in its variability within a factor of 1.5 or so. Filters g, r, i are shown in order from top to bottom of the plot. Night-to-night variations tend to be not only correlated, but of similar size among the filters.

Fig. 3.— Photometric repeatability after correction for atmospheric extinction for each filter, shown as a function of apparent magnitude in that filter. Filters are encoded as different plot symbols: "+"=g, "*"=r, (no symbol)=i, (triangle)=z, and (diamond)=$D51$.

Fig. 4.— First diagnostic plot created by routine bad_hand. The top panel shows, for each image, the difference between the observed magnitudes of 2MASS stars and the expected magnitudes of the same stars, extrapolated from their 2MASS J and K magnitudes. The regular sawtooth pattern represents the combination of systematic errors in the extrapolation and the repetitive cycling of filters that comprise observations at a given pointing. The bottom panel shows the sky brightness measured in the g and i filters, for each image taken using either of these filters. Both are scaled so that an image at the zenith on a moonless night yields a value of unity.

Fig. 5.— Second diagnostic plot created by routine bad_hand. The top panel shows, for each image, the number of 2MASS stars that should have been found in the field, but that failed to match with identified stars. Values greater than a few indicate a failure of the astrometric fitting. The bottom panel shows the rms difference in position on the sky (in arcsec) between stars in the image and matched 2MASS stars. Typical values larger than about 0.2 arcsec indicate trouble.

Fig. 6.— Third diagnostic plot created by routine bad_hand. The top panel shows, for each image, the median "sharpness" statistic computed by DAOphot in fitting the stellar PSFs. The expectation value is zero; absolute values much greater than unity indicate systematically bad stellar images. The bottom panel shows the median (over stars) of the "chi" statistic from DAOphot. The expectation value for this statistic is unity, though median values near zero are fairly common. Very large or small values indicate a failure to fit star images accurately.

Fig. 7.— First diagnostic plot created by routine bad_hand, on a night corrupted by variable cirrus. Note the variations in zero point and in local scatter in the "diff SDSS" plot, and the sudden sharp jumps in the brightness of the moonlit sky. The erratic gaps in the data also suggest that the telescope operator was periodically closing down because of obvious clouds.

Fig. 8.— Typical plot for one filter (g in this case) produced by routine fit_stds. All panels show the residuals after fitting an airmass model to the instrumental magnitudes for an entire nifht of data; each point corresponds to one standard star observed on one image. The top panel shows the residuals plotted against airmass of the observation, the middle panel against stellar $g - r$ color, and the bottom panel against time. Residuals plotted as crosses were used in the fits for extinction; those plotted as small points were discarded, either automatically (as a result of large deviation from the median) or via by-hand editing.

Fig. 9.— Plot of g filter data produced by routine fit_stds on a poor night, when extinction was significantly variable with time.

Fig. 10.— Plot of g data produced by routine plot_refvsx, on a representative (mildly non-photometric) night. The top panel shows extinction as a function of airmass. Points plotted as diamonds were obtained before the meridian transit of the standard star field; thost plotted as triangles were obtained after it. Vertical bars indicate the interquartile spacing of results from the 20 stars used to estimate the extinction. The diagonal dashed line is the result of a robust linear fit to the extinction values, with coefficients tabulated in the upper left corner of the plot. The bottom panel shows residuals around this fit plotted against the time-varying FWHM of the stellar point spread function.

Fig. 11.— First diagnostic plot created by routine kic_qc3. This plot represents all of the stars contained in one database tile, covering an area spanning 1 degree in RA by 1 degree in Dec on the sky. The title gives the location of the tile, and its galactic latitude b. Each point corresponds to one star, and the plotted positions show the residuals (in magnitudes) after subtracting the best-fit model from the photometry for that star, for two chosen colors (eg $g - i$ vs $g - r$, as in the top panel). The tile plotted here (t292+040) is fairly typical of areas in which the model fits are good. The RMS scatter in the residuals is .02 mag or less in all of $g - r$, $r - i$, and $g - D51$. Errors are anticorrelated between $g - r$ and $g - i$, but this tendency is more noticeable and has a different slope in the wings of the distributions than in the cores.

Fig. 12.— Same as Fig. 6, but showing $i - z$ vs $r - i$, and $z - J$ vs $i - z$. Note the larger scatter in these plots (especially in $z - J$, and also the strong negative correlations between these pairs of residuals.

Fig. 13.— Same as Fig. 6, but showing $J-K$ vs $g-i$ and $H-J$ vs $J-H$. Note the change in plot scale; the residuals in the IR colors are much larger than in the visible bands. The $J-K$ and $g-i$ residuals are almost uncorrelated, but one now sees a very significant offset from zero of the mean residual in $g-i$. The two IR colors in the bottom panel have only slightly correlated residuals, but the center of the $J-H$ distribution is also far displaced from zero.

Fig. 14.— Results of a spectroscopic analysis by Soeren Meibom of stars in M67, showing estimated $\log(g)$ vs T_{eff}. The main sequence, subgiant turnoff, and giant branch are all quite clear in these data.

Fig. 15.— Comparison between Meibom's spectroscopically-estimated T_{eff} values (horizontal axis) and those from the photometry-based maximum-likelihood analysis (vertical axis). Aside from one highly divergent point, the RMS difference is 160K.

Fig. 16.— Comparison between Meibom's spectroscopically-estimated $\log(g)$ values (horizontal axis) and those from the photometry-based maximum-likelihood analysis (vertical axis). The RMS difference is about 0.4 dex.

Fig. 17.— Plot number 1 created by kic_qc1, showing the $(g-D51)$ vs $(g-r)$ color-color diagram for one 1-degree-square tile (in this case, with its southwest corner located at $RA = 292°$, $Dec = +40°$). In the top panel, stars classified as giants (with $\log g) < 3.6$) are indicated by heavy sympbols, and dwarfs (with larger $\log(g)$ are plotted as light symbols. In the lower panel, this convention is reversed.

Fig. 18.— Plot number 2 created by kic_qc1. It shows a $(J-K)$ vs $(g-r)$ color-color diagram for the same tile as in the previous Figure. The meaning of the symbols is the same.

Fig. 19.— Plot number 3 created by kic_qc1. The top panel is a log(g) vs T_{eff} diagram for the same tile as in the previous Figure. In this plot the main sequence runs more-or-less horizontally across the bottom, and the giant branch trends to the upper right. The lower panel plots estimated A_V extinction plotted against g magnitude.

Fig. 20.— Plot number 4 created by kic_qc1. The top panel shows the relation between estimated T_{eff} and the observed (not dereddened) $g - r$ color. Usually this relationship is fairly tight, but it becomes fuzzy and acquires odd-shaped features in regions of large interstellar extinction. The bottom panel shows a histogram of the estimated log(Z) for the field. In keeping with the Bayesian prior, values near -0.2 are usually common, and large positive or negative values are rare. This plot hints at a tendency for strongly negative log(Z) to clump near half-integer values; this tendency can be confirmed by plotting results for more stars.

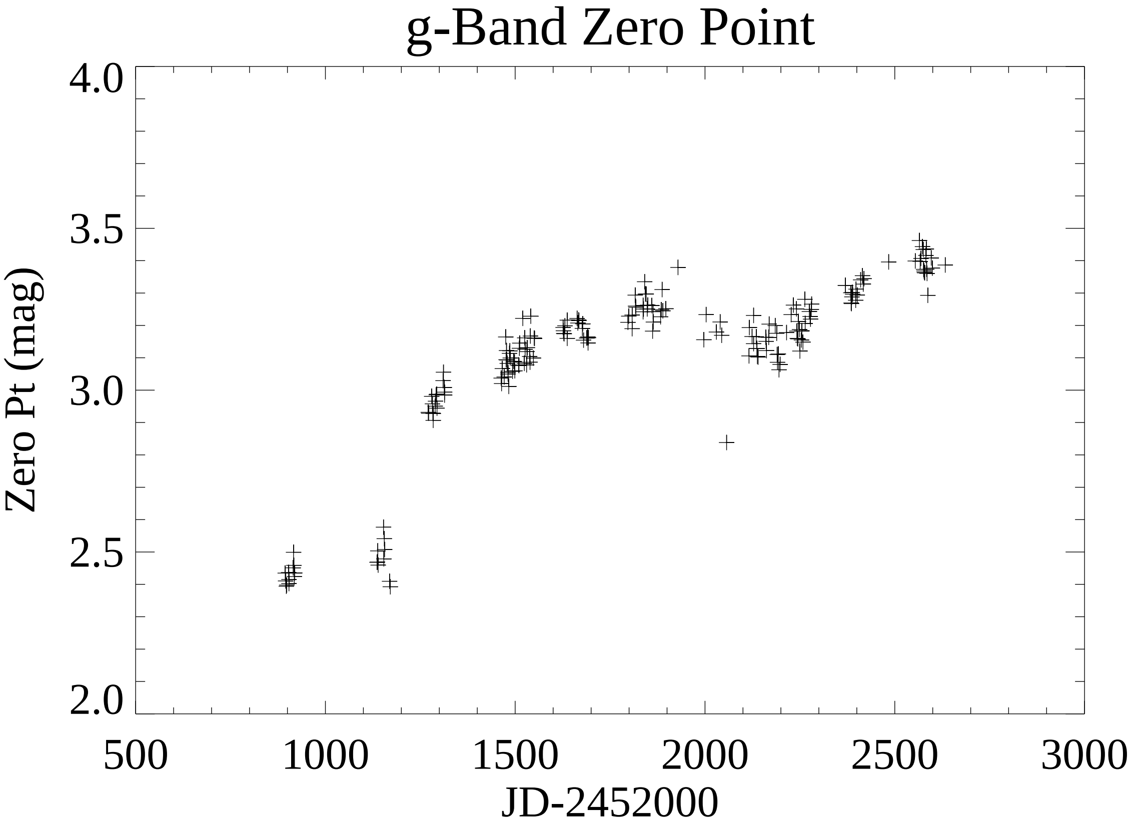
Fig. 21.— Plot number 5 created by kic_qc1. The top panel plots $\log R$ against log(M), where R is the stellar radius and M is the mass, both in solar units. The main sequence is the well-populated cloud of points trending from lower left to upper right. As explained in the text, the giant branch is the artificially-thin and nearly-vertical locus of points passing near $\log M \approx 0.2$. The lower panel shows the positions (RA and Dec) of each star in the tile. These are plotted one point per star, with no selection or indication of brightness or of anything but position.

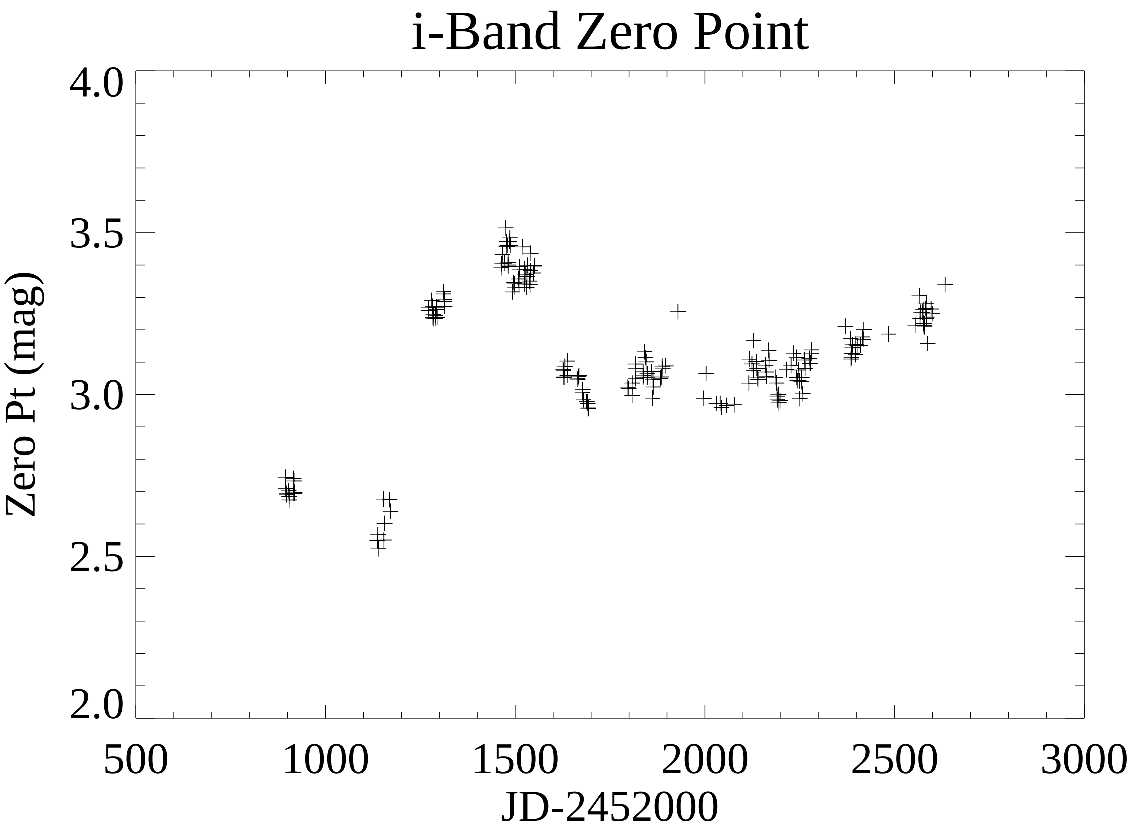
Fig. 22.— Plot number 1 created by kic_qc2. The quantity displayed is the number of stars per unit area within the Kepler field of view. Stars are counted if they have valid magnitudes in filters $\{g, r, i, z, J\}$ and if $r \leq 17$. Spatial resolution is 12 arcmin, and the range of densities plotted is between roughly 0 and 650 per resolution element; the densities in the well-observed part of the field range from about 100 to about 1000.

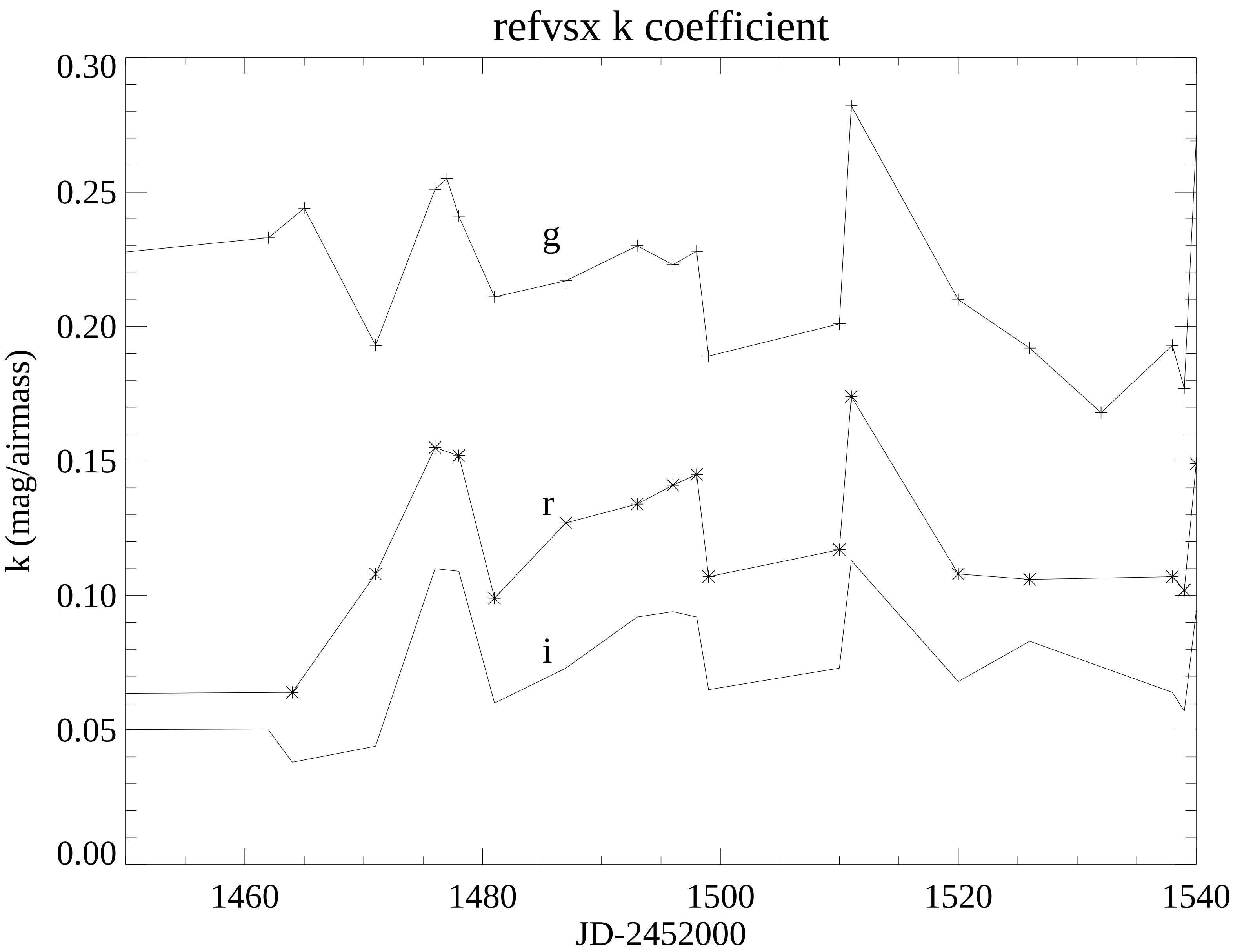
Fig. 23.— Plot number 2 created by kic_qc2. The plotted quantity is the $g - r$ color in each resolution element, with each star weighted equally (independent of brightness). The plotted range is [0.3,1.0].

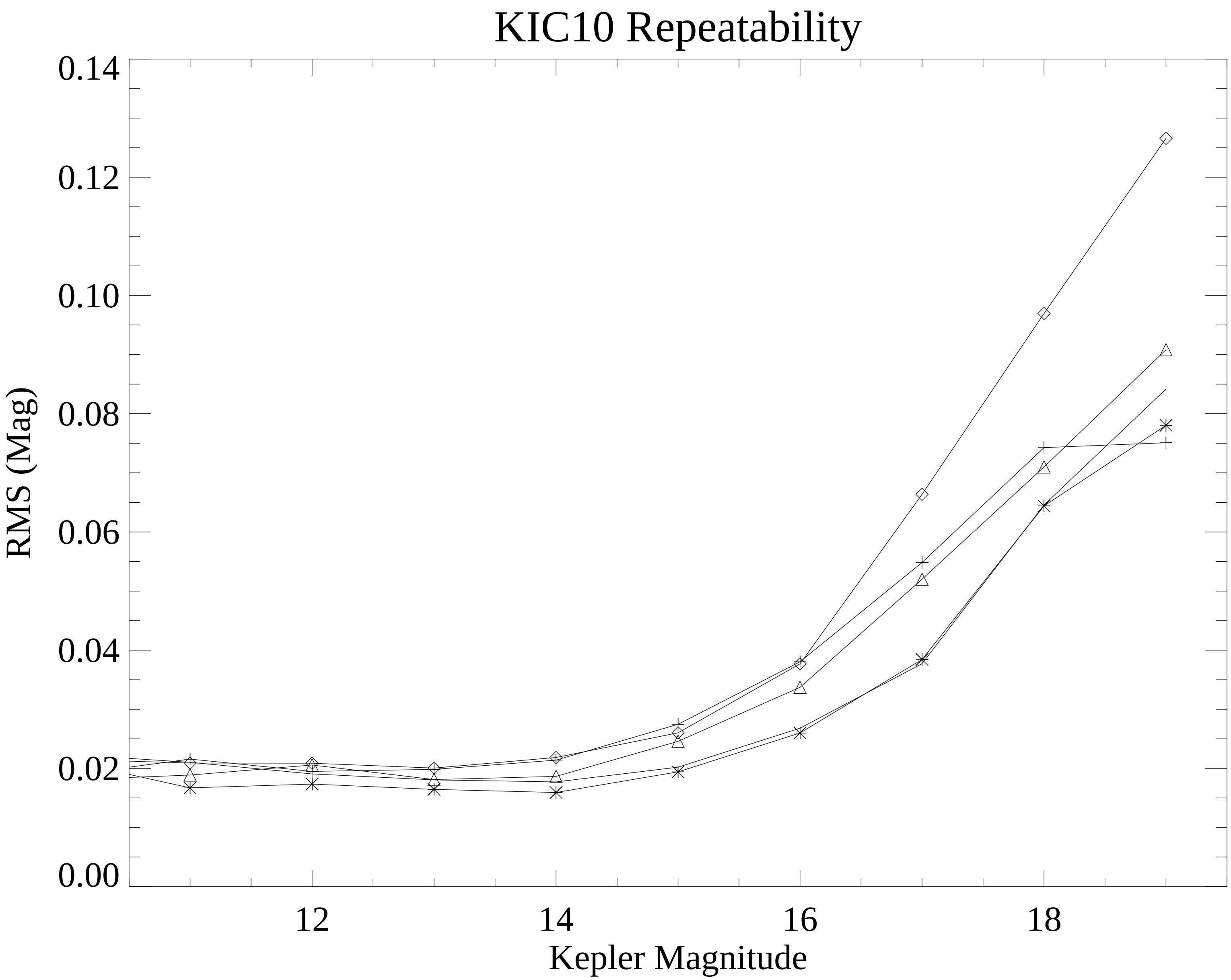
Fig. 24.— Plot number 3 created by kic_qc2. This is the same as the previous plot, except that the plotted quantity is the $i - z$ color, and the plot range is [0.,0.3].

