Quantitative Planetary Image Analysis via Machine Learning

A thesis submitted to the University of Manchester for the degree of PhD in the faculty of Engineering and Physical Sciences

2014

Paul D. Tar

School of Earth, Atmospheric and Environmental Sciences

Contents

Word count: 55,000

List of Figures

List of Tables

Abstract

The University of Manchester

PhD

Quantitative Planetary Image Analysis via Machine Learning

28th March 2014

Over recent decades enormous quantities of image data have been acquired from planetary missions. High resolution imagery is available for many of the inner planets, gas giant systems, and some asteroids and comets. Yet, the scientific value of these images will only be fully realised if sufficient analytic power can be applied to their large scale and detailed interpretation. Unfortunately, the quantity of data has now surpassed researchers' abilities to manually analyse each image, whilst available automated approaches are limited in their scope and reliability. To mitigate against this citizen science projects are becoming increasingly common allowing large numbers of volunteers, using web-based resources, to assist in image interpretation. Yet human involvement, expert or otherwise, introduces additional problems of subjectivity and consistency. This thesis argues that what is required is an objective, quantitative, automated alternative.

This thesis advocates a quantitative approach to making automated measurements from a range of surface features, including varied terrains and the counting of impact craters. Existing pattern recognition systems, and established practices, found within the imaging science and machine learning communities will be critically assessed with reference to strict quantitative criteria. This criteria is designed to accommodate the needs of scientists wishing to undertake quantitative research into the evolution of planetary surfaces, permitting measurements to be used with confidence. A new and unique method of pattern recognition, facilitating the meaningful interpretation of extracted information, will be presented. What makes the new system unique is the inclusion of a comprehensive predictive theory of measurement errors and additional safeguards to ensure the trustworthiness and integrity of results.

The resulting supervised machine learning/pattern recognition system is applied to Monte-Carlo distributions, martian image data and citizen science lunar crater data. Conclusions are drawn that applying such quantitative techniques in practice is difficult, but possible, given appropriately encoded data and application specific extensions to theories and methods. It is also concluded that existing imaging science practices and methods would benefit from a change in ethos towards a quantitative agenda, and that planetary scientists wishing to use such methods will need to develop an understanding of their properties and limitations.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

I would like to thank all those involved in the production of this thesis, including:

- Neil Thacker, Jamie Gilmour and Merren Jones for excellent supervision and expert input;
- Katie Joy, Roberto Bugiolacchi, Ian Crawford, the Moon Zoo team and many citizen scientists for providing lunar science input and raw data;
- Visvanathan, David Hodgetts and Bart Van Dongen, as advisors / examiners;
- Sean Corrigan, Alex Griffiths, Tim Gregory, Hazel Blake, Dayl Martin, Maggie Sliz, Joe Scaife and Pavel Kamenov for their assistance in providing ground-truth crater counts;
- The STFC for providing project funding and The BMVA for travel expenses;
- and Oszkar Tar, Hillary Tar and Beth Marshall for their support during the project

I would like to further acknowledge co-authors of additional outputs, including:

- Neil Thacker, Jamie Gilmour and Merren Jones for contributions to the following abstracts/papers: “Linear Poisson Models: A Pattern Recognition Solution to the Histogram Composition Problem”, Annals of the BMVA 2014, Vol. 1, 2014; “A Quantitative Approach to the Analysis of Planetary Terrains”, Proc. Remote Sensing and Photogrammetry Soc. Conf., 2012; “Automated Quantitative Planetary Measurements”, Proc. European Planetary Science Congress, 2013; “Coalescence and refinement of Moon Zoo crater annotations”, Proc. European Planetary Science Congress, 2014; “Quantification of false positives within Moon Zoo crater annotations”, Proc. European Planetary Science Congress, 2014; “Automated Quantitative Measurements and Associated Error Covariances for Planetary Image Analysis”, Preprint submitted to Advances in Space Research
- Neil Thacker, Jamie Gilmour, Merren Jones, Katie Joy and Adam Mcmahon for contributions to the following grant proposals: “Quantitative Use of Pattern Recognition in the Analysis of Complex Data Distributions”, successfully funded by Leverhulme Trust; “The Application of Quantitative Pattern Recognition to the Analysis of Planetary Image Data” as part of the consolidated STFC proposal “Cosmochemistry and Planetary Science at The University of Manchester” currently under review.

1 Introduction

This thesis is a summary of work undertaken towards the creation of an automated system for the analysis of planetary images. This document combines an understanding of planetary science requirements, statistical methods and imaging science techniques, resulting in a flexible framework for making a range of quantitative measurements from surface imagery. Theoretical considerations and empirical results from simulated and genuine planetary data will be presented. It is intended that the methods, software, results and conclusions developed here will contribute to solving the immediate problem of how to best utilise the vast quantities of image data being returned to Earth from space missions across the solar system.

This chapter will provide an introduction to imaging from space and state the motivations for creating an automated analysis system. It will also explain the requirements of quantitative measurements and articulate key concepts which will be applied throughout this work. These key concepts include: the application of scientific methodologies in the context of imaging science and planetary research; the importance of understanding measurements; and the use of statistics.

1.1 The rise of imaging from space

Humanity saw Earth from a new perspective on 24th October, 1946, when a captured German V-2 rocket lifted the first ever space-bound camera above New Mexico on a short sub-orbital flight [1]. From an altitude of over 65 miles, a 35-millimeter motion picture camera took grainy monochromatic views of cloud tops and desert covering 40,000 square miles of Earth's surface. This was the beginning of imaging from space. The following decades saw the establishment of American, Russian and European space agencies and the development of sophisticated technologies leading to thousands of images being acquired from targets beyond Earth and across the solar system. More recently a new Asian space-race has seen China, India and Japan embark upon ambitious lunar, martian and asteroid exploration programs, whilst established space agencies continue to accumulate data from active missions and plan others.

The science case for collecting image data during space missions will be presented shortly, but first the scale of the resulting image interpretation problem will be emphasised through a brief history of space-based imaging. The following subsections convey a simple message: there is a lot of data and much more will follow[1].

[1] The figures used within the following subsections were mostly taken from official NASA and ESA mission websites containing dynamic content and are therefore not directly referenced. Links and further information can be found at www.nasa.gov, www.esa.eu and hirise.lpl.arizona.edu

1.1.1 Historical images

Starting close to home, throughout the 1950s and 60s robotic probes visited the Moon repeatedly. These included the American Pioneer, Ranger [2], Surveyor [3] and Lunar Orbiter [4] programs and various Russian Luna and Zond missions [5]. Many of these probes had limited or no imaging capabilities. Those that did provided relatively few low to mid-resolution surface images, including Ranger 7 which took around 4,000 frames and Ranger 8 which gave researchers over 7,000, covering Oceanus Procellarum and Mare Nubium regions. The Lunar Orbiter missions provided unprecedented global coverage of the Moon with resolvable details close to 1 meter in some places. These orbiters gathered almost 3,000 images.

Moving out to neighbouring planets, the 1960s and 70s saw many failed Mars missions, but the NASA Mariner [6] fly-bys and Viking orbiter and lander program [7] eventually succeeded. The first close-up views of Mars were captured by Mariner 4 in 1965 giving only 21 low-resolution frames. By the mid-70s and up until 1980 the 2 Viking orbiters gathered over 50,000 images. These covered the entire martian surface at resolutions of 150 to 300 meters, with selected regions mapped with resolutions as fine as 8 meters. Until the 1990s data gained via these missions were the dominant source of researchers' knowledge of Mars.

Expanding outwards, the prolific Voyager program [8] launched twin probes in 1977 to explore the gas giants and their moons. Between them during a series of fly-bys they studied Jupiter and Saturn, with Voyager 2 continuing on to Uranus and Neptune. These fly-by missions provided short windows of opportunity to gather image data. Between the two probes over 33,000 images of Jupiter and its five largest satellites were taken. These remained the most detailed images of the jovian system available until the orbiting Galileo [9] probe reached Jupiter in late 1995. Galileo returned approximately 30 gigabytes of data, including 14,000 images with much improved resolution.

Until the late 1990s these and other missions produced images in their tens of thousands over a period of decades, mainly with resolutions no better than orders of hundreds of meters. However, advances in technology soon provided magnitude changes in the quantity and resolution of data.

1.1.2 Contemporary images

From the early 2000s there have been many high profile missions launched within the inner solar system which at the time of writing are still actively gathering data. These include Mars orbital missions: Mars Odyssey(NASA, launched 2001); Mars Express [10] (ESA, launched 2003); and Mars Reconnaissance Orbiter [11] (NASA, launched 2005), Mercury orbiter mission MESSENGER [12] (NASA, launched 2004), dual asteroid mission DAWN [13] to Ceres and Vesta (NASA, launched 2007), and the Moon mission Lunar Reconnaissance Orbiter [14] (NASA, launched 2009). Between them they have provided planetary scientists with almost complete coverage of Mars, Ceres, Mercury and the Moon, including significant subsets with resolutions down to 0.25 meters. This thesis will make use of Lunar Reconnaissance Orbiter and Mars Reconnaissance Orbiter images in later chapters.

Data and imagery is being continually accumulated from such missions. A NASA mission web page (http://mars.jpl.nasa.gov/mro/) reports that over 25.5 terabytes of data has so far been received from Mars Reconnaissance Orbiter, as of February 2014. Some of this data is made public via the HiRISE project, which currently provides a catalogue of over 30,000 high-resolution images typically 160 megapixels in size. The size of these images corresponds to thousands of times as much data as their historic Viking counterparts.

Meanwhile, the Lunar Reconnaissance Orbiter has accumulated over 192 terabytes of data within its first few years of operation. This includes high-resolution imagery revealing the lunar surface in unprecedented detail. Images are of such high quality that tracks of astronauts' footprints can be seen surrounding Apollo landing sites. In these images the remains of Apollo lunar landing hardware can clearly be seen spanning several pixels.

Comparable quantities of data at similar resolutions continues to be gathered from contemporary missions. Pixel-for-pixel, the total quantity of modern image data equates to millions of historic low-resolution images. Images and other data from NASA missions are archived in the Planetary Data System (PDS) [15] where they are made available to researchers.

1.1.3 Future images

Over the coming years many other robotic probes will begin to contribute to the increasing quantity of available imagery. Probes en-route include: the Rosetta probe [16] (ESA) which will arrive at comet Churyumov-Gerasimenko in late 2014; New Horizons [17] (NASA) will fly-by Pluto mid-2015; DAWN [13] (NASA) will orbit Ceres early 2015; and Juno [18] (NASA) will enter the jovian system mid-2016. Scheduled near-term missions yet to launch include: BepiColombo [19] (ESA/Japan) to orbit Mercury; Chandrayaan-2 [20] (India) lunar orbiter; Luna-Glob 1 [21] (Russia) lunar orbiter; and Chang'e lunar missions (China). Amongst their many instruments, these missions carry a range of imaging equipment.

This brief tour of imaging from space has focused on a limited set of fly-by and orbital mission examples, yet there are many other fly-by and orbital missions, landers and rovers which have imaging science elements. There are also ground and space-based telescopes generating further streams of images, all increasing the total quantity of astronomical data, and all requiring analysis in order to extract scientifically useful information. Examples of potential science applications for this information are described in the following section.

1.2 Science case

Scientific applications for planetary data are numerous and varied. This section will highlight a small number of examples specific to lunar and martian applications providing background science for the use of lunar crater data and martian terrain images in later chapters.

1.2.1 Lunar science

Impact craters can provide a wealth of information regarding the geological evolution of the Moon's crust [22]. Individual craters can be used to study local variations in crustal stratigraphy, including estimates of regolith depth, subsurface composition and subsequent erosion [23]. This information can be inferred from patterns of bright ejecta from young craters [24][25], the presence of boulders excavated during impacts [26][27], morphological features in bench craters which indicate the layering of underlying material [28] and levels of degradation.

Beyond individual impacts the total density of craters, as summarised using Size-Frequency Distributions (SFD), can be used to infer relative surface ages where geological units containing larger numbers of craters are considered older [29]. An SFD summarises the number of impact craters within a region, typically normalised to craters per unit area, binned into geometrically increasing size bands. SFDs have been calibrated against returned lunar samples using radiometric dating methods to provide a tentative absolute model age scale [30]. The study of SFDs can also provide estimates of the changes in cratering rates over time. This can be used to infer the population of small bodies in the inner solar system.

Crater counting is a valuable tool often used in conjunction with other information to study lunar volcanism. This includes estimating the thickness of mare basalt flow units [31], the chronological ordering of exposed basalts [32] and studies of stratigraphy and composition of lava flows [33].

The study of cratering is not limited to lunar science [34], nor is lunar science limited to crater counting. However, given that craters are a common feature to most bodies within the solar system their automated counting is a prime application and the Moon is an ideal testbed for such work.

1.2.2 Martian science

Mars, in comparison to the Moon, is a highly dynamic body [35]. For example, images of Mars contains evidence of: recent and ancient drainage networks; seasonal deposits of carbon-dioxide ice; many fields of large dunes and smaller dune-like features known as Transverse Aeolian Ridges (TARs); past volcanism; and many impact craters. Evidence of these is seen in indicative patterns and textures throughout martian terrains, some only recently discovered through contemporary high-resolution images. Studying these features can aid researchers in their understanding of the evolution of Mars, including changes in geological and atmospheric processes.

There is much interest in drainage systems, with recent seasonal flows being observed within the walls of some craters [36]. Contemporary flows are believed to be concentrated brine, capable of existing in liquid form for short periods in Mars' cold tenuous atmosphere, although there is some debate on this interpretation [37]. Their study may help answer questions regarding the current quantity of water on Mars and the depths at which it might be found within subsurface aquifers. Also, the distribution and orientation of ancient drainage systems can be used to infer Mars' climate history [38] and the timing of some geodynamic events [39].

The record of inactive drainage channels visible in martian images might be exploited to illuminate the global tectonic evolution of Mars. The concept of palaeotopography to reconstruct vertical motions of the lithosphere has been widely used on Earth [40], but much less so in planetary applications [41]. Palaeotopographic reconstructions are based on the assumption that fluids in drainage networks follow the path of steepest descent. The orientation of channels might, therefore, be used to infer the down-slope direction of surfaces during the times channels were active. This can be compared to modern topographic data (i.e. Mars Orbiter Laser Altimeter: MOLA) to reveal any changes.

There is an active carbon-dioxide cycle on Mars which is especially evident around the polar regions [42]. Seasonal eruptions of sublimating CO2 causes: dark patches to appear around dunes; radiating fissures to develop known as 'spiders'; and geysers to eject fans of dark material. The study of these events can help researchers better understand this CO2 cycle. Information about wind speed and direction can also be inferred from the direction and length of fans.

The density, morphology and orientation of dunes can be used to estimate the availability and size of grains [43], wind speeds and wind directions [44]. This information can inform grain transportation models and atmospheric models which in turn can be used to better understand weather and erosive processes. Currently active dunes have been observed with orientations consistent with wind azimuths obtained from atmospheric Global Circulation Models [45] (GCM). Others appear to evolve slowly over hundreds of thousands of years, thereby providing records of past climatic changes [46].

Finally, the study of martian craters can be used to place bounds on the timing of surface modifying events and to place surface changes into chronological order. This has been used to date caldera in volcanic regions.

1.3 Image interpretation

Image data is complex and requires considerable interpretation. The science applications given above entail the identification and quantification of features such as craters, dunes and drainage networks. Historically this has been achieved manually by experts, but alternative approaches are now being explored.

1.3.1 Manual analysis

Manual interpretation dominates the analysis of planetary surface imagery. As an example, a large amount of work has been undertaken by experts to build catalogues of craters. Lunar Orbiter Laser Altimeter (LOLA) data has been used to compile crater statistics on the Moon [47]. Data from Mariner and MESSENGER missions have also been used to compile global crater population information for Mercury [48]. In both cases only craters with diameters equal to or greater than 20 km were included. These were compiled using manual mark-up tools provided by Geographic Information Systems (GIS).

Expertly compiled databases of martian dunes have been created [49] using Thermal Emission Imaging System (THEMIS) and infrared (IR) images. These have been limited to moderate and large sized dune fields with areas above 1 km^2, with results designed for manipulation in GIS packages. Global maps of Saturn's largest moon Titan have also be compiled using data from the Cassini-Huygens mission [50] revealing widespread fields of dunes.

Also, martian drainage networks have been manually mapped using Viking data [51] and more recently using Mars Odyssey THEMIS and IR images [52] for use in climate and hydrological studies.

The above examples constitute a very small sample of manual work undertaken by experts. They are necessarily limited in scope, often utilising relatively low resolution data giving large contextual results. In contrast, fine details in high-resolution images are being mapped with the aid of large numbers of volunteers, or 'citizen scientists', via web-based interfaces. These include the mapping of small lunar craters as part of the Moon Zoo [53] and Moon Mappers projects [54], and the mapping of seasonal carbon-dioxide 'fans' on Mars via Planet Four [55]. These projects have been successful in gathering large quantities of data, with work currently being undertaken to interpret their outputs including this thesis, which utilises Moon Zoo data in chapters 8 and 9.

1.3.2 Automated analysis

An in-depth analysis of automated methods will be provided in chapter 2. Here, only a brief summary of automated approaches to planetary image analysis will be presented.

General terrain classification has been attempted using texture information and multi-spectral data. Early work on terrain classification methods made use of pixel co-occurrence statistics [56]. In this work, the distribution of local pixel patterns was learned, with derived quantities (such as entropy) being used to differentiate between surface types. This texture-based approach utilised the repeating patterns present in different terrains for classification. In contrast, most contemporary work on terrain classification (often from remotely sensed Earth observation data) makes use of multi-spectral data rather than texture. This pixel-based approach utilises the spectral profile of different terrains for classification, perhaps using as few as 3 colour channels [57][58][59][60].

Many crater detection algorithms have been proposed for use on lunar and martian data, most following a common design pattern [61][62][63][64][65][66][67][68][69][70][71]. Typically, raw image data (optical or topographic) is first encoded using a higher-level descriptive format (edge strings, Haar transform, texture descriptors, templates etc.), before being fed into a classifier. Performance is then evaluated in terms of numbers of correct verses incorrect classifications. The reported efficiencies of these algorithms usually range between 60% to 80% for correct detections, with many false positives reported. The extraction of dunes has also been approached using martian data and HOG descriptors [72], yielding similar results to crater detection performance.

The automated extraction of martian valley networks has been attempted using topographic data (MOLA) [73][74][75]. This work is based upon finding paths of steepest descent through digital terrain modes, thereby mapping likely tributaries and main channels of drainage networks. The relatively low resolution of MOLA data results in less detailed maps than can be generated manually, which are often based upon high-resolution visual imagery. Other features have also been extracted from MOLA data, including fault scarps [76].

1.4 Measurements

Planetary scientists analyse image data qualitatively and quantitatively. A qualitative analysis involves giving high level descriptive accounts of data, capturing some essence of the processes being observed. A qualitative approach need not include objective numerical descriptions, nor well-defined measurements, and therefore introduces subjectivity into research. Such an approach can be appropriate for providing context to questions, developing intuitions and communicating concepts. A quantitative analysis, in contrast, requires clearly defined and meaningful measurements to be made. This can facilitate the rigorous testing of hypotheses and can help develop more detailed understandings and mathematical models of processes of interest.

Measurements are fundamental to the quantitative application of The Scientific Method and are a cornerstone of experimental and physical studies. Measurements which may be taken from planetary images include densities of craters, orientations of dunes, lengths of drainage networks and surface areas of selected terrains. Section 3.1.1 will explore the types of measurements applicable to different surface features during an analysis of an automated system's requirements. As this thesis is primarily concerned with making measurements, time will be taken to elaborate upon their use and importance.

A reader familiar with quantitative methods within the sciences may wish to skip ahead to section 1.5. These subsections are included as background reference material for the literature review in chapter 2. There it will be argued that basic scientific principles are often overlooked in imaging science and pattern recognition.

1.4.1 Quantitative measurements and The Scientific Method

The Scientific Method involves the development of theories which capture the behaviour of phenomena under investigation. Such theories must be capable of making predictions about the behaviour of phenomena given specified initial states. A good theory will yield very specific predictions which can be tested through experiments and observations. Results should either support the theory via successful predictions, or cast doubt on the theory through incorrect predictions. It has been suggested by philosophers of science [77] that the more testable a theory is, the better that theory is, and that the advancement of theories occurs most productively through falsification. Historians of science [78] have also highlighted the difficulties of applying theories in practice and noted that established paradigms (consisting of related theories, experimental practices and research traditions) are difficult to overthrow even when there is significant criticism against them. To increase specificity, testability and productivity this thesis advocates a quantitative approach to science.

The quantitative application of The Scientific Method involves developing formalised mathematical theories. This makes theories more specific and testable, but also more difficult to apply and reconcile with experimental results. When making quantitative predictions the initial state of a phenomenon is embodied by the values entered as parameters to theoretical equations. These may represent physical quantities measured from the system under investigation. The numerical outputs of these equations often predict the values expected from experimental measurements. These might be taken from the system under investigation after some action has occurred which the theory was designed to explain. The numerical nature of the quantitative approach makes comparisons between predicted and observed measurements an objective and highly specific process, hopefully leading to easier falsification of inferior theories. The increased specificity gained through a quantitatively measured approach forces researchers to be more precise with their descriptions of phenomena. These advantages are difficult to realise in more qualitative and subjective traditions.

The increased scientific testability and potential falsifiability achieved through the use of quantitative measurements brings about additional complications. Comparing predicted measurements to observed measurements is an objective and specific process, but it is not necessarily simple. Measurements are inexact. Measurements contain uncertainty which can stem from statistical noise and systematic effects inherent to the measuring apparatus. The initial state values acting as inputs into theoretical equations are often themselves the result of measurements and therefore also subject to uncertainty. Given the inexactitude of measurements, it is almost certain that a predicted measurement will not equal its observed counterpart. Quantitative researchers must therefore rely upon some criteria to decide whether or not a theory should be considered corroborated or falsified based upon the evidence.

There are key questions that must be asked when an observed measurement disagrees with what was predicted by theory:

- Was the difference caused by flaws in the experimental design?
- Was the difference caused by noise?
- How significant is the difference?
- Is there sufficient evidence against the theory to consider it falsified?

Part of the difficulty in applying quantitative theories in practice is the need to design an appropriate experiment which does not introduce effects which may invalidate results, such as outliers in datasets or bugs in computer code. Safeguards must be in place to help identify such cases. Assuming that an appropriate experiment has been performed it is then necessary to estimate the amount of noise likely to be affecting the results. Based upon some perturbation model of how noise propagates through the theory, the size of the discrepancy between prediction and observation must be quantified. Finally, if it is highly improbable that the difference between prediction and observation is due to noise alone then the theory must be reconsidered. There are standard statistical tests and techniques [79] which can be applied to assist in this process. Those which may be of value to a quantitative analysis of planetary images are presented next using a hypothetical remote sensing example.

1.4.2 The role of statistics

A statistical model can be considered as a type of theory. For example, a Gaussian Mixture Model (GMM) might constitute a theory for how different surface compositions influence pixel values in hyper-spectral remotely sensed data [57]. Such models are also common in medical applications [80]. The theory (simplified) may say that there are two terrains spread throughout an image with pixel values following two Gaussian distributions which are linearly combined:

$$\mathbf{M}_i = \mathcal{G}(i; \mu_a, \sigma_a)\mathbf{Q}_a + \mathcal{G}(i; \mu_b, \sigma_b)\mathbf{Q}_b \tag{1}$$

where $\mathbf{M}$ is a model of continuous pixel values; $\mathcal{G}$ is a gaussian distribution with mean μ and standard deviation σ; $\mathbf{Q}$ is a weighting quantity; and a and b denote different terrain compositions.

This is a quantitative theory which can be used to predict approximate frequencies of different pixel intensities, given initial state values of μ, σ and $\mathbf{Q}$. This predictive capability provides a means to guard against outliers and other problems which can be tested for using a goodness-of-fit score. Given a set of hypothesised parameters, the theory, $\mathbf{M}$, can be compared to measurements, $\mathbf{H}$. By including an appropriate perturbation model a difference score can be computed:

$$\chi^2_D = \frac{1}{D}\sum_i^N \frac{(\mathbf{M}_i\delta - \mathbf{H}_i)^2}{\sigma_i^2} \tag{2}$$

This is a standard chi-square per degree of freedom function where: D is the number of degrees of freedom in the model, which is typically equal to the number of data points minus the number of estimated parameters; $\mathbf{H}$ is a histogram of observed frequencies; δ is a small interval approximating the integral of the model over histogram bins; and σ_i^2 is the expected variance on the comparison at point i. If a Poisson perturbation model is assumed for the histogram data then $\sigma_i^2 \approx \mathbf{H}_i$. If the theory, initial state values and perturbation models are correct then the goodness-of-fit should equal unity. Fit values larger than unity means that the deviation between predicted and observed frequencies is worse than expected indicating possible problems. The statistical significance (i.e. probability of discrepancy based upon noise alone) might then be estimated using a chi-square distribution if desired. The ability to spot poor fits between theory and practice is a valuable tool for the trustworthy use of quantitative methods.

Assuming a satisfactory fit of the model to the data has been achieved the theory might be applied in some additional analysis. For example, differences between quantities, $\mathbf{Q}$, of terrain types might be compared to some reference data. To make comparisons meaningful, the noise on the quantity measurements must be estimated. If the quantity measurements themselves were estimated using Likelihood (as is often achieved using Expectation Maximisation in GMMs) then the Cramer-Rao Bound (or lower variance bound) can be applied:

$$\frac{1}{\sigma_Q^2} \leq \frac{\partial \ln \mathcal{L}}{\partial \mathbf{Q}} \tag{3}$$

where $\mathcal{L}$ is the Likelihood function maximised to estimate $\mathbf{Q}$ and σ_Q^2 is the estimated variance on the measured quantity. This provides a lower-bound which is close to the correct variance, assuming sufficient numbers of data points were used during the fit. This can easily be extended to provide full error covariance matrices. This link between Likelihood estimated parameters and error estimates is another valuable tool for the application of quantitative methods.

As a final example of the standard tools available for quantitative analysis the use of derived quantities computed using several estimated measurements can be considered. If the ratio of the two terrain compositions was of interest, $r = \mathbf{Q}_a/\mathbf{Q}_b$, the expected noise on the ratio could be estimated using error propagation:

$$[\sigma_r^2] = \nabla \mathbf{C}_Q \nabla^\intercal \tag{4}$$

where ∇ is a matrix containing the partial derivatives of r with respect to the two quantities and $\mathbf{C}_Q$ is the covariance matrix describing the errors on the two quantities. Error propagation can be used to provide a first-order approximation to output errors based upon small perturbations in input for any differentiable function.

Goodness-of-fit checks, Likelihood estimators, variance bounds and the propagation of errors are all common practices within the physical sciences. Indeed, undergraduate physics students are trained in these methods as a standard introduction to data-driven experimental science, i.e. [79]. However, as will be seen in chapter 2 they are rarely applied in imaging science and pattern recognition research. This is despite encouragement from highly popular texts [81] including Numerical Recipes In C.

1.4.3 Assumptions and approximations

Quantitative theories and statistical tools are often only approximative and subject to assumptions [82]. For example, the typical noise model assumed on continuous valued measurements is Gaussian. As such, it is expected that around 1/3 of data points would lie outside of +/- 1 standard deviation error bars. Whilst the central limit theorem can be invoked to justify Gaussian assumptions, limited quantities of data and other real-world factors will often prevent distributions behaving as Gaussians beyond 3 or more standard deviations. Equally, common approximates to Gaussian distributions, such as Poisson distributions with means above 30, must be used with caution. Other common assumptions, such as independence between data points combined using a multiplicative Likelihood function, cannot be taken for granted. However, strict adherence to The Scientific Method, particularly with respect to testable predictions, can guard against difficulties. Only when predictions are being corroborated consistently through empirical evidence can a theory be trusted, and strictly then it can only be trusted within the domain of values upon which it was tested.

A large part of this thesis will make use of Likelihood and predictive error theories, combining Poisson distributions in various quantities. This will involve many assumptions and approximations. As such the corroboration or falsification of error predictions is an important part of this work. Just as the chi-square per degree of freedom goodness-of-fit function can be used to check predicted histogram frequencies, a similar arrangement can be used, over repeated measurements, to check approximate error estimates. With the aid of ground-truth, the residuals between predicted measurements and true values can be normalised to their errors to give a Pull distribution:

$$\Delta_t = \frac{\mathbf{Q}_t - \mathbf{T}_t}{\sigma_t} \tag{5}$$

where Δ_t is the normalised residual for trial t; $\mathbf{Q}$ is an estimated measurement; $\mathbf{T}$ is a ground-truth value; and σ_t is the predicted error on the measurement. The standard deviation of Δ_t should be unity if error predictions are correct. If a Pull distribution is populated using 1,000 samples the width should be unity, plus or minus approximately 0.022, as given by the standard formula for errors on estimated σ:

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}} \tag{6}$$

where σ_σ is the expected error on the sample standard deviation and N is the number of trials. This thesis will aim to achieve this level of agreement when predicting errors, but will allow a degree of tolorance.

The difficulties in modelling uncertainties in practice leads to the acceptance of approximate error theories. As long as error estimates are correct to within a factor of 2 then the majority of uncertainty in measurements is being modelled. This level of agreement is sufficient for avoiding the over-interpretation of data. For example, it is noted in [79] (page 55) that if more points than expected are observed outside of 2σ then 'there is probably some effect at work that you do no understand' but also that 'points should only be condemned after giving them a fair hearing.' It is also noted in [81] (page 660) that '...reasonable experimenters are often rather tolerant of low probabilities...' when accepting a point from an approximate error distribution, and states that 'This is not as sloppy as it sounds.'

1.5 Argument for quantitative automation

There are two key arguments for the automation of planetary image analysis: there is too much data for humans to inspect manually; and automation potentially avoids problems of human subjectivity.

The quantity of data available has been emphasised already in section 1.1. The practical issues involved in interpreting this vast amount of data will only increase with time as additional data is received. The growth in citizen science projects can only constitute a partial solution to this problem. This will be seen in chapters 8 and 9 where the challenges involved in the interpretation of Moon Zoo data will be explored.

An automated system which can identify and measure a range of surface features could free researchers from tedious low-level tasks, such as the counting of individual craters, allowing them to focus on higher-level interpretations. It is often the summary statistics which are of interest to researchers, not the individual data points. Returning to craters for an example, the density of craters, as expressed in SFDs, is of more interest for chronological studies than individual impacts. Automating low-level tasks using a supervised system which can learn though examples then find and measure similar objects, such as craters, in new data would be of great benefit.

Beyond issues of limited human resources, the manual interpretation of images can be highly subjective. This problem has been illustrated in crater counting [83] where large discrepancies have been observed between different experts and even larger discrepancies observed between minimally trained volunteers. In this work, errors in estimated surface ages could be as large as a billion years. This casts doubt upon the reliability of both expert and citizen science performance. An automated alternative could remove this subjectivity. Such a system should be capable of objectively applying a definition of a feature to make consistent and repeatable measurements.

It will be argued in chapter 2 that most proposed automated techniques, such as those noted in section 1.3.2, are not necessarily appropriate for scientific use. This is largely due to their lack of a predictive error theory, leading to problems when measurements need to be compared to predictions, as described in section 1.4.1. It is therefore necessary to construct not just an automated solution, but a quantitative solution. The criteria for a quantitative solution will be discussed next.

1.6 Criteria for a quantitative system

This thesis seeks to provide *quantitative* tools for interpreting image data. It is therefore important to clearly define what is meant by the word quantitative when used in this document. It is also important to define what is meant by non-quantitative, especially because many existing methods produce numerical results which can not be described as being qualitative, but do not meet the criteria defined below to be adequately described as being truly quantitative either. The quantitative criteria given below directly addresses scientific needs.

A quantitative measurement:

1. must be a numerical estimate driven by evidence found within the data;

2. must be accompanied by an error estimate indicating the expected accuracy of the measurement;

3. must be shown in practice to not deviate from the true measurement by more than is predicted by the estimated error;

4. and where possible be supported by additional checks to ensure the trustworthiness of results.

These criteria do not require measurements to be the most accurate in any absolute sense, but they do require measurements to be honest [84][2]. Whilst it is desirable to

[2] here the word ‘honest’ is being used in a technical sense meaning that the number of correct measurements, within errors, should occur as frequently in practice as the error bars predict.

achieve high levels of accuracy, it is the understandability of measurements that is of most concern. From a quantitative perspective, the ability to give confidence intervals (error bars), or p-values, is far more important for interpretation than merely insisting upon the best possible accuracy. Because of this, measurement techniques which *do not* provide error estimates will be considered *non-quantitative*, as they lack the vital information necessary for confidently interpreting results. Measurement techniques which purport to provide error estimates but *do not* corroborate the use of those error estimates in general cases will also be considered non-quantitative, as error estimates can only lead to confident interpretations if they are reliable. This quantitative criteria will be applied whilst reviewing existing methods. It will also be applied when assessing the success or failure of the quantitative system which will be developed during this thesis.

The above criteria will be interpreted in line with the discussion in section 1.4.3, allowing error estimates which are correct to within a factor of 2 to be considered acceptable. If errors are correct to within a factor of two then confidence can be had that the majority of uncertainty in measurements is being accounted for. An error bar which is half as large as it should be still provides protection against the over-interpretation of results, so long as measurements are only trusted to within 2 to 3 standard deviations. However, if measured quantities are to be used as input to further analyses a stricter interpretation of the criteria may be necessary to prevent the possible growth of unaccounted errors.

1.7 Thesis outline

This thesis is divided into 2 parts: Theory and Application. The theory part will develop the necessary methods for making quantitative measurements. The application part will demonstrate the use of the quantitative methods on synthetic and real planetary science data.

Part 1: Theory

Chapter 2 will critically review relevant literature including existing automated planetary analysis methods, pattern recognition systems and performance characterisation techniques. The selected work will be examined with respect to the quantitative criteria of section 1.6 and its usefulness in a scientific context. Chapter 3 will develop a statistical model for approximating the distributions of spatially repeating patterns found in surface images. This model will suggest how texture histograms constructed from planetary images can be described using linear combinations of probability distributions which can be fitted using Likelihood parameter estimation. It will be argued that the model parameters can be linked to a wide range of measurements which may be sought from planetary surfaces. Chapter 4 will explore how statistical perturbations in incoming data are likely to affect the stability of estimated measurements. Theoretical considerations will be used to predict errors that can be tested using Monte-Carlo simulations. Chapter 5 will extend the analysis of errors by considering how combinations of statistical and systematic effects change as a function of the quantity of analysed data, resulting in a fully quantitative pattern recognition system.

Part 2: Application

Chapter 6 will investigate an image encoding scheme based upon BRIEF for constructing texture histograms. Synthetic martian terrains derived from real martian HiRISE data will be used to test the encoded histogram's abilities and limitations when applied to making terrain surface area measurements. Chapter 7 will make improvements to the BRIEF inspired encoding to better match the assumptions made by the pattern recognition system. The improvements will be shown to allow the pattern recognition system to fulfill quantitative criteria on some terrain types, whilst coming close on others. Chapter 8 will begin investigations into the analysis of real planetary science data from the Moon Zoo project. This chapter will prepare raw Moon Zoo data into a form suitable for populating histograms. This will be achieved using a combination of clustering and template matching. Chapter 9 will make use of the Moon Zoo histograms to make estimates of the quantities of false positive and true positive craters amongst the citizen science data. It will be shown that this estimation task can be achieved, fulfilling quantitative criteria. Finally, chapter 10 will summarise both theory and application, highlighting strengths and limitations of the developed theories and encoding schemes. Opportunities for future work will also be identified.

PART 1: THEORY

The chapters found within this part of the thesis will: review technical literature; develop a statistical model for planetary terrains; and derive an error theory for statistical and systematic uncertainties in quantity measurements.

Supporting material, including preliminary work and publications generated from this part, include:

- P.D. Tar, N.A. Thacker, Linear Poisson Models: A Pattern Recognition Solution to the Histogram Composition Problem, Annals of the BMVA 2014, Vol. 1, 2014
- N.A. Thacker, Can we use Pattern Recognition for Science?, Internal Memo, 2010-008, www.tina-vision.net
- P.D. Tar, Quantitative Counting with Bayes Theorem, Internal Memo, 2011-003, www.tina-vision.net
- P.D. Tar, Extended Maximum Likelihood vs Maximum Likelihood, Internal Memo, 2011-004, www.tina-vision.net
- P.D. Tar, N.A. Thacker, Quantitative Prior Estimation and Independent Component Analysis for Linear Poisson Models, Internal Memo, 2012-003, www.tina-vision.net
- N.A. Thacker, Quantitative Pattern Recognition: Warts and All, Internal Memo, 2013-007, www.tina-vision.net
- P.D. Tar, Tutorial: Using Tina Vision's Quantitative Pattern Recognition Tool, Internal Memo, 2014-004, www.tina-vision.net

2 Literature Review

The related fields of imaging science, pattern recognition and machine learning have become dominated by modular building blocks [85]. Many of these modules are available to download as part of code libraries, ready to be incorporated into new algorithms. Other building blocks are collections of standard techniques and resources, such as bench-mark datasets. This paradigm permits the rapid development and evaluation of new applications. An automated planetary analysis system could be approached in this modular fashion by first selecting an input representation (e.g. SIFT, HOG, Wavelets etc.), then a classification module (SVM, RF etc.), and finally feeding in prepared data and ground-truth to conduct a standard evaluation (ROC curve etc.). Indeed, some of the approaches noted in section 1.3.2 resemble this pattern. This chapter will review standard techniques, including those used in the previously noted literature on automated planetary analysis. This chapter will consider:

- popular representations used to encode image information;
- statistical modelling methods, used to summarise data distributions;
- standard supervised classifiers for categorising data;
- and performance evaluation methods.

Each group of methods will first be described and their links to planetary image analysis will be highlighted. The usefulness of the selected techniques will then be critically assessed with reference to the quantitative criteria of section 1.6. This critical assessment will reject building blocks which are unsuitable for a quantitative analysis system, whilst providing a short-list of potentially useful components for further consideration.

2.1 Representations

Raw image data, i.e. pixels, is not always the most convenient form of input to an analysis. The absolute values of pixels often contain little meaning. Rather, it is the structure of the pixels which contains information. There is a general link between the variability in pixel values and their information content, e.g. uniform regions of an image showing little variation contains less information than regions showing many discontinuities. To account for this, representations often extract structures, such as edges and corners, where there are steep local changes in pixel intensities.

Encoded image data can be used in several ways. In simple cases, where the representation provides an output value related to the strength of a signal, the output can be searched for maxima or compared to a threshold. Peaks in outputs, or values above a given threshold, can then be interpreted as locations of features of interest. Alternatively, outputs can be processed using statistical modelling or classification methods, designed to permit more intelligent decisions to be made regarding image content.

The various automated methods noted in sections 1.3.2 make use of several different encoding schemes, which, along with other popular representations, will be described over the next few subsections. But first, some key properties of representations will be discussed as a point of reference.

2.1.1 Properties

The main aim of an image representation is to encode meaningful information about underlying structures, and to do so in a compact way. However, images contain noise, varying illumination conditions, and occlusion issues at boundaries, all of which complicate the extraction of useful information. Alternative schemes can be considered with regards to some key properties related to these issues:

- spatial scope, e.g. the extent of the area being described: local, regional or global;
- invariance to transformations, e.g. scale, rotation etc.;
- robustness and tolerance to noise;
- tolerance to boundaries;
- and completeness, i.e. how much information is retained or discarded.

Most of these are self-explanatory, but scope and completeness may require clarification.

Scope: Representations which describe groups of immediately neighbouring pixels will be considered to be local. Those which describe extended structures or whole objects will be considered to be regional representations. And those which encode entire images will be described as global representations.

Completeness: An image encoding capable of fully describing all salient structures within data is known as a complete representation. If a representation is complete then there should exist an inverse transformation, which can convert an image descriptor back into pixel data, whilst maintaining recognisable structures. A complete representation is less ambiguous than one which discards information.

2.1.2 Image encodings

Grey Level Co-Occurrence Matrices (GLCM): GLCMs were first used for describing the texture of terrains [56] in aerial images. They have also been used in crater detection [65] during a ‘focusing’ stage, where potential crater locations are narrowed down using information about the texture around crater rims. A GLCM is a probability mass function describing the distribution of pixel values which co-occur at a fixed offset from one another. Quantities such as contrast, homogeneity and entropy, are usually derived from GLCMs to give a condensed description of the data.

GLCMs have a local scope, with co-occurring pixels being selected within just a few pixels' distance from each other. They can be made tolerant to small illumination changes and noise by coarsely quantising pixel values, but otherwise have limited invariance, being designed for fixed scales, rotations and absolute pixel intensities. Also, the probability mass functions only describe pure textures, which do not model boundary or occlusion effects. GLCMs are partially complete, providing enough information to reconstruct the relative frequencies of local pixel value occurrences, but not necessarily in the correct spatial order, or over extended regions.

Law's texture filters: Law's texture filters [86] are a series of small convolution kernels designed to give a high response around basic features. Filters exist for describing structures including, for example, ridges, $\{-1, 0, +1\}$, dots, $\{0, 1, 0\}$, and ripples, $\{0, -1, 0, +1\}$. These basic feature detectors can be combined in 2 dimensions to give a range of textural patterns. A feature vector can be defined over these patterns, with each element giving the response to a different filter.

Law's filters have local scope and invariance to regional illumination levels, as pixels are considered relative to their neighbours. However, they are not invariant to scale or rotation. A sufficient number of filters can be combined to give a complete representation of small image patches, but like GLCMs, they can only describe fine image detail rather than extended patterns.

Scale Invariant Feature Transform (SIFT): SIFT has been used for describing rocks in martian images [87] and for experimental navigation of planetary rovers [88]. The SIFT representation [89] identifies structures within local image patches which are stable over many scales. A Difference of Gaussian (DoG) convolution kernel is used at a range of widths to extract edge information at multiple resolutions. The locations and resolutions at which edges are most detectable are extracted as SIFT key points. The use of only stable edge points also makes SIFT robust to small changes in rotation and affine transformations.

Once identified, the immediate pixel neighborhood surrounding each SIFT key point is analysed. Each key point is assigned an orientation, based upon the locally dominant edge direction. A vector of relative edge orientations is then constructed, with entries weighted by edge strengths. These vectors are placed into a database, forming a key point description of the data. To match features to new images key points are extracted then compared, using Euclidean distances, to find nearest neighbors in a reference database.

By limiting itself to strong edge features only, SIFT discards much image information, thereby limiting its completeness. SIFT data can reconstruct structures only up to the level of discrete points, which might be dense, but lacks information about actual pixel values.

Speeded Up Robust Features (SURF): SURF has been used in experimental planetary rover vision systems for mapping paths [90]. The SURF representation [91] was inspired by SIFT, but uses Haar transforms to find stable image points, rather than DoG filters. SURF is reported to be more stable than SIFT, providing greater invariance to small changes. SURF is a local representation, which like SIFT, lacks completeness.

Binary Robust Independent Elementary Features (BRIEF): BRIEF has been used for experimental planetary rover navigation systems [88]. The BRIEF representation [92] encodes images using a long binary string. Each bit of the string is associated with a different pair of pixels. Images are first smoothed, then a brightness comparison is made between end points of each pixel pair. For each pair, if pixel a_i is brighter than pixel b_i, then bit i is set, otherwise bit i is zeroed. The Hamming distance can then be computed between two strings to form a similarity measure. BRIEF is reported to outperform SIFT and SURF on feature matching tasks.

The scope of BRIEF can vary, but might be best described as regional. Whilst it is not invariant to scale or rotation, it is robust to changes in local illumination. And the quantity of pixel pairs provides redundancy which might be exploited to avoid boundaries and occlusion, but this has not been investigated. A sufficiently dense collection of BRIEF pairs can form a complete representation, as it has been shown that sets of pairwise comparisons can be used to reconstruct functions [93], at least up to the rank-order of point (i.e. pixel) values.

Histogram of Oriented Gradients (HOG): The HOG descriptor [94] is a histogram-like structure that records the orientation of edges within a small grid of pixels. HOG has been applied to the detection of dune fields on Mars [72]. Edge orientations are computed from raw pixel data using simple horizontal and vertical derivative masks, i.e. $\{-1, 0, +1\}$, before being combined to give an angle. This was first proposed as a method of describing pedestrians in video sequences.

HOG is a local representation which is invariant to varied illumination conditions and some affine transformations, as long as they roughly preserve edge orientations. And like most other representations, HOG does not take explicit account of boundaries. HOG is only partially complete, as the histogram of gradients discards precise pixel locations, recording only relative frequencies of orientations.

Hough transform, ellipse fittings: Geometric representations can be fitted to edges, derived from Canny or other edge detection methods, to describe basic shapes. These involve geometric equations being solved for particular perimeter points believed to have been generated by a given form. Circular and elliptical Hough transforms [95][96], and other fitting methods, have been applied to finding craters using the top-down knowledge that craters are roughly circular [68][69][71]. Variants on the basic Hough transform can take into consideration uncertainties in extracted edges to improve results [97].

These methods are regional, working on whole objects. Hough transforms and other shape fitting methods can be highly robust to missing data, including potential occlusion and boundary conditions. They are also invariant to many illumination changes, and can naturally scale and transform along their defined parameters, e.g. for setting different radii. However, they assume that points being fitted were generated by the shape being sought, making them less appropriate if other shapes are present, e.g. a ridge could easily be mistaken for a row of circles.

Templates: Scalable crater templates have been used in crater detection algorithms [71]. Templates resembling sought features can be used to find similar structures in images using a match score. Common matching methods include minimising the sum of squared residuals between template and image pixels, maximising the dot product between template and image vectors, and other variants.

Templates are regional representations, which can be made somewhat invariant to illumination changes via normalisation of pixel values, perhaps removing the mean regional intensities. Also, templates constructed from derivative images, rather than direct pixel intensities, can provide additional invariance to local illumination changes by discarding absolute intensity information [98][99]. Templates can be made invariant to transformations through brute-force methods, i.e. applying transformations to the template until a good match can be found. However, templates are especially prone to problems at boundaries, as missing data cannot be matched.

Appearance Models: An appearance model [100] is a flexible template computed from many example features, which can include pixel value and shape information. The modes of feature variability are learned, e.g. scale, elongation, illumination etc. by finding the eigenvectors around a feature's mean appearance using Principal Component Analysis (PCA). Similar templates have been used for crater detection [64][65].

Appearance models are regional representations that can be given enough free parameters to deal with many transformations, including transformations of pixel intensities and object shapes. However, the transformations have to be linearly interpolatable due to their PCA origins. Also, like their more ridged template counterparts, they suffer in cases of occlusion, as missing data cannot be matched.

Fourier Domain: The repeating structures found within textures have been described in the Fourier Domain, where whole images are encoded as sums of sine and cosine waves, rather than individual pixel values. A sum of such waves, combined at different frequencies, can completely describe image data. This is a global representation, where isolated changes in pixels affect the frequency components of all others. Fourier analysis has been proposed as an alternative to directly counting craters in digital terrain models [101], but has not been widely applied in planetary feature detection methods.

Wavelets, Gabor Filters, Haar transforms: Wavelets have been proposed as an alternative to Fourier analysis. Sums of single wave periods from an arbitrary wave form, rather than entire sines or cosines, can be used to describe textures [102]. The single period, or 'mother' wavelet, is rotated and shifted to describe different parts of an image. These can be applied as convolution kernels, similar to Law's texture filters. Gabor filters are sine and cosine wavelets which have been used for texture analysis [103]. The Haar transform is a square profile wavelet which has been used in crater detection [62][63]. Wavelets have also been applied to fault scarps on martian terrain maps [76].

Wavelets can act as local or regional representations, depending upon the application. They can also constitute a complete representation, as just like a Fourier domain description, a sum of wavelets can approximate any function.

2.1.3 Summary

The image representations described above have been included because of their applications to planetary data, their popularity, or their historic significance, but other representations are available. Any representation constitutes only part of a potential analysis system, as encoded data still requires analysis. Modelling and analysis techniques which might be applicable to such encoded data are described next.

2.2 Statistical modelling

It is often desirable to model the distribution of data using multi-variate decomposition techniques. These techniques attempt to describe data using mixtures of some base components, or vectors. Base components are sought that allow weighted combinations to approximate a range of data variability, thus providing a low parameter description of otherwise complex distributions. Linear and non-linear methods exist, the most popular of which are described below.

2.2.1 Data modelling methods

Principal Component Analysis (PCA): PCA [104] computes orthogonal vectors within data space which best accounts for linear variances. PCA is equivalent to forming least squared hyperplane fits to data, and as such errors on data points are assumed to be uniform and isotropic. PCA amounts to finding a projection of the original data onto an alternative coordinate system which yields a diagonal covariance matrix described by:

$$S = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(y_i - \bar{y})^T \tag{7}$$

$$D = W^T S W \tag{8}$$

where S is the original covariance matrix, y_i the data points and N the number of data points. Eigenvectors of the data's covariance matrix are computed, the eigenvalues of which reflect the amount of variance in those directions. These vectors, w_i, are known as principal components and are used to diagonalise S forming D. The least significant principal components can be discarded thereby leaving a lower dimensional representation of the original data. PCA can be an effective method of reducing dimensionality as long as only linear relationships exist between dimensions and there is sufficient variance remaining in the retained principal components.

Kernel PCA [105] is an extension to PCA which attempts to model non-linear manifolds by mapping data points into a new space via a non-linear function, ϕ:

$$S = \frac{1}{N}\sum_{i=1}^{N} \phi(y_i - \bar{y})\phi(y_i - \bar{y})^T \tag{9}$$

The kernel ϕ attempts to flatten out data so regular PCA can subsequently be applied. The kernel function itself contains an arbitrary number of degrees of freedom, the number of which will determine how accurately a manifold can be approximated. Too few degrees of freedom will lead to a rough approximation, whereas too many may cause over-fitting beyond the noise level and therefore poor generalisation to other datasets. The selection of kernel is therefore based on trial and error.

Whereas PCA only attempts to extract the axes occupied by data, Probabilistic PCA is an extension which attempts to describe, in a probabilistic manner, a lower dimensional representation, conditional upon its higher dimensional counterpart and a transformation matrix [106]. This assumes the data is Gaussian distributed on a linear manifold. The parameters of the appropriate transformation can be determined using maximum Likelihood on an iterative basis. This method assumes a high dimensional data vector, x_n, can be transformed into a lower dimensional approximation, y_n via a linear transformation, W:

$$y_n = Wx_n + \mu + \epsilon \tag{10}$$

where μ is the mean to which data points must be translated and ϵ is an isotropic normally distributed error with an expectation of zero. The transformation W is an l by m matrix where m is the dimensionality of the reduced data vector and l the dimensionality of the original data vector. The operating assumptions of PPCA are stronger than those for the similar Factor Analysis [107], as PPCA will only behave well when data has uniform isotropic noise, whereas Factor Analysis allows anisotropy.

Assuming data is first centred to give zero mean and data is normally distributed upon its manifold, PPCA gives the probability of observing a data point as

$$p(y_n|x_n, W, \sigma) \propto N(y_n|Wx_n, \sigma^2 I) \tag{11}$$

Integrating over x gives the marginal distribution

$$y_n \propto N(y_n|0, C) \tag{12}$$

where $C = WW^T + \sigma^2 I$. Assuming all data points are independent, the likelihood of observing a lower dimensional dataset Y conditioned on the transformation parameters can be calculate and maximised to determine the parameters W and σ:

$$p(Y|W, \sigma) = \prod_{n=1}^{N} p(y_n|W, \sigma) \tag{13}$$

$$ML = \arg\max_{W,\sigma} p(Y|W, \sigma) \tag{14}$$

Gaussian Process Latent Variable Models (GPLVMs: GPLVMs [108][109] generalise PPCA by replacing the linear transformation matrix W with an arbitrary function ϕ thereby combining the idea with that of Kernel PCA.

$$y_n = \phi(x_n) + \mu + \epsilon \tag{15}$$

Agreement between data and model is obtained by minimising the Kullback-Leibler divergence in the latent (lower dimensional) space.

Independent Component Analysis: ICA methods [110][111] attempt to provide a set of maximally independent components which can be combined, with a weighting matrix, to reconstruct multivariate data. Unlike PCA, ICA methods do not require extracted components to be orthogonal (e.g. as are eigenvectors), but only require that they are independent when measured by some objective function. Various ICA algorithms exist, differing in their objective functions and optimisation methods. A common application for ICA is Blind Source Separation [112], where multiple unknown independent sources are generating the data, such as multiple voices in an audio recording. Generally, ICA methods assume uniform Gaussian noise and continuous data.

