Tina Memo No. 2014-005
Internal Report

Negative Calibration for Citizen Science Crater Data.

Paul D. Tar, N.A. Thacker.

Last updated
24 / 10 / 2014

ISBE, Medical School,
University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT, UK.

False Negative Calibration for Citizen Science Crater Data

Paul D. Tar, N.A. Thacker

1 Moon Zoo and Expert False Negatives

This document completes a series describing the filtering of Moon Zoo data, starting with Tina Memo 2014-001 - 'Coalescence and Refinement of Moon Zoo Crater Annotations', then Tina Memo 2014-002 - 'Quantification of False Positives within Moon Zoo Crater Annotations'. This document also extends section 9.3 of the thesis 'Quantitative Planetary Image Analysis via Machine Learning'. It is assumed that readers are familiar with the content of these previous reports.

The Moon Zoo filtering pipeline stages described in previous reports had the character of reducing the number of craters in raw crater data. The clustering and refinement stages (2014-001) reduced multiple mark-ups into individual craters, and the linear modelling (2014-002) reduced counts further by removing the effects of false positive contamination. If the original raw data contained every single real crater, plus contamination, then by the end of the previous step the task of filtering Moon Zoo data would be complete. However, not all craters are identified in Moon Zoo datasets which leads to underestimated crater counts in Size Frequency Distributions (SFD). Missing craters are partially caused by inattention of users and partially caused by difficulties in spotting craters which may be heavily eroded etc. Similar problems can exist in expert counts also. This document provides two potential methods of correcting such underestimated counts:

1. How to correct small discrepancies in "expert" counts where at least two independent counts of the same region are available, thereby providing a reasonable approximation to act as ground-truth.
2. How to correct large discrepancies in citizen science crater counts using "expert" ground-truth regions for calibration.

This document investigates these possibility using a subset of Moon Zoo data around the Apollo 17 landing site. The "expert" ground-truth is provided by undergraduate students who marked-up sample regions twice. Once processed, this expert ground-truth is used to attempt correction of Moon Zoo counts for the same regions.

2 Correcting "Expert" Counts for use as Ground Truth

As no absolute ground-truth for crater counting exists, this document defines a ground truth for illustrative purposes based upon undergraduate Earth Science students and their repeatability. However, the processes described below could equally be applied using more experienced crater counters. The ground-truth is estimated as follows:

- Multiple calibration regions of the Apollo 17 site are selected;
- Craters in each region are diligently annotated **twice**;
- The differences between the two annotation attempts are modelled in terms of Binomial statistics;
- Counts of craters are computed taking into account the estimated Binomial efficiencies, thereby correcting for craters likely have been missed;

The method assumes that the only variability in expert counts is caused by missing craters. If expertly identified craters are likely to contain significant quantities of false positive contamination then this method would be invalid.

2.1 Calibration Regions

Moon Zoo data covering LROC NAC images M104311715LE and M104311715RE were selected for testing. Within this data the following regions were chosen for calibration. These regions spanned the full width of the source images in 400 pixel high strips:

1. Image M104311715LE, pixel rows 400 to 800
2. Image M104311715LE, pixel rows 6,000 to 6,400

3. Image M104311715LE, pixel rows 15,200 to 15,600
4. Image M104311715LE, pixel rows 32,000 to 32,400
5. Image M104311715LE, pixel rows 49,500 to 49,900
6. Image M104311715RE, pixel rows 400 to 800
7. Image M104311715RE, pixel rows 6,000 to 6,400
8. Image M104311715RE, pixel rows 15,200 to 15,600

Undergraduates[1] of the School of Earth, Atmospheric and Environmental Sciences, University of Manchester, marked-up each image twice, annotating all craters down to 10 pixels.

2.2 Estimating Efficiencies and Ground Truth using Binomial Assumption

After two attempts at annotating craters in the calibration regions there were a number of craters that were identified twice, (i.e. on both attempts), once (i.e. missed on either the first or second attempt) and some which were never identified (i.e. missed during both attempts). These frequencies will be referred to as F_2, F_1 and F_0 respectively, where the subscript indicates the number of times a crater was identified. Assuming that there is some fixed efficiency (probability), P, of annotation and there are N true number of craters, the number of craters falling into each category should be:

$$F_2 = NP^2 \tag{1}$$

$$F_1 = NP(1-P) + NP(1-P) \tag{2}$$

$$F_0 = N(1-P)^2 \tag{3}$$

The errors on any of these counts can be modelled using:

$$\sigma_F^2 = N\left(\frac{F}{N} - \frac{F^2}{N^2}\right) = F - \frac{F^2}{N} \tag{4}$$

Which assumes Binomial variances, as N is a fixed quantity.