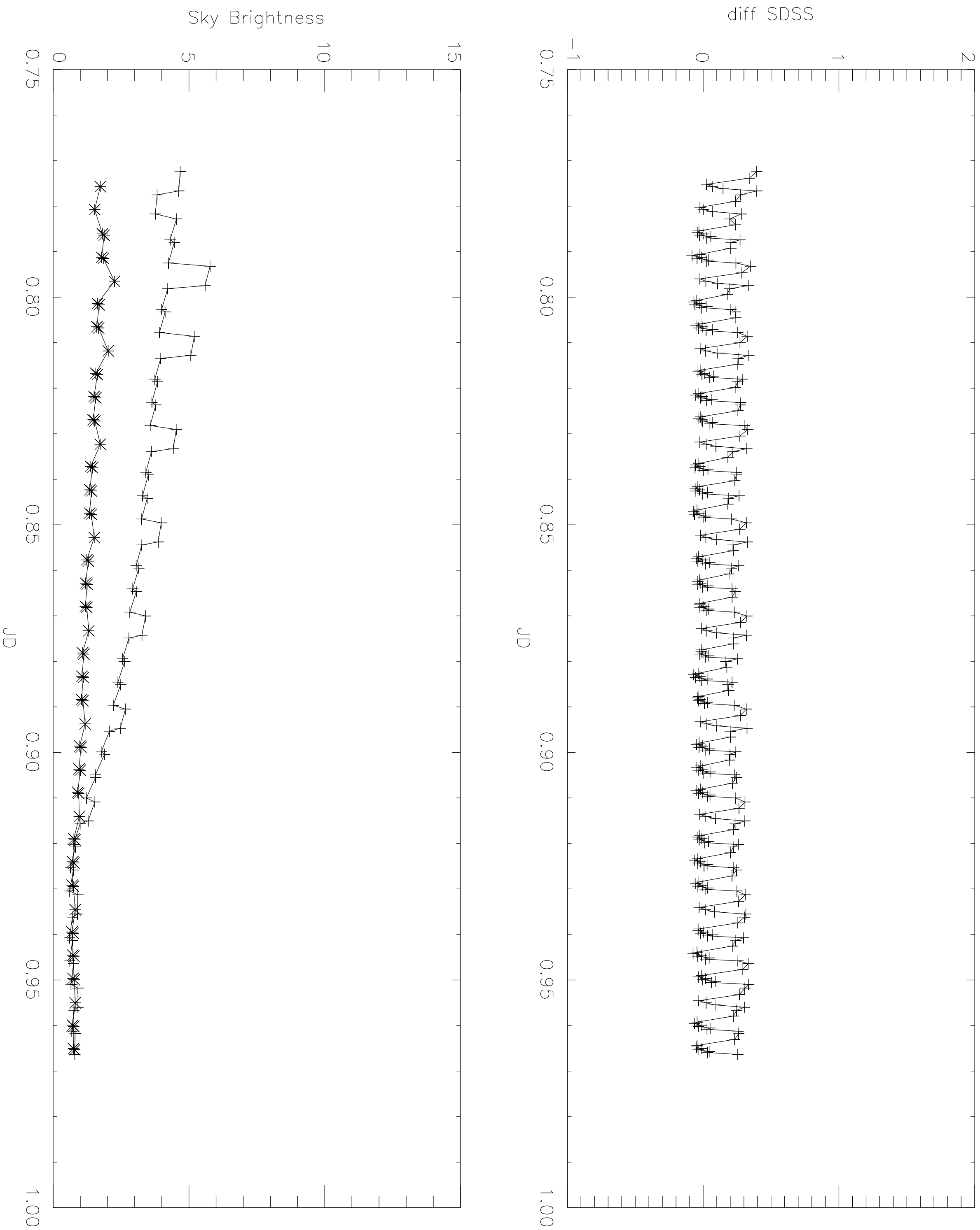
Fig. 25.— Plot number 3 created by kic_qc2. This is the same as the previous plot, except that the plotted quantity is the $J - K$ color, and the plot range is [0.3,0.7].

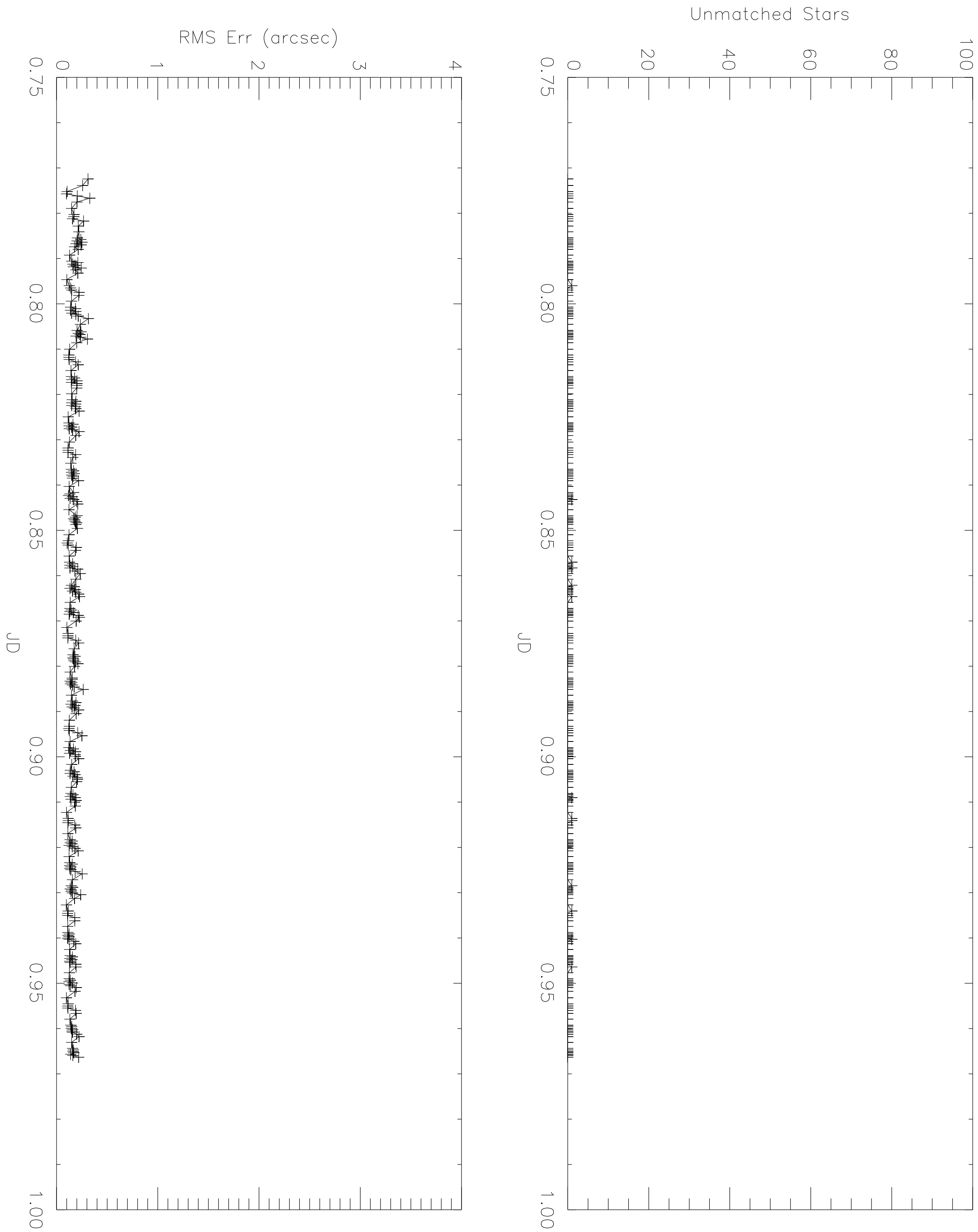

g-Band Zero Point
Zero Pt (mag)
JD-2452000
4.0
3.5
3.0
2.5
2.0
500
1000
1500
2000
2500
3000

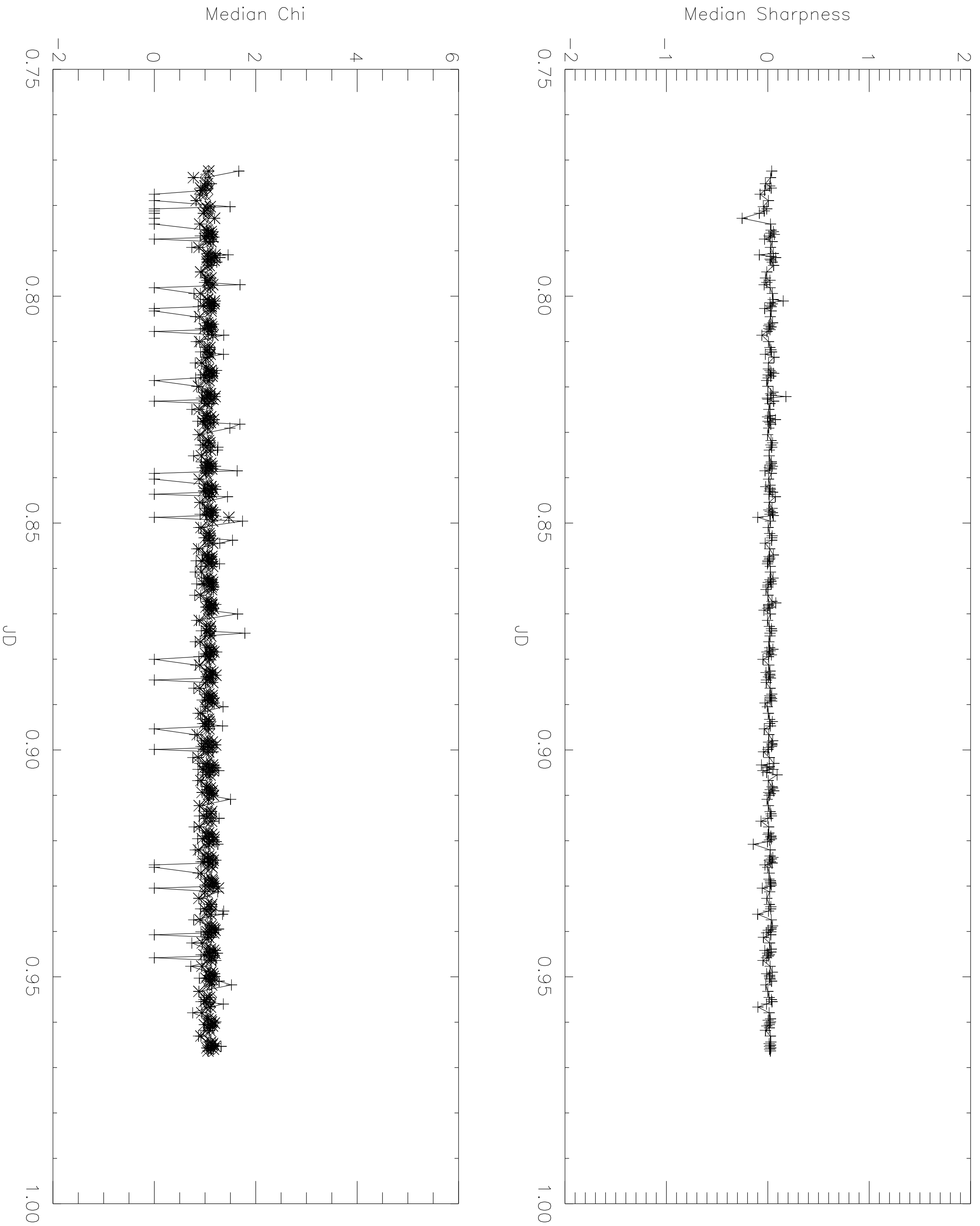

i-Band Zero Point
Zero Pt (mag)
JD-2452000
4.0
3.5
3.0
2.5
2.0
500
1000
1500
2000
2500
3000

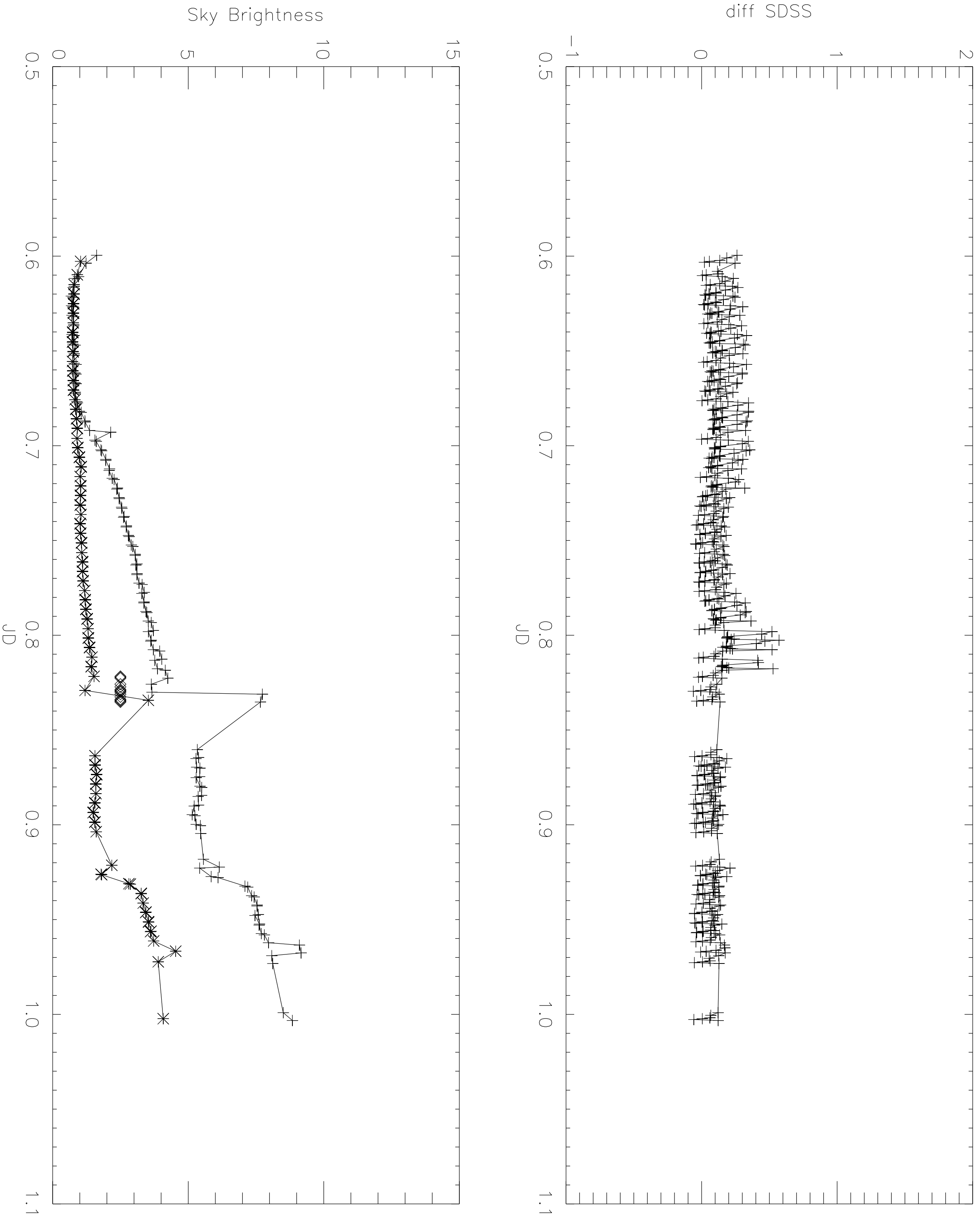

refvsx k coefficient
0.30
0.25
0.20
0.15
0.10
0.05
0.00
k (mag/airmass)
g
r
i
1460
1480
1500
1520
1540
JD-2452000

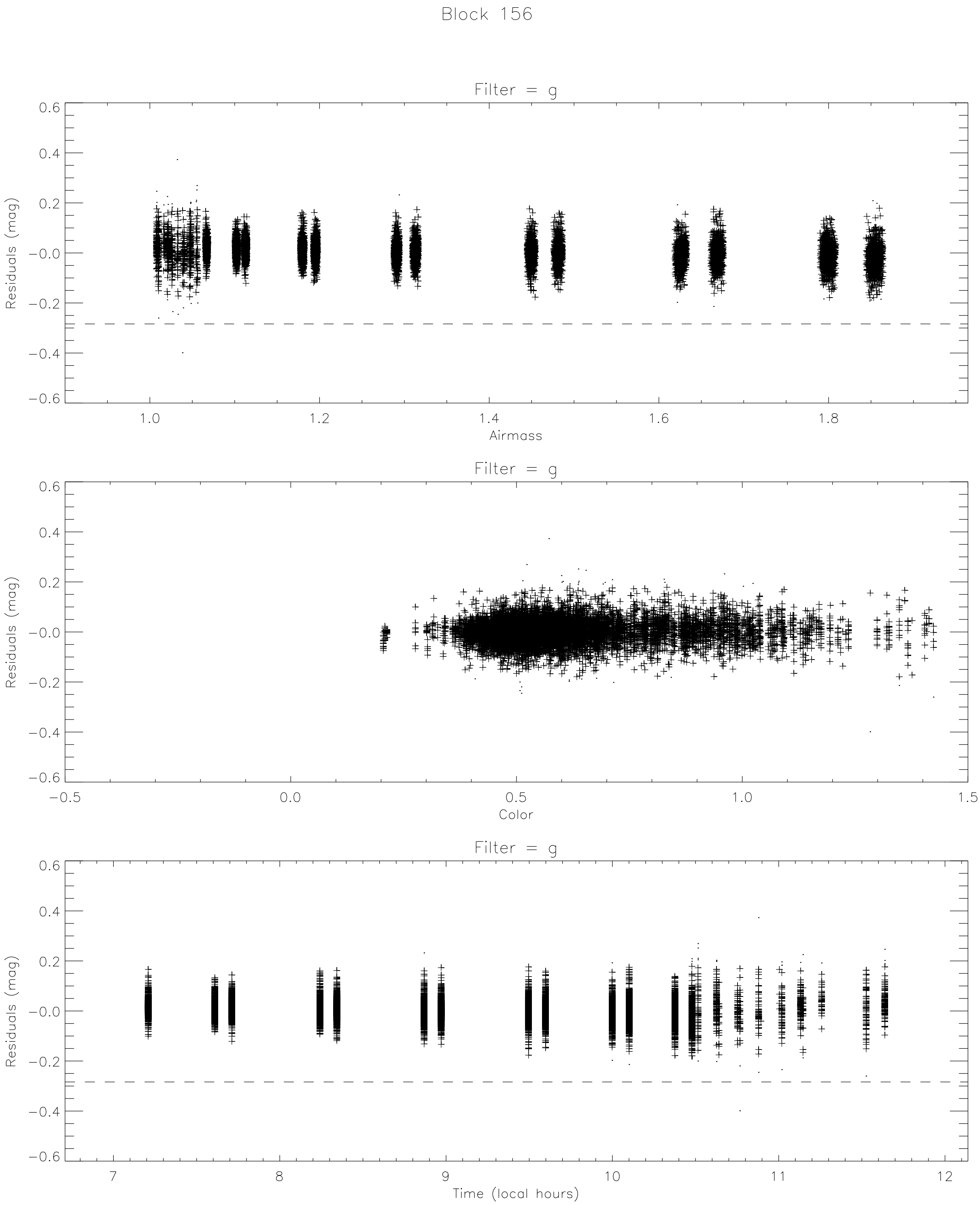

KIC10 Repeatability
RMS (Mag)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
12
14
16
18
Kepler Magnitude


diff SDSS
Sky Brightness
JD
JD
−1
0
1
2
0
5
10
15
0.75
0.80
0.85
0.90
0.95
1.00

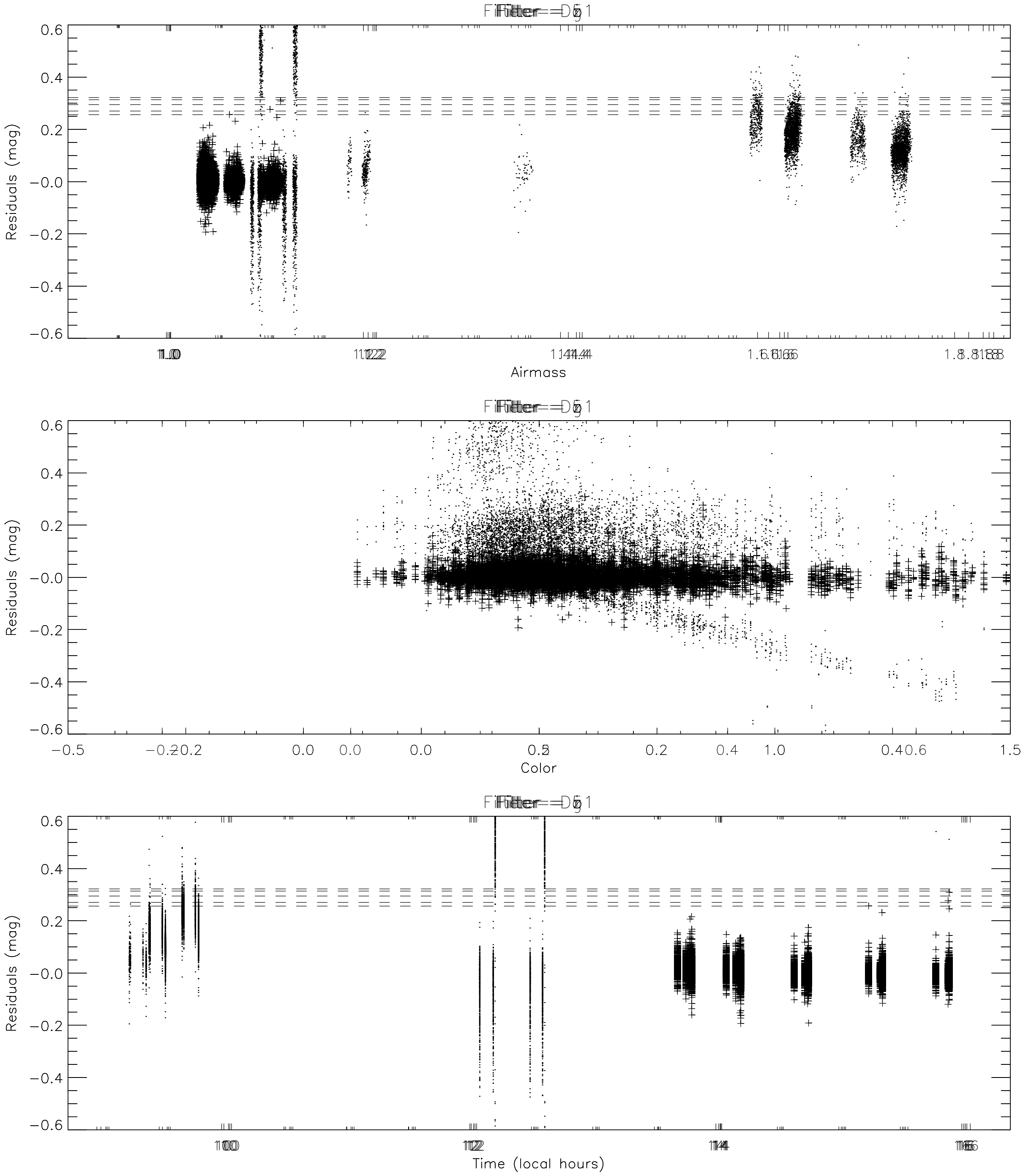

Unmatched Stars
RMS Err (arcsec)
JD
JD
0.75
0.80
0.85
0.90
0.95
1.00
0
20
40
60
80
100
0
1
2
3
4

