First Look at Photometric Reduction *via* Mixed-Model Regression

Eric Dose

Bois d'Arc Observatory, Auburn, Kansas :: eric@bois-darc.org :: https://github.com/edose

Bois d'Arc Observatory
Kansas

presented at the Spring 2016 meeting of the AAVSO, St. Louis, May 6-7, 2016

Mixed models

Mixed-model regression is presented here as a new approach to photometric reduction.

Mixed-model regression originated in the social sciences as an advance in data reduction that begins with normal multivariate regression, then adds "random effects" to account for confounding influence of categories. Mixed-model regression very effectively models measured data having systematic errors or biases by fitting both (1) normal parameters (fixed effects) and (2) categorical parameters (random effects). Mixed-model regression best applies to modeling of substantial data sets that suffer from both per-point **random errors** and per-category **systematic errors.**

Astronomical photometric data reduction can be posed as *exactly* such a modeling problem. At 2014's SAS meeting I presented early simulation evidence that mixed-model regression could isolate and account for—that is, effectively remove—60-80% of certain photometric systematic errors, especially for:

per-image errors, due for example to exposure time jitter or passing cirrus; and

per-catalog-star errors, due to minor errors in catalog or sequence magnitudes.

The simplest generally useful regression on comp-star Instrumental Magnitudes $\mathbf{I_V}$ (the dependent variable, here in V filter),

Catalog Magnitudes $\mathbf{M_V}$ (known slope=1),
Transform $\mathbf{T_V}$ and Color Index **CI**,
Airmass **A** (fixed effect parameter), *and*
Image ID **IID** (a systematic "random effect")

would be expressed in the R language's "lme4" package as:

I_V ~ offset(M_V) + offset(T_V*CI) + A + (1 | IID)

yielding the extracted coefficients : intercept (the zero-point), extinction in V, and per-image random error. One begins to see the value of mixed modeling of photometric data when one sees that the per-image random error is fixed to be equal for all comp stars in the image; that is, it serves exactly the same purpose as ensembles of comp stars, with the advantage that one can plot the per-image random error vs time (a "cirrus plot") to evaluate :

drift in sky transparency,
shutter jitter,
dew forming on the optics, etc

for each image, across the same or varied targets, all night long.

In practice, I build a model on all the comp stars observed in an entire night, across all fields of view. This yields so many regression degrees of freedom that it's straightforward to extract additional parameters. For example, an extended model using physically realistic parameters might be:

I_V ~ offset(M_V+T_V*CI) + (1 | IID) + A + X + Y + "Vignette"

where additional parameters X and Y are the X- and Y-positions of the individual stars' centroids on the CCD (to allow for consistent asymmetric vignetting), and "Vignette" is the star's squared distance from the CCD center (to remove any consistent parabolic vignetting not removed by flat-calibration).

My **current mixed-model workflow**, beginning with a table of all comp, check, and target instrumental magnitudes (aperture) and metadata from fully calibrated images:

1. Collect all the night's eligible (high S/N, non-saturated) comp-star data into one R data frame.
2. Perform a mixed-model regression for each filter,
3. Examine each of 14 diagnostic plots, investigate outlier images or star observations,
4. Omit images and/or observations as warranted.
5. Repeat steps 2-4 in each filter until convergence.
6. Use the mixed model to "predict" Instrument Magnitudes for all check stars (as unknowns) and target stars while fixing M = 0 and CI = 0. Assign the offset from measured instrumental magnitude as best estimate of the true, untransformed magnitude.
7. Impute targets stars' (time-dependent) V-I Color Index from V & I measurements, solve the simultaneous equations; apply transformations using imputed Color Index values.
8. Evaluate using final diagnostic plots and scripts to check for consistency. Curate and/or average results if warranted, then auto-generate, proof, and submit AAVSO-format report.

This approach characterizes individual images only by their cirrus-effect terms and perhaps image airmasses. Comp stars are assumed to have known, constant magnitudes and colors. The model currently assumes constant extinction all night, but a mid-2016 version will extend the model to allow constrained, smoothly varying extinction curves over several hours.