F_1 and F_2 are both known quantities from the undergraduate annotations. The superset containing all annotations from both mark-up attempts contains $F_1 + F_2$ craters. But the correct number of craters should include F_0 also: $N = F_0 + F_1 + F_2$. If these frequencies are modelled as bins in a Binomial distribution, the known frequencies can be used to estimate annotation efficiency, P, and also the "true" number of craters, N.

The success probability (counting efficiency) can be given in terms of the observed frequencies:

$$P = \frac{2F_2}{F_1 + 2F_2} \tag{5}$$

With errors given by error propagation:

$$\sigma_P^2 = \left(\frac{dP}{dF_1}\right)^2 \sigma_{F_1}^2 + \left(\frac{dP}{dF_2}\right)^2 \sigma_{F_2}^2 \tag{6}$$

$$\sigma_P^2 = \frac{P^4}{4F_2^2}\sigma_{F_1}^2 + \frac{P^4 F_1^2}{4F_2^4}\sigma_{F_2}^2 \tag{7}$$

The ground-truth counts, N, corrected for missing craters, F_0 can also be given:

$$N = \frac{(F_1 + 2F_2)^2}{F_2} \tag{8}$$

[1]Sean Corrigan, Alex Griffiths, Tim Gregory, Hazel Blake, Dayl Martin, Maggie Sliz, Joe Scaife and Pavel Kamenov

With errors given by:

$$\sigma_N^2 = \left(\frac{dN}{dF_1}\right)^2 \sigma_{F_1}^2 + \left(\frac{dN}{dF_2}\right)^2 \sigma_{F_2}^2 \tag{9}$$

$$\sigma_N^2 = \left(\frac{F_1}{4F_2} + 1\right)^2 \sigma_{F_1}^2 + \left(-\frac{F_1^2}{4F_2^2}\right)^2 \sigma_{F_2}^2 \tag{10}$$

Efficiencies and ground-truth crater counts were estimated from the undergraduate mark-up on a regional and crater size basis.

2.3 Results

Figure 1 shows the difference between using the total number of identified craters ($F_1 + F_2$) verses the estimated total number after correction for Binomial efficiency losses. The Undergrad Superset blue bars show $F_1 + F_2$ for different regions and crater sizes. The Undergrad Corrected red bars show the estimated ground-truth, N, which includes the uncounted F_0 contributions. As can be seen, the corrected counts are only ever a small percentage larger than the simpler superset counts.

Figure 2 shows the estimated identification success efficiencies for each region. This plot shows that, on average, individual attempts to count craters identified 78% of the total number of craters present. After two attempts to count all craters, on average, 95% of craters were identified. The efficiencies across the different regions are all roughly equivalent, being within 2 to 3 standard deviations away from one another.

3 Moon Zoo False Negative Calibration

The above Binomial model for correcting expert counts is inapplicable in Moon Zoo data because here the number of attempted counts is not a fixed quantity. Some regions have been seen by more users than others and there are no guarantees that each user has attempted to annotate all craters. Because of this it is necessary to apply a more empirical approach. Corrected Moon Zoo counts are estimated as follows:

- Expert ground-truth is compared to Moon Zoo crater counts within calibration regions;
- Corrective scaling factors are computed: one overall scaling factor and crater size specific factors;
- The factors are applied back to the calibration regions to double-check their effectiveness;
- If effective, the same factors are used to correct counts in similar independent regions.