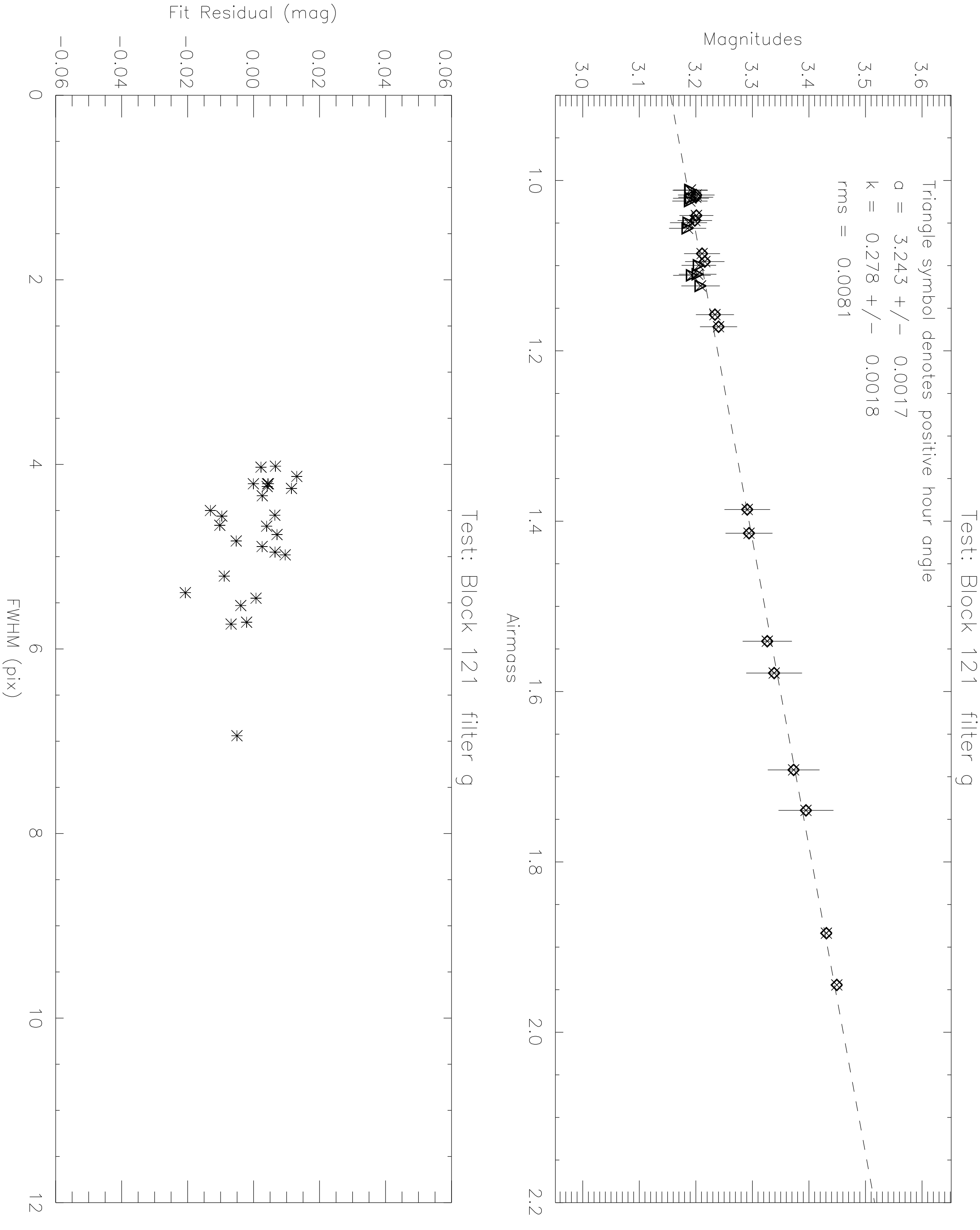

Median Sharpness
Median Chi
JD
JD
0.75
0.80
0.85
0.90
0.95
1.00
−2
−1
0
1
2
−2
0
2
4
6

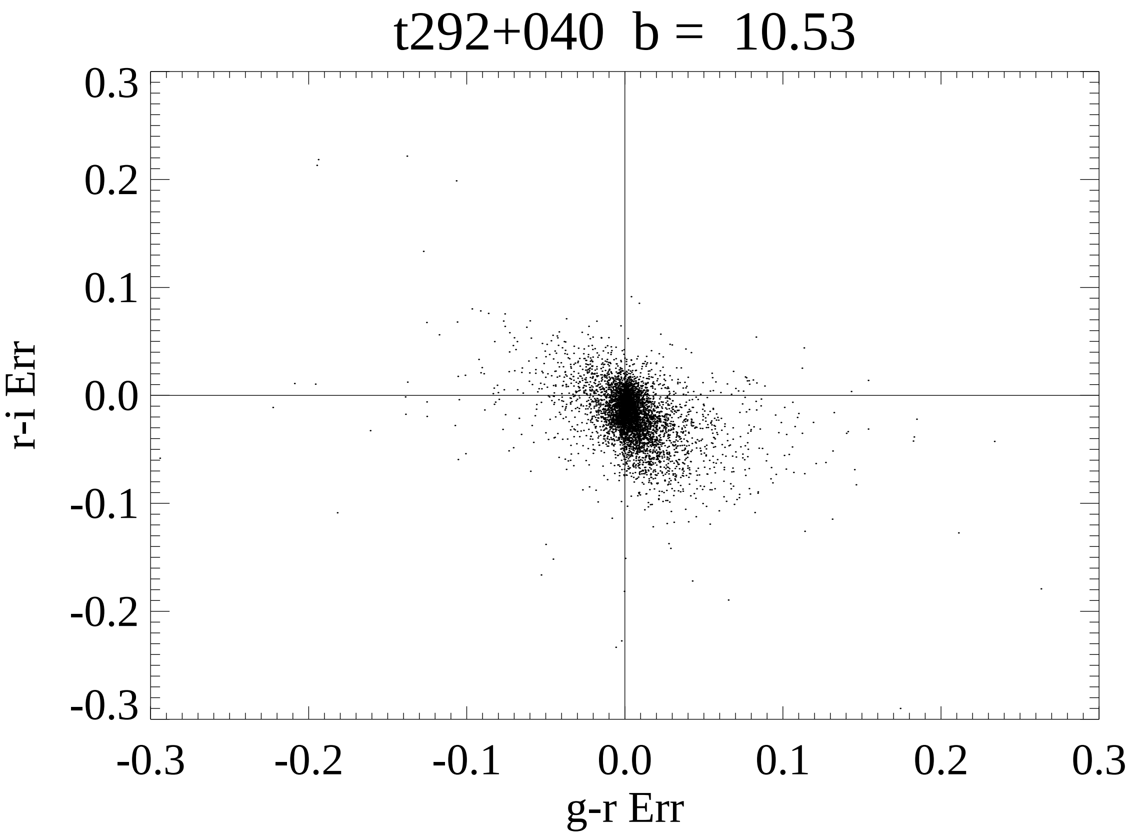

diff SDSS
Sky Brightness
JD

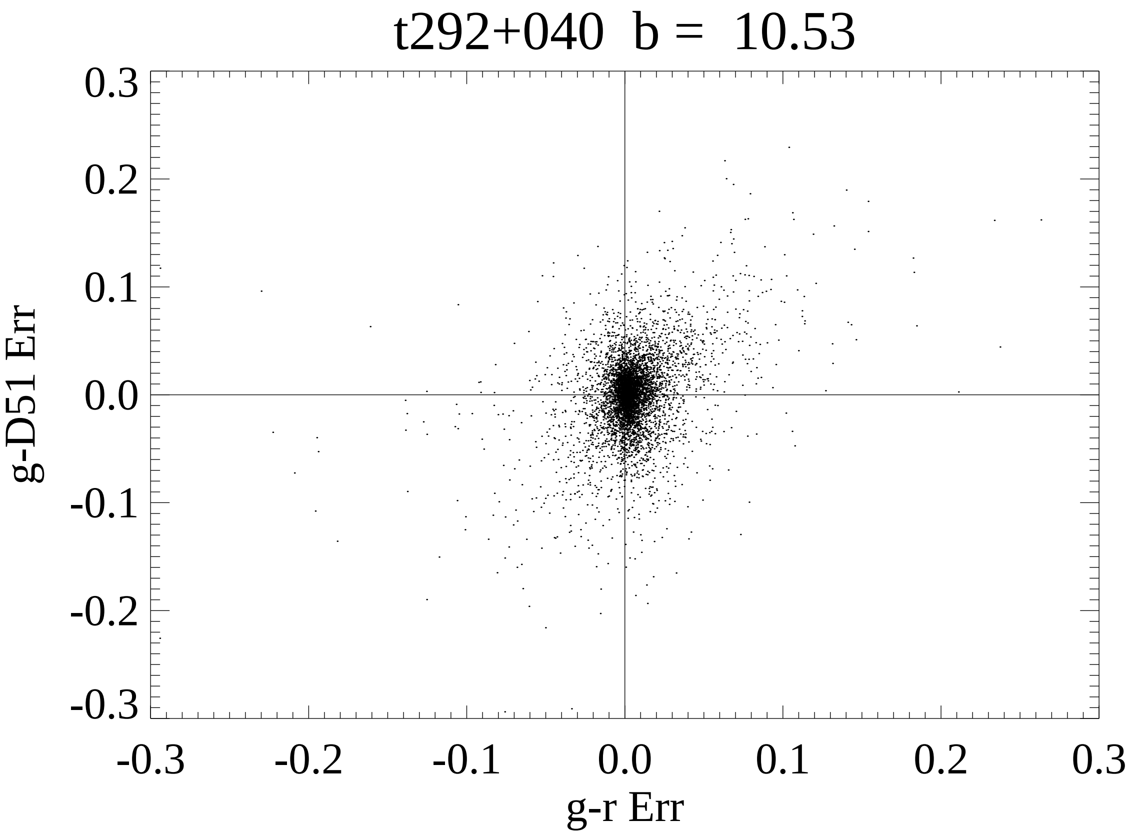

Block 156
Filter = g
Residuals (mag)
Airmass
Filter = g
Residuals (mag)
Color
Filter = g
Residuals (mag)
Time (local hours)

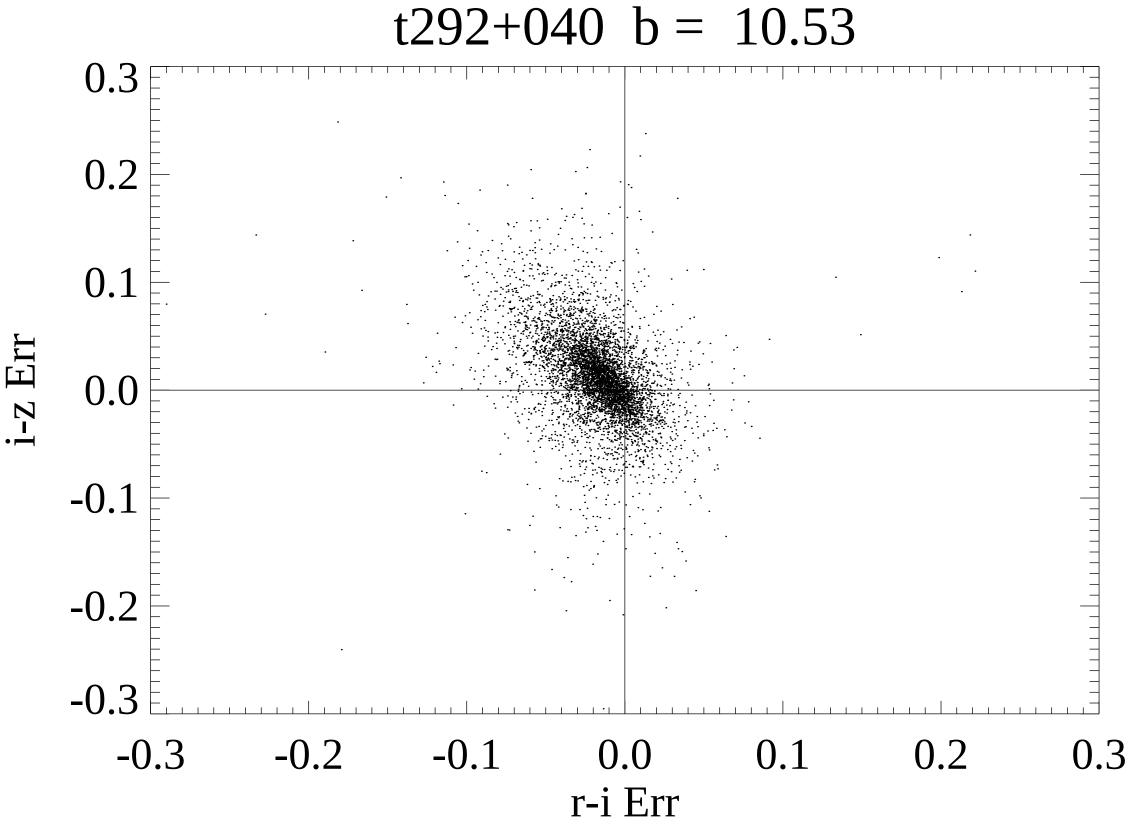
Block 150

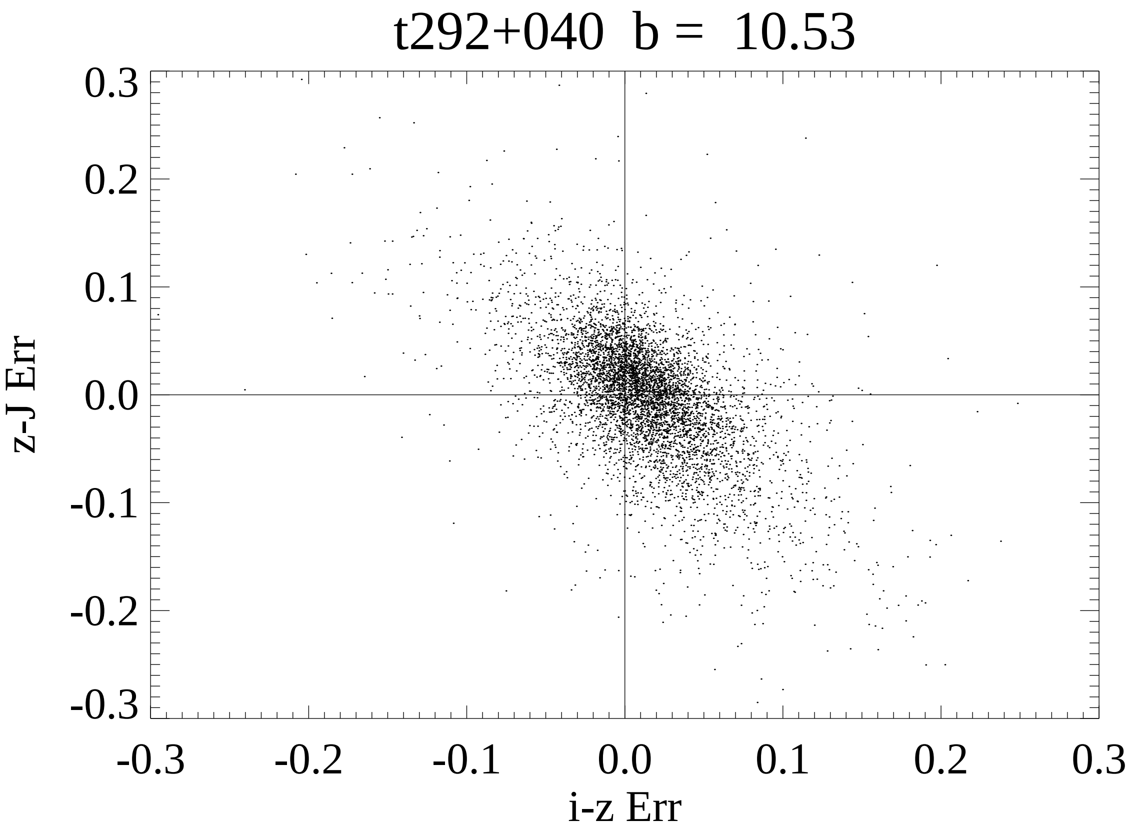


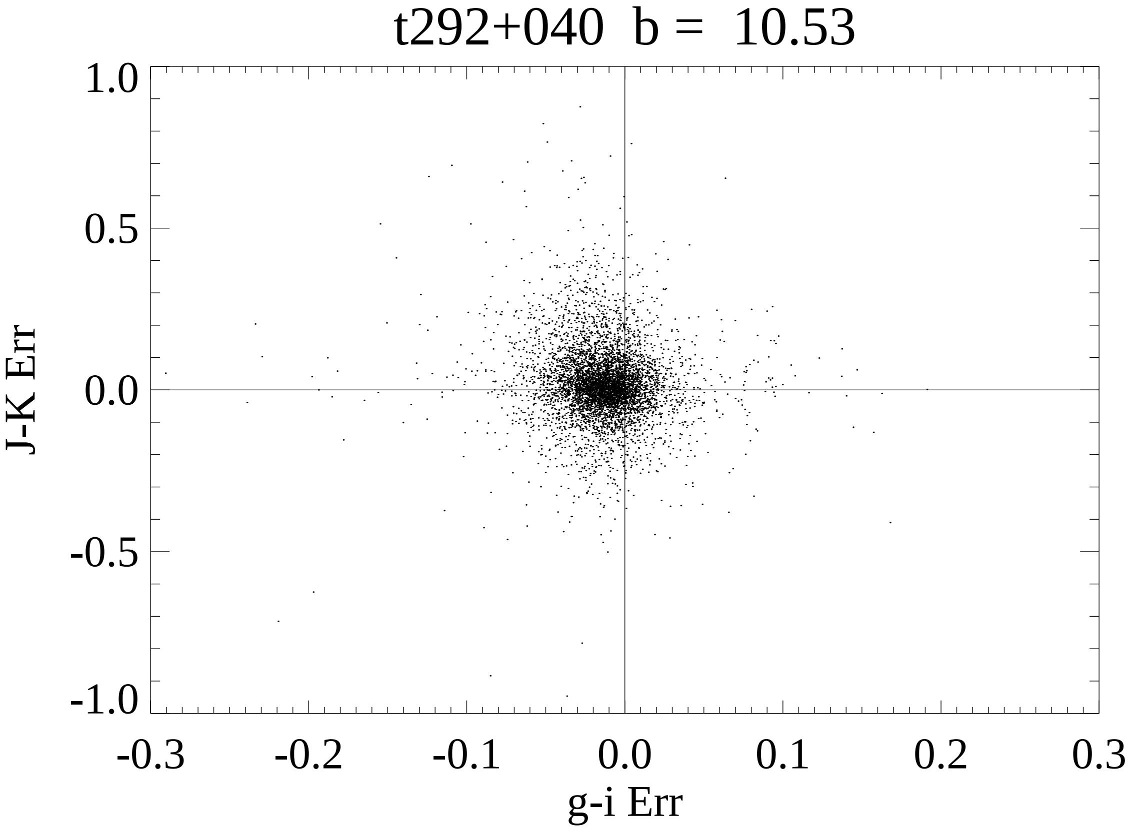

Test: Block 121 filter g
Triangle symbol denotes positive hour angle
a = 3.243 +/- 0.0017
k = 0.278 +/- 0.0018
rms = 0.0081
Magnitudes
Airmass
Test: Block 121 filter g
Fit Residual (mag)
FWHM (pix)