This workflow is highly automated via extensive R scripts. A typical night's observing session will generate 1000-3000 comp star observations and 50-200 reportable magnitudes, which usually requires about 1.5-2 hours' reduction effort from raw FITS images to finished AAVSO report.

In the optical system used in the examples to upper right (Edge C14 + SBIG 6303E), calibrated images typically yield X and Y coefficients of less than 5 mMag edge-to edge, and the Vignette coefficient is typically well under 10 mMag edge-to-edge. Experiments with artificially biased and/or uncalibrated images show that up to 400 mMag of vignetting bias is removed accurately by one parabolic term.

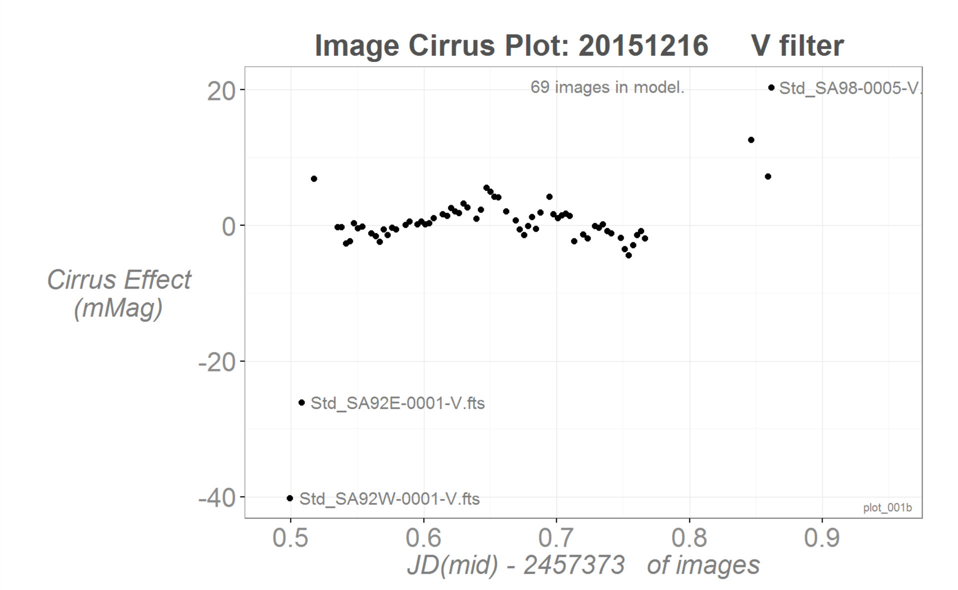
R Language

The present work is written in the free, open-source R language. R has become the reference statistics framework globally, popular outside North America for 25 years, notably in advanced statistics, molecular biology, commercial analytics, and quantitative finance and high-frequency trading strategies.

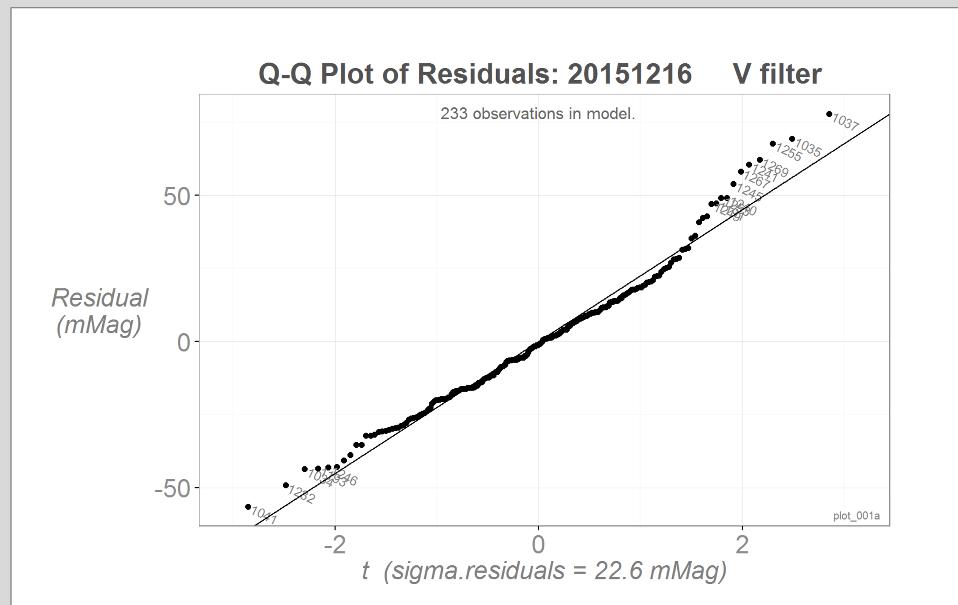
R is famous for its steep learning curve, but once learned it yields very fast, iterative exploration of one's own data, and rapid prototyping of new statistical algorithms. Vectors, matrices, and data tables are native variable types, and thousands of open-source packages are available.