3.1 Estimating Scaling Factors and their Stability

A scaling factor, s, can be applied multiplicatively to a Moon Zoo crater count to approximately correct for missing data:

$$s = \frac{N}{m_0} \tag{11}$$

where within a calibration region N is the ground truth frequency within an SFD bin, as determined ealier; and m_0 is the predicted frequency of true positives from filtered Moon Zoo data, taken from 2014-002. A corrected count for future data can then be given by:

$$c = ms \tag{12}$$

where m is a false positive corrected count from a different area. The error on this new count, σ_c^2, is given by:

$$\sigma_c^2 = s^2\sigma_m^2 + m^2\sigma_s^2 \tag{13}$$

where σ_m^2 is the variance on m, as predicted by 2014-002; and σ_s^2 is the variance on the scaling factor, given by:

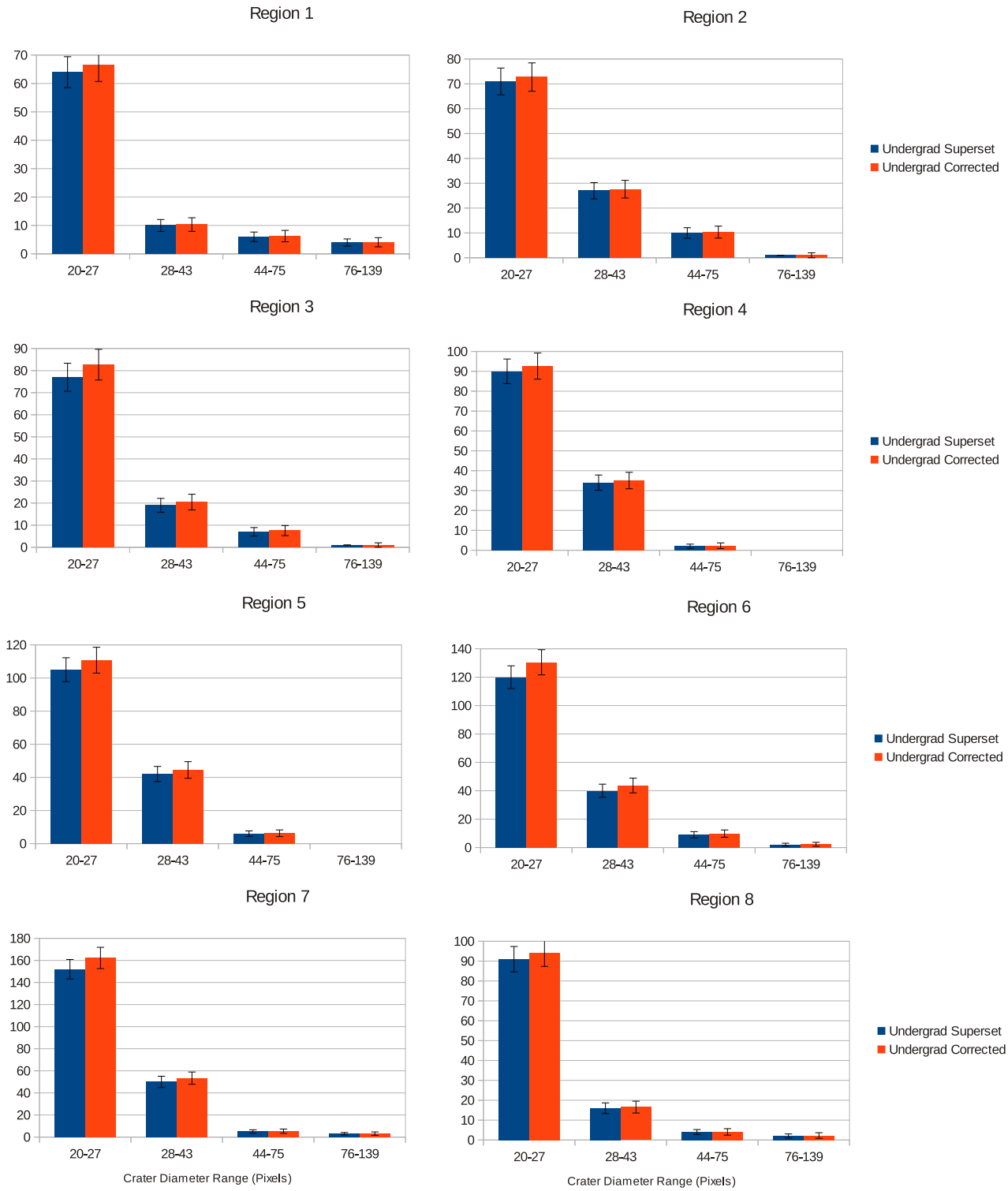

Figure 1: The 8 plots correspond to the 8 Moon Zoo calibration regions. Blue bars indicate the number of craters identified by undergraduates after two annotation attempts. The red bars indicate the estimated "true" number of craters when corrections are made for Binomial efficiency losses.