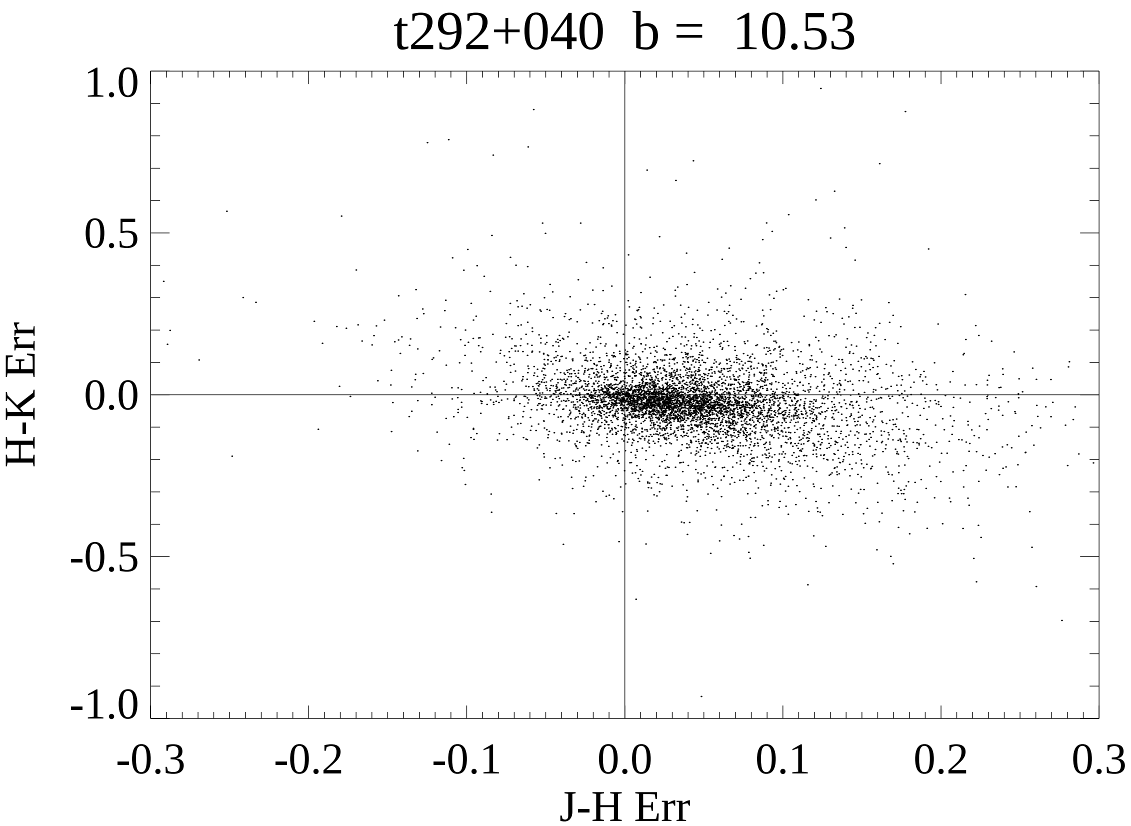

t292+040 b = 10.53
r-i Err
g-r Err
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3

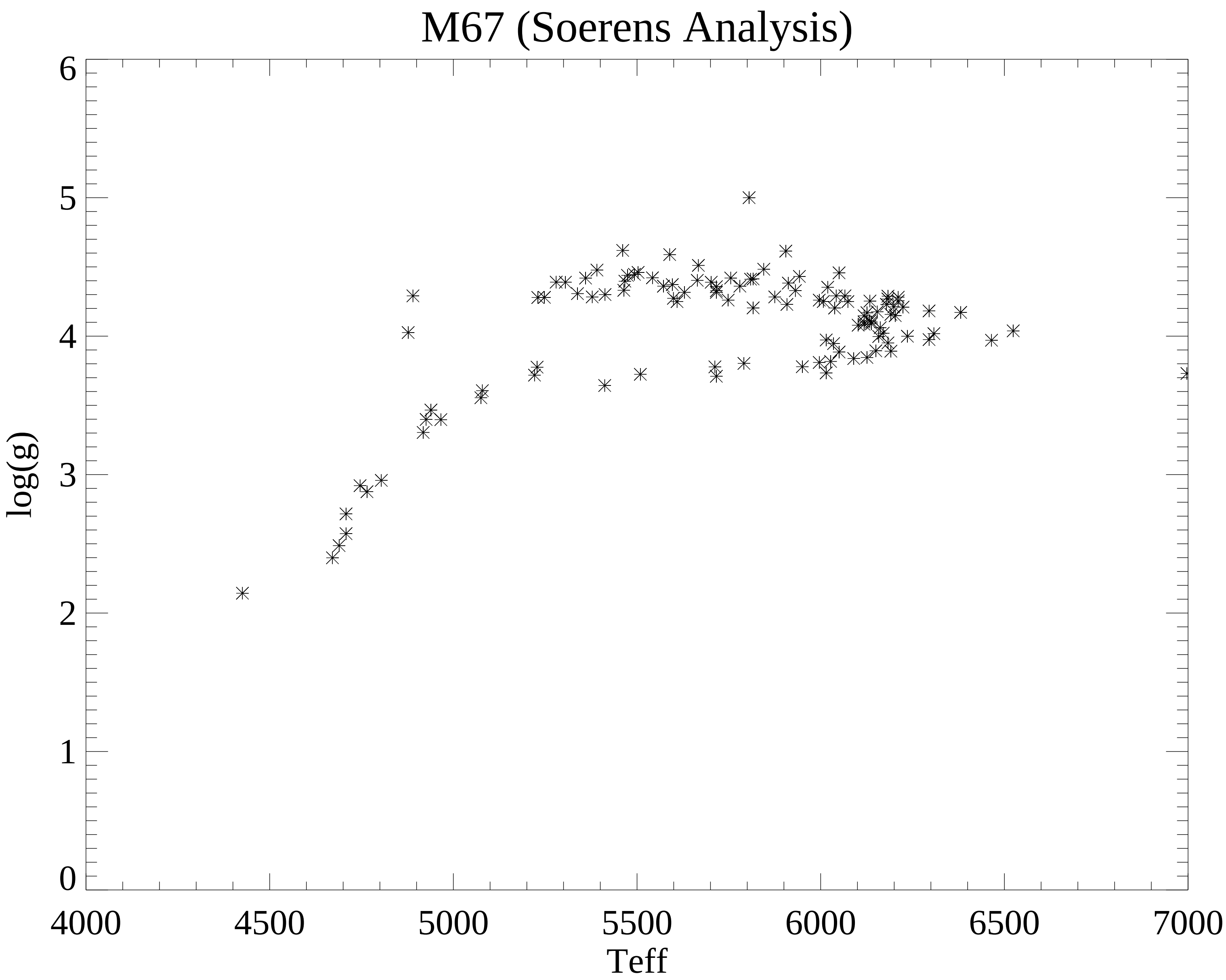

t292+040 b = 10.53
g-D51 Err
g-r Err
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3

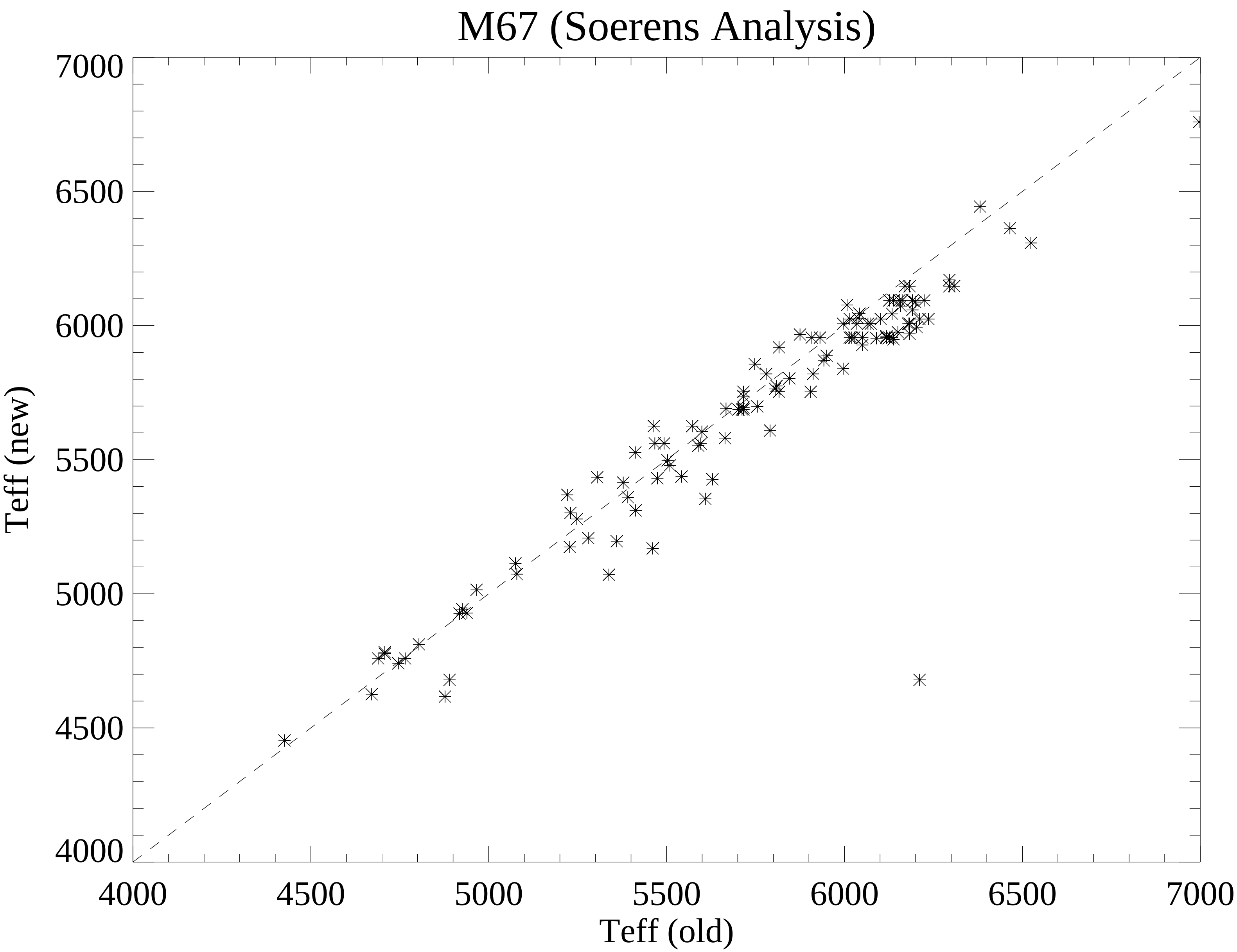

t292+040 b = 10.53
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
i-z Err
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
r-i Err

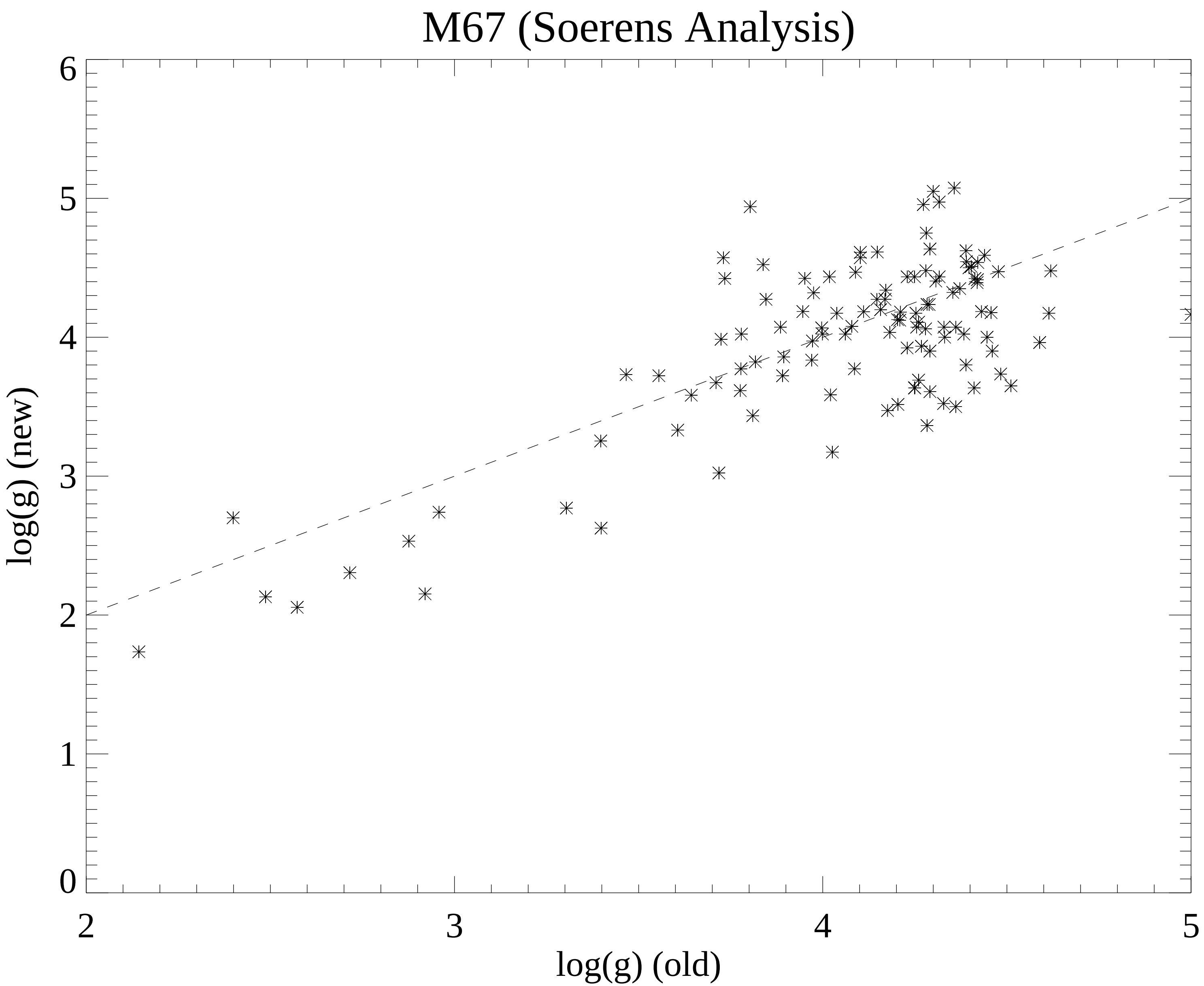

t292+040 b = 10.53
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
z-J Err
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
i-z Err

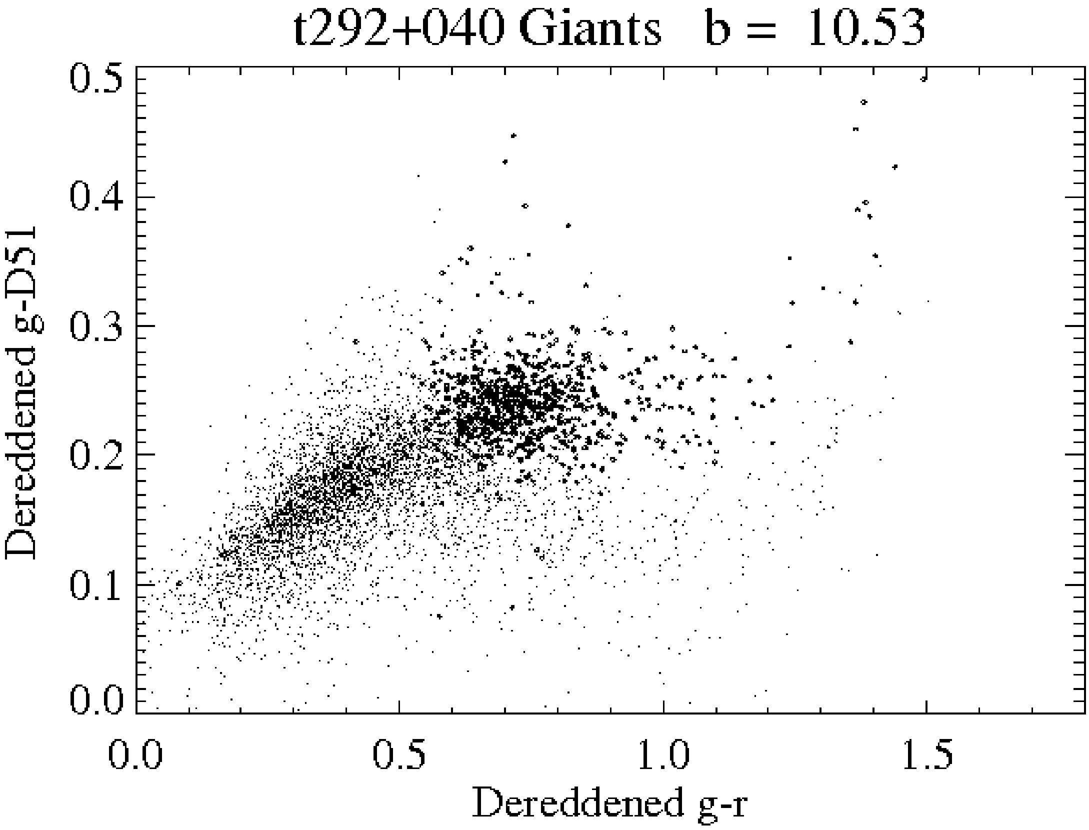

t292+040 b = 10.53
J-K Err
g-i Err
1.0
0.5
0.0
-0.5
-1.0
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3

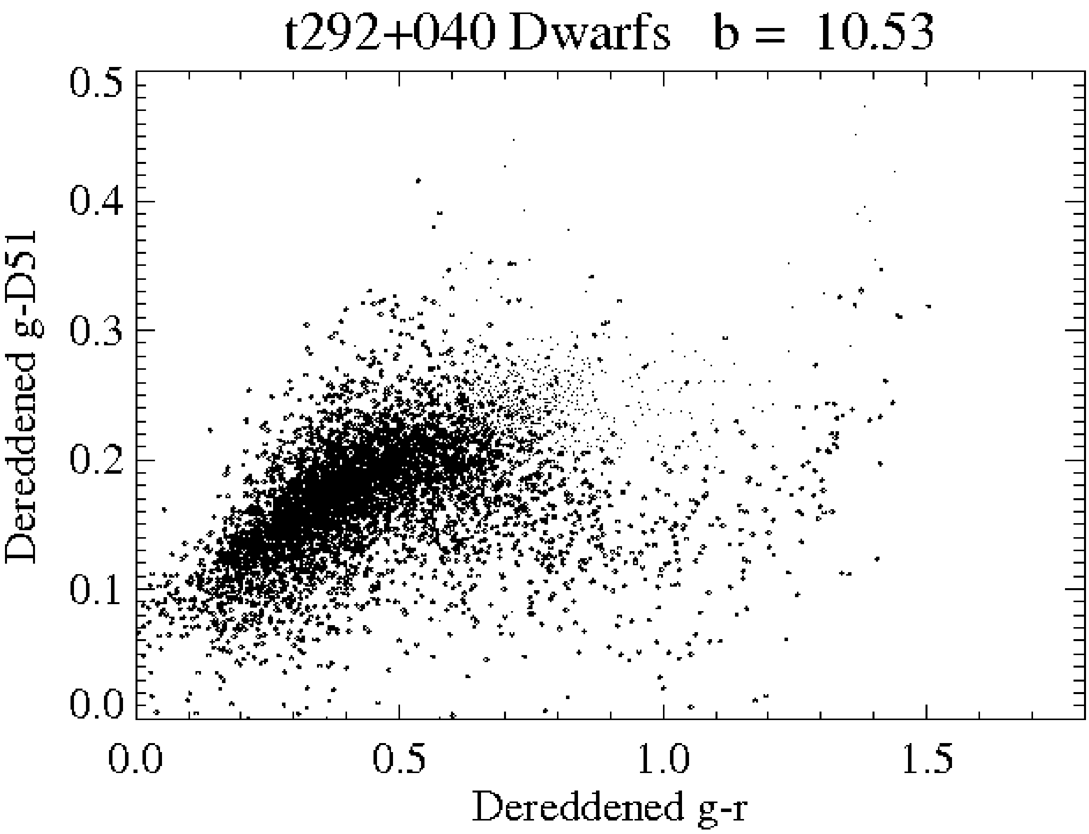

t292+040 b = 10.53
H-K Err
J-H Err
1.0
0.5
0.0
-0.5
-1.0
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3

