Tina Memo No. 2016-010
Internal Report

Automated Feature Quantification for Planetary Surfaces verses Human Subjectivity.

P.D.Tar and N.A. Thacker.

Last updated
4 / 05 / 2016

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Automated Feature Quantification for Planetary Surfaces verses Human Subjectivity

1 Introduction

Millions of high resolution images are available covering the Moon, Mars, Mercury and asteroids Ceres and Vesta, thanks to dedicated orbital missions and landers, e.g. [17]. In addition, many thousands of images from fly-bys of all other major bodies in the solar system are available. Despite the scale of the datasets, the analysis of planetary images is still a largely manual process involving a combination of experts and citizen scientists [5][9]. Attempts have been made to automate this process to both increase analytical capacity and improve objectivity, but none have achieved any wide acceptance or standards. This report explores the possibility of building upon Linear Poisson Models (LPM) to construct an automated system for the identification and counting of features within planetary images which meet scientific criteria for quantitative use. A detailed overview of LPMs and planetary image analysis can be found in [14]. This follow-up report details the current feasibility of creating an automated analysis tool which can learn to recognise a wide range of features, including craters and dunes. The report is organised into a series of questions regarding the properties and potential uses for such a system. This is followed by empirical evidence of subjectivity as a barrier to the creation and acceptance of any supervised learning system.

2 Defining the problem

What does this report mean by an automated feature counting system for planetary images?

Primarily, this document considers extended features in image data, i.e. patterns of pixels, rather than spot pixel values. Various mixture modelling and machine learning approaches have been applied to image data at a per-pixel level, where pixel intensities in different channels map onto some meaningful physical interpretations. For example, white matter verses grey matter in MRI medical scans show as two populations of pixel values. Planetary image data, however, is more complex. Features, such as impact craters and dunes, are extended over many pixels. Lines, ridges, textures and patterns of light and dark combine to give the appearance of features indicative of different geological processes which [12], unlike MRI data, cannot be mapped to simple populations of pixel intensities. Elevation maps are also available, providing 3D data, but most high resolution data is in the form of 2D photographic images. This report limits itself to considering 2D planetary surface images in the visible or near-visible light spectrum.

An automated feature counting system would ideally be a trainable system which can learn the appearance of features of interest, rather than being fixed in purpose for specific patterns. Even within a specific class of feature there is variability, such as craters with different morphologies and on different terrains with different surface properties. In these cases it would also be useful to have the capabilities to learn the valid forms of variation a feature can exhibit. This report therefore considers a useful system to be one which can be trained, via example, to identify and quantify arbitrary features. It is acknowledged that some features can be more clearly defined than others.

Once trained, a system should be capable of searching for those features in future images to produce counts of those features, e.g. for the creation of crater Size-Frequency Distributions. These counts should be considered as scientific measurements and therefore be subject to a level of quantitative rigor that can ensure the validity of any scientific conclusions drawn from them.

What is wanted from an ideal automated feature counting system for quantitative science?

If quantitative science is an objective, then a proposed system would ideally have the following properties:

- be objective
- be consistent
- be unbiased

- be statistically efficient
- have known error characteristics
- above all, be more reliable than humans

Objective: Human subjectivity is a well-known hindrance with regards to making observations. No reasonable person with common sense would recommend estimating length measurements by eye when an appropriately sized and well calibrated ruler was available. Nor would one expect temperature measurements of the oceans to be gathered by placing a hand into the water; a thermometer is clearly preferable. Requesting that an automated feature quantification system be objective is a prerequisite for science.

Consistent: If there is an objective truth, e.g. a 'correct' count of impact craters within a defined region of the Moon, contradictory conclusions regarding this count should not be possible if measurements are made appropriately. For instance, conclusions should not change if images are rotated or mirrored, or if illuminated from left or right. A consistent tool for science should not allow for multiple conflicting measurements. It is acknowledged that unfavorable viewing conditions can be prohibitive, e.g. in the case of total occlusion of features or saturation of pixel values due to illumination. This report considers such issues as missing data, and therefore would not expect useful consistent measurements to be achieved in such cases. But, if data is present, it should not be contradictorily interpreted.

Unbiased: Ideally, an estimated feature count should on average be correct. A systematic under or over estimation, especially one which varies as a function of the data, complicates the interpretation of the measurements. A known bias my be acceptable if it is easily corrected for.

Statistically efficient: The statistical spread of repeated measurements should be as small as possible, or at least sufficiently small to permit changes in measurements to be identified at the level of the effects likely to be seen. Efficiency entails making best use of the data available, e.g. matching data distribution assumptions to the actual behaviour of data. Generally speaking, the less well approximated a data distribution is, the poorer the results will be. Further, the more assumptions which are violated by the data, the poorer the results will be.

