Bad Pixel Mapping

Roger M. Smith*[1], David Hale[1] and Peter Wizinowich[2]

[1]Caltech Optical Observatories, 1200 E. California Blvd, MC11-17, CA 91125, USA;
[2]W. M. Keck Observatory, 65-1120 Mamalahoa Hwy, Kamuela, HI 96743, USA.

ABSTRACT

Bad pixels are generally treated as a loss of useable area and then excluded from averaged performance metrics. The definition and detection of "bad pixels" or "cosmetic defects" are seldom discussed, perhaps because they are considered self-evident or of minor consequence for any scientific grade detector, however the ramifications can be more serious than generally appreciated. While the definition of pixel performance is generally understood, the classification of pixels as useable is highly application-specific, as are the consequences of ignoring or interpolating over such pixels. CMOS sensors (including NIR detectors) exhibit less compact distributions of pixel properties than CCDs. The extended tails in these distributions result in a steeper increase in bad pixel counts as performance thresholds are tightened which comes as a surprise to many users.

To illustrate how some applications are much more sensitive to bad pixels than others, we present a bad pixel mapping exercise for the Teledyne H2RG used as the NIR tip-tilt sensor in the Keck-1 Adaptive Optics system. We use this example to illustrate the wide range of metrics by which a pixel might be judged inadequate. These include pixel bump bond connectivity, vignetting, addressing faults in the mux, severe sensitivity deficiency of some pixels, non linearity, poor signal linearity, low full well, poor mean-variance linearity, excessive noise and high dark current. Some pixels appear bad by multiple metrics. We also discuss the importance of distinguishing true performance outliers from measurement errors. We note how the complexity of these issues has ramifications for sensor procurement and acceptance testing strategies.

Keywords: bad pixel, operability, histogram, Teledyne H2RG

1. INTRODUCTION

Bad pixel mapping has not attracted much discussion in the literature, even though the concept is widely invoked, and has quite important consequences when setting detector specifications. Perhaps this is because the concept is deceptively simple: some pixels work so poorly that it is better to abandon their use completely. If a pixel is never used, it can legitimately be excluded from averaged performance metrics. For example the average dark current for the whole frame may exclude "hot pixels", or the average QE may exclude the effect of disconnected or masked pixels. This exclusion is only legitimate if the selection criteria are consistently applied to both bad pixel mapping and the calculation of those averages.

The definition of bad pixels, *and detector specifications in general,* will differ significantly among manufacturers and users. The metrics used (dark current, QE, linearity etc.), measurement methods, and the specific thresholds will vary. We propose that the remedy for the resulting confusion cannot be the standardization of methods for setting specifications, since the vendor and customer have fundamentally different motivations *and capabilities*.

The manufacturer's definition of what constitutes a "bad pixel" is necessarily somewhat arbitrary, and represents a compromise between what is technically feasible and what the *typical* customer needs. An essential feature is that the specifications be stable over time, particularly when defining bad pixels, so that the definition of cosmetic grades is stable and can thus be used as a benchmark for defining yields, and making yield predictions that are essential to set pricing.

* rsmith@astro.caltech.edu, phone 1-626-395-8780. SPIE 9154-16 (Montreal, 2014)

High Energy, Optical, and Infrared Detectors for Astronomy VI, edited by Andrew D. Holland, James Beletic, Proc. of SPIE Vol. 9154, 915417 · © 2014 SPIE · CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2057694

The user has quite different needs: performance specifications in general, and the "bad pixels" criteria to be discussed here, must be adapted for the scientific objectives, which are usually stated in terms of post-calibration performance. To derive detector performance requirements (expressed in terms of raw data metrics such as linearity, crosstalk, noise etc.), one must understand the calibration methods and their accuracy, and the way in which errors propagate to the science metrics (e.g. centroiding accuracy or systematic error in galaxy ellipticity).

Given the subtlety and complexity of the calibration and test process, and the cost of pushing technological limits, it is usually impractical to impose project specific requirements on the manufacturer, beyond choosing between performance options already offered. It is more helpful to all concerned to recognize that it is reasonable and necessary for the vendor to use fairly stable test methods and criteria, which only evolve in response to market trends, while the customer implements a more detailed acceptance test process whose purpose is to show that the sensor performance is adequate for the application.

This acceptance testing is dependent on (and can foster) the development of calibration methods, and modeling efforts required to understand how the pixel-to-pixel distribution in detector characteristics interacts with the distribution of signals in a typical scene to influence scientific productivity. Exposure times and observing strategies also influence the outcome.

2. DEFINITIONS

Bad pixel mapping is a subset of the specification problem described above. We believe that it is reasonable and necessary for manufacturers to adopt different criteria to users. To avoid confusion, we propose different terms for manufacturer and user specifications:

> ***Inaccessible pixels***: In CCDs, pixels may be rendered inaccessible by deep charge traps, hot pixels, or even a clock electrode failure. In CMOS (including NIR) detectors some pixels may be inaccessible due to multiplexor faults.
>
> ***Inoperable pixels*** are either *inaccessible* or produce no useful signal. In hybrid NIR detectors this may be due to a bump bond failure or gross non-linearity. We propose extending the definition of inoperable to include any pixel *designated as bad by the manufacturer* due to gross malfunction. (*In)operability* is expressed as fraction of all pixels.
>
> ***Bad pixels*** are the superset of those deemed *inoperable* by the manufacturer and those, which the user chooses to exclude in the belief that they would degrade the science, in spite of calibration efforts.
>
> ***Unilluminated pixels*** may lie behind a mask or be intentionally unconnected to the light sensing layer (e.g. reference pixels in the Teledyne HxRG detectors), or may simply lie outside the field of view of the optics. When discussing *bad pixel fraction*, an application specific metric, it is useful to only consider areas of the sensor, which are accessible and illuminated.

3. SAMPLE SCIENCE CASES

3.1 Deep Widefield Photometric Survey

In widefield broadband imaging surveys, it is common to combine multiple images that have been acquired with sufficiently different pointing that outlier detection using the data itself is generally sufficient to prevent errors due to bad pixels. Bad pixel mapping is thus relatively unimportant except as a predictor of observing efficiency. Bad pixel mapping may still come into play when subtler data quality problems (e.g. charge smearing or image persistence) create systematic image quality problems rather than clearly identifiable outliers.

3.2 Widefield survey for transient events

In this example, the *changes* from visit to visit are the key science signal, so bad pixel mapping is important to identify potential false positives. For a survey with a constant exposure time it is relatively straightforward to identify pixels whose signal is too low or noisy to be trusted. The threshold for bad pixel classification may vary with sky brightness

level. As a simplification, the worst-case scenario is usually adopted since the difference in the number of pixels rejected is small.

3.3 NIR tip tilt sensor for Adaptive Optics.

We describe this example in considerable detail below because it represents a case where bad pixel mapping is important, yet best strategy is far from obvious. The Keck-1 Adaptive Optics System has a NIR low order wavefront sensor, TRICK[1] which reimages the entire field of view of the AO system, picking off the (H or K_s) passband not being used by the science instrument, OSIRIS. One or more regions of interest (typically 4x4 in size) are read out at frame rates typically between 50 Hz and 1 kHz. There is presently no capability to produce differential motion between the science and wavefront sensor focal planes so a guide star falling on a bad pixel cannot be relocated without also shifting the science field.