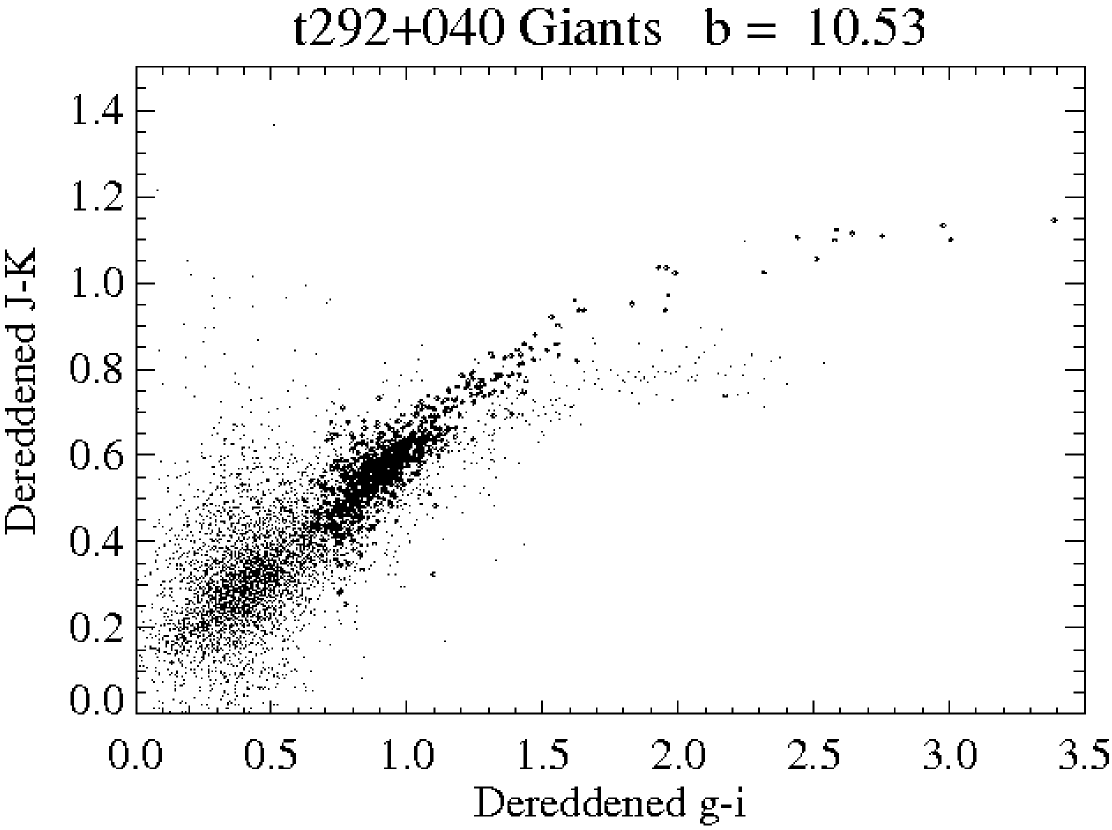

M67 (Soerens Analysis)
log(g)
0
1
2
3
4
5
6
Teff
4000
4500
5000
5500
6000
6500
7000

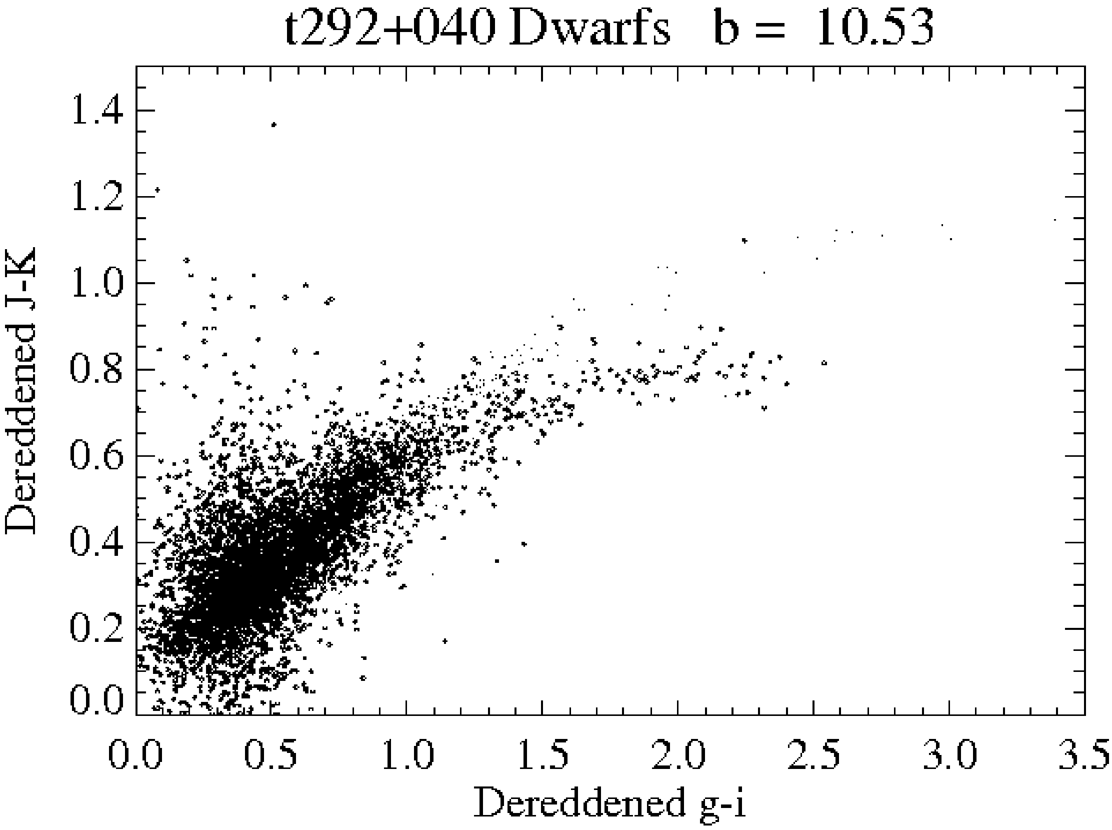

M67 (Soerens Analysis)
Teff (new)
Teff (old)
4000
4500
5000
5500
6000
6500
7000

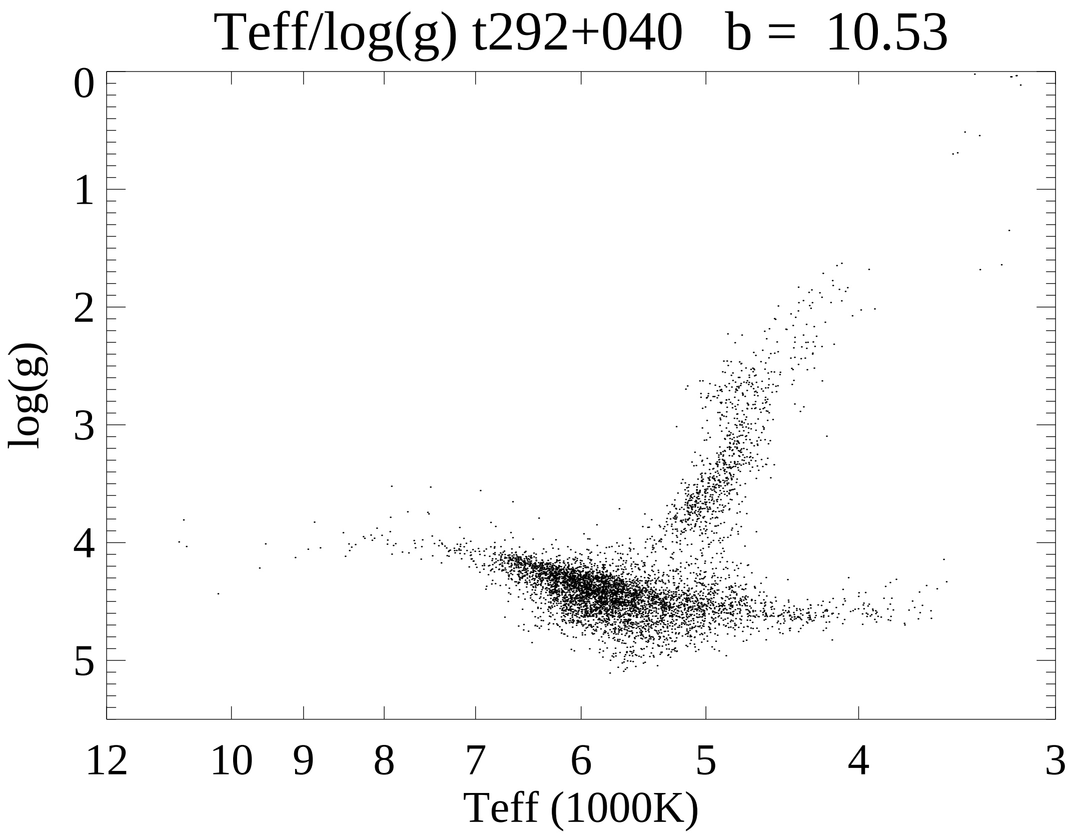

M67 (Soerens Analysis)
log(g) (new)
log(g) (old)
0
1
2
3
4
5
6
2
3
4
5

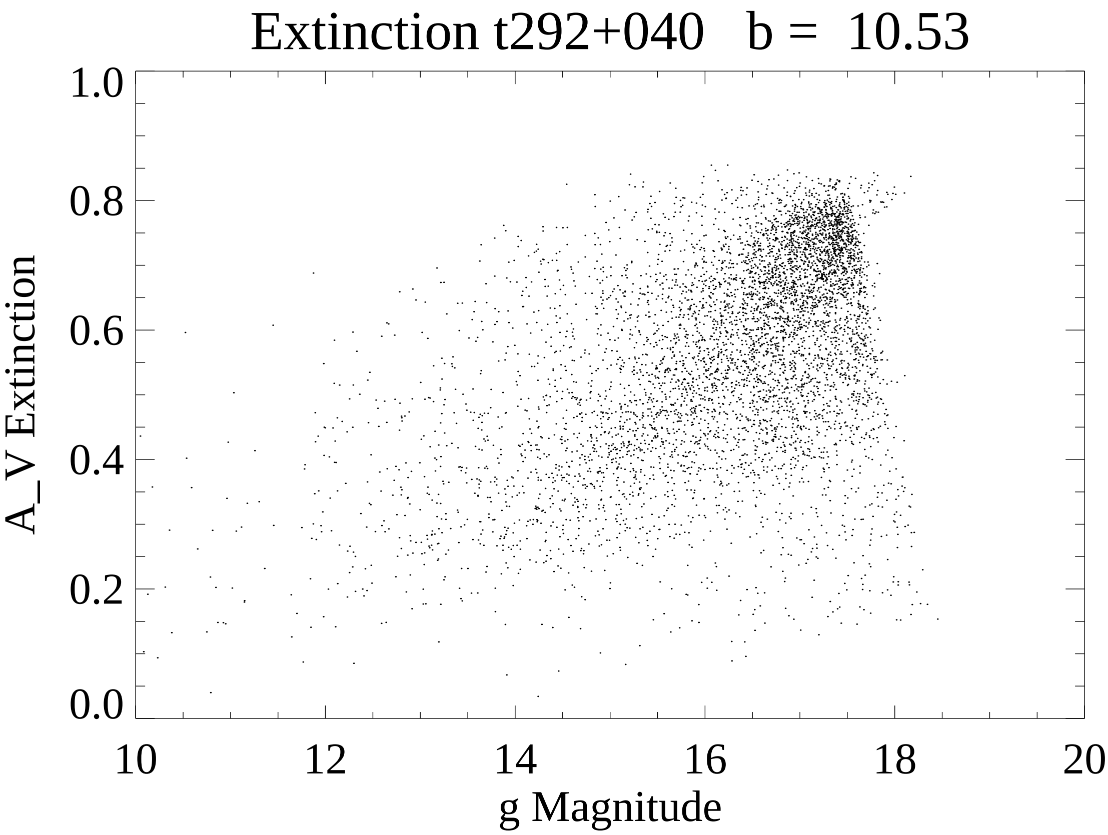

t292+040 Giants b = 10.53
Dereddened g-D51
Dereddened g-r
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.5
1.0
1.5

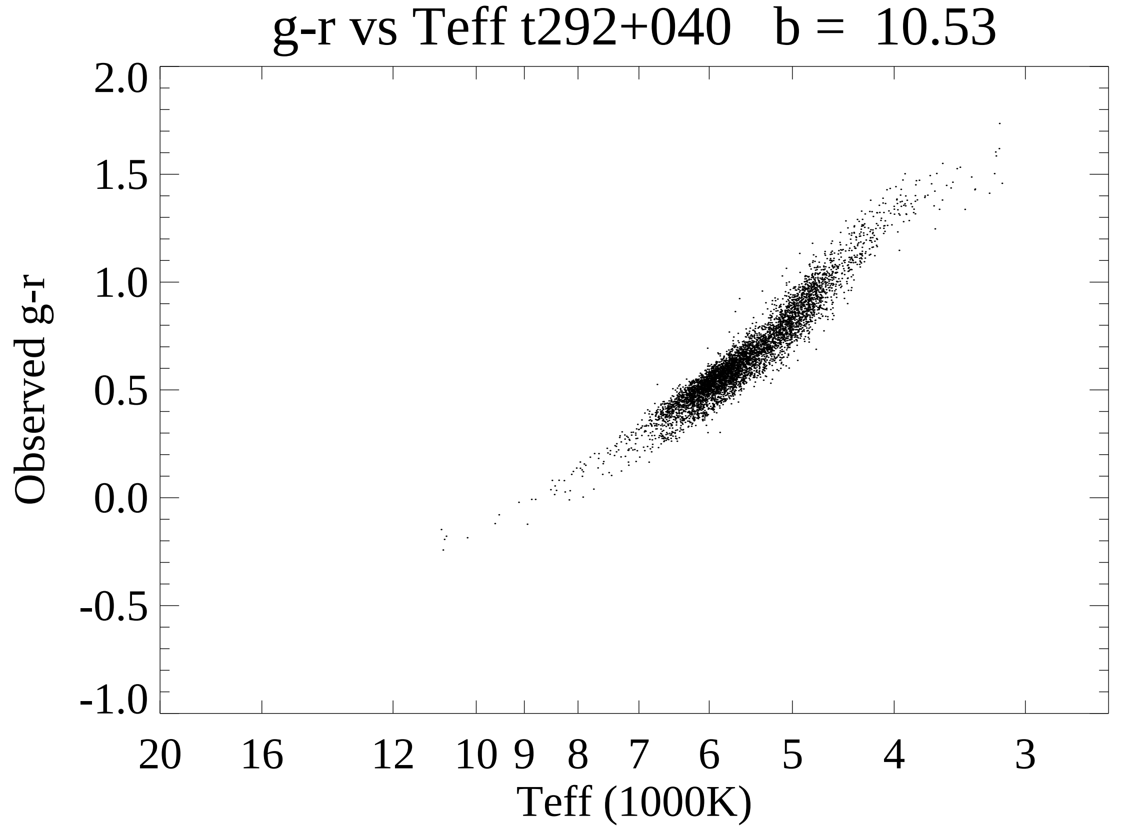

t292+040 Dwarfs b = 10.53
Dereddened g-D51
Dereddened g-r
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.5
1.0
1.5

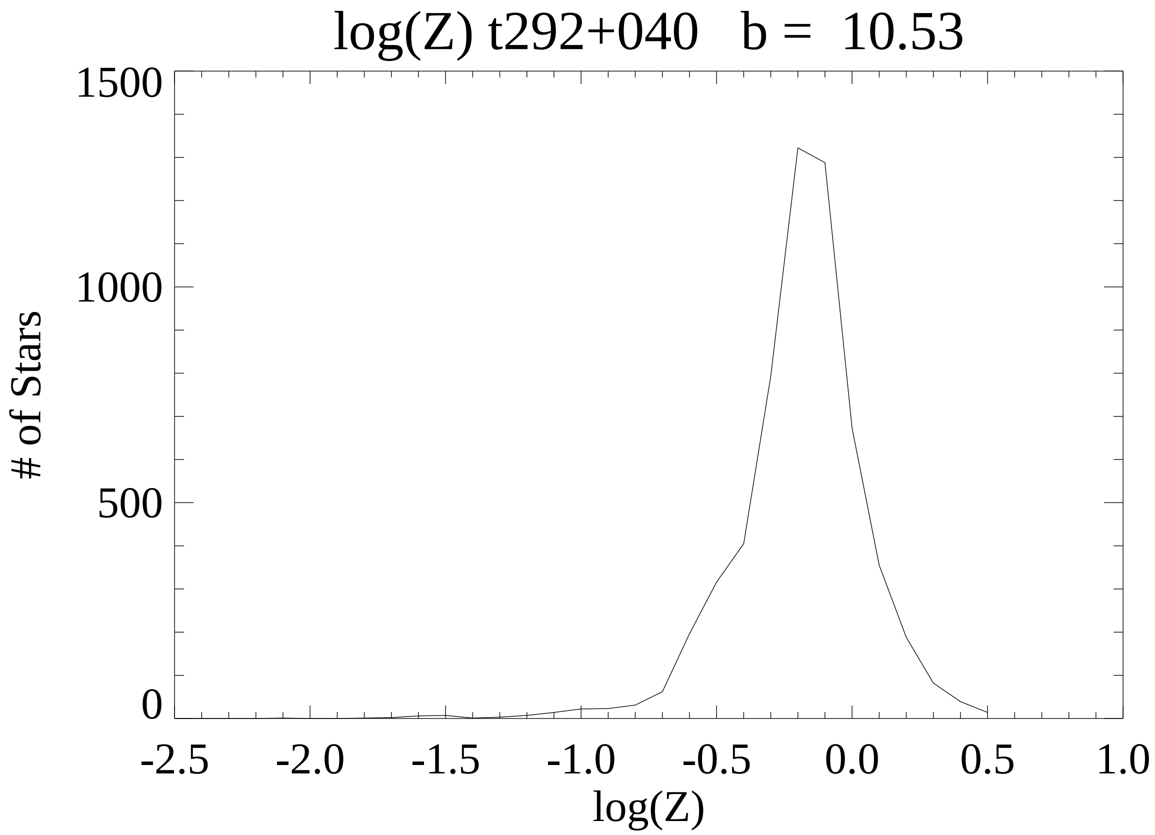

t292+040 Giants b = 10.53
Dereddened J-K
Dereddened g-i
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.5
1.5
2.0
2.5
3.0
3.5

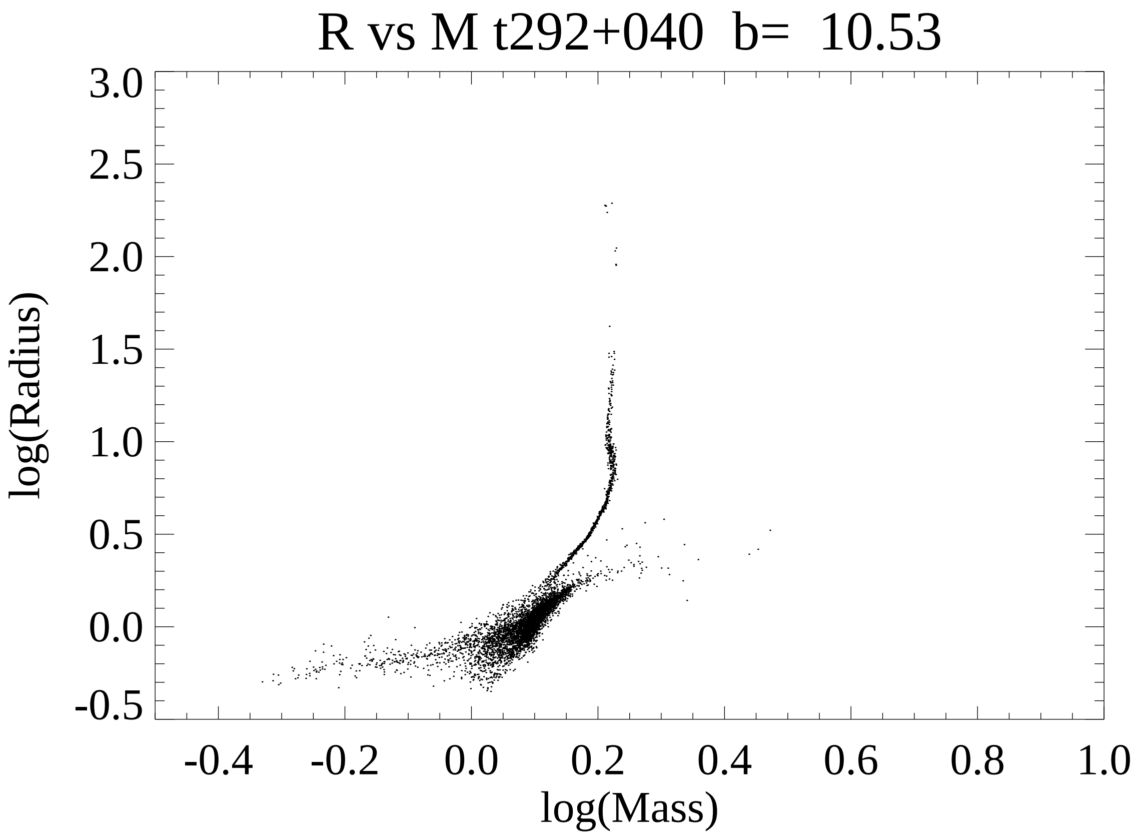

t292+040 Dwarfs b = 10.53
Dereddened J-K
Dereddened g-i
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.5
1.5
2.0
2.5
3.0
3.5

