Improved background subtraction for the Sloan Digital Sky Survey

Michael R. Blanton[1], Eyal Kazin[1], Demitri Muna[1],

and

Benjamin Weaver[1]

ABSTRACT

We describe a procedure for background-subtracting Sloan Digital Sky Survey (SDSS) imaging in a manner that improves the resulting detection and photometry of large galaxies on the sky. Our method treats each drift scan run as a whole and fits a smooth function to variation of the sky background to a heavily masked and binned image of the run. We have run our procedure on the full Data Release 7 for the Northern Galactic Cap. In this region, we have generated 1 deg by 1 deg mosaics for general distribution. As a test of quality, we compare the photometry of point sources in these mosaics relative to the SDSS catalog. **and find?** More critically, we have also tested the effect of our background subtraction on the photometry of large galaxies by inserting fake galaxies into the raw data and attempting to measure them after background subtraction. **and find?** These techniques and these mosaics will provide the basis of upcoming studies of the local Universe.

1. To-do list

mrb: run montage_image on a few patches and make QA

mrb: run montage_image on a few patches with large galaxies and compare to us

mrb: create residual qa test for each run

mrb: make example mosaics

dm: create web front-end and clipping routines for access

bw: build v5_6 for testing purposes

eak: simple comparison of fake nearby to us

eak: dimage comparison of fake nearby to us

eak: photo v5_4 comparison of fake nearby to us

eak: photo v5_6 comparison of fake nearby to us

2. Why improve the sky-subtraction?

Background subtraction of astronomical images is probably a formally impossible task. The ideal treatment of the data gathered by a detector would be to explain (at reasonable χ^2) the counts in each

[1] Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place, New York, NY 10003

pixel, using a physical model of the sky brightness and its gradients, the telescopic optics and scattered light properties, the astronomical and other sources of light, and the sensitivity, noise and other properties of the detector. However, this task is impractical and likely intractable as well, given the detail needed in such a model, the time variability of conditions and instruments, and the probable necessity of treating all the data simultaneously. For this reason, most practical applications of background subtraction tend to focus on simple approximations that are tractable as well as close to correct for the problems of interest.

For the Sloan Digital Sky Survey (York et al. 2000), the largest existing survey of the sky to date, the standard photometric pipeline takes just such a practical and accurate approach (`photo`; Lupton et al. 2001). In the version released with Data Release 7 (DR7; **?**), calculates the median-smoothed background on a scale of 100×100 arcsec and subtracts that from the image before object detection and measurement. For faint point source photometry in this data set — that is, not a confusion-limited one — this approach is highly accurate. In almost all cases, any diffuse light is subtracted away (whatever its source) while the point source itself is left untouched. For these sources, it is not necessary to fully model the background — just to separate point sources from any diffuse sources. The resulting fluxes of stars are generally repeatable at well below the percent level; that is, as well as can be expected given the vagaries of determining the overall calibration of the data set (Padmanabhan et al. 2007a).

However, for galaxies the approach that `photo` uses is more troublesome. While galaxies of small enough angular extent are treated similarly to the stars, larger galaxies can be substantially affected. For typical galaxies this method of sky subtraction causes an underestimate in their flux, size, and concentration (Blanton et al. (2005)). This problem becomes particularly evident for the nearby brightest cluster galaxies (Bernardi et al. 2007; Lauer et al. 2007).

In §3 below we present a different method for sky subtraction that is more accurate for large galaxies (while retaining most of the accuracy for point sources). Our method begins with the estimate of the sky background from a processed SDSS imaging run. It builds masks around all bright detected sources, and any known sources close to but outside the edge of the imaging run. Then it fits a smooth spline to the unmasked data, applying appropriate constraints to regularize the problem in the presence of the heavy masking.

In §4 we describe how we take sky subtracted SDSS fields and mosaic them into single image. We have produced 1 deg by 1 deg FITS format mosaics over the entire Northern Galactic Cap covered by the SDSS, and provide access to snap-shots up this area on the web.

In §5, we test the resulting photometry of these images in two ways. First, we compare aperture photometry of point sources to the standard SDSS aperture photometry, to quantify how much degradation our mosaicking procedure and sky subtraction have introduced.

Second, we insert bright galaxies into the raw data and reanalyze the data completely. We then compare our background-subtracted mosaicked images to the original fake data to evaluate the fidelity of our procedure. We also evaluate the performance of two versions of the standard SDSS pipelines (one that corresponds to the SDSS DR7 and another that is intended for DR8).

The only other publicly available mosaicking and background subtraction facility for SDSS that we are aware of is the results of the Montage package distributed by IRSA **cite**. We compare their methodology to ours in §4, concluding that their background subtraction procedure in principle could be superior to ours, but that their "drizzle" approximation to interpolation is likely to cause errors for PSF photometry. In §5.1, we demonstrate that one the PSF photometry is indeed degraded relative to the original SDSS photometry for Montage mosaics. In §5.2, we compare our bright galaxy photometry to that from Montage

images, **finding what?** As we describe below, these problems are likely not fundamental to the Montage background subtraction method.