Known error characteristics: Systematic and statistical sources of uncertainty (see above two points) need to be understood for scientific conclusions to be reached with known levels of confidence. The distribution of errors must be known in order to place confidence intervals (error bars) on plots, or to reject hypotheses at specific p values.

Basically, an automated feature quantification system should aim to replace the human visual system and replace it with a objective, consistent 'thermometer' for features, with well understood error characteristics.

Can Linear Poisson Models meet these requirements?

Regarding objectivity, LPMs can objectively apply a given definition to new datasets, using its goodness-of-fit criteron (a χ^2 per degree of freedom) to check if the new data conforms sufficiently well to the trained model [11]. This definition, however, must be provided from somewhere which may introduce subjectivity through choice of training examples.

LPMs have been shown to be consistent. The quantity of false positive crater detections in MoonZoo data was successfully estimated (based upon bootstrap resampling from 'defined' ground truth) using different quantities and ratios of training to testing data. During this work, different templates and template matching methods were also used, all producing estimated levels of false positive contamination which were equivalent within predicted errors. The full report can be found here [15]. Additionally, ADC measurements were consistently made over a range of possible histogram binnings, providing equivalent estimates of changes in tumor growth within errors. This report can be found here [16].

LPMs may produce biases away from 'true' values if a training dataset is skewed towards particular atypical examples of desired features. However, repeated training and testing in Monte Carlo shows no evidence of any net biases in LPM applications. But, if trained once then applied multiple times, there is a known systematic bias towards the specific instances of noise found within the training set. But, this noise in the modelling processes is addressed through a systematic error term in the predictive error theory.

LPMs are statistically more efficient than some other methods, applied to some datasets. Evidence of this can be seen in ADC change measurements, where changes could be seen with statistical significance in individual tumors, whereas other methods require cohorts of tumors to achieve the same levels of accuracy [16].

LPMs have known error characteristics, including statistical, systematic and goodness-of-fit estimates. However,

this relies upon input data having independent Poisson behaviour, with classes of feature combining linearly. Violations of these properties can lead to unknown error behaviour.

There has been little opportunity to directly pitch LPMs against humans. Additionally, it is difficult to define 'better than' if humans are the ultimate source of training examples. LPMs will behave deterministically on identical data, whereas humans will not. The consistency with which LPMs can be applied should be better than humans.

What use would such a planetary feature counter be put to? What about counting craters?

Planetary images contain a great many features, from dunes, river channels and ice deposits, to intricate patterns formed by sublimating CO2. Each feature has its own set of properties, including how it was formed, what information it maintains regarding past and present processes etc. The orientation of dunes can reveal wind direction, for example. To explore the challenges of creating an automated system the topic of crater counting will be used as a case study.

Impact craters are found on almost all solar system bodies, especially those with tenuous or no atmospheres where low levels of erosion allow craters to persist for billions of years. Planetary scientists would like to use information hidden within the distribution of these craters to infer the absolute or relative ages of different surfaces. To achieve this, the Size-Frequency Distribution (SFD) of craters are analysed [8]. The most basic SFDs plot a histogram of crater sizes, with diameters increasing along the x-axis, and the number of craters falling into respective size bands plotted on the y-axis. The convention adopted in the crater counting literature, however, is more complex. It has become tradition to use reverse-cumulative plots of relative counts, normalised to some unit of surface area and presented on a log-log scale. The error bars on these plots assume crater counts follow Poisson statistics, i.e. the variance on a count is approximately equal to the count itself: $\sigma_n^2 =< n >\approx n$

The shape of SFDs are a function of many factors, including:

- the historic and current population of impactors (e.g. asteroids and comets) ;
- the length of time the surface has been exposed to the impactors;
- smaller secondary cratering, caused by ejecta from larger primary impacts;
- erosion of craters due to weathering, infilling, micrometeorites etc.;
- and large-scale resurfacing events, such as lava flows;

For homogeneous airless surfaces, where the dominant factors affecting SFDs are just the impactor population and age of surface, empirical Production Functions (PF) have been estimated [4]. These PFs are usually approximated by straight lines or polynomials. The goal of a PF is to give a low parameter description of the relative frequency of different sized craters expected to be observed in different regions. These functions have been estimated for many bodies, including the Moon and Mars. Chronology Functions (CF) [6] also exist which map between SFDs and absolute surface ages. These have been calibrated using radiometric dating of material returned from the Moon and have been extrapolated to other bodies. Besides absolute chronology, the relative order of geological events may be inferred. Assuming younger surfaces are less heavily cratered than older ones, adjacent surfaces can be ranked [7]. This is appropriate for lava flows for example.