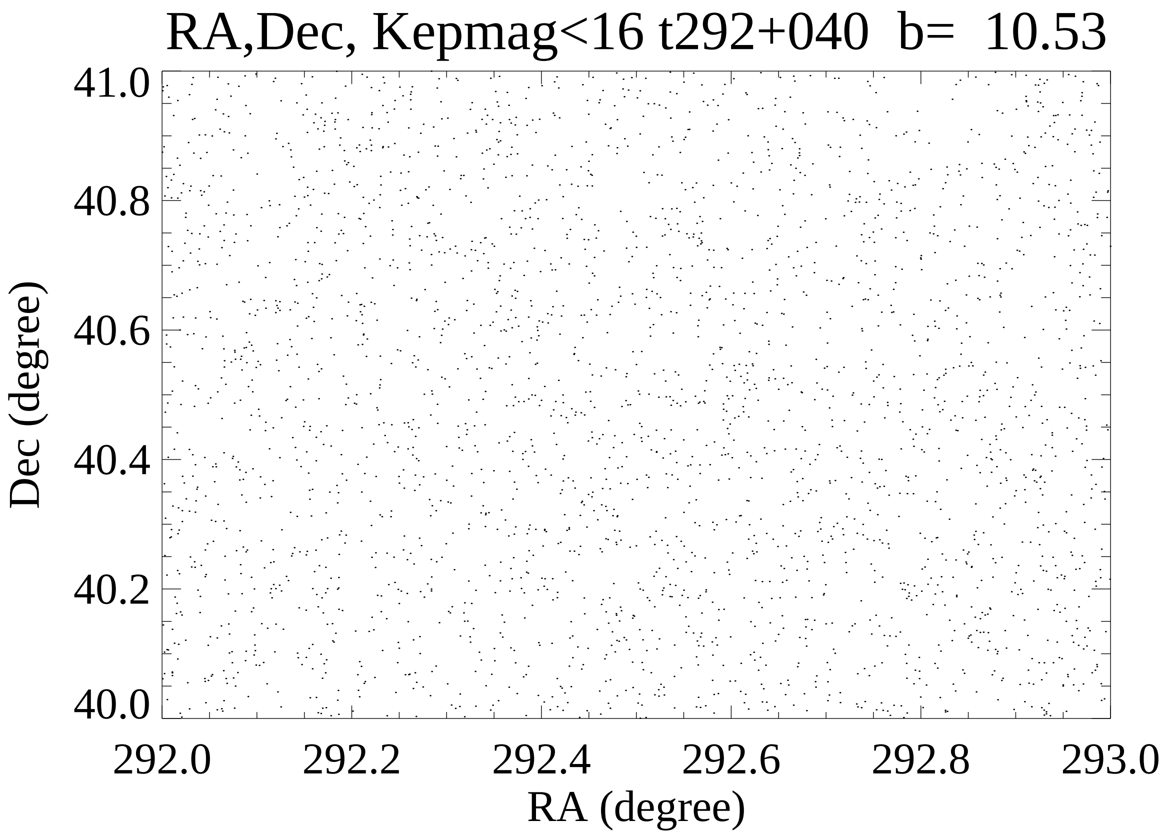

Teff/log(g) t292+040   b = 10.53
log(g)
0
1
2
3
4
5
12
10
9
8
7
6
5
4
3
Teff (1000K)


Extinction t292+040   b = 10.53
A_V Extinction
1.0
0.8
0.6
0.4
0.2
0.0
10
12
14
16
18
20
g Magnitude


g-r vs Teff t292+040   b = 10.53
Observed g-r
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
20
16
12
10
9
8
7
6
5
4
3
Teff (1000K)


log(Z) t292+040   b = 10.53
# of Stars
1500
1000
500
0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
log(Z)


R vs M t292+040 b= 10.53
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
log(Radius)
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
log(Mass)


RA,Dec, Kepmag<16 t292+040 b= 10.53
41.0
40.8
40.6
40.4
40.2
40.0
Dec (degree)
292.0
292.2
292.4
292.6
292.8
293.0
RA (degree)

Appendix 1: IDL Code

13. General Comments

All of the programs for this project are written in IDL. I did all of the algorithm development, and most of the coding. A substantial number of the routines in the oldest part of the code base were however written mostly by Don Kolinski, at HAO/NCAR. The failings and bad choices in this software suite are, of course, entirely mine.

All of the routines have a comment block at the beginning and other comments in line. As usual, however, the comments have not always kept up with the code changes, so some caution is required in interpretation. Quite a lot of code was replaced, bypassed, or otherwise rendered obsolete. In most of these cases the old executable statements have simply been commented out, so the original intent may be understood. These chunks of not-really-executed code can be a trap for the unwary, however.

I have attempted to make the variable names reasonably consistent throughout, but I have not gone back to enforce strict name identity for variables that mean the same thing in different routines.

Many routines find the home directory in which the database directories are located by reading the environment variable $ZDBASE. The zdb keyword that is commonly found in calling sequences provides a way to override this directory choice if desired.

14. Dependencies Among Routines

There follows the software dependency tree for the delivery (July28-Aug12 2008) version of the KIC. Who calls whom is indicated by indentation.

The list does not include IDL builtin functions; does not include Goddard database routines, even though some of these may have been altered for special purposes (mostly, though not exclusively, to allow a "quiet" option that dumps less verbiage to the screen), these routines (both altered and pristine) are listed at the end; does not include IDL Astrolib routines; does not include trivial wrappers to execute main routines on lists of many files.

The source code for these routines may be found at www.cfa.harvard.edu/kepler/kic/sourcecode/.

14.1. Main processing pipeline

```
fit2stardb3
    rd_sdss_q
run_obs2db4
    dec2seg
```

```
    obstilecount
    rebuild_obs2db
    seg2dec
obs2db4
    icamera
    rd_obs2
    rmbaddat
    obs2stardb3
        mk_starid
        tilenames
    obsdiagnos
        icamera
        rd_global
    obs2obsdb
tostd_segment
tostds_deseg
block_hand
bad_hand
    rd_obsdiag
        get_words
new_std_field
    tilenames
cull_refstar
    tilenames
fitstds1
    rd_global
    resid_plot2
        editbox
plot_refvsx
fitstars
    rd_global
stellar_parms1
    guessvals1
    kepler_mag
    err2mass
    find_prop1
        grid_likeli
            linked_props2
                starmass
                tg2vbmv
            colors_blk_intrp
                xform_colors
            red_vector
            hr_blk_intrp
            prob_dist
            prob_z
                rd_nord
            red_vector
                xform_redden
        lstsqr
```

- linked_props2
- parmblkintrp
- star_radius
- tilenames
- lum_r
- dbdump
  - bfpack

14.2. Stellar atmospheres, photometry, Bayesian priors

- narrow_set
  - rd_global
- phot_models
  - cphot
    - rd_castelli
      - get_words
    - rd_basel
    - redden
    - ccd_filtprofile
      - filtprofile
- m67_color_fit
  - m67_basel_colors
    - m67_props
      - rd_specdat
    - xform_colors
- gmr_adjust
- colors_adjust
- basel_index1
- color_fillin1
- color_block
- colblock_smooth
- mk_hrblock
  - prob_hr
    - rd_hip
  - vbmv2tg
- hrb_adjust
- masstable
  - rd_girardi

14.3. Diagnostic routines and plots of results

- kic_qc1
  - red_vector
    - xform_redden
- kic_qc2
  - getallphot
- kic_qc3

red_vector
xform_redden
pointing_eval1.pro
tilenames
rd_pointing1
print_pointing

Appendix 2: Database Schemas

15. Images

```
#title
Images table for Kepler photometry program, 4th generation.
#maxentries
50000
#items
Imid          C*24    Input image name
JD            R*8     Julian Date, decimal days
Filter        C*9     Filter name with trailing blanks
RA            R*8     RA (decimal hours)
Dec           R*8     Dec (decimal degrees)
Exptime       R*4     Exposure time (s)
Xmid          R*4     Airmass at image center
Obj           I*2     Object type (bias,dark,dome fl,sky fl,object)=(1-5)
Focus         R*4     Telescope focus
FWHM(4)       R*4     PSF FWHM for each chip
Bias(4)       R*4     Bias for each chip
Gain(4)       R*4     Gain for each chip
Apercor(4)    R*4     Aperture correction (mag) for each chip
Cldflg        I*2     Cloud flag
Astflg        I*2     Astrometry flag
Fitflg        I*2     Chi/Sharp fitting parameters flag
Camera        I*2     Camera index (4-shooter, MiniCam, KeplerCam)
#formats
ENTRY         I6      ISeq
Imid          A24     Image_Name
Filter        A9      Filter
RA            F12.7   RA,(hr)
DEC           F12.7   Dec,(deg)
Exptime       F7.2    Exptim
Xmid          F7.3    Xmid
Obj           I4      Obj
Focus         F7.3    Focus
FWHM          F7.2    FWHM,(pix)
Bias          F7.1    Bias,(ADU)
Gain          F7.2    Gain,e-/ADU
Apercor       F7.3    Ap_corr,(mag)
Cldflg        I4      CldF
Astflg        I4      AstF
Fitflg        I4      FifF
Camera        I4      Cam
#index
JD            sort
RA            sort
```

16. Blocks

```
#title
Blocks table for Kepler photometry program.
#maxentries
1000
#items
JDmin        R*8    Julian Date, decimal days, first image in block
JDmax        R*8    JD for last image in block
Iseqmin      I*4    Smallest Image sequence number in block
Iseqmax      I*4    Largest Image sequence number in block
Valid(7)     I*2    Which colors are represented in this block
A(7)         R*4    Zero point for each filter
B(7)         R*4    Transformation coefficient for each filter
C(7)         R*4    Color*airmass coefficient for each filter
K(7)         R*4    Linear extinction coefficient for each filter
Chip2(7)     R*4    Chip2 zero point rel to chip1
Chip3(7)     R*4    Chip3 zero point rel to chip1
Chip4(7)     R*4    Chip4 zero point rel to chip1
Qual(7)      R*4    Quality of fit for each filter
Comment      C*240  Comments
#formats
ENTRY        I6     BlkID
JDmin        F14.5  JD min
JDmax        F14.5  JD max
Iseqmin      I6     Ismin
Iseqmax      I6     Ismax
Valid        I4     Val
A            F8.4   A,(Zpt)
B            F8.4   B,(Trans)
C            F8.4   C,(2ndOrd)
K            F8.4   K,(Extinct)
Chip2        F8.4   Chip2,(Zpt)
Chip3        F8.4   Chip3,(Zpt)
Chip4        F8.4   Chip4,(Zpt)
Qual         F9.3   Quality
Comment      A240   Comment
#index
JDmin        sort
Iseqmin      sort
```

17. Stars

```
#title
Stars table for Kepler photometry program, 2nd generation.
#maxentries
100000
#items
StarID       C*21   StarID
RA           R*8    RA, hours, J2000.0
DEC          R*8    Dec, degrees, J2000.0
```

dra	R*4	RA Corr,(hours)
ddec	R*4	Dec Corr,(hours)
u	R*4	u,(SDSS)
sig_u	R*4	u_err
g	R*4	g,(SDSS)
sig_g	R*4	g_err
r	R*4	r,(SDSS)
sig_r	R*4	r_err
i	R*4	i,(SDSS)
sig_i	R*4	i_err
z	R*4	i,(SDSS)
sig_z	R*4	z_err
D51	R*4	D51
sig_D51	R*4	D51_err
Gred	R*4	Gred
sig_Gred	R*4	Gred_err
J	R*4	J,(2MASS)
H	R*4	H,(2MASS)
K	R*4	K,(2MASS)
ID2MASS	I*4	ID_2MASS
Var	I*2	Var
Crwd	I*2	Crwd
JD	R*8	JD
Source	I*2	Src
BadFlag(7)	I*2	If not zero, no clean input data

#formats

ENTRY	I6	Entry
StarID	A21	StarID
RA	F12.7	RA,(hr)
DEC	F12.7	Dec,(deg)
dra	F11.7	RA Corr,(hr)
ddec	F11.7	DecCorr,(deg)
u	F7.3	u,(SDSS)
sig_u	F7.4	u_err
g	F7.3	g,(SDSS)
sig_g	F7.4	g_err
r	F7.3	r,(SDSS)
sig_r	F7.4	r_err
i	F7.3	i,(SDSS)
sig_i	F7.4	i_err
z	F7.3	z,(SDSS)
sig_z	F7.4	z_err
D51	F7.3	D51
sig_D51	F7.4	D51_err
Gred	F7.3	Gred
sig_Gred	F7.4	Gred_err
J	F7.4	J,(2MASS)
H	F7.4	H,(2MASS)
K	F7.4	K,(2MASS)
ID2MASS	I10	ID_2MASS

```
Var            I5     Var
Crwd           I5     Crwd
JD             F14.5  JD
Source         I5     sRC
BadFlg         I4     BadF
#index
RA             sorted
DEC            sort
g              index
Source         index
```

18. Observations

```
#title
Observations table for Kepler photometry program.
#maxentries
250000
#items
StarID         C*21   StarID
RA             R*8    RA, hours, J2000.0 (corresp to 'stars' table)
Dec            R*8    Dec, degrees, J2000.0 (corresp to 'stars' table)
ISeq           I*4    Image sequence number
Filter         C*9    Name of filter, with trailing blanks
mag            R*4    Magnitude in this filter
err            R*4    Estimated error in mag
X              R*4    Airmass for this star
chip           I*2    4-shooter chip number
xcen           R*4    x- center coord of star on chip (pix)
ycen           R*4    y- center coord of star on chip (pix)
sky            R*4    Estimated sky background for star
sharp          R*4    DAOPHOT sharp parameter
chi            R*4    DAOPHOT goodness-of-fit parameter
dra            R*4    RA diff betw this obs and standard (s of arc)
ddec           R*4    Dec diff betw this obs and standard (s of arc)
#formats
ENTRY          I6     Entry
StarID         A21    StarID
RA             F12.7  RA,(hr)
Dec            F12.7  Dec,(deg)
ISeq           I6     ISeq
Filter         A9     Filter
mag            F7.3   mag
err            F7.4   err
X              F7.3   Airmass
chip           I4     Chip
xcen           F8.2   xcen,(pix)
ycen           F8.2   ycen,(pix)
sky            F8.3   Sky,(ADU)
sharp          F7.3   Sharp
chi            F7.3   Chi
dra            F7.3   RA disp,(asec)
```

```
ddec         F7.3   Dec dis,(asec)
#index
RA           sorted
DEC          sort
ISeq         sort
mag          index
err          index
X            index
sky          index
dra          index
ddec         index
```

19. Properties

```
#title
Star properties table for Kepler photometry program.
#maxentries
100000
#items
StarID       C*21   StarID
RA           R*8    RA, hours, J2000.0
DEC          R*8    Dec, degrees, J2000.0
teff         R*4    Teff, K
err_teff     R*4    Teff_err, K
logz         R*4    log(Z)
err_logz     R*4    log(Z)_err
logg         R*4    log(g)
err_logg     R*4    log(g)_err
ext          R*4    A_V, mag
err_ext      R*4    A_V_err, mag
radius       R*4    Radius, R_sun
kepmag       R*4    Kepmag, mag
uflag        I*2    uflg
m2flag       I*2    2Mflg
zflag        I*2    zflg
oflag        I*2    Oflg
chisq        R*4    Chisq
#formats
ENTRY        I6     Entry
StarID       A21    StarID
RA           F12.7  RA,(hr)
DEC          F12.7  Dec,(deg)
teff         F6.0   Teff, K
err_teff     F5.0   Teff_err, K
logz         F6.2   log(Z)
err_logz     F5.2   log(Z)_err
logg         F6.2   log(g)
err_logg     F5.2   log(g)_err
ext          F6.2   A_V, mag
err_ext      F5.2   A_V_err, mag
```

radius	F5.2	Radius, R_sun
kepmag	F7.3	Kepmag, mag
uflag	I5	uflg
m2flag	I5	2Mflg
zflag	I5	zflg
oflag	I5	Oflg
chisq	F8.2	Chisq
#index		
RA	sorted	
DEC	sort	

Appendix 3: Configuration and Input Data Files

Many of the routines that implement the algorithms described here require input from ascii tables or from IDL save files. These files are described briefly here. Copies of the files used for the processing of KIC10 are found in www.cfa.harvard.edu/kepler/kic/ancillary_files/. In the IDL code, complete paths to these data files are usually hard-wired.

glob_parms1.asc, glob_parms2.asc, glob_parms3.asc. These files contain camera-dependent (hence date-dependent) information about average extinction coefficients in each filter.

images4.dbd, blocks.dbd, stars3.dbd, observ.dbd, properties2.dbd. These files define the schemas of the principal IDL databases. Printed versions of them can be found in Appendix 2 above.

obs2db.log. This file contains logging information about each image ingested into the database structure. It is written by obs2db4, and includes diagnostic information computed in routine obsdiagnos, and eventually read and used by routine bad_hand.

4s_survey_c2. This file tells the routines that compute synthetic photometry where to find transmission profiles for filters, and which filters should be used in computing the colors that are compared with observed photometry.

sdssu_prof.asc, sdssg_prof.asc, sdssr_prof.asc, sdssi_prof.asc, sdssz_prof.asc, D51prof.asc, Gred_prof.asc, J_prof.asc, H_prof.asc, Ks_prof.asc. These files contain the filter profiles that are used in computing synthetic photometry.

colblock_cas4_smth.sav, xform_color_cas4.sav. The first of these files contains synthetic colors for all filters, extrapolated to be complete on a specified grid of $\{T_{eff}, \log(g), \log(Z)\}$, and smoothed. The second contains coefficients in the linear color transformation that is applied to force agreement between the model colors and those observed for M67.

hipparcos.txt, nordstrom.txt. These files are used as input to create the Bayesian prior probability distribution that relates to the distribution of stars in the HR diagram.

hrblk_mod2.sav. This file contains the aforementioned Bayesian prior distribution relating to the distribution of stars in the HR diagram.

propsblock.sav. This file contains the tables that allow routine parmblkintrp to calculate estimates of stellar mass, luminosity, absolute r magnitude, $B-V$, $V-R$, and bolometric correction as functions of $\{T_{eff}, \log(g)\}$.

guessvals_in.sav. This file contains data that allow estimation of missing photometry based on observed values of $g-r$ or $r-i$. The interpolated data are used to avoid crashes while estimating stellar parameters.

processing.txt. This file contains the mapping between date of observation (which here is the civil date at the start of the observing night, not the UT date which is invariably 1 day larger) and the block number which sometimes shows up in both the IDL code and the diagnostic plots.

pointing_eval_080729.asc Contains number of visits and RMS scatter of r magnitude and various colors for each of the 1600 pointings that cover the entire Kepler field.

Appendix 4: Known Defects in KIC10

20. Summary

The Stellar classifications provided in KIC10 suffer from several known systematic defects that should be considered when using them. This document describes the most important of these, explains their source when this is known, and illustrates the problems with samples from the data. The effects described here are as follows:

(1) KIC10 T_{eff} values have systematic disagreements with other T_{eff} estimates that apply to the same stars. (Of course, these other estimates also disagree systematically with each other.) Though these disagreements are usually less than 50K (?), for the worst cases they may exceed 200K.

(2) The KIC10 classifications tend to give $\log(g)$ too large for subgiant stars, especially those hotter than about 5500K (?). This leads to underestimates of the radii of this subset of stars, typically by factors of 1.5 to 2.

(3) Some giant stars are classified in a displaced giant sequence, lying a few hundred degrees cooler or with $\log(g)$ about 0.5 dex higher than the actual giant branch.

(4) For $T_{eff} \leq$ 4200K, KIC10 estimates of $\log(Z)$ are systematically positive (average values about +0.3), although the average metallicity in the solar neighborhood is about −0.1 to −0.2. This effect vanishes rather abruptly at $T_{eff} \geq$ 4200K.

(5) Some localized regions of the Kepler field are sampled only by photometry taken under poor photometric conditions, with erratic classifications as a result.

(6) Parts of the Kepler field near the galactic plane have large and irregular interstellar extinction and reddening. This leads to relatively large scatter in the resulting classifications.