3. A model fit to the SDSS sky background

3.1. The SDSS imaging data

The SDSS has taken *ugriz* CCD imaging of $> 10^4$ deg^2 of the sky (York et al. 2000; **?**). Automated software performs all of the data processing: astrometry (Pier et al. 2003); source identification, deblending and photometry (Lupton et al. 2001); photometricity determination (Hogg et al. 2001); and calibration (Fukugita et al. 1996; Smith et al. 2002; Padmanabhan et al. 2007b).

As Gunn et al. (2006) describe, the SDSS focal plane has 30 imaging CCDs evenly spaced (with gaps) across it, six "camcols" of each filter (u, g, r, i and z). Each exposure is taken in drift scan mode, such that a point in the sky passes through each CCD sequentially with a gap of a minute or so. The native pixel scale on the CCDs is approximately 0.396 arcsec; when we refer to pixels in this paper it is this native scale we will mean. When we discuss these images, we will refer to the position in the scan direction as the "pixel row" or y, and the position perpendicular to the scan direction as the "pixel column" or x.

Figure 1 shows this geometry for part of run number 1336 in the r-band, showing each camcol horizontally and the gaps in between each one. Thus each such run can be considered to consist of thirty long rectangular images, one in each filter and camcol. In the standard SDSS pipeline, each camcol is further broken into multiple "fields," each with a width of 1361 pixels in the scan direction.

The image shown is approximately the standard SDSS pipeline's estimate of the background light used for photometry, binned 8×8. It is equal to the actual counts in areas where there were no detected objects, and is equal to the background estimate plus appropriate noise in areas where there were detected objects. This determination is made before any calibration except for the flat-field correction — which is only a function of pixel column since the observations are drift scans. It is about 2.3 deg wide and represents about 25 minutes of exposure time. The units are raw counts — because of the varying gains among the CCDs two of camcols mostly saturate in this particular stretch.

SDSS's background light estimation consists of a median-smoothed image on a scale of 100×100 arcsec. Clearly Figure 1 demonstrates that bright objects have a substantial amount of their light removed during background subtraction — the background shows clearly the presence of bright stars. We will proceed by fitting a much smoother function to this SDSS sky estimate in order to model the clear variation with the background with time but not oversubtract the brightest objects.

3.2. Masking the data

The first step is to mask the data appropriately. Clearly whatever smooth function we use, we cannot allow it to be strongly affected by the presence of bright stars. We begin by defining an initial mask around known bright objects.

First, we identify any objects detected in the SDSS catalog with $m < 15$ in the filter in question. We create a mask of size 32×32 native SDSS pixels around the faintest such objects, growing with decreasing magnitude to 1600×1600 pixels at $m = 12$ (and constant with magnitude for brighter objects). This mask

handles stars inside the imaging run well.

However, galaxies within the imaging run are frequently shredded into fragments by the SDSS deblender, and it is more complex to determine what area to mask based on the SDSS catalog alone. Therefore, we also identify any galaxies from the Third Reference Catalog (RC3; de Vaucouleurs et al. 1991) that land within the imaging run and build masks around those objects that are 1000×1000 pixels.

Finally, the SDSS filters suffer from internal reflection at their edges, causing stars just outside the frame to scatter light onto the CCD. Our model is too smooth to subtract these internal reflections, and we do not want its large scale features affected by these local phenomena. Therefore, we mask these reflections by identifying any stars from the Tycho-2 catalog that lie just outside the frame and masking a rectangular region 1500×900 pixels centered on the star (the reflection has a larger extent in the pixel column direction than in the row direction).

Occasionally SDSS fields have such a bright object inside of them that the SDSS pipeline fails to reduce them at all. We mask the entirety of such fields from our fits.

Figure 2 shows this initial mask for the run shown in Figure 1. White areas are those to be used in the fit, while black areas are those to be ignored. The mask is intentionally very conservative — we want to minimize the flux that is incorrectly assigned to the sky background. Furthermore, as described in the next section, we apply further iterative masking based on the results of our fit.

3.3. Spline model fit

In order to fit the data, we first further bin it from 8×8 to 8×680 — that is, we heavily bin it in the pixel row direction. We keep the resolution in the pixel column direction because the sky ends up tracking some residual flat-fielding errors. When rebinning, we account for the weights.

The spline model that we fit to the final binned data is as follows:

$$f(\mathrm{camcol}, x, y) = S(\mathrm{camcol}) \times Y(\mathrm{camcol}, y) \times X(\mathrm{camcol}, x, y), \tag{1}$$

where x and y are the pixel column and row positions. $S(\mathrm{camcol})$ is an overall scaling factor for each camcol. $Y(\mathrm{camcol}, y)$ is a second-order b-spline for each camcol, with break-points spaced approximately per SDSS field (1361 pixels). Thus, it expresses the overall variation of the sky over time during the scan. $X(\mathrm{camcol}, x, y)$ is a second-order b-spline in two dimensions, with break-points in the x direction spaced once every 8 pixels and in the y direction spaced every 40 fields. Thus, it allows for a rapid variation with column (which accounts for flat-fielding errors) but allows that pattern to vary slowly over time.