When calculating the centroid, relatively little weight is given to the 12 pixels lying on the perimeter of the 4x4 region while very large weight is given to the inner 2x2 pixels, since the PSF is Nyquist sampled. Thus a bad pixel occurring on the perimeter can be replaced by an interpolated value or the gain for that pixel can be set to zero. A bad pixel in the inner 2x2 pixels has a significant impact on centroid estimation and is best moved to the perimeter by repointing slightly. When bad pixels are too close together, the shift required on the science sensor may become too large and some system performance degradation must be accepted.

For a bad pixel fraction β, at random locations the probability of needing to move the telescope to avoid a bad pixel in the central 2x2 pixels is

$$P_m = 1 - (1-\beta)^4 \approx 4\beta \quad \text{for } \beta << 1$$

while the probability of needing to interpolate over one or more pixels on the perimeter of a 4x4 region (sometimes preceded by a "move") is

$$P_i = 1 - (1-\beta)^{16} \approx 16\beta \quad \text{for } \beta << 1$$

Those accustomed to CCDs, which typically have very few randomly located bad pixels, are often surprised to see CMOS (including NIR detectors) with bad pixel fraction in the 1-4% range. It often comes as a surprise that the distance between bad pixels is then only 5 to 10 pixels.

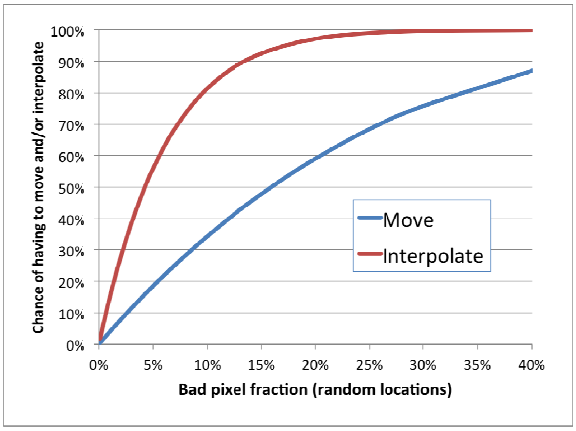

Figure 1: Probabilities of finding one or more bad pixels within a 2x2 region or a 4x4 region as a function of bad pixel fraction, β, assuming no clustering. These correspond to the probabilities of having to move the guide star by one pixel or interpolate over at least one bad pixel lying in the perimeter.

4. BOTTOM UP OR TOP DOWN?

The usual way to define bad pixels is to follow the lead of the manufacturer, who has no alternative but to describe performance in terms of conventional detector metrics such as sensitivity and dark current. The impact on science metrics like astrometry, photometry or image shear is then predicted with a suitably sophisticated model. This is the "bottom up" approach.

An alterative approach is to test the sensor in a scene projector, which allows the signal and observing cadence to be emulated. The planned calibration and analysis methods or a close proxy for them can be used to determine whether a given pixel delivers the required performance. For example, if transiting exoplanets were to be observed with a NIR detector, then a grid of spots could be projected and Allan Variance curves could be measured at different target locations[2]. Tests of the efficacy of decorrelation when the image is dithered (to emulate pointing jitter) can be used to classify individual pixels as good or bad.

This top down approach is feasible but is rarely employed. Its virtue is that it tests the device in the way it will be used. The problem, however, is that when a pixel delivers inferior performance, it may not be apparent why, without reverting to the bottom up approach. The top down approach, or "end to end test" is better employed to validate the modeling required to understand propagation of detector properties to science metrics. Hereafter we only consider the bottom up method wherein pixel properties are measured directly.

5. BAD PIXEL CRITERIA

The user classifies a pixel as "bad" when the inclusion of that pixel in the data set would degrade the outcome of the experiment more than excluding it. The definition of a bad pixel is dependent on the application, the calibration methods, accuracy required and even the observing conditions. In the example to be described (NIR tip-tilt sensing) the metric of interest is the centroid motion of a natural guide star. In other applications the appropriate metrics might be photometry, ellipticity or astrometry. As is often the case, the modeling of the impact of different pixel defects was not fully mature at the time that initial bad pixel maps were generated so some judgment calls had to be made.

These basic detector performance metrics were measured:

- Quantum efficiency
- Dark current (hot pixels)
- Linearity
- Read noise

There are many other performance metrics that could be used, such as inter-pixel crosstalk (IPC) and image persistence or reciprocity failure but these are more subtle effects, which are probably not relevant when measuring *changes* in centroid location with the only-moderate precision required in our application. We investigated conversion gain (e-/ADU) as a possible metric but, as discussed below, more work is needed to show whether these outliers are real.

Initially, the thresholds for designating a pixel as "bad" were set to 3σ above or below the mode as appropriate. For Normally (Gaussian) distributed data only 0.3% of samples would lie above this. By choosing this conventional threshold for outlier detection, it was our hope that only true outliers would be flagged and that there would be very few false positives due to statistical errors. Unfortunately most CMOS sensors and particularly the NIR arrays in question, exhibit quite extended tails in their distributions. For most of the parameters that we measured, the core of the distribution was normally distributed only because it was dominated by statistical errors: one often sees an inverted parabola around the mode in the log-scaled histograms shown below. Only the extended tails appeared to be the real fluctuations in pixel performance.

For read noise, setting bad pixel threshold to "mode+3σ" resulted in labeling ~4% of pixels as bad: this would have requiring repointing 15% of the time and interpolating at least half of the time. Repointing this frequently is undesirable particularly given that there may be several guide stars and thus an even higher probability of encountering a bad pixel. Furthermore, in long exposures differential atmospheric refraction moves the guide star relative to the science field. Given that there are a range of guide stars and that signal can be increased by decreasing frame rate (at the expense of closed loop bandwidth), it seemed preferable to tolerate higher noise or select a different guide star rather than move this often.

Therefore we made the *arbitrary decision* to invert the process and set the noise thresholds to values that produce ~ 1% bad pixel fraction in the accessible area. These thresholds (shown in Table 2) can be used to predict worst-case performance as a function of source brightness for each of the frame rates. Yet another approach would be to define the noise threshold to be some *multiple* of the mode.

However one chooses to define the bad pixel threshold, assessment of the impact of these decisions requires system modeling in which pixel performance is drawn from the various distributions (or maps), and the signal modeled must be drawn from a representative distribution of source brightnesses. This modeling effort was outside the scope the camera

team's funding (at Caltech) and was left to the system integration team (at Keck). Though somewhat arbitrary, Table 1 and Table 2 represent the inner and outer bounds of the choices that might reasonably be made.