21. T_{eff} Scale

Comparison of KIC10 and spectroscopic (from Meibom, private communication) T_{eff} estimates for 1315 field stars shows good agreement for temperatures near 5500K, but also reveals that the KIC temperatures are systematically cooler for both higher and lower temperatures (Fig. A1). The origin of these differences is not clear, since both analyses are derived from the same (Castelli & Kurucz 2004) model atmospheres, and since both appear to provide adequate fits to their respective input data sets. Possibly there is an inconsistency between the model treatment of spectrum lines and that of continua. This is speculation, however; it will take further work to understand the origin of this difference.

It is important to note, however, that the spectroscopic temperatures are themselves open to question, and indeed are clearly in error for some parameter ranges. The points that are plotted as diamonds enclosing plus signs in Fig. A1 are all those having $g-r$ colors redder than 1.2. For these stars the T_{eff}-color relation remains tight if one uses the KIC10 values of T_{eff}, but breaks down if one uses the spectroscopic T_{eff} estimates. This behavior casts doubt on the spectroscopic T_{eff} estimates, at least for stars that are cool enough.

22. Subgiant gravity

For temperatures above roughly 5500K, none of our photometric colors provide information about surface gravity. Accordingly, for hotter temperatures the maximum likelihood analysis has no basis to choose any $\log(g)$ other than the one that is most probable a priori, which corresponds to the center of the main sequence, near $\log(g) = 4.5$. Stars on the giant branch almost all have T_{eff} small enough so that their gravities can be measured; so the radii of true giants appear to be estimated fairly accurately. The result of these two behaviors can be seen in Figs. A2 and A3, which compare $\log(g)$ estimated by the KIC10 methods with those obtained by other means, for the open cluster M67 and for Soeren Meibom's spectroscopic sample of 1315 field stars. For M67, I have estimated $\log(g)$ from isochrone fitting in the $(g - r)$ vs g diagram; for the field stars, Meibom estimated $log(g)$ based on matching model atmosphere spectra to observations obtained with the Hectochelle spectrograph. In both cases, agreement is pretty good for the bulk of the objects, but for a significant subset of (mostly fairly hot) stars, the KIC-derived gravities are systematically larger, sometimes by as much as 1 dex.

Given the available photometric data, this problem is essentially unavoidable. The information needed to distinguish between F- and early G-type dwarfs and subgiants is not present in the data, and there is no way to obtain sensible results without it. (Biasing the prior probabilities towards lower gravities, for instance, results in more subgiants in the HR diagram, but with no guarantee that the new alleged subgiants are in fact those with low gravity.

23. Spurious Displaced Giant Branch

Dennis Stello has analyzed KIC data relating to stars in the open cluster NGC 6819, comparing the KIC results with those derived from photometry by Stetson et al. (2003). Several stars identified as giant-branch members of the cluster appear displaced to the right in a T_{eff}-r-magnitude diagram (Fig. A4, center panel). From other information it seems clear that these stars are being identified as dwarfs, and in order to show the same colors at higher gravity, their T_{eff} estimated values must be decreased by several hundred K. The reason for this failure is not yet clear; more work will be needed to understand it.

24. High $\log(Z)$ at low T_{eff}

A plot of $\log(Z)$ vs T_{eff} for a large randomly-sampled group of stars (Figure A5) shows a number of peculiarities. For T_{eff} below about 4200K, the estimated $\log(Z)$ distribution begins to spread and bifurcate, and below 3800K, virtually all stars show $\log(Z)$ greater than +0.5, which is the highest metallicity represented in the Castelli models that I use. This behavior has not been investigated in detail, but it seems likely that it results from a mismatch between the model and observed color dependences at low temperatures, in the sense that (other astrophysical evidence notwithstanding), high-metallicity stars provide the best fit to the observations.

One can also observe clustering of $\log(Z)$ around integral and half-integral values of $\log(Z)$, for $\log(Z) \leq -1$. These are the tabulated values of $\log(Z)$; the concentration of estimated values near the tabulated ones presumably indicates a failure of the interpolation and fitting code that optimizes the posterior probability. This will be investigated further.

25. Regions with Poor Photometry

A small fraction of the pointings that make up the data for KIC10 have only one usable visit, with that one occurring on a night of questionable sky clarity. Notable examples of such areas can be found along a strip near declination +36 degrees. (The usual observing strategy was to place the telescope at some particular declination, and then sweep out a swath of pointings by stepping west to east. One pointing with substandard data therefore tends to be accompanied by others at the same declination.) Locations with unusually poor photometry can be identified most easily by consulting the file pointing_eval_080729.asc, and checking for a small number of visits or for unusually large RMS scatter in the r magnitude or in the colors. Note, however, that (sensibly enough) pointings that have only one visit show RMS values of zero, hence there is little information about their reliability.

26. Extinction and Reddening

Regions at low galactic latitude are prone to have large and spatially nonuniform extinction and reddening. The model of extinction that is employed in the Bayesian likelihood maximization contains no small-scale structure, so it is unable to deal properly with localized large deviations from the mean extinction. the result is systematic misclassification of many stars, a scattered and confused relation between T_{eff} and color, and other failings. Examples of such difficulties are shown in Figures A6 and A7, both of which relate to a tile lying near galactic latitude +5°. One possible improvement over the present scheme would be to scale the model extinction according to the measured total line-of-sight extinction to each star (as measured by Schlegel et al. (1998), say). A strategy such as this would probably account for some of the spatial variation in extinction, but of course would do nothing about the irregular distribution of dust along the line of sight.

Appendix Figure Captions

Fig. 26.— T_{eff} values for (mostly main-sequence) field stars as estimated in KIC10 (on the y-axis) plotted against those estimated by Meibom's spectrographic analysis (on the x-axis). Plus symbols overplotted with diamonds indicate stars with $g - r$ colors redder than 1.2

Fig. 27.— log(g) estimated using Bayesian likelihood maximization (y-axis), plotted against log(g) estimated from Meobom's spectra (x-axis), for stars in the open cluster M67.

Fig. 28.— Same as for Fig A2, but for a sample of 1315 field stars with spectroscopic classifications by Meibom. These stars were selected among those classified as dwarfs in KIC7; this selection appears to have been fairly successful, insofar as there are few obvious giants. For assessing the fidelity of the Bayesian method, it would however have been better if there were more giants. Stars overplotted with diamonds are those with $g-r$ colors redder than $g - r$=1.2, as in Fig. A1.

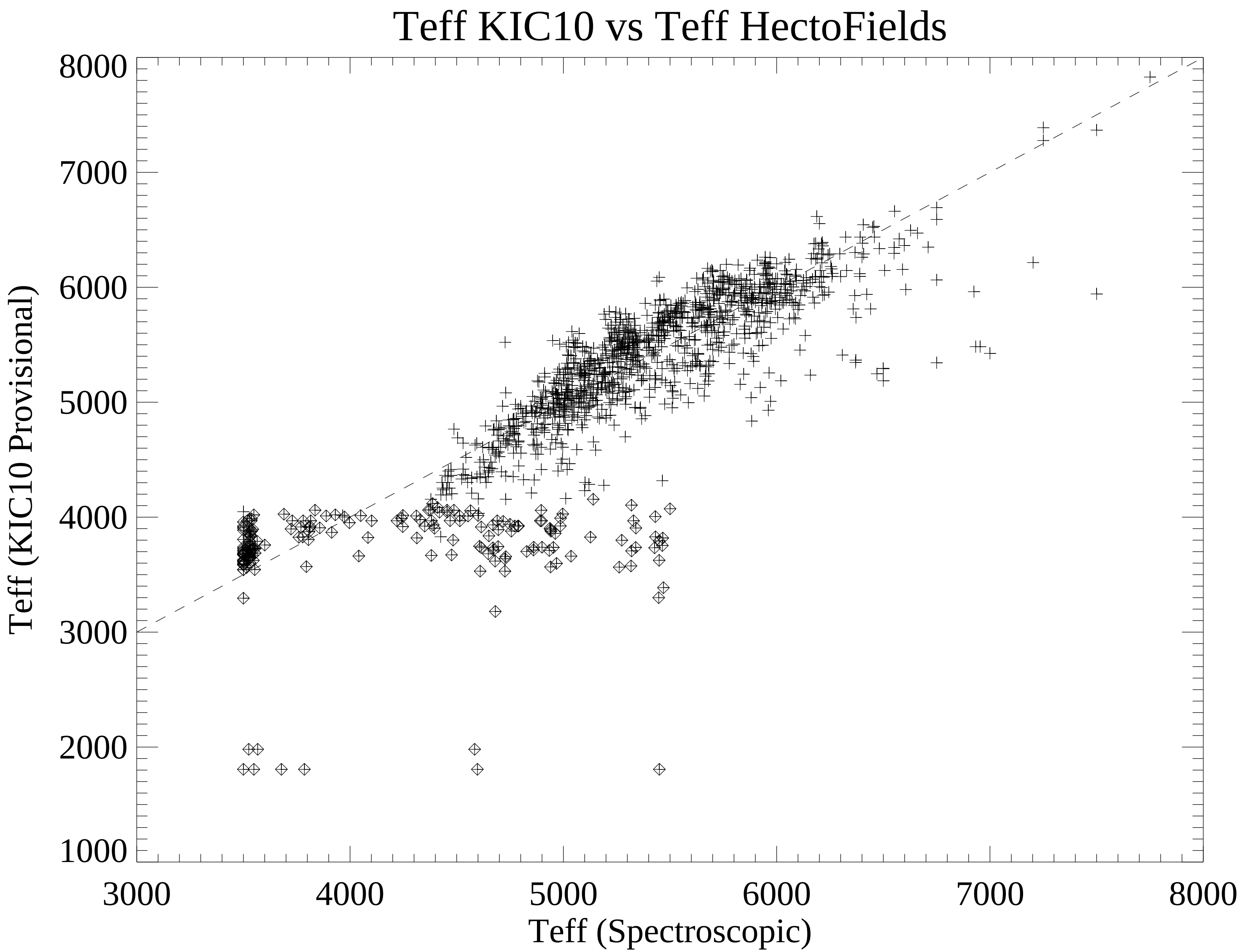
Fig. 29.— Plot of log(Z) against T_{eff}, illustrating the strong tendency of low-temperature stars to be classified with high metallicity, and moreover the tendency for low-Z stars to be ascribed integral- or half-integral values of log(Z), independent of T_{eff}.

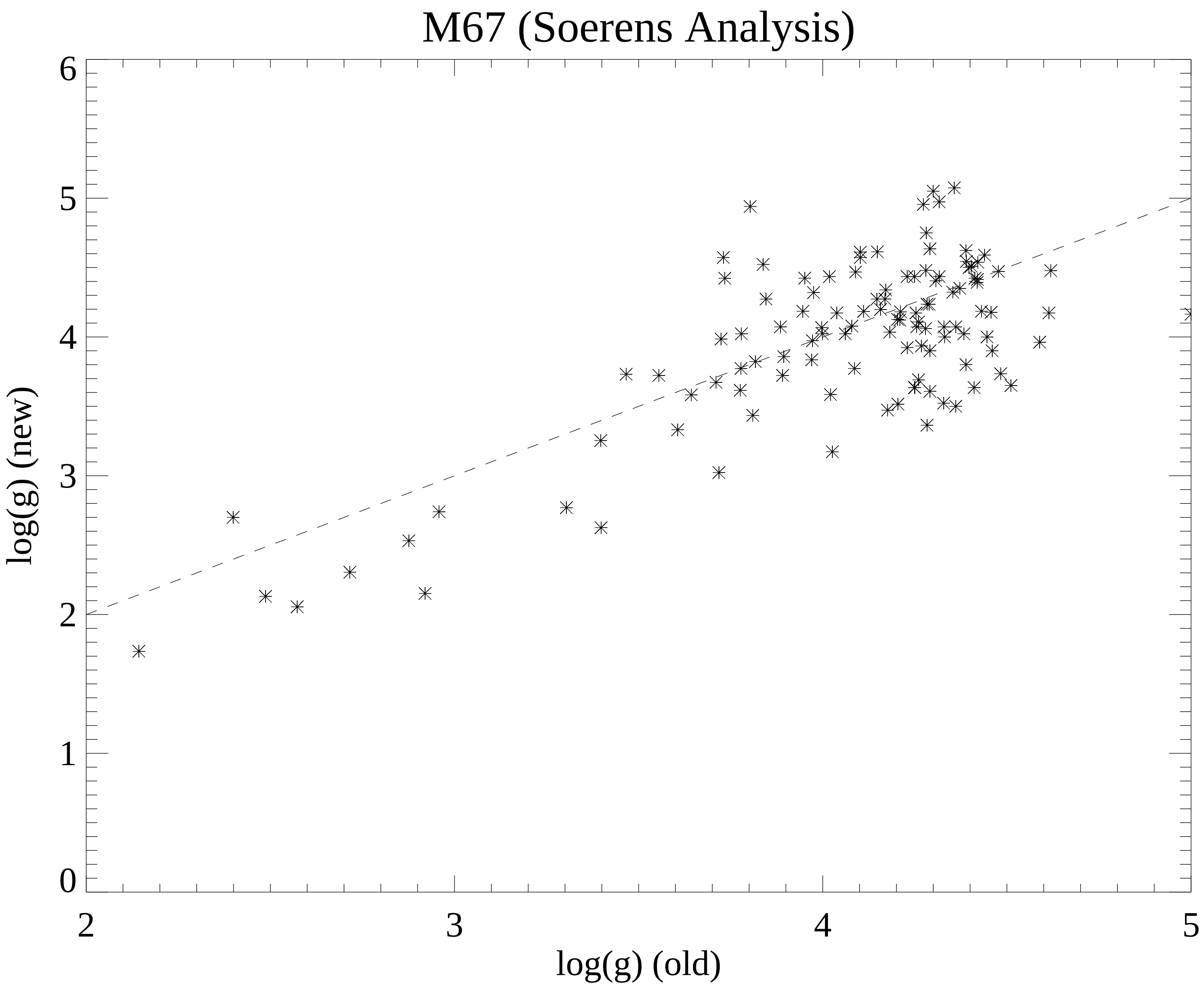
Fig. 30.— Analysis of stars within NGC 6791 by Dennis Stello, showing a "second giant branch" displaced from the true giant branch by a few hundred K in a T_{eff}-r-magnitude plot (center panel).

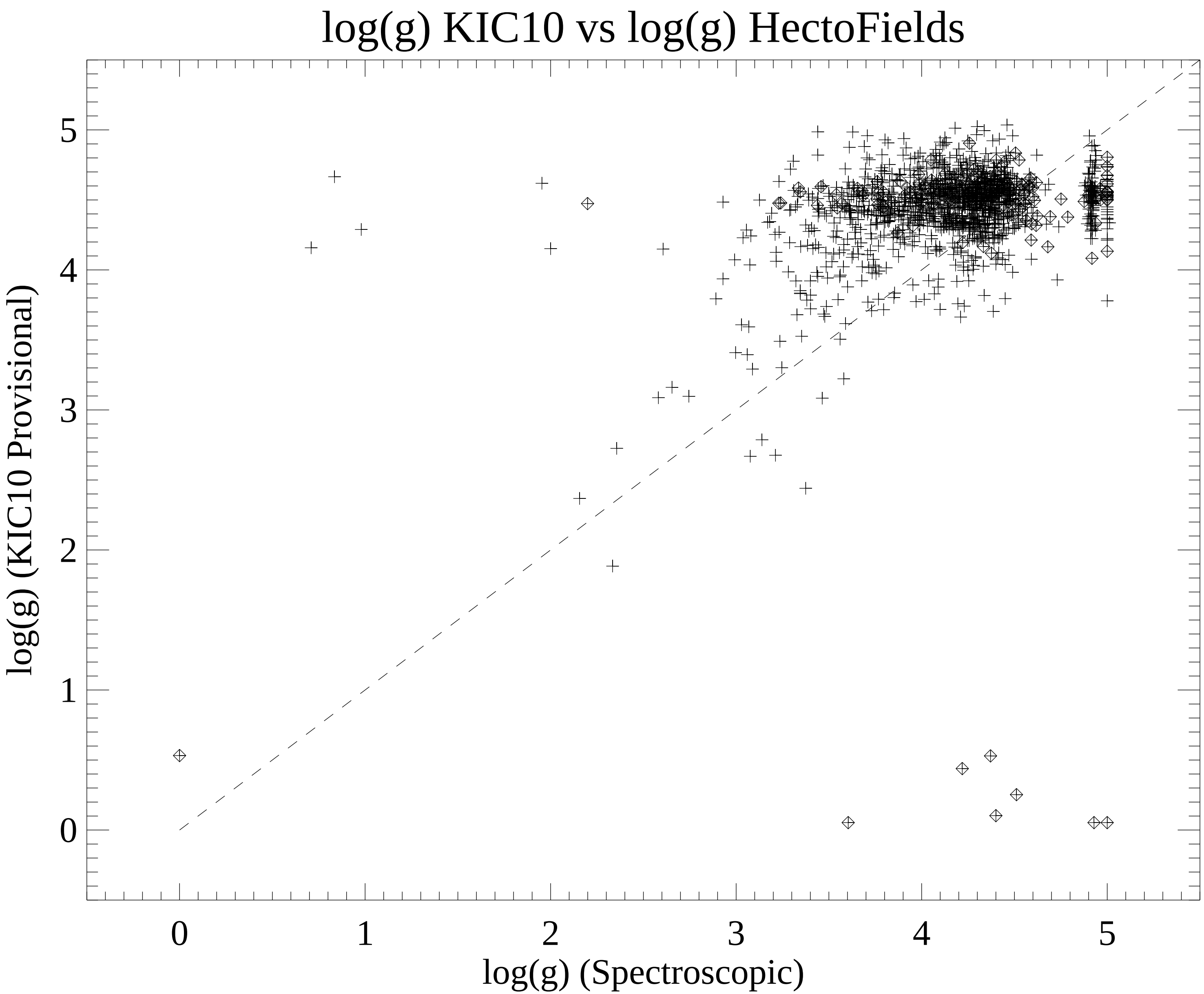
Fig. 31.— First plot from kic_qc1, showing giant-dwarf separation in the tile t299+040. Notice the broad and fuzzy domains in color-color space occupied by the observed stars, partly as a result of problematic de-reddening using the assumed model of interstellar extinction.

Fig. 32.— Fourth plot from kic_qc1. Note the two parallel sequences of stars in the T_{eff}-$g-r$ plot (top panel). This is evidence that one value of the extinction parameter κ_r is inadequate

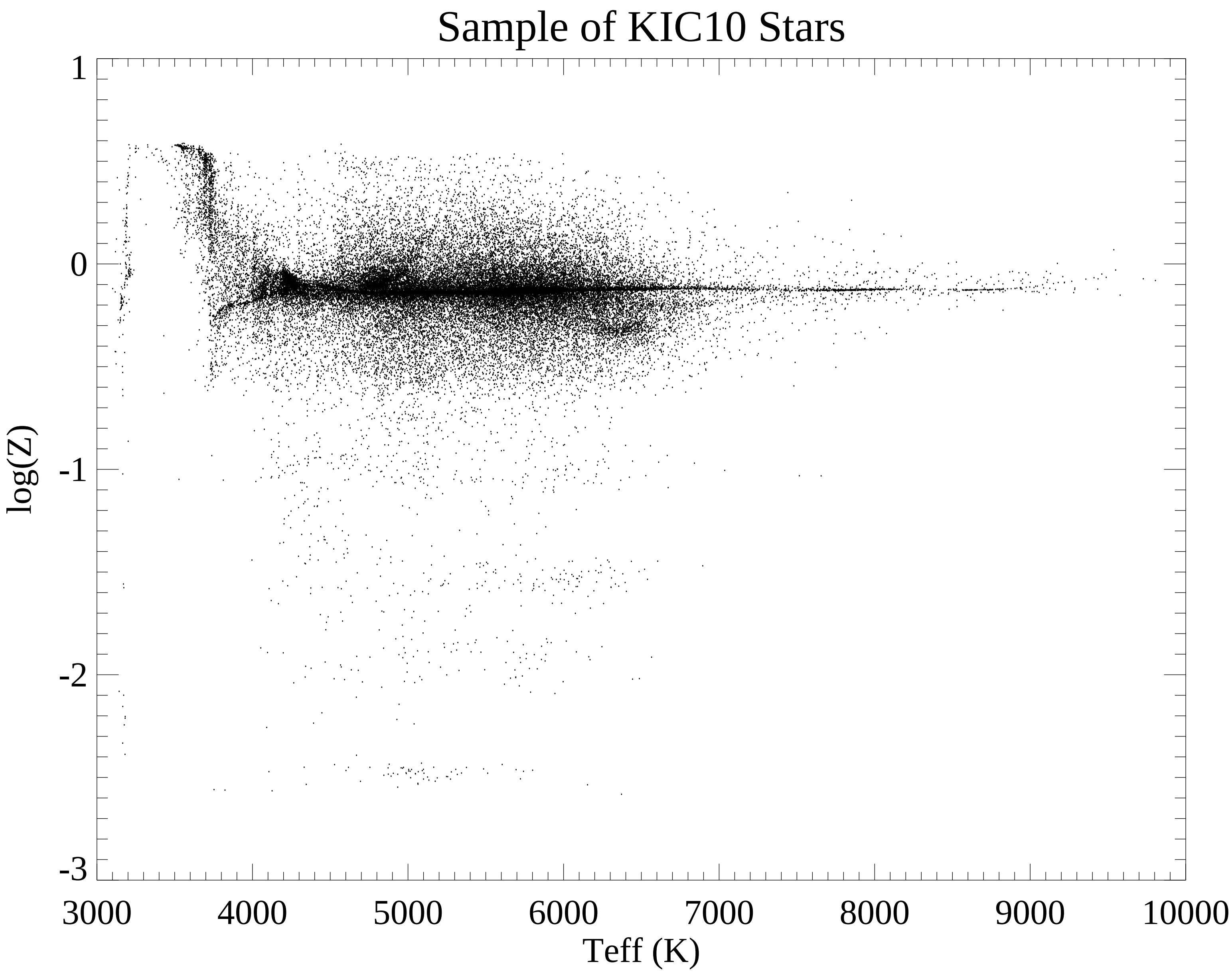
to describe the colors of the stars observed in this tile.