To fit the data, we (approximately) perform χ^2 minimization of the data with respect to the model. In detail, we first determine S by taking a median across each camcol.

Second, we divide each camcol by S and perform χ^2 minimization to fit Y alone. We quadratically couple the spline parameters strongly in the row direction to keep the fit smooth. In addition, we allow a small quadratic coupling between the camcols in order to interpolate more smoothly over masked data. These quadratic couplings retain the linearity of the χ^2 fit.

Third, we divide the data by S and Y, and finally perform a second χ^2 minimization, this time for X. Again, we allow coupling between the spline parameters. In this case, we allow weak coupling between the parameters in the row direction but very strong coupling in the column direction, to keep X as smooth as it

can be. In the χ^2 minimizations for Y and X, we use the mask weights, but no inverse variance weighting.

We have a special condition for CCDs that have two amplifiers rather than one. For such CCDs, half the pixels are read out by one device and half by another; therefore, we do not couple the pixels across the divide, and indeed we observe that the gains of the amplifiers do drift at the fraction of the percent-level over the course of each run.

Finally, given a fit function $f(\mathrm{camcol}, x, y)$ we estimate an rms dispersion σ around the fit, and then mask all regions that are greater than 2σ from the fit (in addition to the initial mask described in §3.2). We then iterate the fit four more times.

An example final fit is shown in Figure 3. This model is a non-parametric, not physical — consequently, for sufficiently diffuse features (many arcminutes in size) our method will by design oversubtract the light. However, Section §5.2 below shows that it is sufficiently good to allow accurate measurements of most nearby galaxies.

3.4. Residual tests

For each run, we have performed a simple residual test by considering random unmasked patches 32×32 pixels in size. Figure 4 shows the distribution of the mean fluxes in such patches across all SDSS runs, as well as the residuals as a function of run number. **need to make this and comment on it**

These residual tests are adequate insofar as they test the performance of our fitting procedure. However, they do not provide a direct estimate of the effects of sky subtraction errors on the resulting galaxy photometry. We will test that more thoroughly in §5.2.

4. Mosaicking the SDSS

4.1. Generating mosaics

The point of our sky subtraction is to improve measurements of large objects on the sky. Many such objects overlap the edges of SDSS fields (and some are bigger than or comparable to the size of a single field). Thus, to recover their photometry we must mosaic together multiple fields after the background subtraction. Here we describe our procedure for doing so.

For each desired mosaic we specify a desired World Coordinate System header (WCS; **cite**). We identify all SDSS fields that overlap the desired area, and pick the minimal set of photometric fields that cover that area.

Then we prepare each field for mosaicking. Starting with the raw SDSS data, we apply the flat field determined by the ubercalibration procedure (Padmanabhan et al. 2007b). Using a procedure similar to that used by the SDSS pipeline, we identify cosmic rays (Lupton et al. 2001). We interpolate over saturated pixels and cosmic rays using simple linear interpolation in the x direction. Then we estimate the sky by evaluating the function described by Equation 1 at each pixel, and subtract it. In the sky estimate, we account for any difference in the flat field as originally used by the SDSS photometric pipeline, and as determined by ubercalibration. Finally, we apply the photometric calibration using the results of Padmanabhan et al. (2007b) for each field.

To resample each image we evaluate the position of each desired pixel within each original field. Then we interpolate the original field's image to the desired set of locations, using the well-known bicubic interpolation approximation to the sinc function (**cite**). This resampling method is known to work well for Nyquist-sampled images, a condition SDSS virtually always satisfies. In practice, we use the built-in IDL utilities `polywarp` and `poly_2d` to perform this resampling.

We evaluate the weighted average of the flux from all the input images at each output pixel. The weights are unity throughout most of each input image, but are apodized smoothly to zero at the edges. Thus, the transition between regions which overlap two fields are relatively smooth. Any output pixels that have no overlapping input images are set to zero.

4.2. Example mosaics

Figure 5 show several example mosaics created with this procedure. **make and describe**

4.3. Distribution of mosaics

Although our procedures allow us to evaluate any arbitrary mosaic, for ease of distribution we have created a set of 1×1 deg mosaics across the entire observed Northern Galactic Cap, or almost 8000 deg^2. We have created a web interface that allows users to easily extract sections of these mosaics.

how to access data

4.4. Comparison to Montage

The only other publicly available mosaicking and background subtraction facility for SDSS that we are aware of is the results of the Montage package distributed by IRSA **cite**. We will describe the quantitative differences between their results and ours below. Here we discuss methodological differences.

The first major difference is in their method for background subtraction. While in our case we rely on a smooth fit to regions with no objects, Montage fits a smooth additive term within each run to minimize the differences between it and other overlapping images. In regions with many overlapping images, this procedure can be far superior to ours. Over much of the SDSS area the only overlapping images are at the north and south edges of each field, and how the Montage algorithm behaves in that regime is a quantitative question. We compare our results to theirs for large galaxies in §5.2 below.