ICA algorithms have been proposed based upon Likelihood [113], which determine the most probable set of base components, given training data. Other methods exist which employ the minimisation of Mutual Information measures [114], Negentropy, or maximising the kurtosis of extracted components[115]. Non-linear variants of ICA [116] can use the kernel methods of Kernel PCA to accommodate non-linear relationships.

2.2.2 Summary

The techniques described above are primarily designed to summarise or compactly describe data. They are not necessarily designed for making decisions or scientific measurements. A method of classification will be required for identifying features of interest as part of any measurement process. Common classification approaches are described in the next section.

2.3 Classifiers

This section will limit itself to a discussion of supervised classification methods, with the requirement that an automated planetary analysis system should be capable of learning arbitrary features through example. Popular classifiers will be described in terms of general properties and assumptions made about target data, rather than their finer implementation details.

Two broad categories of supervised classifiers will be investigated: fixed decision boundary methods; and data density methods. But first the theoretically optimal way to perform classification will be provided as a reference point.

2.3.1 A Bayes Optimal classifier

The job of a classifier is to assign a label, k, to a data point, X, indicating to which category it most likely belongs. The data point is often a vector of quantities, or 'feature vector', $X = \{x_1, x_2, \ldots, x_m\}$, defining a multi-dimensional pattern space. If there are n possible categories to which observation X may belong, the label, k, with the highest probability, $P(k|X)$, must be the optimal choice:

$$k = \arg_{k \in [1,2,\ldots,n]} \max P(k|X) \tag{16}$$

$$P(k|X) = \frac{P(X|k)P(k)}{P(X)} \tag{17}$$

where $P(k|X)$ is the probability that class k was the source of observation X; $P(X|k)$ is the probability of observing X within class k; $P(k)$ is the probability of observing a data point from class k, be it an X or any other value; and $P(X)$ is the probability of observing the value X, be it from class k or any other class, i.e. $P(X) = \sum_k^n P(X|k)P(k)$. $P(k)$ is conventionally referred to as the prior probability of k, yet this is not necessarily always known a-priori. Equation (17) is known as Bayes Theorem, and the label given by equation (16) is known as being Bayes Optimal. Equation (16) implicitly defines boundaries with the pattern space of X. In a two class case this boundary lies along the point where $P(1|X) = P(2|X)$, as seen in figure 1.

A Bayes Optimal classifier can be constructed in two ways. Firstly, the mapping between X and classes k can be learned directly, i.e. via the use of annotated data to sample from the distribution $P(k|X)$. Alternatively, the individual class conditional probability distributions, $P(X|k)$, can be learned then combined later, with the $P(k)$ terms estimated from elsewhere. In the first instance, learning $P(k|X)$ directly, there is

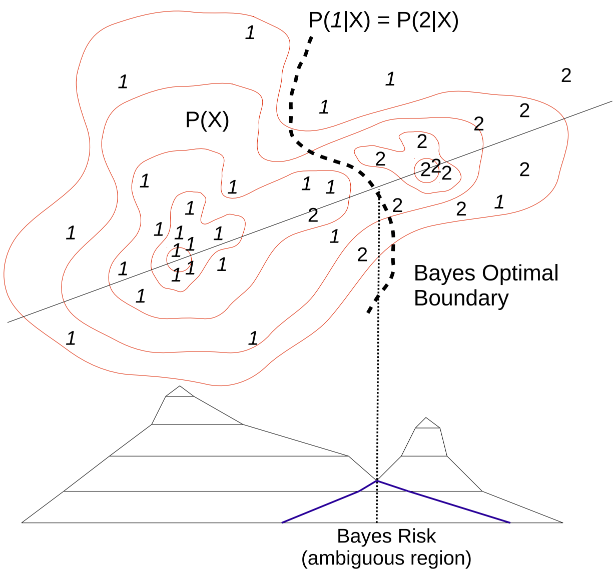


Figure 1: 2D two class data density with Bayes Optimal decision boundary.

a very strong assumption that future data will be statistically equivalent to the training data in every respect, i.e. $P(X|k)$, $P(X)$ and $P(k)$ will always be fixed. This assumption of stationary data distributions has important consequences and will be returned to in section 2.5.1. In the latter instance, learning $P(X|k)$, provides additional flexibility, only requiring that $P(X|k)$ be fixed.

The aim of fixed decision boundary methods is to learn the divide between classes along the line of Bayes optimality, i.e. the boundary following the points at which $P(k|X)$ is split equally between all possible categories. This falls into the category of learning $P(k|X)$ directly (or at least just near the boundaries). The aim of a data density method is to learn the entire probability distribution of the data then apply Bayes Theorem directly. This falls into the category of learning individual class distributions, $P(X|k)$, then estimating priors from elsewhere.

2.3.2 Fixed decision boundary methods

Many popular classifiers make use of training data to approximate a Bayes optimal boundary between categories. Boundary methods have a significant advantage over density methods in terms of reduced training data requirements. Approximate boundaries can be learned using fewer data points than is needed to learn entire probability densities.

Fisher's Discriminant: Fisher's Discriminant provides a method to linearly separate classes of data using hyperplanes. The technique assumes that data from each class follows a Gaussian distribution, typically in multiple dimensions, with means and covariances which can be sampled from training data. These Gaussian distributions are equivalent to the $P(X|k)$ terms in Bayes Theorem. The method provides a closed-form solution to finding an approximately optimal linear boundary between two classes, assuming the prior terms, $P(k)$, remain fixed. Similar classifiers have been utilised in automated terrain classification [56].

Linear Discriminant Analysis: Linear Discriminant Analysis (LDA) is a simplification of Fisher's Discriminant. It too provides a closed-form solution to an approximately optimal boundary. However, the simplification comes at the price of additional assumptions. LDA assumes that the Gaussian distributed classes are also parameterised using the same covariance matrix.

Support Vector Machine: Support Vector Machines (SVM) [117], like Fisher's Discriminant and LDA, attempt to find hyperplanes which can linearly separate data classes. However, unlike Fisher's Discriminant and LDA, they do not assume Gaussian distributed data. Plus, in cases where data is not easily separable, SVMs can transform the pattern space, using non-linear functions, with the aim of maximising separability in an alternative domain. The added complexity means that an iterative, rather than closed-form solution is required. SVMs have been applied to terrain analysis [58][59], crater detection [61] and dune field identification [72].

Decision Trees: A Decision Tree decomposes data points, X, into constituent elements, x_i, then makes a series of hierarchical classification decisions based upon each element. At each decision, depending upon the value of x_i, a Decision Tree will follow a particular branch within the hierarchy until a leaf node is found. Training data is used to construct trees with the most probable class labels being assigned to each leaf. This is equivalent to constructing a large number of local classifiers, which together approximate a more complex decision boundary. These trees make few assumptions regarding the distribution of data, except that the distribution is smooth.

A Random Forest (RF) [118] is a large collection of alternative Decision Trees. These are constructed from random subsets of the data, using different subsets of feature vector elements. The classification assigned to a data point fed into an RF is then taken to be the one which most trees agree with, i.e. a vote. Decision Tree methods have been applied to terrain classification [59] and crater detection [67].

Boosting: Boosting [119] is a method designed to fuse large numbers of 'weak' classifiers together, thereby providing higher accuracy classifications via linearly weighting results. Several alternative boosting algorithms have been proposed which take classifiers with very low success rates, with possibly only very weak correlations between outputs and correct classifications. As each new weak classifier is added, the algorithms re-evaluate the current weightings to focus on the more difficult cases, increasing their weights whilst reducing the weights of cases which have been largely solved. This results in a piece-wise linear decision boundary. The performance of Boosted classifiers has been investigated to give bounds on error rates via numerical and theoretical approaches [120].

Boosting is often used to combine the outputs from many classifiers to build a weighted voting system. Boosting has been applied to crater detection [62][63].

Neural Networks: Neural networks can be trained to learn a mapping between data points, X, and labels, k, using a layered architecture of artificial neurons [121]. Elements of a data point are used as input to the first layer, and outputs of the last layer can be a set of binary outputs, one per class. A neuron at layer i takes a weighted sum of outputs from neurons at layer $i-1$. The result is then fed through a non-linear 'activation function' (usually sigmoid) before feeding into the next layer. A back-propagation algorithm, based upon a downward derivative search, can be used to set the weighting coefficients between neurons to approximate any smooth function. Neural networks have been applied to terrain analysis [59][60] and crater detection [65][66].

2.3.3 Density methods

Density methods attempt to model the total probability density of a pattern space, with the aim of estimating probabilities of classification labels. For a given data point, the label with the highest probability of being correct can be assigned. Alternatively, the probabilities can be manipulated directly, e.g. for weighted summing of quantities. In general, density methods directly implement Bayes Theorem, with terms approximated using parametric models.

Naïve Bayes: A Naïve Bayes classifier simplifies the estimation of class conditional probabilities, $P(X|k)$, by assuming each element of the feature vector, X, is an independent variable. Making this assumption allows individual elements to be modelled separately, without need to consider potential correlations between them:

$$P(k|X) = \frac{\prod_i P(x_i|k)P(k)}{P(X)} \tag{18}$$

The individual terms, $P(x_i|k)$, might then be sampled directly from training data, or approximated using an appropriate parametric form, e.g. Gaussian distributions. The prior terms are then often selected subjectively.

Naïve Bayes classifiers are often used as a base-line for comparisons, having been superseded by empirically more successful fixed boundary methods. Examples of this can be seen in [122] and [123], where Random Forests and SVMs are shown to outperform Naïve Bayes for astronomical applications involving the classification of galaxies.

Subjective Bayes: The origins of prior probabilities, $P(k)$, are a matter of debate, with the Subjectivist school of thought asserting that probabilities cannot be objectively defined, and thus can only represent 'degrees of belief' [124]. In contrast, this thesis adopts a Frequentist definition of probabilities, and therefore takes the view that a prior probability should be proportional to the frequency with which an event of a particular class occurs.

The Subjectivist view on priors leads to arbitrary prior selections that need not be linked to physically meaningful quantities. A Bayesian classifier may be constructed using such subjective priors, which can lead to mathematically consistent solutions by not violating Kolmogorov's axiomatic definition of probabilities [125], but does not relate to physical measurements.

Gaussian Mixture Models (GMM) : Gaussian Mixture Models are used to describe data distributions which either contain linearly combined Gaussian classes, or can be closely approximated by them. They are commonly used in remote sensing [57] and medical imaging [80] applications.

GMMs are usually fitted to data via the iterative Expectation Maximisation (EM) [126] algorithm, which utilises Bayes Theorem as an update function [127]. This is a two step process starting with approximate initial estimates of Gaussian parameters, μ_k, and σ_k. Likelihood estimates of these parameters are computed by iterating the steps of calculating the expectation of the data, given the parameters, then maximising the expectation by updating the parameters.

In the first step, the normalisation (i.e. integral) of the Gaussians is estimated by summing posteriori probabilities, $\mathbf{Q}_k = \sum_X P(k|i)$, for Gaussian classes k, given the data. The relative normalisations from each class provide new estimates for the prior probabilities, $P(k) = \frac{\mathbf{Q}_k}{\sum \mathbf{Q}}$. These feed in to the GMM to give the expected model, i.e. equation (1).

In the second step, the posteriori probabilities are used to update the Gaussian parameters by computing weighted means and sigmas, e.g. $\mu'_k = \frac{1}{\mathbf{Q}_k} \sum_i P(k|i)\mathbf{H}_i$. Where the standard formula for the sample mean of observations $\mathbf{H}$ is weighted with $P(k|i)$. And similarity, for σ'_k, the standard deviation is computed, with weighted observations.

Upon convergence, Bayes Theorem can be used to construct a classifier to label data points according to their posteriori probabilities of belonging to the relevant Gaussian class.

2.3.4 Summary

The outputs of classifiers, either boundary-based or density-based, might be used for making simple measurements, such as counting features. To use any resulting measurements with confidence it is necessary to understand the performance of the chosen classifier. Non-trivial analyses will be affected by sources of uncertainty leading to perturbations in outputs. These perturbations can be investigated theoretically and empirically, as described next.

2.4 Performance evaluation

Following the conventional modular development paradigm, once a candidate representation (or representations) have been connected to a candidate classifier (or classifiers), the performance of the overall system must be quantified. Unless a system is perfect then there will be classification errors, which will lead to errors in measurements. The importance of these errors was noted in section 1.4.1. A range of empirical tools exist for evaluating these effects, allowing competing systems to be directly compared to each other. For these tools to be applied, a sufficient quantity of ground truth data is required, e.g. images which have been annotated to indicate 'correct' classifications. Only with knowledge of the expected true answers can mistakes in classifications be identified.

Empirical approaches usually begin with simple statistics, such as error rates. Once gathered, these error rates can be combined over different system parameter settings, thereby giving a more comprehensive performance assessment. Finally, such tests can be performed multiple time for different cuts of the ground-truth data to test the repeatability of a system's performance. These steps will be discussed in more detail in the following section.

Theoretical approaches to performance can also be applied to some systems, assuming they are formulated in an appropriate way. These have already been noted in section 1.4.2, where Likelihood estimators can be examined using the CRB (lower variance bound), and differentiable functions can be examined using error propagation. The application of these theoretical approaches to performance evaluations will also be discussed after the empirical methods.

The choice between empirical verses theoretical approaches can be reduced to knowledge (or lack of knowledge) of algorithms and data properties, or 'black-box' verse 'white-box'. The most appropriate style of evaluation is a matter of debate [150].

2.4.1 Empirical performance evaluations

Empirical performance evaluations are often necessary, as modular building blocks are used as 'black boxes', with few or no assumptions made about their inner workings [129]. The intentional lack of knowledge presumed about components makes predicting performance difficult, or impossible. However, given one key assumption, that training and testing data is strongly representative of real world data, allows (in theory) the performance of a system to be understood.

Simple performance statistics: Most classification systems are evaluated by counting the number of times they assign correct verses incorrect class labels to data points. This is most easily achieved if classifiers are reduced to binary choices, which for multi-class systems can be translated to 'this class or another'. For a two class case, there are four possible classification outcomes:

- true positive (tp) classifications describe cases where a ‘positive’ data point is correctly identified as belonging to the ‘positive’ class;
- true negative (tn) classifications describe cases where a ‘negative’ data point is correctly identified as belonging to the ‘negative’ class;
- false positive (fp) classifications describe cases where a ‘negative’ data point is incorrectly identified as belonging to the ‘positive’ class;
- and false negative (fn) classifications describe cases where a ‘positive’ data point is incorrectly identified as belonging to the ‘negative’ class.

The positive and negative classes have been placed in quotation marks, as these can arbitrarily be switched, depending upon which class is to be defined as which. Viewed this way, there are only two cases, giving a True Acceptance Rate (TAR) and False Acceptance Rate (FAR) [130].

Given a ground-truth dataset, half the data might be used for training and the other half used for evaluation. The raw values of tp, tn, fp and fn can then be counted, giving an account of the system’s performance over a set of experiments. Often this is as far as evaluations go, including those conducted for the noted automated methods of section 1.3.2.

Confusion matrix: A confusion matrix [131] can be used to tabulate the simple statistics given above using a 2 by 2 grid. The rows can indicate predicted class labels, whereas the columns can indicate actual true class labels. Each cell is then populated with the corresponding statistic:

		Actual class	
		+	-
Predicted	+	tp	fp
	-	fn	tn

In multi-class cases, a confusion matrix can be extended to record the relationships between all class labelling error combinations:

		Actual class		
		k=1	k=2	k=3
Predicted	k=1	t1	f1	f1
	k=2	f2	t2	f2
	k=3	f3	f3	t3

Other useful values can be derived from these matrices, including ‘sensitivity’, which is the fraction of true positives found within the positive column, and ‘specificity’, which is the fraction of true negatives in the negative column:

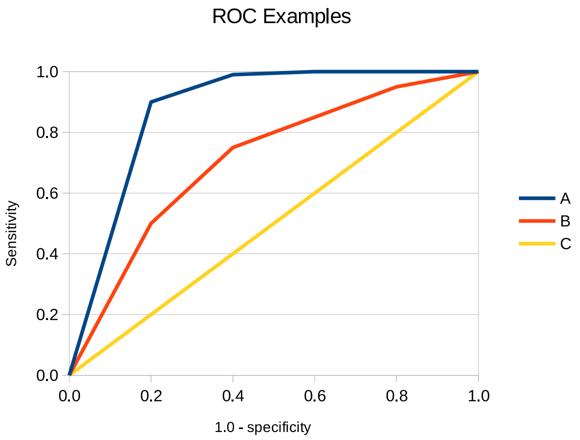


Figure 2: Example ROC curves: Algorithm A performs best, giving largest area under curve, where as algorithm C performs no better than chance.

$$sensitivity = \frac{tp}{tp + fn} \tag{19}$$

$$specificity = \frac{tn}{tn + fp} \tag{20}$$

The above values can be used to profile the performance of a system as various parameters are adjusted, as will be described next.

Receiver Operating Characteristic (ROC) curves: Most classifiers and image representations have tuning parameters which will affect the performance of the overall system. The effects of such parameters can be plotted using a Receiver Operating Characteristic (ROC) curve [132][133]. Such curves plot sensitivity against specificity (or 1 minus specificity), as the value of a tuning parameter is scanned. An example ROC curve can be seen in figure 2. If an ROC curve forms a perfect diagonal, then the system it describes performs no better than chance. The area beneath an ROC curve will be larger for better performing systems. Beyond characterising the overall performance of a system, an ROC curve can be used to determine optimal tuning parameter settings [134]. ROC curves constructed for different algorithms can be compared directly, with the aid of standard datasets.

Cross validation: The performance evaluations described above assume a finite quantity of ground-truth data which has been divided into a single training and a single testing set. However, variability within the data will make any such evaluations just one outcome from a distribution of possible performances. Cross validation can be used to investigate this distribution by performing repeated evaluations upon different cuts, or folds, of the finite ground-truth [135][136]. For example, the total data can be divided into 6 equal parts, with 6 different performance evaluations being conducted for each combination of

5-parts training to 1-part testing. In the extreme, this can be conducted as a leave-one-out evaluation where all data, except a single data point, is used for training, then the remaining data point is used for testing. The benefits of cross validation is the ability to give a range of performance indicators, hopefully placing realistic bounds on performance in the presence real-world data variability.

Bootstrap re-sampling: Bootstrap re-sampling involves randomly selecting data points, with replacement, to simulate larger quantities of testing data [137][138]. This assumes each data point is independent and identically distributed (i.i.d.). If used in combination with ground-truth, bootstrap re-sampling resembles a rigorous form of cross validation, allowing performance to be evaluated using far more data than is actually available. If used without ground-truth, re-sampling can still be used to assess the statistical repeatability of results, but not necessarily potential systematic effects.

If a pattern recognition system is considered as a statistical estimator, the values estimated will have properties such as variance and possible biases. These properties come from the fact that a given dataset is only a sample of a wider population, and if an independent, yet equivalent, sample of data was analysed the results would vary. The properties of estimators describe the relationship between sample and population, where the population itself is generally **unknown**. In bootstrap re-sampling, the re-sample becomes the target data, and the original sample becomes the population, which **is known**. In bootstrap re-sampling, properties of the original sample, via inspection of re-sampled results, are extrapolated to infer properties of the population.

Monte-Carlo: Monte-Carlo methods [139] can be used to provide arbitrary quantities of training and testing data, with objective ground-truth. This is achieved by artificially generating data, using an assumed statistical model, believed to be representative of real-world data. For a Monte-Carlo dataset to be useful, the distributions, correlations and perturbations simulated must closely resemble those found in reality. This requires a very good understanding of the data, beyond what is required in cross validation or bootstrap re-sampling.

2.4.2 Theoretical performance evaluations

Theoretical methods require a good understanding of the inner-workings of an algorithm, i.e. they are white-box methods. Unlike empirical alternatives, theoretical approaches can make predictions, which can then be tested empirically in line with The Scientific Method. Advocates of a theoretical approach have analysed numerous low-level algorithms using statistical perturbation models [140][141]. It has been suggested that this approach can both characterise performance and also be used for parameter tuning [142]. Statistical methods for analysing performance have been described already in section 1.4.2. Here, only examples of their use in selected applications will be given.

Cramer Rao Bound (CRB) / Lower variance bound: The CRB has been applied to estimate the accuracy of edge and corner detectors in 2D data [143], and also 3D points extracted from stereo data [144]. It has also been used to estimate the variance on weighting matrices for ICA [145][146][147].

Error propagation: The use of error propagation in computer vision [148][149][150] has been demonstrated on tasks such as the extraction of 2D points, measuring distances between them, and determining the accuracy of parameters of fitted shapes [151]. It has also been applied to assess the performance of multi-stage 3D shape extraction from 2D projections and motion [152], and in the use of linear shape models [153]. A comprehensive example of theoretical error analysis can also be found in [154] where propagated location uncertainty assists in target recognition.

With reference to some of the techniques listed earlier, error propagation has been used to investigate the effects of noise in Hough transforms [155] and also iterative algorithms [156], such as EM.

2.4.3 Summary

Performance characterisation techniques can be, and have been, applied to various combinations of image encoding, modelling and classification modules. The success or failure of module combinations and any performance measures extracted from them must be framed in the context of making scientific measurements. The following sections will critically address this key issue.

2.5 Applicability to quantitative measurements

A system appropriate for quantitative use must be capable of producing measurements fulfilling the criteria of section 1.6. As an example measurement, this section will consider the common process of counting the number of data points associated with a specific class. The next chapter will argue that this process can be linked to many measurements sought from planetary images.

Counting might be achieved by summing data points which have been assigned a specific class label, with the hope that false positive and false negative results will roughly cancel out. This label counting method is a common approach in counting applications, e.g. [159][160][161][162]. Alternatively, counting could be achieved by summing over all data points, weighting each by the probability of a specific class, with the hope that the weighting will account for labeling uncertainty. This approach can be found within GMMs [80] and other data density approaches to counting [167].

The building blocks and evaluation methods presented above will be critically examined below. The quantitative criteria of section 1.6 will each be applied in turn, with inappropriate components being rejected.

2.5.1 Criterion: a measurement must be a numerical estimate driven by evidence found within the data

Representations: Image evidence needs to be represented, yet there are no clear right or wrong methods for encoding such information. Different representations encode more or less evidence, provide more or less invariance to transformations, and are more or less robust to the effects of noise and clutter. All of the representations listed in section 2.1, or closely related variants, have been used to encode planetary image data and are therefore potential candidates. What is required, from a quantitative counting perspective, is an encoding which provides easy translation between image descriptors and physical counts or surface areas. Also, the statistical properties of the encoding must be understandable, so they can be matched to the assumptions made by any modelling and classification methods applied to them. Ideally, a complete representation which is robust to noise and illumination changes is desirable. For now, no decision will be made regarding representations. Modelling considerations in chapters 3, 4 and 5 will better inform representation selection for use in part 2 of this thesis.

Modelling: Statistical modelling methods, such as those noted in section 2.2, tend to be designed to summarise data, or to transform data into a lower dimensional form with particular properties, e.g. decorrelated. Model parameters might be correlate with physical measurements, such as the principal components of a PCA model of crater being linked to size, elongation etc. However, their use is subject to assumptions, as already explained. If data contains linear relationships and uniform isotropic noise, then basic PCA and ICA methods might be appropriate. The use of more complex data and non-linear modelling methods should only be approached if their is evidence that the modelling method matches the properties of the target data.

Label counting using decision boundaries: All fixed decision boundary methods listed in section 2.3 can be eliminated from consideration immediately via the first quantitative criterion. This is due to the strict representativeness assumptions required for the optimal positioning of decision boundaries. A boundary may be optimal for the evidence found within training data, i.e. for the relative quantities of classes within the training set, yet be suboptimal when applied to new datasets containing different quantities of classes [168]. This can be seen in figure 3, illustrating that a change in relative quantities, i.e. prior probabilities, must logically lead to biases which are a function of incoming data. Such biases will change the ratio of false positive and false negative results, and this bias will be a function of the (unknown) quantities of classes which are being counted. For the criterion to be met, decision boundaries must be capable of shifting to new optimal locations on a dataset by dataset basis, driven by evidence in the data being measured.

Fixed boundary methods might be defended with arguments that they should only be applied with representative data, e.g. the assumption of fixed $P(X|k)$ and fixed $P(k)$

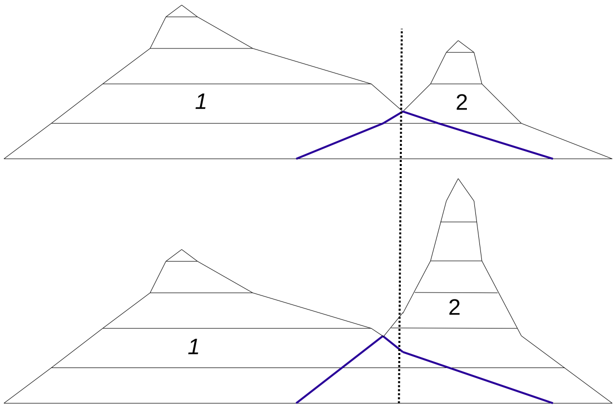


Figure 3: Top: original decision boundary using training data of figure 1. Bottom: application of trained boundary when class 2 has increased in quantity, i.e. $P(X|2)$ is fixed, but $P(2)$ has increased. Here, the boundary is no longer optimal, i.e. at the point where $P(1|X) = P(2|X)$.

must not be violated. However, these methods are commonly (and thus mistakenly) used for counting applications, e.g. [159][160][161][162]. In such applications, the problem domain assumes that quantities of classes are unknown and varying from dataset to dataset, otherwise they would not need counting. Yet, this assumption is directly opposed to assumptions of representativeness made by fixed decision boundaries, and therefore estimated quantities cannot be considered reliable.

Probabilistically weighted counting: Posteriori probabilities, $P(k|X)$, have been estimated for SVMs [163][164] using auxiliary data structures. These include histograms (binning), kernels and sigmoid functions, which attempt to map the distances between data points and the decision boundary to probability estimates. Similar work has been performed for decision trees [165] and boosting [166], where generating honest probabilities requires calibration against empirical training data. Whilst these probabilities may be of value when representative data is analysed, they still suffer the problems noted above. Calibrated probabilities are based upon evidence in training data, not evidence found in data under analysis, which, as argued above, is unrepresentative for counting applications. Using probabilities from fixed boundary methods must also be considered non-quantitative.

Some density methods can be eliminated via this first quantitative criterion also. Naïve Bayes classifiers, which do not take into account correlations in incoming data, cannot be driven by evidence unless elements within a feature vector are genuinely independent. Subjectivist Bayesian classifiers cannot be driven by evidence either, as they do not require that prior probabilities are linked (in a Frequentist sense) to the underlying data. The use of subjective probabilities in science has been rejected by philosophers of science [77]. These methods must also be considered non-quantitative and therefore eliminated from consideration.

GMMs may be of value for quantitative measurements, assuming data classes are

Gaussian distributed. They dynamically adjust to provide the best fit to incoming data, objectively setting prior probabilities to match the proportions of classes under analysis. GMMs have been shown to outperform some alternative classifiers in the segmentation of infrared imagery[133]. These methods might allow quantitative counting using multi-spectral remote sensed data, but their Gaussian assumptions will limit applicability to visual spectrum images containing texture and extended features.

2.5.2 Criterion: a measurement must be accompanied by an error estimate indicating the expected accuracy of the measurement; and Criterion: a measurement must be shown in practice to not deviate from the true measurement by more than is predicted by the estimated error

Representations: Measurement errors begin with noise in data. Whilst the selection of a representation is being deferred for now, there are general notes which should be considered. A useful representation must have predictable error characteristics, otherwise no meaningful error estimates could be produced. Ideally, to simplify the analysis of errors, data points presented using a given encoding would have independent errors which are either uniform, or at least a simple function of the data. Preprocessing techniques, such as smoothing, introduce local correlations in pixel noise. Contrast enhancing techniques, such as histogram equalisation, can introduce non-uniform errors. Such issues must factor into any error analysis if quantitative criteria are to be fulfilled.

Modelling: Modelling methods which can be linked to Likelihood, e.g. PCA and Likelihood-based ICA, are amenable to theoretical error assessments, assuming the model is appropriate for the data. However, more complex modelling methods, involving kernels and non-linear transformations, are less appropriate in the presence of noise. For example, kernel methods, which apply non-linear transformations to data also apply the same non-linear transformations to noise. The resulting noise is no longer uniform or isotropic, and therefore violates the assumptions made by linear modelling methods. Error propagation could be used to track the sources of uncertainty in non-linear cases, however, it would be preferable to have a simple error model initially, only increasing the complexity if it was demanded by the application.

Classifiers: None of the classifiers listed in section 2.3 contain explicit support for the production of measurement error bars and are therefore not immediately quantitative. Nor has work been undertaken to add this functionality in any of the automated methods listed in section 1.3.2. However, GMMs, which are Likelihood fits to data, are amenable to statistical analysis, as was described in section 1.4.2. A Likelihood system based upon more flexible mixture models would also have this property. Other methods may be forced to rely upon empirical evaluations.

Empirical performance evaluations: The error rates and ROC curves described in section 2.4 are not the same as error bars on measured quantities. Nor are empirical error rates guaranteed predictors of future errors, for the same representativeness arguments apply here as they did to fixed boundary classifiers: the error rates observed during testing are specific to the distribution of the testing set, which is not necessarily equivalent to future data.

In defence of the conventional approach, a mapping might be learned between error rates and error bars, and these error bars could, in principle, be calibrated using (very) large quantities of Monte-Carlo or verified testing data. However, this work does not appear to be undertaken in counting applications, e.g. [159][160][161][162], and has not been undertaken in methods listed in section 1.3.2.

Bootstrap methods might be employed using re-sampling of data under analysis to estimate the statistical spread of measurements for incoming data, but without ground-truth such an approach would not reveal any potential systematic effects. Also, the i.i.d. assumptions of re-sampling may be invalid if there are unknown correlations between data points. Again, if training and future data (simulated, bootstrapped, or otherwise) are not representative then unknown systematic biases may be introduced making this approach unworkable.

Monte-Carlo methods too suffer from issues of representativeness, as simulated data must mimic real-world data in sufficient detail to be of value. However, if data is well enough understood then Monte-Carlo can be a valuable assessment tool, especially for testing theoretical results. For example, the theoretical performance of Canny edge detection [157] isn't realised in Monte-Carlo [158].

Theoretical performance evaluations: Statistical tools, such as CRB and error propagation (section 1.4.2) can provide confidence intervals on estimated values, assuming the model of the data and its perturbations are correct. This is one reason why an ideal image encoding would have the desired properties given above in section 2.5.1.

Theoretical approaches to determining errors have the advantage of revealing the functional forms of measurement uncertainties. An empirical evaluation of a system may reveal **how** a system behaves, but a theoretical evaluation can reveal **why** it behaves as it does. Unfortunately, systems lacking Likelihood origins or differentiable forms, such as most of the classifiers listed above, are not immediately amenable to this approach.

2.5.3 Criterion: a measurement, where possible, should be supported by additional checks to ensure trustworthiness

Representativeness: It has been noted several times that many methods will fail, or be suboptimal, if there are discrepancies between training data, testing data and real-world data. This can affect trained classifiers, statistical modelling methods, and empirical and theoretical evaluation methods. These problems can be summarised as the problem of representativeness. A hugely useful tool would be one which corroborated representativeness, allowing unrepresentative data to be flagged to users as untrustworthy. This must also be achievable when no ground-truth is available, as this is the case with real-world data.

Representativeness will be defined here as a statistical equivalence. If two datasets are statistically equivalent then the parameters of their distributions should be the same, to within the limits of the expected errors. Any correlations within the data should also be the same, within the limits of expected covariances. Tests for representativeness must therefore have access to such knowledge, instantly excluding methods which do no record such information, as described next.

Classifiers: There is insufficient information at a decision boundary to test if data points are, on average, too close or far away from a boundary to be considered representative. There is more information in density-based methods, but in practice representativeness is not confirmed as a standard part of any method presented in this review. Published work tends to, assume, rather than corroborate distributions and errors, and therefore cannot be considered quantitative by the criteria put forth by this thesis.

Ad-hoc methods have been applied to test the conformity of some models, but with no predictions as to what the conformity scores should equal when data is representative. This is in contrast to chi-squared per degree of freedom methods, which are expected to give a value of unity, or other conventional approaches such as Fisher-Exact, which have precise predictions of how data should appear.

2.6 Summary

Upon critically examining common building blocks, a modular plug'n'play approach to the development of a quantitative system appears unlikely to succeed, i.e. it is not sufficient to arbitrarily select and apply a standard input representation, a standard classifier, and then evaluate using a standard ROC curve. This is due to many available modules and methods failing quantitative criteria in one or more ways. A black-box approach is therefore rejected.

However, there are potential avenues for investigation which may lead to quantitative results. In particular, GMMs estimated using EM provide Likelihood solutions, amenable to goodness-of-fits and error analyses, as has already been explained in section 1.4.2. Unfortunately, they are limited by their distribution assumptions. Yet an appropriate ICA model might allow non-Gaussian distributions to be extracted from more complex data, which might be used as substitutes for Gaussians in a GMM-style approach. Furthermore, the construction of appropriate Monte-Carlo simulations and bootstrapped samples might be used for corroborating quantitative methods, including predicted probabilities and error estimates, if the representativeness of data can be assured. Incorporating these ideas into a quantitative system will require a white-box mentality, where the statistical properties of data and inner-workings of algorithms must be understood for a cohesive system to be engineered[169].

The next chapter will consider the requirements of an automated planetary analysis system in greater detail. It will expand upon the potential avenues noted above, leading to a flexible system which will be comprehensively tested in part 2.

3 A Statistical Model for Planetary Image Data

The texture of terrains or the structure of surface features found within planetary images may be described as sets of repeating elementary patterns. Representations for describing such patterns were discussed in chapter 2, such as Law's texture filters and Haralick's grey level co-occurrence matrices. The quantitative measurement of terrain surface areas, or the counting of individual features, requires an understanding of the statistical distributions of these patterns. It will be argued that taking measurements mainly involves the counting of patterns associated with classes of interest, whilst accounting for correlations, variations and perturbation processes which introduce uncertainty. This chapter will describe a statistical model for capturing the distribution of these patterns parameterised in terms of the relative abundance of correlated groups. A Likelihood parameter estimation method will be provided for fitting the model to training and testing data, and a goodness-of-fit function will be presented for checking the appropriateness of the model when applied to possibly contaminated data. Monte-Carlo simulations will be used to test the resulting model fitting algorithm and the goodness-of-fit function over a range of different distributions and quantities of data.

3.1 Properties of planetary images

Any modelling decisions made must facilitate the quantitative analysis of the data consistent with criteria prescribed in section 1.6. This must include an understanding of what needs to be measured within an image and what uncertainties may exist. An understanding of the properties of planetary images is required before a statistical model can be selected. Consideration must be given to the types of surface features that may occur, how stable those features are, how distinctive they might be and what types of measurements might be taken from them. Interactions between different features must be considered, including how the composition of a terrain may change from place to place, how features may overlap and what types of boundaries may form between them.

To better understand the requirements of a statistical model for planetary terrains a list of example surface features will be examined. Although not a complete list, those features included will span a wide cross-section of properties which a statistical model must cater for. This list can be found in table 1 and images of the features listed can be seen in figure 4.

3.1.1 Features and their measurements

Most of the measurements listed in the final column of table 1 can be linked to the counting of pixels. For example, the size (radius) of a crater is linked, via πr^2, to the number of pixels which make up a crater's area, which can be counted. The lengths of fissures, drainage channels and dune crests are proportional to the number of pixels along their

Feature	Description	Variability	Interactions	Measurements
Craters	Distinctive circular structures associated with impact events.	Very large range of sizes, with different morphologies between smallest and largest craters. Change with age as erosive processes smooth appearance. Some rare craters are highly elliptical.	Can be nested or overlapping, with new craters sharply destroying parts of old craters. Destroy older local features.	Size Frequency Distributions (SFD): counts of craters falling into different size bands.
Dunes	Mounds of loosely bound grains usually formed by wind.	Highly variable shapes, including stars, domes and crescents. Have different orientations, linked with wind direction. Appear in a range of sizes.	Can smoothly merge into one another forming larger dunes, chains and linear features. Can engulf other features.	Wavelength, frequency and density of dunes. Orientation. Surface area.
Fissures	Generic term for faults, cracks or fractures caused by the deformation of a surface. Often linear.	Range of lengths, widths and orientations.	Can merge to form larger fissures. Can be branching.	Length and density. Orientation. Surface area affected.
Drainage networks	Coalescing channels formed typically by flowing water in a down hill direction.	Range of lengths and widths. Can contain complex dendritic patterns or simple linear channels. Channels range in curvature.	Can merge together hierarchically into large extended structures. Can cut through older features.	Length and density. Orientation. Surface area of drainage basin.
Martian ‘spiders’	Radial fissures growing from gas eruptions.	Variable size and shape, with greater or fewer numbers of radiating branches.	Can merge together into complex networks of fissures.	Counts of eruptions. Sizes and surface areas affected.
Martian chaos terrain	Irregular groupings of large flat topped blocks divided by depressions and valleys.	Highly variable extended regions containing different sized mesas, buttes and collapsed areas.	Complex matrix of intersecting boundaries.	Surface area.

Table 1: Example features found within planetary images

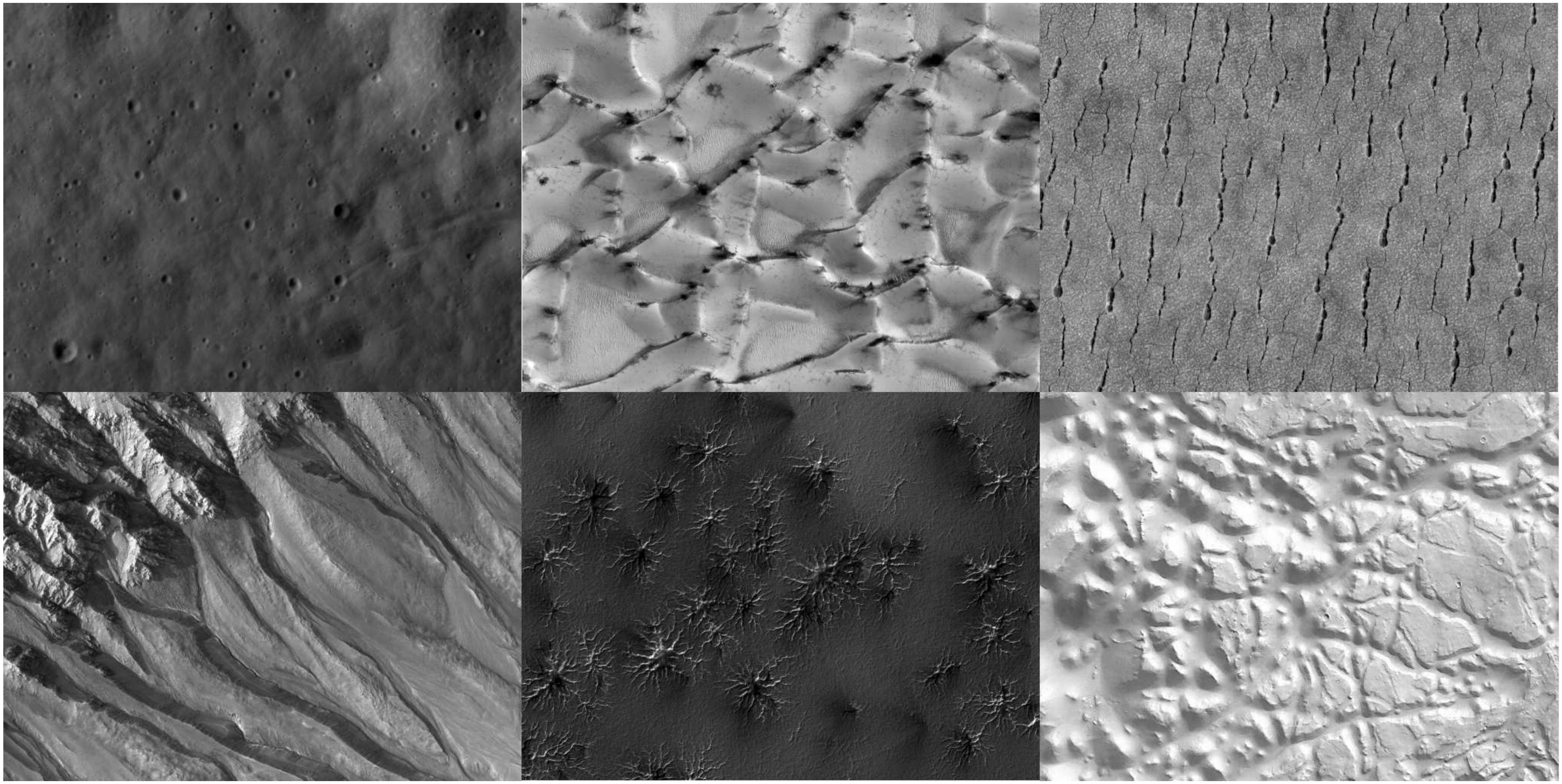


Figure 4: Features found within planetary terrain images. Top row, from left to right: Craters; Dunes; Fissures. Bottom row, from left to right: Drainage network; Martian ‘spiders’; Martian chaos terrain. Images courtesy of NASA, Lunar Reconnaissance Orbiter and Mars Reconnaissance Orbiter.

edges, which again can be counted. Also, measurements of densities and frequencies are just counts divided by areas or lengths. As long as there is a mechanism allowing pixels to be attributed to particular features then most measurement problems can be solved by estimating quantities of related pixels.

There is little information contained within individual pixels so contextual information must be used to determine which pixels belong to which type of feature. Ridged or deformable templates, or edge strings may be used to describe whole features. However, the wide ranging sizes and shapes of features noted in table 1 suggest that an essentially infinite number of variants may exist, possibly making a wholesale approach to feature identification impractical. Alternatively, features can be viewed as collections of smaller fixed sized patterns appearing in correlated groups. For example, an extended feature like a drainage network may be viewed as a chain of small, often repeating, patterns forming the various channels and tributaries. The structure of an entire terrain might be describable in terms of such correlated groups, where each elementary part can occur multiple times and each part can have a finite set of variants. Features might then be described as distributions of patterns with indicative probability functions that could be learned. This view may also provide a solution to the measurement of orientation by assigning different orientations to different pattern distributions.

Quantitative measurements require estimates of errors. These errors must be affected by perturbations in distributions and potential ambiguities between similar patterns in different groups. The occurrence of features can be seen as random events. Using examples from table 1, in the case of craters these are random impact events, or in the case of fissures these are random breakdowns of weaknesses in surfaces under tension. It may be reasonable to assume that these occur with frequencies following a Poisson distribution. The appearance of individual patterns within features might also be attributed to Poisson processes, such as small random collapses at the edges of mesas or crater rims. The modelling of such perturbations must feed into the statistical description of the data and their effects upon measurements must be quantifiable.

Another source of uncertainty might be found in patterns that occur in more than one group. A distribution describing a drainage channel feature will likely overlap patterns describing fissures, as on some scale a cut in a surface looks like a cut in a surface. As a further example, similar confusions may occur between meandering channels and patterns found on a crater rim. These sources of confusion must also feed into statistical descriptions and resulting measurements. The probability distributions described in the previous paragraph could be used to quantify levels of confusions.

3.1.2 High levels of variation within features

Moving on to issues of variability, it may have been implied above that each feature should be associated with a single distribution of patterns, but this may not be the case. For example, the variability in the length of fissures might be described using two correlated pattern groups: the terminating end points patterns, which might appear in fixed quantities regardless of length; and the central segment patterns, which will be more abundant in longer fissures. Similarly, drainage network patterns may be divided into several sub-groups based on bifurcation pattern (distributary versus tributary) or channel sinuosity, all of which may appear in different networks in different proportions. A statistical model must be capable of identifying and combining such sub-groups in the most appropriate way to describe a terrain being analysed.

Additional variabilities caused by the interaction of features at occluding boundaries might be partially accounted for in terms of the pattern groups described in the previous paragraph. Table 1 noted ways in which features could interact, such a dunes engulfing other features or craters overlapping one another. If features were described using single distributions then these occlusion events would cause holes to appear within those distributions. However, if the crests and faces of dunes were grouped separately, and segments of crater rims were grouped separately, a dune overlapping a crater may be described with low quantities of a corrupted face and rim segment, whilst other distributions can be left uninterrupted.

3.1.3 Overview

In summary, an entire terrain image might be described as being a collection of correlated pattern groups and sub-groups, associated with features and parts of features, all appearing in different quantities in different regions. The pixels within patterns may be counted to make most types of measurements, with the errors on those measurements being affected by assumed Poisson perturbations and pattern ambiguities.

At a low level, the above description is consistent with the idea that the 'independent components' of an image are pixel patterns describable with edge filters [170]. The spatial distribution of these components have previously been modelled statistically using Markov Random Fields [171][172]. The resulting generative models can produce visually convincing reproductions of textures, but are less concerned with making quantitative measurements of textured regions. Here, what is required is a method for identifying and counting pixels belonging to particular classes of texture.

In the next section, these observations will be used to place the most general possible requirements and constraints on the design of a statistical model.

3.2 Model requirements

It will be assumed that any surface image can be encoded as a finite set of discrete patterns. A pattern, X, can be viewed as a feature vector describing local image structure. Rather than specifying a particular encoding, the representation will be kept abstract for now, allowing multiple possible encodings to use the same statistical model. However, a set of properties will be assumed about the patterns based upon the observations made in the previous section. It is a requirement of the model that the following properties are appropriately incorporated:

- a pattern appears as an independent Poisson event;
- the expected frequency of patterns are surface feature dependent;
- surface features may contain arbitrary distributions of patterns;
- a pattern can appear in more than one class of feature;
- and patterns sum linearly giving finite surface areas which can be related to physical measurements.

Any representation which satisfies the above properties can be substituted for X. A requirement of the model is to describe distributions of such patterns in a way that facilitates quantitative measurement. Taking into consideration the properties of images observed in the previous section the model must be able to:

- identify correlated groups of patterns and flexibly describe their arbitrary probability functions;
- associate groups of patterns to user defined surface features for which measurements are sought;
- and combine distributions in proportions appropriate for describing complicated highly variable terrains.

A training algorithm must determine how many components (pattern groups / probability functions) are required to describe the data and how each should be distributed. Once trained, resulting model components must be able to describe new incoming data, which can be assumed to contain the same distributions, but in unknown proportions. The measurements sought (areas, feature counts etc.) need to be easily extracted from the model by combining quantities of related components. The errors on measurements must also be computable from the model, accounting for the Poisson randomness of patterns and possible pattern ambiguities.

Whilst these requirements have been specified for planetary images, they are rather general and could describe other data sources where groups of Poisson events are generated. The following sections provide a potential solution to the modelling of this type of data.

3.3 Linear histograms

Discrete patterns occurring as independent Poisson events can be sampled into a histogram which can be represented as a vector, $\mathbf{H} = \{H_{X=1}, H_{X=2}, \ldots, H_{X=m}\}$, with an element for each pattern bin. The non-parametric nature of histograms is appropriate, as the model requirements specify that arbitrary distributions must be describable. The flexibility provided by a histogram means any distribution can be represented, regardless of its complexity, assuming a sufficient quantity of data can be supplied to populate the bins.

It was explained earlier that within any given image there will be a variety of surface features, and within any surface feature there may be a variety of related subgroups of patterns. It was also noted that any particular pattern found within one group may be found in others too. The recorded frequency in a histogram bin must then be explained by the accumulation of patterns from various sources, with each source being a function of the physical processes modifying a planet's surface. It is assumed that each pattern represents a fixed surface area, and that counting these areas is the basis for most measurements. This requires a linear additive model, as each pattern must contribute a fixed quantity to a planet's surface irrespective of their order or origin. The frequency recorded in a histogram bin can then be linked to n different processes by:

$$\mathbf{H}_X = \sum_{k=1}^{n} R(X|k) \tag{21}$$

where $R(X|k)$ is a function generating distributions of related patterns; X is a specific pattern; and k is a label indicating a particular pattern group. Assuming that the frequency of a pattern grows linearly with the quantity of surface features, the function R can be approximated by a weighted probability mass function (PMF):

$$R(X|k) \approx P(X|k)\mathbf{Q}_k \tag{22}$$

where $P(X|k)$ is the PMF for pattern group k and $\mathbf{Q}_k$ is the quantity of k in the data. An entire histogram can then be modelled as a linear system with components for each related pattern group:

$$\mathbf{H} = \mathbf{PQ} + \mathbf{e}_H \tag{23}$$

where $\mathbf{P}$ is an m by n matrix describing the PMFs of n components with elements $\mathbf{P}_{ij} = P(X = i|k = j)$, i.e. the probability of an entry in bin X given component k; $\mathbf{Q}$ is a column vector of n quantities corresponding to the amount of each component present within the histogram; and $\mathbf{e}_H$ is a column vector of noise assumed to be independent Poisson perturbations consistent with the histogram's formation. The inverted formulation, which is appropriate for making quantity measurements, is then given by:

$$\mathbf{Q} = \mathbf{P}^{-1}\mathbf{H} + \mathbf{e}_Q \tag{24}$$

where $\mathbf{P}^{-1}$ is an n by m matrix with elements consistent with Bayes Theorem[3], $\mathbf{P}_{ij}^{-1} = P(k = i|X = j)$, i.e. the probability that component k was the source of an entry in bin X; and $\mathbf{e}_Q$ is noise on the quantities. How PMFs and weighting quantities can be estimated will be addressed shortly, but first their meaning and significance in making measurements will be elaborated upon.

3.4 Components, quantities and measurements

It was stated in the properties of planetary images that measurements, such as surface areas, relate to the quantities of patterns found within each class of interest. The quantities of patterns within each class, in turn, relate to linear model components, $\mathbf{P}$, and the quantities by which they are weighted, $\mathbf{Q}$. In the simplest case each class of feature may contain a single highly stable distribution of patterns describable using one fixed PMF per class. Assuming each pattern, X, covers a fixed surface area, the total image area accounted for by each class of feature will be directly proportional to the weighting quantities, $\mathbf{Q}$.

Most features, however, will exhibit a complex range of allowable variations, as described earlier in the properties of planetary images. In these cases it is insufficient to describe a class of terrain using only one fixed PMF. Instead, a set of related PMFs can be used per feature making each class a nested subset of the linear model. Each sub-model can contain as many components as are necessary to describe the modes of variation present in that class of terrain. All components are placed into the matrix, $\mathbf{P}$, side-by-side, with their interpretation requiring knowledge of which components are associated with which class of feature. From the point of view of measurements, this means summing related quantities together to give total class quantities:

$$\mathbb{Q} = \mathbb{K}\mathbf{Q} + \mathbf{e}_{\mathbb{Q}} \tag{25}$$

where $\mathbb{Q}$ is a column vector containing o quantities of the o feature classes for which measurements are sought; $\mathbb{K}$ is an o by n mapping matrix with mutually exclusive binary elements indicating which model components belong to which class of feature; and $\mathbf{e}_{\mathbb{Q}}$ is the noise on the total class quantities. The elements of $\mathbb{K}$ are then:

$$\mathbb{K}_{ij} = \delta(k_j \in K_i) \tag{26}$$

where k_j is the jth model component and K_i is the ith class for which quantity measurements are sought; and the function δ is 1 when component k is within the subset K

[3] It will be seen later and in subsequent chapters that $\mathbf{P}^{-1}$ is a function of the noise terms $\mathbf{e}_H$ and $\mathbf{e}_Q$.

and 0 otherwise.

An expert user must define the individual classes of feature for which measurements are sought by providing exemplars. A training algorithm must use the exemplars to extract linear sub-models for each class to construct the components of **P** and set the mapping from components to classes, $\mathbb{K}$. The extraction of linear components may be recognised as a form of Independent Component Analysis which is commonly found in pattern recognition. i.e. a set of maximally independent, non-orthogonal, model components are extracted sufficient for describing variations in the data. Here, a linear decomposition is required suitable for histogram models. Once trained, an algorithm must fit the model to new images in order to measure the component quantities, **Q**, and the class quantities sought, $\mathbb{Q}$. The next few sections describe how a Likelihood approach can be applied to compute the best fitting parameters yielding the most probable statistical model during both training and making measurements.