But as a flat-script exploratory framework, R is really limited to code bases of perhaps 1000 lines of code. More complex workflows need the stricter code organization provided by more object-oriented languages like C# or Python. So, most extensions of this work are being written in Python 3 (especially as Python scripts can call R functions, which is easing the transition and preserves statistical rigor).

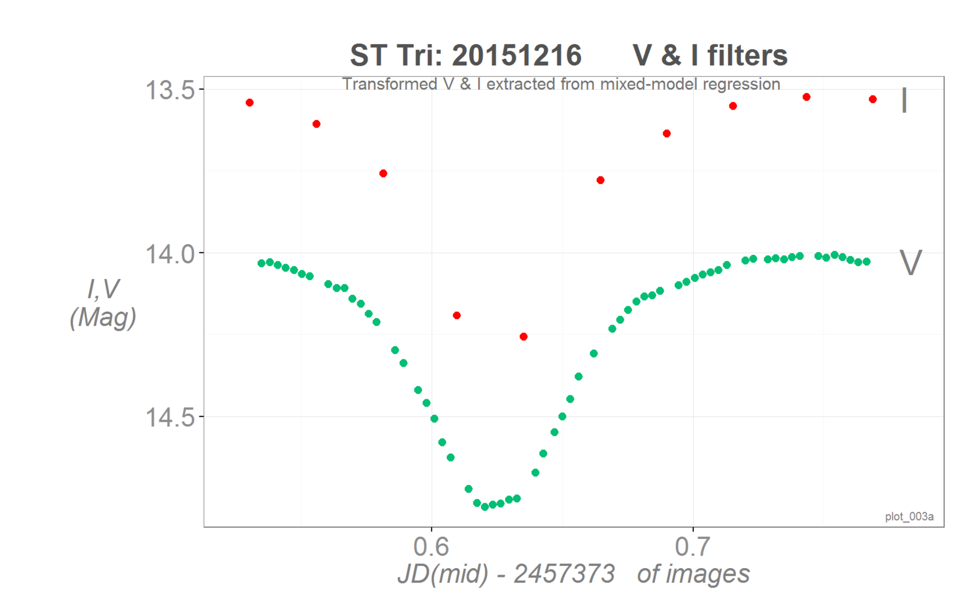
First experimental results



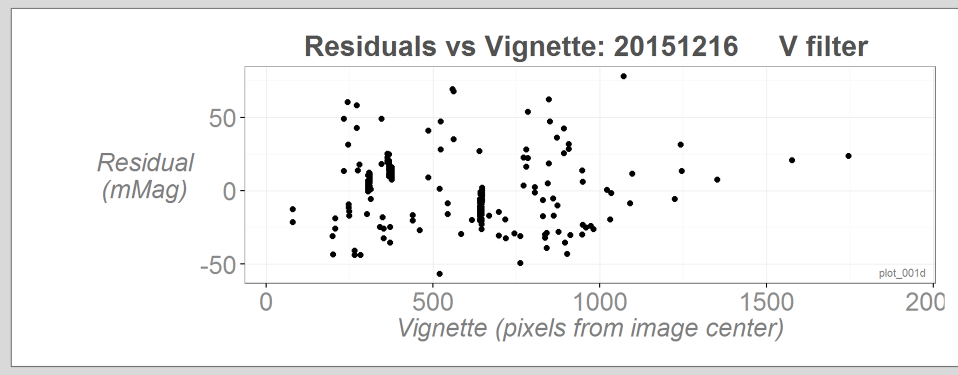
A **"cirrus plot"** (per-image systematic error vs time), the image-to-image magnitude variation of all ensemble comp stars. This plot emerges as a "random effect" from mixed-model regression on all comp-star observations in one filter. Early variations are from dusk twilight; late variations are from clouds moving in.