Table 1: bad pixel fraction as a function of frame rate for noise < mode+3σ; sensitivity < 25% of mode; dark current < 5000 e-/s/pix at 100C; intensity ratio (after dark subtraction) < 2.6 for 3 fold exposure time increase. TRICK frame rates are typically 50 Hz to 1 kHz. [Conversion gain 1.15 e-/ADU]

Frame Rate	Threshold [ADU]	Bad pixel fraction
8110 Hz	12.05	4.2%
4055 Hz	9.14	4.1%
2028 Hz	7.12	4.2%
1014 Hz	5.88	4.1%
507 Hz	5.30	3.9%
300 Hz	5.12	3.7%
100 Hz	5.18	3.3%
50 Hz	5.66	3.0%
20 Hz	7.01	2.5%
10 Hz	8.64	2.3%

Table 2: noise thresholds to obtain ~1% overall bad pixel fraction given sensitivity < 25% of mode; dark current < 5000 e-/s/pix at 100C; intensity ratio < 2.6 for 3 fold exposure time increase. [Conversion gain 1.15 e-/ADU]

Frame Rate	Modal Noise [ADU]	Mode+3σ [ADU]	Threshold for ~1% loss [ADU]
8110 Hz	8.52	12.1	20.7
4055 Hz	6.39	9.1	16.2
2028 Hz	4.85	7.1	13.7
1014 Hz	3.85	5.9	12.7
507 Hz	3.31	5.3	12.8
300 Hz	3.06	5.1	13.3
100 Hz	2.62	5.2	15.8
50 Hz	2.58	5.7	18.7
20 Hz	2.76	7.0	25.7
10 Hz	2.84	8.6	34.6

6. THRESHOLD SELECTION

6.1 Accessibility

The mask shown in Figure 2 was made by applying a threshold at 25% of the modal value to a flat field. This identified pixels that lie outside the circular illuminated area, or which have very low sensitivity. Clustered low sensitivity pixels were marked as "inaccessible" since these were concentrated in two areas at the bottom of the frame of this engineering grade detector. Lines 1361 to 1490 were also marked as inaccessible since prior tests had shown that the line pointer does not advance correctly through this range of rows even though windowed readout functions correctly on either side.

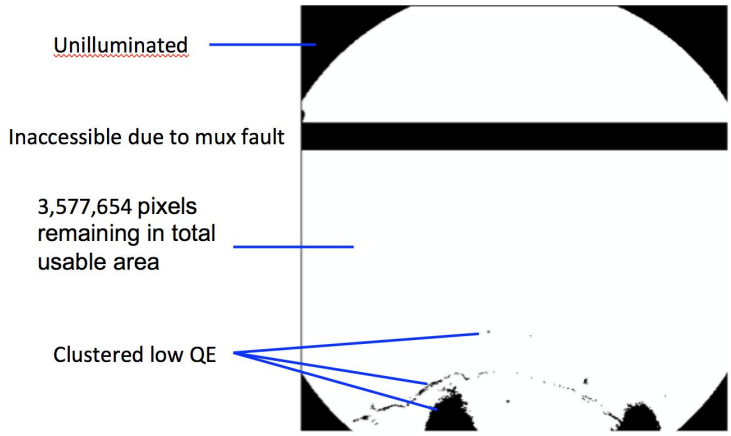

Figure 2: Inaccessible pixels are marked with zeroes while accessible pixels are marked with ones.

6.2 Dark current

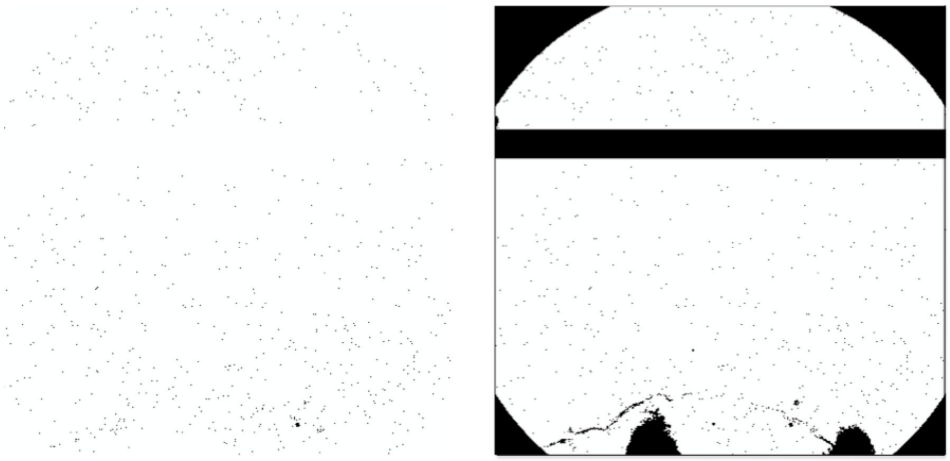

Figure 3: The hot pixel mask (>5000 e-/s/pix at 105C) is shown at left while the combination of inaccessible and hot pixels is shown at right.

At the high frame rate used for this application, dark current is not a significant problem. This is fortunate since the particular detector used has a higher than typical hot pixel population requiring operation at very modest detector bias (DSUB-VReset =175mV). The "Orca" Joule-Thompson cooled made by Advanced Research Systems was selected since it featured a non flammable refrigerant as required by the Keck Safety Office, but its cold head only reached 95K to 100K. With several contact resistances in series the detector temperature was typically 105K. Figure 3 shows that there are a large number (0.37%) of randomly distributed hot pixels that exceed even a 5000e-/s/pixel threshold at 105K. In setting this high threshold we were conscious that the hot pixels would be flagged in the noise maps. The purpose of

the additional test for hot pixels is to identify pixels that might saturate prematurely or fail to subtract out due to time dependence of dark current due to thermal drifts, etc.

The hot pixel map is an example where the vendor and user criteria differ greatly. Many of these hot pixels will freeze out at lower temperatures. Many will come back at lower frame rates.

6.3 Sensitivity

Pixels with low sensitivity are still useable if they are linear, and have acceptable noise and dark current. A histogram of such pixels (Figure 4) shows a low tail that is probably due to surface contamination or bump bond connectivity failures. Setting the threshold at 25% eliminates a comparatively small fraction of pixels (0.07%).

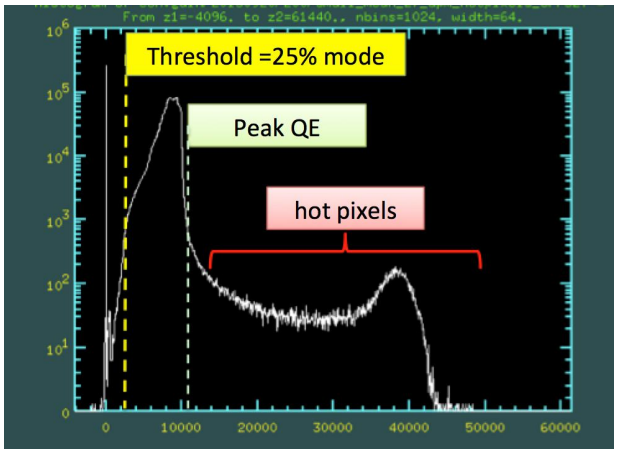

Figure 4: Histogram of flat field after removing pixels that are deemed unacceptable for other reasons (unilluminated, inaccessible, high dark current, non-linear, noisy).