3.5 Likelihood parameter estimation

Populating a histogram through the counting of independent Poisson events is a common model. The statistical Likelihood used to describe this scenario is usually attributed to Fermi in the form of Extended Maximum Likelihood [79], which in terms of the linear model given by equation (23) becomes:

$$\ln \mathcal{L} = \sum_X \ln \left[\sum_k P(X|k)\mathbf{Q}_k \right] \mathbf{H}_X - \sum_k \mathbf{Q}_k \tag{27}$$

where $P(X|k)$ (i.e. **P**) and **Q** are parameters to be estimated. This Likelihood can be optimised during training and also during the taking of measurements. During training, when the components are unknown and need to be learned, the Likelihood function can be maximised to jointly estimate **P** and **Q** for multiple example histograms. During measurements, assuming **P** is representative of the incoming data, **H**, the Likelihood can be maximised to only estimate the quantities. In both cases an Expectation Maximisation(EM) algorithm can be applied to perform the optimisation.

Whilst other estimation methods are available it is appropriate to use Likelihood for quantitative measurements, as the link between Likelihood and error estimates is well understood. Chapter 4 will show how the stability of Likelihood estimated parameters can be determined. This is essential for the quantitative criteria prescribed in the introductory chapter (section 1.6), and the model requirements stated in this chapter (section 3.2), to be met.

The estimation of **Q** will be considered first, as this is the simpler of the two parameters to estimate and is also required for the subsequent estimation of **P**. The estimation of **P** will be considered afterwards, where multiple exemplars of each class are used to constructs the nested linear models using an EM version of an Independent Component Analysis

(ICA).

3.5.1 Quantity estimation: Q

In cases where $\mathbf{P}$ is known the EM algorithm can be used to optimise (27) with respect to free parameters, $\mathbf{Q}$. EM iteratively updates the elements of $\mathbf{Q}$ by weighting them with the current estimate of the posteriori probability $P(k|X)$, (i.e. $\mathbf{P}^{-1}$), starting from some initial (perhaps randomised) estimate at iteration $t = 0$:

$$\mathbf{Q}_{(t)} = \mathbf{P}^{-1}_{(t-1)}\mathbf{H} \tag{28}$$

where $\mathbf{Q}_{(t)}$ is the quantity vector estimate at time t; and $\mathbf{P}^{-1}_{(t-1)}$ is the last estimate of the posteriori probabilities. For a single quantity, k, this update is implemented using Bayes Theorem [121][127]:

$$\mathbf{Q}_{k(t)} = \sum_X \frac{P(X|k)\mathbf{Q}_{k(t-1)}}{\sum_l P(X|l)\mathbf{Q}_{l(t-1)}}\mathbf{H}_X \tag{29}$$

where $P(X|k) = \mathbf{P}_{i=X,j=k}$; and the Bayesian 'prior' is substituted for the quantity being estimated[4]. Upon convergence the quantities, $\hat{\mathbf{Q}} = \mathbf{Q}_{(t=\infty)}$, provide the maximum Likelihood solutions assuming the PMFs of the model components were correct for the incoming histogram data.

This process can be summarised in pseudo-code:

```
1. Randomly initialise the weighting quantity vector $\mathbf{Q}$
2. For incoming histogram $\mathbf{H}$:
    (a) Compute equ. (28) via repeated use of equ. (29) for each component, $k$
    (b) Observe the change in $\mathbf{P}^{-1}$ and $\mathbf{Q}$
    (c) Repeat until $\mathbf{P}^{-1}$ and $\mathbf{Q}$ converge
3. Compute total class quantities, $\mathbb{Q}$, using equ. (25)
```

3.5.2 Component PMF estimation: P

Each surface terrain requires a set of PMFs to be estimated which, when linearly combined, can approximate the distribution and modes of variation of the patterns indicative of each class. This requires multiple example histograms of terrains to be provided by an expert user, $\{\mathbf{H}_{(r=1)}, \mathbf{H}_{(r=2)}, \ldots\}$. These examples must contain mixtures of components in different proportions so that the modes of variation can be observed. For notational

[4] The estimation of quantities as a sum of probabilities seen in equation (29) is consistent with the method of counting described in section 2.5, where honest probabilities are summed as an alternative to counting decisive labels.

convenience the description that follows will be presented as if there is only a single class of feature. During real training the algorithm should be applied separately to sub-matrices associated with each class of feature being learned.

The number of components required to fully describe the data will generally be unknown. A model selection process based upon a goodness-of-fit function will be presented in the next section, but for now it will be assumed that there are n unknown components in a class. Given a class composed of n unknown PMFs, and given a set of N ($n \ll N$) independent exemplar pattern histograms of that class, the EM algorithm can be applied to perform an ICA appropriate for histogram data.

Initial estimates of component PMFs (perhaps randomised) are generated at iteration 0 giving $\mathbf{P}_{(t=0)}$. Weighting quantities are estimated (via estimation process of section 3.5.1) for each of the N examples giving modelled approximations:

$$\mathbf{H}_{(r)} \approx \mathbf{M}_{(r)(t)} = \mathbf{P}_{(t)}\mathbf{Q}_{(r)(t)} \tag{30}$$

where all histogram models (i.e. each r) share a common definition of components $\mathbf{P}_{(t)}$; and each histogram has its own estimate of weighting quantities, $\mathbf{Q}_{(r)(t)}$. In line with the EM algorithm the estimated posteriori probabilities, $\mathbf{P}^{-1}$, are used to provide a new estimate of the contribution to each histogram, r, from each component, k:

$$R_{r(t)}(X|k) = P_{r(t-1)}(k|X)\mathbf{H}_{X(r)} \tag{31}$$

This results in a set of approximate data distributions, one for each exemplar, which must be combined to give a new joint approximation of each PMF. It can be shown that each exemplar's estimated distribution has approximate Poisson properties allowing them to be treated as sampled histograms, and therefore allowing them to be combined through addition. These approximate Poisson properties can be seen through a two step argument:

1) The variance on $R_r(X|k)$ includes a Binomial contribution, as each bin entry has a probability of $P(k|X)$ of being included in the sample and a probability of $1 - P(k|X)$ of being excluded. This variance is given by:

$$\mathbf{H}_{X(r)}P(k|X)[1 - P(k|X)] = \mathbf{H}_{X(r)}P(k|X) - \mathbf{H}_{X(r)}P(k|X)^2 \tag{32}$$

2) There is also an independent Poisson contribution from the originating histogram:

$$\left(\frac{\partial R_r(X|k)}{\partial \mathbf{H}_{X(r)}}\right)^2 \mathbf{H}_{X(r)} = P(k|X)^2\mathbf{H}_{X(r)} \tag{33}$$

which cancels with the second term of the Binomial contribution giving:

$$\sigma^2_{R_r(X|k)} = P(k|X)\mathbf{H}_{X(r)} = R_r(X|k) \tag{34}$$

resulting in Poisson variance allowing the individual example histograms to be safely combined conserving their Poisson characteristics:

$$R(X|k) = \sum_r R_r(X|k) \tag{35}$$

$$\sigma^2_{R(X|k)} = R(X|k) \tag{36}$$

An understanding $\sigma^2_{R(X|k)}$ is essential for meeting the quantitative criteria of computing error bars on final measurements and allowing measurement errors to be corroborated empirically. These estimates will be used in a following section to corroborate the goodness-of-fit of the model to data, and again in chapter 5 where errors on measurements caused by inaccuracies in the model will be addressed.

The estimates of $R(X|k)$ are normalised to create a new common estimate of $\mathbf{P}_{(t)}$:

$$\mathbf{P}_{i=X,j=k,(t)} = P_t(X|k) = \frac{R_{(t)}(X|k)}{\sum_r \mathbf{Q}_{k(r)(t)}} \tag{37}$$

This new common estimate of $\mathbf{P}$ is used to reestimate quantities and posteriori probabilities for exemplar histograms and the process continues until convergence giving $\hat{\mathbf{P}} = \mathbf{P}_{t=\infty}$, maximising (27) consistent with the convergence theorem of EM [121][127]. To avoid the risk of converging into a local minimum the algorithm can be restarted multiple times from different random PMF initialisations. Once a set of PMFs have been estimated for a given class the mapping matrix, $\mathbb{K}$, must be updated to indicate which components are associated with that type.

The algorithm can be summarised in pseudo-code:

```
1. For each class $\{K = 0, K = 1, \dots, K = o\}$, gather $N$ exemplar histograms of class
   $K$ $\{\mathbf{H}_{(r=1)}, \mathbf{H}_{(r=2)}, \dots, \mathbf{H}_{(r=N)}\}$
2. Randomly initialise common definition of $\mathbf{P}$
3. Using $n$ components, while $\mathbf{P}$ has not converged:
   (a) For each histogram $\mathbf{H}_{(r)}$:
       i. Iterate equ. (28) and equ. (29) until $\mathbf{P}^{-1}_{(r)}$ and $\mathbf{Q}_{(r)}$ converge
       ii. Use $\mathbf{P}^{-1}_{(r)}$ to estimate individual component contributions, $\hat{R}_r$, using equ.
           (31)
   (b) Sum and normalise each component to give new common estimate of $\mathbf{P}$ using
       equ (37)
4. Update component-class mapping matrix, $\mathbb{K}$, to indicate which components belong
   to class $K$
```

3.6 Goodness of fit

The importance of testing theories by corroborating predictions was emphasised during the introductory chapter. The linear Poisson histogram model described in this chapter constitutes a theory for how patterns within planetary images should be distributed, and therefore should be capable of making testable predictions. In particular, a set of fitted model parameters (**P** and **Q**) forms the basis of hypotheses for how many entries should be observed in each histogram bin within incoming data. This naturally leads from equations (21) and (22) to give predicted bin frequencies:

$$\mathbf{H}_X \approx \mathbf{M}_X = \sum_{k=1}^{n} P(X|k)\mathbf{Q}_k \tag{38}$$

where $\mathbf{M}_X$ is the bin frequency predicted by the model for pattern X, which should be approximately equal to the observed frequency, $\mathbf{H}_X$, within the limits of Poisson sampling statistics. This approximation can be tested by comparing the distribution of model-data residuals to the expected Poisson errors. For a given dataset, a ratio of unity between observed and predicted residuals is corroboration that the model is successfully describing the incoming histogram, whilst a ratio above unity indicates that the model is inappropriate for that particular data. This test can be performed using a modified chi-squared statistic designed for histograms.

A standard chi-squared per degree of freedom function is given by:

$$\frac{1}{D}\sum_{X}\frac{(\mathbf{M}_X - \mathbf{H}_X)^2}{\sigma_X^2} \tag{39}$$

where σ_X^2 is the expected variance on the residual between the model and observed data at pattern X; and D is the number of degrees of freedom in the model. A standard chi-squared per degree of freedom assumes Gaussian residuals, not Poisson residuals as found in histograms. Whilst a Poisson can be approximated with a Gaussian at large sample sizes, the discrepancies between the distributions at low sample sizes may cause problems, especially for poorly populated histograms. A square-root transform [174] can be performed on histogram bins to transform the residuals into something better approximating a Gaussian with uniform width of $\sigma^2 = \frac{1}{4}$. This property of the square-root transform can be seen via error propagation:

$$\begin{aligned}\sigma^2_{\sqrt{\mathbf{H}_X}} &= \left(\frac{\partial\sqrt{\mathbf{H}_X}}{\partial\mathbf{H}_X}\right)^2 \sigma^2_{\mathbf{H}_X} \\ &= \left(\frac{1}{2}\mathbf{H}_X^{-\frac{1}{2}}\right)^2 \mathbf{H}_X \\ &= \frac{1}{4}\mathbf{H}_X^{-1}\mathbf{H}_X = \frac{1}{4}\end{aligned} \tag{40}$$

where the Poisson error $\sigma^2_{\mathbf{H}_X} \approx \mathbf{H}_X$. A chi-squared per degree of freedom test on the transformed data:

$$\frac{1}{D}\sum_X \frac{\left(\sqrt{\mathbf{M}_X} - \sqrt{\mathbf{H}_X}\right)^2}{\sigma^2_X} \tag{41}$$

can form the basis of a model selection scheme during training and a success indicator when making measurements. The variance to which each residual is normalised will be different in the two cases. These uses and how they are normalised are described below.

3.6.1 Model selection

The Likelihood parameter estimation algorithm described in section 3.5 assumed that the number of components required to model a class of feature was known a-priori. It was also noted that in general this would not be the case. The problem of determining the most appropriate number of components can be solved using the Chi-squared per D degrees of freedom function, χ^2_D:

$$\chi^2_{(m-n)} = \frac{1}{m-n}\sum_X \frac{\left(\sqrt{\mathbf{M}_X} - \sqrt{\mathbf{H}_X}\right)^2}{\frac{1}{4}} \tag{42}$$

where m is the number of bins in the histograms; n is the number of components in the model; and the model-data residuals are assumed to have a variance driven only by the incoming data training histograms, which via the square-root transform becomes a constant value of $\frac{1}{4}$. The model selection problem can be solved by executed the training algorithm multiple times, incrementing n until χ^2_D reaches unity. However, some care must be taken, as if n is too large giving a goodness-of-fit less than unity, the model will become over-trained, i.e. additional components will be modelling noise specific to the training examples which may not generalise to future datasets.

3.6.2 Measurement success indicator

During training, whilst the model components are being defined, only noise from incoming histograms are considered in the χ^2_D function. However, once trained, the sampling noise from those histograms become systematic errors in the extracted $\mathbf{P}$ components. When making measurements by fitting the model to new incoming data these errors will change the expected model-data residuals. The goodness-of-fit can be normalised to these residuals by approximating the errors on the PMFs and incorporating them into the function. The errors on PMFs are easy to calculate, as they are just Poisson histogram errors from equation (35) which have been scaled to give histograms of unit integral, i.e. probabilities. Error propagation then gives:

$$\sigma^2_{P(X|k)} \approx \left(\frac{\partial P(X|k)}{\partial R(X|k)}\right)^2 \sigma^2_{R(X|k)} \tag{43}$$

$$= \frac{R(X|k)}{(\sum_r \mathbf{Q}k(r))^2} \tag{44}$$

The errors on the square-root transformed predicted model frequencies, again via error propagation, become:

$$\sigma^2_{\sqrt{\mathbf{M}_X}} = \sum_k \left(\frac{\partial\sqrt{\mathbf{M}_X}}{\partial P(X|k)}\right)^2 \sigma^2_{P(X|k)}$$

$$= \sum_k \frac{\mathbf{Q}_k^2}{4\mathbf{M}_X}\sigma^2_{P(X|k)}$$

The goodness-of-fit is then given by:

$$\chi^2_{(m-n)} = \frac{1}{m-n}\sum_X \frac{\left(\sqrt{\mathbf{M}_X} - \sqrt{\mathbf{H}_X}\right)^2}{\frac{1}{4} + \sigma^2_{\sqrt{\mathbf{M}_X}}} \tag{45}$$

Which should be unity when the model is applied to new data composed from sources representative of those seen during training. Any unforeseen contamination in new data, including highly irregular instances of known features, should cause larger residuals than are predicted by the model and therefore produce fits above unity. A data set that fails the chi-squared per degree of freedom test cannot be trusted and any measurements derived from it should be disregarded.

3.7 Monte-Carlo testing

The testing of actual measurements is avoided until chapter 4. The focus here is on the model's ability to produce honest probability distributions, i.e. do PMFs reflect the true frequencies with which pattern events occur? The goodness-of-fit function can be used to check this, as it is a direct test of predicted frequencies. A Monte-Carlo simulation was created to generate histogram data following the pattern properties given in section 3.2. This simulation was used to test the statistical model's ability to describe incoming data using a range of different distributions, numbers of components and quantities of classes. These tests used the goodness-of-fit functions to confirm that the training algorithm could successfully extract representative PMFs and that those PMFs could be combined to describe new incoming distributions whilst spotting problems caused by simulated contamination. These tests included:

- demonstrating that the goodness-of-fit function could be used for model selection, i.e. it reaches unity when the correct number of components have been extracted during training;

- confirming that extracted PMFs could model new incoming data, i.e. the goodness-of-fit reaches unity when model weights have been estimated using EM;
- and showing that contamination can be spotted, i.e. the goodness-of-fit does not reach unity when unrepresentative data is presented.

Training and testing histograms were generated by scaling reference distributions (in the form of PMFs) by randomly selected known quantities. The scaling quantities were all real valued uniform random numbers between 1,000 and 1,000,000. For each histogram bin, a random Poisson number generator was used to draw independent samples from each component before summing them to give total histogram frequencies.

Model selection was tested for histograms containing 64, 4096 and 16384 bins. Data was synthesised for each sized histogram containing known linear combinations of predefined reference distributions. The aim of testing was to confirm that the EM ICA algorithm could successfully extract sufficient approximations to all of the predefined distributions so as to achieve a goodness-of-fit of unity. This was performed for increasingly complex models containing between 1 to 10 linear components. Figure 5 plots the goodness-of-fit for different numbers of extracted components for the 4096 bin histograms. These results are indicative of the other histogram sizes.

The 64 bin reference distributions were coded by hand and can be seen in figure 8. These distributions were used as templates for creating 10 different components by shifting the distributions along the x-axis. For example, figure 9 shows the approximate distributions extracted during 64 bin tests at the point where 3 components per model were generated. In contrast, the 4096 and 16384 bin histogram component reference distributions were randomly generated[5], giving a complex range of synthetic data.

During each iteration the number of training histograms was increased ensuring there were always 10 times as many example histograms as components in the model, i.e. in figure 5 there were 10 training examples at the left end of the plot, rising to 100 at the right end.

The analysis of new data was tested by fitting components extracted using the EM ICA training algorithm to previously unseen histograms via the EM model fitting algorithm. This was done for all sizes of histogram and numbers of components. Each fit was performed multiple times using a different ratio of training to testing data. Results can be seen in figure 6 where the relative quantity of training data ranged from 0.01 to 100 times that of the testing data.

[5] The random reference distributions were created using a texture sampling scheme which will not be presented until chapter 6. To summarise, the distributions were sampled from martian image data using 12 bit and 14 bit BRIEF-like descriptors, involving an initial random selection of pixels making the organisation of the histogram bins essentially random. These experiments should not be confused with those presented later. Here, the sampled reference distributions became the source of other histograms, whereas in chapter 6 all histograms were sampled from images.

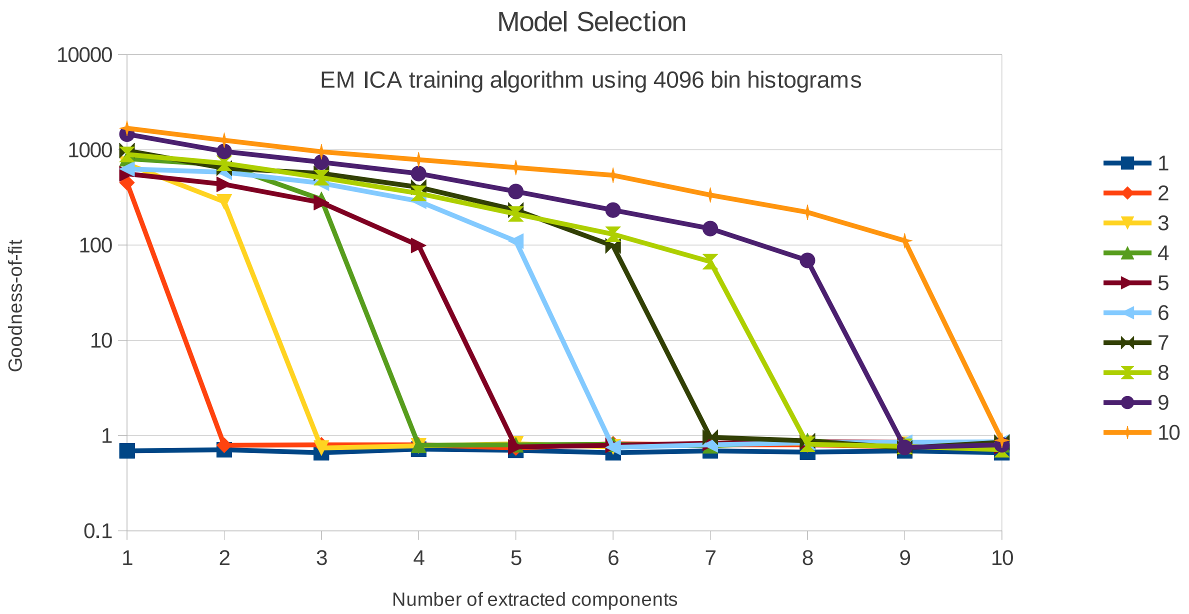


Figure 5: This plot shows that by using the χ^2_D goodness-of-fit function the EM ICA algorithm can extract an appropriate number of linear components to describe a set of example training histograms. Each curve represents a Monte-Carlo data source generated with between 1 and 10 components. Each curve crosses unity at the most appropriate point, i.e. when the number of extracted components equals the number of generating components.

Finally, the goodness-of-fit was used as a success indicator to confirm that unrepresentative testing data containing random outliers could be detected. This was achieved by replacing a set percentage of incoming histogram bins with uniformly distributed random numbers within an order of magnitude of the original data. This was performed for all sized histograms containing 10 components. Results can be seen in figure 7.

3.8 Discussion

The training algorithm, quantity estimation algorithm and goodness-of-fit functions all performed as predicted. The use of the χ^2_D goodness-of-fit function proved to be effective at selecting the necessary number of components to describe a set of example histograms. In all cases the fit approximately equaled unity when the number of extracted components equaled the true number of generating components, as seen in figure 5. The learned PMFs were all successfully fitted to new representative data, again reaching a fit of approximately unity upon convergence of the EM algorithm, as seen in figure 6. Finally, when random entries were made in a proportion of the bins to simulate contamination of data, the goodness-of-fit function grew above unity, as seen in figure 7. The process of extracting PMFs using EM from an initial random seed can be seen in figure 10, which plots the evolution of a single component over time as the EM algorithm converges.

An inspection of extracted components reveals some discrepancies between true gener-

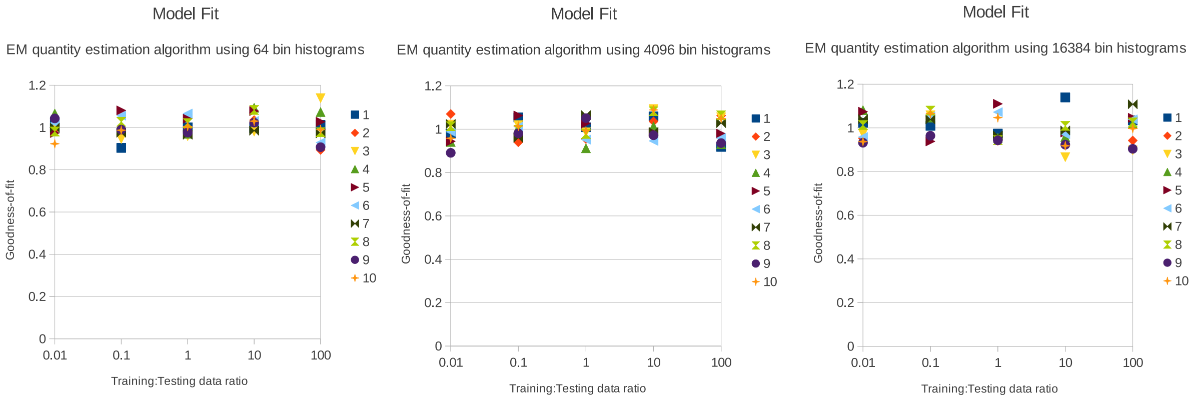


Figure 6: This plot shows that the EM quantity parameter estimation algorithm can successfully fit previously learned components to new unknown mixtures of the same components in the case when they are representative. Each data point corresponds to a model of different complexity of between 1 and 10 components. It can be seen that the goodness-of-fit function remains stable around unity for varying quantities of training and testing data.

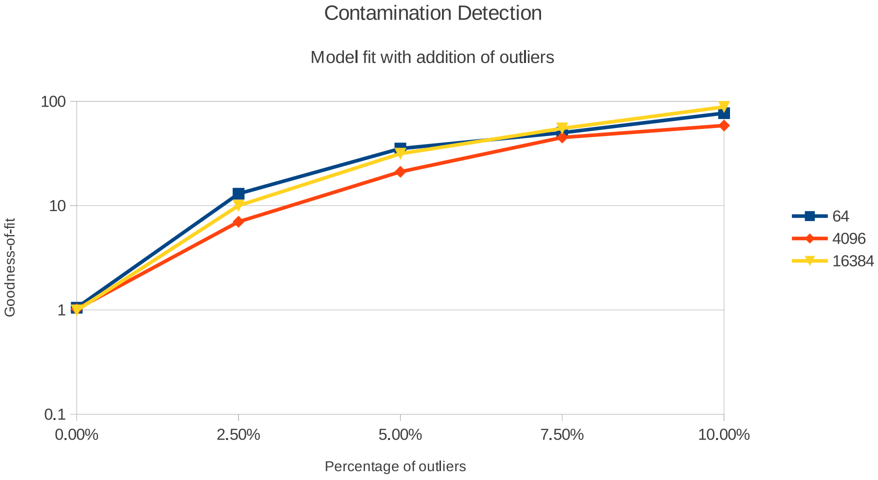


Figure 7: This plot shows that the χ^2_D goodness-of-fit function can successfully detect contamination in data as outliers are purposefully introduced in varying quantities to otherwise representative data.

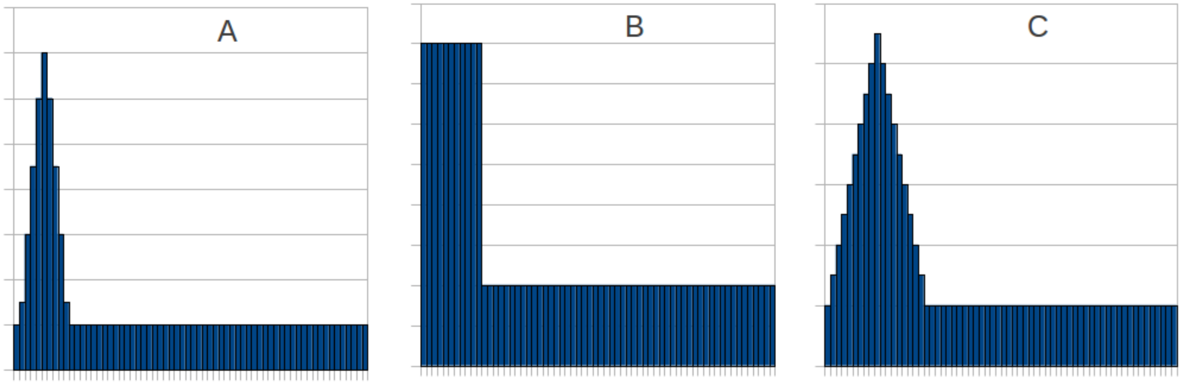


Figure 8: These are example distributions used as generators within the Monte-Carlo simulation. Shifted and scaled versions of these distributions were used to generate the 64 bin models.

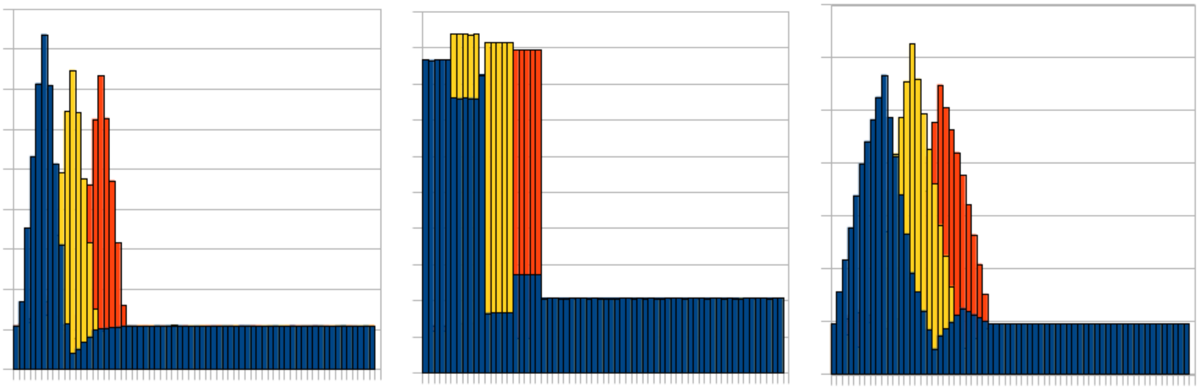


Figure 9: These are examples of components extracted from 64 bin training histograms created during the Monte-Carlo simulations.

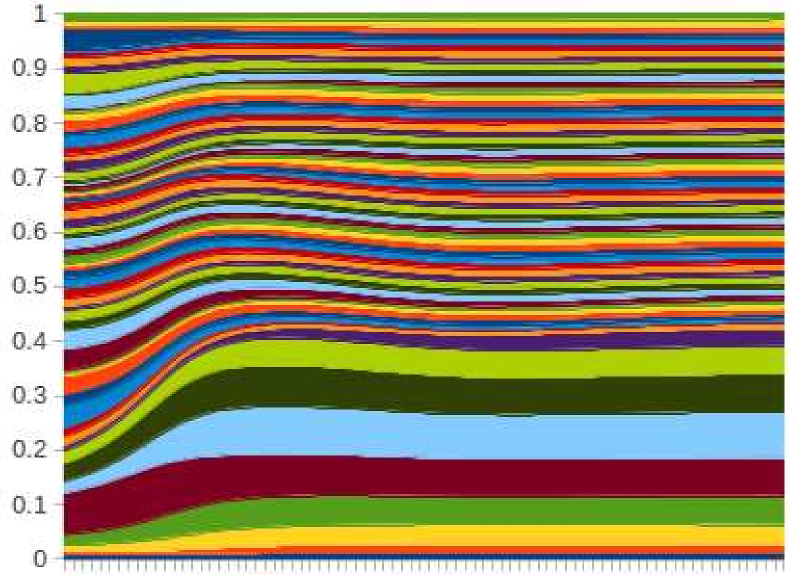


Figure 10: This plot shows the evolution of a randomly initialised 64 bin PMF during the execution of the EM ICA algorithm. Time runs positively along the x-axis. The algorithm can be seen to converge quickly to reach a stable estimate of the underlying data generating component.

ating PMFs and those determined using the EM ICA training algorithm. Figure 8 shows examples of generating distributions, scaled and shifted linear combinations of which were used to generate a series of training example mixtures. Figure 9 shows the PMFs extracted from those mixtures, which resemble the originating distributions, but with some unexpected peaks and troughs. The mismatch between generators and extracted components can be explained in terms of the span of linear subspaces if PMFs are viewed as vectors defining a manifold in an m dimensional histogram space. Each example histogram forms a point in the histogram space. Together, all examples define a hyperplane (assuming a linear model is the correct model) upon which all possible mixtures of components lie. As long as the extracted components lie upon the same hyperplane and can span the same subspace as the original generators then a linear model constructed from them will be capable of describing similar mixtures. From a quantitative perspective, what is important is the ability to predict bin frequencies within known errors. The χ^2_D function has corroborated predicted frequencies modelled by extracted components, assuming Poisson bin errors. The extracted components can, therefore, be considered to generate valid statistical descriptions of the data, despite the apparent discrepancies upon visual inspection.

The above argument defending the use of alternative, but similar, component definitions than the original generators becomes more complex when other constraints are considered. In general, the span of a vector subspace includes points reachable using negative coefficients. However, the weighting quantities found within histogram components must be non-negative. The consequence of this is that suitable training data must provide example mixtures with coefficients spanning a full range of allowable variations. In particular, examples containing the lowest expected quantity of each component must be present in training data, providing representative samples of the most extreme mixtures likely to be found in future. If a future dataset is encountered containing an unprecedentedly low quantity of a component there is no guarantee that the trained components can interpolate far enough to reach the mixture's location on the histogram manifold.

Assuming appropriate training examples are given the statistical models produced appear to be valid. Monte-Carlo data is, however, artificially representative and the mixture models generated during the above tests were forced to exhibit the properties that planetary surface patterns are assumed to have. In practice, real data may not behave this well. The goodness-of-fit functions do provide a safeguard against unrepresentative data, but the modified chi-squared per degree of freedom will not be capable of spotting all problems with data. Whilst it can highlight larger than expected residuals, it cannot spot correlations between residuals. If a linear model has not successfully separated all correlated patterns into separate components, or if there are highly localised correlations between bins, it is still possible for the goodness-of-fit to reach unity despite the violation of the assumption of independent Poisson noise within each histogram bin. Only testing on real data can confirm the appropriateness of the statistical model for practical use.

3.9 Summary

This chapter presented a linear Poisson model capable of describing the distribution of patterns that may be found within planetary images. Rather than selecting a specific representation for image patterns, this chapter assumed a set of reasonable properties which might be expected of image patterns and a statistical model was then created to address them. A Likelihood parameter estimation process, based upon the EM algorithm, has been presented for extracting correlated groups of patterns using a histogram ICA. The same Likelihood was optimised for fitting those groups to new incoming data. The link between the quantities of pattern groups and measurements was also made. Monte-Carlo testing was conducted, with an emphasis on achieving approximations of histogram distributions within expected Poisson perturbations. Satisfactory approximations to data distributions were achieved, with corroboration from a modified chi-squared per degree of freedom goodness-of-fit function, showing that predicted bin frequencies generated by the model matched, within errors, actual observed bin frequencies in simulated data.

The limitations of the method were highlighted, including the inability of the model to interpolate to negative model coefficients, making it important for training data to exhibit a full range of variations for appropriate model components to be extracted. The limitations of the goodness-of-fit function were also noted, concluding that by itself it was incapable of confirming all the assumed properties of the data.

The artificially representative testing, via Monte-Carlo, has confirmed that the modelling method is appropriate, as long as the patterns found within real data abide by the properties assumed in this chapter. The end goal of producing real measurements from real data will require much more testing. This testing must include an understanding of measurement errors, not just individual model bin errors. The understanding of measurement errors will begin in the next chapter where the stability of model weighting quantities will be addressed.

4 Statistical Error Estimation

The previous chapter provided a method for describing distributions of patterns found within planetary terrain images using a linear Poisson model based upon sampled histograms. The weighted components of that model were linked to groups of patterns found within surface features for which measurements were sought. It was argued that most measurements could be viewed in terms of counts of pixels and that such counts were proportional to the quantity parameters, $\mathbf{Q}$, which could be estimated using Likelihood. Those quantity estimates were associated with errors, $\mathbf{e}_Q$, in equation (24), which must be understood for the quantitative criteria prescribed in section 1.6 to be fulfilled. This chapter will begin to explore these errors using the Cramer Rao Bound to place a lower bound on the errors expected to be present on these parameters, which in turn will place a lower bound on associated measurement errors.

Within this chapter it will be shown that a covariance matrix can be computed for predicting the statistical spread of estimated quantities. A theoretical analysis will reveal the relationships between statistical errors, underlying Poisson processes and ambiguities found between model components. These covariances will be combined using error propagation, consistent with equation (25), to give covariances on class quantities, $\mathbb{Q} + \mathbf{e}_{\mathbb{Q}}$. Predicted errors will be corroborated using Monte-Carlo simulated data, showing the applicability of the lower bound over a range of distributions and quantities of data.

4.1 Cramer Rao Bound

The Cramer Rao Bound (CRB) provides a lower bound on the variance of Likelihood estimated parameters. This bound can be applied to the Likelihood function of equation (27) giving an estimate of the stability of the $\mathbf{Q}$ parameters:

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \mathbf{Q}_i \partial \mathbf{Q}_j} \geq \mathbf{C}_{ij}^{-1} \tag{46}$$

where $\mathbf{C}_{ij}^{-1}$ is the inverse covariance between quantities $\mathbf{Q}_i$ and $\mathbf{Q}_j$. Computing the first and second order derivatives gives:

$$\frac{\partial \ln \mathcal{L}}{\partial \mathbf{Q}_j} = \sum_X \frac{P(X|j)\mathbf{H}_X}{\sum_k P(X|k)\mathbf{Q}_k} - 1 \tag{47}$$

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \mathbf{Q}_i \partial \mathbf{Q}_j} = \sum_X \frac{P(X|i)P(X|j)\mathbf{H}_X}{[\sum_k P(X|l)\mathbf{Q}_k]^2} \tag{48}$$

then the similarity with Bayes Theorem can be used to give:

$$\mathbf{C}_{ij}^{-1} \approx \frac{\sum_X P(i|X)P(j|X)\mathbf{H}_X}{\mathbf{Q}_i \mathbf{Q}_j} \tag{49}$$

From this the covariance matrix, $\mathbf{C}$, can be computed using a standard matrix inversion algorithm.

4.1.1 Anticipated properties

It was stated in section 3.2 that errors on measurements must account for Poisson perturbations and possible ambiguities between similar patterns. Before inspecting the covariance, $\mathbf{C}$, to find if these issues are addressed, logical arguments will be used to identify two key properties which should be seen.

Firstly, if there is no ambiguity between components the only source of variance should be from the independent Poisson perturbations assumed to exist in the generation of patterns. The underlying generator of patterns, $R(X|k)$, was introduced in equation (21) and estimated in equation (35). For a single quantity, $\mathbf{Q}_k$, the associated generator should give:

$$\mathbf{Q}_k = \sum_X R(X|k) \tag{50}$$

where the quantity of component k is an accumulation of all patterns associated with k. As the contribution from each pattern is assumed to be an independent Poisson quantity, the variance on the total quantity should be the sum of the variances of the individual parts:

$$\sigma^2_{unam} = \sum_X \sigma^2_{R(X|k)} = \sum_X < R(X|k) > \tag{51}$$

which predicts an unambiguous variance, σ^2_{unam}, equal to the expectation of the quantity:

$$\sigma^2_{unam} =< \mathbf{Q}_k > \tag{52}$$

giving the lowest possible variance attainable due to fundamental properties of the data.

Secondly, if there is ambiguity between patterns, i.e. there is a finite probability that a pattern could have originated from more than one source, then logically the variances on the related quantities should increase. This should be expected, as the possibility of classification mistakes must grow with the possibility of confusion between patterns giving:

$$\sigma^2_{unam} \leq \sigma^2_{\mathbf{Q}_k} \tag{53}$$

where $\sigma^2_{\mathbf{Q}_k}$ is the variance on estimated quantity $\mathbf{Q}_k$.

4.1.2 Lower bound properties

An inspection of the model component covariance, **C**, reveals that the CRB does have the expected properties. These are best seen by studying just the variance terms in the covariance matrix. The diagonal terms of the inverse covariance reduce to:

$$\mathbf{C}^{-1}_{kk} \approx \frac{\sum_X P(k|X)^2 \mathbf{H}_X}{\mathbf{Q}^2_k} \tag{54}$$

giving a variance estimate for a single component of:

$$\sigma^2_{\mathbf{Q}_k} \approx \frac{\mathbf{Q}^2_k}{\sum_X P(k|X)^2 \mathbf{H}_X} \tag{55}$$

The Poisson uncertainty in the estimation of a single model quantity can be seen by assuming that the associated component is completely unambiguous, i.e. that each pattern, X, within the component has a posteriori probability of exactly zero or one. The sum over patterns in the denominator then becomes equivalent to simply summing all patterns guaranteed to be part of the component, i.e. $\sum_X R(X|k)$ from equation (50). This gives an unambiguous variance estimate of:

$$\sigma^2_{unam} \approx \frac{\mathbf{Q}^2_k}{\mathbf{Q}_k} = \mathbf{Q}_k \tag{56}$$

consistent with the expected property that the lowest possible variance is driven by the Poisson processes at work on the level of individual patterns. This can only increase as ambiguities between patterns grow, i.e. that each pattern, X, has a posteriori probability somewhere between zero and one, consistent with the second expected property that variances should go up with ambiguity:

$$\mathbf{Q}_k \leq \frac{\mathbf{Q}^2_k}{\sum_X P(k|X)^2 \mathbf{H}_X} \tag{57}$$

In summary, the CRB theoretically accounts for the sources of uncertainty assumed to be within the data and exhibits the properties that would be expected of an estimator of errors for model component quantities.

4.2 Class quantity covariance

The covariance, **C**, is an estimate of the statistical spread of individual model component weights. What is of real interest is the statistical spread of the quantities of classes, $\mathbb{Q}$, computed via equation (25). This can be estimated using error propagation, which for the simple case of summed independent quantities becomes:

$$\mathbf{C}_{\mathbb{Q}} = \mathbb{K}\mathbf{C}\mathbb{K}^{\intercal} \tag{58}$$

where $\mathbb{K}$ is the component to class mapping matrix of equation 26.

4.3 Monte-Carlo testing

The process of applying the CRB then combining related errors together via error propagation was tested as a predictor of errors on class quantities using Monte-Carlo simulated data. Two types of error were investigated: the statistical spread of estimated quantities around mean estimates; and the total spread of quantities around true values. The first, purely statistical test, assessed the stability of repeated estimated quantities. The second test assessed the ability to estimate true quantities allowing for the possibility of systematic effects caused by mismatches between modelled components and incoming data.

The Monte-Carlo simulation developed in the previous chapter was reused to test error predictions. Histogram components previously extracted from the 64 bin, 4096 bin and 16384 bin experiments were grouped into 3 classes each containing 3 components. Known random quantities of each component were repeatedly combined to generate a series of new histograms. The EM model fitting algorithm was used to fit the extracted components to this new incoming data and estimated quantity parameters, $\mathbf{Q}$, were inspected. These quantities were summed into class quantities, $\mathbb{Q}$, with the differences between true class values and estimated values being measured over multiple independent trials.

The training histograms were constructed using uniform random real valued quantities between 1,000 and 1,000,000. The incoming testing histograms were generated using a range of relative quantities from 0.01 to 100 times the training amounts, consistent with the experiments of the previous chapter. Each ratio of training to testing data was repeatedly tested using different random quantities over 1,000 trials. At each trial the difference between the true class quantities and estimated class quantities were divided by the predicted standard deviation (computed by using the CRB on each components, then error propagation to combine into class errors) and recorded in a Pull distribution as described in section 1.4.3. If errors were predicted correctly with no bias then the standard deviation of the Pull distribution around zero should equal unity, i.e. the mean ratio of observed to predicted errors should be one. If the spread of the errors were predicted correctly but with a bias then the standard deviation of the Pull distribution about the mean should equal unity, but the standard deviation about zero would be larger, i.e. the predicted errors would describe the repeatability of the measurements, but the mean difference between true and estimated values would be non-zero.

Figure 11 plots the standard deviations of the Pull distributions about their means for each histogram size and relative quantity of testing data. This plot shows measurement repeatability, excluding any possible bias. Figure 12 plots the standard deviations of the Pull distributions about their zero points, including the effects of any bias. If successful, the data points within both of these plots would cluster around unity within statistical

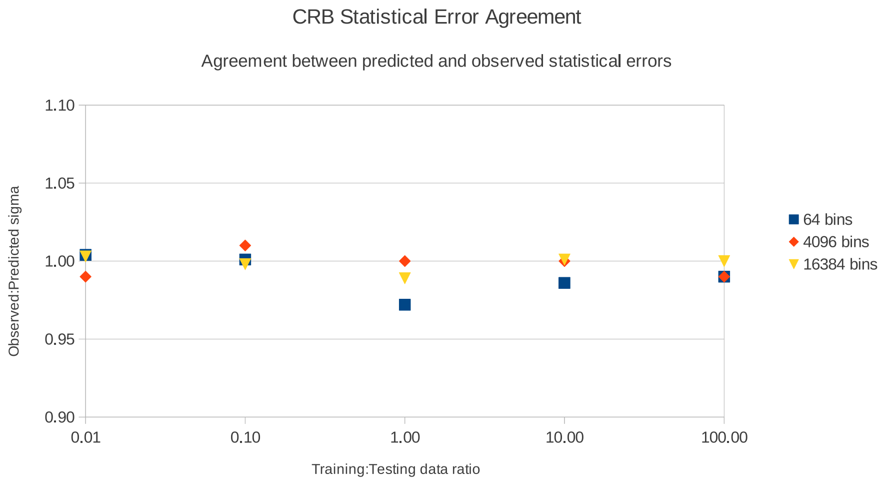


Figure 11: This plot compares CRB error estimates to the observed spread of estimated quantities for varying ratios of training to testing data. The y-axis shows the ratio of observed to predicted errors.

limits of computing the Pull distributions, i.e. +/- 0.022 from equation (6) of section 1.4.3.

4.4 Discussion

Figure 11 shows that the CRB performs well as a predictor of statistical fluctuations in estimated quantities. For all models and quantities of data tested there was high agreement between widths of estimated quantity distributions (around mean values) and those predicted. However, as a predictor of errors around true quantities the CRB performs increasingly poorly as the relative quantity of testing data grows. Figure 12 shows that the CRB becomes unusable as a predictor of true errors once the quantity of incoming data exceed that which was used during training.

The discrepancy between the spread of estimated quantities and residuals from true values can be explained by systematic effects present in trained PMFs. These systematic effects were first noted in section 3.6.2, where the goodness-of-fit function of equation (45) had to be modified to account for modelling errors. The consequence of these systematic effects are to bias the Likelihood estimation process, as Likelihood assumes that the model is error free. If the CRB method is to be used for constructing error bars on summary outputs for practical applications then the quantity of data analysed must be kept small. This limitation is too restrictive for the large scale analysis of planetary images suggesting either the histogram based model or error estimation approach should be reconsidered.

A histogram is the most generic form of representation for a statistical model and has been justified with reference to the properties of planetary images. However, it is also the least accurate form of model due to the large number of parameters which must to be estimated, i.e. each individual bin frequency. This equally applies to their unit normalised

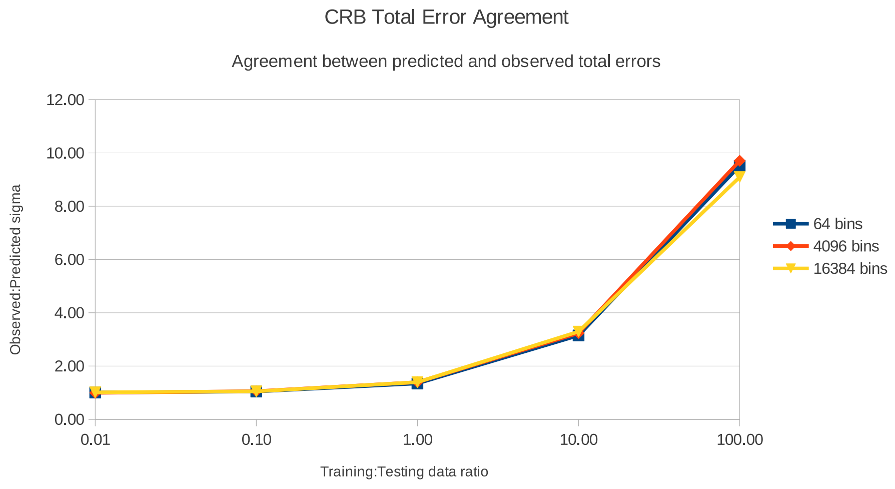


Figure 12: This plot compares CRB error estimates to the observed total error from true values on estimated quantities for varying ratios of training to testing data. The y-axis shows the ratio of observed to predicted errors.

counterparts - the PMFs which are used as model components. A parametric alternative, such as a Gaussian mixture model, requires far fewer parameters (mean, standard deviation and scaling) per component, potentially providing far more accurate models given the same number of data points. If a parametric model could be estimated to high levels of accuracy then CRB error predictions may be sufficient by themselves, or at least any systematic effects could be minimised. However, this is predicated upon finding a valid parametric form, understanding its error characteristics and then corroborating its appropriateness on test data, all with no guarantee of success. Arguably, it is better to develop an understanding of the worst-case systematic errors, allowing them to be quantified and factored in to error predictions, rather than seeking a new model. The next chapter will develop this latter option. For quantitative use it is sensible to provide this worst-case error estimate, as it is less likely to lead to over-interpretation of measurements. Unless it is shown that greater accuracy is required then there is no justification for increasing the complexity of the model at this stage.

4.5 Summary

A standard method for estimating errors on maximum Likelihood computed parameters has been applied to the non-parametric linear Poisson models with some success. The repeatability of estimated parameters, i.e. spread around mean estimates, has been shown to be effectively predicted using the CRB. However, systematic effects caused by sampling noise during training precludes the use of the CRB alone as a predictor of errors around true quantities.

The next chapter will provide a more comprehensive analysis, incorporating statistical and systematic errors, to widen the scope of the linear modelling method to larger relative quantities of data.

5 Systematic Error Estimation

The previous chapter provided a method for predicting statistical errors on quantities of classes, $\mathbb{Q}$, estimated for linear Poisson histogram models. It was shown that a combination of the Cramer Rao Bound and error propagation could be used to compute a covariance matrix, $\mathbf{C}_{\mathbb{Q}}$, which successfully predicted the spread of estimated quantities around mean estimated values. However, total deviations around true values were seen to be greater than predicted due to biases caused by systematic modelling errors. This chapter will address the issue of systematic errors in model components by analysing how sampling noise introduced during training propagates though the EM algorithm. It will also be shown that the same approach can be used to give an alternative to the CRB for quantifying the effects of statistical errors from incoming data. Monte-Carlo testing will be used to corroborate the error estimation theory for a range of quantities and distributions.

5.1 EM Error propagation

The CRB method presented in the previous chapter might be considered as a top-down approach to error estimation though an analysis of the Likelihood function being optimised. This chapter applies a bottom-up approach by first identifying the underlying sources of uncertainty then approximating, to first order, the effects small changes in those sources have on estimated quantities through the repeated application of the EM algorithm. This approach involves the following steps:

- identify the sources of uncertainty in inputted data;
- approximate the effect those sources of uncertainty have on the EM update function;
- approximate the amplification of errors during the iterative convergence of the EM algorithm.

These three steps will be explored in the following sections.

5.1.1 Sources of uncertainty

Two sources of uncertainty can be identified in the linear Poisson histogram models described in chapter 3. These are:

1. Poisson sampling noise contained within exemplar training histograms;
2. and Poisson sampling noise found in incoming data under analysis.

The EM ICA training algorithm takes a set of sampled histograms as input. From these histograms a set of linear components are extracted approximating the underlying data generating distributions, $R(X|k)$, which are contaminated by sampling noise, as explained in section 3.6.2. The normalised versions of these distributions, $P(X|k)$, are computed in equation (37) forming the basis of the linear model components. The noise in these components is believed to be the cause of the biased quantity estimates observed in the previous chapter. This will be referred to as the systematic error. The extracted component distributions, $R(X|k)$, will be treated as sampled histograms in the following sections, which is permissible given their origin and Poisson behaviour. For compactness and to emphasise their histogram nature they will be denoted:

$$\mathbf{H}_{X|k} = R(X|k) \tag{59}$$

where $\mathbf{H}$ is a histogram vector; X is a bin within the histogram; and k is the component the histogram describes.

Whilst not the main focus of this chapter, statistical errors will also be included for completeness. The sampling noise in incoming histograms are the cause of statistical fluctuations in estimated quantities which were successfully predicted using the CRB in the previous chapter. The bottom-up approach to error estimation in this chapter will be shown to be consistent with the previous CRB method. The incoming data histogram will continue to be denoted by $\mathbf{H}_X$.

5.1.2 Single EM step error

For the purposes of error propagation it is convenient to consider the EM update function in terms of a single histogram bin, X, and every other bin which is not X, which will be denoted $\bar{X}$, e.g. $\mathbf{H}_{\bar{X}} = \sum_{Y \neq X} \mathbf{H}_Y$. It is also convenient to consider components in a similar notation using k and all other components, $\bar{k}$, e.g. $\mathbf{Q}_{\bar{k}} = \sum_{l \neq k} \mathbf{Q}_l$. This grouping of terms will assist in the computation of derivatives. The update function can then be stated in terms of incoming histogram, $\mathbf{H}_X$, (data) and extracted components, $\mathbf{H}_{X|k}$, (model) bins giving:

$$\mathbf{Q}'_k = P(k|X)\mathbf{H}_X + P(k|\bar{X})\mathbf{H}_{\bar{X}} \tag{60}$$

$$= \frac{\left(\frac{\mathbf{H}_{X|k}\mathbf{Q}_k}{\mathbf{H}_{X|k}+\mathbf{H}_{\bar{X}|k}}\right)\mathbf{H}_X}{\left(\frac{\mathbf{H}_{X|k}\mathbf{Q}_k}{\mathbf{H}_{X|k}+\mathbf{H}_{\bar{X}|k}} + \frac{\mathbf{H}_{X|\bar{k}}\mathbf{Q}_{\bar{k}}}{\mathbf{H}_{X|\bar{k}}+\mathbf{H}_{\bar{X}|\bar{k}}}\right)} + \frac{\left(\frac{\mathbf{H}_{\bar{X}|k}\mathbf{Q}_k}{\mathbf{H}_{X|k}+\mathbf{H}_{\bar{X}|k}}\right)\mathbf{H}_{\bar{X}}}{\left(\frac{\mathbf{H}_{\bar{X}|k}\mathbf{Q}_k}{\mathbf{H}_{X|k}+\mathbf{H}_{\bar{X}|k}} + \frac{\mathbf{H}_{\bar{X}|\bar{k}}\mathbf{Q}_{\bar{k}}}{\mathbf{H}_{X|\bar{k}}+\mathbf{H}_{\bar{X}|\bar{k}}}\right)} \tag{61}$$

where $\mathbf{Q}'$ is the updated quantity; and $\mathbf{Q}$ is the previous quantity. Uncertainty from the two sources of error can be propagated through this update function by considering how small changes in the inputs affect the estimated vector of quantities. As the two sources are independent their contributions to the covariance can be derived separately then summed:

$$\mathbf{C}_{EMStep} = \mathbf{C}_{data} + \mathbf{C}_{model} \tag{62}$$

$$\mathbf{C}_{ij(data)} = \sum_X \left[\left(\frac{\partial \mathbf{Q}'_i}{\partial \mathbf{H}_X} \right) \left(\frac{\partial \mathbf{Q}'_j}{\partial \mathbf{H}_X} \right) \sigma^2_{\mathbf{H}_X} \right] \tag{63}$$

$$\mathbf{C}_{ij(model)} = \sum_X \left[\sum_k \left(\frac{\partial \mathbf{Q}'_i}{\partial \mathbf{H}_{X|k}} \right) \left(\frac{\partial \mathbf{Q}'_j}{\partial \mathbf{H}_{X|k}} \right) \sigma^2_{\mathbf{H}_{X|k}} \right] \tag{64}$$

where $\mathbf{C}_{data}$ is the statistical contribution from the incoming histogram data; and $\mathbf{C}_{model}$ is the systematic contribution from the training exemplar histograms used to construct the component models.

