Icarus 247 (2015) 77–80

Contents lists available at ScienceDirect

Icarus

journal homepage: www.elsevier.com/locate/icarus

Note

Improved techniques for size–frequency distribution analysis in the planetary sciences: Application to blocks on 25143 Itokawa

I. DeSouza [a], M.G. Daly [a,*], O.S. Barnouin [b], C.M. Ernst [b], E.B. Bierhaus [c]

[a] The Centre for Research in Earth and Space Science, York University, Toronto, Ont., M3J 1P3, Canada
[b] The Johns Hopkins University, Applied Physics Laboratory, Laurel, MD 20723-6099, USA
[c] Lockheed Martin, Space Exploration Systems, Denver, CO 80201, USA

ARTICLE INFO

Article history:
Received 25 June 2014
Revised 8 September 2014
Accepted 6 October 2014
Available online 19 October 2014

Keywords:
Asteroid Itokawa
Asteroids, surfaces
Geological processes
Data reduction techniques
Asteroids

ABSTRACT

The analysis of size–frequency distributions is common for studying planetary bodies with applications for craters and blocks. However, the common method of using a linear regression on cumulative block distributions is subject to systematic errors that can lead to an underestimation of uncertainties and/or a biasing of the slope. The power-law fitting procedure proposed by Clauset et al. is applied for the first time to a block or crater dataset: the global block survey of Asteroid 25143 Itokawa. Along with new results, a discussion of the importance on the preparation and presentation of block distribution statistics is also given in the context of asteroid block populations. Finally, different block sizing methods are evaluated and demonstrate the advantages of a mass-driven approach rather than a size-driven approach for a power-law fitting.

1. Introduction

The analysis of size–frequency distributions is widely used in the planetary sciences to understand geological processes. Examples of these applications include studies of asteroid populations and their collisional histories (e.g. Davis et al., 2002; Bottke et al., 2005; Mazrouei et al., 2014), cratering processes and aging (e.g. McEwen and Bierhaus, 2006) and block distributions (e.g. Thomas et al., 2002; Küppers et al., 2012), to name a few. The basis of these analyses is that many natural phenomenon follow power law distributions of the form

$$N \propto R^{-\alpha}$$

where N is the number of objects expected in a size range R where α is commonly referred to as the slope of the distribution or the scaling parameter. Differences in the scaling parameter are used to differentiate between geological processes. The methodology used for the analysis often makes use of the linearization

$$\log N \propto -\alpha \log R.$$

Two analysis approaches have been detailed by the Crater Analysis Techniques Working Group (1979) – differential (or relative) analyses and cumulative analyses. Differential analyses are performed by grouping the objects to be counted (i.e., blocks or craters) into size bins. Cumulative analyses are performed by calculating the number of objects greater than a given size. Additionally, the cumulative distribution, which is the integral of the differential distribution, can be plotted without binning and often shows a smoother behavior with increasing size. This smoother behavior often makes cumulative distributions the preferred choice for comparative studies.

As a specific example, block populations generally appear to follow these power-law distributions where the probability of selecting a block of a certain size, x, is given as $P(x) \propto x^{-\alpha}$ where α is the scaling parameter mentioned previously. If the differential distribution follows such a power-law, the cumulative distribution function (CDF) will also follow a power law as they are related through an integral. The scaling parameter is related to the slope of the cumulative distribution by $S = \alpha + 1$, where S is the slope of the CDF. Uncertainties for these block counts are calculated assuming Poisson statistics with an uncertainty on N blocks given by $\pm\sqrt{N}$.

An example of an unbinned cumulative distribution for blocks on Itokawa is shown in Fig. 1 (Mazrouei et al., 2014). Blocks were defined as rocks and features with distinctive positive relief that are larger than a few meters in size. The blocks were fitted with ellipses, where the semi-major and semi-minor axes provided long and short axes for each block. Most blocks larger than 5 m along their major-axis were mapped but blocks below 6 m are not fully represented due to image resolution limitations. Therefore, 6 m was the lower size limit used in calculating the least-squares fit to these data. This limit was chosen through qualitative evaluation of several linear fits to the cumulative distribution. For this study, the unknown vertical axis of each block was assumed to be equal to the horizontal minor axis of the block. The diameter of a sphere whose volume was equivalent to that of an ellipsoid defined by the length of these minor and major axes was computed and used to quantify the measured block size. This approach was used to approximate block size by a mass-related value for ease in comparison with laboratory studies. A total of 1433 blocks were measured over the entire surface area of 0.4011 km^2. Here the slope of -3.5 ± 0.1 is taken to be representative of the geological process that created and modified the $\geqslant 6$ m block population.