To make this map it was necessary to take data using a region of interest so that the frame rate was high enough to avoid misleading results due to hot pixels saturating. A 2048x15 region was moved iteratively across the whole focal plane. This method was also needed so that ROI readout could begin beyond lines 1361 to 1490 where the line pointer does not increment correctly.

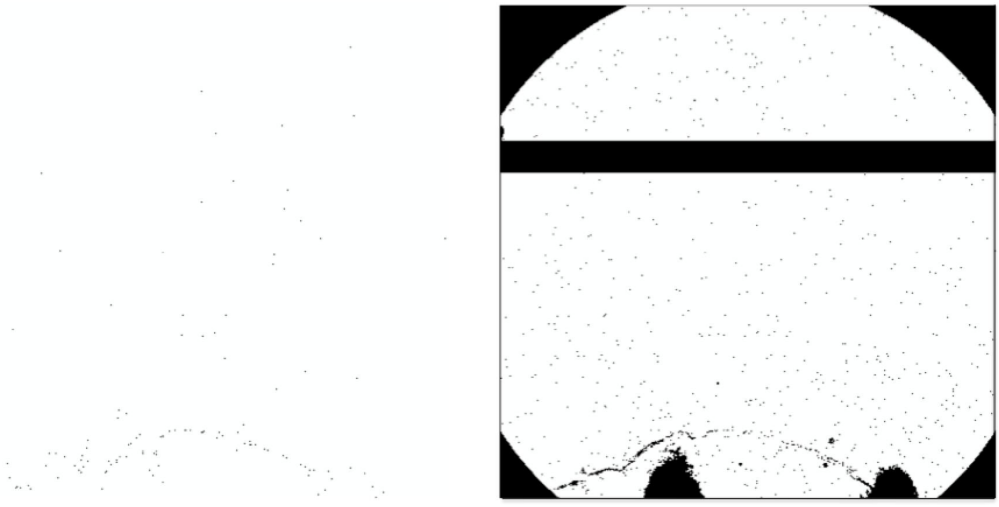

Figure 5: Insensitive pixel mask (<25% of mode) is shown at left. The combination of inaccessible, hot and insensitive pixels is shown at right.

6.4 Linearity

To test linearity a ratio image was computed for a three-fold exposure time increase. Dark frames *with matched exposure times* were subtracted prior to computing the image ratio. The mean signal level in the bright frame was 13,650 e-, so any pixel with less well capacity than this will also be flagged as non-linear. Some pixels with extreme dark current suffer sufficient dynamic range to show up as non-linear notwithstanding the dark frame subtraction. Pixels with more than 15% signal deficit in the brighter image are flagged as non-linear. At 0.27% this is the second largest pixel loss mechanism. Figure 7 shows that these non-linear pixels are randomly located. The threshold is set well below the normally distributed histogram core, which is probably dominated by measurement errors. Repetitive measurements would reduce these statistic errors so that the linearity requirement could be tightened, however at 15% the error in the centroid *motion* estimate due to non-linearity is probably already small.

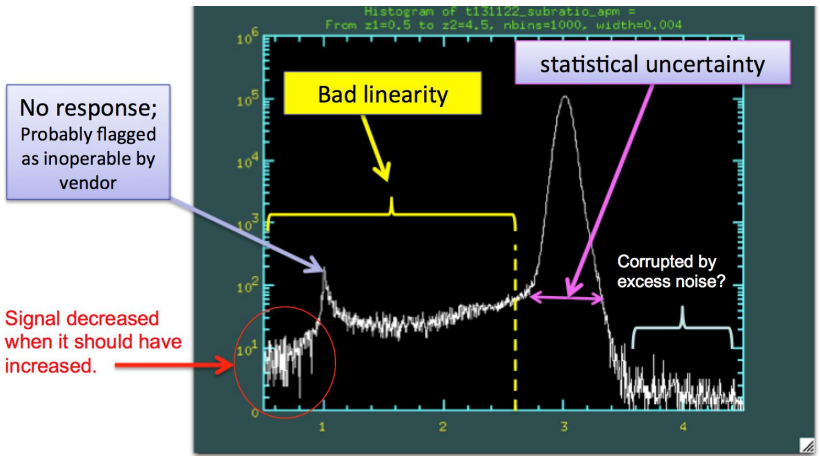

Figure 6: Histogram of ratio of dark-subtracted flats with 3 fold difference in exposure time.

Again, the mapping of the whole detector was time consuming since a region of interest had to be moved across the whole detector to build up the map iteratively. (The data acquisition and analysis was automated of course.)

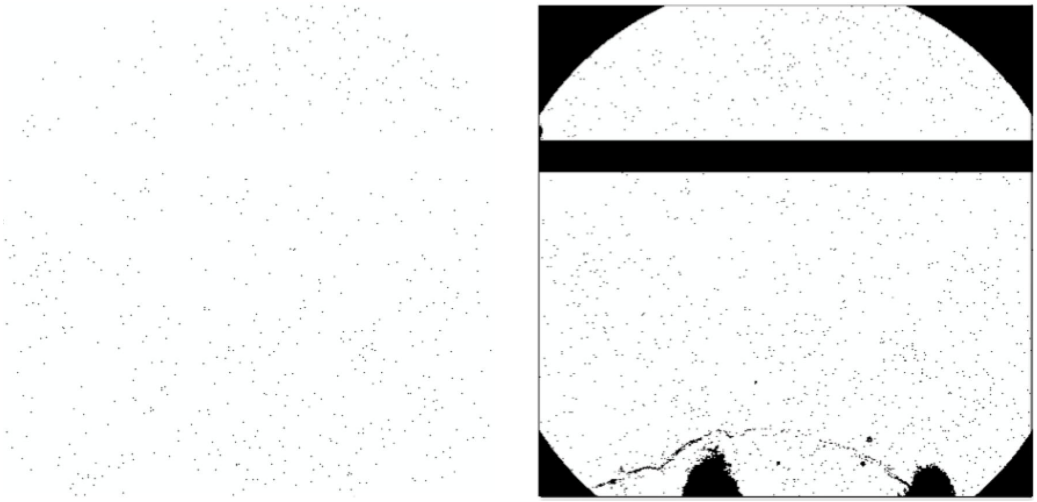

Figure 7: Non-linearity mask.

The non-linearity mask is shown at left (15% signal deficit). The combination of inaccessible, hot and insensitive and non-linear pixels is shown at right.

6.5 Read noise

Noise outliers in three-transistor-per-pixel (3T) CMOS pixel detectors such as the Teledyne H2RG can be caused by high dark current at low frame rates, whereas noise outliers caused by Random Telegraph Signal (RTS) in the pixel

buffer MOSFET can appear at any frame rate and the same RTS noise can be concealed at other frame rates. This leads to a strong frame-rate dependence of the noise outlier map.