The statistical contribution is straightforward giving:

$$\mathbf{C}_{ij(data)} = \sum_X P(i|X)P(j|X)\mathbf{H}_X \tag{65}$$

In contrast, the systematic contribution involves relatively complex derivatives and is best approached in parts. The derivative of a quantity with respect to an exemplar histogram bin can be divided into two independent terms:

$$\frac{\partial \mathbf{Q}'_j}{\partial \mathbf{H}_{X|k}} = \frac{\partial P(j|X)\mathbf{H}_X}{\partial \mathbf{H}_{X|k}} + \frac{\partial P(j|\bar{X})\mathbf{H}_{\bar{X}}}{\partial \mathbf{H}_{X|k}} \tag{66}$$

Defining the total quantity of training data for component k as $\mathbf{T}_k = \mathbf{H}_{X|k} + \mathbf{H}_{\bar{X}|k}$, in the cases where $j = k$ the two terms are given by

$$\begin{aligned}
\frac{\partial P(k|X)\mathbf{H}_X}{\partial \mathbf{H}_{X|k}} &= \frac{\left(\frac{\mathbf{H}_{X|k}\mathbf{Q}_k}{\mathbf{T}_k} + \frac{\mathbf{H}_{X|\bar{k}}\mathbf{Q}_{\bar{k}}}{\mathbf{T}_{\bar{k}}} \right) \frac{P(\bar{X}|k)\mathbf{Q}_k}{\mathbf{T}_k} - \left(\frac{\mathbf{H}_{X|k}\mathbf{Q}_k}{\mathbf{T}_k} \right) \frac{P(\bar{X}|k)\mathbf{Q}_k}{\mathbf{T}_k}}{\left(\frac{\mathbf{H}_{X|k}\mathbf{Q}_k}{\mathbf{T}_k} + \frac{\mathbf{H}_{X|\bar{k}}\mathbf{Q}_{\bar{k}}}{\mathbf{T}_{\bar{k}}} \right)^2} \mathbf{H}_X \\
&= \frac{P(X|\bar{k})\mathbf{Q}_{\bar{k}}P(\bar{X}|k)\mathbf{Q}_k\mathbf{H}_X}{\mathbf{T}_k[P(X|k)\mathbf{Q}_k + P(X|\bar{k})\mathbf{Q}_{\bar{k}}]^2} \times \frac{P(X|k)}{P(X|k)} \\
&= \frac{P(k|X)P(\bar{k}|X)P(\bar{X}|k)\mathbf{H}_X}{\mathbf{T}_k P(X|k)}
\end{aligned} \tag{67}$$

and

$$\begin{aligned}
\frac{\partial P(k|\bar{X})\mathbf{H}_{\bar{X}}}{\partial \mathbf{H}_{X|k}} &= \frac{-\left(\frac{\mathbf{H}_{\bar{X}|k}\mathbf{Q}_k}{\mathbf{T}_k} + \frac{\mathbf{H}_{\bar{X}|\bar{k}}\mathbf{Q}_{\bar{k}}}{\mathbf{T}_{\bar{k}}} \right) \frac{P(\bar{X}|k)\mathbf{Q}_k}{\mathbf{T}_k} + \left(\frac{\mathbf{H}_{\bar{X}|k}\mathbf{Q}_k}{\mathbf{T}_k} \right) \frac{P(\bar{X}|k)\mathbf{Q}_k}{\mathbf{T}_k}}{\left(\frac{\mathbf{H}_{\bar{X}|k}\mathbf{Q}_k}{\mathbf{T}_k} + \frac{\mathbf{H}_{\bar{X}|\bar{k}}\mathbf{Q}_{\bar{k}}}{\mathbf{T}_{\bar{k}}} \right)^2} \mathbf{H}_{\bar{X}} \\
&= -\frac{\mathbf{H}_{\bar{X}|\bar{k}}\mathbf{Q}_{\bar{k}}P(\bar{X}|k)\mathbf{Q}_k\mathbf{H}_{\bar{X}}}{\mathbf{T}_{\bar{k}}\mathbf{T}_k[P(\bar{X}|k)\mathbf{Q}_k + P(\bar{X}|\bar{k})\mathbf{Q}_{\bar{k}}]^2} \\
&= -\frac{P(\bar{k}|\bar{X})P(k|\bar{X})\mathbf{H}_{\bar{X}}}{\mathbf{T}_k}
\end{aligned} \tag{68}$$

giving

$$\frac{\partial \mathbf{Q}'_k}{\partial \mathbf{H}_{X|k}} = \frac{P(k|X)P(\bar{k}|X)P(\bar{X}|k)\mathbf{H}_X}{\mathbf{T}_k P(X|k)} - \frac{P(\bar{k}|\bar{X})P(k|\bar{X})\mathbf{H}_{\bar{X}}}{\mathbf{T}_k} \tag{69}$$

In the case where $j \neq k$ the same terms become

$$\begin{aligned}
\frac{\partial P(j|X)\mathbf{H}_X}{\partial \mathbf{H}_{X|k}} &= \frac{-\left(\frac{\mathbf{H}_{X|j}\mathbf{Q}_j}{\mathbf{T}_j}\right)\frac{P(\bar{X}|k)\mathbf{Q}_k}{\mathbf{T}_k}}{\left(\frac{\mathbf{H}_{X|k}\mathbf{Q}_k}{\mathbf{T}_k}+\frac{\mathbf{H}_{X|\bar{k}}\mathbf{Q}_{\bar{k}}}{\mathbf{T}_{\bar{k}}}\right)^2}\mathbf{H}_X \\
&= -\frac{P(X|j)\mathbf{Q}_j P(\bar{X}|k)\mathbf{Q}_k\mathbf{H}_X}{\mathbf{T}_k[P(X|k)\mathbf{Q}_k + P(X|\bar{k})\mathbf{Q}_{\bar{k}}]^2} \times \frac{P(X|k)}{P(X|k)} \\
&= -\frac{P(j|X)P(k|X)P(\bar{X}|k)\mathbf{H}_X}{\mathbf{T}_k P(X|k)}
\end{aligned} \tag{70}$$

and

$$\begin{aligned}
\frac{\partial P(j|\bar{X})\mathbf{H}_{\bar{X}}}{\partial \mathbf{H}_{X|k}} &= \frac{P(\bar{X}|j)\mathbf{Q}_j P(\bar{X}|k)\mathbf{Q}_k\mathbf{H}_{\bar{X}}}{\mathbf{Q}_k[P(\bar{X}|k)\mathbf{Q}_k + P(\bar{X}|\bar{k})\mathbf{Q}_{\bar{k}}]^2} \\
&= \frac{P(j|\bar{X})P(k|\bar{X})\mathbf{H}_{\bar{X}}}{\mathbf{Q}_k}
\end{aligned} \tag{71}$$

giving

$$\frac{\partial \mathbf{Q}'_j}{\partial \mathbf{H}_{X|k}} = \frac{P(j|\bar{X})P(k|\bar{X})\mathbf{H}_{\bar{X}}}{\mathbf{Q}_k} - \frac{P(j|X)P(k|X)P(\bar{X}|k)\mathbf{H}_X}{\mathbf{T}_k P(X|k)} \tag{72}$$

Substituting these results back into the covariance calculation provides an estimate of the error on quantities after performing a single EM step.

5.1.3 Error amplification

If the EM algorithm is seeded with the ground truth, i.e. $\mathbf{Q}$ is set to match the proportions of the data generating processes which actually generated incoming data, then noise in the model and data will cause convergence to occur some distance away from the ground truth. Beginning at the true values for the prior quantities, an iteration of the EM update function will cause these values to change. This error will bias the next and subsequent estimates amplifying the initial effect, making the final error potentially much larger. In general, the true values with which to seed the EM algorithm are not available, but the convergence point (assuming there are no local minima) will be the same distance away from the true values irrespective of where the algorithm started. This insight supports a possible approach for estimation of the deviation. This section will approximate the deviation using a convergent geometric series, and so generate a linear approximation for the amplification process, suitable for use in error propagation.

If $\boldsymbol{\Delta}$ is a vector of quantity errors evaluated at a particular time, t, then the accumulation of error from one step to the next is given by

$$\mathbf{\Delta}|_t \approx (\mathbf{\Delta}|_{t-1})^T \nabla \mathbf{Q}|_{t-1} \tag{73}$$

where $\nabla \mathbf{Q}$ is the Jacobian

$$\nabla \mathbf{Q}_{ij}|_{t-1} = \frac{\partial \mathbf{Q}_i|_{t-1}}{\partial \mathbf{\Delta}_j} \tag{74}$$

The diagonal terms of the Jacobian are given by

$$\nabla \mathbf{Q}_{ii} = \frac{\partial \mathbf{Q}_i}{\partial \mathbf{\Delta}_i} \tag{75}$$

$$= \sum_X \frac{P(X|i)[P(X|i)\mathbf{Q}_i + P(X|\bar{i})\mathbf{Q}_{\bar{i}}] - P(X|i)^2 \mathbf{Q}_i}{[P(X|i)\mathbf{Q}_i + P(X|\bar{i})\mathbf{Q}_{\bar{i}}]^2} \mathbf{H}_X \tag{76}$$

where $P(X|i)\mathbf{Q}_i + P(X|\bar{i})\mathbf{Q}_{\bar{i}} \approx \mathbf{H}_X$ giving

$$= \sum_X \frac{P(X|i)\mathbf{H}_X - P(X|i)^2 \mathbf{Q}_i}{\mathbf{H}_X^2} \mathbf{H}_X \tag{77}$$

$$= \sum_X P(X|i) - \frac{P(X|i)^2 \mathbf{Q}_i}{\mathbf{H}_X} \tag{78}$$

which via Bayes Theorem becomes

$$\nabla \mathbf{Q}_{ii} = \sum_X P(X|i) - P(X|i)P(i|X) \tag{79}$$

Similar treatment gives the off-diagonal terms

$$\nabla \mathbf{Q}_{ij} = \frac{\partial \mathbf{Q}_i}{\partial \mathbf{\Delta}_j} \tag{80}$$

$$= \sum_X \frac{-P(X|i)\mathbf{Q}_i P(X|j)}{[P(X|j)\mathbf{Q}_j + P(X|\bar{j})\mathbf{Q}_{\bar{j}}]^2} \mathbf{H}_X \tag{81}$$

$$= \sum_X \frac{-P(X|i)\mathbf{Q}_i P(X|j)}{\mathbf{H}_X^2} \mathbf{H}_X \tag{82}$$

$$= \sum_X -P(X|j) \frac{P(X|i)\mathbf{Q}_i}{\mathbf{H}_X} \tag{83}$$

i.e.

$$\nabla \mathbf{Q}_{ij} = \sum_X -P(X|j)P(i|X) \tag{84}$$

Assuming the derivative computed at any time is approximately equal, i.e. $\frac{\partial \mathbf{Q}_i|_{t-1}}{\partial \mathbf{\Delta}_j} \approx \frac{\partial \mathbf{Q}_i|_t}{\partial \mathbf{\Delta}_j}$, such that for all t then $\nabla \mathbf{Q}|_t \approx \nabla \mathbf{Q}$, then the error accumulation from the first step onwards becomes

$$\mathbf{\Delta}|_1 \approx \mathbf{\Delta}|_0 \nabla\mathbf{Q} \tag{85}$$

$$\mathbf{\Delta}|_2 \approx \mathbf{\Delta}|_1 \nabla\mathbf{Q} \approx \mathbf{\Delta}|_0 \nabla\mathbf{Q}^2 \tag{86}$$

$$\mathbf{\Delta}|_t \approx \mathbf{\Delta}|_0 \nabla\mathbf{Q}^t \tag{87}$$

The total vector change in the quantity $\mathbf{Q}$ is then given by

$$\mathbf{\Delta} \approx \sum_{t=0}^{\infty} \mathbf{\Delta}|_t = \mathbf{\Delta}|_0 + \sum_{t=1}^{\infty} \mathbf{\Delta}|_0 \nabla\mathbf{Q}^t \tag{88}$$

$$\mathbf{\Delta} \approx \mathbf{\Delta}^T|_0 \left[\mathbf{I} + \sum_{t=1}^{\infty} \nabla\mathbf{Q}^t\right] \;=\; \mathbf{\Delta}^T|_0 \left[\mathbf{I} - \nabla\mathbf{Q}\right]^{-1} \tag{89}$$

so that the total error amplification can be approximated by a single linear process

$$\mathbf{\Delta} \approx \mathbf{\Delta}^T|_0 \mathbf{A} \tag{90}$$

where the amplification matrix is

$$\mathbf{A} = \left[\mathbf{I} - \nabla\mathbf{Q}\right]^{-1} \tag{91}$$

Finally, the covariance matrix for the quantity vector can be given by scaling the one step covariance by the amplification matrix:

$$\mathbf{C} = \mathbf{A}^{\intercal} \mathbf{C}_{EMStep} \mathbf{A} \tag{92}$$

This provides the component covariance, which can be used to compute the class covariance, $\mathbf{C}_{\mathbb{Q}}$ using equation (58).

Whilst this formulation looks considerably different from the original CRB covariance estimate the two are consistent. It will be shown numerically that when model error terms are all zero the error propagation version described above produces equivalent results to the CRB version (figure 15). Importantly, this confirms that the approach taken provides an appropriate approximation of the EM amplification effect.

5.2 Monte-Carlo testing

Tests were performed to ensure that the new error predictions could successfully estimate measurement errors, appropriately accounting for both statistical and systematic effects. Tests were also performed to confirm that the error propagation approach to computing the statistical component of the error was equivalent to the previous CRB version.

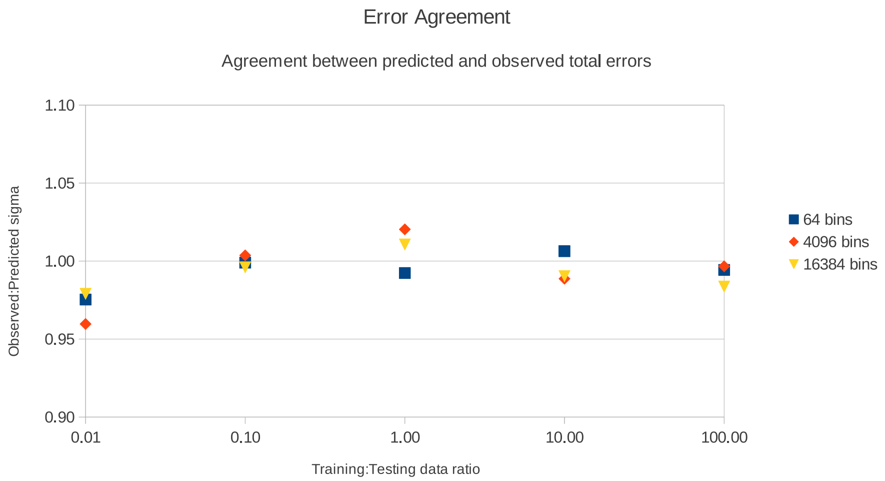


Figure 13: This plot compares total error estimates, incorporating both systematic and statistical sources of uncertainty, with observed errors around true values for varying ratios of training to testing data. The y-axis shows the ratio of observed to predicted errors. Note the slight over-estimated errors at 0.01.

To ensure consistency with previous experiments the histograms, components, classes and quantities used in the previous chapter (section 4.3) were reused following the same methodology. The experiments conducted previously were repeated for the error propagation approach, first including both statistical and systematic error predictions, then only including the statistical errors.

The ratio of observed to predicted total errors (stat + sys) can be seen in figure 13, which can be directly compared to figure 12 of chapter 4. The change in relative contributions from statistical and systematic effects to the total error as the ratio of training to testing data increases can be seen in figure 14. Finally, the relationship between statistical errors computed using the CRB verses the same errors computed using error propagation can be seen in figure 15.

5.3 Discussion

Total errors around true quantities were successfully predicted by the bottom-up error propagation approach. Figure 13 shows close agreement between predicted and observed deviations from ground truth, in contrast to figure 12 from the previous chapter where the CRB failed to account for additional errors caused by inaccuracies in model components. The marginal over-estimation of errors at very low quantities of data might be explained by the truncation of assumed Gaussian noise at zero, as negative estimates are not possible. The contribution to the total error from systematic effects grows with the quantity of data being analysed. Figure 14 shows that systematic errors grow to dominate the uncertainty in estimated quantities when the amount of incoming data exceeds that which was used during training. An inspection of the covariance calculations reveals that statistical errors

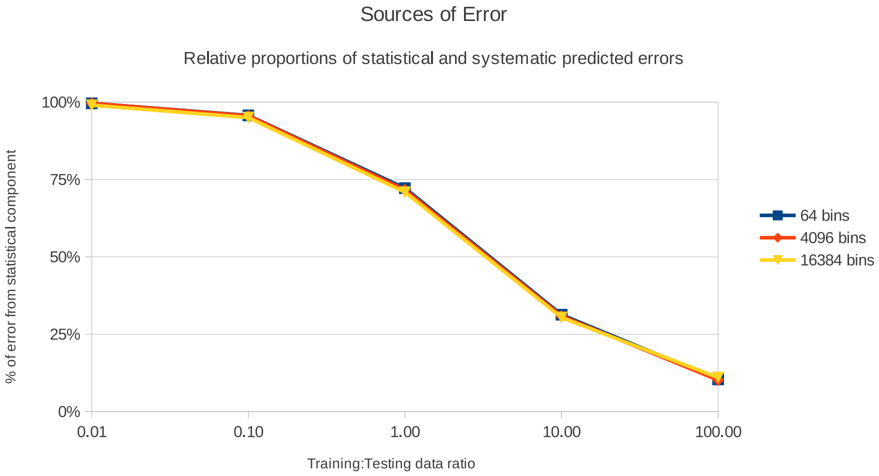


Figure 14: This plot shows how contributions from statistical and systematic sources to the total error change in proportion as the relative quantity of testing data increases. The contribution from the statistical component is below the curve.

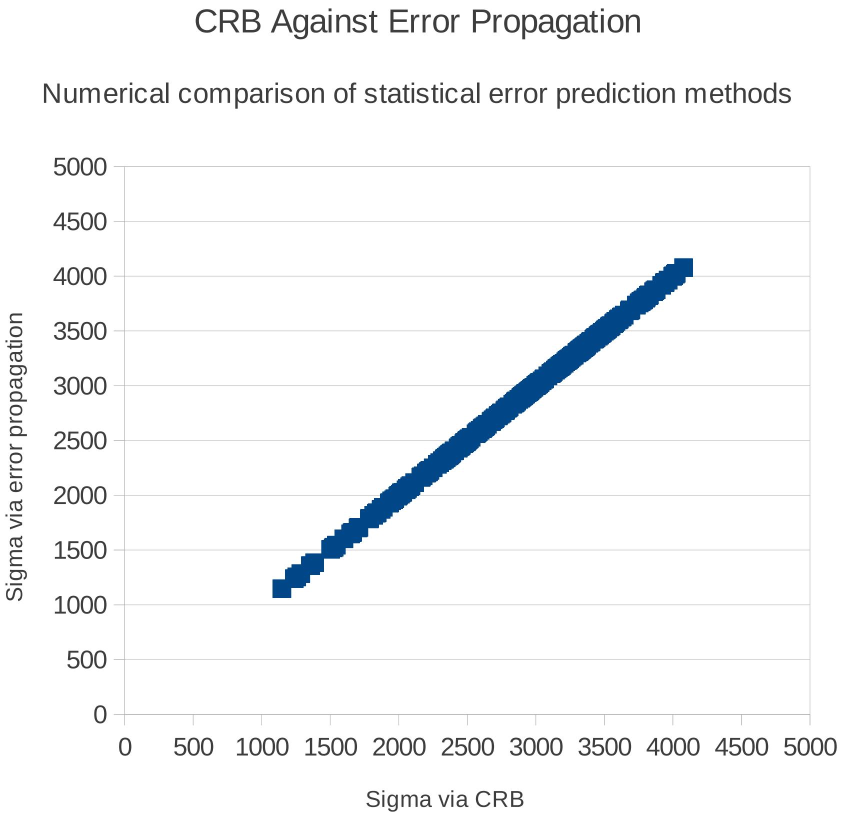


Figure 15: The plot shows the equivalence between the CRB predicted statistical error and the error propagation alternative over a range of quantities.

grow with the square-root of estimated quantities, whereas systematic components grow linearly.

The bottom-up approach is more complex than the CRB method, but can be seen to be numerically equivalent for predicting statistical errors. Figure 15 confirms, over 300 trials, that errors estimated using the CRB match statistical errors computed using error propagation on the EM algorithm. Whilst the approach presented in this chapter has a wider domain of applicability, for applications involving small relative quantities of data it may be preferable to implement the simpler CRB method.

The Likelihood estimation process with goodness-of-fit validation, combined with an understanding of the statistical and systematic errors on estimated quantities, constitutes (in theory) a completely quantitative system. The Monte-Carlo simulations conducted during chapters 3 and 4, and those undertaken in this chapter show that the quantitative criteria specified during the introductory chapter, section 1.6, have been fulfilled. Assuming the behaviour of real data follows the properties given in section 3.2, which were the basis of Monte-Carlo data, then the methods presented so far should be capable of making quantitative measurements from planetary images.

It was noted in section 4.4 that a histogram model was likely to be the least accurate, giving the worst systematic errors, due to the large number of parameters estimated during training. This worst-case could be improved upon, given an appropriate parametric alternative. However, this thesis will continue to use the simple non-parametric model. But, as a challenge to the pattern recognition and machine learning communities, a detailed theoretical analysis of errors should be undertaken for alternative statistical models (parametric or otherwise). Moreover, those analyses should include empirical corroboration to ensure that theoretically more accurate models are capable of describing the data, i.e. they should have their own goodness-of-fit criterion and produce error predictions matching observed error distributions. Some alternatives which could be investigated include models constructed from different linear ICA algorithms (with alternative cost functions), those which use kernel methods to address non-linearities (e.g. Kernel PCA, Gaussian Process Latent Variable Models etc.) and those which form fixed decision boundaries rather than data densities (e.g. SVM, Boosting, Random Forests etc.). Only with such an understanding of errors can these alternatives meet the quantitative criteria set by this thesis.

5.4 Summary

This chapter has explained how errors around true quantities can be predicted by considering underlying sources of error, including modelling errors caused by sampling noise during training and incoming sampling noise from histograms being analysed. These sources of error were propagated through the EM update function, then amplified to account for the iterative nature of the EM algorithm. The resulting error predictions, corroborated by

Monte-Carlo studies, have provided a sufficiently detail understanding of linear Poisson histogram models to allow quantitative measurements to be taken from real data, assuming that data behaves as expected. The remaining chapters of this thesis will work towards applying the methods developed thus far to increasingly realistic data.

PART 2: APPLICATION

The chapters found within this part of the thesis will: demonstrate how martian terrains can be measured; show the filtering of citizen science crater data; and draw overall conclusions.

Supporting material, including preliminary work and publications generated from this part, include:

- P.D. Tar, N.A. Thacker, M.A. Jones and J.D. Gilmour, A Quantitative Approach to the Analysis of Planetary Terrains, Proc. Remote Sensing and Photogrammetry Soc. Conf., 2012

- P.D. Tar, N.A. Thacker, J.D. Gilmour and M.A. Jones, Automated Quantitative Planetary Measurements, Proc. European Planetary Science Congress, 2013

- P.D. Tar, N.A. Thacker, Coalescence and refinement of Moon Zoo crater annotations, Proc. European Planetary Science Congress, 2014

- P.D. Tar, N.A. Thacker, Quantification of false positives within Moon Zoo crater annotations, Proc. European Planetary Science Congress, 2014

- P.D. Tar, N.A. Thacker, J.D. Gilmour and M.A. Jones, Automated Quantitative Measurements and Associated Error Covariances for Planetary Image Analysis, Preprint submitted to Advances in Space Research

- P.D. Tar, The Application of Appearance Models to Martian Impact Craters, Internal Memo, 2010-011, www.tina-vision.net

- P.D. Tar, N.A. Thacker, A Quantitative Representation for the Segmentation of Martian Images, Internal Memo, 2011-002, www.tina-vision.net

- P.D. Tar, N.A. Thacker, Towards a Quantitative Analysis of Martian Terrains, Internal Memo, 2011-005, www.tina-vision.net

- P.D. Tar, N.A. Thacker, A Connected Blob Image Representation for Poisson Linear Models, Internal Memo, 2012-006, www.tina-vision.net

6 Martian Terrains: BRIEF Representation

A statistical model for planetary image data based upon linearly combined histograms has been developed over the previous chapters. An error theory has also been included for predicting the accuracies with which model parameters can be estimated. It was argued that these parameters could be linked to measurements sought from surface imagery, so the estimation of parameters, and their errors, could constitute a quantitative measurement system satisfying the criteria of section 1.6. However, this assumed an appropriate encoding could be found for representing patterns within data that had the properties stated in section 3.2. Image encoding schemes must now be investigated and their statistical properties analysed to ensure that they are appropriate as input for the algorithms described thus far.

The wide ranging science applications noted in section 1.2.2 suggest the need for a generic method for analysing varied martian terrains. To accommodate this a texture-based approach will be adopted allowing arbitrary terrains to be decomposed in terms of repeating local patterns. This chapter proposes a simple representation for local image structures which can be used to populate linear histogram models with texture information. The representation will conveniently provide a one-to-one correspondence between model weighting quantities and surface area measurements, from which many other measurements can be derived. The performance of the representation will be examined using synthetic martian terrain images created from real martian data. A martian terrain simulator tool will be developed allowing unlimited quantities of test data to be generated with known ground-truths. A range of experiments will reveal the limitations of the image encoding, resulting in a better understanding of the properties of planetary images and an appreciation of the difficulties involved in making statistical theories operate successfully in practice.

6.1 Local BRIEF descriptors

Several image representations were discussed in section 2.1 of the literature review providing numerous possible encodings from which histograms could be constructed. An appropriate representation must be readily adaptable to provide small pattern vectors, X, forming the independent Poisson histogram bins expected by the model. One such representation is BRIEF [92], which has been shown to be effective in object matching tasks and is scalable to small patches allowing local structures to be encoded in short binary strings. Inspired by BRIEF, the representation adopted in this chapter is based upon circular local image descriptors consisting of η pixel pairs selected within a ι pixel radius. The endpoints of each pair are initially selected at random as a set of offsets relative to the origin of the sampling disc. All subsequent descriptors use this definition of pairs providing a deterministic way of sampling pixels from source images. From the

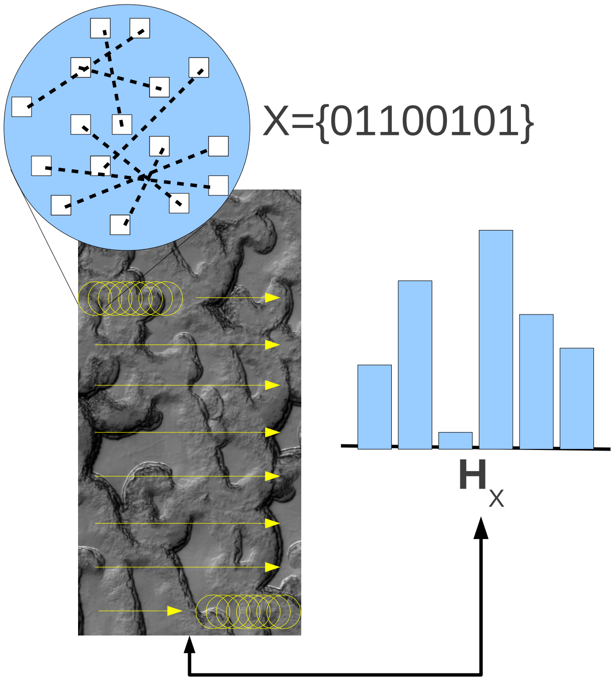


Figure 16: This illustration shows an 8 bit local BRIEF descriptor constructed by comparing pairs of pixels within a circular patch. The origin of the patch is shifted across the image one pixel at a time until every pixel has acted as the BRIEF descriptor's origin. The model histogram, H_X, is populated with the frequency of occurrence of each type of BRIEF pattern.

pixel pairs, $[(\alpha_1, \beta_1), \ldots, (\alpha_\eta, \beta_\eta)]$, an η-bit binary pattern is constructed, $X = [x_1, \ldots, x_\eta]$, where a bit, x_i, is set if the corresponding α_i pixel is brighter than the corresponding β_i pixel by a given threshold, i.e. $x_i = \delta(\alpha_i - \beta_i > \theta)$. This representation is BRIEF-like, differing only by the introduction of a threshold which may be interpreted as a statistical hypothesis test on the brightness difference in comparison to image noise, avoiding the need to smooth images which entails loss in image detail. Populating a histogram then proceeds by computing overlapping BRIEF descriptors, exhaustively covering each location in an image by shifting the origin of the BRIEF sampling patch one pixel at a time. Each unique pattern can be assigned to a histogram bin with the frequency of occurrence of each pattern recorded. Figure 16 illustrates this process.

Besides the simplicity and convenience of the BRIEF encoding, there are other reasons for believing this style of representation is appropriate for planetary terrain images. It was shown in [93] that sets of pair-wise comparisons can be used to reconstruct an underlying function up to a rank-order, therefore a sufficiently dense set of pixel pairs can be used as a complete representation of local patches. Also, relative pixel comparisons discard absolute pixel values, which in planetary images are subject to illumination effects and local albedo, therefore the encoding gives a level of invariance and reduces the number of patterns which must be learned. Finally, local BRIEF descriptors provide a one-to-one mapping between model quantities, $\mathbf{Q}$, and surface area measurements, as each descriptor can be associated with a single pixel at the origin of the circular sampling patch.

It is acknowledged that the encoding scheme selected is just one of many possibilities. What is important is that whichever representation is used, the patterns produced have the necessary properties for the histogram models to work correctly. The analysis of errors in quantity estimation (section 4.1.2) implies theoretically that the performance of a representation is determined by its ability to produce the least ambiguous patterns. The difference between one appropriate representation[6]and another should only manifest in larger or smaller errors on quantity measurements, but the quantitative criteria of section 1.6 does not insist upon best possible accuracy. It is reasonable to first seek a representation that can be used quantitatively, and only if it is shown later that greater accuracy is required should effort be spent finding a more sophisticated solution.

6.2 Generating test data: Martian terrain simulator

The Monte-Carlo testing of previous chapters compared predicted errors to empirically observed error distributions by repeating measurements over thousands of independent trials over wide ranges of quantities. Performing such tests on martian terrain data would require prohibitively large quantities of ground-truth. A practical and highly flexible alternative is described below which can synthesize arbitrarily composed terrains using samples from real martian data.

The strategy for generating large quantities of independent martian terrains with known ground-truth can be summarised by the following points:

1. Gather large images containing examples of martian terrains;
2. Using smoothing kernels with different widths, separate the low and high frequency spatial components of the terrain images;
3. Divide the example images into many small tiles;
4. Provide a labelled template showing the desired layout for a synthetic martian image;
5. Randomly select tiles (with replacement) for the terrain types prescribed by the template;
6. Slightly perturb the shape of each tile through stretching;
7. Following the template, form a composite image using the tile's high-frequency spatial components;
8. Add uniform Gaussian noise to pixel values in the high-frequency image;

[6] An appropriate representation is one which provides independent Poisson bins which can be linearly combined. A representation which does not provide these properties is inappropriate for the method and should not be used at all.

Dataset A:	(0) EPS 023675 0930	(1) EPS 024889 2605	(2) EPS 024926 2525
Dataset B:	(0) EPS 017810 1850	(1) EPS 019243 2550	(2) EPS 024984 2610
Dataset C:	(0) EPS 023661 0931	(1) EPS 017866 2855	(2) EPS 024979 2550
Dataset D:	(0) EPS 023738 0915	(1) EPS 023744 1775	(2) EPS 024927 2555
Dataset E:	(0) EPS 020558 0930	(1) EPS 022543 0950	(2) EPS 024949 2440
Dataset F:	(0) EPS 023767 0925	(1) EPS 024662 2555	(2) EPS 024950 1870
Dataset G:	(0) EPS 017866 2855	(1) EPS 019104 1740	(2) EPS 024899 2540
Dataset H:	(0) EPS 023729 0935	(1) EPS 022882 2030	(2) EPS 024991 2540
Dataset I:	(0) EPS 023621 0970	(1) EPS 023676 0925	(2) EPS 024983 2160
Dataset J:	(0) EPS 018948 2250	(1) EPS 024724 1760	(2) EPS 024883 1525

Table 2: HiRISE datasets used in testing

9. Following the template, form a composite image using the tile's low-frequency spatial components;

10. Heavily smooth the low-frequency image;

11. Add together the high and low frequency images giving the final synthetic terrain;

Combining the low and high frequency components separately minimises discontinuities between contrasting terrains which may cause artifacts at tile boundaries. Stretching tiles randomly by up to 10% in both x and y directions and adding additional noise ensures each simulated terrain is unique, giving independent statistical samples.

30 martian images were taken from the HiRISE project [11] (samples can be seen in figure 17) each of which was divided into 200 rectangular tiles providing the raw material for the terrain simulator. An example template and resulting synthetic terrain images can be seen in figure 18 and figure 19. The 30 HiRISE terrains were grouped into 10 triplets, as given in table 2, so that 10 different 3-class problems could be tested.

The synthetic images are unrealistic in the respect that the types of terrains combined are unlikely to occur adjacent to one another in genuine martian images. Also, the boundaries between the synthetic terrains are more distinctive than those likely to be found on Mars. However, the underlying patterns form textures which do occur in practice. The use of synthetic images, derived from real martian data, thus permits valid testing of the statistical modelling method and error estimation theories.

6.3 Statistical properties of BRIEF histograms

The need to find an image encoding which satisfies the statistical properties expected by the developed methods has been emphasised several times. Tests must be performed on BRIEF histograms to check if those properties are fulfilled. Section 3.6.2 provided a goodness-of-fit test to confirm that a linear model could be successfully fitted to data. However, this goodness-of-fit only provided a single valued summary giving a mean estimate of the behaviour of model-data residuals. The goodness-of-fit function therefore

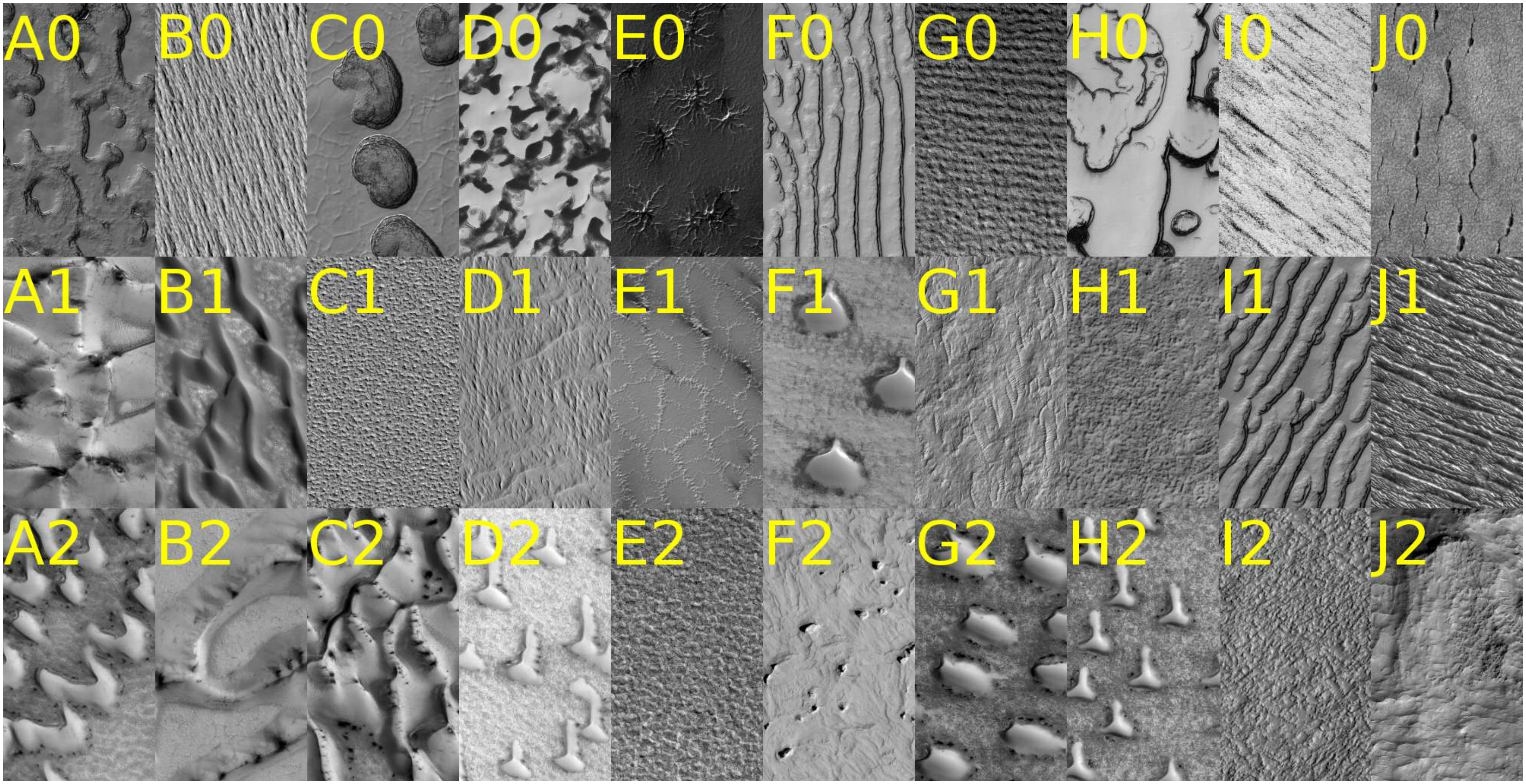


Figure 17: Samples of the 30 HiRISE images used as source data for simulated martian terrains.

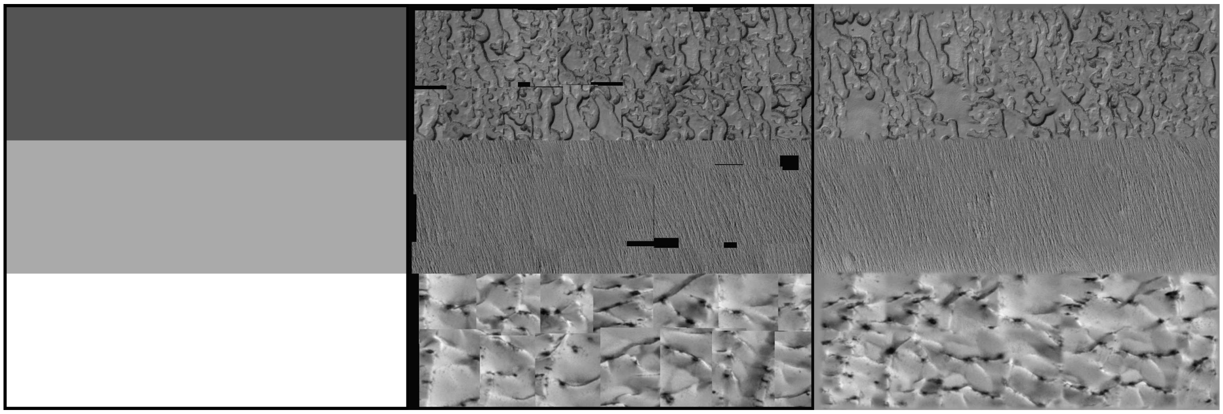


Figure 18: Synthetic martian terrain generator data. Left: Terrain template image determining layout of synthetic image. Middle: Composite image formed without high and low spatial feature separation and without smoothing causing sharp discontinuities at tile boundaries. Right: Smoothly composited image with high and low spatial features combined separately.

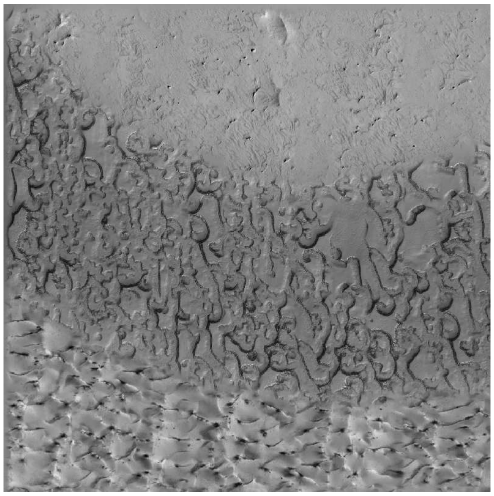


Figure 19: Example of synthetic terrain following more complex boundaries than previous figure.

cannot be used alone to confirm that all of the statistical properties of the BRIEF encoding are appropriate. Instead, a correlation matrix can be estimated from N multiple model fits allowing each residual, i.e. each pattern, to be inspected. Such a matrix can test if individual patterns follow Poisson statistics and are independent from one another.

An m by m residual correlation matrix, ρ, with elements for each possible pair of X patterns can be constructed to give the identity matrix under ideal conditions. If the BRIEF encoding provides truly independent Poisson bins, as is required by the model, then there should be no correlations giving expected off-diagonal terms of zero. Expected diagonal terms of unity can be achieved by normalising each residual to its predicted error. Significant variations away from the identity matrix would be evidence that a representation did not have the necessary properties stated in section 3.2. The elements of the correlation matrix can be computed using:

$$\rho_{XY} = \frac{1}{N}\sum_{r=1}^{N} \frac{\delta_{Xr}\delta_{Yr} + cov_{dof}(\delta_{Xr}, \delta_{Yr})}{\sigma_{\delta_{Xr}}\sigma_{\delta_{Yr}}} \tag{93}$$

where $\delta_{Xr} = \sqrt{\mathbf{H}_{X(r)}} - \sqrt{\mathbf{M}_{X(r)}}$ is the residual between data and model at bin X for sample r, consistent with equation (45); $\sigma_{\delta_{Xr}} = \sqrt{\frac{1}{4} + \sigma^2_{\sqrt{\mathbf{M}_{X(r)}}}}$ is the predicted standard deviation of the residual, also consistent with equation (45); and $cov_{dof}(\delta_{Xr}, \delta_{Yr})$ is a degree of freedom correction. The correction term is required as the estimation of quantities, $\mathbf{Q}$, during model fitting minimises residuals, removing additional variation which would otherwise be observed if the true values of $\mathbf{Q}$ were used. The degree of freedom correction reintroduces this missing variance and can be computed using error propagation:

$$cov_{dof}(\delta_X, \delta_Y) = \sum_i \sum_j \left(\frac{\partial \delta_X}{\partial \mathbf{Q}_i}\right)\left(\frac{\partial \delta_Y}{\partial \mathbf{Q}_j}\right) cov(\mathbf{Q}_i, \mathbf{Q}_j) \tag{94}$$

where

$$\frac{\partial \delta_X}{\partial \mathbf{Q}_i} = \frac{P(X|i)}{2\sqrt{\mathbf{M}_X}} \tag{95}$$

This is a generalisation of the method used in [153] where covariances from eigenvector shape models required correction.

Measuring significant changes away from the ideal identity matrix requires knowledge of the stability of the correlation coefficients. Figure 20 confirms the ideal correlation coefficients using Monte-Carlo data with 9 components and 64 bins. The correlation matrix is shown as a 2D plot, with grey levels indicating coefficient values. This figure also gives the ideal probability distributions of the coefficients when computed from 1,000 Monte-Carlo samples, which can be compared to future correlation matrices computed from equal samples of BRIEF patterns.

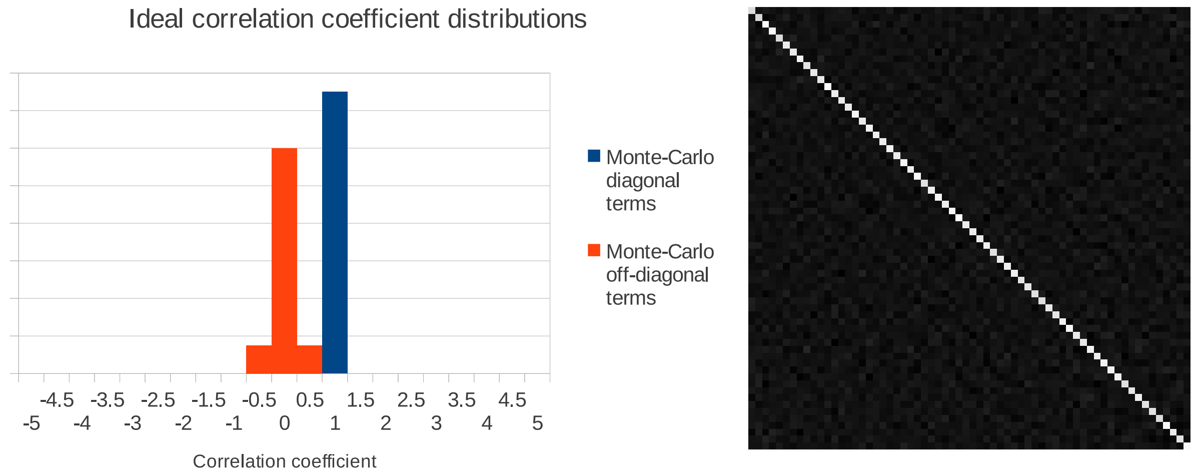


Figure 20: Left: distribution of diagonal and off-diagonal correlation coefficients when computed using 1,000 samples for use as a base-line against which future correlations can be compared. Right: Monte-Carlo correlation matrix summary showing diagonal terms near unity and off-diagonal terms near zero.

6.4 Synthetic terrain testing

Tests were conducted using synthetic images to assess the estimation method's ability to make surface area measurements when images were composed from different terrain classes. Before this could be achieved model components for each class of terrain were extracted from histograms created using a range of BRIEF parameters. The sampling circle size parameter, ι, was held fixed at 8 pixels, whilst the number of pixel pairs, η, was tested at 8, 10 and 12 to give increasing spatial resolutions. The brightness comparison threshold, θ, was tested at values of 3, 5 and 7 times the simulated image noise. Models of the 30 HiRISE textures were created for each set of BRIEF parameters. The models were derived from synthetic images 2048x4096 pixels in size. A 10 row by 10 column rectangular grid was used to divide each image into 100 regions, with individual BRIEF histograms created per region. These histograms were fed into the EM ICA training algorithm, just as Monte-Carlo histograms were used previously in chapter 4.3.

Figure 21 shows a per-terrain breakdown of the model selection process for BRIEF parameters of 8 pixel pairs with a threshold of 5, with model convergence for other parameters following similar curves. Figures 22 and 23 summarise the model selection process for other parameters by plotting mean curves for ease of comparison. In all cases the best fitting models were achieved by reaching approximately 6 components. These figures should be compared to figure 5 of chapter 3 which shows model selection on Monte-Carlo histogram data when ideal data is provided.

To test surface area measurements the trained models, using 6 components per class, were fitted to BRIEF histograms sampled from test images created using 3 strips of different quantities of each terrain following the 10 groups listed in table 2. Each group was

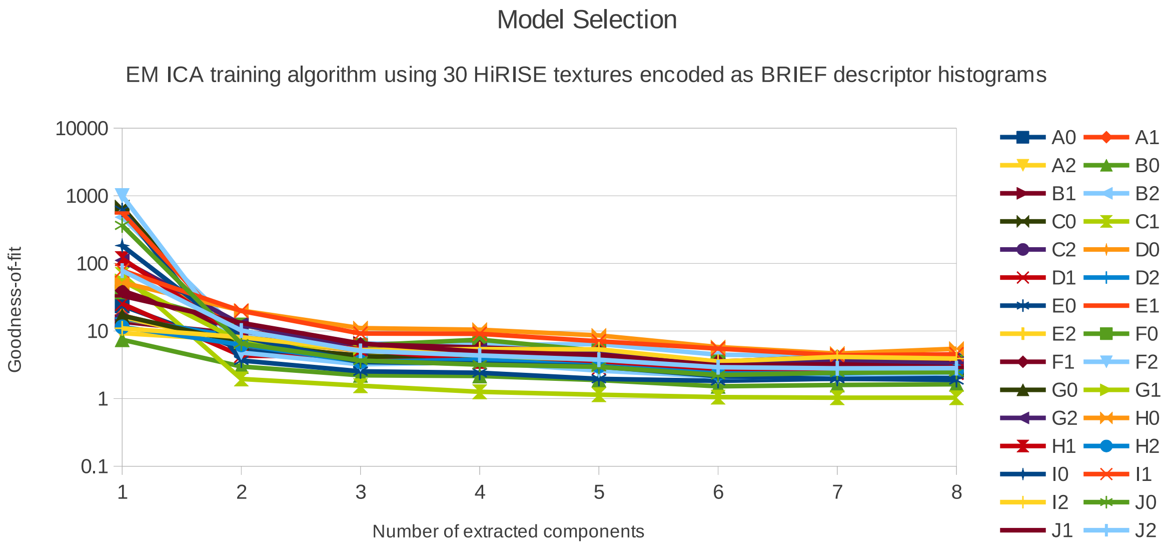


Figure 21: This plot shows the convergence of the goodness-of-fit function when used to select an optimal number of model components to describe the distribution of BRIEF descriptors. Each curve represents a terrain from table 2. The BRIEF parameters used were 8 pixel pairs within an 8 pixel radius with a threshold of 5 times the image noise. This plot is indicative of other BRIEF parameters.

tested using images with dimensions of 2048x6144, 2048x12288, and 2048x24576 pixels. 100 independent images for each group and quantity of data were generated, with estimated class quantities, $\mathbb{Q}$, compared to ground truth surface areas at each trial. Predicted errors were compared to observed deviations from ground truth and recorded using Pull distributions, consistent with previous Monte-Carlo testing. The proportional size of the errors in comparison to the estimated areas were also recorded to test how accurately surface areas could be measured. Figure 24 shows a per-terrain breakdown of predicted to observed error agreements for BRIEF parameters of 8 pixel pairs with a threshold of 5 over a range of image sizes. This figure also summarises error agreements over other parameters. The proportional size of the predicted errors are summarised for different parameters in figure 25.

Correlation matrices were generated from a subset of the data (with fixed parameters of 8 pixel pairs with a threshold of 5 times image noise tested on images of 2048x12288 pixels) generated from 1,000 model fits allowing direct comparison to Monte-Carlo results. Results from these tests are summarised in figure 26.