Given the importance of crater statistics, an automated feature counting system must be capable of producing outputs appropriate for the construction of SFDs, the fitting of functions to SFDs (i.e. error distributions must be known), and for making statistically valid comparisons between SFDs.

What are the key limitations of current crater counting methodology?

Subjectivity: A crater, being a roughly circular depression, can be confused with any other feature which appears as a roughly circular depression. Highly eroded shallow craters in particular can be difficult to tell apart from unrelated undulations in a terrain. Shadows and poorly illuminated regions can make it more difficult to confidently identify what is and what is not an impact crater. A crater counter who is confident in their personal definition of a crater can still make mistakes. The week's workloads, personal factors, or even time of day can make people more or less attentive. Not seeing an obvious crater due to being distracted will add variability between counts. Studies which compare expert and non-expert crater counts show wide variations between individuals [10].

In addition to simply identifying craters, there are subjective choices in selecting which craters to count and which not to count. As noted above, the empirical production functions and associated chronology functions used in association with SFDs assume homogeneous surfaces, unaffected by significant numbers of secondary impacts and partial resurfacing events. This can encourage crater counters to subjectively identify and pre-filter primaries from secondaries. It also encourages crater counters to subjectively identify SFD intervals which are believed to be unaffected by resurfacing. Whilst there are sensible criteria for making these decisions (such as ignoring crater clusters associated with ejecta rays, and selecting SFD intervals to avoid clear discontinuities in distributions), these decisions are, ultimately, personal judgement-calls. This situation limits the quantitative validity of any resulting summary estimates.

Poisson error assumption overlooks other sources of uncertainty: The cratering process, where rare impact events occur within a continuous time frame, should be well modelled by the Poisson distribution. Indeed, the Poisson distribution was designed precisely for such data. There is no theoretical problem in considering the true number of craters on a surface as being a Poisson random variable. However, the crater count in a given size interval is not the 'true' number of craters in a region, but a subjectively counted number of craters, biased by human perception and attention. This additional uncertainty has yet to be fully incorporated into error bars on SFDs. The consequences of this include potential over-interpretation of data, underestimation of errors on fitted parameters, and confusion as to the 'correct' statistical methods to adopt when quantitatively analysing SFDs.

Under-use and misuse of figures of merit: There is confusion within the field of crater counting as to how figures of merit should be applied, especially in light of additional variability from human subjectivity and attentional issues. The majority of papers in the crater counting literature fail to provide meaningful goodness-of-fits for assessing the adequacy of models when fitted to target data. Despite long established recommendations [1], figures of merit such as K-S tests and chi-squared tests [2] are inappropriately used, or not used at all. It would appear that when they are used, the K-S test is preferred (perhaps due to the cumulative plotting convention), despite the benefits of a well-formed chi-squared test. And if chi-squared tests are used, their use on cumulative data violates the independence assumptions upon which the method is based. To compound matters, any confidence intervals computed from such tests do not take into account the additional uncertainties noted above, again risking over-interpretation. Together, the lack of usage on the one hand and the lack of good error modelling on the other limits the quantitative validity of fitted results.

As an example, the following paragraph is taken from [10] explaining their use of K-S tests to compare SCFDs by different crater counters and software, created from a common set of Lunar surface regions.

"This test finds the maximum vertical separation between the two curves for the full range of x-values (we use crater diameter), normalized to a cumulative value of 1.0. This maximum is then compared with expected values to determine the probability (P-value) that the null hypothesis (the distributions are the same) is rejected. We used a P-value of > 0.05 to accept the null hypothesis. A P-value of ≤ 0.05 is used to state that two populations are likely different, and a P-value of ≤ 0.01 is interpreted to reject the null hypothesis and state that they are different; 0.05 is the most common value used in statistics but the smaller value was used because the K-S test does not take error bars into consideration."

A K-S test is designed to quantify the level of agreement between two distributions which are assumed to be equivalent. This null hypothesis can be rejected if the size of differences between distributions cannot be easily explained by differences arising by the sampling errors alone. The standard probability tables for this test are computed accordingly, and therefore are only appropriate if the data being compared conforms to Poisson (strictly multi-nomial) distributions. Given the purpose and properties of the K-S test, its use in [10] is inappropriate for three related reasons:

- The assumed $\sqrt{(N)}$ errors placed upon crater counts within the paper describe the natural perturbations in crater **production**, not crater **counting**. As such, it is only a reasonable error model for true crater counts in **independent** identically distributed (i.i.d.) surface regions.
- Any two crater count distributions being compared were taken from identical areas of the lunar surface, so by definition the production process had to be identical.
- The actual counting errors between counters within the paper are shown to be greater than Poisson, with deviations between experts reaching to more than 20% even for counts approaching 1,000 craters. This additional variability is not considered in the K-S test P-value look-up tables.