The second major difference is in their method for "resampling" of images. They do not perform a normal resampling but instead use a generalized version of the "Drizzle" algorithm. This algorithm computes the fraction of each original pixel that overlaps each output pixel, and distributes the original pixel's flux accordingly. Relative to properly resampling a Nyquist-sampled image, such methods do a poor job at maintaining the original PSF (or even at producing an image with a well-defined PSF). **refs??** They are also clearly irreversible operations even in the noiseless case, demonstrating that information is lost and the image is consequently degraded. We compare our results to the Montage results for point sources in §5.1 below, and believe that Montage's relatively poor performance in those tests is due to their drizzling procedure. However, it is unlikely that these issues affect our main interest here, which is large galaxies on

the sky.

5. End-to-end tests of the sky background

5.1. PSF photometry

comparison of PSF photometry to SDSS results

how does Montage do?

5.2. Galaxy photometry: isolating the effect of the background

how much does sky subtraction affect the photometry

direct comparison to Montage for some bright galaxies

5.3. Deblended galaxy photometry

Comparisons of mock data to dimage measurements

Comparisons of mock data to photo v5.4, v5.6

6. Summary

basic model

outline of tests

better than Montage: but Montage method a good idea

Funding for the creation and distribution of the SDSS Archive has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS Web site is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium (ARC) for the Participating Institutions. The Participating Institutions are The University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, the Korean Scientist Group, Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

REFERENCES

Bernardi, M., Hyde, J. B., Sheth, R. K., Miller, C. J., & Nichol, R. C. 2007, AJ, 133, 1741

Blanton, M. R., Eisenstein, D., Hogg, D. W., Schlegel, D. J., & Brinkmann, J. 2005, ApJ, 629, 143

de Vaucouleurs, G., de Vaucouleurs, A., Corwin, H. G., Buta, R. J., Paturel, G., & Fouque, P. 1991, Third Reference Catalogue of Bright Galaxies (Volume 1-3, XII, 2069 pp. 7 figs.. Springer-Verlag Berlin Heidelberg New York)

Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., & Schneider, D. P. 1996, AJ, 111, 1748

Gunn, J. E. et al. 2006, AJ, 131, 2332

Hogg, D. W., Finkbeiner, D. P., Schlegel, D. J., & Gunn, J. E. 2001, AJ, 122, 2129

Lauer, T. R., Faber, S. M., Richstone, D., Gebhardt, K., Tremaine, S., Postman, M., Dressler, A., Aller, M. C., Filippenko, A. V., Green, R., Ho, L. C., Kormendy, J., Magorrian, J., & Pinkney, J. 2007, ApJ, 662, 808

Lupton, R. H., Gunn, J. E., Ivezić, Z., Knapp, G. R., Kent, S., & Yasuda, N. 2001, in ASP Conf. Ser. 238: Astronomical Data Analysis Software and Systems X, Vol. 10, 269

Padmanabhan, N. et al. 2007a, ArXiv, astro-ph/0703454

Padmanabhan, N. et al. 2007b, MNRAS, 378, 852

Pier, J. R., Munn, J. A., Hindsley, R. B., Hennessy, G. S., Kent, S. M., Lupton, R. H., & Ivezić, Ž. 2003, AJ, 125, 1559