6.5 Discussion

The tests performed using BRIEF histograms yielded mixed results. The shapes of plotted curves were generally consistent with what would be expected from theory. However, the overall scale of many curves suggest that the BRIEF representation is not an appropriate

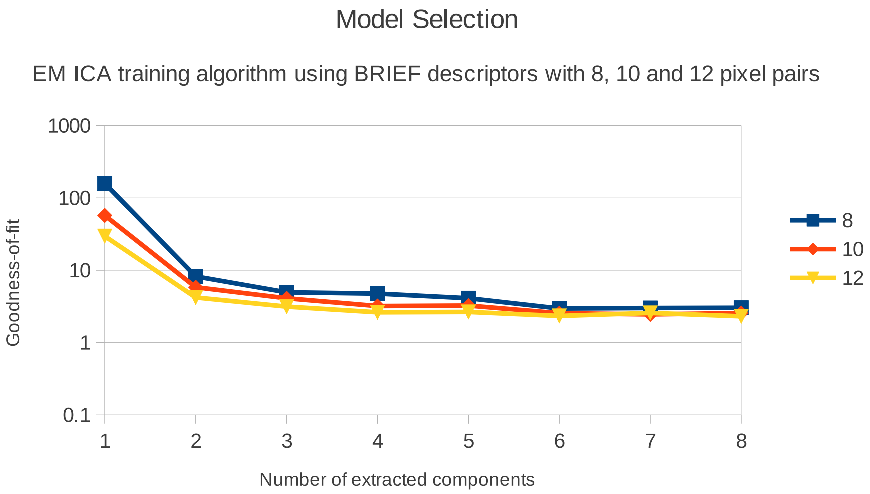


Figure 22: This plot shows the convergence of the goodness-of-fit function across a range of BRIEF parameters. Each curve represents the mean curve of 30 HiRISE terrains modelled using 8, 10 and 12 pixel pairs. The threshold was fixed at 5 time the image noise.

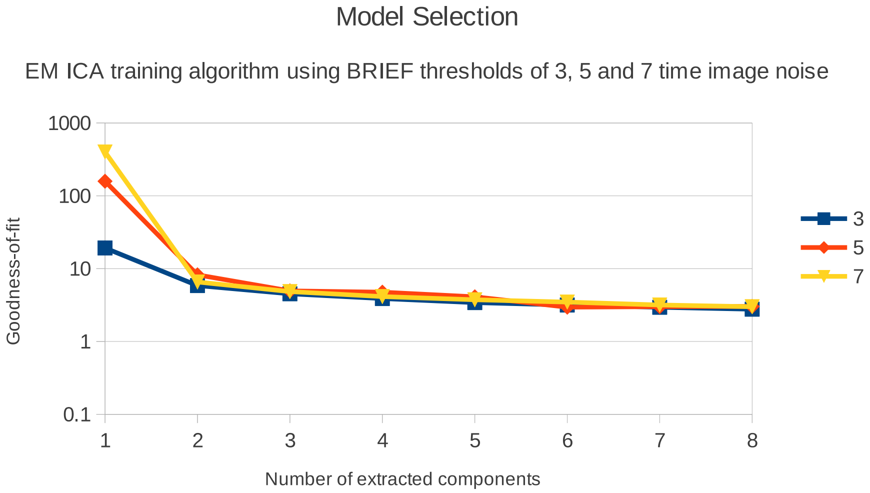


Figure 23: This plot shows the convergence of the goodness-of-fit function across a range of BRIEF parameters. Each curve represents the mean curve of 30 HiRISE terrains modelled using thresholds of 3, 5 and 7 times image noise. The number of pixel pairs was held fixed at 8.

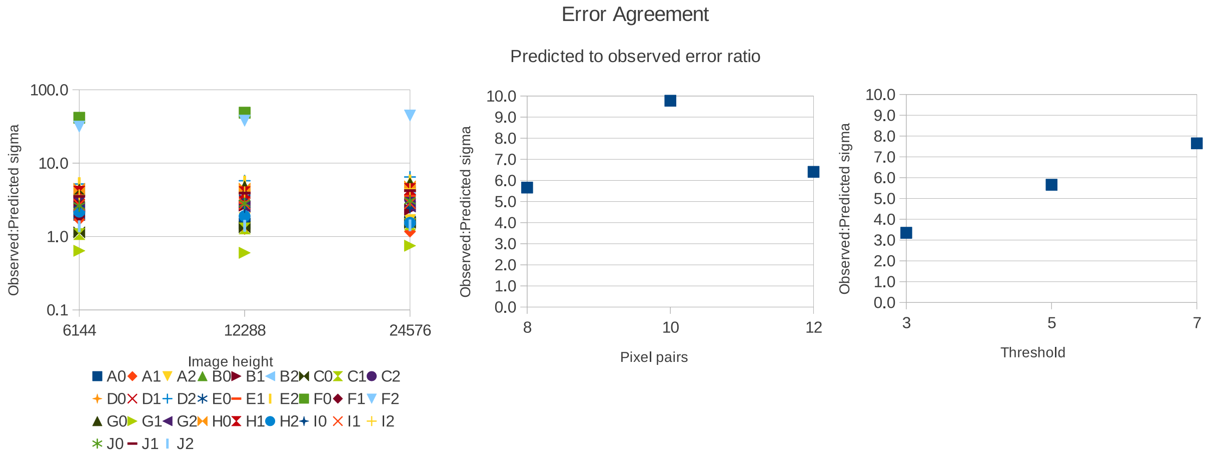


Figure 24: These plots show the agreement between predicted and observed surface area measurement errors over a range of parameters, with the ratio of observed to predicted errors indicated on the y-axis. Left: A breakdown of all terrains' error agreements as a function of image size, where the width of each image was held fixed at 2048 pixels. Centre: The mean error agreements as a function of the number of pixel pairs in the BRIEF representation. Right: The mean error agreements as a function of the BRIEF threshold.

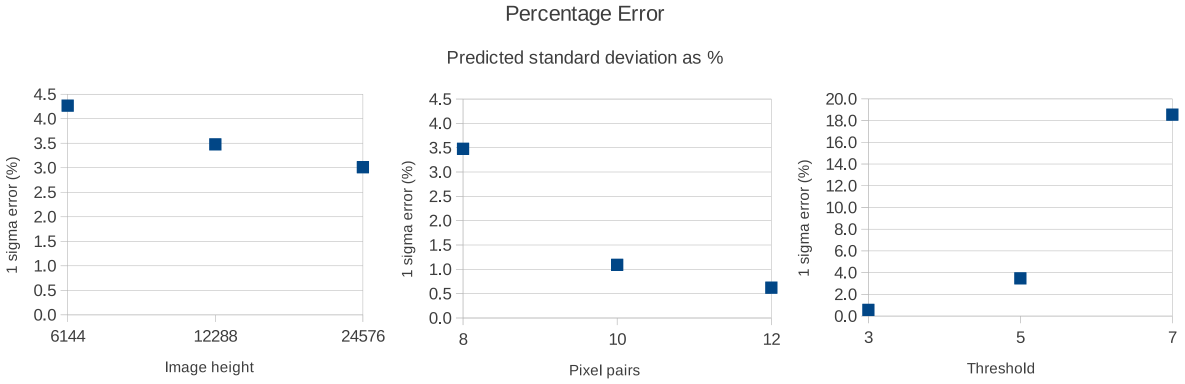


Figure 25: These plots show the mean sizes of the predicted surface area measurement errors as a function of parameters. Left: the percentage area measurement error as a function of image size. Centre: the percentage area measurement error as a function of the number of BRIEF pixel pairs. Right: the percentage area measurement error as a function of the BRIEF threshold.

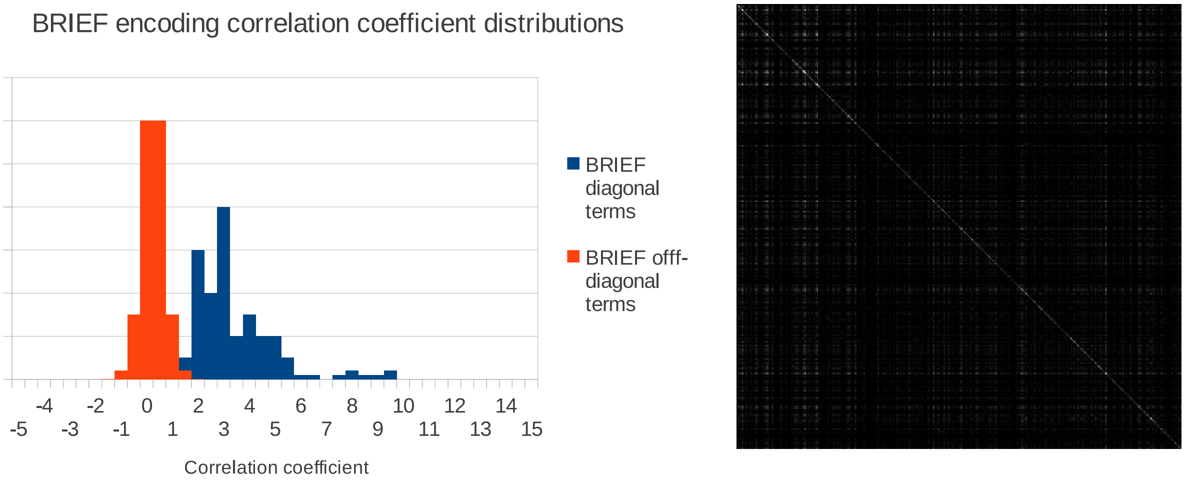


Figure 26: Left: distribution of diagonal and off-diagonal correlation coefficients when computed using 1,000 samples of BRIEF histograms. Right: BRIEF Correlation matrix summary showing diagonal terms above unity and off-diagonal terms with greater than expected spread around zero.

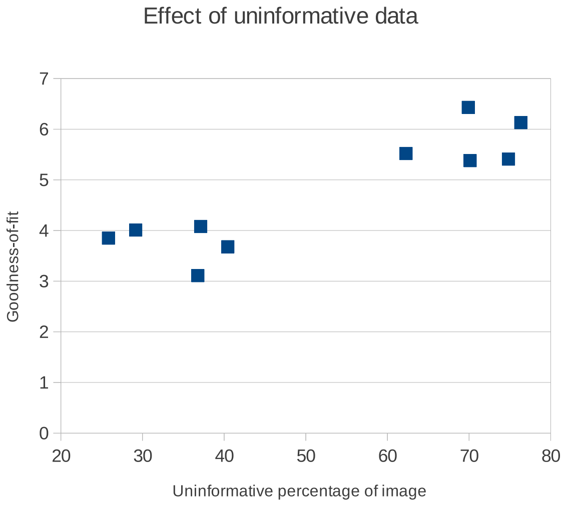


Figure 27: This plots shows the link between poor model fits when fitting to new incoming data and the quantity of data which is uninformative. Each point represents a model fit to a different terrain using BRIEF with 8 pixel pairs and a threshold of 5 times image noise.

encoding scheme for use as input into linear Poisson histogram models. The general behaviour of the histograms will be discussed first, followed by an analysis of the problematic scaling of results.

6.5.1 Behaviour consistent with theory

During training, the goodness-of-fit improved monotonically as the number of extracted components increased. This behaviour was consistent across all terrain types as seen in figure 21. This was to be expected, as each additional component increased the model's ability to describe the data and therefore reduced the model-data residuals. Similar results were observed for other BRIEF parameters, with only marginally different convergence rates.

During testing of area measurements the predicted to observed error ratios remained consistent at fixed (yet biased) values for each terrain across different quantities of data, as seen in figure 24. This showed that the statistical and systematic components of the predicted errors, which are both functions of data quantity, produced self-consistent (yet biased) estimates as their relative contributions to total errors changed. This general behaviour was to be expected and had been observed in Monte-Carlo tests as seen previously in figures 11 and 13.

The average size of the predicted errors changed as a function of the quantity of data and BRIEF parameters as seen in figure 25. As the quantity of data (image height) increased the proportional error decreased consistent with the proportional reduction in statistical error noted in section 5.3 which stated that statistical errors grew with the square-root of quantities and systematic errors grew linearly, i.e. giving an overall sub-linear growth. The average size of predicted errors also decreased as the number of pixel pairs increased consistent with the theory that errors are smaller for less ambiguous patterns, as explained in section 4.1.2. As the number of pixel pairs increased more information was encoded about local structure therefore each pattern became more discriminating, i.e. less ambiguous. Finally, average errors increased as the BRIEF threshold increased, again consistent with levels of ambiguity with larger thresholds being less discriminating of subtle changes in pixel grey levels.

Despite the general behaviour of plots following what would be expected from theory, the overall scale of some plots highlighted issues: the goodness-of-fits and error agreement ratios should have reached unity, but they did not. These issues are discussed next.

6.5.2 Problematic behaviour

The ICA training algorithm was incapable of extracting sufficiently descriptive linear components to achieve model goodness-of-fits of unity. Extrapolating the model selection curves of figure 21 suggests that extracting a greater number of components would be of no benefit. The apparent floor in attainable model fits could have been caused by model-data residuals quantised above the level predicted by the counting of individual Poisson events. This might be explained if BRIEF descriptors systematically appeared in pairs, or larger multiples, causing underlying Poisson events to be double counted. Further evidence of a double counting effect can be seen in the correlation matrices, summarised in figure 26 (in comparison to the expected correlations of figure 20), which reveal diagonal terms significantly above 1. The larger than expected diagonal terms suggest that individual pattern events are being scaled upwards by some unknown factor consistent with Poisson events being recorded multiple times. Off-diagonal terms, some of which were significantly away from zero, also suggest that correlated clusters of patterns were being generated in batches. The possible double counting of events is not addressed in the statistical model, nor are the correlations between patterns. These violations of model assumptions are a likely cause of error predictions failing to match errors observed empirically, as seen in figure 24.

It was argued in section 3.1.1 that patterns within planetary images are likely to follow Poisson statistics due to the physical systems from which they arise, e.g. it is reasonable to assume that impact events which cause craters are Poisson distributed, with some mean number of impacts expected per unit of time. However, it may have been naïve to believe that fixed-sized overlapping BRIEF descriptors would trivially correspond to the patterns produced through these processes. Indeed, the evidence suggests that if Poisson events are behind the generation of the data then individual events do not correspond to individual BRIEF samples, but rather they correspond to groups of related samples. This observation, if correct, explains discrepancies between predictions and observations, and could provide the insights needed to design an improved image encoding better approximating independent Poisson events. An improved representation would be one that grouped together all patterns related to individual events so that they could be counted individually, mitigating the problems described above.

Evidence of further model assumption violations may be seen in uninformative data. Featureless parts of images containing empty space results in the generation of large numbers of BRIEF descriptors with all bits set to zero. An increase in uninformative descriptors coincide with increasingly poor goodness-of-fits when models are fitted to new incoming data, as seen in figure 27. The counting of non-features, i.e. uninformative descriptors, cannot easily be attributed to Poisson events. Also, as uninformative data may appear in almost all terrains it is difficult to attribute uninformative descriptors to appropriate classes raising some difficult philosophical questions. For example, should empty space between dunes be counted as part of the surface area of the dunes, or should such space be attributed to some other background classification? Or, at a sparse boundary between dunes and boulders, at which point should the empty space around the boundary be attributed to the boulder field rather than the dune field? It may be beneficial to exclude uninformative regions in general analysis tasks, leaving their interpretation to the needs of specific applications.

6.6 Summary

An image encoding based upon local BRIEF descriptors was proposed for converting image data into pattern histograms suitable for linear Poisson histogram models. It was hoped that each BRIEF pattern, X, would behave as an independent Poisson event attributable to individual image locations so that estimated model quantities could be used as surface area measurements. Synthetic martian images were generated in large quantities to test area estimates and predicted errors against known ground truths and empirically observed measurement repeatability. Unfortunately the method failed because of the encoding's inability to capture the underlying Poisson events believed to be responsible for the generation of the data. It could be argued that the data generators are not Poisson and/or not linear, but this conclusion would require the statistical modelling methods developed to be abandoned. Instead it will be assumed that linearly combined Poisson processes can be used to describe martian image data and the insights gained will be used to develop an improved encoding better approximating the required properties.

7 Martian Terrains: Poisson Blob Representation

The techniques for modelling pattern distributions and computing errors presented in chapters 3, 4 and 5 were developed under the assumptions that histograms are linearly composed, with independent Poisson bins. The previous chapter suggested the use of a fixed sized BRIEF-like representation for translating images into such histograms so the techniques could be applied to make surface area measurements for martian terrains. Unfortunately, the simple BRIEF-like representation proved ineffective, leading to suspicions that spatial correlations were not being correctly accounted for, thereby violating the assumption of independence.

An alternative representation will be presented in this chapter which groups related image data points together. This representation will combine neigbouring BRIEF-like patterns, forming connected 'blobs', better approximating independent Poisson events. However, as blobs can be irregular in shape and size it will be seen that the new encoding requires additional interpretation to convert model quantities, $\mathbf{Q}$, into meaningful area measurements with adjusted error estimates. Two alternative approaches to the calculation of area covariances will be presented then both areas and error estimates will be tested using Monte-Carlo simulated blob histograms and simulated martian images.

7.1 Blobs and Areas

In the previous chapter X bins represented fixed sized BRIEF patterns taken at every image location, with each location being treated as an individual Poisson event. This resulted in a convenient one-to-one correspondence between estimated model quantities, $\mathbf{Q}$ and physical surface areas. However, evidence was observed that BRIEF patterns came in clusters, suggesting that Poisson processes operating on a higher level were double counted in the form of multiple similar X and correlated X patterns. The proposed Poisson blob representation replaces each fixed size X with groups of connected image points sharing a common BRIEF pattern, thereby removing the double counting of individual patterns and reducing the magnitude of correlations between them, but not necessarily removing the correlations entirely. Entries are then made in histogram bins on a blob-by-blob basis. As blobs encompass multiple image locations the representation, $X = \{\pi, \gamma\}$, must encode two pieces of information: π, the BRIEF descriptor common to all image locations within the blob; and γ, a size band indicating the size of the blob. The size bands are organised into logarithmic (base-2) bins spanning a large dynamic range of possible blob sizes whilst keeping the pattern space reasonably small. Figure 28 illustrates how image pixels are converted into Poisson blobs. In addition to the X encoding, the location and precise size (in pixels) of each blob can be recorded separately for use in area calculations. Finally, to remove problems associated with uninformative space all patterns containing BRIEF descriptors with all bits set to zero are excluded.

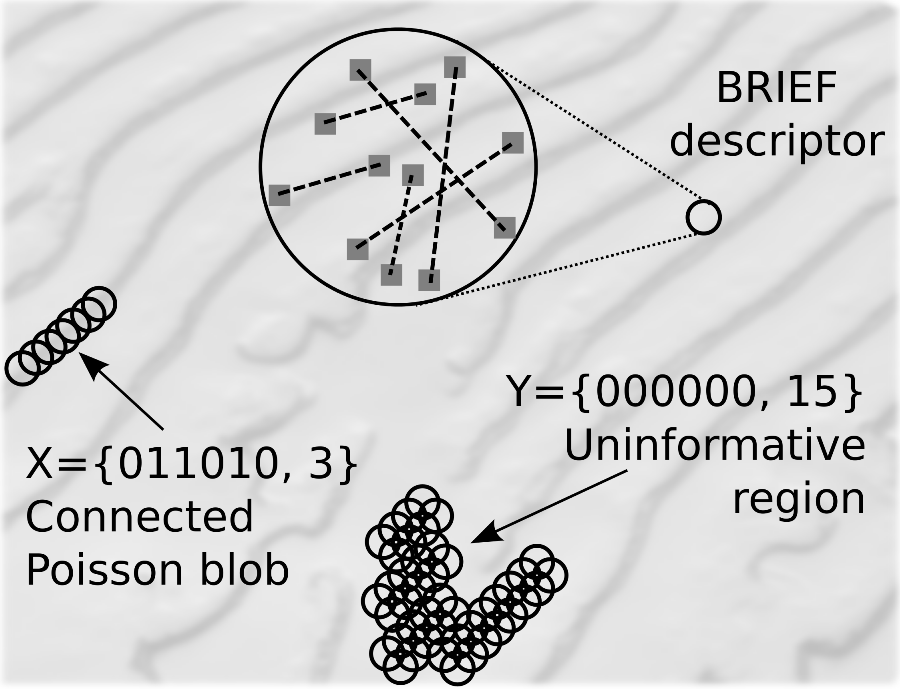


Figure 28: The Poisson blob representation groups together adjacent image locations which share a common BRIEF pattern. Here, blob $X = \{\pi, \gamma\}$ contains a chain of correlated locations sharing the BRIEF pattern $\pi = 011010$ placed into size bin $\gamma = 3$. Blob Y is a large connected region of empty space with pattern $\pi = 000000$ in size band $\gamma = 15$.

The model component weighting quantities, $\mathbf{Q}$, must now be interpreted as blob counts rather than surface areas. The estimated areas, $\mathbf{A}$, covered by component k then becomes:

$$\mathbf{A}_k = \sum_X P(k|X) a_X \tag{96}$$

$$a_X = \sum_d a_d \delta(X_d = X) \tag{97}$$

where $\mathbf{A}_k$ is the area estimate for component k; a_X is the total area covered by blobs of type X; d is a specific blob (with pattern X_d); a_d is the specific area of blob d; and $\delta(X_d = X)$ is 1 if the pattern at d matches X and is zero otherwise. These per-component area estimates can be combined, analogously to quantities in equation (25), to give total class area estimates:

$$\mathbb{A} = \mathbb{K}\mathbf{A} \tag{98}$$

where $\mathbb{K}$ is the mapping matrix between components and their classes.

7.2 Area Errors

Chapter 5 provided a covariance matrix, $\mathbf{C}$, for estimated quantities, $\mathbf{Q}$, from which an area covariance matrix, $\mathbf{C}_A$, can be computed for estimated areas, $\mathbf{A}$. Two methods are considered:

- dividing the covariance, $\mathbf{C}$, across individual area terms then re-accumulating an area covariance by scaling the individual terms;
- and applying conventional error propagation to the sources of uncertainty (estimates of $\mathbf{Q}$ and blob sizes a_X) found within area estimation calculations.

Once known, the component area covariances can be combined, analogously to quantity covariances in equation (58), to give total class area covariances:

$$\mathbf{C}_{\mathbb{A}} = \mathbb{K}\mathbf{C}_{\mathbf{A}}\mathbb{K}^{\intercal} \tag{99}$$

providing the total predicted accuracies on terrain surface area measurements. The approaches to computing $\mathbf{C}_{\mathbf{A}}$ are described in the next two sections.

7.2.1 Per X bin covariance scaling approach

As explain in chapter 5, the covariance for quantities is composed of statistical and systematic errors, $\mathbf{C}_{data}$ and $\mathbf{C}_{model}$. The statistical errors come from sampling fluctuations in incoming data and systematic errors are due to errors in estimated model components. As a consequence, statistical errors are independent over individual blobs, while systematic model errors are identical between blobs of the same X. In terms of area contributions, the statistical effects can be viewed as operating as a sum over individual blobs: $\sum_d P(k|X)a_d$; whereas the systematic effects can be viewed as operating as a sum over blob types: $\sum_X P(k|X)a_X$.

For systematic errors, logically it must be possible to write the total covariance as a sum over X:

$$\mathbf{C}_{model} = \sum_X \mathbf{C}_{X,model} \tag{100}$$

And for statistical errors, it must be possible to write the total covariance as a sum over d:

$$\mathbf{C}_{data} = \sum_d \mathbf{C}_{d,data} \tag{101}$$

The challenge is then to estimate from $\mathbf{C}_{model}$ and $\mathbf{C}_{data}$ the term-by-term contributions, $\mathbf{C}_{X,model}$ and $\mathbf{C}_{d,data}$, and how much they should be scaled by.

Beginning with the systematic error, $\mathbf{C}_{X,model}$ can be interpreted as the contribution to the covariances arising due to **all equivalent** X patterns and giving rise to the contribution $P(k|X)\mathbf{H}_X$ to total $\mathbf{Q}_k$. We can write this as a measurement and covariance for each contribution:

$$\mathbf{Q}_X = \mathbf{H}_X \mathbf{P}_X \pm \mathbf{C}_{X,model} \tag{102}$$

where $\mathbf{P}_X$ is a vector of posterior probabilities $\{P(k=1|X), P(k=2|X), \ldots, P(k=n|X)\}^\intercal$. By simple scaling, the contributions to the total area $P(k|X)a_X$ must therefore have an associated covariance of $\mathbf{C}_{X,model}(a_X/\mathbf{H}_X)^2$.

It follows that the total covariance of systematic errors on $\mathbf{A}$ is the sum of the errors on each independent area estimate, given by:

$$\mathbf{C}_{\mathbf{A},model} = \sum_X \frac{a_X^2 \mathbf{C}_{X,model}}{\mathbf{H}_X^2} \tag{103}$$

While the statistical errors are given by:

$$\mathbf{C}_{data} = \sum_X \mathbf{C}_{X,data} = \sum_d \frac{\mathbf{C}_{X_d,data}}{\mathbf{H}_{X_d}} \tag{104}$$

which is the uncertainty in the final area estimate associated with each individual blob, i.e.

$$\mathbf{Q}_d = \mathbf{P}_X \pm \mathbf{C}_{data}/\mathbf{H}_{X_d} \tag{105}$$

and so continuing as above

$$\mathbf{C}_{\mathbf{A},data} = \sum_d \frac{a_d^2 \mathbf{C}_{X_d,data}}{\mathbf{H}_{X_d}} \tag{106}$$

Both components of the error can now be recombine to estimate the total uncertainty on areas $\mathbf{A}$:

$$\mathbf{C}_{\mathbf{A}} = \mathbf{C}_{\mathbf{A},data} + \mathbf{C}_{\mathbf{A},model} \tag{107}$$

The above expressions can be tested for consistency by checking the covariance estimates for the case when blobs are pixel, in which case $\mathbf{C}$ should equal $\mathbf{C}_{\mathbf{A}}$. So, letting $a_X \;=\; \mathbf{H}_X$ and $a_d = 1$:

$$C_{\mathbf{A},model} \;=\; \sum_X \frac{\mathbf{H}_X^2 \mathbf{C}_{X,model}}{\mathbf{H}_X^2} \;=\; \sum_X \mathbf{C}_{X,model} \;=\; \mathbf{C}_{model} \tag{108}$$

and

$$C_{\mathbf{A},data} \;=\; \sum_d \frac{\mathbf{C}_{X_d,data}}{\mathbf{H}_{X_d}} \;=\; \sum_X \mathbf{C}_{X,data} \;=\; \mathbf{C}_{data} \tag{109}$$

as expected.

7.2.2 Error propagation approach

To apply error propagation the area measurement calculation of equation (96) can be rewritten using Bayes Theorem:

$$\mathbf{A}_k = \sum_X \left(\frac{P(X|k)\mathbf{Q}_k}{\mathbf{M}_X} \right) a_X \tag{110}$$

$$= \mathbf{Q}_k \sum_X \frac{P(X|k) a_X}{\mathbf{M}_X}$$

where the sources of errors are noise in $\mathbf{Q}$ and noise in areas a_X; $P(X|k)$ is the probability of X given histogram component k; and $\mathbf{M}_X$ is the modelled frequency of X. The uncertainties stemming from elements of $\mathbf{Q}$ are expected to be the dominant source of error, as the total blob areas for each X, a_X, should be relatively stable given the large number of blobs within an image. However, at very low sample sizes this additional error may become noticeable. Error propagation can be applied giving an area covariance including both sources:

$$\mathbf{C}_{\mathbf{A}} = \nabla_Q \mathbf{C} \nabla_Q^T + \sum_X [\nabla_{a_X} \otimes \nabla_{a_X}^T] \sigma_{a_X}^2 \tag{111}$$

where ∇_Q is the matrix of partial derivatives:

$$\nabla_{Q,ij} = \frac{\partial \mathbf{A}_i}{\partial \mathbf{Q}_j} \tag{112}$$

and ∇_{a_X} is the vector of derivatives:

$$\nabla_{a_X,k} = \frac{\partial \mathbf{A}_k}{\partial a_X} \tag{113}$$

For the case when $i = j$ is:

$$\nabla_{Q,ij} = \sum_X \frac{P(X|j) a_X}{M(X)} \tag{114}$$

$$= \sum_X \frac{P(X|j) Q(j) a_X}{\mathbf{M}_X \mathbf{Q}_j} = \sum_X \frac{P(j|X) a_X}{\mathbf{Q}_j} \tag{115}$$

$$= \frac{\mathbf{A}_j}{\mathbf{Q}_j} \tag{116}$$

and is zero for $i \neq j$ forming a diagonal matrix. The other terms are given by:

$$\begin{aligned} \nabla_{a_X,k} &= \mathbf{Q}_k \frac{P(X|k)}{\mathbf{M}_X} \\ &= P(k|X) \end{aligned} \tag{117}$$

The variance on an a_X can be estimated by summing the individual independent blob variances:

$$\sigma^2_{a_X} = \mathbf{H}_X \sigma^2_{a_{Xd}} = \sum_{d,\delta(X_d=X)} (a_d - < a_{Xd} >)^2 \tag{118}$$

which can be done as above using sample variances. Alternatively, assuming a uniform size-band distribution (which may be justified if the logarithmic size bands are narrow enough to be locally flat) the blob variances can be computed using:

$$\sigma^2_{a_{Xd}} = \frac{1}{12}(\gamma_{X_l} - \gamma_{X_u})^2 \tag{119}$$

where γ_{X_l} and γ_{X_u} are the lower and upper bounds of the size bin γ for blob type X.

7.3 Monte-Carlo and synthetic terrain testing

The additional theory required to convert model quantities into terrain surface area measurements was first tested using Monte-Carlo generated blob histograms. The Monte-Carlo histogram generating methods of chapter 4 were amended to create histogram bin frequencies with accompanying blobs. The size fields, γ, and blob areas, a_d, were set to cover a range of possible conditions: blobs of fixed finite sizes; blobs of uniformly distributed random sizes; and blobs with random size distributions as a function of X. Fixed finite blob sizes of 1 and 10 pixels were tested; random blob sizes uniformly selected between 1 and 10 were tested; and X specific random blob sizes between 1 to 10, 11 to 20, etc. for increasing X were tested.

The blob histograms were created with 9 components spread over 64 pattern bins, using the same hand-coded distributions as found in chapter 4. Different quantities of training and testing data were used, with predicted errors repeatedly computed over 1,000 trials per experiment. The per-X covariance method (section 7.2.1) and the error propagation method (section 7.2.2) for computing errors were both tested. Pull distributions were used, consistent with previous Monte-Carlo experiments, to test agreement between observed and predicted area errors for the different ratios of training and testing data.

Monte-Carlo experiments were also conducted using the error propagation method (section 7.2.2) to test the relative contributions and stabilities of the sources of uncertainty at very low sample statistics, i.e. uncertainties in quantities and uncertainties in blob sizes. The effects of blob size uncertainties were tested by predicting errors with and without

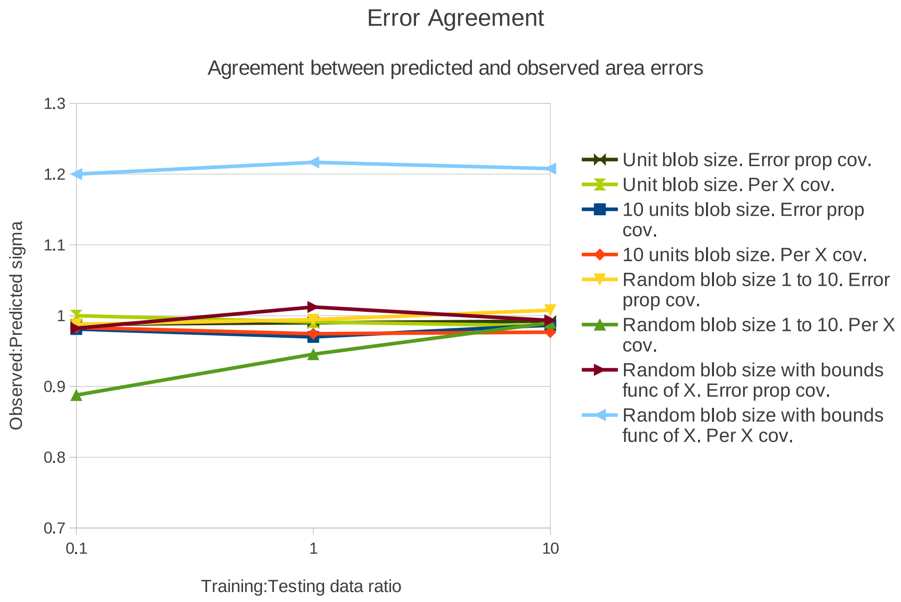


Figure 29: This plot shows the agreement between observed and predicted blob Monte-Carlo area errors over a range of quantities using the per X covariance method and the error propagation method. Under all circumstances (fixed sized blobs, random sized blobs) the error propagation method produces good predictions. However, the per X covariance method fails to produce good predictions when random blob sizes are simulated.

blob size variance contributions, i.e. with noise only propagated from $\mathbf{Q}$ and with noise propagated from both $\mathbf{Q}$ and blob areas, a_X.

Once corroborated in Monte-Carlo, blob histograms were sampled from HiRISE data. To provide continuity of testing the synthetic martian terrain image experiments of the previous chapter (section 6.4) were repeated using the Poisson Blob representation. To allow a direct comparison to be made the tests were conducted under the same conditions and parameter setting as BRIEF. Area covariances were computed using the error propagation method of section 7.2.2 and ground-truth areas were amended to exclude all uninformative regions. The improvements in model selection and error predictions were reported using the same methods as earlier.

7.4 Discussion

This discussion will be divided into three parts covering: the additional theory required to convert quantity estimates into area measurements; the improvements gained through the use of Poisson blobs rather than BRIEF; and the limitations of the improved image encoding.

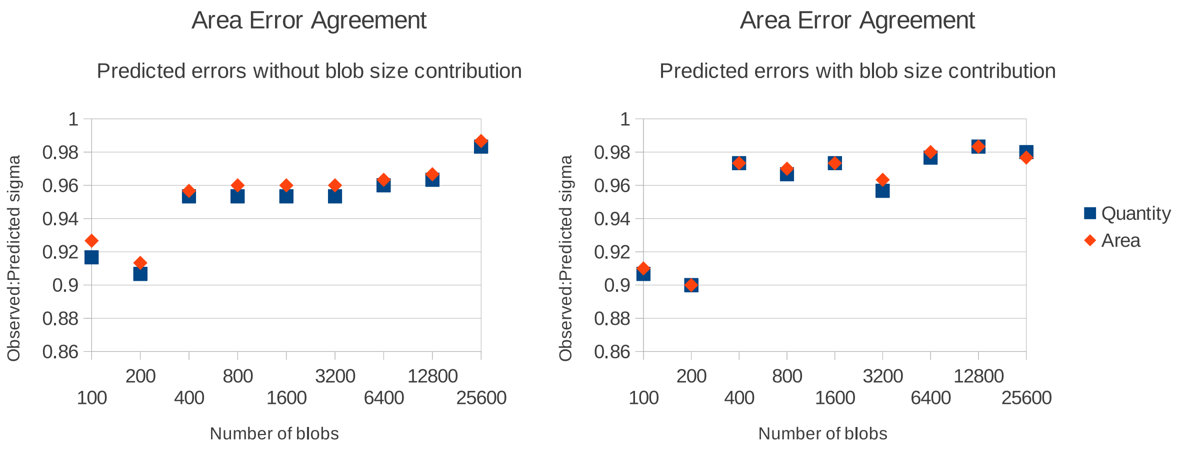


Figure 30: These plots show the negligible, yet systematic, underestimation of area errors when blob size uncertainty is omitted from the error propagation approach. Left: errors without random blob size contribution. Right: errors with random blob size contribution.

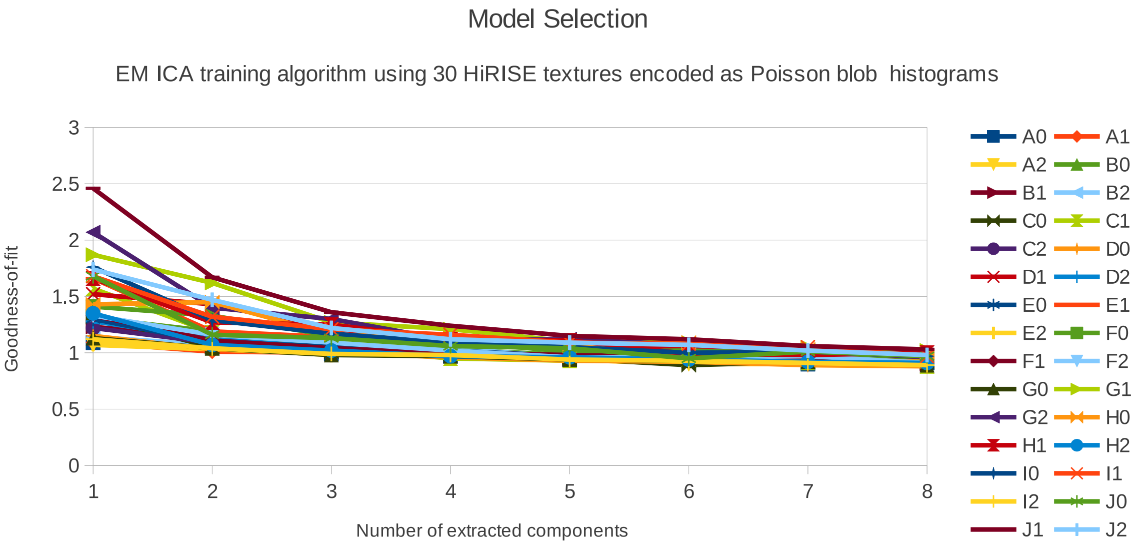


Figure 31: This plot shows the convergence of the goodness-of-fit function when used to select an optimal number of model components to describe the distribution of Poisson blobs. Each curve represents a terrain from table 2. The underlying BRIEF parameters used were 8 pixel pairs within an 8 pixel radius with a threshold of 5 times the image noise. This plot is indicative of other parameters. See figure 21 for comparison with BRIEF encoding.

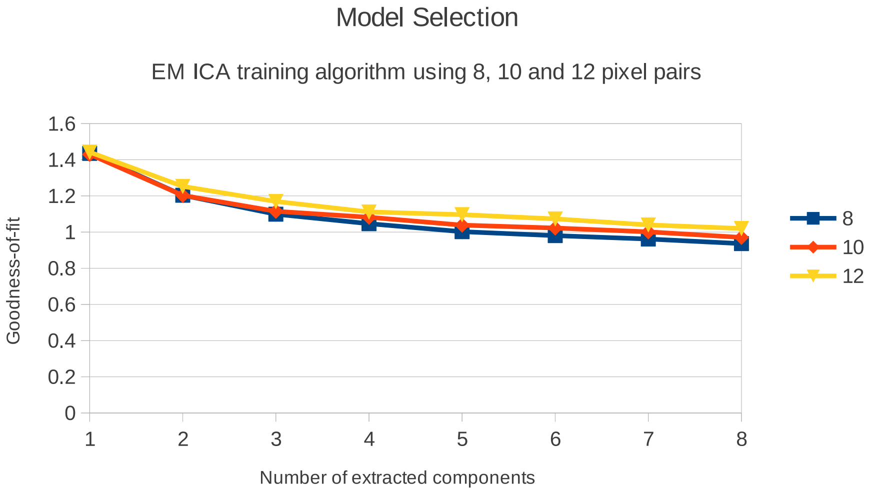


Figure 32: This plot shows the convergence of the goodness-of-fit function across a range of parameters. Each curve represents the mean curve of 30 HiRISE terrains modelled using 8, 10 and 12 pixel pairs. The threshold was fixed at 5 time the image noise. See figure 22 for comparison with BRIEF encoding.

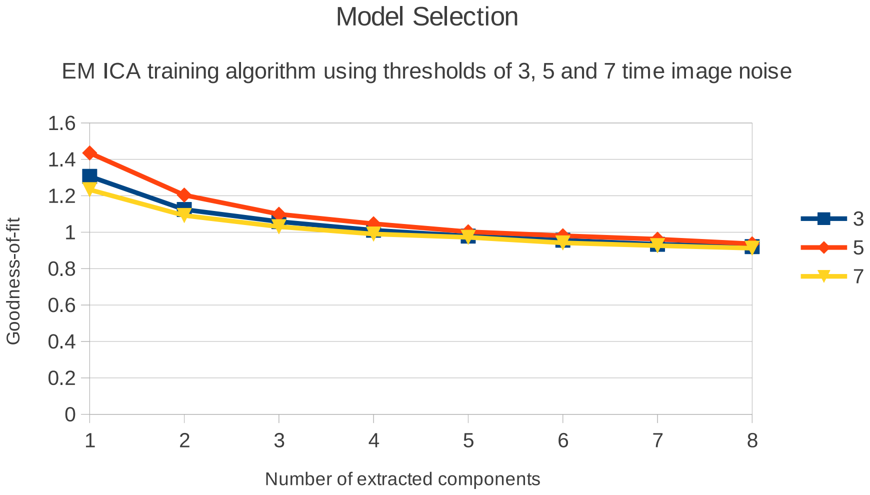


Figure 33: This plot shows the convergence of the goodness-of-fit function across a range of parameters. Each curve represents the mean curve of 30 HiRISE terrains modelled using thresholds of 3, 5 and 7 times image noise. The number of pixel pairs was held fixed at 8. See figure 23 for comparison with BRIEF encoding.

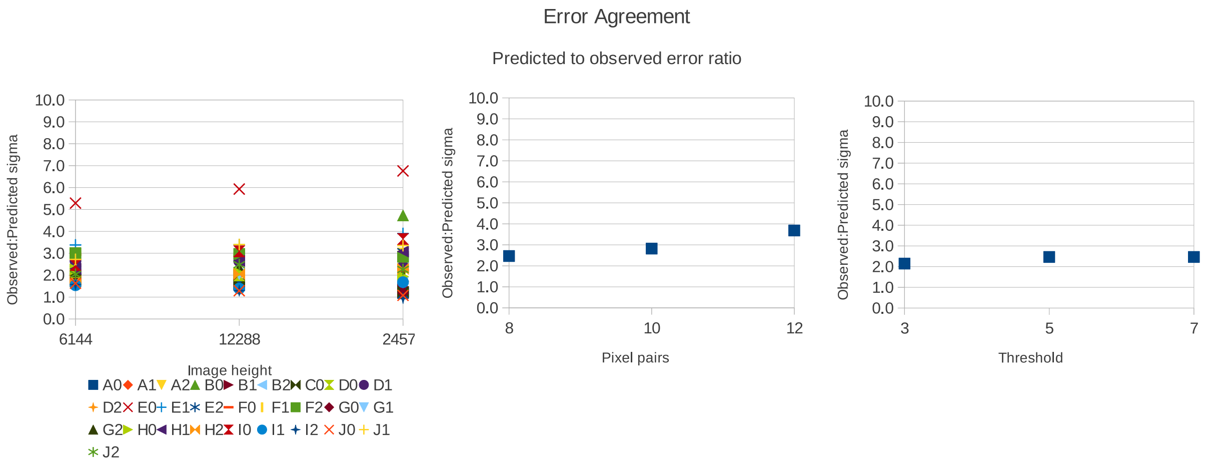


Figure 34: These plots show the agreement between predicted and observed surface area measurement errors over a range of parameters, with the ratio of observed to predicted errors indicated on the y-axis. Left: A breakdown of all terrains' error agreements as a function of image size, where the width of each image was held fixed at 2048 pixels. Centre: The mean error agreements as a function of the number of pixel pairs. Right: The mean error agreements as a function of threshold. See figure 24 for comparison with BRIEF encoding.

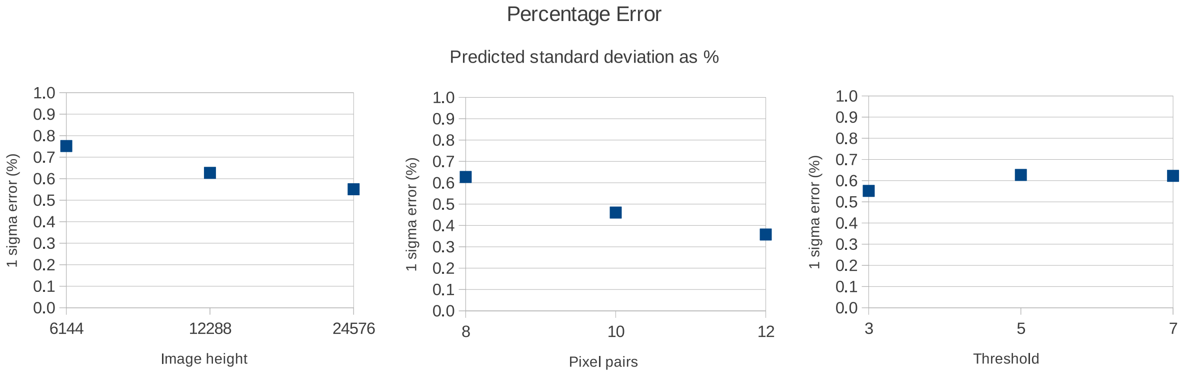


Figure 35: These plots show the mean sizes of the predicted surface area measurement errors as a function of parameters. Left: the percentage area measurement error as a function of image size. Centre: the percentage area measurement error as a function of the number of pixel pairs. Right: the percentage area measurement error as a function of threshold. See figure 25 for comparison with BRIEF encoding.

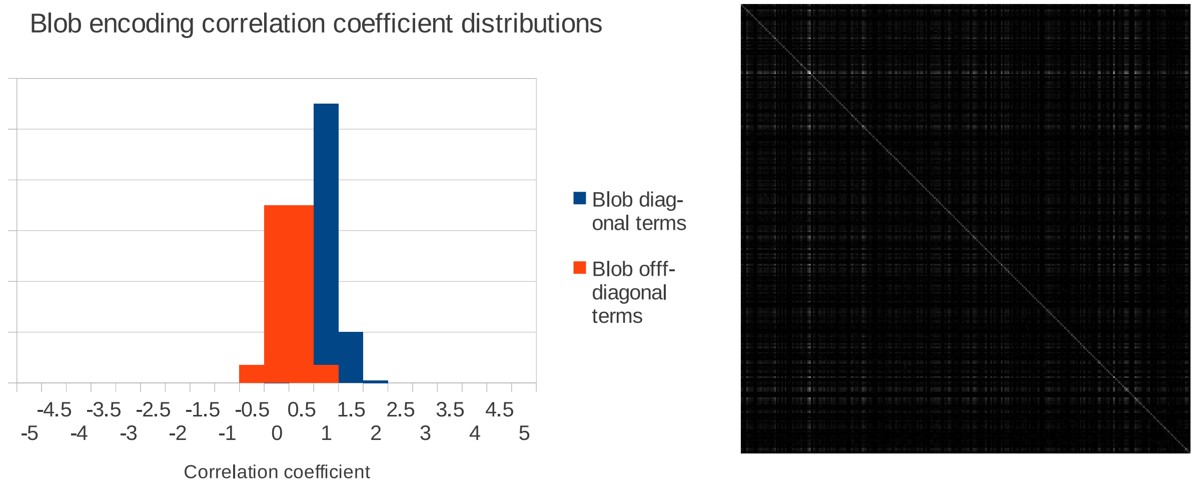


Figure 36: Left: distribution of diagonal and off-diagonal correlation coefficients when computed using 1,000 samples of Poisson blob histograms. Right: Correlation matrix summary showing diagonal terms close to unity and off-diagonal terms with slightly greater than expected spread around zero. See figure 26 for comparison with BRIEF encoding.

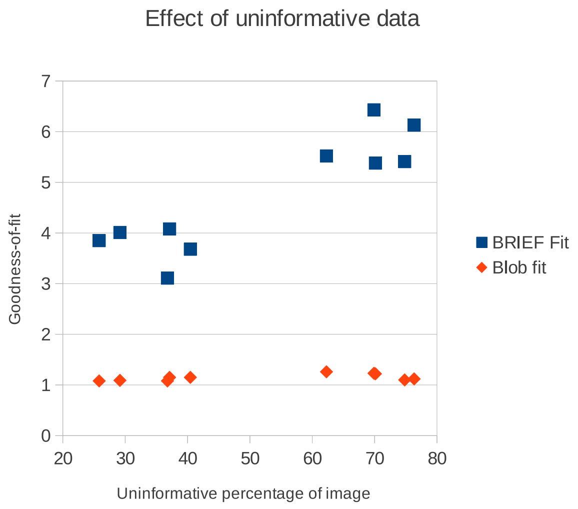


Figure 37: This plots shows the link between model fits when fitting to new incoming data and the quantity of data which is uninformative. The curves compare the original BRIEF and Poisson blob representations. Each point represents a model fit to a different terrain using 8 pixel pairs and a threshold of 5 times image noise.

7.4.1 Area measurements from blobs

The per-X area covariance method (section 7.2.1) only worked in limited cases, whereas the error propagation method (section 7.2.2) worked generally. Monte-Carlo blob histograms showed that the per-X covariance scaling method for area error predictions failed in the most realistic test, that of random blob sizes as a function of X. This can be seen on the light blue curve in figure 29. This failure in the method, despite the logical arguments for the approach, might be explained by missing terms which should be present in the per-X covariances. The approach to computing the per-X covariances, $\mathbf{C}_X$, took a top-down short-cut by assuming it was valid to divide the total covariance over the independent area terms. Correctly derived per-X covariances from the bottom-up, following a scheme similar to that used in section 5.1.2, may have proven more effective. Despite this, in many cases the approach did work, but can not be guaranteed for general use. In contrast, the error propagation approach performed well under all circumstances, hence its adoption in later tests.

The Monte-Carlo blob testing at very low sample statistics was primarily intended to observe the importance of blob area error contributions. However, this testing also revealed the breakdown of the error theory when quantities of blobs dropped below a limit[7] which can be seen in figure 30. Conveniently this is a safe mode of failure, as the predicted errors were overestimated at low samples, avoiding the possibility of accidental over-interpretation of results. This overestimation can be explain as predicted area errors are assumed to be Gaussian, with tails which may straddle zero. When low quantities are observed in practice they are truncated at zero, making true error distributions narrower than predicted.

The effect of the random spread of individual blob sizes within any given X bin is shown to be negligible in figure 30. Here, the effect of omitting this additional variance from the area error estimates ($\sum_X [\nabla_{a_X} \otimes \nabla_{a_X}^T] \sigma_{a_X}^2$) can be seen to give a systematic underestimate of error in comparison to the corresponding blob quantity error. However, this error is within the noise of the ability to measure the effect and is only evident from its systematic nature (with the red diamond area points always being above the blue square quantity points).

7.4.2 Poisson blob improvements

Trained models of martian images constructed using Poisson blobs proved to be far more descriptive than those produced using BRIEF. During model selection, the goodness-of-fit approximately reached unity for all terrains using between 3 to 8 components. A comparison between figure 21 (BREIF) and figure 31 (Poisson blob) shows that where BRIEF was limited to goodness-of-fits of around 5, blob fits could reach the ideal score

[7] The 400 sample limit is likely a function of the 64 pattern bins. This limit will be different for other datasets depending upon the complexity of the model and number of patterns.

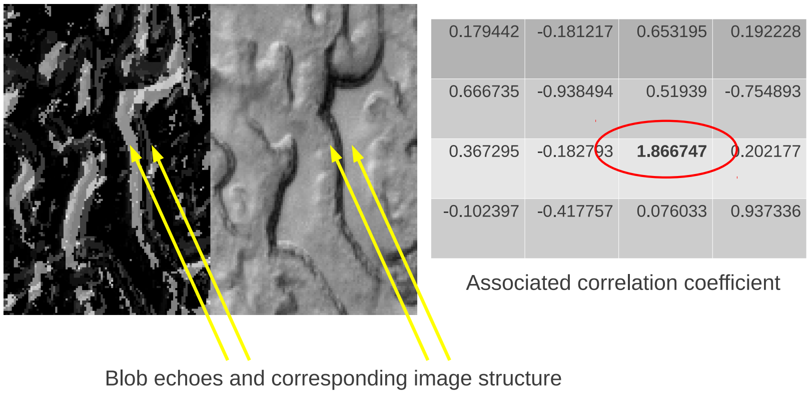
of unity suggesting that the the double counting effects previously observed had been significantly reduced. Similar results were obtained using a range of parameters, as seen in figures 32 and 33.

Surface area measurements were computed with predicted errors typically between a factor of 2 to 3 of observed errors across most terrains and parameters. Whilst Poisson blob error predictions continue to be underestimated, they are an improvement upon BRIEF error predictions which were typically underestimated by between a factor of 3 to 10, as seen by comparing figure 34 (Poisson blob) with figure 24 (BRIEF). Evidence also shows that the improved image encoding scheme produces less ambiguous patterns, as across all parameters the percentage measurement errors on average were less than 1 percent in contrast to BRIEF which gave a range of errors up to 10 percent or greater, depending upon the parameters used. This can be seen by comparing figure 35 (Poisson blob) with figure 25 (BRIEF). This improvement in accuracy could be attributed to the removal of ambiguity resulting from the exclusion of uninformative regions, as seen in figure 37. It may also be attributed to the increased amount of information encoded about image content via the addition of a size field allowing groups of similar BRIEF patterns to be differentiated based upon their local abundance.