We argue that the only reason why a K-S test should be used in reproducibility studies is to calibrate the probability look-up tables used to assign P-values to observer discrepancies. As these calibrated P-values will presumably be used later for comparison of independent surfaces, this requires a further step to include the Poisson cratering process. Similar issues arise in the identification and counting of any feature, be they dunes, craters or otherwise.

Might Linear Poisson Models overcome current crater counting limitations?

The problem of subjectivity in defining craters cannot be fixed by applying LPMs, but as noted earlier, a given definition can be applied objectively. This can be seen in [15]. However, with regards to levels of uncertainty and figures of merit, LPMs can improve matters. The estimated quantity of features counted by a LPM considers the underlying Poisson variability in true feature counts, as well as additional variability due to ambiguity between classes of feature. Furthermore, the LPM error theory is developed sufficiently well as to form a good basis for constructing new hypothesis tests for the conformity of different counts to one another.

3 Training data

But what about the training data? Can an objective system be created using subjective examples?

If an agreed standard definition for planetary features could be achieved and applied objectively then LPMs might become feasible. However, such standards would be needed for different illumination conditions, for different terrain types and for different planets with different rates and style of erosion. The logistical and political barriers to this problem in an established crater counting community could be prohibitive. However, for personal use, an individual or group of individuals may wish to apply their own standards. This raises the question of how much agreement can be reached amongst individuals. Robbins et al. demonstrated clearly that experts and citizen scientists disagreed widely on how many craters were present in common areas in lunar images, showing a lack of reproducability. Several problems have been identified, including the issue of "roll-off", where marking up down to a minimum crater size can cause poor reproducibility around this threshold. However, they did not go as far as to check the repeatability of individuals.

Can the repeatability of individuals be tested?

To help answer this question, undergraduates from the school of Earth, Atmospheric and Environmental Sciences, Manchester, and an expert crater counter were asked to repeatedly annotate regions of the Apollo 17 landing site. The smallest crater size was thresholded above the minimum mark-up size to avoid roll-off. Each undergraduate and the expert were asked to annotate the regions twice, at least 24 hours apart.

In order to gather a sample of crater annotations, a custom macro was written for the ImageJ software package that allowed a crater to be highlighted by clicking on 3 points around its rim. The macro computed a circle from these points with the coordinate of the centre and the diameter being recorded into a text file. Repeatability data was gathered from selected regions of a NAC image. A crater counter using the ImageJ macro manually annotated within the regions over the course of two sessions. The sessions were set a day apart to reduce attentional and learning effects. The clustering method used to coalesce MoonZoo data [13][3] was applied to merge the repeated annotations together. The clustered outputs were used to determine the number of single and double counted craters, F_1 and F_2, respectively.

The underlying population of craters in the study is fixed by the selection of NAC image regions, providing a stable cohort from which to sample. Assuming a fixed probability, P (see Appendices A and B), that a crater will be annotated during a single counting session within this cohort (and $1 - P$ that it will not), the number of craters annotated in both sessions (F_2), in only one session (F_1), and in neither session (F_0) can be predicted:

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

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

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

where N is the predicted number of total craters that would be counted eventually by the counter. The repeatability data give binomial sample estimates F_1 and F_2 from which P and N can be estimated:

$$P = \frac{2F_2}{F_1 + 2F_2} \tag{4}$$

$$N = \frac{(F_1 + 2F_2)^2}{4F_2} \tag{5}$$

with errors on the estimated count (due to the observer) given by:

$$\sigma_N^2 = \left(\frac{F_1}{2F_2} + 1\right)^2 var(F_1) + \left(1 - \frac{F_1^2}{4F_2^2}\right)^2 var(F_2) \tag{6}$$

Additional independant variation (consistent with two Poisson generation processes) needs to be added when considering differences between two different regions.

Using the above equations, a crater counting efficiency can be estimated and a total crater count complete with error estimate ($N \pm \sigma_N$) is provided. Note that the estimates given are for the fixed crater cohorts defined by the NAC regions. These efficiencies could be computed for different types of terrain and illumination conditions.

Pairs of crater counts were provided by undergraduate students (see acknowledgements) from the School of Earth, Atmospheric and Environmental Sciences, and also an expert (R. Bugiolacchi) crater counter. Craters were annotated down to 10 pixel diameters for the following NAC image 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

Binomial efficiencies were estimated for each strip for both undergraduate and expert annotations.