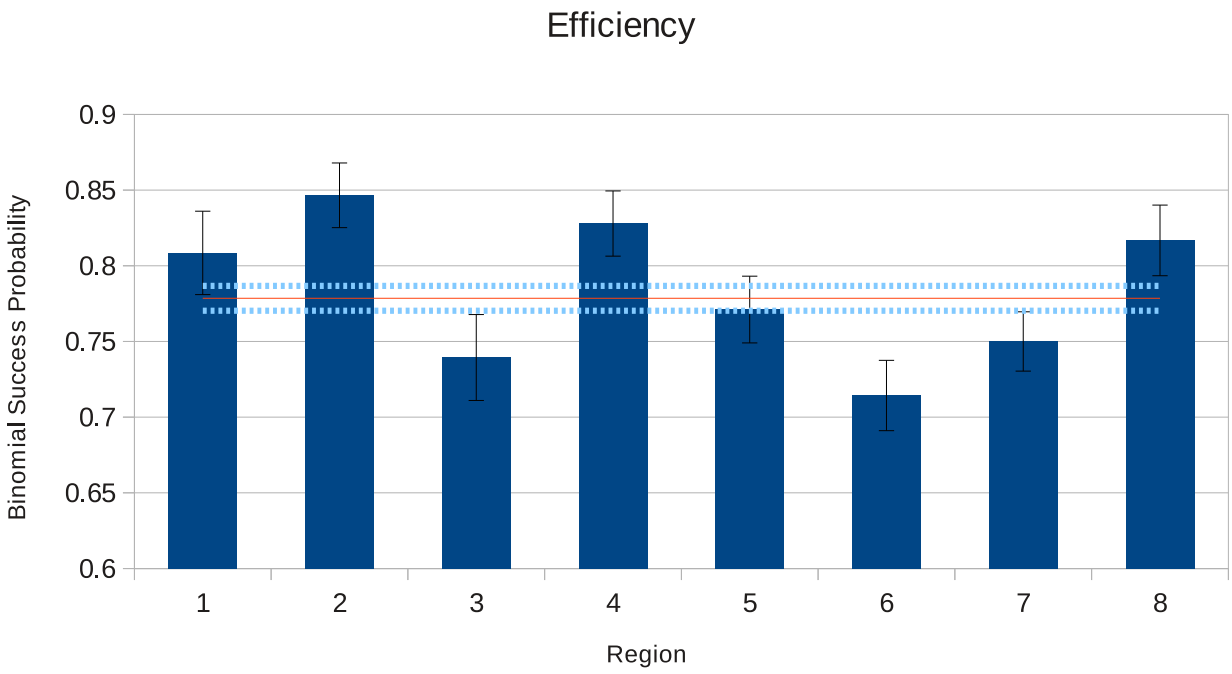

Figure 2: The 8 bars correspond to the estimated identification efficiencies for the 8 calibration regions. The red line shows the estimated overall efficiency, taking into account all regions jointly.

$$\sigma_s^2 = \frac{1}{m_0^2}\sigma_N^2 + \frac{u^2}{m_0^4}\sigma_{m_0}^2 \tag{14}$$

where σ_N^2 is the estimated error on the ground-truth, given earlier.

Scaling factors were computed for Moon Zoo data jointly across all selected calibration regions. One overall scaling factor and size-specific scaling factors were both tested. As a best-case scenario, these factors were applied back to the calibration regions themselves, as they would be the most representative.

3.2 Results

Figure 3 shows the effects of applying one overall scaling factor to all counts in all regions verses the effects of applying size specific scaling factors to different crater size bands. Figure 4 show the size of these scaling factors.

If calibration was successful, in Figure 3 the blue ground truth bars should be statistically equivelent to the calibrated Moon Zoo bars. A chi-square per degree of freedom reveals this is not the case:

- $\chi_D^2 = 23.9$ between uncalibrated and ground truth counts;
- $\chi_D^2 = 6.2$ between overall calibrated and ground truth;
- $\chi_D^2 = 2.9$ between size-specific calibrated and ground truth.