An inspection of the correlation matrices and distribution of correlation coefficients shows another improvement, with diagonal and off-diagonal terms moving closer to what would be expected if model assumptions were not being violated. The spread of both terms is smaller than for BRIEF, especially the diagonal terms as seen by comparing figure 36 (Poisson blob) with figure 26. Whilst all correlations have not been completely removed it is clear that the Poisson blob representation is much closer in behaviour to what is required by the modelling method than the BRIEF alternative.

7.4.3 Poisson blob limitations

Despite the many improvements, on average all terrain surface area measurements were less accurate than theory would predicted by a factor of 2 to 3. The residual correlations between blobs were the likely cause of these discrepancies, suggesting that there was still some form of Poisson event double counting occurring. Figure 38 illustrates this problem by representing blob types as grey levels positioned relative to underlying image features. It can be seen in this figure that ‘echoes’ of features occur in the form of similarly shaped adjacent blobs, always occurring together in a spatially correlated manner. Upon inspection, these spatial correlations between echoes often correspond to the peaks in off-diagonal correlation matrix terms. These effects could be a systematic result of the blob extraction process. The BRIEF descriptors from which blobs are constructed are sampled within a small radius (8 pixels in the experiments) such that the image regions used to construct each blob overlap one another. The echoes observed in the figure are consistent with parts of features straddling the circumference of the BRIEF sampling discs, resulting



0.179442	-0.181217	0.653195	0.192228
0.666735	-0.938494	0.51939	-0.754893
0.367295	-0.182793	**1.866747**	0.202177
-0.102397	-0.417757	0.076033	0.937336

Associated correlation coefficient
Blob echoes and corresponding image structure

Figure 38: This illustration shows the link between correlated spatial features and correlation matrix elements. The split image on the left shows Poisson blobs sampled from a subset of a martian terrain image and how they correspond to the originating structures. The 'echoes' seen match the strong correlations in the associated correlation matrix.

in blobs occurring some distance away from the discs' origins.

Removing the echoes may not be necessary to fix the correlation problems. Instead, a sub-sampling approach could be taken where blobs are only recorded in histograms if they are a sufficient distance away from one another to be no longer correlated. Such an approach would require further work to convert model quantities into surface areas.

7.5 Summary

This chapter has presented an improved image encoding scheme designed to reduce the correlated effects observed when local BRIEF descriptors were used to populate linear model histograms. It has been shown that Poisson-like behaviour is better approximated when BRIEF patterns are clustered into blobs providing a more appropriate input into the algorithms developed thus far. Model fits, error predictions and measurement accuracies all improved when the blob representation was used to sample martian terrain images. However, the method became more complex, as additional computation was required to convert linear model quantities into surface area measurements with associated error predictions.

Despite the improvements, residual correlations still prevented perfect results leading to error predictions which were around a factor of 2 to 3 away from those observed in practice. But the source of these correlations has been identified and a sparse sub-sampling approach has been proposed to fix the problem.

The experiments on synthetic martian images used over the previous two chapters have been of value for selecting an appropriate image encoding method, but they remain somewhat artificial. The next chapter will attempt to apply the methods developed to more realistic data to make quantitative crater Size Frequency Distributions from lunar images.

8 Lunar Crater Counting: Moon Zoo Part 1

The experiments performed in previous chapters have been limited to artificial problems. The Monte-Carlo histograms from chapters 3 to 5 were designed to generate ideal data, conforming fully to the assumptions made by the algorithms applied to them. The synthetic martian terrains, found within chapters 6 and 7, were also constructed to give idealistic surfaces, drawn from a common pool of image tiles to ensure that training and testing images were representative. Idealised Monte-Carlo data has been successfully analysed, yet the more realistic simulated martian terrain data raised problems with underestimated errors. So far it is not clear whether any genuine planetary science data is amenable to the types of analysis proposed.

Work now will attempt to demonstrate the utility of the developed methods on a real planetary science problem, whilst avoiding the pitfalls of a fully automated analysis of complex terrains. The methods will be applied to the filtering of citizen science crater data from the Moon Zoo project, taken around the Apollo 17 landing site. This will be a real analysis task involving the counting of lunar craters, where an extensive list of potential candidate craters has already been provided by non-expert human users, albeit incomplete and with some contamination.

Martian terrain analysis work showed that image data has to be carefully encoded to make it suitable for analysis via linear histogram modelling techniques. The raw Moon Zoo data and associated images also require encoding before analysis. This provides an opportunity to test alternative image representations that may be more appropriate than BRIEF or Poisson blob formats for describing isolated features (lunar craters), as opposed to extended connected regions (martian terrains), which suffer problematic spatial correlations. It will be seen that this encoding task for Moon Zoo data is non-trivial. To make the task more manageable this work will be divided across the next two chapters. This current chapter will focus on preprocessing raw Moon Zoo crater data to create a set of outputs appropriate for populating linear histogram models. The subsequent chapter will focus on the analysis of preprocessed craters in order to estimate the number of true positives, false positives and false negative entries.

8.1 Moon Zoo project

Moon Zoo is a citizen science project allowing members of the public to assist in the interpretation of lunar images [53]. Volunteers are asked to identify features within selected images by highlighting regions using a graphical interface embedded inside a website[8]. A large database of mouse clicks has been accumulated via this website providing the locations and sizes of features which researchers hope to use to answer planetary science questions. Aims of the project include identifying impact craters and novel features,

[8] www.moonzoo.org

and determining the relative abundance of boulders. The counting of impact craters is of particular interest and will be the focus of the next two chapters. The ultimate goal will be to take Moon Zoo data and convert it into crater Size-Frequency Distributions (SFDs)[29]. However, candidate craters identified by non-expert citizen scientists contains duplication, contamination, missing craters and systematic effects which must be quantified before useful SFDs can be created.

The automated counting of craters was noted in section 1.2.1 as being a prime application. A semi-automated solution which combines citizen science data and lunar imagery is a step towards the end goal of full automation, which avoids the need for an initial global image search for potential candidate craters. It is hoped that experience gained here will provide valuable insights into how a fully automated solution could be achieved.

8.1.1 Properties of Moon Zoo crater data

Raw data

The data used within the following chapters was supplied by the Moon Zoo science team in pixel coordinates, relative to two lunar images captured via the Narrow Angle Camera aboard NASA's Lunar Reconnaissance Orbiter [14]. These relate to images M104311715LE and M104311715RE, which contain relatively sparse populations of craters in comparison to the heavily cratered lunar highlands, such that few craters intersect one another. The crater data contains the coordinates of centres and the radii of candidate craters identified by multiple different users. Many craters within this data are highlighted multiple times by different users. This gives correlated groups of candidates, corresponding to the same true crater, with slightly perturbed centres and sizes indicative of the human accuracy attainable using the Moon Zoo interface. Amongst these candidates will be a number of false positives, caused by users erroneously highlighting craters in ambiguous images, or purposefully introducing errors though acts of cyber-vandalism. There are over 40,000 candidate craters in total. A sample of these candidates can be seen in figure 39.

Systematic effects

There are systematic effects found in the data due to a minimum candidate crater size and default crater size imposed by the Moon Zoo graphical interface. This results in a bias towards craters of these fixed sizes through two mechanisms: craters below the minimum crater size are highlighted with erroneously large radii, as users attempt to identify craters which are too small; and larger craters are highlighted with the erroneously small default radii, as users forget to adjust the crater size appropriately. These effects occur on several scales, as Moon Zoo users are presented with images at various fixed zoom levels. Figure 40 shows the candidate crater size distributions, with and without the biased values.

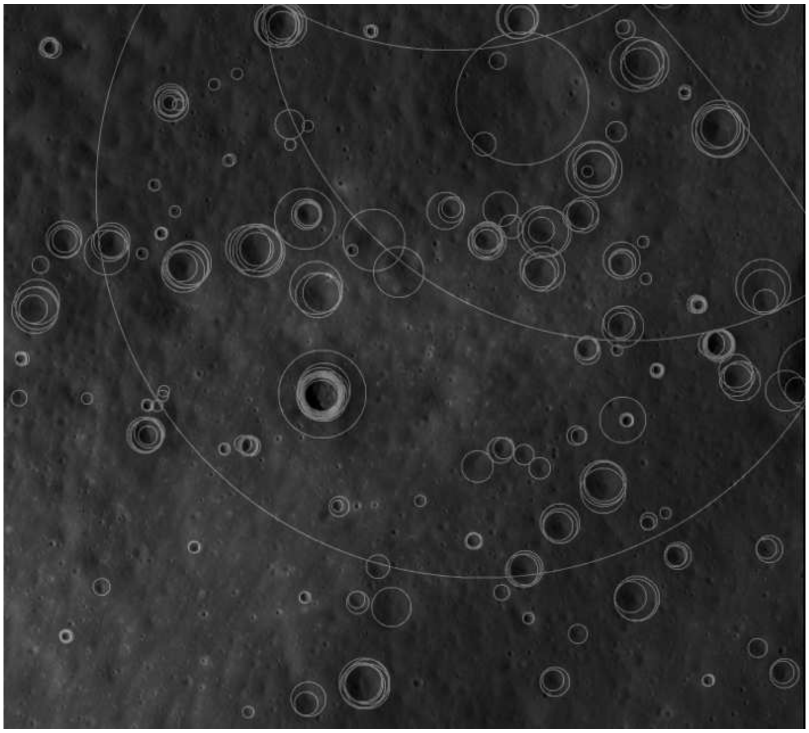


Figure 39: A sample of raw Moon Zoo crater data overlain onto corresponding LROC NAC image frame of lunar surface. Each circle represents a potential crater. Note the clusters of candidates where multiple users have marked the same craters several times.

Missing data

There are also real craters existing in source images which have not been highlighted by users and therefore do not appear in the Moon Zoo database. Some of these false negatives will be due to craters being difficult to spot. Others will be due to random fluctuations in user behaviour and image selection, as Moon Zoo presents images randomly and does not insist that every crater is identified by each volunteer. The missing craters could bias results, giving systematic underestimates in crater counts.

Questions

The challenges involved in converting this data into useful SFDs can be summarised with the following questions:

- Can multiple mark-ups be reliably coalesced into uniquely identifiable craters?
- Can the parameters of candidate craters be corroborated against image data, e.g. to spot default crater size effects?
- Can the quantities of false positive contamination and true positive craters be estimated using linear histogram models?
- Can the above quantities be measured consistently within predicted errors under different conditions?

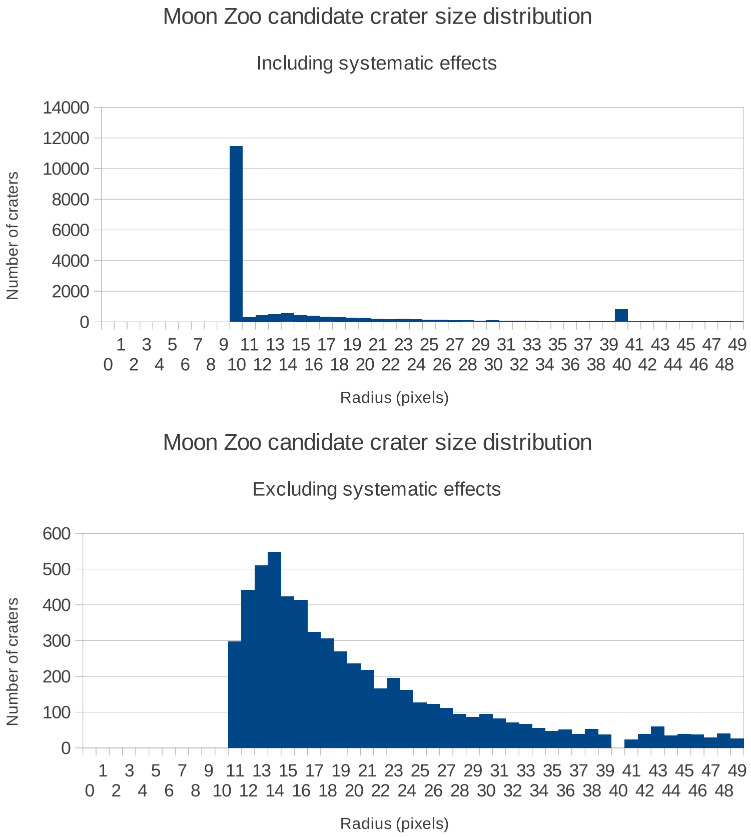


Figure 40: Top: Candidate Moon Zoo crater radii distribution, including biases due to default crater size. Bottom: Radii distribution with systematically biased crater size bins removed.

- Can the quantities of false negative missing craters be accounted for?

These questions naturally lead to a processing pipeline, each stage of which will require the application of appropriate statistical methods.

8.2 Crater processing pipeline

The following set of processing stages will be investigated over the following two chapters:

1. Coalescence: the multiple x, y and radius parameters from different users will be clustered to give a consensus as to the location and size of individual craters (chapter 8).

2. Refinement: x, y and radius parameters for each crater will be refined by searching locally within images for optimal matches against crater template images (chapter 8).

3. Linear Modelling: linear histogram models will be applied to histograms of template match scores to estimate crater counts and background contamination, with associated error estimates (chapter 9).

4. False Negative Calibration: any underestimation in SFDs caused by missing craters will be corrected for by calibration against preprepared ground truth SFDs (chapter 9).

The first step must aim to find a deterministic clustering method which will produce consistent, uniquely identifiable craters. To perform this task optimally, the algorithm employed must make efforts to approximate a Likelihood solution, thereby giving the most probable set of unique craters in light of the mouse click evidence available. The second step must aim to reconcile as many discrepancies as possible between candidate craters and underlying image evidence. To perform this task optimally, the algorithm employed must make efforts to model the appearance of typical Moon Zoo craters and match them to candidate craters, thereby giving the most probable parameters for the image evidence available. Both of these steps will be investigated during the remainder of this chapter.

The linear modelling step must aim to construct well-behaved histograms for true and false positive craters, which can be fitted to Moon Zoo data in order to estimate the number of genuine craters within the data. The calibration step must compare SFDs produced by the above pipeline with those prepared under expert supervision. These two steps will be investigated separately in the next chapter.

8.3 Step 1: Coalescence

When a single crater is highlighted by multiple users, a cluster of related candidate craters appears in the data. Each candidate will be slightly different, with a spread of centres and radii corresponding to users' abilities to measure crater locations and scales. The accuracy may also be affected by the degradation of craters, with older craters having less well-defined rims and therefore less precisely measured parameters. This randomness around true crater parameters prevents a trivial solution in which identical candidates are merged. Random parameter errors mean the identicality of candidate craters can only be considered probabilistically. This section will investigate a method to assign such noisy clusters to individual craters using an approximate Likelihood solution, given a small set of assumptions about the data. It will be assumed that:

- there is an unknown number of true craters, each associated with zero or more candidate mark-ups;
- each candidate is independent;
- **any** candidate could belong to **any** true crater;
- the noise on a candidate's x, y and radius parameters are Gaussian distributed;
- the probability of multiple true craters heavily overlapping is negligible.

The method described in the following subsections will attempt to find the most probable set of true craters, given the proximity of candidates to one another and the overlap between their Gaussian distributed parameters. At this point the problem of false positives and false negatives will be ignored, as will the systematic bias in radii due to the Moon Zoo minimum crater size.

8.3.1 A Likelihood solution

An algorithm for approximating a Likelihood solution to the clustering problem can be developed by first considering an individual candidate crater in isolation, then considering the effects of adding further candidates. If there is only a single candidate with parameters (x_1, y_1, r_1) and errors $(\sigma_x, \sigma_y, \sigma_r)$, the probability of finding a true crater within the range (x_l, y_l, r_l) to (x_u, y_u, r_u) is simply

$$P_{single}(crater) = \int_{r_l}^{r_u} \int_{y_l}^{y_u} \int_{x_l}^{x_u} \mathcal{N}(x, y, r; x_1, y_1, r_1; \sigma_x, \sigma_y, \sigma_r) dx dy dr \qquad (120)$$

where $\mathcal{N}$ is a 3D normal distribution with means x_1, y_1, r_1 and widths σ_x, σ_y, σ_r. If there is only a single true crater and a single candidate then the most probable parameters of the true crater are those where $P_{single}(crater)$ peak. This will trivially correspond to the single candidate's originally assigned values. Now, if a second candidate crater is added with parameters (x_2, y_2, r_2) there are two alternative interpretations. The first interpretation assumes that there is still only a single true crater, in which case the probability of finding this one true crater within the same range becomes

$$P_{and}(crater) = \left(\int_{r_l}^{r_u} \int_{y_l}^{y_u} \int_{x_l}^{x_u} \mathcal{N}(x, y, r; x_1, y_1, r_1; \sigma_x, \sigma_y, \sigma_r) dxdydr\right) \tag{121}$$

$$\times \left(\int_{r_l}^{r_u} \int_{y_l}^{y_u} \int_{x_l}^{x_u} \mathcal{N}(x, y, r; x_2, y_2, r_2; \sigma_x, \sigma_y, \sigma_r) dxdydr\right)$$

where the two independent candidates can be combined multiplicatively, consistent with the 'and' relation in probability theory. This 'and' relation is appropriate, as in this interpretation candidate 1 **and** candidate 2 both originate from the same true crater. The most probable parameters of the one true crater can then be found by searching for the largest peak in $P_{and}(crater)$ (assuming there is one clear peak). Alternatively, the second interpretation assumes that there could be any number of true craters, such that both candidates could correspond to the same crater or two different craters. In this second interpretation the probability of finding any true crater within the above range becomes

$$P_{or}(crater) \propto \left(\int_{r_l}^{r_u} \int_{y_l}^{y_u} \int_{x_l}^{x_u} \mathcal{N}(x, y, r; x_1, y_1, r_1; \sigma_x, \sigma_y, \sigma_r) dxdydr\right) \tag{122}$$

$$+ \left(\int_{r_l}^{r_u} \int_{y_l}^{y_u} \int_{x_l}^{x_u} \mathcal{N}(x, y, r; x_2, y_2, r_2; \sigma_x, \sigma_y, \sigma_r) dxdydr\right)$$

where the two independent candidates are combined additively, consistent with the 'or' relation in probability theory. This 'or' is appropriate, as in this interpretation either candidate crater could have originated from one true crater **or** another. In this case the probability is only proportional, as the total normalisation becomes dependent upon the unknown number of true craters. Both the 'and' and 'or' interpretations can be extended to any number of candidates in the Moon Zoo data. Either interpretation will be correct for some parts of that data and not others, as clearly some candidates are associated with the same true crater, yet there are many different true craters.

On average, within the Moon Zoo data, candidates will belong to different craters, making the 'or' interpretation preferable. Plus, despite Likelihood functions generally being multiplicative, additive components are not unprecedented. For example, the EM Likelihood function of equation (27) contains a summation to account for data originating from one class or another. An 'or' based Likelihood may be statistically less efficient, but should still result in valid estimates when summed over all candidates, i:

$$\mathcal{L} = \sum_i \left(\int_{r_{il}}^{r_{iu}} \int_{y_{il}}^{y_{iu}} \int_{x_{il}}^{x_{iu}} \mathcal{N}(x, y, r; x_i, y_i, r_i; \sigma_x, \sigma_y, \sigma_r) dxdydr\right) \tag{123}$$

Peaks found within this function can be interpreted as belonging to individual craters. It must be noted, however, that closely overlapping true craters may become erroneously combined into single peaks. But, the Moon Zoo data is quite sparse, with the explicate assumption of rarely interacting craters already being made at the start of section 8.3. An implementation of this function would resemble the accumulation of weighted votes within a parameter space, like a probabilistic Hough transform.

Given the Moon Zoo data, the only unknowns remaining in the Likelihood formulation are the values of σ_x, σ_y and σ_r, and their relationships to crater parameters, e.g. are they all constants, are they all the same etc.? Estimating these values will be the topic of the next subsection in order to inform the development of a more concrete Hough transform-like algorithm.

8.3.2 Parameter accuracy

Incorporating equation (123) into a working clustering algorithm requires knowledge of the accuracies of crater x, y and radius parameters. This knowledge will determine how a Hough transform parameter space should be constructed and how much weight should be given to entries.

The most straightforward method of measuring the accuracies of these parameters might be to manually cluster a set of candidates believed to be associated with a single crater, then compute sample standard deviations directly. Unfortunately, this approach is problematic because there are relatively few mark-ups in each cluster. There are as few as one, and up to approximately a dozen candidates per cluster, based upon visually inspecting candidates overlaid onto lunar images such as in figure 39. This is far fewer than the 30+ samples recommended by standard statistical texts, which would give very poor σ estimates. It is therefore necessary to combine samples from multiple clusters, ideally without having to manually (and subjectively) cluster the data first. This can be achieved by sampling the deviations between all candidates, irrespective of their true cluster, then removing the background contamination introduced by not separating the clusters first.

Let $X = \{x_1, x_2, \ldots, x_n\}$, $Y = \{y_1, y_2, \ldots, y_n\}$, and $R = \{r_1, r_2, \ldots, r_n\}$ be vectors containing the x coordinates, y coordinates and radii parameters of n candidate craters, respectively. If there was only 1 candidate per crater and assuming that changes in crater density due to surface age differences will be averaged over wide areas, the X and Y vectors should be approximately uniformly distributed. The R vector will be distributed according to a size-frequency distribution, with fewer large radii and many more smaller radii, resembling an exponential decay curve. However, as there are multiple candidates per crater these distributions will become 'lumpy', with local clusters interrupting the basic distribution shapes. Each 'lump' will correspond to a different cluster, all of which can be aligning to a common origin before plotting deviations, superimposing the parameter error distributions on top of the natural parameter distributions.

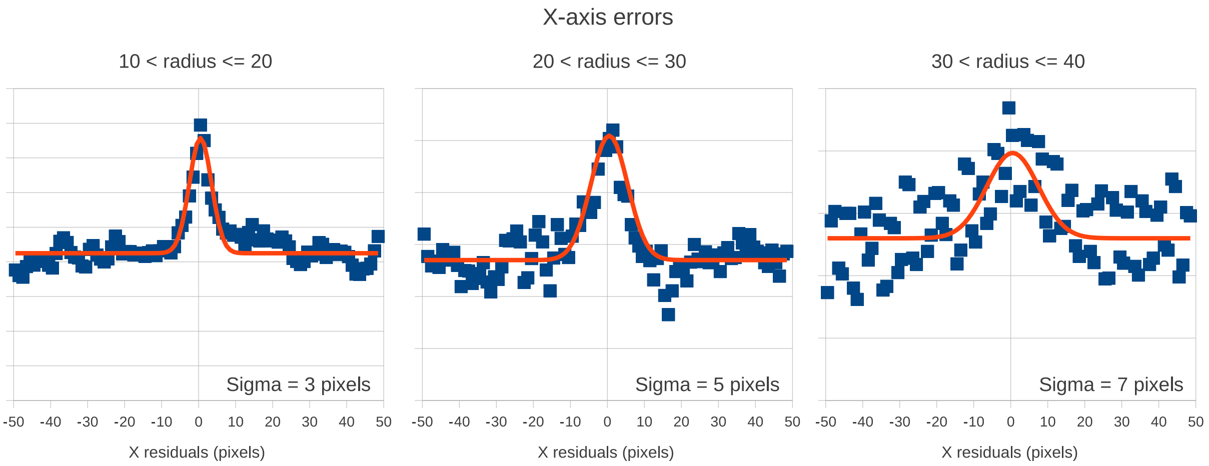


Figure 41: Distribution of X parameter accuracies as a function of crater size, with mean crater radii of 15, 25 and 35 pixels, respectively. The Gaussians fitted all have widths approximately 0.2 times the crater radii.

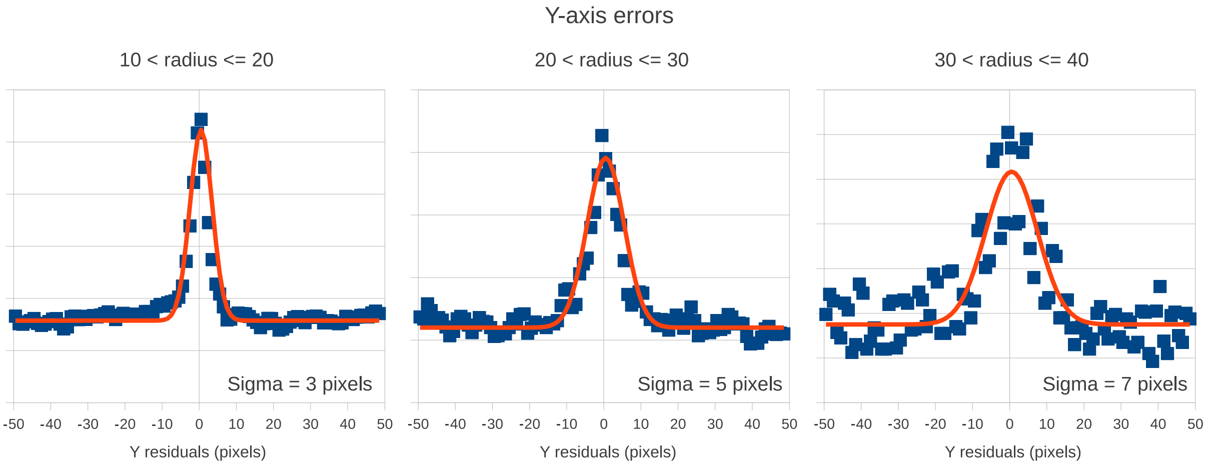


Figure 42: Distribution of Y parameter accuracies as a function of crater size, with mean crater radii of 15, 25 and 35 pixels, respectively. The Gaussians fitted all have widths approximately 0.2 times the crater radii.

More precisely, the distribution of the differences between each combination of candidate, e.g. the distribution of $x_1 - x_2$, $x_1 - x_3$, to $x_1 - x_n$ followed by $x_2 - x_3$, $x_2 - x_4$, to x_n etc., giving a total of $\frac{n(n-1)}{2}$ comparisons, can be plotted. Plotting the differences this way has the effect of making each candidate the origin of the distribution, i.e. aligning the 'lumps'. This can be done for various cuts of the data. The same can be done for Y and R vectors.

This process was performed using the Moon Zoo data with cuts only allowing entries to be plotted that are within a few diameters distance from the origin to reduce the uniform/SFD backgrounds, e.g. if $x_1 - x_2$ was larger than twice the diameter of either crater 1 or 2 then it was not included in the plot. Figures 41, 42 and 43 show these distribution for each parameter, with plots given for three ranges of crater size. The size bands investigated covered radii between 10 to 20, 20 to 30, and between 30 to 40 pixels.

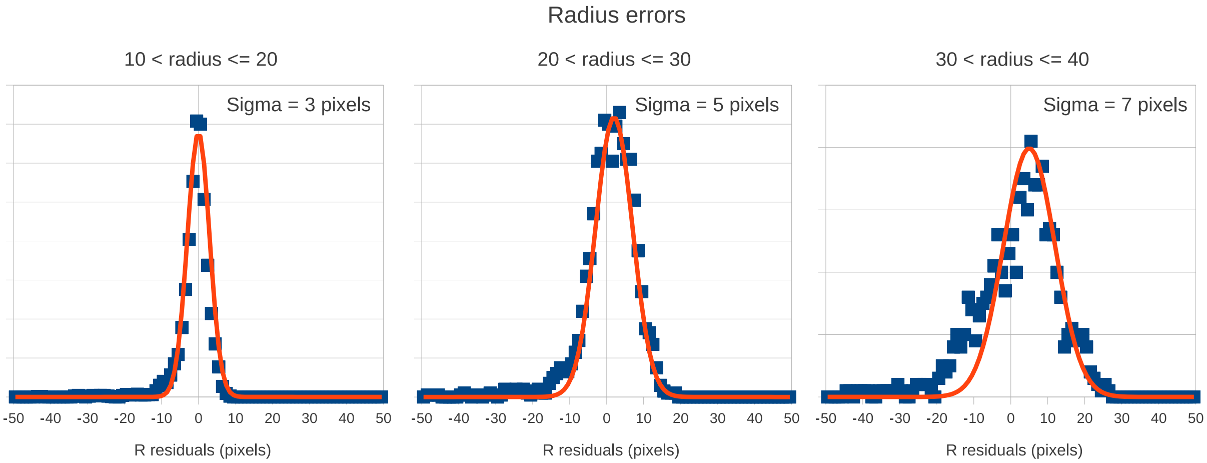


Figure 43: Distribution of R parameter accuracies as a function of crater size, with mean crater radii of 15, 25 and 35 pixels, respectively. The Gaussians fitted all have widths approximately 0.2 times the crater radii.

Figures 41 and 42, showing the X and Y distributions, both contain Gaussian central peaks sitting on top of a uniform background. Figure 43, showing the R distributions, contain Gaussian central peaks which have been slightly skewed by the exponential background of size-frequencies.

In each case Gaussian curves have been roughly fitted manually[9], with standard deviations of 3, 5 and 7 pixels for respective size bands. These correspond to average parameter errors with standard deviations which are 20% of a candidate's radius. The plots provide evidence of a link between measurement accuracies and crater radii making the error characteristics of the crater parameter space non-uniform:

$$\sigma_x = c_x r \tag{124}$$

$$\sigma_y = c_y r \tag{125}$$

$$\sigma_r = c_r r \tag{126}$$

where c_x, c_y and c_r appear to be the same constant of 0.2. Note that the observed behaviour is an average across all types of crater, not accounting for states of degradation.

The observation that errors are radially dependent has important consequences for the construction of a clustering algorithm reliant upon the proximity of points. Whilst it is possible to use Pythagorean arguments for the closeness of two points, their statistical closeness (i.e. the probability of their closeness) must be argued in terms of measurement

[9] Background contamination from spatial and size distributions of craters makes it difficult to reliably apply a fitting routine to estimate curve widths. Upon inspection, the manually fitted curves are convincing enough for the purposes of estimating parameter errors.

accuracies. If a parameter space is uniform and isotropic then a given length will have the same statistical significance no matter where it is measured in that space making euclidean comparisons compatible with their statistical counterparts. If errors are not uniform or isotropic this will not generally be the case, such that the same distance in one part of the space could have a difference significance than the same distance in another. Ideally, a parameter space would be chosen which provides complete uniformity of statistical significance in all directions allowing euclidean distances to be compared directly. This is known as a homoscedastic space [82]. The non-uniform errors found within the Moon Zoo crater data, however, forms a hetroscedastic space, which complicates the coalescence problem.

8.3.3 Coalescence algorithm

This subsection will now combine the Likelihood function $\mathcal{L}$ (equation 123) and knowledge of parameter errors to form a probabilistic Hough transform [95][96] style clustering algorithm.

The first step is to construct a quantised parameter space containing x, y and radius dimensions to act as the Hough transform space. Then rather than explicitly implementing equation (123), point entries can be accumulated in the parameter space for each candidate before being smoothed with a Gaussian kernel with widths commensurate to the accuracies of the crater parameters. This smoothing approximates the normally distributed values of equation (123). The Hough parameter space can then be searched for peaks which can be taken as the most likely candidates for the crater parameters being sought.

In light of the non-uniform errors the actual Hough parameter space must be smoothed carefully. Whilst x and y errors vary as a function of radius, within any given radial slice σ_x and σ_y are uniform and isotropic (at 0.2 of the radius), such that a 2D Gaussian smoothing kernel can be convolved with the slice giving the desired effect. However, a vertical slice through different radii will have smoothly varying σ_r from top to bottom preventing a simple convolution with a fixed width kernel. Fortunately, a simple transformation exists for converting the proportional radial errors into a homoscedastic space:

$$r'_i = \frac{\ln r_i}{c_r} \tag{127}$$

where the transformed variable r' is forced to have constant unit width errors. The updated parameter space (x, y, r') is a preferred space for the Hough transform permitting easy smoothing via convolution whilst maintaining the statistical significance of distances in the radius dimension.

The width of the bins in the Hough transform space should be selected for consistency with the Rayleigh Criterion [173], i.e. the bin widths can be no wider than half the distance at which adjacent peaks are resolvable, which for the Moon Zoo data is approximately half of the smallest measurement error. The smallest crater candidates within the Moon Zoo data are 10 pixels in radius, making the smallest measurement errors approximately 2 pixels (0.2 proportional error).

Once populated, the bins can be searched for peaks corresponding to the most likely craters. The collection of peak parameters can be converted back into the data's native parameter space, from r' to r, using the inverse transformation. For practical implementation reasons the resolution of the parameter space can be limited. However, a 3D quadratic interpolation scheme can be used to estimate parameters to sub-bin level accuracy.

8.3.4 Testing

Clustering mistakes can occur when two or more distinct craters are coalesced despite being separate. This will result in some craters being unaccounted for, introducing additional false negatives. This may occur when the craters are very close to one another, intersecting or even nesting. This particular mode of failure is not expected to occur often, as it was assumed at the start of section 8.3 that the probability of craters interacting was negligible. Mistakes can also occur when two or more mark-ups of a single crater are highly dissimilar leading to multiple craters being generated when only a single crater exists. This mode of failure will introduce additional false positives. A lack of objective ground-truth accompanying the Moon Zoo data precludes testing clustering efficiency directly on the target data. However, the data generation mechanism (excluding systematic effects due to default crater sizes and false positives) is sufficiently simple to construct a realistic Monte-Carlo to empirically evaluate clustering performance. Subsequent clustering of real Moon Zoo data can then be performed with performance estimated by comparison to clustering efficiencies in a similar Monte-Carlo, e.g. similar size image, number of craters, number of mark-ups etc.

Monte-Carlo data was generated by randomly defining true crater locations and radii within different set ranges. For each defined crater a number of random mark-ups was drawn from a Normal distribution with widths set to 20% of the radius, consistent with the accuracies measured from the real data. Datasets were generated for different numbers of craters, crater size ranges, image sizes and different numbers of multiple mark-ups per crater. Each dataset was entered into the Hough transform clustering algorithm, with the number of output clusters compared to the correct number of craters.

8.3.5 Discussion

Figure 44 shows the efficiency of clustering over a range of conditions. There are general trends showing that for sparse data (low crater densities) containing few multiple mark-ups (less than 10 candidates per true crater) there is an initial over estimate of craters with some counts growing up to 10% above ground truth. The additional peaks in the Hough transform space can be explained by there being insufficient evidence populating local regions preventing candidates from the same crater to be smoothed together effectively. Close to this point, for sparse data containing larger numbers of mark-ups (greater than 10 candidates per true crater) then efficiencies between 90% to 100% are achievable, as there are few interacting craters and sufficient local evidence to smoothly merge candidates giving distinctive peaks. Figure 46 shows results attainable in this range of data. However, as the coverage of craters increases the opportunity for craters to interact increases, leading to a decrease in efficiency as adjacent craters begin to be inappropriately merged. As craters begin to saturate the data, efficiencies can be seen to drop below 20%. The sub-figures all show a decrease in efficiency as a function of increases in crater density. An increase in density can be due to increasing total crater counts, decreasing image sizes, increasing numbers of mark-ups per crater, and increasing the size of craters. Finally, figure 48 shows how the number of mark-ups per crater relate to Hough peak heights, which will be useful for estimating the number of times candidates in Moon Zoo data are highlighted by users.

The Apollo 17 Moon Zoo crater data was clustered using the same method, with and without the biased crater size bins. There were a total of 23,957 candidate craters in the original data, which was reduced to 9,057 craters after coalescence. There were a total of 11,419 candidate craters when biased crater sizes were excluded, which was reduced to 4,350 craters after coalescence. These craters were distributed over a large image of approximately 5,000 by 50,000 pixels. Figure 47 shows the non-cumulative size-frequency distributions for these two clustered results. Figure 45 plots the distribution of peak heights when biased sizes are excluded. Finally, figure 49 shows a small sample of clustering results overlaid on to the Apollo 17 NAC image region. The candidate crater densities (candidate crater count divided by image area) of 9.6×10^{-5} and 4.4×10^{-5}, including and excluding biased bins, respectively, are on the low end of the Monte-Carlo testing ranges, suggesting that clustering is being performed with good efficiency (very few genuine craters will be lost), with some double counting of craters due to some failures in merging. Visual inspection of craters confirms this in sample regions. This provides an answer to the first question posed in section 8.1.1, that it is possible to reliably cluster multiple mark-ups. Any failed mergers could potentially be fixed during the subsequent pipeline stages. Failed merger leading to new false positives or false negatives can be quantified in the final steps which will be covered in the next chapter.

8.4 Step 2: Refinement

Systematic effects, user subjectivity and inefficiencies in clustering can combine to produce inconsistencies between coalesced crater parameters and actual evidence in underlying image data. Craters marked using the minimum/default crater sizes and craters which have only been marked up by one or two users might be particularly prone to systematic and subjective effects. A refinement stage can enforce consistency by matching a set of crater

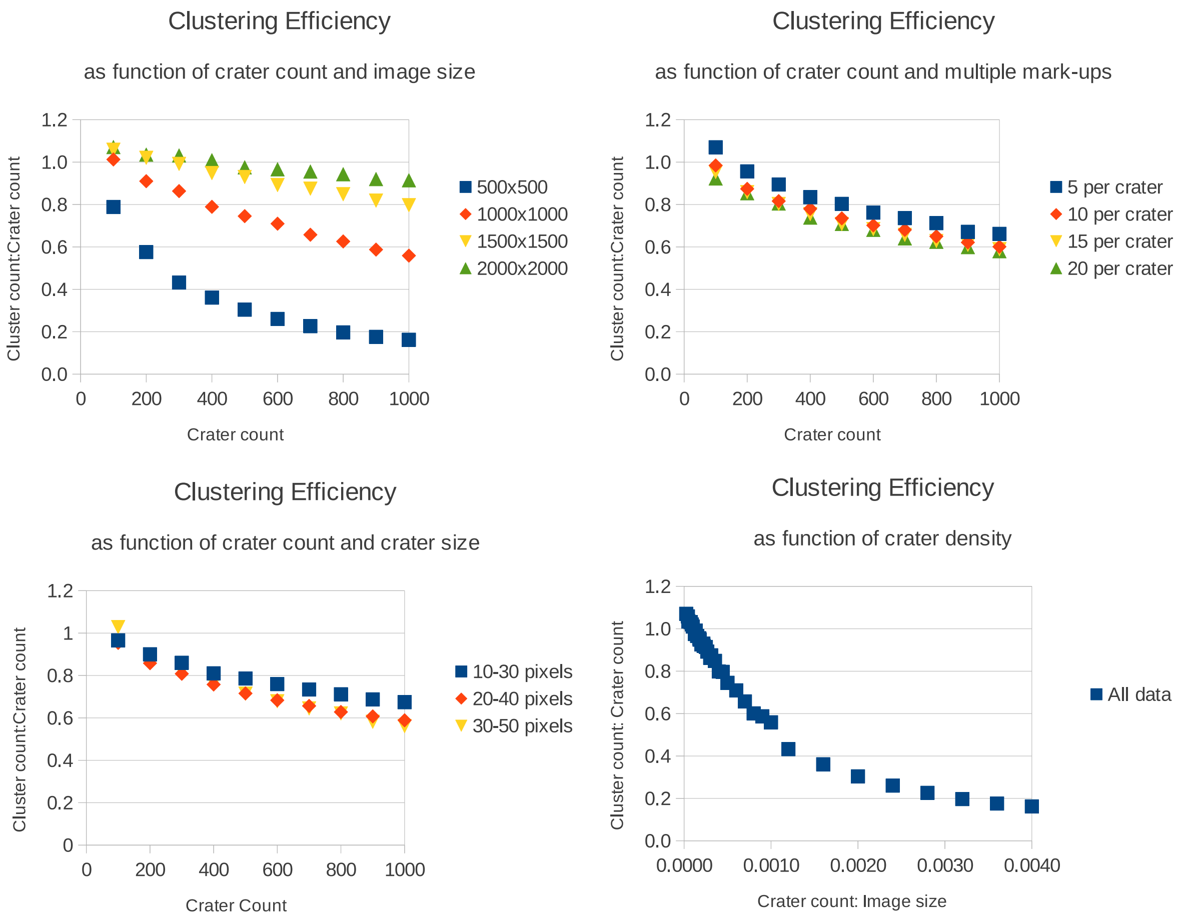


Figure 44: The efficiency of the Hough transform clustering algorithm. Top left: Clustering efficiency (cluster count divided by true crater count) as a function of true crater count, with curves for 4 sizes of image. Top right: Clustering efficiency as a function of true crater count, with curves for different numbers of candidates per true crater. Bottom left: Efficiency as a function of true crater count, with curves for different crater size distributions. Bottom right: Efficiency as a function of crater density (crater count divided by image area).

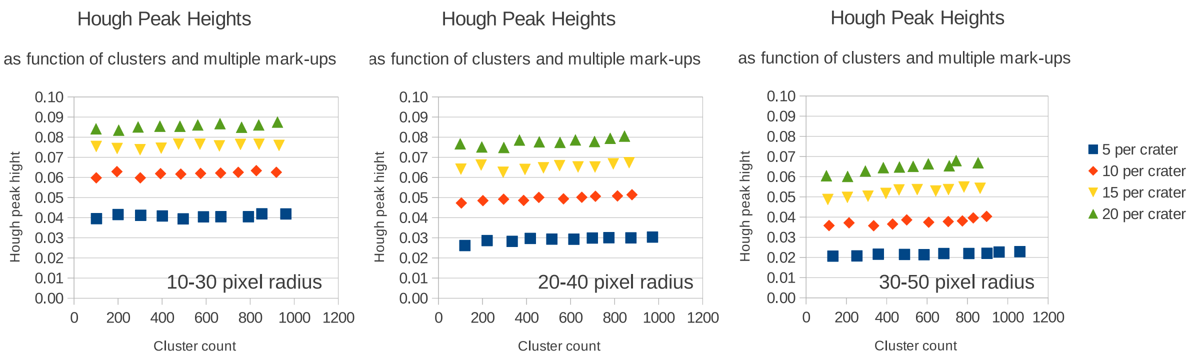


Figure 45: Typical peak heights within Hough transform parameter space for different ranges of crater size

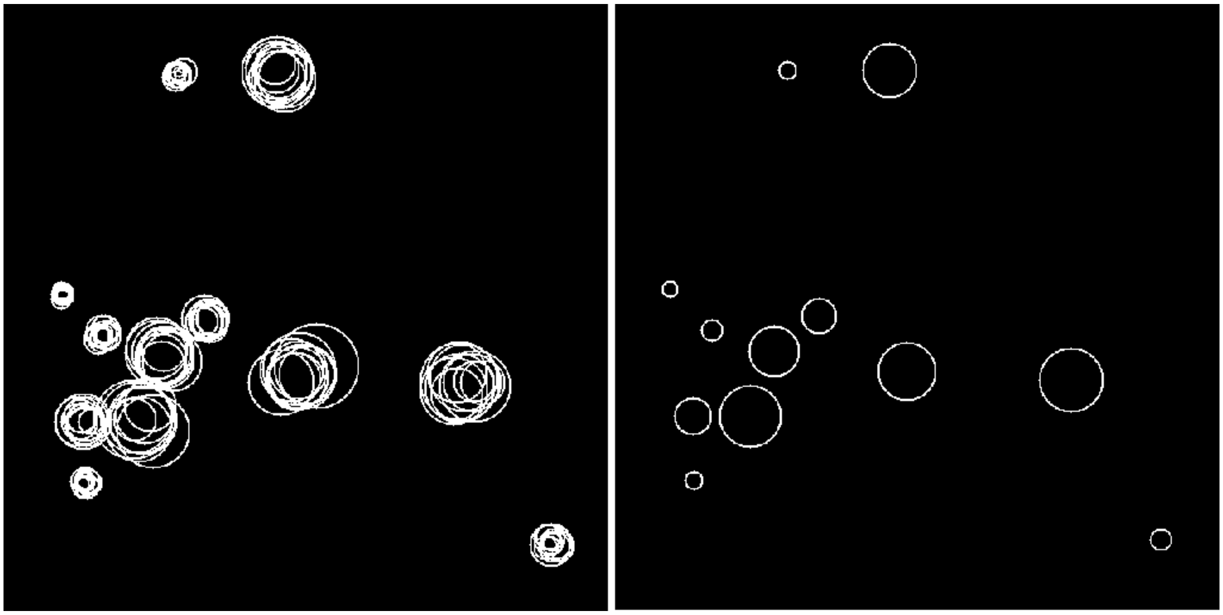


Figure 46: Example before and after clustering of Monte-Carlo data for sparse craters with 10 candidates per crater. Efficiencies close to 100% can be achieved with data in this range.

image templates to candidates, locally searching centres and radii to find optima within some match score. For the approach to work it requires reliable, representative templates to be defined. Crater templates must accommodate differences in local albedo and illumination, morphology and states of degradation. Ideal templates would also accommodate crater interactions and occlusion conditions, but this will be omitted due to the assumption of rarely interacting craters. These issues will be addressed in the following subsections.

8.4.1 A Likelihood solution

The refinement of crater locations though the matching of templates can be linked to Likelihood by adopting the standard assumption of uniform Gaussian noise on pixel values. Given a template of pixels, a, and an image patch centred and scaled to a specific x, y, and radius, b, the residuals between each pixel, $a_i - b_i$, should also be distributed as a Gaussian, $e^{-(a_i-b_i)^2}$. Incorporating this into a Likelihood suggests that a sum of squares is an appropriate function for computing template match scores:

$$\mathcal{L}_{match} \propto \prod_i e^{-(a_i-b_i)^2} \tag{128}$$

$$\ln \mathcal{L}_{match} \propto -\sum_i (a_i - b_i)^2 \tag{129}$$

Finding peaks in this Likelihood, or a function closely approximating it, can therefore be defended as being a good choice for a template matching algorithm. However, this Likelihood interpretation assumes the only sources of template matching errors are in the

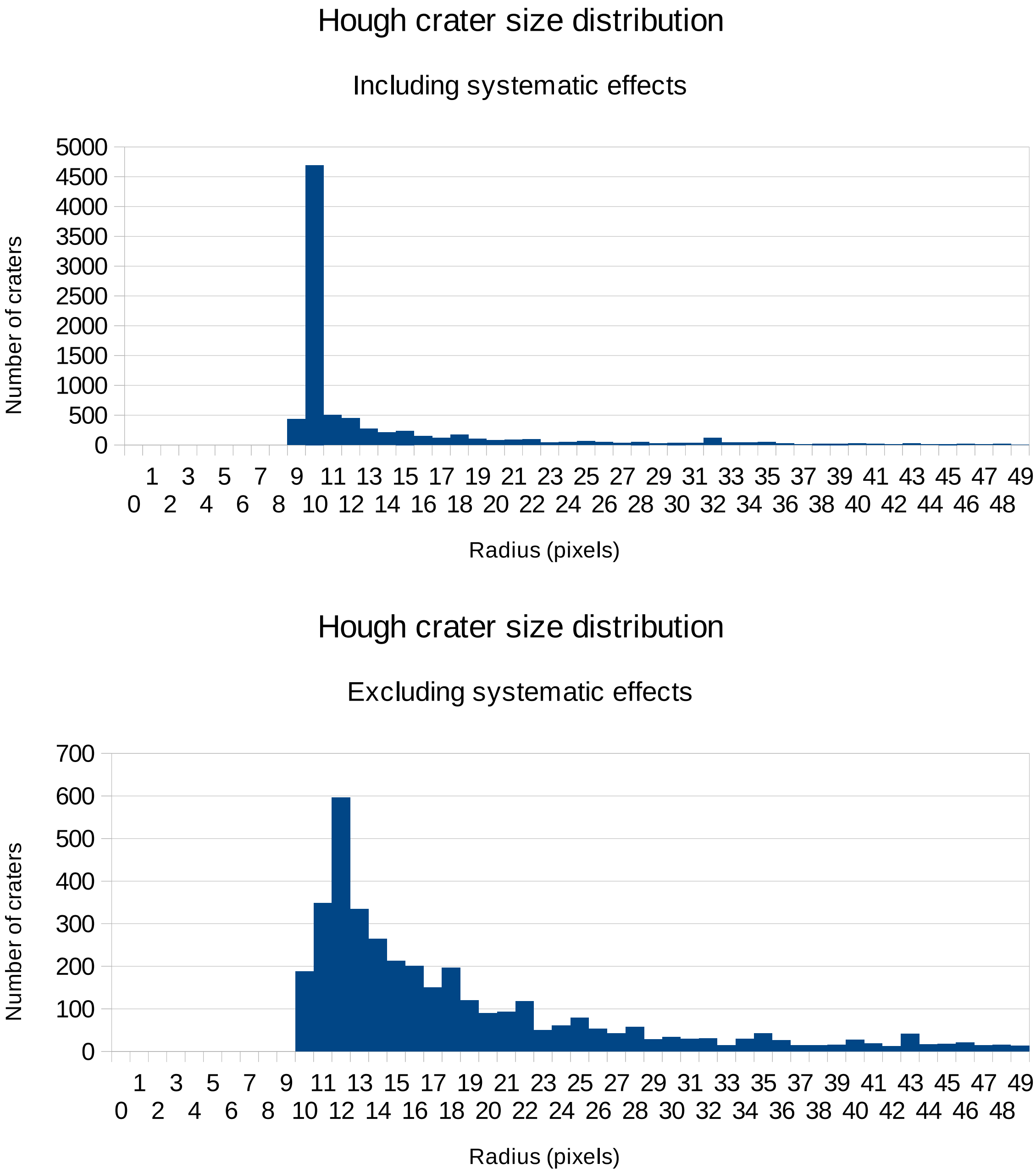


Figure 47: Crater size distribution of Moon Zoo data after Hough coalescence step.

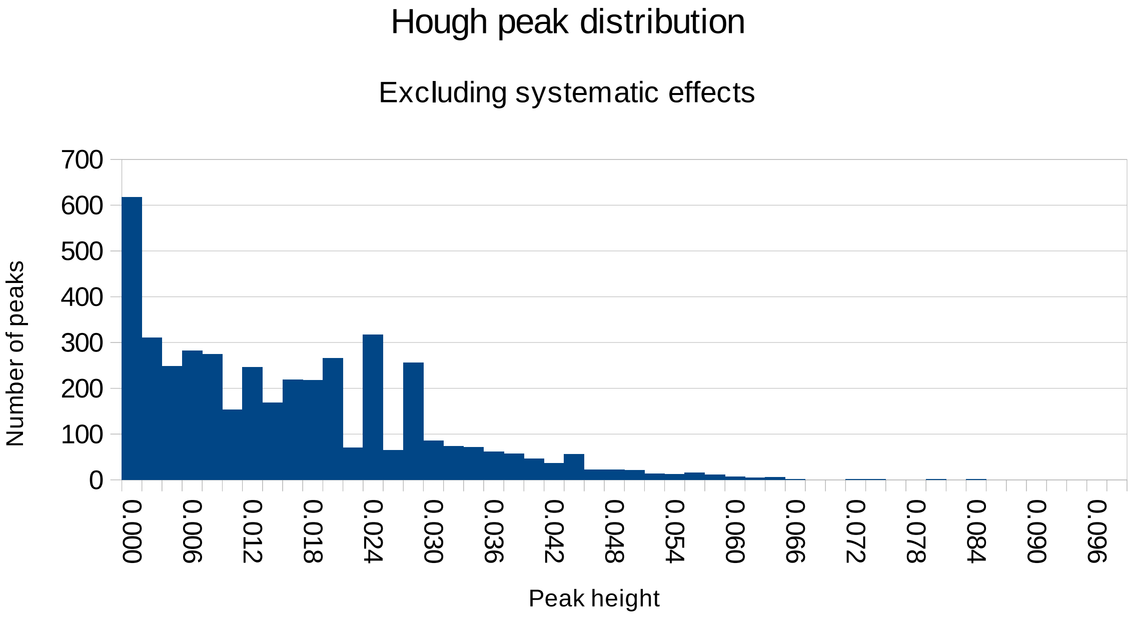


Figure 48: Distribution of Hough peak heights in Moon Zoo coalesced data.

pixel values themselves. In reality craters contain additional sources of uncertainty due to local conditions, morphology and erosion. Steps to mitigate against this will be suggested during the selection of templates and match scores.

8.4.2 Template selection

The first requirement for the generation of template craters is to find a source of example craters. Given the availability of clustered candidate craters in the Moon Zoo data, it is convenient to make use of the Moon Zoo data itself to guide template construction, as opposed to using expertly prepared examples. It is reasonable to assume that craters highlighted by multiple Moon Zoo users are more likely to be true craters than those which have been highlighted only once. Under this assumption templates can be created only from craters which have been highlighted more than a given minimum number of times.

Given the variability in crater appearances steps should be taken to make templates as representative as possible, i.e. to remove additional variations beyond just pixel Gaussian noise. This might be achieved by only retaining the most invariant attributes of a crater. Firstly, craters can be scaled to a known fixed size to better align their rims. Templates' mean grey levels can be subtracted to remove regional illumination and albedo effects within the vicinity of the candidates which could otherwise cause large discrepancies between templates are target craters. Derivative images can also be used to remove more localised illumination effects, as the relative differences between closely adjacent pixel values will be less affected by larger scale illumination or albedo changes across the diameter of a crater.