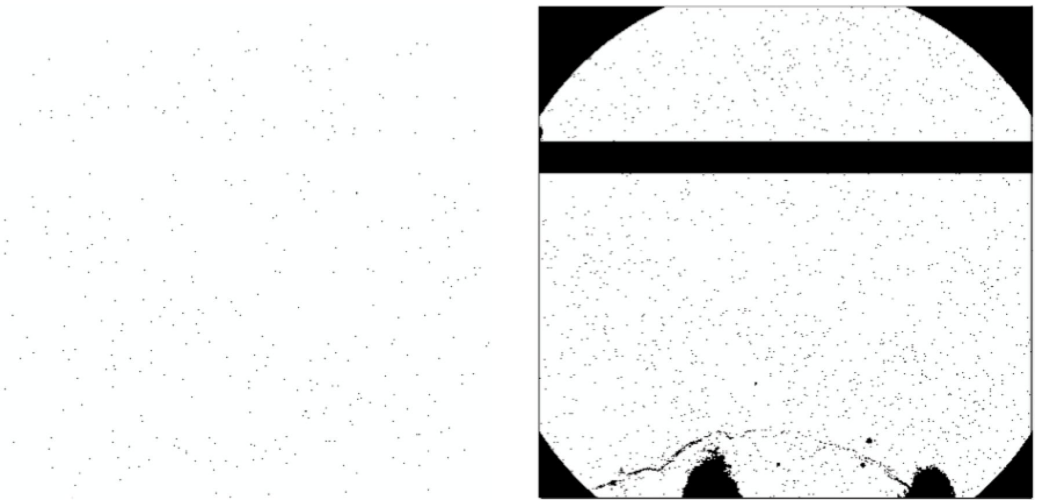

Figure 8: 100 Hz noise outlier map is shown at left. The combination of inaccessible, hot, insensitive pixels, non-linear and noisy pixels is shown at right.

RTS is a bi-stable FET gain, switching between states randomly but with some characteristic frequency. In the TRICK camera, exposures are synthesized by coadding successive non-destructive reads then taking the difference of coadded groups[3]. Pixel resets occur at a much lower cadence. RTS seldom produces errors when its switching frequency lies well below the pixel rate (chances of a transition during an exposure is low). When the RTS switching rate is faster than the frame rate, it is strongly attenuated by sample averaging. Thus the location of high noise pixels on the detector (Figure 9) can change significantly even if the number of noise outliers does not.

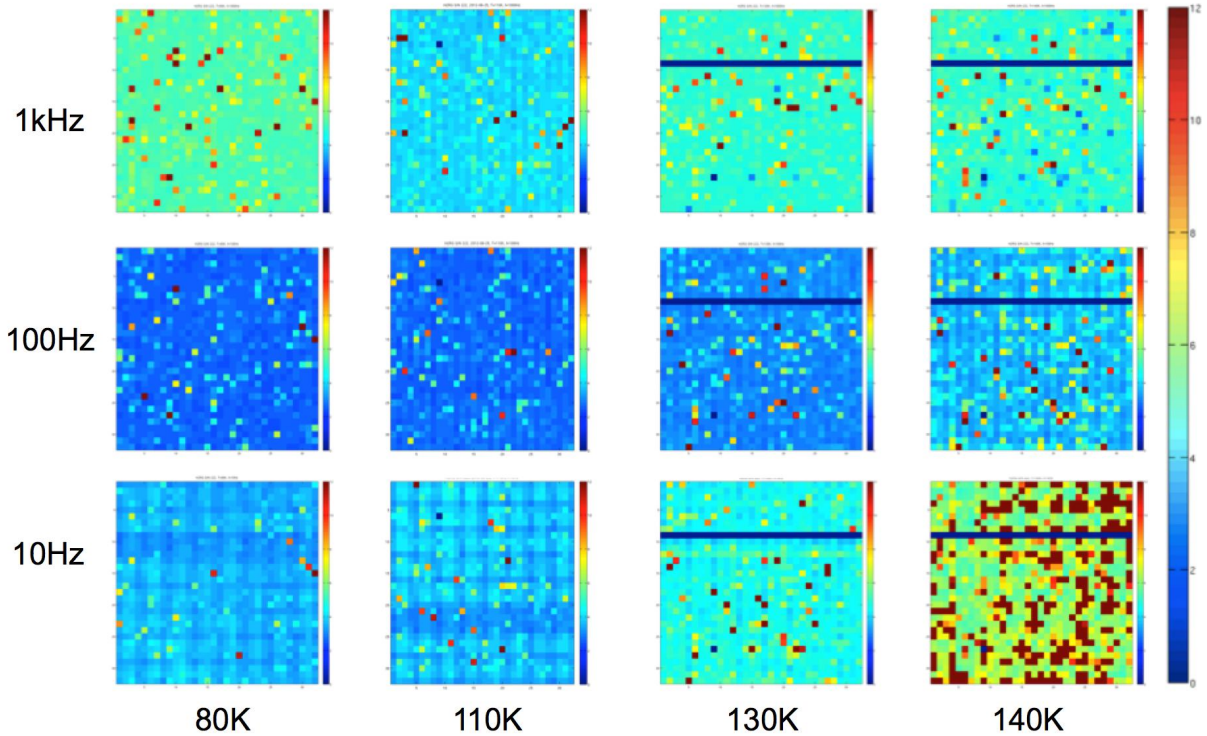

Figure 9: Section of noise map showing how spatial distribution of noise is strongly dependent on both frame rate and temperature. [3]

Whereas the linearity histogram breaks clearly into a normally distributed core (width probably set by sampling statistics) and outliers, the noise histograms shown Figure 10 through Figure 13 are highly skewed and extend on the high side with no clear delineation between core and outliers. The extended high side and steep slope on the low side are typical of noise processes: the choice of bad pixel threshold is thus highly subjective and clearly requires simulation to test the impact on science. While the high-side of the histogram may appear to be similar across the frequency range (Figure 10 to Figure 13), examination of the noise maps (Figure 9) shows that pixels are moving in and out of the tail as frame rate changes.

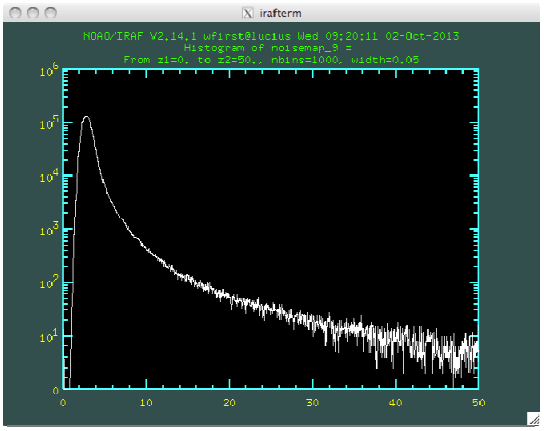

Figure 10: Noise (e-) histogram at 20 Hz frame rate.

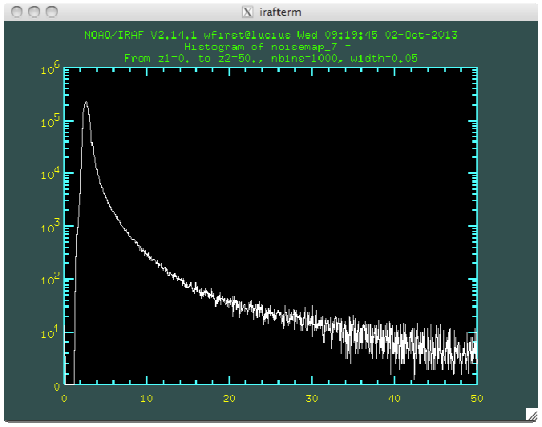

Figure 11: Noise (e-) histogram at 100 Hz frame rate.