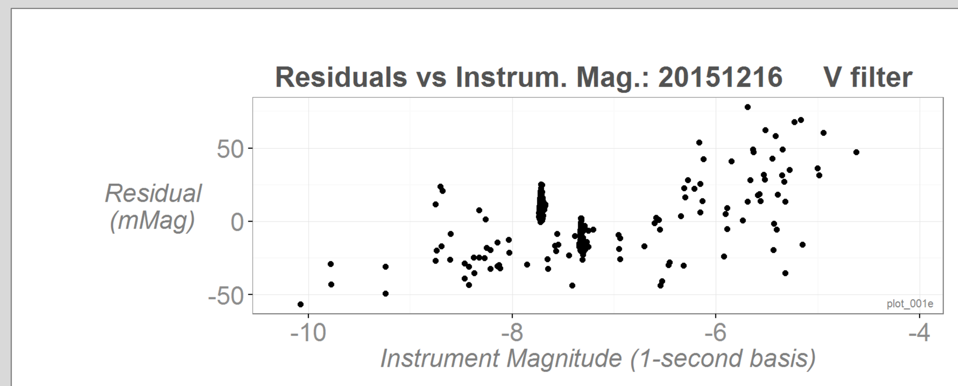


This classical **"Q-Q" plot** emerges naturally from mixed-model regression, plotting residuals sorted by size and against sorted normal distribution t-values. It exposes non-Gaussian errors and aids in responsibly removing comp-star outliers.



First light curves extracted via mixed-model regression: eclipser ST Tri imaged December 16 2015 in 14" SCT.





An extreme example of (uncorrected) systematic error as exposed by the mixed-model regression. This bias is caused by inaccurate sky-background estimation. Early corrections have since been applied, and methods are in further development *(SAS presentation, June 2016)*.

Next steps

The statistical power yielded by including all eligible comp-star observations in each night (1000-3000 in my recent Mira sessions) suggests several possibilities for extension:

1. Continue to test and improve this model and approach with new photometric data.
2. Develop and rigorously test new algorithms for robust estimates of sky background bias, to extend this photometric model to fainter sources.
3. Test (via Analysis of Variance) several new parameters, especially for time-dependent extinction coefficients, i.e., hourly rather than nightly.
4. Flag and report AAVSO comp stars estimated via at least 3 nights' data to have questionable catalog magnitudes (facilitated by the new model's consistency requirements.
5. Investigate combining multiple nights' data (from one optical rig) into one photometric model, adding a few random-effect terms to cover nightly differences.

References

1. Source code for this work is publicly available at: *https://github.com/edose/Photometry*
2. Andrew Gelman, Jennifer Hill, (2007), "Data Analysis Using Regression and Multilevel/Hierarchical Models", Cambridge University Press.
3. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015), "Fitting Linear Mixed-Effects Models Using lme4", *Journal of Statistical Software*, 67(1), 1-48. Description and usage of lme4 mixed-model package for the R language.
4. Eric Dose (2014), "Towards Millimagnitude Photometry", presentation at SAS 2014 Symposium, Ontario, California.
5. R Language, *https://www.r-project.org/*. RStudio environment for R (recommended, free) *https://www.rstudio.com/products/RStudio/*
6. Brian D. Warner (2006), "A Practical Guide to Lightcurve Photometry and Analysis", Springer Science.