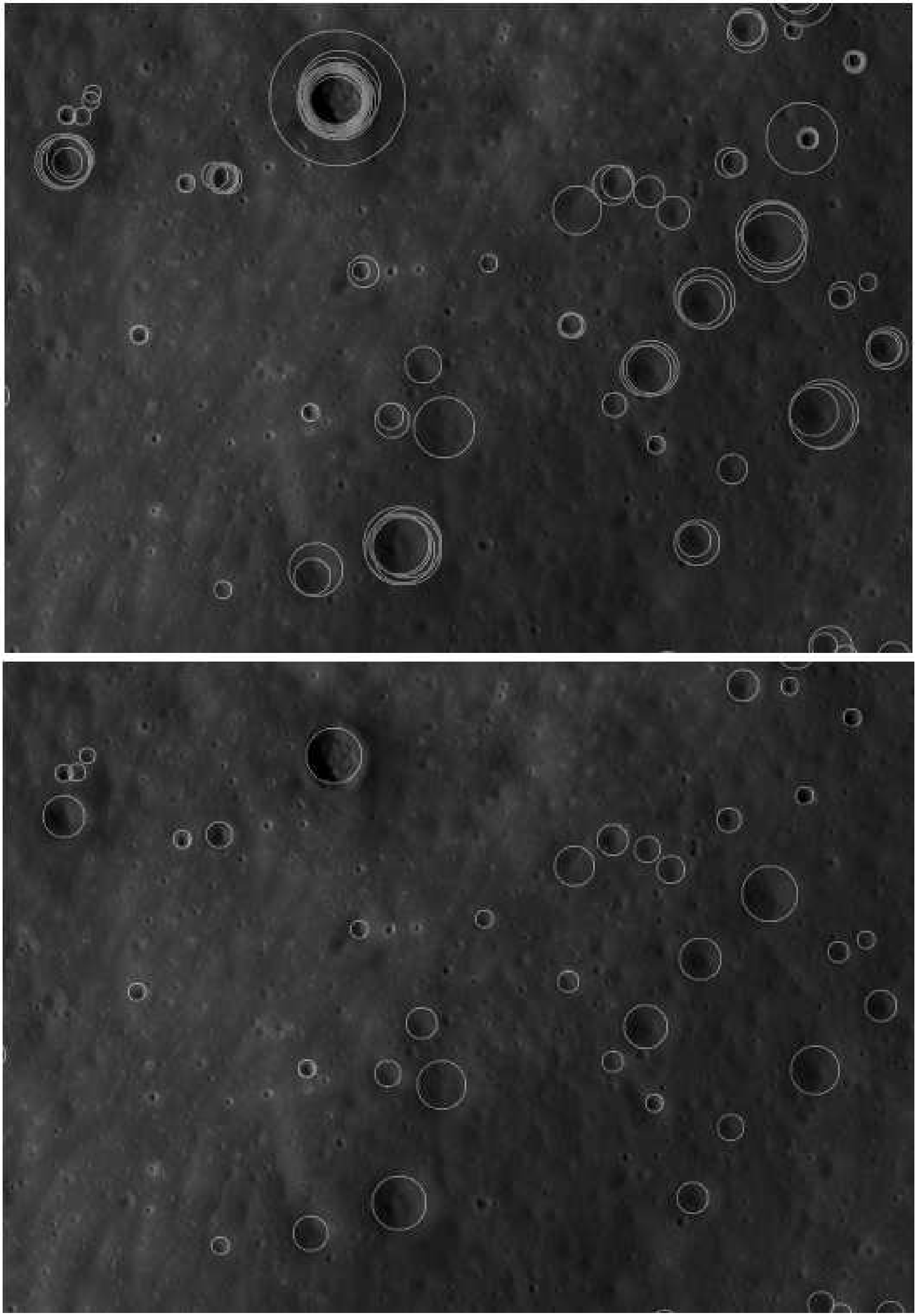


Figure 49: Sample from Moon Zoo Apollo 17 site craters. Top: before coalescence. Bottom: after coalescence.

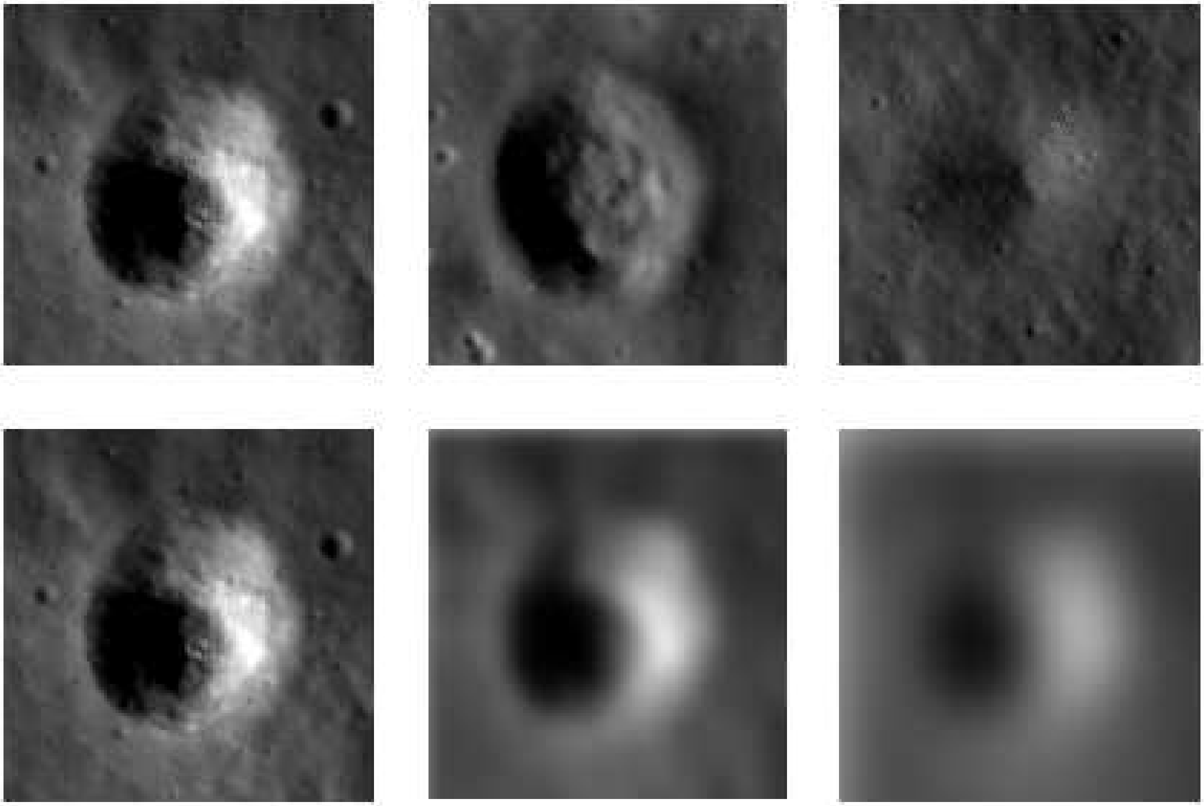


Figure 50: Similarities between physical crater degradation and Gaussian smoothing of crater images. Top row: Apollo 17 site craters showing different levels of degradation. Bottom row: Approximation of degradation levels by successive smoothing of a relatively clear crater.

One of the largest modes of crater variation is the state of degradation. Over time, weathering via micrometeorites and mass wasting causes the contours of a crater to become less well-defined. Visually, degraded craters are similar to fresh crater images which have been smoothed, as can be seen in figure 50. Using this observation, a template can be compared to craters in varying states of degradation using a free parameter which controls levels of smoothing. Refining a crater's position would then involve a search over x, y, radius and smoothing parameters. The addition of a smoothing parameter not only could be used to improve template fits, but could also be used to give a degradation index which could be valuable to researchers.

Two types of fixed sized templates will be investigated: a grey level template modelling an average crater's appearance, with mean grey level subtracted; and a pair of gradient templates, again an average, but computed using x and y derivative images. Figure 51 shows example templates crated from the Moon Zoo data from craters which were marked by 3 or more users. These template use 60 by 60 pixel patches, including a crater of radius 20 and a margin of 20 pixels. The gradient templates are expected to be the most invariant, as they model the relative structure of craters, as opposed to the grey level templates which might still be susceptible to local changes in illumination.

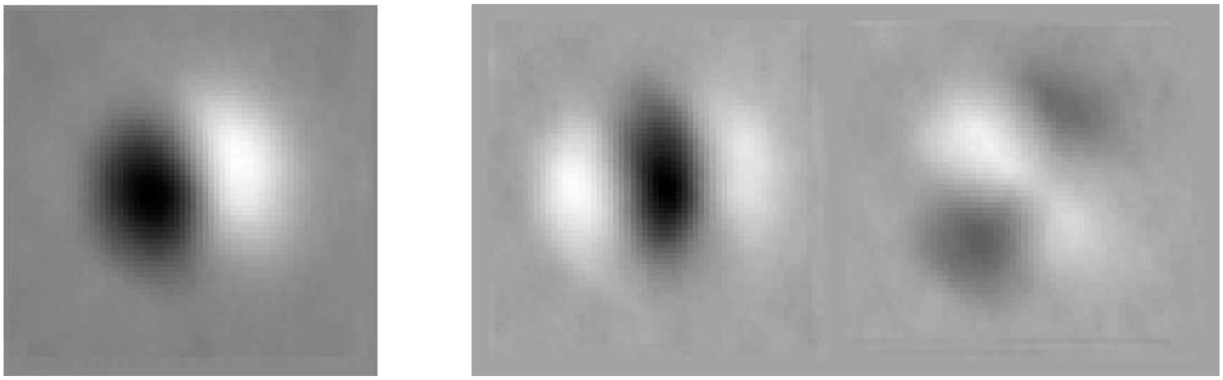


Figure 51: Left: Mean grey level crater template derived from Moon Zoo data. Right: Combined horizontal and vertical gradient (x, y derivative) template.

8.4.3 Match score selection

The match scores which will be analysed are: a mean sum of squared errors (MSE) and a per-pixel normalised dot product (DP). Given a template vector, a, and image vector, b, containing N pixel samples, these match scores become:

$$S_{MSE} = \frac{1}{N} \sum_i (a_i - b_i)^2 \tag{130}$$

$$S_{DP} = \frac{a \circ b}{\|a\| N} \tag{131}$$

It was noted in section 8.4.1 that under standard Gaussian noise assumptions a function proportional to the sum of squared residuals will have peaks consistent with a likelihood solution. The S_{MSE} match score clearly fits this framework.

The dot product style match score contains terms which are sums of products, $a \circ b = \sum_i a_i b_i$, normalised to different values. Similar sums can be seen in the expanded version of the S_{MSE}:

$$S_{MSE} = \frac{1}{N} \sum_i a_i^2 + \sum_i b_i^2 - 2 \sum_i a_i b_i \tag{132}$$

where the first two summation terms will be relatively constant over local areas, with the dot-product style final summation term performing the real template comparison. This observation suggests that the MSE and DP match scores should peak at approximately the same locations, with the MSE score disregarding more of the regional variations. These two match scores will also differ in their error characteristics.

Under the standard assumes that image noise is uniform, normally distributed and independent between samples, error propagation can be applied to these match scores to predict their stability. Assuming the template is error free and there is uniform noise of σ on the image being matched, the error on a match score, σ_S, can be approximated using:

$$\sigma_S^2 = \sum_i \left[\frac{\partial S}{\partial b_i}\right]^2 \sigma^2 \tag{133}$$

This will be performed below for the different matching functions.

MSE errors

The mean squared error match score can be broken into individual terms:

$$S_{MSE} = \frac{1}{N}\sum_i (a_i - b_i)^2 = \sum_i s_i \tag{134}$$

$$s_i = \frac{a_i^2 + b_i^2 - 2a_i b_i}{N} \tag{135}$$

with derivatives for each term given by:

$$\frac{\partial s_i}{\partial b_i} = \frac{2b_i - 2a_i}{N} \tag{136}$$

These can be inserted into the formula for error propagation to give:

$$\begin{aligned} \sigma_S^2 &= \sum_i \left[\frac{2b_i - 2a_i}{N}\right]^2 \sigma^2 \\ &= \frac{4\sigma^2}{N^2}\sum_i (b_i - a_i)^2 \end{aligned} \tag{137}$$

As can be seen, this match score's stability is functionally dependent upon the incoming data. Given a fixed sized template, the errors on a match to that template will vary from image location to image location. This might suggest that this match score will perform with different efficiencies in different parts of the Moon Zoo data, potentially leading to some poorly refined craters.

DP errors

The per-pixel normalised dot product can be analysed following the same pattern as above, beginning with exposing the inner workings of the match score:

$$\begin{aligned} S_{PNDP} &= \frac{a \circ b}{\|a\| N} \\ &= \frac{u}{v} = \frac{\sum_i a_i b_i}{N\sqrt{\sum_i a_i^2}} \end{aligned} \tag{138}$$

Then the derivatives of numerator and denominator, respectively, are given by:

$$\frac{\partial u}{\partial b_i} = a_i \tag{139}$$

$$\frac{\partial v}{\partial b_i} = 0 \tag{140}$$

Applying the quotient rule again gives:

$$\frac{\partial S}{\partial b_i} = \frac{vu' - uv'}{v^2}$$

$$= \frac{N a_i \sqrt{\sum_i a_i^2}}{\left[N\sqrt{\sum_i a_i^2}\right]^2}$$

$$= \frac{a_i \sqrt{\sum_i a_i^2}}{N \sum_i a_i^2} \tag{141}$$

Which when substituted back into the formula for error propagation gives:

$$\sigma_S^2 = \sum_i \frac{a_i^2 \sum_i a_i^2}{\left[N \sum_i a_i^2\right]^2} \sigma^2 \tag{142}$$

This template match score has a stability which is independent of the pixel values contained within the image being matched. Errors on this score will be the same regardless of where it is applied. It is only dependent upon the template and the uniform image noise. This could suggest a more stable, better behaved metric.

8.4.4 Refinement algorithm

The template matching refinement algorithm will utilise both styles of template and match score, testing each possible combination of:

- Grey level template with MSE match score;
- Gradient template with MSE match score;
- Grey level template with DP match score;
- and Gradient template with DP match score.

The algorithm begins with the construction of templates, followed by a candidate by candidate comparison where a brute-force search computes match score values for a range of local x, y and radius parameters. This is repeated for different levels of image smoothing to allow for differences in crater degradation. The refined parameters are those which correspond to the best overall match scores, leading to updated values for x, y and radius parameters. An addition output is the smoothing parameter indicating the level at which the candidates achieved their best matches. The details of template construction and searching are given below.

Template construction

There are thousands of candidate craters in Moon Zoo data which could be used to generate a mean crater template, but this data also contains false positives which could introduce contamination. To avoid this the candidates are first filtered to only include those which have been marked-up by at least 3 users. This is determined by the candidate's peak hight in the coalescence Hough transform space.

The candidates that make the cut are scaled to the dimensions of the desired template, which during subsequent testing becomes 60 by 60 pixels, with the candidate centered with a 20 pixel margin. The mean regional grey level of each example is computed then subtracted. From here, the process forks to create one mean template and one compound gradient template.

The mean template is created simply by computing a mean image from all of the examples. The gradient template is created by computing two derivative images per example: one by calculating pixel value differences at a 4 pixel horizontal offset; the other by using a 4 pixel vertical offset. Mean horizontal and vertical gradient images are computed from all of the examples, then tiled together next to one another making a single gradient template 120 by 60 pixels in size. The grey level and gradient templates can be seen in figure 51.

Local search

The refinement search is repeated multiple times for different levels of image smoothing, moving through 16 logarithmic smoothing levels. This starts with no smoothing, then smoothing with a Gaussian kernel 2 pixels wide, followed by width increases of 20% each iteration. The logarithmic scale allows a wide range of smoothing levels to be tested emulating the degradation of small to large craters.

For each smoothing level and each candidate, a quantised x, y and radius parameter space is constructed covering values +/- 3 standard deviations (60% of radius) around each candidate's initial parameter values. The parameter space is similar to the Hough transform parameter space of section 8.3.3, but without any parameter transformations (i.e. using r, not r'). For each parameter bin an image region corresponding to those parameters is extracted, scaled and normalised to give grey level and gradient images compatible with the mean templates. Template match scores are computed and recorded.

After every smoothing level has been tried each candidate's parameter values are updated to those corresponding to the overall best x, y, radius and smoothing where the match scores are optimal. If there is no clear peak in match scores then the candidate's original parameter values are retained. The sub-bin location of peaks in the quantised search space can be estimated using 3D quadratic interpolation, as used previously to find Hough transform peaks.

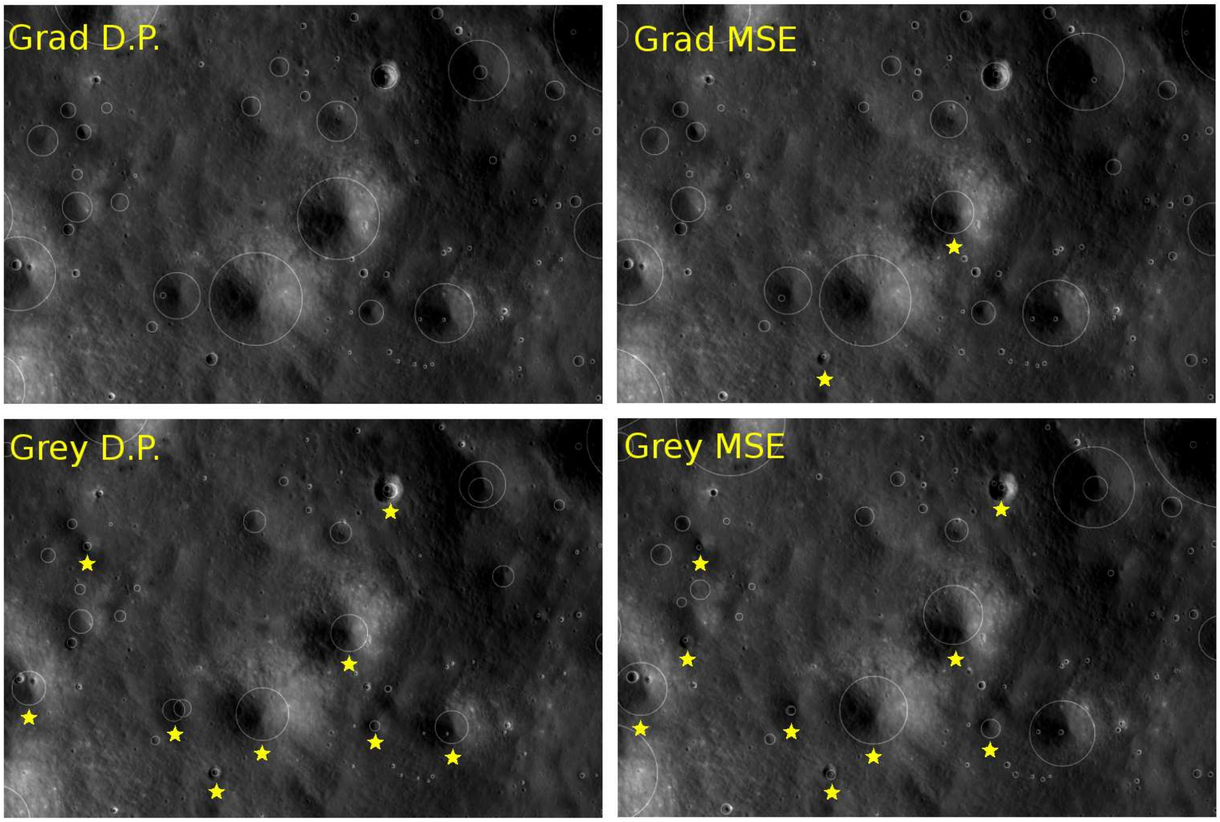


Figure 52: Sample of refined craters using gradient template with dot-product (Grad D.P.), Gradient template with MSE (Grad MSE), Grey template with dot-product (Grey D.P.) and Grey template with MSE (Grey MSE). Yellow stars in Grad MSE, Grey D.P and Grey MSE images indicate clear mistakes in comparison to the Grad DP bench-mark.

8.4.5 Testing

The refinement algorithm was tested on the clustered Moon Zoo data outputted from step 1 (section 8.3), with match scores and types of template tried in all combinations. Examples of refined locations and scales for each combination can be seen in figure 52.

A lack of definitive ground truth, with precise x, y and radius parameters, makes objectively assessing the success of the algorithm problematic. Whilst an expert may be able to state, with confidence, that a candidate represents a true crater in the vicinity of the candidate's parameters, there will always be uncertainty as to that crater's true centre and size. Given these difficulties, during testing the refinement results can be visually inspected. The match score and template combination yielding the fewest 'obvious' errors can then be selected as the best. Obvious errors include offsets and sizes outside the 20% mark-up error, however, this manual checking is a clear source of subjectivity.

8.4.6 Discussion

An inspection of refined candidates suggests that using a gradient image template matched using the pixel normalised dot product gives the best overall results. The samples seen in figure 52 illustrate this within a region of the Moon Zoo data. This result is consistent with the arguments given in sections 8.4.2 and 8.4.3 regarding the use of invariant information. The gradient images discard low frequency grey level changes across craters, focusing more on the local structure. The DP match score also discards low frequency information. It would appear that it is the higher frequency spatial details which provide most information about a crater's parameters. This is also consistent with reports from stereo matching algorithms using similar derivative images and dot product comparisons [98][99].

Returning to the questions of section 8.1.1, in particular the second: Can the parameters of candidate craters be corroborated against image data? The answer would appear to be yes. Although, due to subjectivity and a lack of exact parameter ground truth, the refinement algorithm might be better considered as a definitional process for providing output for the next pipeline stage. The refined match score distributions will be used in the next chapter where they will be used to train linear histogram models.

8.5 Summary

This chapter began by stating the goals and challenges involved in the processing of Moon Zoo crater data, with the ultimate aim of generating crater size-frequency distributions. The properties of Moon Zoo data was described, including multiple mark-ups per true crater, contamination from false positives, missing craters and systematic biases in certain crater sizes.

A four stage processing pipeline was proposed to convert the raw Moon Zoo data into SFDs, with the first two steps being designed, implemented and tested for this chapter. These steps coalesced multiple mark-ups into individual crater candidates and refined the parameters of those candidates to better match evidence of craters within the associated lunar images. A Likelihood justification was provided for each step, where coalescence was achieved using a Hough transform-like clustering algorithm and refinement was achieved using a combination of templates and match scores. The performance of the coalescence algorithm was assessed using Monte-Carlo, with conclusions drawn that the sparseness of craters in the Moon Zoo data would lead to few mistakes during clustering. The performance of the refinement algorithm was assessed subjectively, but produced observable differences in size and location distributions, with conclusions drawn that matching is best performed with gradient templates and dot product style comparisons.

The input into the first step of the processing pipeline contained over 40,000 candidates covering two NAC images. The output of this chapter, which will feed into steps 3 and 4 of the pipeline, contains approximately 20,000 clustered and refined craters. This reduced dataset will still contain false positives, true positives and false negatives. Some of the false positives and false negatives would have persisted in the data from the outset, but additional false positives and negatives could have been introduced due to inefficiencies in the first steps of the processing pipeline.

The next chapter will complete the Moon Zoo processing pipeline by implementing and testing steps 3 and 4: Linear modelling to estimate true verses false positive craters; and false negative estimation to correct for missing craters. The final three questions posed in section 8.1.1 will then be answered: Can the quantities of false positive contamination and true positive craters be estimated using linear histogram models? Can the above quantities be measured consistently within predicted errors under different conditions? And can the quantities of false negative missing craters be accounted for?

9 Lunar Crater Counting: Moon Zoo Part 2

The previous chapter described the goals of the Moon Zoo project and explained the challenges involved in interpreting the citizen science crater data it produced. To briefly review, raw Moon Zoo data contains multiple mark-ups detailing the potential locations and sizes of thousands of lunar impact craters from selected regions of the Moon. Amongst those mark-ups are genuine craters (true positives) which have been correctly identified by Moon Zoo users and bogus craters (false positives) which have been incorrectly identified. There are also missing craters (false negatives) which exist within the lunar images but have not been marked-up by any user and therefore do not appear within the crater data. Further to the raw location and size information, evidence of potential craters can be supported by analysing the associated image regions by searching for crater-like structures.

A four stage filtering pipeline was proposed in the previous chapter to reduce the Moon Zoo data to create size-frequency distributions. This pipeline contains the steps:

1. Coalescence: the clustering of related mark-ups into individual craters via a Hough transform-like algorithm;

2. Refinement: the refinement of candidate crater parameters via a local brute-force image search using crater templates;

3. Linear Modelling: the application of linear histogram models to estimate the quantity of true and false positive craters, via analysing the distribution of template match scores;

4. False Negative Calibration: the correction of underestimated SFDs caused by missing craters via calibration against preprepared ground truth SFDs.

The first two of these steps were implemented and tested in the previous chapter. This chapter will complete the processing pipeline by implementing and testing the final two steps, and answering the final questions of section 8.1.1:

- Can the quantities of false positive contamination and true positive craters be estimated using linear histogram models?

- Can the above quantities be measured consistently within predicted errors under different conditions?

- Can the quantities of false negative missing craters be accounted for?

9.1 Moon Zoo reduced crater data

The source of data used within this chapter is the output of pipeline step 2 described in the previous chapter (section 8.4). This data includes the following information per candidate crater:

- the refined coordinates of each crater centre within the pixel coordinate system of the source image;
- the refined radius of each crater in pixels;
- the peak hight of the Hough transform cluster associated with each crater;
- the smoothing kernel width required to best match the crater image with a mean crater template;
- the grey image template mean squared error match score (MSE of equ (130));
- the gradient image template mean squared error match score (MSE of equ (130));
- the grey image template pixel-normalised dot product (DP of equ (131));
- and the gradient image pixel-normalised dot product (DP of equ (131)).

It is the match score values which will be of most use in this chapter.

9.1.1 Ground truth

The pipeline steps investigated in this chapter are designed to quantify two complementary effects, that of contamination from false positives and of inefficiencies due to false negatives. In addition to the above data different ground truths are required for investigating these different effects.

To perform experiments and to quantify contamination from false positives a ground truth is required which labels all candidates in the reduced dataset as being either true or false craters. It is necessary to divide the candidates into these classes so that appropriate training data can be provided. It is also necessary to know the true quantities so that estimates from linear modelling can be checked against known values. To estimate the quantity of missing craters in size-frequency distributions an alternative ground truth is required in the form of known size-frequency distributions. These SFDs will assume that there are no false positives, simplifying the analysis.

The ground truths will necessarily be somewhat subjective, but efforts should be made to match expert definitions as closely as possible. The ground truths used within this chapter were created by visually inspecting craters in the dataset. This was undertaken by non-expert crater counters with some experience of inspecting lunar images (the thesis

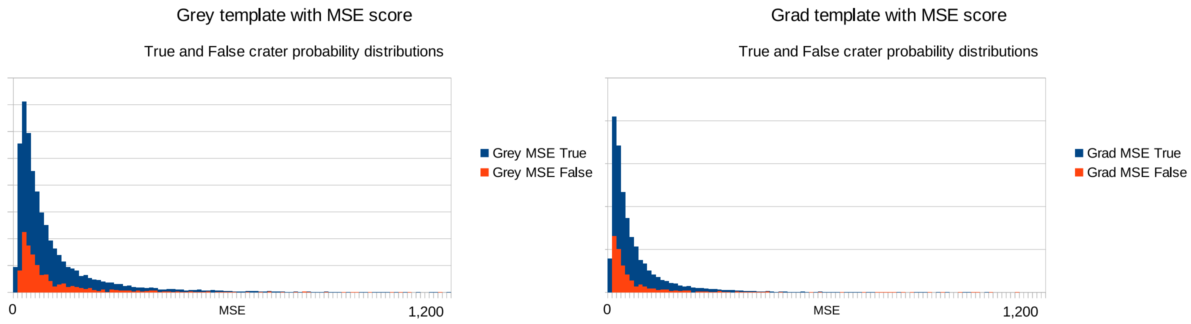


Figure 53: Left: Mean Squared Error match score distribution computed using grey level image template. Right: MSE match score distribution computed using gradient image template.

author and group of School of Earth, Atmospheric and Environmental Sciences undergraduates), followed by a subsample being verified by a lunar expert.

The reduced dataset ground truth of 20,565 potential craters includes 6,291 assigned as false positives and 14,274 as true positives, giving a Moon Zoo user false positive error rate of approximately 31%. The SFD ground truths were constructed for 8 regions of the data, with craters counted down to the minimum crater size of 10 pixels in radius.

9.2 Step 3: Linear modelling

The problem of measuring the quantity of true craters verses bogus craters can be viewed as a two class categorisation problem in which a subset of true and false positives can be used as training data then applied to independent subsets for making measurements. Various combinations of match score histograms can be used in the process.

Unlike the martian terrain histograms populated from densely sampled BRIEF (chapter 6) or Poisson blob (chapter 7) descriptors, match score values are continuous and sparsely sampled from non-overlapping spatial regions. This potentially avoids problems with correlated residuals which have previously lead to underestimated error predictions. This also provides the opportunity to investigate previously untested aspects of theory by constructing histograms with different axises and different binning resolutions, which should produce statistically equivalent measurements differing only by the size of the error bars.

The following subsections will investigate the use of linear histogram models in this context primarily to solve the Moon Zoo false positive problem, but will also test consistency over varying sampling conditions.

9.2.1 Match score distributions and histogram selection

A combination of 1 dimensional and 2 dimensional histograms are used in this chapter. The 1 dimensional histograms are constructed for each individual match score. The 2 di-

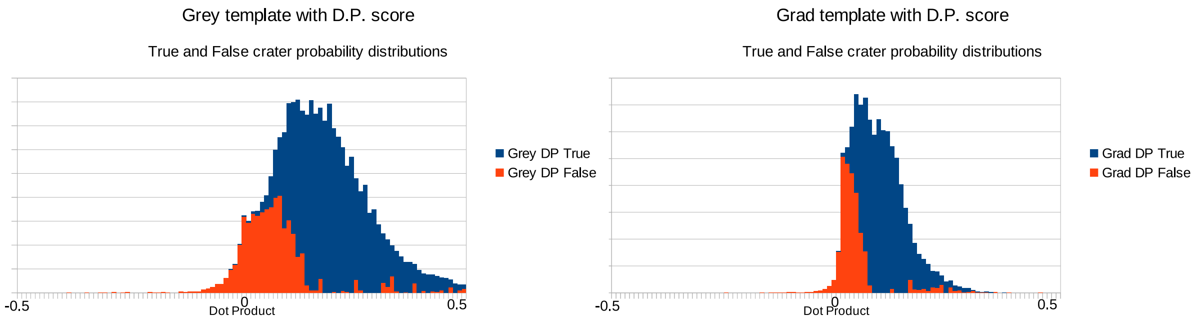


Figure 54: Left: Dot Product match score distribution computed using grey level image template. Right: DP match score distribution computed using gradient image template.

mensional histograms combine 2 complementary match scores. Each match score contains different information about the underlying image, as described in the previous chapter in section 8.4.3. The histograms tested are then:

1D histograms:

- Grey templates with MSE match scores (Grey MSE);
- Gradient templates with MSE match scores (Grad MSE);
- Grey templates with DP match scores (Grey DP);
- and Gradient templates DP match scores (Grad DP)

2D histograms:

- Grey MSE vs Grey DP;
- Grey MSE vs Grad DP;
- Grad MSE vs Grey DP;
- Grad MSE vs Grad DP;
- Grey MSE vs Grad MSE;
- Grey DP vs Grad DP.

The individual match score distributions can be seen in figure 53 and figure 54. The MSE style match scores have an intuitive shape, with true craters accumulating near MSE values of zero (zero being a perfect template match) and false craters forming an overlapping distribution away from zero. The DP style match scores also take an intuitive form, with a peak at low values straddling zero and going negative for false positives and higher values for true positives.

According to the error theories developed in chapters 4 and 5, assuming the data can be sufficiently approximated using a linear combination of independent Poisson bins (following properties of section 3.2), any valid histogram description of the data should yield statistically valid measurements. Estimated quantities of false and true positives should be consistent, within errors, with ground truth, and predicted measurement accuracies should match observed accuracies over repeated trials. The measurement from different histograms should differ only by the size of the final error bars, with less ambiguous representations giving better accuracies than histograms with large overlapping classes. The best results are therefore anticipated from most separable match score distributions.

Given that match scores are continuous values there is also choice available for the binning of each histogram. This is in contrast to the BRIEF and Poisson blob histograms used previously during martian terrain analysis experiments where binning was dictated by discrete patterns with no clear adjacency relation between them. With match scores, the choice of a wider binning will result in better populated bins with lower relative per-bin Poisson errors, but with potentially more overlap between classes. Whereas a greater number of narrower bins may provide better separability at the boundaries between classes, i.e. less ambiguity at overlaps, but at the expense of fewer entries per bin. Achieving best results will therefore be a function of the selected match score(s) and their sampling.

9.2.2 Populating histograms

The histograms listed in the previous section must be described in terms of linear subcomponents before they can be fitted to new data for the task of estimating false and true positive quantities of craters. This requires multiple exemplar histograms to be sampled for both true and false positive classes in order to conduct the model building histogram Independent Component Analysis. Once extracted, the linear model components must be fitted to large numbers of independent histograms and done so repeatedly in order to examine actual error distributions to corroborate error predictions. Both of these processes requires large quantities of data. Previously, when constructing and testing linear models for martian terrain analysis, there was an infinite quantity of training and testing data available via the terrain simulator Monte Carlo. Unfortunately this is not the case with selected Moon Zoo data, where the 20,000+ craters limits the total quantity of independent samples available. To solve this the Moon Zoo data can be repeatedly reused by sampling with replacement.

To keep the sampling with replacement realistic the sampling can be performed on a regional basis, as opposed to a crater-by-crater basis. This strategy is important for maintaining possible regionally dependent variations in match score distributions. To sample N true positive craters a rectangular image region containing at least N true positives can be selected uniformly at random. The first N true positives found within that region, from left to right, top to bottom, can be entered into the histogram being populated. The same scheme can be applied for false positives. The sampling with replacement can be achieved by allowing regions to overlap an unlimited number of times. To ensure uniqueness of the samples a small amount of additional noise can be added to each sampled match score value. This additional noise, significantly less than the binning resolution, should not adversely change the shape of the distributions.

9.2.3 Testing

During each trial 10,000 craters were selected in the form of 10 rectangular regions sampled with replacement. The 10 histograms were then used to train a linear model. For each trial these models were fitted to different quantities of testing data, with approximately one quarter being false positives to match the demographics of the raw Moon Zoo data. The estimated quantities of true and false positive craters were then extracted from the model using equation (25), which map directly onto counts of craters.

After each trial the difference between known ground-truth values and estimated values were divided by the predicted error and recorded in pull distributions in line with previous experiments (e.g. section 6.4). The predicted accuracies were also recorded as percentage errors on measured quantities. 1,000 repeated measurements were taken per trial, again consistent with previous martian tests.

The building and fitting of linear models was tested under varying conditions, including the measuring of different quantities of data across each type of histogram and the testing of different histogram binnings. Both 1D and 2D histograms were tested using 0.01, 0.10, 1.00, 10.00 and 100.00 times as much testing data as training data. 1D histograms were also tested using 4 to 256 histogram bins spanning the range of the match score distributions, which for MSE scores ranged between 0 and 1,200 and for DP scores ranged between -0.2 to +0.5.

9.2.4 Discussion

The models built show that match score distributions contain relatively low amounts of regional variability. Linear models of both true and false positives can be constructed using as few as 10 regions each, with only 3 to 4 linear components per class required to describe data with sufficient goodness-of-fit. Figure 61 shows the model selection curves for selected histograms, using 10 regions and increasing numbers of components. These are in contrast to the relatively large amounts of variability found in previous martian terrain models, with model selection curves for comparison plotted in figures 21 and 31.

Each experiment confirmed that it is possible to quantify the amounts of true and false positive craters within predictable accuracies using the linear modelling techniques. This can be seen in figures 55 and 59 where the ratio of predicted to observed errors approximately equals unity consistently across different ratios of training to testing data. Further more, there are no apparent problems with under estimated errors suggesting that the

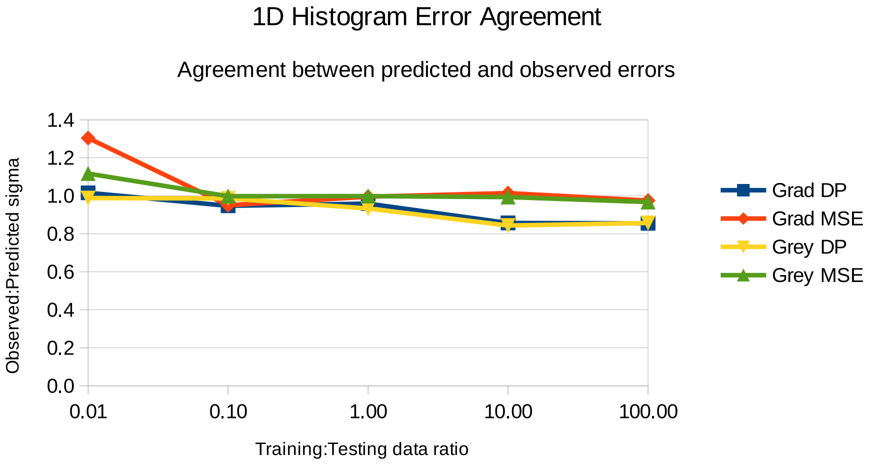


Figure 55: Corroboration that predicted measurement errors are seen in practice when linear models are constructed and fitted using 1D match score histograms. The x-axis indicates the relative quantities of training and testing data. The y-axis shows observed errors over 1,000 trials per point divided by the predicted errors.

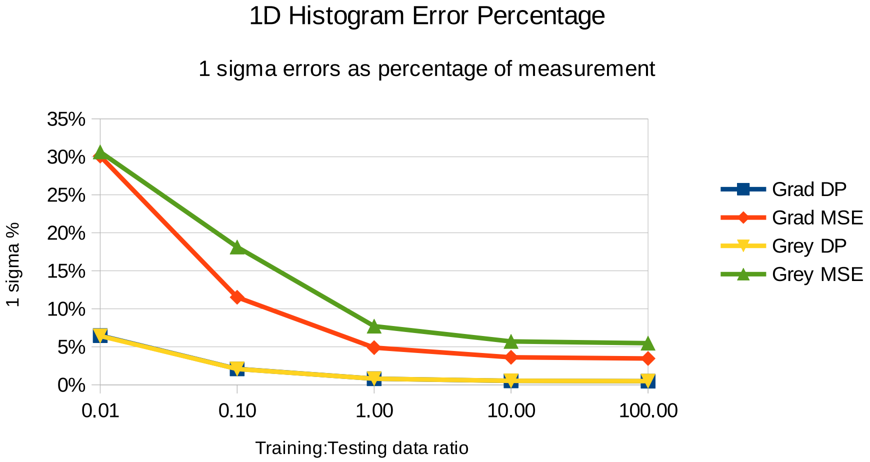


Figure 56: Measurement errors as percentage of measured quantities when using 1D match score histograms. The x-axis indicates the relative quantities of training and testing data. The y-axis shows one standard deviation of predicted accuracies as a percentage of the measurement.

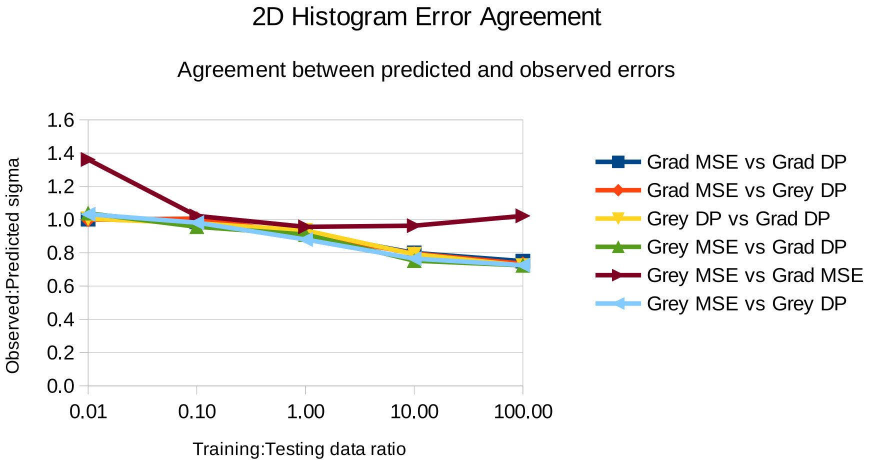


Figure 57: Corroboration that predicted measurement errors are seen in practice when linear models are constructed and fitted using 2D match score histograms. The x-axis indicates the relative quantities of training and testing data. The y-axis shows observed errors over 1,000 trials per point divided by the predicted errors.

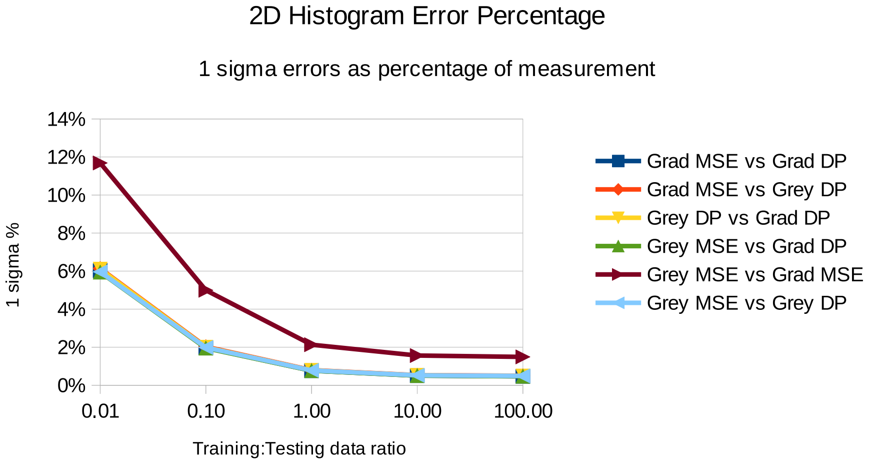


Figure 58: Measurement errors as percentage of measured quantities when using 2D match score histograms. The x-axis indicates the relative quantities of training and testing data. The y-axis shows one standard deviation of predicted accuracies as a percentage of the measurement.

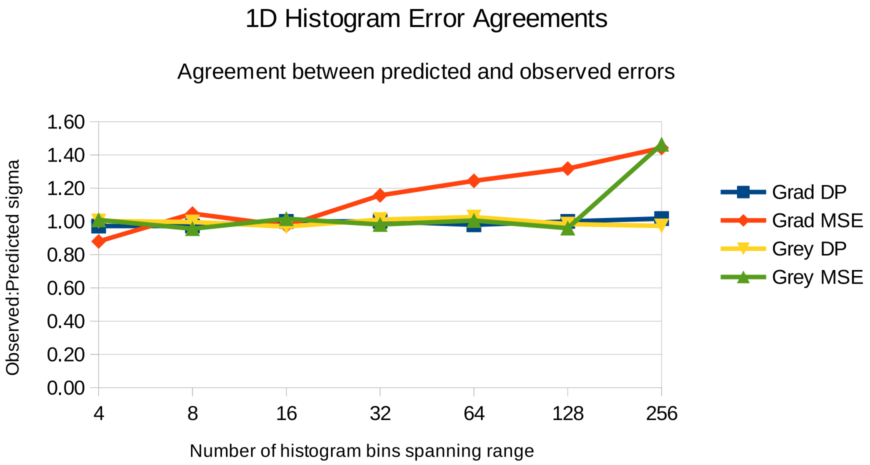


Figure 59: Agreement between observed and predicted errors when using different binnings with 1D match score histograms. The x-axis indicates the number of bins spanning the range of match score values (0 to 1,200 for MSE and -0.2 to +0.5 for DP scores). The y-axis shows observed errors over 1,000 trials per point divided by the predicted errors.

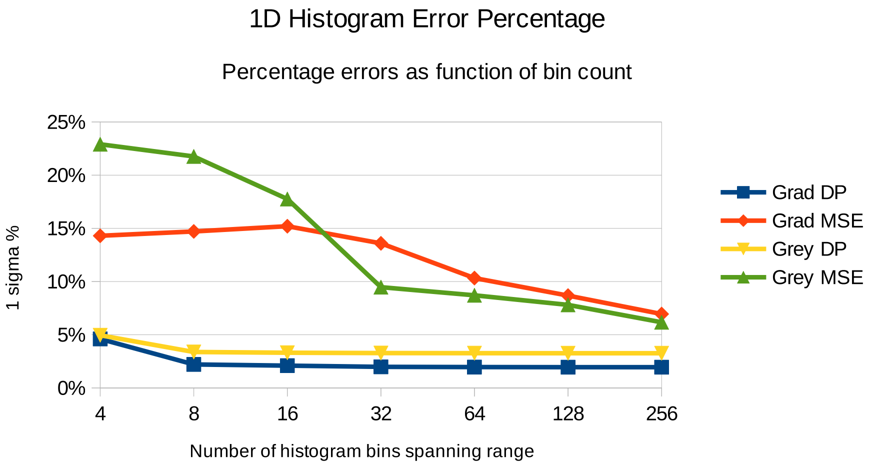


Figure 60: Measurement errors as percentage of measured quantities when using 1D match score histograms with different binnings. The x-axis indicates the number of bins spanning the range of match score values (0 to 1,200 for MSE and -0.2 to +0.5 for DP scores). The y-axis shows one standard deviation of predicted accuracies as a percentage of the measurement.

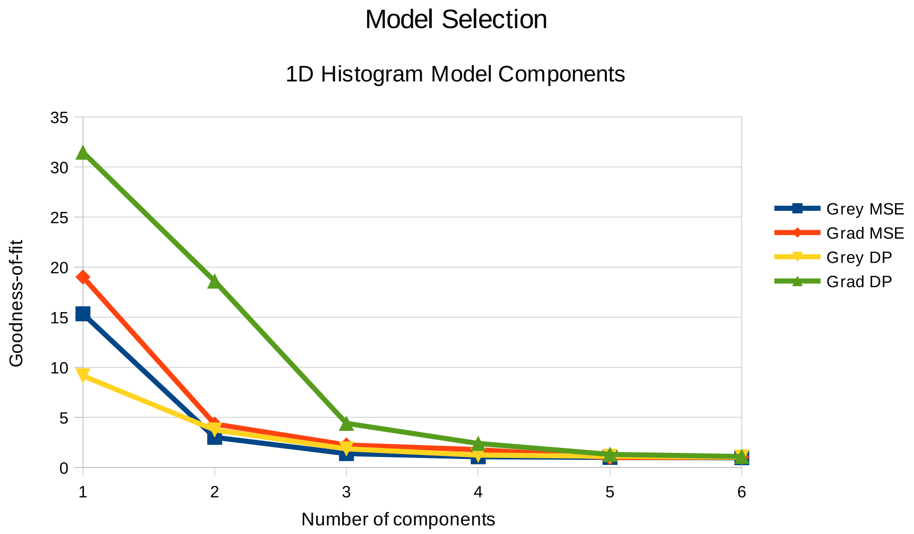


Figure 61: Model selection curves for 1D histograms. Each curve shows an alternative match score. The number of extracted components increases along the x-axis. The y-axis gives the goodness-of-fit, which should reach unity.

spatially correlated residual problems previously observed during martian terrain analyses do not occur in this new context. The behaviour of the predicted errors can be seen in figures 56 and 58 to be consistent with theory and previous experiments, with percentage errors reducing as the quantity of data increases. The poorest accuracies were seen at relatively low quantities on the 1 dimensional MSE type histograms, with percentage errors around 30%. The best accuracies were seen at relatively high quantities on the 1 dimensional DP type histograms and 2 dimensional histograms, which incorporated the DP type information, with best percentage errors around 0.5%. This is consistent with the relative levels of ambiguity between MSE and DP distributions, where visually it is clear that there is less overlap between the true and false positive distribution in the DP match scores, as seen previously in figures 53 and 54. Consistent results were also observed for different binning resolutions, with percentage errors varying across the plot in figure 60, generally improving as the number of bins increases until bins become too narrow and underpopulated.

Importantly, irrespective of the match scores sampled or the binning selected, statistically valid measurements were achieved under all conditions. It was emphasised in introductory chapters (e.g. section 1.6) that a quantitatively successful method need not be the most accurate in absolute terms. In the context of filtering Moon Zoo citizen science data the developed linear modeling and quantity estimation methods must be considered successful and trustworthy for scientific use, assuming the other issues with Moon Zoo data (false negatives, subjectivity of ground truth etc.) are appropriately addressed. With predictable accuracies it becomes possible to automate the process of comparing crater counts from different geological units with confidence. This leads to practical considerations of usability in terms of spotting significant crater count differences. Whilst any histogram and binning will give valid measurements, a researcher comparing crater counts will only be able to measure significant differences if the quantities of craters differ by at least a couple of standard deviations of the measurement accuracy. For science, this practical criteria should supersede conventional pattern recognition thinking that best ROC results represent the best algorithms. Rather, it should be acknowledged that quantitative error predictions are essential, and the absolute levels of accuracy required for an application only need to be commensurate with the differences expected to be measured in the target data. One consequence of this for crater counting is that counts are fundamentally limited to best-case Poisson errors, therefore there will be a minimum region size associated with any analysis which hopes to spot differences in crater counts. Very small regions containing very few craters will necessarily have large relative errors, and no algorithm can improve upon this.

Whilst achieving better than Poisson errors on quantity estimates is impossible according to theory, the quantification of Moon Zoo data could potentially be improved towards this accuracy limit by improving the crater templates used to gather match score information. The grey level and gradient image templates developed in the previous chapter had few degrees of freedom to adapt to target craters. An improved model would maintain the x, y and radius parameters, but could dispense with the smoothing parameter and instead extract modes of crater variation using either an eigenvector model, as has been used in various proposed crater detection algorithms [64][65], or perhaps with an ICA model, possibly reusing some linear modeling code already developed for histograms[10]. Further, to better differentiate between craters and common false positive features additional templates could be constructed for ridges, hillocks and other ambiguous entities. The match scores from these additional templates can be treated as new histogram dimensions in the hope of providing wider separation between classes.

9.3 Step 4: False negative calibration

The following subsections will attempt to answer the question ‘Can the quantities of false negative missing craters be accounted for?’

Previous pipeline stages had the character of reducing the number of craters in the dataset. The clustering reduced multiple mark-ups into individual craters and the linear modelling reduced the dataset to (probabilistic) true verses false craters. If the original raw data contained every single real crater, plus contamination, then by step 3 the task would be complete. However, there are missing craters in the raw data which will cause underestimated SFD bins. It is hoped that the efficiency of Moon Zoo users is relatively

[10] Whilst pixel grey levels cannot be treated as Poisson distributed quantities, appropriate modifications to the linear modelling theory could potentially lead to a nested linear model for the hierarchical interpretation of complex data.

consistent across the data. If this is so, a single correction factor can be determined and applied to boost SFD bin counts to approximately correct levels.

9.3.1 Testing

8 random regions of the data were selected for calibration. These regions spanned the full width of the source images in 400 pixel high strips. The regions selected were:

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 of the School of Earth, Atmospheric and Environmental Sciences, University of Manchester, marked-up each image twice, annotating all craters down to 10 pixels. The annotations were then checked by a lunar expert[11]. The double mark-ups were clustered using the coalescence method of step 1 (section 8.3), then used to construct SFDs. Poisson errors were assumed on final bin counts. The same regions were analysed using refined data from step 2 (section 8.4), followed by linear modelling of step 3 (section 9.2), before also being converted to SFDs. Predicted errors from step 3 were assumed on final bin counts. Figure 62 shows the comparative results.

9.3.2 Discussion

Unfortunately, for smaller craters of less than 20 pixels, there were large discrepancy between ground-truth and filtered SFDs. These discrepancies were also regionally dependent, preventing the use of a single correction factor. However, above 20 pixels there was good agreement between filtered Moon Zoo SFDs and ground-truth, albeit at low sample sizes with large relative errors.

There is weak evidence of a link between efficiencies and total numbers of craters, with efficiency going down as the total number of craters goes up, as seen in figure 63. If this relationship is genuine then there is an intuitive interpretation: citizen scientists can only

[11] The names of these undergraduates and lunar expert are noted within the acknowledgements.

work at a finite speed, and within a given time-span, a heavily cratered region will be less well counted than a sparsely populated one.