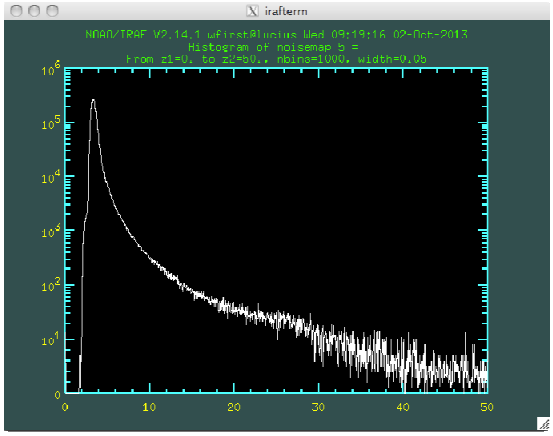

Figure 12: Noise (e-) histogram at 507 Hz frame rate.

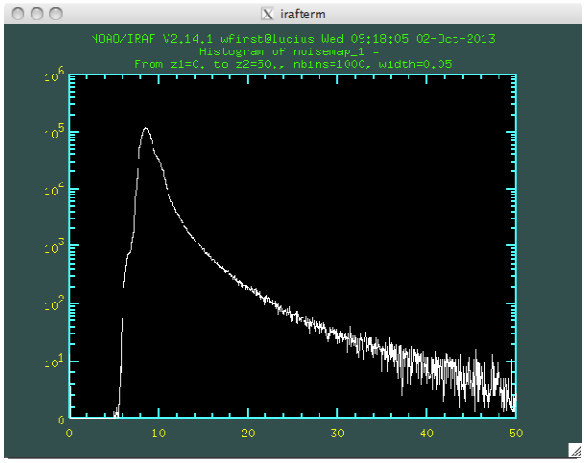

Figure 13: Noise (e-) histogram at 8110 Hz frame rate (CDS = 2 reads per exposure)

6.6 Criteria not used

In this application we have chosen to ignore well capacity since signals are generally faint and pixels with particularly low well capacity will be identified in the linearity test. Inter-pixel crosstalk, image persistence, reciprocity failure, intrapixel response variation, lateral charge diffusion, gain and offset stability may be critical in other applications requiring high precision but are ignored here.

6.7 Results, and how to state them

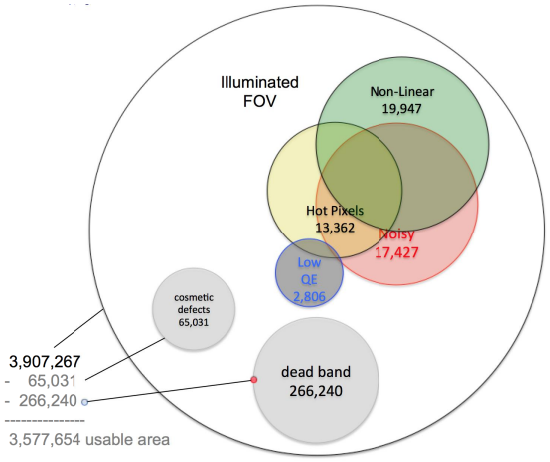

Figure 14: Venn diagram, approximate.

Some pixels will be marked as bad in more than one metric. When looking at histograms to identify the bad pixel threshold, it is helpful to first mask out all sources bad data that are already known, and then to look at the incremental contribution of a particular phenomenon. In this process, the exact classification according to "cause" becomes dependent on the order of the screening process but the cumulative counts are not altered unless the process promotes a different choice of threshold. It is preferable to first examine parameters where the histogram clearly has a defined core and outliers, before progressing to parameters where the distribution is more continuous. It is for this reason that we chose hot pixels, linearity, sensitivity and noise in that order. An approximate Venn diagram such as Figure 14 can be helpful to visualize the classification but the cumulative distribution (Table 3) is less work to produce and less subject to ambiguity.

Table 3: Incremental and cumulative bad pixel counts. Fractions are with respect to usable area of 3,577,654 pixels. Incremental bad pixels are uniquely removed by a given mask, after the previous masks have been applied. Noise is frame rate dependent: this is the100 Hz case.

Cause	Incremental	Incremental %	Cumulative Bad	Cumulative bad %
Hot Pixels	13,362	0.37%	13,362	0.37%
Low QE	2,541	0.07%	15,903	0.44%
Linearity	9,282	0.26%	25,185	0.70%
100 Hz Noise	6,414	0.18%	31,599	0.88%

7. TESTS FOR ACCURACY

7.1 Per-Pixel Photon Transfer Curves

A measurement of variance versus mean signal, often called a Photon Transfer Curve, is a fundamental test to show that a sensor is performing correctly. This test is sensitive to non-linearity, well capacity, gain instability and subtleties such as signal dependent pixel to pixel crosstalk or charge redistribution. It would therefore seem like an excellent candidate for bad pixel mapping.

We constructed per-pixel photon transfer curves by acquiring sample up the ramp data for regions at all locations on the sensor, as follows. Record single-channel, full-frame images in "sample up the ramp" 15 samples per ramp, each producing a 2048×2048 pixel frame. We concatenate these frames into a single file which we call a filmstrip due to the obvious analogy. Repeat to record 500 such filmstrips. From these 15 frame filmstrips, produce 14 CDS images for each: frame1-2, 1-3, 1-4, … 1-15. This produces 500 CDS images at each of 14 different signal levels. Use IRAF's *imcombine* to produce the mean and standard deviation of the film strips which can then be separated into 14 mean frames and 14 standard deviation frames. For each pixel we now have a set of mean/variance points based on a time series of 500 samples, for each of 14 different signal levels. Fit a line to these points in mean/variance space: the slope of mean versus variance is conversion gain (e-/ADU), with one value per pixel location. Arranging this into an image produces a pixel map of system gains. The fitting routine also produces R^2, the "coefficient of determination", a statistical measure, which indicates how well the data fit a straight line.