Smith, J. A., Tucker, D. L., et al. 2002, AJ, 123, 2121

York, D. et al. 2000, AJ, 120, 1579

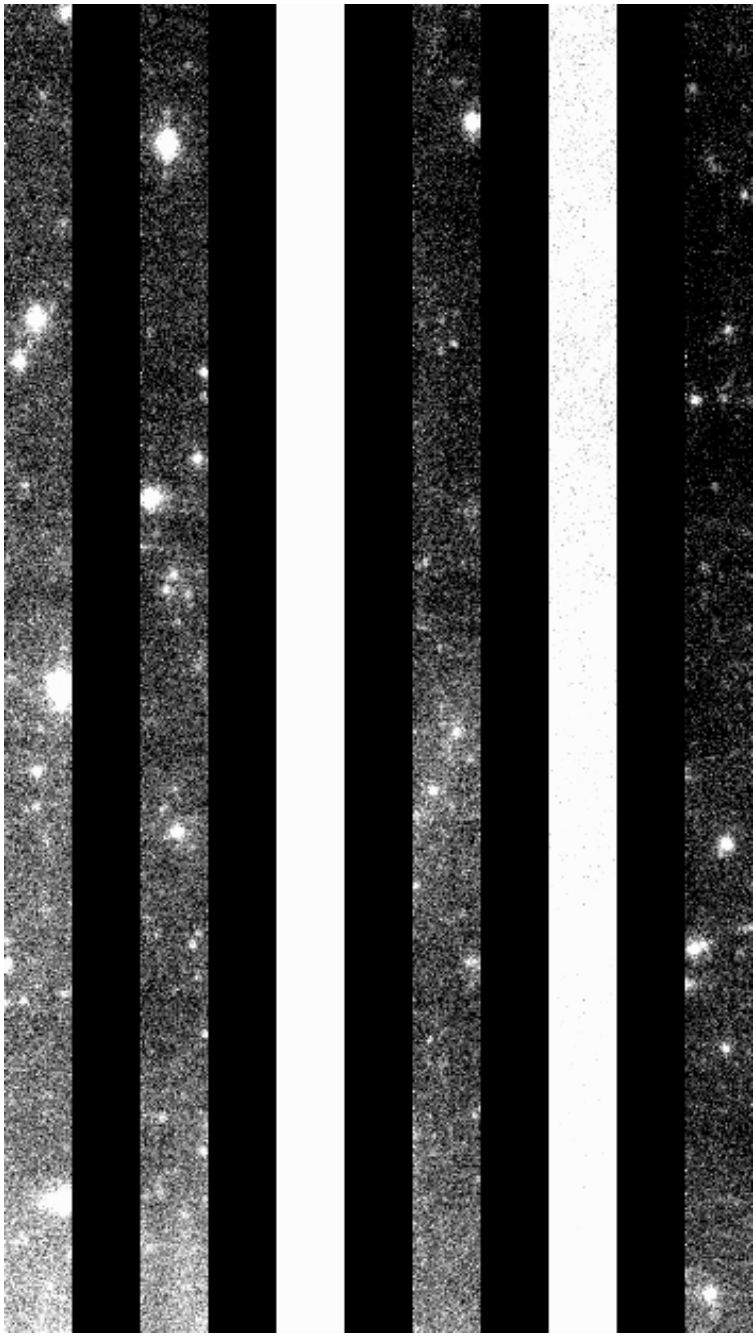

Fig. 1.— Data that our sky model is fit to, for part of an SDSS drift scan run. The vertical direction is the scan direction (y), the horizontal direction is perpendicular to the scan direction (x). This image is the 8×8 binned flat-fielded SDSS data. Areas where no objects were detected show the original data. Areas where objects were detected are replaced by the background sky estimate plus noise. Each of the six vertical stripes represents a "camcol" and the black areas in between are not covered by a CCD. The units of the image are raw counts, which therefore reflect the relative gains of the different CCDs.

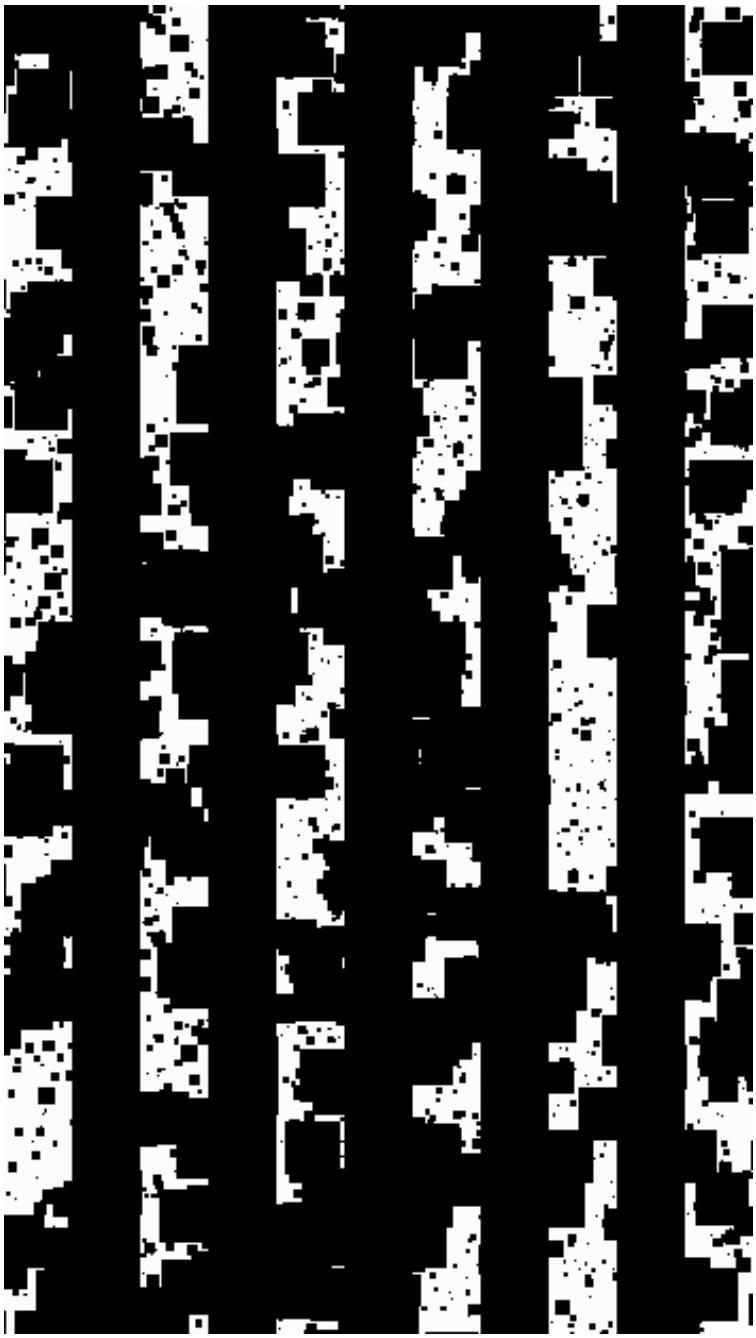

Fig. 2.— Mask applied to the data, for the same area of sky as shown in Figure 1. White areas contribute to the fit, black areas do not.