How repeatable were they?

Figure 2 shows the range of efficiencies achieved by expert and undergraduate crater counters in the selected test regions. After a single attempt to annotate, between 70% to 85% of what the counters considered to be craters were identified. There was no clear difference between expert and undergraduate repeatability. However, studying Figures 3 and 4 shows that the undergraduates tended to count more craters than the expert, with counts deviating by up to 50% or more. These large deviations are corroboration of similar deviations found in [10]. It shows that experts and undergraduates have different personal definitions of craters, but apply those different definitions with approximately the same level of repeatability.

Within the range of crater quantities studied, these plots suggest only a negligible number of additional craters will be counted (under the personal definition) after two attempted counts, with the estimated total counts, N, being only percentage-level different from the sum of F_1 and F_2.

4 Discussion

Whilst we have assumed that the probability of detecting craters can be summarised with a single variable (P), it is instructive to consider a more general model in order to understand how this analysis may fail. The inability to identify the same set of craters could be due to at least two different causes, efficiency (some are overlooked) and subjectivity (some are difficult to recognise). To describe these mechanisms the detection model needs a minimum of two additional terms (figure 1), and an observed crater count $C = PN$ becomes

$$C' = (N_T - N_A)P_E + (N_F + N_A)P_S \tag{7}$$

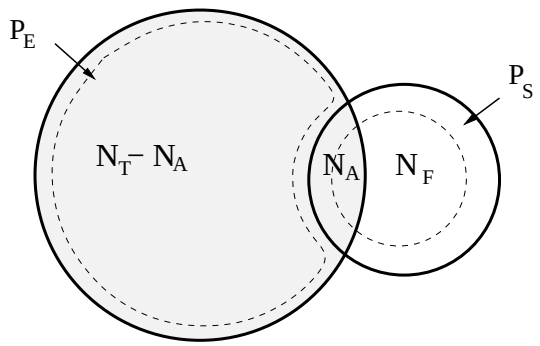

Figure 1: The Venn model of crater mark-up efficiencies when considering ambiguous N_A and false positive N_F craters. If ambiguous and false positive craters are truly indistinguishable then they must have the same probability of detection P_S. Meanwhile, the subset of clearly recogniseable true craters $N_T - N_A$ must have their own detection efficiency P_E to prevent equation(7) from reverting to the original mathematical form.

where C' is now a combination of true craters N_T and ambiguous true craters N_A counted at some level of efficiency P_E, added to a number of false craters N_F (indistinguishable from N_A true craters) counted at a level of efficiency P_S.

This new model reverts to the previous one when $P_E = P_S$ with $N = N_T + N_F$. A perfect crater counter would count 100 percent of N_T and N_A with $N_F = 0$. However, the value of all terms except N_T will vary from person to person depending upon their definition of a crater and how attentive they are to the task. We know that in practive $N_F > 0$. Even for experts, it would be unwise to assume that $P_E = 1$, $N_F = N_A$ and $P_S = 0.5$, in order to make $C' = N_T$, as we cannot expect that $P_S = 0.5$ and the proportions of N_F and N_A may vary between regions in data.

Results from this particular repeatability study would suggest that crater counters would benefit by repeating their own counts twice and correcting for observer efficiency (equation (5)). This corrected count is then equivalent to using $P_E = P_S$ in equation(7) (see Appendix B for details), i.e.

$$N = \frac{(F_1 + 2F_2)^2}{4F_2} \approx N_T + N_F$$

Provided $P_E \approx P_S$, using N as the estimated number of craters eliminates entirely the P_E,P_S and N_A terms in C', but assumes that there are no false positives being counted under the crater definition adopted by the counter ($N_F \approx 0$). As this is unlikely to be true for non-experts this explains the additional steps needed to remove false positive counts from citizen science data [15]. Whether this approach would be enough to eliminate significant differences between experts is a problem which needs further study[1]. We have also ignored for now the effects of small sample counting bias, in the interest of transparency.

In summary, the repeatability of individuals, driven by their counting efficiency, P, accounts for some variability in counts. This variability can be accounted for by limited attention and subjectivity of the counters (P_E and P_S). The larger discrepency between different counters can be accounted for by variable definitions of crater appearence (N_F and N_A). Once combined, these variations explain the overall poor reproducability observed by ourselves and others. Consequently, it is unlikely that expert counters are currently estimating N_T. The assumptions needed to relate equation (5) to equation(7) are not too unrealistic, but suggest this would be estimating $N_T + N_F$ and not N_T.