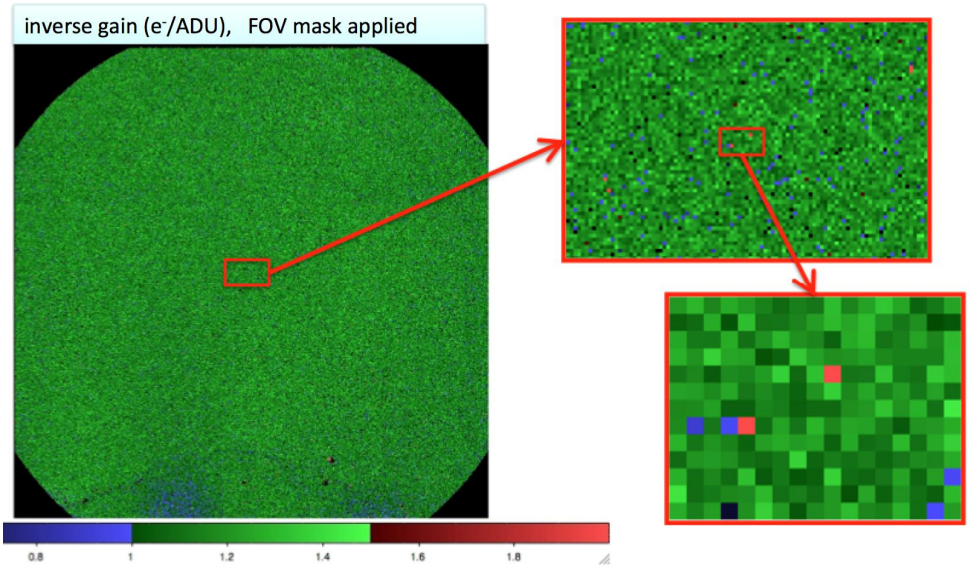

Figure 15: Per-pixel conversion gain (e-/ADU) map.

The conversion gain map shown in Figure 15 contains spatial correlations, which align with low QE areas at the bottom of the frame and are probably real. The histogram (Figure 16) reveals sub-populations with extreme values. This highlights a problem with less direct metrics: it is less clear what constitutes "good" and "bad".

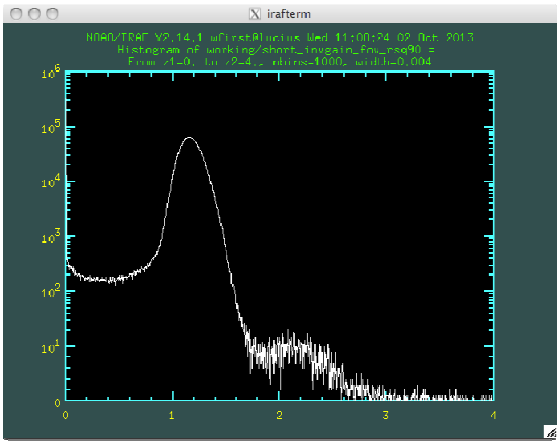

Figure 16: Per-pixel conversion gain (e-/ADU) histogram using all 500 samples

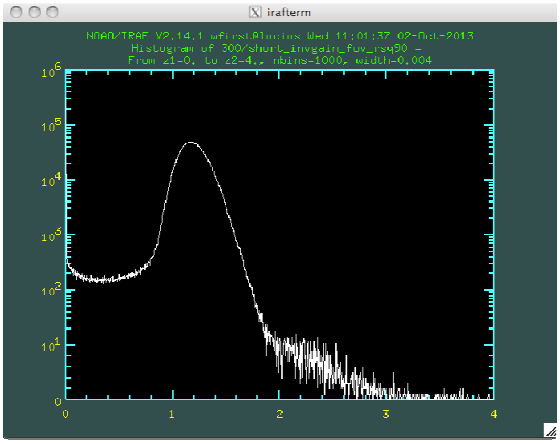

Figure 17: Per-pixel conversion gain (e-/ADU) histogram using 300 samples

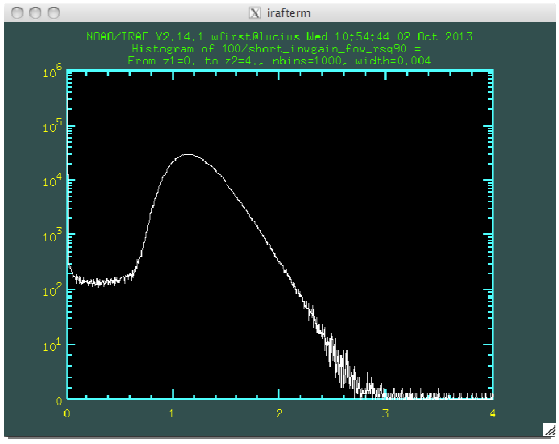

Figure 18: Per-pixel conversion gain (e-/ADU) histogram using 100 samples

7.2 Statistical accuracy

To test the contribution of statistical errors to dispersion in the parameter being measured one can construct histograms using subsets of the data, then plot sigma as function of number samples used. If the width of the histogram is that of the intrinsic distribution in the parameter being measured with little statistical error, then the width will not change (much). On the other hand, if the dominant source of histogram width is statistical error, then sigma will scale as square root of the number of samples.

Figure 16 through Figure 18 show the conversion gain histograms for progressively fewer samples. (This test could have been performed using any of the metrics in this paper.) Figure 19 plots histogram width versus number of samples with the square root scaling for comparison. Clearly the core of the histogram is dominated entirely by statistical error so it is better to use the mean conversion gain rather than the per pixel value.

It remains to be determined whether the conversion gain outliers are merely unusual or malfunctioning. High e-/ADU represents lower than average electronic gain, which is feasible. e-/ADU near zero implies many ADU per electron, but extreme electronic gain is infeasible since the source follower in the H2RG mux can only produce gain less than unity. It is more likely that the variance estimate has been corrupted by some other noise source and that the conversion gain value is erroneous and/or the pixel is malfunctioning in some other way.

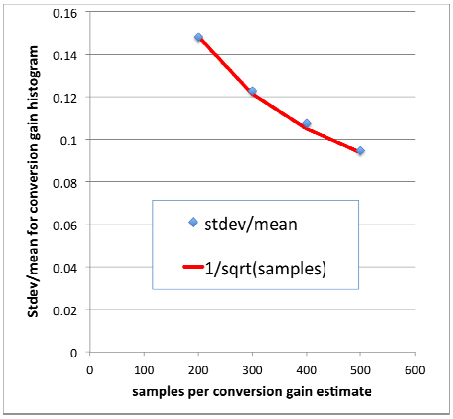

Figure 19: Standard deviation of the conversion gain estimate scales as 1/√samples indicating that the core of the histogram is dominated by stochastic errors and that the underlying dispersion in conversion gain itself is negligible.

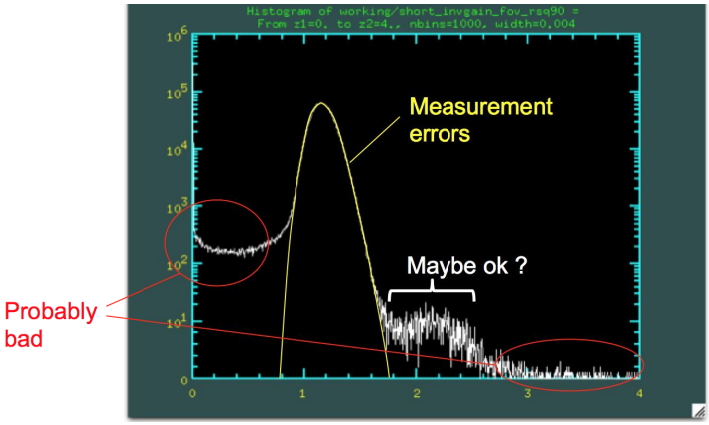

Figure 20: This histogram has been made after screening for bad pixels, but still shows significant number of new outliers. The core of the histogram is dominated by measurement errors that would force bad pixel thresholds to be set at 0.8 e-/ADU and 1.8e-/ADU. High e-/ADU corresponds to low electronic gain, which does not necessarily make a pixel unusable, however low e-/ADU implies implausibly high electronic gain or excess noise implying that the pixel is probably defective.