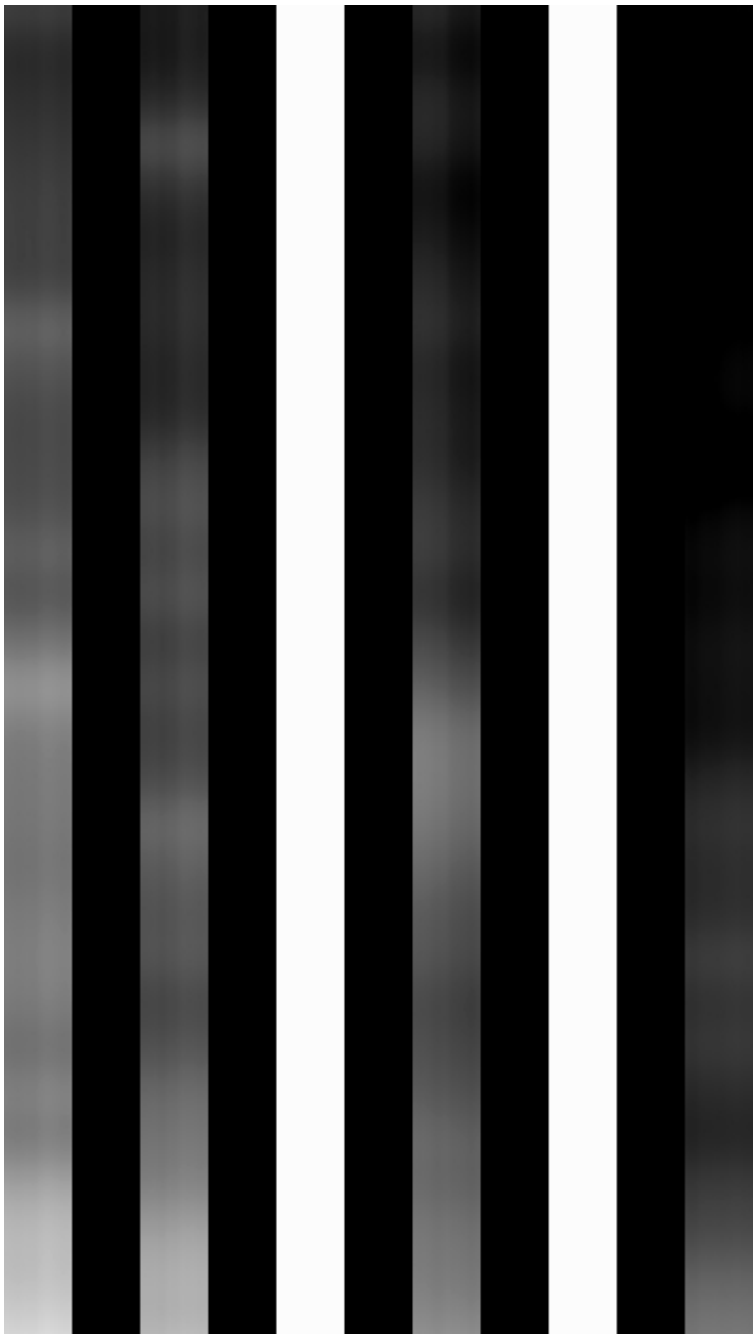

Fig. 3.— Final model sky fit, for the same area of sky as shown in Figure 1. This model represents an evaluation of Equation 1.