4 Discussion

Binomially corrected ground truth from undergraduates and their estimated efficiencies show a roughly consistent mark-up rate across the different regions and different undergraduates.

Moon Zoo results using a single scaling factor show better agreement between Moon Zoo and expert SFDs for small crater sizes. However, the overall fixed scaling is inappropriate for some of the larger craters. Plus, the method fails in region 7, where the smallest size bin remains underestimated by a significant amount.

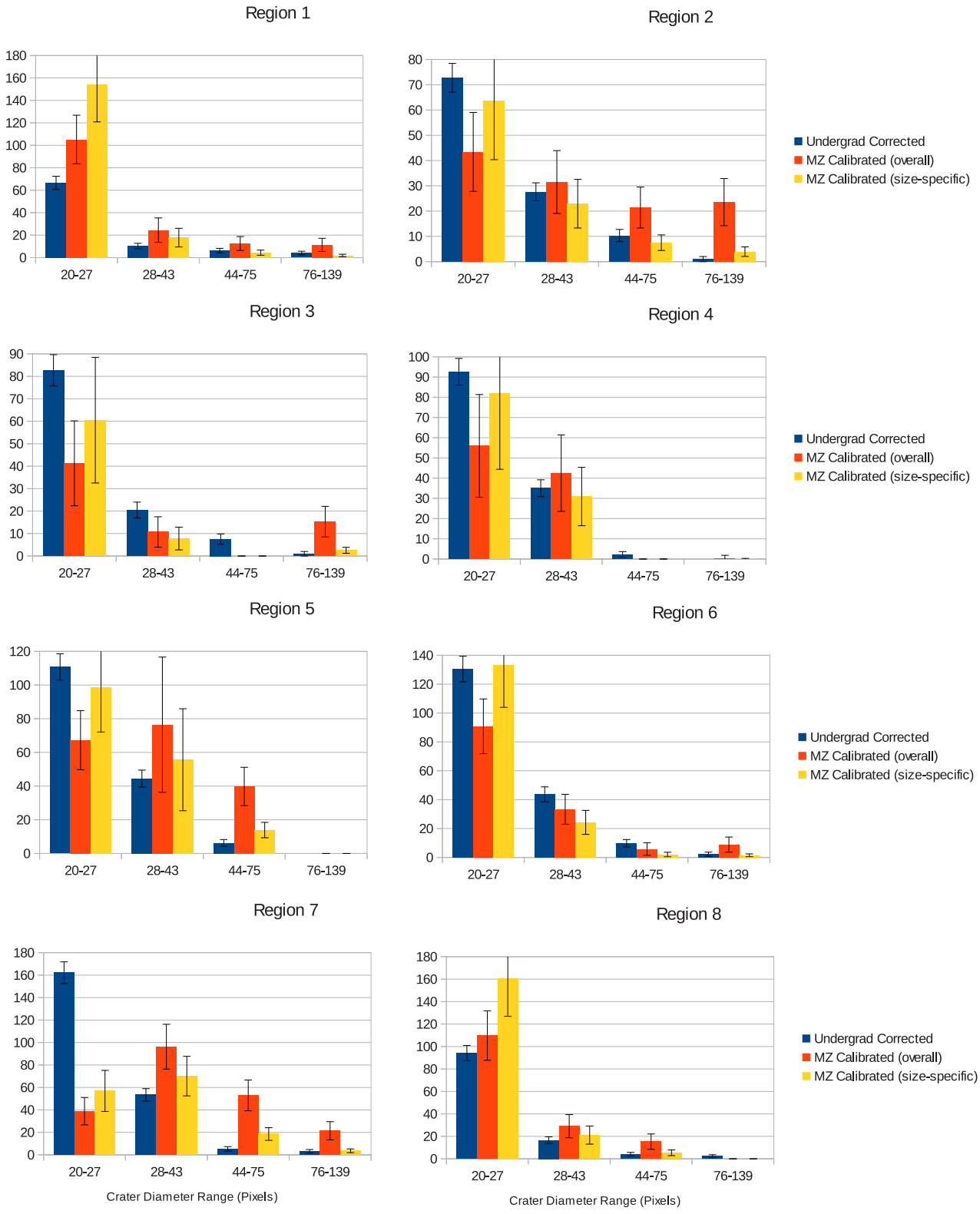

Figure 3: Blue bars represent ground truth counts which the calibrated counts attempt to replicate. Red bars represent calibrated Moon Zoo counts which have been scaled using a single overall scaling factor. Yello bars represent calibrated counts using size-specific scaling factors.