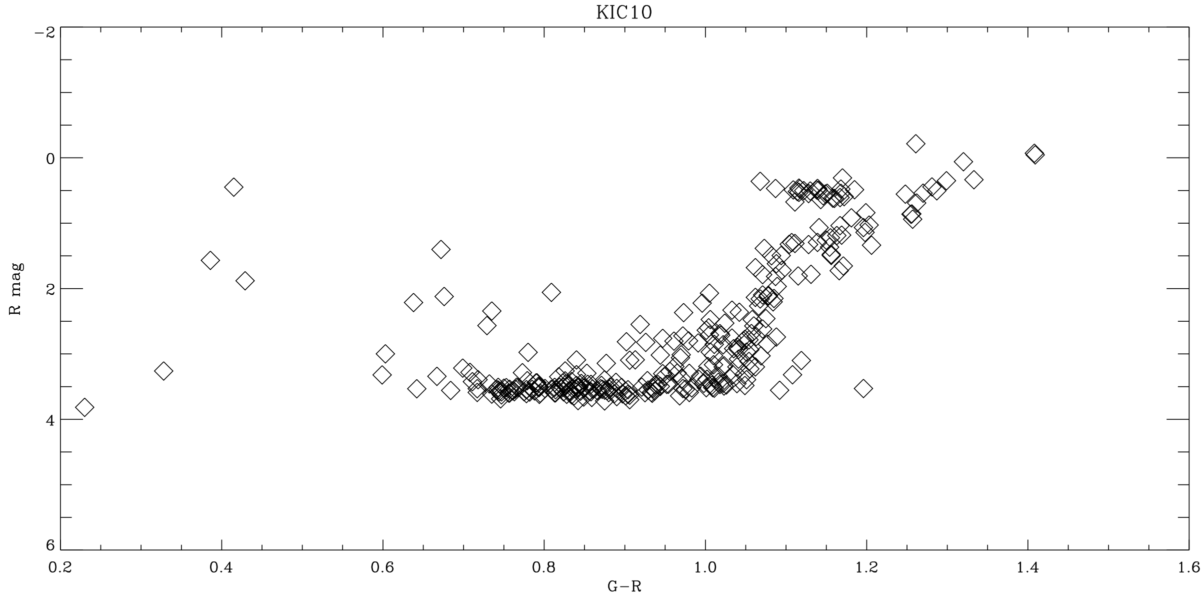

Teff KIC10 vs Teff HectoFields
Teff (KIC10 Provisional)
Teff (Spectroscopic)
8000
7000
6000
5000
4000
3000
2000
1000
3000
4000
5000
6000
7000
8000

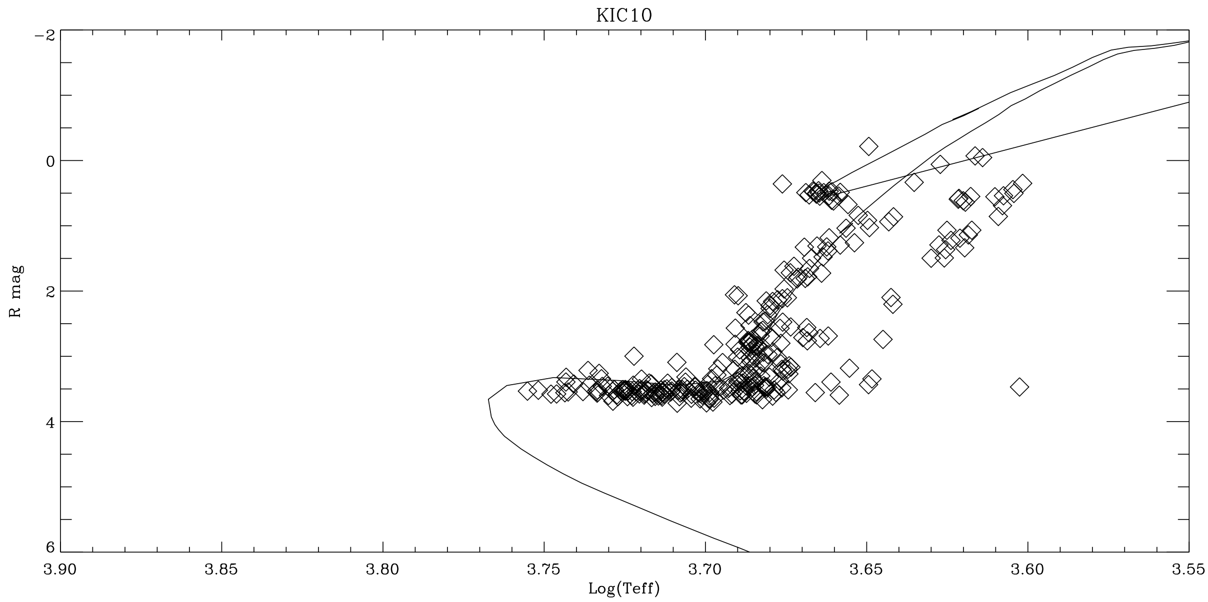

M67 (Soerens Analysis)
log(g) (new)
log(g) (old)
0
1
2
3
4
5
6
2
3
4
5

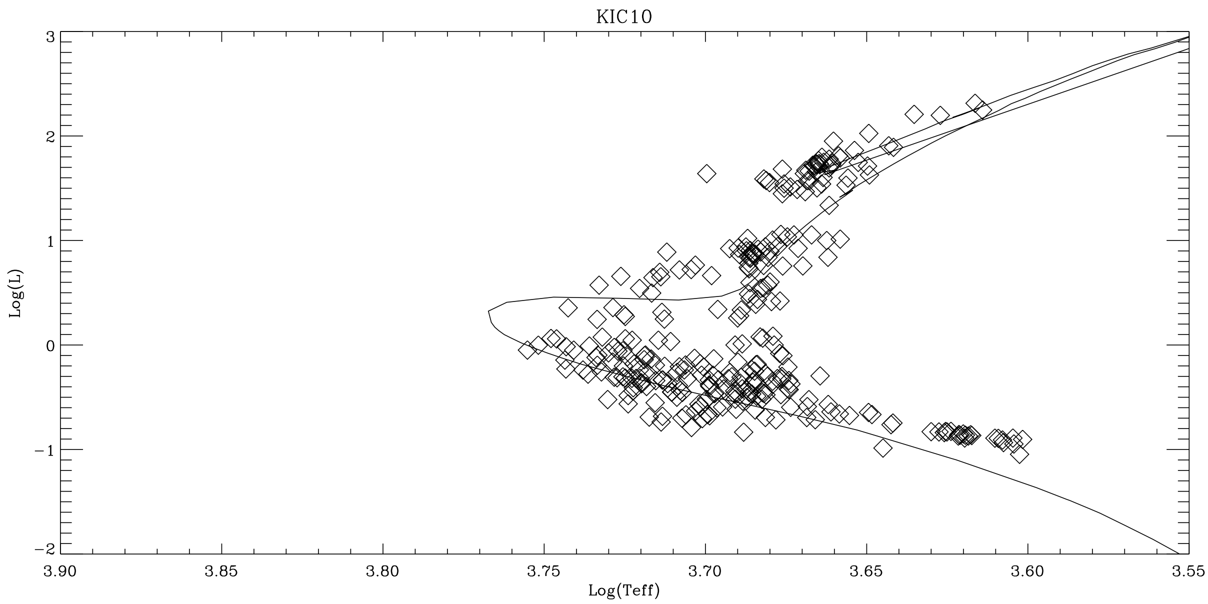


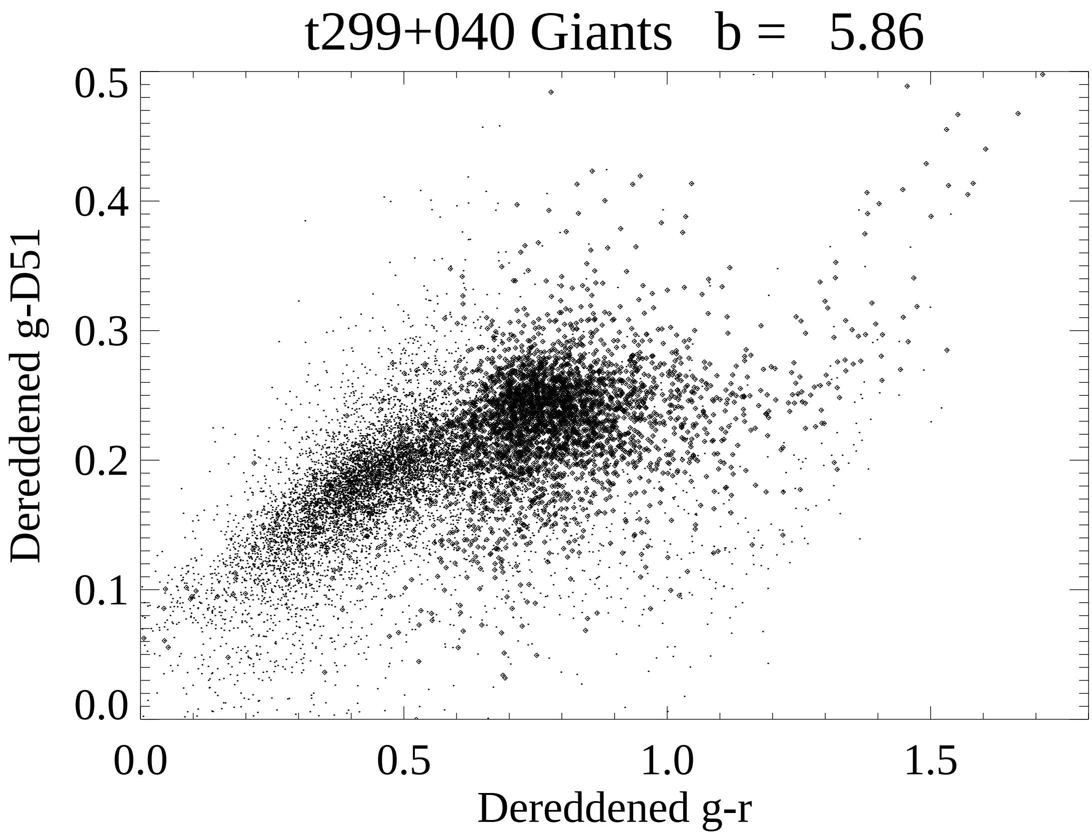
log(g) KIC10 vs log(g) HectoFields
log(g) (KIC10 Provisional)
log(g) (Spectroscopic)
0
1
2
3
4
5

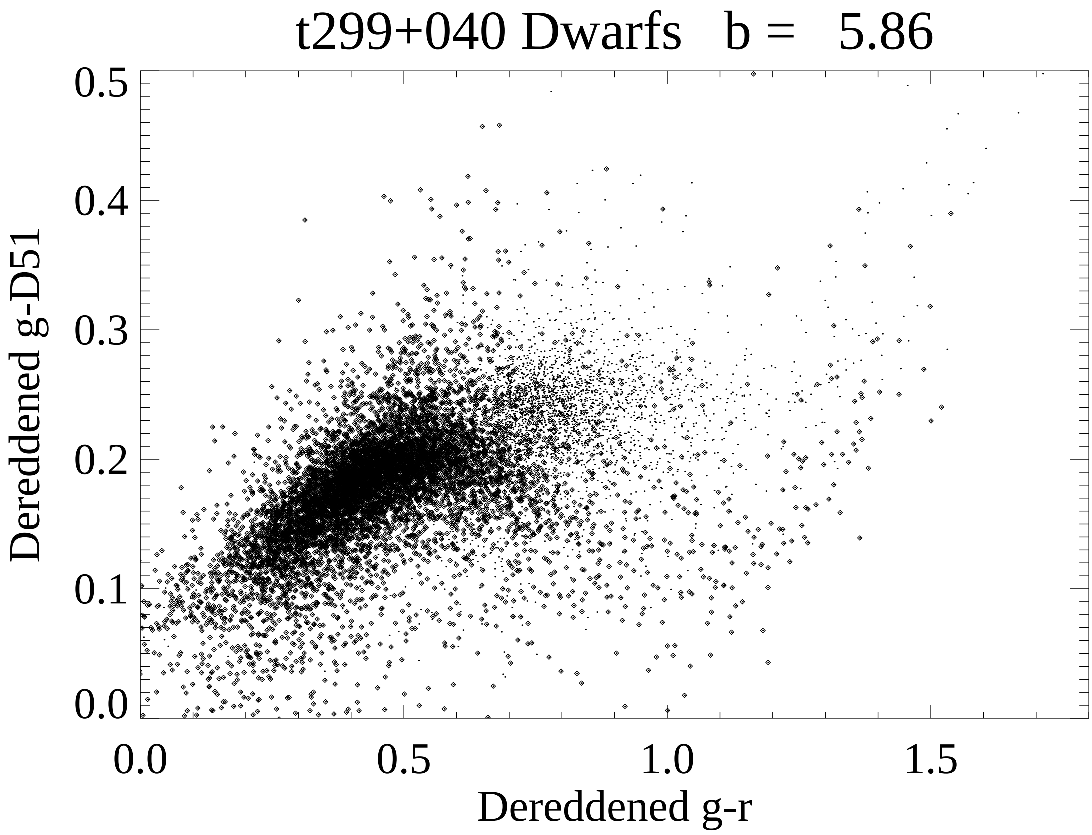

Sample of KIC10 Stars
log(Z)
1
0
-1
-2
-3
Teff (K)
3000
4000
5000
6000
7000
8000
9000
10000

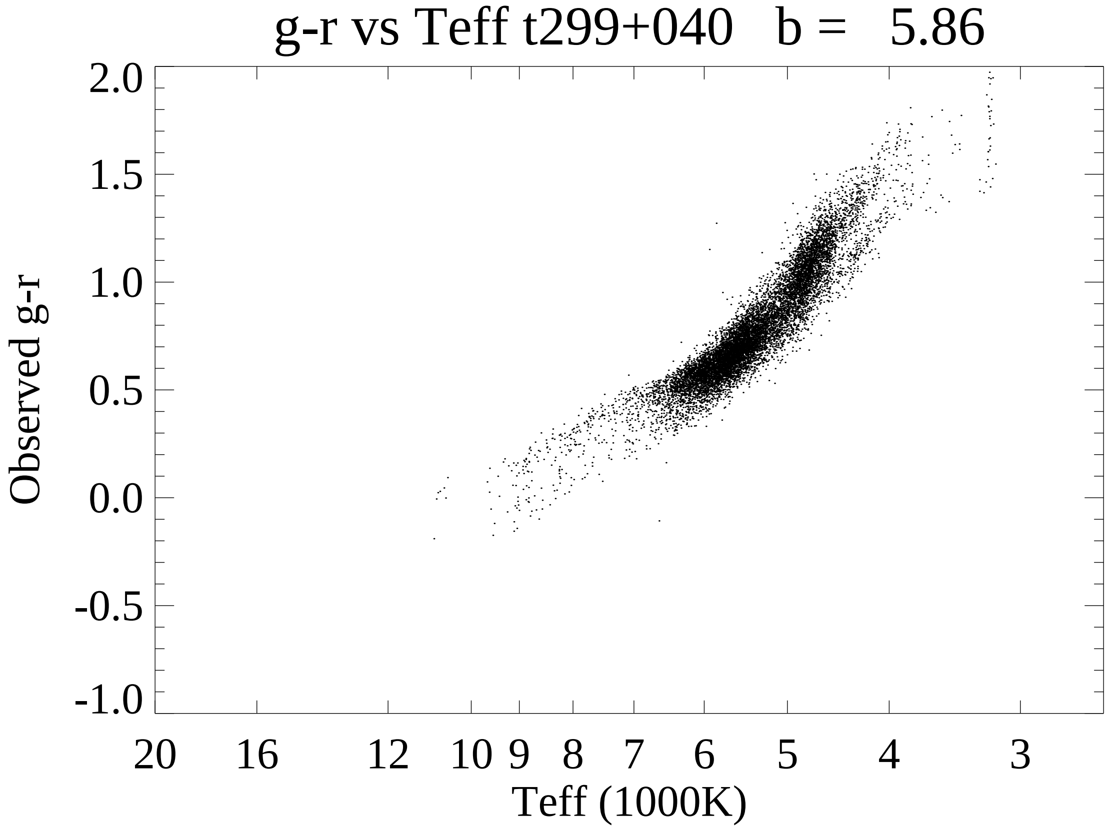

KIC10
R mag
G−R

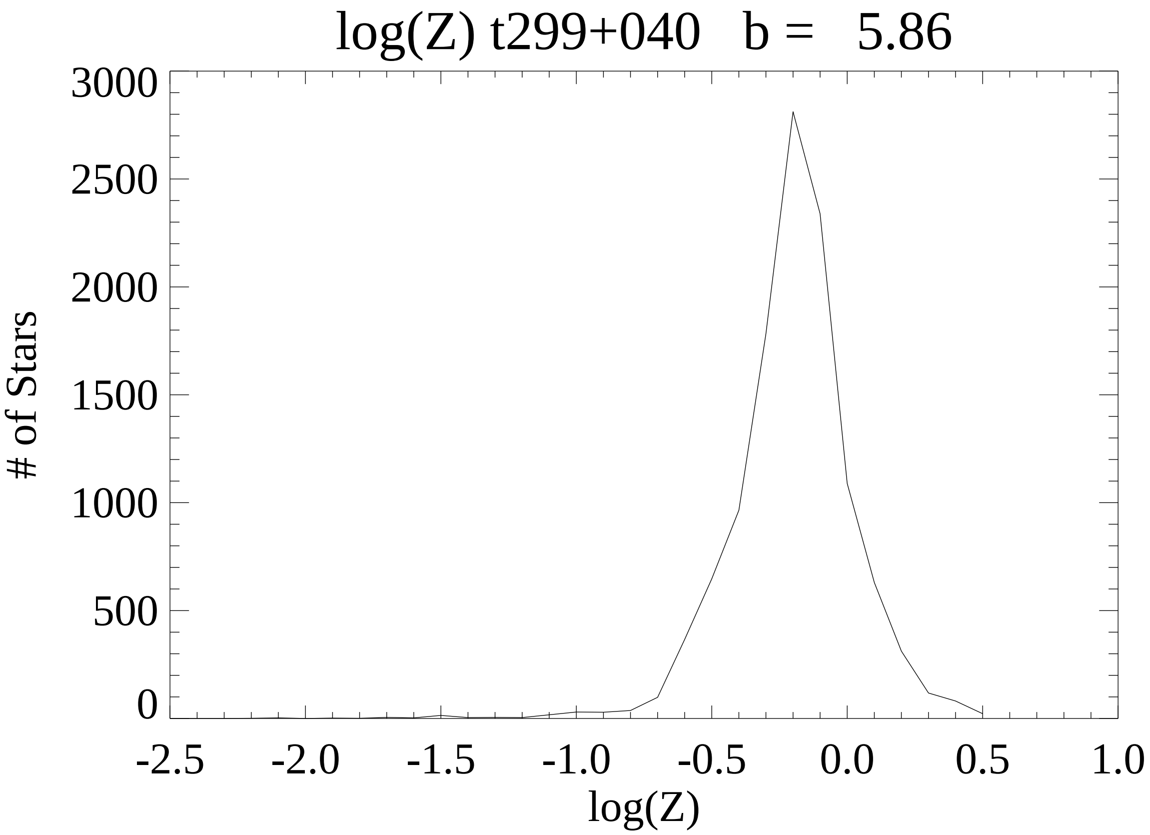

KIC10
R mag
Log(Teff)


KIC10
Log(L)
Log(Teff)


t299+040 Giants b = 5.86
Dereddened g-D51
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.5
1.0
1.5
Dereddened g-r


t299+040 Dwarfs b = 5.86
Dereddened g-D51
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.5
1.0
1.5
Dereddened g-r


g-r vs Teff t299+040 b = 5.86
Observed g-r
Teff (1000K)
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
20
16
12
10
9
8
7
6
5
4
3


log(Z) t299+040 b = 5.86
# of Stars
log(Z)
3000
2500
2000
1500
1000
500
0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0