5 Conclusions

Our data show that even for the simplest of feature recognition tasks, such as crater counting, there is significant variation due to the limited repeatability of individuals when comparing their own counts from the same surface. Further, the wider disagreement between different individuals shows evidence that they work from different definitions of what a crater is [10]. This may be interpretted as either efficiency (attention) or subjectivity (ambiguity). The evidence points to these variations being the dominant source of error in crater counts, being larger than

[1] If differences remain it would be possible to extend the approach based instead on equation (7), but this would require observers to categorise craters as potentially ambiguous during the process of mark-up so that equation (5) could be applied seperately to the two groups.

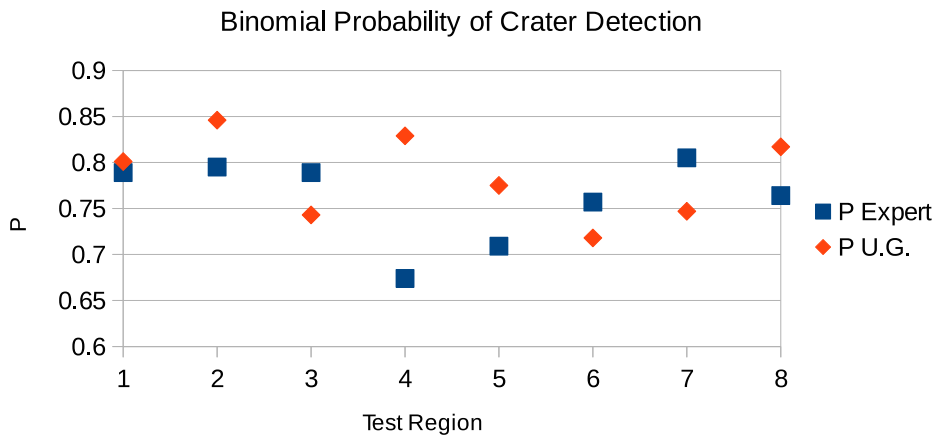

Figure 2: The Binomial success probabilities, P, of expert and undergraduates identifying craters

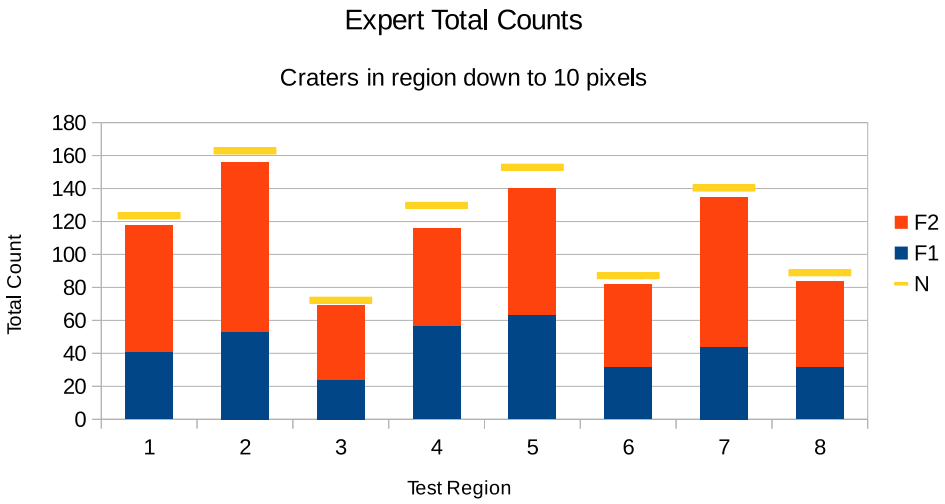

Figure 3: Single and double mark-up frequencies and corrected N for expert counts.

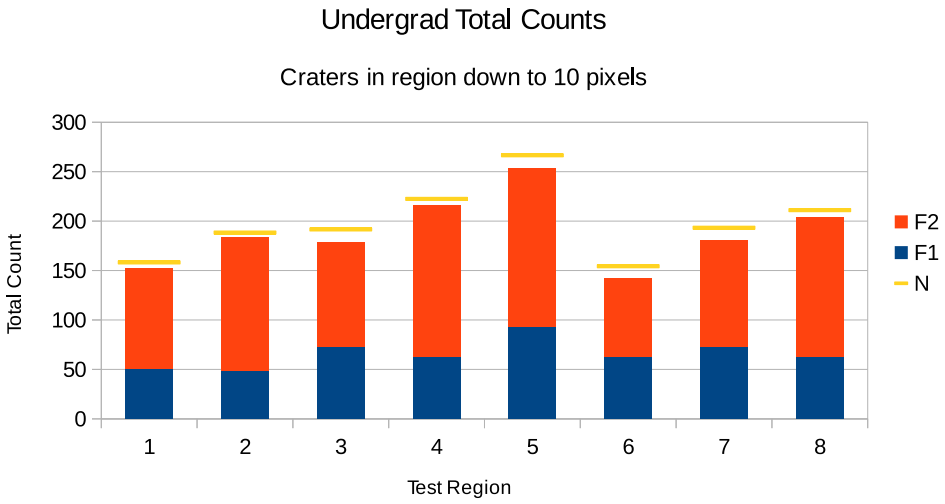

Figure 4: Single and double mark-up frequencies and corrected N for undergraduate counts.