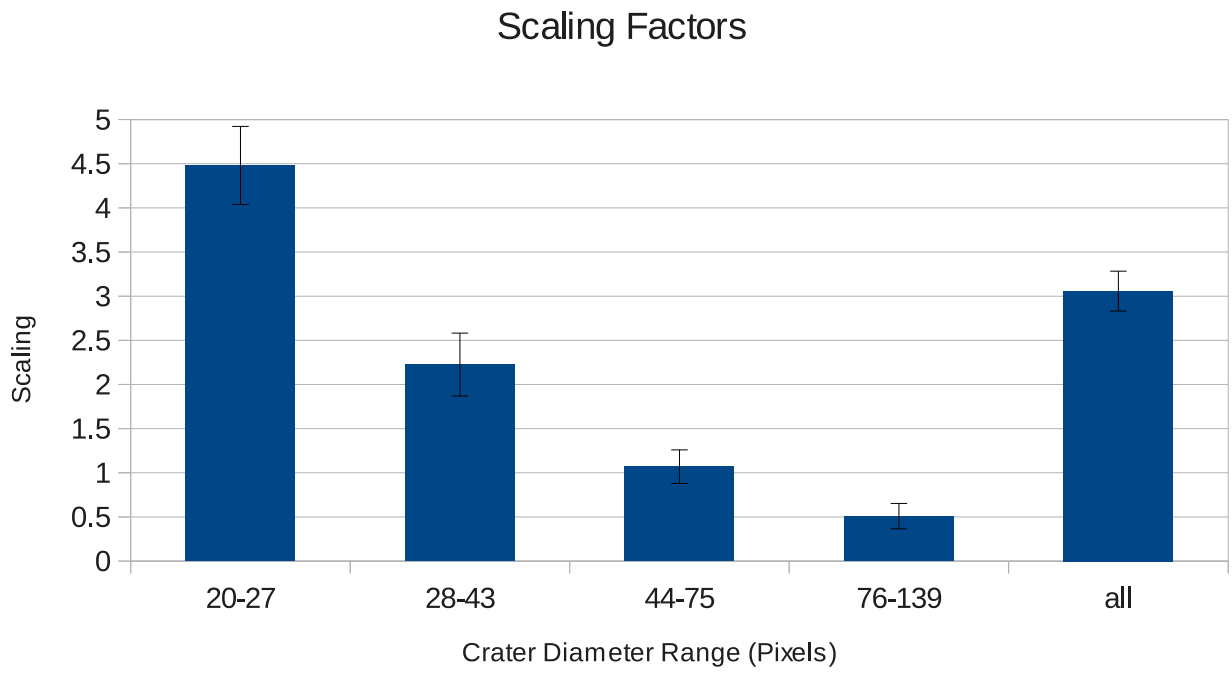

Figure 4: The scaling factors computed per size range and overall scaling.

Individual scaling factors proved more effective, where highest agreement between Moon Zoo and expert SFDs was achieved. Only one bin failed to meet its ground truth count (again, regions 7's smallest bin), being several standard deviations away from agreement.

In both cases, the calibrated counts have large relative errors. Proportionally, the uncertainty on Moon Zoo SFD bins can be as large as 50%. This is primarily due to the small sample sizes available and could be improved by gathering more data. At these levels of error, making quantitative use of results would be difficult. Plus, the inability to correct region 7's small size bin suggests that region specific scalings would also be required.

4.1 Summary

This document completes the process of filtering raw Moon Zoo crater data into Size Frequency Distributions. The false negative calibration stage described provides a framework for making future corrections, but at present is limited by the sample made data available.

To improve SFDs to usable levels of certainty, either more data must be gathered to reduce the total number of false negative results, or a greater number of calibration regions might be used to determine more appropriate local scalings. A combination of both methods will likely be required.

A limiting factor in the use of Moon Zoo data will be sample size. Any calibration method must make use of sufficient numbers of craters to ensure usable errors in SFDs.