With further investigation it might be possible to determine a relative calibration based upon user behaviour. However, there is a very simple solution to the false negative problem: images can be left on-line for much longer to ensure few craters are missed.

9.4 Summary

Within this chapter it has been shown that linear histogram models can be used to quantify the amount of contamination in Moon Zoo crater data from false positives. This was achieved with real-world data to within predictable accuracies. This demonstration shows that the methods developed within this thesis have practical utility for at least some real-world problems.

The Moon Zoo processing pipeline developed over the previous two chapters still only constitutes a semi-automated crater counting system. There is an obvious extension to the above method involving a global image template search for all crater-like features, which is usually the starting point of other proposed automated crater detection algorithms. However, such a system could suffer from false negatives, where genuine yet atypical craters fail to secure decent template match optima. Scaling up to this type of system is the next logical step and should be the focus of future work.

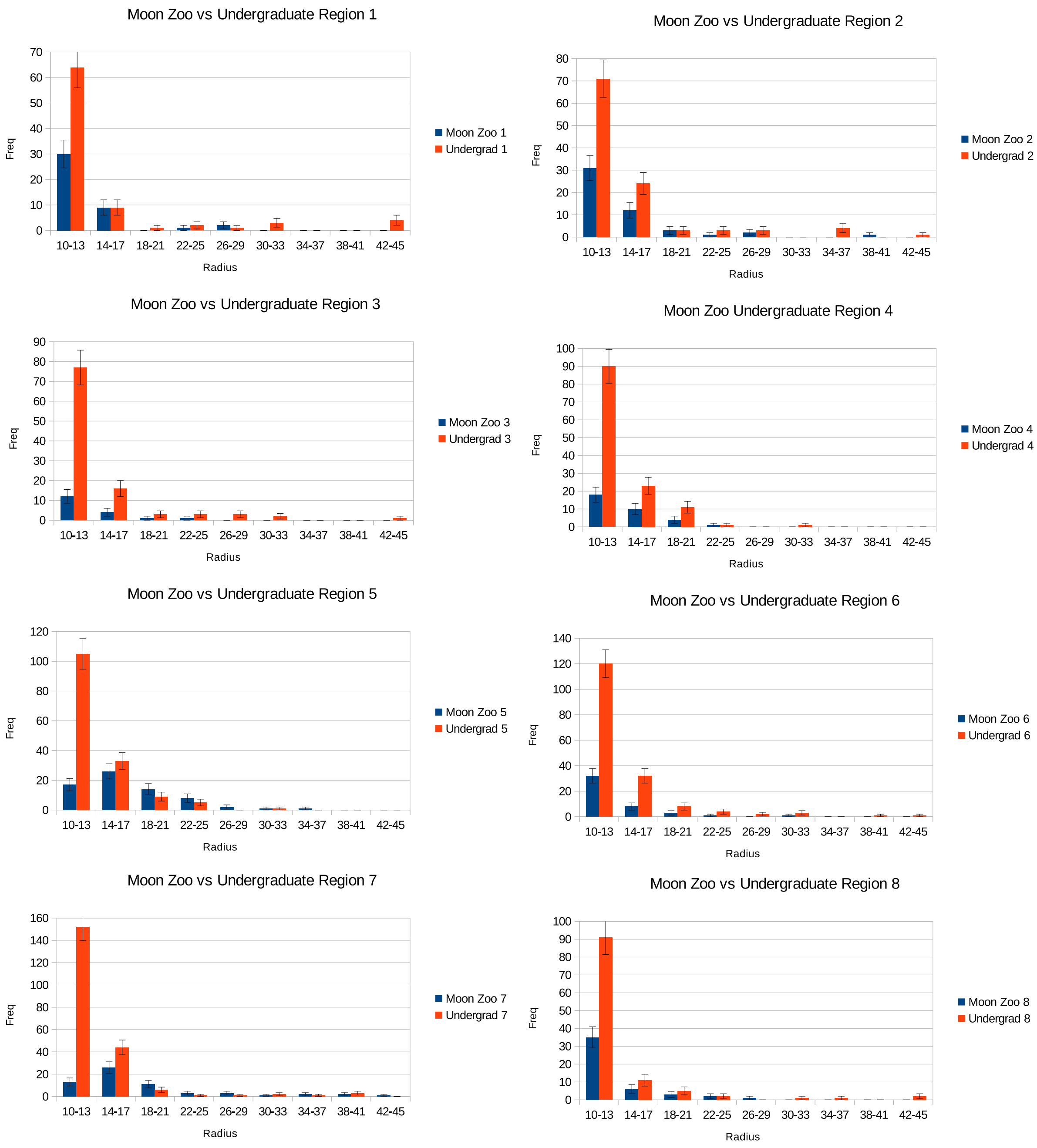


Figure 62: Comparative SFDs for 8 randomly selected regions of data.

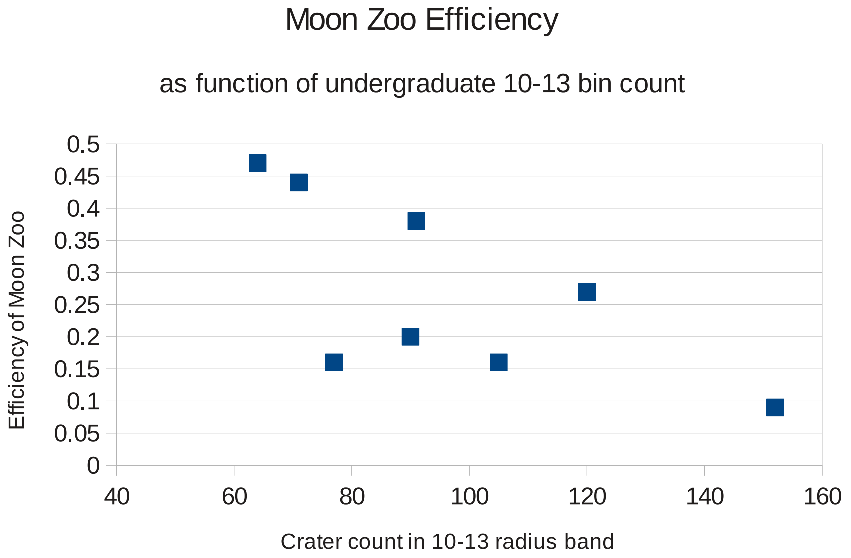


Figure 63: Relationship between crater counting efficiency within the smallest size bin and the total number of ground-truth craters in those size bins.

10 Conclusions

This thesis began by arguing in chapter 1 for the need of an automated solution to the analysis of planetary images. This need was motivated by an unmanageable accumulation of data, a limited supply of experts and the inherent subjectivity of humans. From the outset the needs of researchers wishing to make objective measurements was emphasised and a set of quantitative criteria stated against which the success of any proposed automated methods could be compared. These criteria specify that measurements are driven by evidence, accompanied by error estimates, and that measurements in practice do not deviate by more than their predicted accuracies. The criteria also encouraged the use of additional tools for corroborating that measurements can be trusted. A review of literature in chapter 2 concluded that the conceptual building blocks necessary to fulfil the criteria existed, but were rarely combined into scientifically useful systems and that empiricism was often favoured over theoretical considerations of measurement errors.

Within the first half of the thesis theoretical considerations and assumptions about data generation lead to the development of a flexible histogram-based pattern recognition system capable of learning complex and variable distributions. The basics of a histogram-based statistical model was provided in chapter 3, followed by the development of a predictive error theory in chapters 4 and 5. During these chapters a new method of Independent Component Analysis was created specifically for histogram data as part of a training algorithm and an Expectation Maximisation algorithm was created for fitting extracted components to new data for the purposes of making measurements. The Cramer Rao Bound (lower variance bound) and error propagation were applied to estimate the statistical and systematic uncertainties within measured quantities. These methods were tested using Monte-Carlo simulated histograms showing that for ideal data, which abided by the assumed data properties, the developed system fulfilled all of the quantitative criteria.

The second half of the thesis took on the challenge of applying the pattern recognition system to more realistic data. This began with synthetic martian terrains derived from HiRISE images in chapters 6 and 7, then moving on to lunar craters with assistance from the Moon Zoo team and their many citizen scientists in chapters 8 and 9. Different image encoding schemes were used as input to the system for the tasks of measuring martian terrain surface areas and counting lunar craters. Experiments revealed the inherent difficulties in making a new statistical method work in practice, as various experiments unveiled problems with the assumed data properties. These problems were examined using a goodness-of-fit function, residual correlation matrices and repeated measurements compared against ground-truth values. This resulted in a better understanding of image properties and the limitations of the method. Success was achieved in chapter 9, where an appropriate image encoding allowed the quantitative measurement of true positive verses false positive craters in Moon Zoo data.

Now in this concluding chapter the main findings will be summarised. The strengths and limitations of the developed methods will be explored and opportunities for future work identified.

10.1 Theory summary

The quantitative pattern recognition theory developed centres around a linear additive model which combines weighted probability mass functions to approximate arbitrary histogram distributions. The constituent PMFs are extracted from training data using a form of ICA, leading to a flexible model capable of being fitted to new data containing varying proportions of subcomponents. Measurements are extracted by fitting the model to new data and relating model parameters to physical quantities. An error theory accompanies this model for making first-order approximations to how perturbations in input data affects the stability of measured quantities. This error theory provides measurement error covariances appropriate for scientific use, assuming the data to which the method is applied meets certain criteria.

10.1.1 Findings

Early insights into the behaviour of measurements came from inspecting error covariance calculations and testing error predictions using Monte-Carlo techniques. The main findings were:

- statistical perturbations in measurements can be modelled upon assumptions of Poisson noise in incoming histogram data;
- there are systematic effects in measurements which can be modelled by Poisson noise in training data which becomes fixed in the model;
- statistical errors grow proportionally to the square-root of the measured quantities: $\sigma_{stat} = \alpha\sqrt{\mathbf{Q}_k}$;
- systematic errors grow proportionally to measured quantities: $\sigma_{sys} = \beta\mathbf{Q}_k$;
- and the levels of proportionality, α and β, are functions of the ambiguity between classes which grows as the level of overlap between component distributions increases.

One consequence of these findings leads to a lower bound on attainable accuracies. In the best possible case where there is no ambiguity between classes, i.e. $P(X|k)$ is either exactly 1 or 0 for all X and k, then the expected error on measured quantities is Poisson, i.e. $\sigma^2_{\mathbf{Q_k}} =< \mathbf{Q_k} >$, and the systematic term goes to zero. In the general case of ambiguity the statistical component dominates the total error at small relative quantities of data and the systematic component grows to dominate at large quantities. This change is relative to the quantity of training data where: 10 times as much training data than testing data leads to dominant statistical effects; 1 to 1 training to testing data leads to roughly half statistical/half systematic effects; and 10 times as much testing data than training data leads to systematic effects dominating. These changes in effects motivated the x-axis of various plots throughout later chapters, where different ratios of training to testing data were tested to confirm that error predictions worked in both extreme cases.

10.1.2 Strengths

The strengths of the modelling and estimation methods include:

- the non-parametric choice of histogram models allows any distribution to be approximated;
- the ability to construct histograms using arbitrary numbers of linear components allows many modes of variation to be modelled;
- because the entire data density is being modelled (unlike decision boundary methods) per bin predictions of pattern frequencies can be produced, $\mathbf{M}_X$, and these can be tested;
- and the error theories give case-by-case data-driven error predictions (unlike ROC alternatives).

The choice of histogram models was purposeful, as few assumptions could be made regarding the shape of pattern distributions in planetary terrains. Similarly, there are few assumptions which can be made regarding how distributions may vary making the ability to add additional components as required highly advantageous. This flexibility ensures that a wide range of data can be modelled. Plus, to ensure data is being modelled sufficiently well the bin-by-bin predictions of frequencies, $\mathbf{M}_X$, can be compared to data, $\mathbf{H}_X$, to construct goodness-of-fit functions (equ 45) and correlation matrices (equ 93) for corroboration. Yet the biggest strength is the ability to produce full error covariance predictions for measured quantities, which is believed to be a unique feature of this style of supervised pattern recognition system.

10.1.3 Limitations

The limitations of the modelling and estimation methods include:

- the assumed properties of data must be met;
- Monte-Carlo testing is artificial;
- the method represents a worse-case scenario requiring large numbers of parameters to be trained.

- training and testing requires large quantities of data;
- and ICA extracted model components might not span the full range of possible data variability.

Some major limitations of the theory emerge from the assumed data properties stated in section 3.2, which require histogram data to contain independent Poisson bins forming distributions describable using linear combinations of base components. Monte-Carlo testing did corroborate the method's validity, but the simulated data used was specifically constructed to ensure that necessary properties were fulfilled. It can not be assumed that the methods will work more generally on all histograms.

Other major limitations of the method are linked to the worse-case nature of histograms as a representation for distributions. Whilst it is true that any distribution can be approximated with a histogram, sufficient data is required to populate it. A valid parametric description with few parameters could be more accurate, even with far less data, than a histogram alternative where each bin requires estimation. Histograms were selected for flexibility, simplicity and understandability, but a quantitative system could benefit from parametric components if the application demanded it.

Lots of data is required to both train and test histogram models. Underpopulated bins can lead to problems as seen in sections 5.3 and 7.4.1 where approximations break down. Holes may appear in distributions when insufficient data is provided. Gaussian approximations to Poisson variances (as is assumed in error propagation) and square-root transforms to improve approximations (as is used in the goodness-of-fit function) both fail eventually at low enough statistics. Even with well-populated histograms, if an insufficient number of exemplar histograms are provided to the ICA training algorithm the full range of data variability might not be modelled. This issue was explained in section 3.8, as extracted components can only describe histogram subspaces accessible via positive coefficients. Only the variability seen in training data can be expected to be modelled correctly when applied to unseen data.

10.2 Application

Beyond the basic Monte-Carlo studies where histograms were synthesised directly, the methods were tested using histograms derived from real images. These images included simulated martian terrains constructed using martian HiRISE data and lunar crater images partially annotated by Moon Zoo users. Inspired by the BRIEF representation, local BRIEF descriptors and correlated BRIEF blobs approximating Poisson events were used to encode martian terrains. Grey level and gradient template images with a smoothing parameter were used to represent lunar craters. These different image representations were used to measure martian surface areas and to count true lunar craters in the presence of false positive contamination.

10.2.1 Findings

Applying the developed methods to realistic data provided many insights into the difficulties of appropriately encoding images into usable histograms. These difficulties include:

- real histograms are more complex than an assumed accumulation of independent Poisson events;
- if Poisson events are driving the data there is no guarantee that there is a one-to-one correspondence between patterns (i.e. individual histogram bins) and the generating events;
- the goodness-of-fit function alone cannot spot all problems with properties of incoming data;
- a correlation matrix can be used as an effective safety net for corroborating several required data properties;
- and additional layers of theory may be required for converting estimated model parameters to useful measurements.

Real histograms populated using the BRIEF (ch. 6) and Poisson Blob (ch. 7) representations proved to be more complex than assumed. The properties of section 3.2 were violated in multiple ways. Double counting and highly localised correlations between histogram bins suggested that any underlying Poisson events generating the image data caused clusters of related patterns to appear together. This resulted in poor model fits for BRIEF histograms and underestimated error predictions for both BREIF and Poisson Blob histograms. However, a combination of goodness-of-fit and correlation matrices were shown to be effective at identifying such violations making it possible to spot issues and disregard untrustworthy estimates.

Attempts to encode image data in a more Poisson-like way, i.e. Poisson Blobs, improved agreement between theory and practice but at the cost of complicating the analysis. The irregular shape of the blobs required additional steps to be taken to convert model parameters to measurements, ie. from blob counts to surface areas. This included extensions to the error theories.

It was found that there was a real planetary science application and image representation amenable to quantitative analysis. The successful estimation of true and false positive craters in Moon Zoo data demonstrated the value of the method.

10.2.2 Strengths

In practice there are several strengths of the method:

- quantitative criteria can be met for many martian terrain types;

- quantitative criteria can be met for Moon Zoo crater data;

- high levels of accuracy are attainable when making some measurements (within +/- fractions of a percent);

- equivalent measurements can be taken using different encodings, assuming the data is well-behaved;

- the method scales from small to large histograms using few or many components.

Despite discrepancies between predicted and observed measurement errors during martian terrain surface area tests some terrains could be measured to within a factor of two of these errors, fulfilling quantitative criteria. Amongst these results, accuracies as good as +/- 0.5% were achieved under some parameter settings of the Poisson blob representation. Very good agreement was achieved between predicted and observed measurement errors when quantifying true and false positive Moon Zoo craters, again achieving accuracies as good as +/- 0.5% when using dot-product style template match scores. The practical implications of this is the ability to measure small percentage differences between measurements to several standard deviations' significance, given the right encoding.

As predicted by theory, and observed in practice, for well-behaved data the same measurements can be taken using different information resulting in statistically equivalent estimates, differing only by the size of their errors. Experiments using Moon Zoo data, using different binning and match scores, demonstrated this clearly. This strength of the method emphasises the importance of measurement reliability and consistency rather than just absolute levels of accuracy as is common in the computer vision field.

10.2.3 Limitations

The most significant limitations of applying the system in realistic data are due to violations of assumptions made about the data, i.e. independent Poisson bins which can be linearly combined:

- it is difficult to find an appropriate encoding for image data which gives the necessary properties;

- the goodness-of-fit function alone cannot spot all problems with data;

- correlation matrices require large quantities of data to populate;

- uninformative image regions which contain no information are problematic for interpretation;

- and the selected image encodings don't fully address issues of occlusion and boundaries.

For the system to work correctly it was important to supply data which was well-behaved, i.e. met the properties of section 3.2. Multiple failures of the system were observed when using more realistic data, as was explained above in the Findings section. Amongst the least well-behaved areas of data was uninformative image regions, as there was no information within them to differentiate between empty space belonging to different features. It also extended over wide regions, making it difficult to interpret as individual Poisson events. Uninformative space was excluded from analyses from Poisson Blob experiments in chapter 7.

Given the difficulties in formatting data appropriately, it was especially important to provide mechanisms for spotting difficulties so that end users could be given honest information about the analysability of their data. The Chi-squared per degree of freedom goodness-of-fit function could spot larger than expected residuals in BRIEF histograms, yet was incapable of spotting residual correlations in Poisson Blob histograms. A correlation matrix was successful at spotting problematic correlations, but required multiple model fits to generate. The quantity of data required to construct such correlations matrices excludes their use for checking individual fits. The Method-of-Runs [79] might be a workable solution to spotting correlations if the adjacencies between bins is well understood, but for BRIEF and Blob representations there was no simple way in which this method could have been applied.

Finally, the image encodings largely avoided issues of occlusion. Martian terrains were simulated with simple boundaries and Moon Zoo craters rarely overlapped. In more complex terrains boundaries and interactions between features might not be so easily ignored.

10.3 Future work

The methods developed could be extended in several ways, including:

- accommodating correlations between bins;
- investigating improved image encodings;
- nesting the model to create a hierarchical description of data;
- and optimising image region selection via minimising measurement uncertainty.

The thrust of future work should be aimed at widening applicability of the methods to more realistic data. Part of this includes accommodating correlations between bins. This will be required if martian terrain surface area measurements are to be improved. This might be achieved simply by spotting highly correlated bins and merging them together, or a more sophisticated solution could be to extend the theory to incorporate knowledge of bin covariances. Alternatively, a data whitening stage might be considered.

Improved image encodings might mitigate against violations of data assumptions. An adaptive representation which can learn the structure of underlying Poisson events (if they exist at all) would be ideal. These goals of learning representations may be similar in scope to sparse coding, e.g. [175][176] and Deep Learning, e.g. [177][178]. Extending the theory to include nested models, where sets of estimated quantities from one layer act as input to the next, might be an integral part of such a system. For example, an adapted form of the histogram ICA learning algorithm might learn correlations between pixels to form appearance models. The parameters of such models might then be learned for classification of features. This hierarchical organisation might move closer to a quantitative Deep Learning framework.

Despite all attempts to model image data, there is always the possibility in complex images that things will appear that just can not be analysed successfully. The work presented in this thesis took a naive approach to region selection, simply choosing random rectangles or entire images as sources. A more intelligent system may subdivide an image into regions which optimise either the level of error on estimated quantities, or optimises the goodness-of-fit in each region. Such a system would aim to mask off problematic data, including boundary conditions and occlusions. A fine-grain region selection might even operate on individual descriptors (BRIEF or otherwise) to maximise the probability of patch classification based upon local evidence.

Improvements may lead to wide ranging applications, beyond image analysis. The goal of an improved system should be to facilitate, as far as possible, the analysis of truly arbitrary histogram data. Indeed, at the time of writing funds have been awarded by the Leverhulme Trust (grant RPG-2014-019) to undertake this work in which further planetary science data and complex mass spectra will be analysed. A proposal submitted to STFC is also under consideration for additional planetary science applications. Finally, the Moon Zoo team are considering adopting the developed methods for their data analysis needs.

10.4 Conclusion

Many existing computer vision systems are constructed using modular code blocks freely available to download. Modules are available for encoding images, decomposing data, classifying features and visualising results. In general such blocks are pieced together into pipelines with little statistical understanding of how such blocks interact. The end results are often only analysable empirically, leading to after-the-event evaluations of performance and 'shoot-outs' to determine which components should be considered 'state-of-the-art'. Under this ethos, success is usually measured in terms of 'best attainable accuracy' rather than measurement reliability and understandability. This leads to the adoption of new building blocks only if those modules equal or outperform rivals. From the outset this thesis took a different approach. The empiricist's plug'n'play black-box methodology which dominates the field[85] was rejected in favour of a data-driven statistical approach with 'state-of-the-science' taking centre stage. Here, the goal was always to achieve statistically valid, meaningful and honest measurements that could be used with confidence. In this regard, this work is more akin to research in the physical sciences than in the computer science tradition. The original goals have been largely achieved, but much work will be required to widen applicability to other real science problems.

This work (along with a small but growing list of others e.g. [80][153][179][180]) represents a potentially new paradigm, or at least a new sub-topic, for computer vision research: Quantitative Vision. In Quantitative Vision algorithms are designed specifically to cater for the properties of the target data. This necessitates a detailed understanding of distributions, correlations and perturbations, i.e. statistics. In Quantitative Vision the performance of an algorithm must be understandable, allowing predictions to be made which can, and must, be tested before the algorithm can be applied with confidence in scientific applications. As a challenge to the computer vision and pattern recognition communities it would be of great value if researchers re-evaluated existing algorithms under the quantitative criteria of section 1.6. This will require black-box methods to be opened up to statistical scrutiny with questions being asked such as:

- Can we predict the performance of our algorithm ahead of time?
- Can we understand the origins and nature of any statistical or systematic errors in our results?
- How far must our algorithm's performance deviate from predictions before we conclude that the algorithm or our understanding is flawed?
- What does our new algorithm assume about the data?
- How can we test that incoming data meets the assumptions made by our algorithm?
- How do we know that our model is truly describing our data?
- Can we relate our algorithm to probability theory? And if so, can we show that our results are in principle the most probable given the evidence?

In the case of the quantitative pattern recognition system developed in this thesis there are positive answers to all of the above questions.

Planetary scientists wishing to adopt automated analysis methods should too be asking these questions, and they must also be prepared to learn the fundamentals of how and why such methods work. The limited range of applications approached in this thesis revealed the difficulties in applying quantitative methods correctly. Planetary scientists wishing to use such techniques more generally will be required to make extensions to theory and algorithms as necessary, which will require the development of statistical and computer programming skills.

References

[1] A. Foerstner, James Van Allen: The First Eight Billion Miles, University of Iowa Press, p76, 2007

[2] R.C. Hall, Lunar Impact: A History of Project Ranger, NASA History Series, Scientific and Technical Information Office, NASA, 1977

[3] R. Choate et al., Lunar surface mechanical properties, Surveyor Program Results, NASA SP-184, Wash., DC, 1969

[4] B.K. Byers, Destination Moon: A History of the Lunar Orbiter Program, NASA History Series, Scientific and Technical Information Office, NASA, 1977

[5] W. Shelton, Soviet space exploration - the first decade, Arthur Barker Ltd., Unnumbered, London, England, 1969.

[6] S.A. Collins, Mariner 6 and 7 pictures of Mars, NASA, SP-263, Wash., D.C., 1971.

[7] Soffen, G. A., and C. W. Snyder, First Viking mission to Mars, Science, 193, 759-766, Aug. 1976

[8] B. Evans, D.M. Harland, NASA's Voyager Missions: Exploring the Outer Solar System and Beyone, Springer, ISBN 1-85233-745-1, 2003

[9] M. Meltzer, Mission to Jupiter: A History of the Galileo Project, NASA History Series, Scientific and Technical Information Office, NASA, 2007

[10] A. Wilson, Mars Express: A European Mission to the Red Planet, ESA Publications Division, ISBN 92-9092-556-6, 2004

[11] A.S. McEwen et al., Mars Reconnaissance Orbiter's High Resolution Imaging Science Experiment (HiRISE), Journal of Geophysical Research, vol 112, E5, 2007

[12] Hawkins SE et al. (2007), The Mercury dual imaging system on the MESSENGER spacecraft, Space Science Reviews, vol 131, 1-4, 247-338

[13] Russell C et al. (2007), Dawn Mission to Vesta and Ceres - Symbiosis Between Terrestrial Observations and Robotic Exploration, Earth Moon and Planets, vol 101, 65-91

[14] M.B. Houghton et al., Misson Design and Operations Considerations for NASA's Lunar Reconnaissance Orbiter, IAC 07, 2007

[15] J.S. Hughes, S.K. McMahon, The Planetary Data System: A Case Study in the Development and Management of Meta-Data for a Scientific Digital Library, Lecture Notes in Computer Science, Second European Conference on Research and Advanced Technology for Digital Libraries, 335-35, 1998

[16] K.H. Glassmeier, et al., The Rosetta Mission: Flying Towards the Origin of the Solar System, Space Science Reviews, 128: 1-21, 2007

[17] J. Spencer et al., Finding KBO flyby targets for New Horizons, Earth Moon Planets, 92, 483-491, 2003

[18] S.J. Bolton, The Juno Mission, International Workshop on Instrumentation for Planetary Missions, 2012

[19] J. Benkhoff, The BepiColombo Mission to Explore Mercury - Overview and Mission Status, 44th Lunar and Planetary Science Conference, 2013

[20] J.N. Goswami, M. Annadurai, Chandayaan-2 Mission, 42nd Lunar and Planetary Science Conference, 2011

[21] V. Shevchenko, Russian project Luna-Glob: goals and status, Geophysical Research Abstracts, Vol. 10, 2008

[22] D. Stoffler et al., Cratering History and Lunar Chronology, Reviews in Mineralogy & Geochemistry, Vol. 60, pp 519-596, 2006

[23] D.E. Wilhelms, The Geologic History of the Moon, U.S. Geological Survey professional Paper 1348, ISBN 978-1495919855, 2014

[24] H. Otake, H. Mizutani, Subsurface Chemistry of the Imrium Basin Inferred from Clementine UVVIS Spectroscopy, Earth Planets Space, 58, 1499-1510, 2006

[25] D.A. Rogers, Crustal compositions exposed by impact craters in the Tyrrhena Terra region of Mars: Considerations for Noachian environments, Earth and Planetary Science Letters, 301, 353-364, 2011

[26] G.D. Bart, H.J. Melosh, Distribution of boulders ejected from lunar craters, Icarus, Vol 209, 2, 337-357, 2010

[27] Kickapoo Lunar Research Team, G.Y. Kramer, Stratified Ejecta Boulders as Indicators of Layered Plutons on the Lunar Nearside, Proceedings 44th Lunar and Planetary Science Conference, 2013

[28] B.B. Wilcox et al. Constraints on the depth and variability of lunar regolith, Meteoritics & Planetary Science, Volume 40, 5, 695-710, 2010

[29] G. Neukum, et al. Cratering Records in the Inner Solar System in Relation to the Lunar Reference System. Space Science Reviews, 96:5586, 2001

[30] J.B. Plescia, M.S. Robinson, New Constraints on the Absolute Lunar Crater Chronology, 42nd Lunar and Planetary Science Conference, 2011

[31] H. Hiesinger et al. Lunar mar basalt flow units: Thickness determined from crater size-frequency distributions, Geophysical Research Letters, Vol. 29, No. 8, 2002

[32] R. Bugiolacchi, J. E. Guest, Compositional and temporal investigation of exposed lunar basalts in the Mare Imbrium region. Icarus, Vol. 197, Issue 1. pp. 1, 2008

[33] R. Bugiolacchi, P. D. Spudis, J. E. Guest., Stratigraphy and composition of lava flows in Mare Nubium and Mare Cognitum. Meteoritics & Planetary Science 41, Nr2, pp. 285-304(20), 2006

[34] K. Zahnle et al., Cratering rates in the outer Solar System, ICARUS, 163, 263-289, 2003

[35] N. Barlow, Mars: An Introduction to its Interior, Surface and Atmosphere, Cambridge Planetary Science, Cambridge University Press, ISBN 978-0-521-85226-5, 2008

[36] N.L. Lanza et al., Evidence for debris flow gully formation initiated by shallow subsurface water on Mars, Icarus 205, 103-112, 2010

[37] A. Kereszturi et al., Comparison of possible recent water or brine related flow features on Mars, 43rd Lunar and Planetary Science Conference, 1787, 2012

[38] R.P Irwin III et al. Topographic influences on development of Martian valley networks, Journal of Geophysical Research, 116, E02005, 2011

[39] R.J. Phillip et al., Ancient Geodynamics and Global-Scale Hydrology on Mars, Science, 291, 2587-2591, 2001

[40] L. Liu, M. Gurnis, Dynamic subsidence and uplift of the Colorado Plateau, Geology, 38, 663-666, 2010

[41] R.J. Isherwood et al., The volcanic histroty of Olympus Mons from paleo-topography and flexural modelling, Earth and Planetary science Letters, 363, 88-96, 2013

[42] S. Piqueux et al., Sublimation of Mars's southern seasonal CO2 ice cap and the formation of spiders, Journal of Geophysical Research, Vol. 108, No. E8, 5084, 2003

[43] K.S. Edgett, P.R. Christensen, The Particle Size of Martian Aeolian Dunes, Journal of Geophysical Research, Vol. 96, No. E5, 22765-22776, 1991

[44] R.A. Bagnold, The Physics of Blown Sand and Desert Dunes, Springer, ISBN 978-94-009-5684-1, 1974

[45] P.H. Stone, J.S. Risbey, On the limitations of general circulation climate models, Geophysical Research Letters, Vol. 17, Issue 12, 2173-2176, 1990

[46] E. Gardin et al., Dune fields on Mars: Markers of climatic changes?, Planetary Dunes Workshop: A Record of Climate Change, 7022, 2008

[47] S.J. Kadish et al., A Global Catalog of Large Lunar Craters (¿20KM) from the Lunar Orbiter Laser Altimeter, 42nd Lunar and Planetary Science Conference, 2011

[48] C.I. Fassett et al., The global population of large craters on Mercury and comparison with the Moon, Geophysical Research Letters, Vol. 38, L10202, 2011

[49] R.K. Hayward, K. F. Mullins, L. K. Fenton, T. M. Hare, T. N. Titus, M. C. Bourke, A. Colaprete, and P. R. Christensen, Mars Global Digital Dune Database and initial science results, J. Geophys. Res., 112, E11007, 2007

[50] A. Garcia et al., Global mapping and characterization of Titans dune fields with Cassini: correlation between RADAR and VIMS observations, 44th Lunar and Planetary Science Conference, 2013

[51] M.H. Carr, The martian drainage system and the origin of valley networks and fretted channels, J. Geophys. Res., 100, 7479-7507. 1995

[52] B.M. Hynek, M. Beach, M.R.T. Hoke, Updated global map of Martian valley networks and implications for climate and hydrologic processes, Journal of Geophysical Research, 115, E9, 2010

[53] K. Joy et al., Moon Zoo: citizen science in lunar exploration, Astronomy & Geophysics, Vol. 52, 2, 2.10-2.12, 2011

[54] S.J. Robbins et al., Cataloging the Moon with the CosmoQuest Moon Mappers Citizen Science Project, 43rd Lunar and Planetary Science Conference, 2012

[55] C.J. Hansen et al., Mars' Seasonal Fans measured by Citizen Scientists, European Planetary Science Congress, Vol. 8, EPSC2013-855-1, 2013

[56] R.M. Haralick et al., Textural Features for Image Classification, IEEE Transactions on systems, man and cybernetics, Vol. SMC-3, No. 6, pp 610-621, 1973

[57] L. Wei et al., Hyperspectral Image Classification Using Gaussian Mixture Models and Markov Random Fields, Geoscience and Remote Sensing, Vol. 11 I. 1, 2014

[58] C. Brooks, K. Iagnemma, Terrain Classification and Classifier Fusion for Planetary Exploration Rovers, Proceedings of the 2007 IEEE Aerospace Conference, 2007

[59] C. Huang, L. S. Davis, and J. R. G. Townshend. An assessment of support vector machines for land cover classification. International Journal of Remote Sensing, 23(4), 2002

[60] H. Bischof, W. Schneider, and A. J. Pinz. Multispectral classification of landsat images using neural networks. IEEE Transaction on Geoscience and Remote Sensing, 30(3), 1993

[61] G. Salamuniccar, S. Loncaric, Application of machine learning using support vector machines for crater detection from Martian digital topography data, 38th COSPAR Scientific Assembly, 2010

[62] W. Ding et al., Automatic Detection of Craters in Planetary Images: An Embedded Framework Using Feature Selection and Boosting, CIKM 2010 Proceedings of the 19th ACM international conference on Information and knowledge management Pages 749-758, 2010

[63] L. Bandeira et al., Automatic Detection of Sub-KM Craters using Shape and Texture Information, Proc. 41st Lunar and Planetary Science Conferece, 2010

[64] M.C. Burl et al, Automated Detection of Craters and Other Geological Features, Proc 6th International Symposium on Artificial Intelligence and Robotics & Automation in Space: i-SAIRAS 2001, 2001

[65] J.R. Kim, et al., Automated Crater Detection, A New Tool for Mars Cartography and Chronology, Photogrammetric Engineering & Remote Sensing, Vol 71, No. 10, pp 1205-1217, 2005

[66] J.I. Simpson et al., 3D Crater Database Production on Mars by Automated Crater Detection and Data Fusion, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B4. Beijing 2008

[67] A.W. Bauer, et al. Machine Cataloging of Lunar Craters from Digital Terrain Model, 42nd Lunar and Planetary Science Conference, 2011

[68] Y. Sawabe et al., Automated detection and classification of lunar craters using multiple approaches, COSPAR (Advance in Space Research), 2005

[69] M. et al., Autonomous Craters Detection from Planetary Image. In Proceedings of the 2008 3rd International Conference on Innovative Computing Information and Control (ICICIC), pages 443, 2008

[70] N.D. Kamarudin et al. An Overview of Crater Analyses, Tests and Various Methods of Crater Detection Algorithm, Frontiers in Environmental Engineering (FIEE) Volume 1 Issue 1, December 2012

[71] P.G. Wetzler et al., Learning to Detect Small Impact Craters, Application of Computer Vision, Seventh IEEE Workshopson WACV/MOTIONS '05 Volume 1, 2005

[72] L. Bandeira et al., Detection of dune fields on Mars using HOG features and a SVM classifier, RecPad2009. 2009

[73] L. Wei, T.F. Stepinski, Computer-generated global map of valley networks on Mars, Journal of Geophysical Research, Part E: Planets, Vol. 114, No. E11, E11010, 2009

[74] L. Wei, T.F. Stepinski, Topographically derived maps of valley networks and drainage density in the Mare Tyrrhenum quadrangle on Mars, Geophysical Research Letters, Vol. 33, No. 18, 2006

[75] L. Molloy, T.F. Stepinski, Automatic mapping of valley networks on Mars, Computers and Geosciences, Vol. 33, No. 6, pp 728-738, 2007

[76] D.A. Vaz et al. Automatic Detection and Classification of Fault Scarps on MOLA data, Centro de Geofisica, Universidade de Coimbra, Lunar and Planetary Science Conf, 2006

[77] K. Popper, The Logic of Scientific Discovery (reprint), Routledge, ISBN 978-0-415-27843-0, 2009

[78] T.S. Kuhn, The Structure of Scientific Revolutions (reprint), 4th Edition, The University of Chicago Press, ISBN 978-0-226-45812-0, 2012

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

[80] P.A. Bromiley, N.A. Thacker, Multi-dimensional Medical Image Segmentation with Partial Volume and Gradient Modelling, Annals of the BMVA, Vol. 2008 no. 2 pp. 1-22, 2008

[81] W.H. Press et al. Numerical Recipes in C The Art of Scientific Comptuting, Second Edition, Cambridge University Press, ISBN 978-81-85618-16-6, page 656, 2009

[82] J.W. Osborne, E. Waters, Four Assumptions Of Multiple Regression That Researchers Should Always Test, Practical Assessment, Research, and Evaluation, 8(2), 2002

[83] Robbins, S.J. et al., The Variability of Crater Identification Among Expert and Community Crater Analysts. Planetary Crater Consortium, 4, #1305, 2013

[84] A.P. Dawid, Probability forecasting, Encyclopedia of statistical science, Vol. 7, pp. 210-218, Wiley, 1986

[85] E.R. Davies, Editorial: The Laziness Syndrome, The Newsletter of the British Machine Vision Association and Society for Pattern Recognition, Vol. 24, No. 1, 2013

[86] K. Laws, Textured Image Segmentation, Ph.D. Thesis, Internal Report 940, Image Processing Institute, University of Southern California, 1980

[87] C. Gui et al., A SIFT-Based Method for Matching Desired Keypoints on Mars Rock Targets, i-SAIRAS: International Symposium on Artificial Intelligence, Robotics and Automation in Space

[88] M. Lourakis et al., Autonomous Visual Navigation for Planetary Exploration Rovers, Proceedings of the 12th Symposium on Advanced Space Technologies in Automation and Robotics, ESA/ESTEC, Noordwijk, the Netherlands, May 15-17, 2013

[89] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60, 2, pp. 91-110, 2004

[90] T.D Barfoot et al., Place Revisiting for Planetary Rovers: An Enabling Technology and Field Testing of Three Mission Concepts, ICRA 2013, 2013

[91] H. Bay et al, SURF: Speeded Up Robust Features, Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346359, 2008

[92] M. Calonder et al., BRIEF: Binary Robust Independent Elementary Features, ECCV 2010, Lecture Notes in Computer Science Volume 6314, pp 778-792, 2010

[93] P. Kisilev, D. Freedman, Parameter Tuning by Pairwise Preferences, Proceedings of the British Machine Vision Conference, 2010

[94] N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection, Computer Vision and Pattern Recognition, 2005

[95] R.S. Stephens, Probabilistic approach to the Hough transform, Image and Vision Computing, Volume 9, Issue 1, The first BMVC 1990, February 1991, Pages 66-71, ISSN 0262-8856, 1990

[96] H. Kalviainen et al. Probabilistic and non-probabilistic Hough transforms: overview and comparisons, Image and Vision Computing, Volume 13, Issue 4, 1995

[97] B. Bascle, X. Gao and V. Ramesh, Parametric and Non-parametric methods for linear extraction, Proceedings of Workshop on Statistical methods in Video Processing, SMVP 2004, LNCS 3247, pp. 175-186, 2004

[98] R. Lane et. al, Stretch Correlation as a Real-Time Alternative to Feature Based Stereo Matching Algorithms, Image and Vision Computing, 12, 4, pp 203-234, 1994

[99] S. Crossley et. al, Improving Accuracy, Robustness and Computational Efficiency in 3D Computer Vision, IVC, 22, 5, 399-412, 2004

[100] T.F. Cootes et al. Active Appearance Models, in Proc. European Conference on Computer Vision 1998, Vol. 2, 484-498, Springer, 1998

[101] S. Morita et al., Approach to Crater Chronology with Fourier Transform of Digital Terrain Models, 41st Lunar and Planetary Science Conference, 2010

[102] N. Sebe, M.S. Lew, Wavelet-based Texture Classification, International Conference on Pattern Recognition (ICPR'00), vol III, pp. 959-962, 2000

[103] Jain A, Unsupervised Texture Segmentation Using Gabor Filters, Pattern Recognition Vol 24, No 12, 1167-1186, 1991

[104] Jolliffe I T, Principle Component Analysis, New York Springer, University of Geneva, ISBN: 0-387-96269-7, 1986

[105] Scholkopf B et al. Kernel Principle Component Analysis, Max Planc Institute, Kybernetik, Spemannstr, 1998

[106] Tipping M, Bishop C, Probabilistic Principal Component Analysis, Microsoft Research, Cambridge UK, 1999

[107] H.H. Harman, Modern Factor Analysis 2nd edition, University of Chicago Press, 1967

[108] Lawrence N, The Gaussian Process Latent Variable Model, Department of Computer Science, University of Sheffield, 2006

[109] N. Lawrence, Probabilistic Non-Linear Principle Component Analysis with Gaussian Process Latent Variable Models, Jou. Mach. Learn. Res., 6, 1783-1816, 2005

[110] P. Comon, Independent Component Analysis - a new concept? Signal Processing 36, 287-314, 1994

[111] A. Hyvarinen, Survey on Independent Component Analysis, Neural Computing Surveys 2, 94-128, 1999

[112] C. Jutten, J. Herault, Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture, Signal Processing, 24:1-10, 1991

[113] A. Belouchrani, J.F. Cardoso, Maximum likelihood source separation by the expectation-maximization technique: deterministic and stochastic implementation, Proc. NOLTA, 49-53, 1995

[114] P. Pajunen, Blind source separation using algorithmic information theory, Neurocomputing, 22:35-48, 1998

[115] J.F. Cardoso, Source Separation using Higher Order Moments, Proc. ICASSP'89, 2109-2112, 1989

[116] E. Oja, Nonlinear PCA criterion and maximum likelihood in independent component analysis, Proc. Int. Workshop on Independent Component Analysis and Signal Separation, 143-148, 1999

[117] I. Steinwart, A. Christmann, Support Vector Machines, Springer, ISBN 978-0-387-77241-7, 2008

[118] T. Ho, Random Decision Forest, 3rd Int'l Conf, on Document Analysis and Recognition, 278-282, 1995

[119] Y. Freund, Boosting a weak learning algorithm by majority, Proceedings of the Third Annual Workshop on Computational Learning Theory, 1990

[120] W. Li, X. Gao, Y. Zhu, V. Ramesh, and T.E. Boult, On the Small Sample Performance of Boosted Classifiers, Proceedings of CVPR conference, pp. 574-581, vol. 2, 2005

[121] C.M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 66 ff, 1995

[122] J. Calleja, O. Fuentes, Automated Classification of Galaxy Images, Lecture Notes in Computer Science, Vol. 3215, pp.411418, 2004

[123] S. Kasivajhula, N. Raghavan, H. Shah, Morphological galaxy classification using machine learning, Monthly Notices Royal Astron. Soc., Vol. 8, pp.1-8, 2007

[124] W.M. Bolstad, Introduction to Bayesian Statistics Second Edition, John Wiley & Sons, ISBN 978-0-470-14115-1, 2007

[125] M.R. Spiegel et al., Probability and Statistics Third Edition, McGraw Hill, ISBN 978-0-07-154425-2, page 5, 2009

[126] C. F. Jeff Wu, On the convergence properties of the EM algorithm, The Annals of Statistics, 11:95103, 1983

[127] B.D. Ripley, Appendix A in Pattern Recognition and Neural Networks, Cambridge University Press, 1996

[128] W. Foerstner, 10 Pros and Cons Against Performance Characterisation of Vision Algorithms, Proceedings of ECCV Workshop on Performance Characteristics of Vision Algorithms, 1996. Also in Machine Vision Applications, 9 (5/6), pp.215-218, 1997

[129] N.A. Thacker et al., Performance Characterisation in Computer Vision: A Guide to Best Practices, CVIU, 109, 305-334, 2008

[130] P. Courtney, N.A. Thacker, Chapter: Performance Characterisation in Computer Vision: The Role of Statistics in Testing and Design, Imaging and Vision Systems: Theory, Assessment and Applications, Jacques Blanc-Talon and Dan Popescu (Eds.), NOVA Science Books, ISBN 1-59033-033-1, 2001

[131] R. Kohavi, F. Provost, On Applied Research in Machine Learning, In Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, Vol. 30, 1998

[132] T. Fawcett, ROC Graphs: Notes and Practical Considerations for Researchers, Pattern Recognition Letters, 27(8):882891, 2004

[133] G.S. Rees, W.A. Wright, P. Greenway, ROC Method for the Evaluation of Multi-class Segmentation/Classification Algorithms with Infrared Imagery, Proc. BMVC 2002, vol. 2 pp537-546, 2002

[134] L. Costa et al., Tuning Parameters of Evolutionary Algorithms Using ROC Analysis, 2nd International Workshop on Practical Applications of Computational Biology and Bioinformatics, Advances in Soft Computing Volume 49, 2009, pp 217-222, 2008

[135] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, International Joint Conference on Artificial Intelligence (IJCAI), 1995

[136] B. Efron, R. Tibshirani, Improvements on cross-validation: The .632 + Bootstrap Method, Journal of the American Statistical Association 92 (438): 548560, 1997

[137] B. Efron, R. Tibshirani, An Introduction to the Bootstrap, London: Chapman & Hall, 1993

[138] K. Cho et al., Performance Assessment Through Bootstrap, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 19, No. 11, 1997

[139] G.S. Fishman, Monte Carlo: Concepts, Algorithms, and Applications, New York: Springer, ISBN 0-387-94527-X, 1995

[140] V. Ramesh, R. M. Haralick. Random perturbation models and performance characterization in computer vision, In Proceedings of Conference on Computer Vision and Pattern Recognition CVPR92, pages 521527, 1992

[141] V. Ramesh, R. M. Haralick, A. S. Bedekar, X. Liu, D. C. Nadadur, K. B. Thornton, and X. Zhang. Computer vision performance characterization. In RADIUS: Image Understanding for Imagery Intelligence, pages 241282, 1997

[142] V. Ramesh, R. M. Haralick. Automatic tuning parameters selection for feature extraction sequence. In Proceedings of Computer Vision and Pattern Recognition, pages 672677, 1994.

[143] K. Rohr, Landmark-Based Image Analysis: Using Geometric and Intensity Models, Chapter 3. Performance Characterization of Landmark Operators, Springer, 2001

[144] D.E. Clark, S. Ivekovic, The Cramer-Rao Lower Bound for 3-D State Estimation from Rectified Stereo Cameras, ICIF 2010, pages 1-8, 2010

[145] B. Loesch, B. Yang, Cramer-Rao Bound for Circular Complex Independent Component Analysis, Latent Variable Analysis and Signal Separation Lecture Notes in Computer Science Volume 7191, pp 42-49, 2012

[146] Tichavsky, P., Koldovsky, Z., Oja, E.: Performance analysis of the FastICA algorithm and Cramr-Rao bounds for linear independent component analysis. IEEE Trans. on Sig. Proc. 54(4) (April 2006)

[147] Ollila, E., Kim, H.-J., Koivunen, V.: Compact Cramr-Rao bound expression for independent component analysis. IEEE Trans. on Sig. Proc. 56(4) (April 2008)

[148] R. M. Haralick. Performance Characterization in Computer Vision, chapter Propagating Covariance in Computer Vision, pages 95114. Computational Imaging and Vision. Kluwer Academic Publishers, ISBN-13: 978-0792363743, ISBN-10: 0792363744, 2000

[149] R.M. Haralick, On the Use of Error Propagation for Statistical Validation of Computer Vision Software, (with Xufei Liu and Tapas Kanungo), IEEE Pattern Analysis and Machine Intelligence 27, No. 10, pp. 1603-1614, 2005

[150] W. Foerstner, Diagnostics and Performance evaluation in Computer Vision, Workshop on Robust Computer Vision, 1994

[151] S. Yi et al., Error propagation in machine vision, Machine Vision and Applications, 7:93-114, 1994

[152] Z. Sun, V. Ramesh, A.M. Tekalp, Error Characterization of the Factorization Method, Computer Vision and Image Understanding, 82, 110137, 2001

[153] H. Ragheb et al., Quantitative Shape Analysis with Weighted Covariance Estimates for Increased Statistical Efficiency, Frontiers in Zoology, 10(16), 2013

[154] Gang Liu, Automatic Target recognition using location uncertainty, Phd Dissertation, Department of Electrical Engineering, University of Washington, 2000

[155] Q. Ji, R.M. Haralick, Error Propagation for the Hough transform, Pattern Recognition Letters, Volume 22, pp. 813-823, 2001

[156] Jinyi Qi, A unified noise analysis for iterative image estimation, Physics in Medicine and Biology, vol. 48, No. 21, 2003

[157] J. F. Canny. A computational approach to edge detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(6):679698, 1986

[158] E. R. Davies, Machine Vision: Theory, Algorithms, practicalities, London Academic press, 2nd edition, 1997

[159] T. Markiewicz et al. White Blood Cell Automatic Counting System Based on Support Vector Machine, Adaptive and Natural Computing Algorithms, Lecture Notes in Computer Science, Vol. 4432, pp 318-326, 2007

[160] F. Zhu et al., A New Method for People-Counting Based on Support Vector Machine, Asia-Pacific Conference on Information Processing, vol. 1, pp 109-112, 2009

[161] L. Fan, Y, Liu, Automate fry counting using computer vision and multi-class least squares support vector machine, Aquaculture, Vol. 380383, Pages 9198, 2013

[162] H. Kuba et al., Automatic Particle Detection and Counting by One-Class SVM from Microscope Image, Advances in Neuro-Information Processing, Lecture Notes in Computer Science, Vol. 5507, pp 361-368, 2009

[163] J. Drish, Obtaining Calibrated Probability Estimates from Support Vector Machines, CiteSeerX, 1998

[164] J. Platt, Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In Advances in Large Margin Classifiers, 1999

[165] B. Zadrozny, C. Elkan, C., Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. Proceedings of the Eighteenth International Conference on Machine Learning, 2001

[166] A. Niculescu-Mizil, R. Caruana, Obtaining Calibrated Probabilities from Boosting, Proc 21st Conf. Uncertainty in AI, 2005

[167] V. Lempitsky, A. Zisserman, Learning To Count Objects in Images, Advances in Neural Information Processing Systems 23, 1324-1332, 2010

[168] P. Latinne et al., Adjusting the Outputs of a Classifier to New a Priori Probabilities May Significantly Improve Classification Accuracy: Evidence from a multi-class problem in remote sensing, Proceedings of the Eighteenth International Conference on Machine Learning, 2001

[169] M Greiffenhagen, D Comaniciu, H Niemann, V Ramesh, “Design, analysis, and engineering of video monitoring systems: an approach and a case study”, Proceedings of the IEEE 89 (10), 1498-1517, 2001

[170] A.J. Bell, T.J. Sejnowski, The ‘independent components’ of natural scenes are edge filters, Vision Research, 37:3327-3338, 1997

[171] S. C. Zhu, Y. Wu, D. Mumford, “Filters, Random Fields and Maximum Entropy (FRAME): Towards a unified theory for texture modeling”, International Journal of Computer Vision, vol. 27 (2), pp. 107-126, 1998

[172] J. Portilla and E. P. Simoncelli, A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients, International Journal of Computer Vision, vol. 40 (1), pp. 49-71, 2000.

[173] P.F. Meilan, M. Garavaglia, Rayleigh Resolution Criterion for Light Sources of Different Spectral Composition, Brazilian Journal of Physics, Vol. 27, No. 4, 1997

[174] F.J. Anscombe, The transformation of Poisson, binomial and negative-binomial data, Biometrika 35, 3-4, 1948

[175] B.A. Olshausen, D.J. Field, Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?, Vision Research, Vol. 37, No. 32, pp 3311-3325, 1997

[176] H. Lee et. al, Efficient sparse coding algorithms, Neural Information Processing Systems Conf. (NIPS), 2006

[177] L. Arnold et al., An Introduction to Deep Learning, European Symposium on Artificial Neural Networks (ESANN 2011), 2011

[178] H. Larochelle et al., Exploring strategies for training deep neural networks, The Journal of Machine Learning Research, 10:140, 2009.

[179] P.A. Bromiley et al., Bayesian and Non-Bayesian Probabilistic Models for Magnetic Resonance Image Analysis, Image can Vision Computing, Special Edition: The use of Probabilistic Models in Computer Vision, 21, 851-864, 2003

[180] N.A. Thacker et al., B-Fitting: A Statistical Estimation Technique with Automatic Parameter Selection, Proc. BMVC, 283-292, 1996