the Poisson variability generally assumed for the 'true' counts. The efficiency of a single expert observer can be quantified and addressed using the double counting approach suggested here. This leaves the subjective differences between counters as the weakest link in any attempt to construct a quantitative automated system.

In principle, a pattern recognition approach based upon training examples could be used as the basis for automated feature counting. In particular, examples of LPMs abilities to estimate quantities within predicted levels of systematic and statistical errors on well-behaved data have been demonstrated [15]. The consistency of results has been demonstrated using different ranges of data binning and encoding [16]. And finally, the statistical efficiency of the method in comparison to some simpler alternatives have been demonstrated on some datasets [16]. However, as with all supervised learning systems, an objective and consistent definition of the features of interest (ground truth) must first be provided.

Evidence from other researchers and our own experiments show that human subjectivity is the dominant source of uncertainty within crater counts. We argue that similar issues are likely to exist in the manual identification and quantification of any planetary features. In summary, whilst the technology to objectively search for features and provide quantitatively useful outputs is developing, their potential for use still relies upon planetary researchers resolving their definitional differences and providing good training data sets.

Acknowledgements

We thank University of Manchester undergraduate BSc(Hons) and MEarthSci (Hons) Geology with Planetary Science students Sean Corrigan, Alex Griffiths, Tim Gregory, Hazel Blake, Dayl Martin, Maggie Sliz, Joe Scaife and Pavel Kamenov for their assistance in providing crater counts. We also thank R. Bugiolacchi for expert crater counts. We thank Dr M. Jones for supervising the students in this work and Dr K. Joy for additional support. We'd also like to thank the Leverhulme Trust for providing project funding through RPG-2014-019 and the Science and Technology Facilities Council for support through grants ST/M001253/1 and ST/J001643/1.

Appendix A: Assuming Equal Probabilities

The crater markers were instructed to mark up all the craters they could see, so it is reasonable to expect they achieve this with equal efficiency on each attempt for simplicity. However, if we do not believe that both attempts are equivalent (for example one mark-up was terminated prematurely), we could perform the analysis in a way which allows different probabilities (P_1 and P_2) on each attempt. If the number of craters seen only once in each case is recorded separately, $F_1 = F_{11}+F_{12}$, then this information can be used to solve for the separate probabilities and the number of missed craters (F_0), i.e.

$$F_0 = \frac{F_{11}F_{12}}{F_2}, \qquad N = F_0 + F_{11} + F_{12} + F_2$$

and

$$P_1 = \frac{F_2 + F_{11}}{N}, \qquad P_2 = \frac{F_2 + F_{12}}{N}$$

Note that two equavalent attempts should have equivalent crater estimates in accordance with binomial statistics, allowing us to test this assumption if needed.

Appendix B: Combinations of Binomial Samples

In circumstances where the data have been generated by two different processes, corresponding to N_a and N_b events and P_a P_b probabilies we would have single detections given by

$$F_1 = 2(P_a N_a(1-P_a) + P_b N_b(1-P_b))$$

which we might say is equivalent to a single generator

$$= 2P(N_a + N_b)(1-P)$$

but then double detections are not consistent with this definition of P

$$F_2 = P_a^2 N_a + P_b^2 N_b \neq P^2(N_a + N_b)$$

Putting $f = N_b/(N_a + N_b)$ and $P_b = P_a + \Delta$ we find that the ratio $N/(N_a + N_b)$ can be written as

$$N/(N_a + N_b) \;=\; \frac{P_a^2 + 2fP_a\Delta + f^2\Delta^2}{P_a^2 + 2fP_a\Delta + f\Delta^2}$$

It then clear that as the upper term must always be less than or equal to the lower (because $0 \leq f \leq 1$), we can deduce $N \leq (N_a + N_b)$. This problem disappears if $P_a = P_b$, so that we can only use equation 5 to approximate $N_a + N_b$ on the assumption that $P_a \approx P_b$ is sufficiently close.