7.3 Reproducibility

As another test of fidelity, a data set can be split into two halves to test for reproducibility of any given metric. A correlation plot of "second measurement versus the first" will contain one point per pixel. With ~4 million points on the plot this is best viewed as an image with frequency on the "intensity" axis. The use of a log scale for frequency highlights the relatively small number of outliers. Points should lie along the diagonal and be clustered about the mean value of the parameter. Elongation along the diagonal reflects the combined effect of statistical error and dispersion in the underlying parameter, while the dispersion in the orthogonal direction will be only dependent on statistical errors.

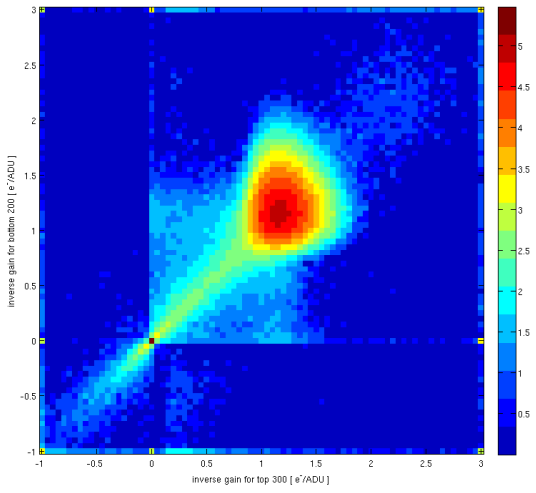

Figure 21: correlation plot for conversion gain (e-/ADU) of all accessible pixels measured using first 300 data points versus that measured using remaining 200 data points, prior to masking out any "bad" pixels. Note the implausible negative and very low values.

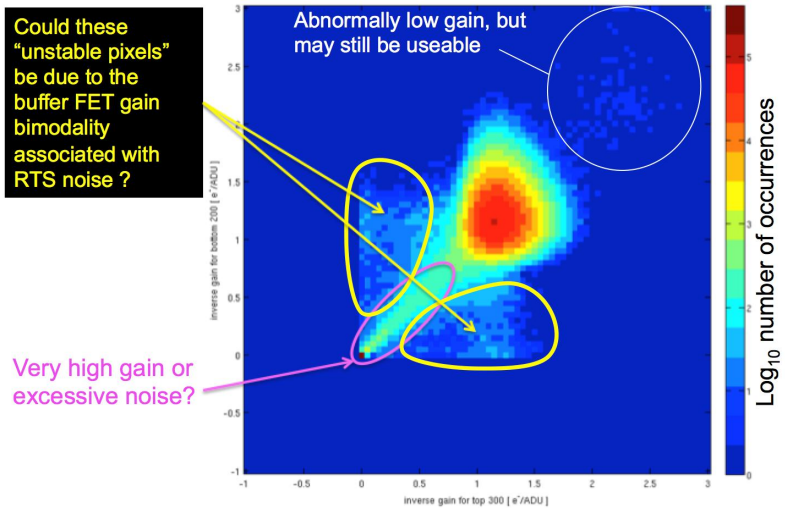

Figure 22: correlation plot for conversion gain (e-/ADU) of all accessible pixels measured using first 300 data points versus that measured using remaining 200 data points. Negative conversion gain values were eliminated by masking out hot, non-linear, low sensitivity and noisy pixels, but a small population with very low conversion gain remains. *This test reveals that some of these pixels are intermittently low!*

Points with negative conversion gain were removed by applying the bad pixel masks above, but a subset of pixels remains whose conversion gain falls well below that expected from statistical errors, in one or both measurements. We are inclined to believe that these are malfunctioning pixels, but have not yet marked them as bad since we have not determined whether this intermittency is a property of these particular pixels or is a low rate failure of the measurement method that could affect any pixel.

> We apply the equivalent of the Hippocratic Oath to bad pixel mapping:
>
> *Don't mark a pixel as bad, unless it is consistently and unambiguously bad.*

8. DISCUSSION

Clearly bad pixel mapping is complicated, and requires great care to achieve adequate precision. The assessment of the effects of pixel designation as "bad" is a subset of the larger problem of understanding how detector performance impacts science. While the modeling of effects of detector parameters expressed as averages might be straightforward in principle, taking into account histogram shapes and alternative strategies for dealing with outliers requires a Monte Carlo simulation or complicated integrals.

In the absence of comprehensive modeling, arbitrary choices for bad pixel thresholds must be made. Even when somewhat arbitrary, the existence of a "reference case" allows some appreciation of the problem and establishes a vehicle for the detector engineers to communicate with the calibration team and users. At the same time it is important that all parties understand that other choices can be made.

Customers are usually better off performing their own tests based on application specific criteria than to impose their test criteria on the vendor who may or may not be equipped to perform the more detailed testing and analysis required. The vendor's testing should be adequate to prescreen and possibly to grade candidate devices. Changes in vendor test protocols should be avoided where these threaten the vendor's ability to determine whether their process-control is delivering consistent results.

The risk that devices will not satisfy science requirements is best mitigated by more detailed testing and analysis, *performed by the instrument team.*

For large/critical procurements, testing of early devices and prompt acceptance testing during production can mitigate significant risk. Regrettably this modeling is often only done late in the project by the calibration team, when it in fact it is needed to set specifications (including bad pixel criteria).

Image sensors for scientific imaging are complex and not fully understood at the limits of precision. For critical procurements such as space missions attempting high precision measurements, end to end testing will be indispensible to validate the models, needed to set up the acceptance test protocols: sample devices must be illuminated with represented scenes, operated as planned for the application. Planned calibration must be performed and data analyzed in representative ways to test whether the models are both correct *and complete.* Bad pixel definitions will be a natural by product.

REFERENCES

[1] Wizinowich, P. L., Smith, R. M., Biasi, R., Cetre, S., Dekany, R. G., Femenia-Castella, B., Fucik, J. R., Hale, D., Neyman, C. R., Pescoller, D., Ragland, S., Stomski Jr., P. J., Andrighettoni, M., Bartos, R., Bui, K., Cooper, A., Cromer, J. L., Hess, M. J., James, E., Lyke, J. E., Rodriguez, H., Stalcup Jr., T. E., "A near-infrared tip-tilt sensor for the Keck I laser guide star adaptive optics system," Proc. SPIE 9148 (2014).

[2] Clanton, C., Beichman, C., Vasisht, G., Smith, R., Gaudi, B. S., "Precision near infrared photometry for exoplanet transit observations: I - ensemble spot photometry for an all-sky survey," PASP, Vol. 124, No. 917, 700-713 (July 2012).

[3] Smith, R. M., Hale, D., "Read noise for a 2.5μm cutoff Teledyne H2RG at 1-1000Hz frame rates," Proc. SPIE 8453 (2012). http://authors.library.caltech.edu/37253/1/84530Y.pdf