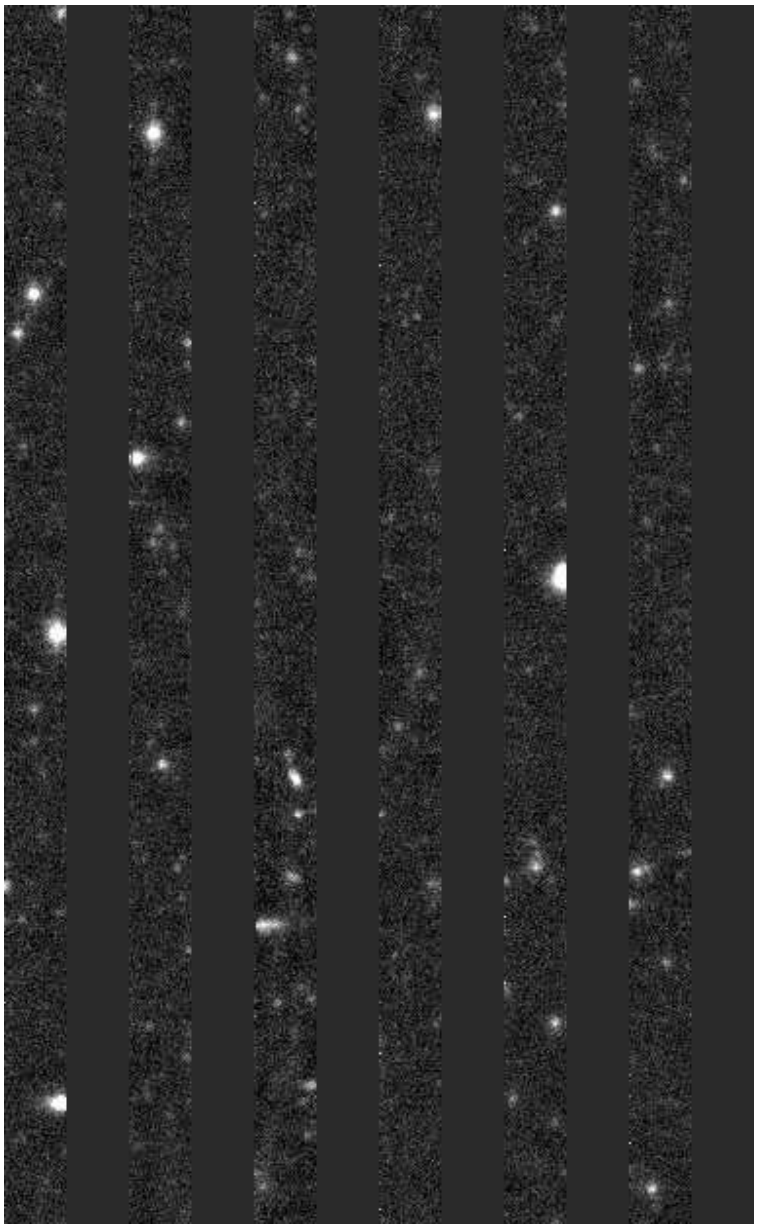

Fig. 4.— The residuals of the data from the model (literally Figure 1 minus Figure 3).