In the specific case of crater counting by non-experts ($N_b = N_A + N_F$ being the ambiguous and false positive craters), typical values might be; $P_a = 0.8$, $\Delta = -0.2$, $f = 0.3$ then when using equation (5), $N_a + N_b$ will be underestimated by 1.15%.

In the general case, samples which are multiple combinations of binomial systems (P_i) cannot be correctly analysed using a simplified binomial model (P). Ideally, in order to apply this procedure, each distinguisable group associated with a distinct probability of detection needs to be counted and corrected individually. This makes application of re-count approaches difficult in the real world as in some cases we may not even know there are different groups[2].

References

[1] R. Arvidson, J. Boyce, C. Chapman, M. Cintala, M. Fulchignoni, H. Moore, G. Neukum, P. Schultz, R. Strom, A. Woronow, and R. Young. Standard techniques for presentation and analysis of crater size-frequency data. *Icarus*, 37:467–474, 1979.

[2] R.J. Barlow. *Statistics: A Guide to the use of Statistical Methods in the Physical Sciences*. John Wiley and Sons, UK, 1989.

[3] R. Bugiolacchi, S. Bamford, P. Tar, N. Thacker, I.A. Crawford, K.H. Joy, P. Grindrod, and C. Lintott. The moon zoo citizen science project: Preliminary results for the apollo 17 landing site. *Icarus*, 271:30–48, 2016.

[4] B.A. Ivanov, G. Neukum, W.F. Bottke Jr, and W.K. Hartmann. The comparison of size-frequency distributions of impact craters and asteroids and the planetary cratering rate. *Asteroids III*, 2002.

[5] K. Joy, I. Crawford, P. Grindrod, C. Lintott, S. Bamford, and A. Cook. Moon zoo: citizen science in lunar exploration. *Astronomy & Geophysics*, 52 (2):2.10–2.12, 2011.

[6] G.G. Michael and G. Neukum. Planetary surface dating from crater size-frequency distribution measurements: Partial resurfacing events and statistical age uncertainty. *Earth and Planetary Science Letters*, 294:223–229, 2010.

[7] G. Neukum, B. Ivanov, and W.K. Hartmann. Cratering records in the inner solar system. *Chronology and Evolution of Mars*, pages 55–86, 2001.

[8] G. Neukum, B. Konig, and J. Arkani-Hamed. A study of lunar impact crater size-distributions. *The Moon*, 12:201–229, 1975.

[9] S.J. Robbins, I. Antonenko, P.L. Gay, C. Lehan, and J Moore. Cataloging the moon with the cosmoquest moon mappers citizen science project. *43rd Lunar and Planetary Science Conference*, 43, 2012.

[10] S.J. Robbins, I. Antonenko, M.R. Kirchoff, C.R. Chapman, C.I. Fassett, R.R. Herrick, K. Singer, M. Zanetti, C. Lehan, D. Huang, and P.L. Gay. The variability of crater identification among expert and community crater analysts. *Icarus*, 234:109–131, 2014.

[11] P. D. Tar and N. A. Thacker. Linear poisson models: A pattern recognition solution to the histogram composition problem. *Annals of the BMVA*, 2014(1):1–22, March 2014.

[12] P. D. Tar, N. A. Thacker, J.D. Gilmour, and M.A. Jones. Automated quantitative measurements and associated error covariances for planetary image analysis. *Advances in Space Research*, 56(1):92–105, 2015.

[2]The classic application is in capture/recapture for animal populations, which will not work correctly if two indistinguishable groups of animals within one sample have different capture probabilities (e.g. due to learning effects such as heightened awareness or availability of free food).

[13] P. D. Tar, N. A. Thacker, and MoonZoo Team. Coalescence and refinement of moon zoo crater annotations. *EPSC 2014*, 2014, 2014.

[14] P.D. Tar. Quantitative planetary image analysis via machine learning (thesis). *University of Manchester*, 2014.

[15] P.D. Tar, R. Bugiolacchi, N.A. Thacker, J.D. Gilmour, and MoonZoo. Estimating false positive contamination in crater annotations from citizen science data. *Tina Memo (www.tina-vision.net, submitted to Earth, Moon and Planets)*, 2016, 2016.

[16] P.D. Tar and N.A. Thacker. The application of linear poisson models to changes in adc measurements. *Tina Memo (www.tina-vision.net)*, 2015, 205.

[17] C.R. Tooley, M.B. Houghton, R.S. Saylor, C. Peddie, D.F. Everett, C.L. Baker, and K.N. Safdie. Lunar reconnaissance orbiter mission and spacecraft design. *Space Science Reviews*, 150 (1-4):23–62, 2010.