Fig. 5.— **make some examples here**

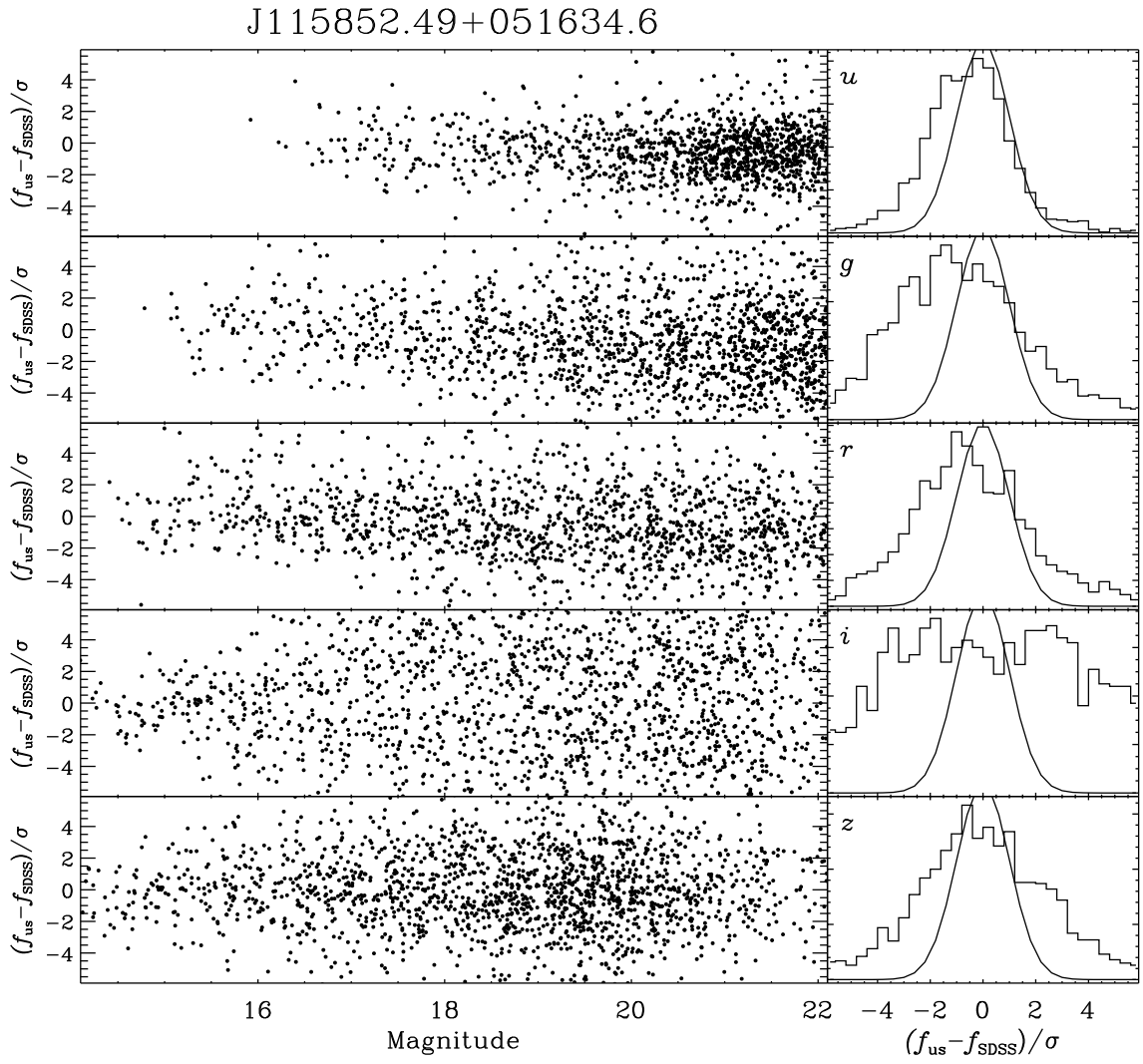

Fig. 6.— **montage QA plot**

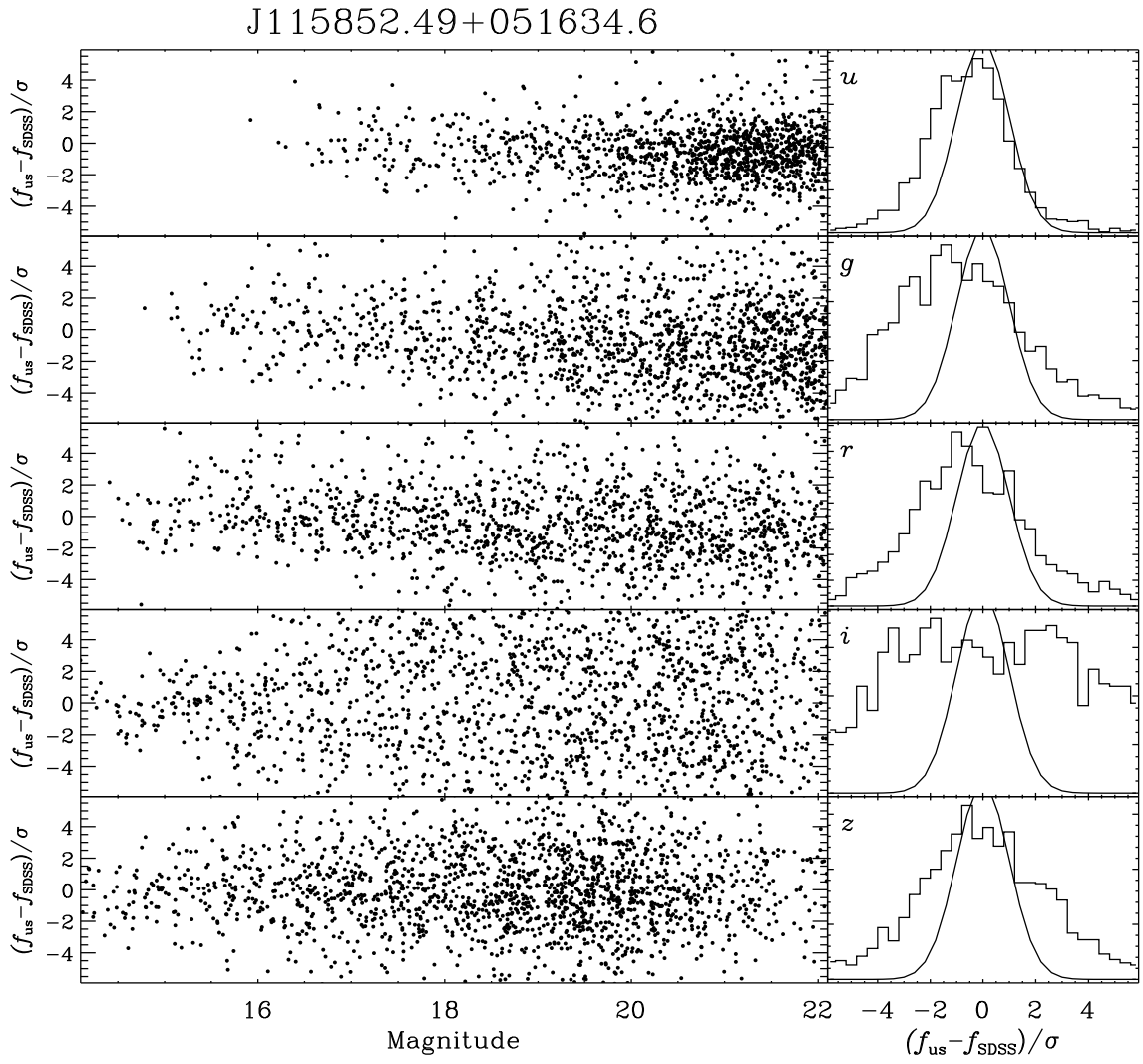

Fig. 7.— **montage QA plot**