2. Limitations of the standard analysis technique

Inherent in the standard least-squares fitting to a dataset that is commonly used in the literature (e.g., Thomas et al., 2002; Michikami et al., 2008; Küppers et al., 2012; Mazrouei et al., 2014), is the assumption of the statistical independence of the dataset on a point-by-point basis. For a CDF, the assumption of independence fails as each point is related to the previous points, with the amount of correlation

[*] Corresponding author.
E-mail address: dalym@yorku.ca (M.G. Daly).

http://dx.doi.org/10.1016/j.icarus.2014.10.009
0019-1035/© 2014 Published by Elsevier Inc.

decreasing with separation between the points. For example, for a size-ordered, unbinned dataset of the type shown in Fig. 1, where the size of item i is less than the size of item $(i+1)$, the CDF is calculated by the recursive relation:

$$N(i) = N(i+1) + 1,$$

where $N(i)$ is the number of items of size greater or equal to the size of the ith item. Clearly as a relationship exists that relates each value of $N(i)$ to any other value with differing i, the values of N are not independent and the least-squares approach and its subsequent error analysis is not correct. The assumption of statistical independence will result in an under-estimation of the uncertainties in the least-squares fit to the data set.

Additional problems exist with the standard approach. For example, Fig. 1 exhibits a roll-over at smaller scales. This is generally a feature of these datasets as, at some scale, all such distributions will eventually depart from linear. The departure can be caused by incompleteness of the dataset as in Fig. 1, where image coverage and resolution limitations are the cause of the roll-over. Alternatively, the roll-over can be caused by a change in the physical processes that are dominant at those scales. The location of the start of the roll-off is generally chosen by visual inspection, based on image resolution or semi-quantitatively by inspection of various fits. This approach causes two potential problems. The first is due to the biased nature of this process – two investigators with identical datasets are not guaranteed to achieve identical results. Additionally, the smaller sizes in the distribution tend to control the fit due to the smaller uncertainties associated with the larger block count – making the selection of this point critical to the fit results. An analytical method for choosing the minimum block size of the distribution and assessing the uncertainty in its determination would improve both uncertainty estimation of the fit and intercomparison between analyses.

Bierhaus (2004) points out a number of additional problems with typical analysis approaches and recommends a differential approach. The first problem is due to a common non-ideal implementation of linearizing the distribution. Although a good estimate of the uncertainty in a count is symmetric about the count, when the problem is linearized, these uncertainties are not symmetric. Most fitting routines assume Gaussianly distributed symmetric errors leading to erroneous results and therefore, a non-linear fit in non-log–log space is recommended. Bierhaus (2004) also states that this can also be violated for counts < 10 where the Poisson distribution no longer approximates a Gaussian distribution. Another point that should be mentioned but is not discussed in the remainder of this paper is that of systematic errors. Bierhaus (2004) concludes that these uncertainties can dominate at the small scales and high counts where Poisson statistics provides high confidence. This is a consequence of statistical errors $\sim \sqrt{N}$ and systematic errors $\sim N$. As systematic error sources differ greatly depending on the data available, how it was collected and how it was assessed – we leave this as a cautionary reminder for future studies. The Bierhaus (2004) method is recommended over fitting to the CDF but has the disadvantage of several requirements: selection of bin sizes, selection of the minimum bin size and an approach for dealing with empty bins.

The CDF, by virtue of being statistically related on a point-to-point basis can mask distributions that are not good fits to power-laws. In fact, it is this well-behaved nature of a CDF that makes them popular. Clauset et al. (2009), who recognized the above problems with CDF analysis, provides examples from various disciplines that appear as good linear fits to a CDF but the underlying distributions are non power-law distributions such as log-normal, or exponential distributions. Additionally, they provide a methodology for fitting these distributions that accounts for the problems previously identified. In this paper, we present a summary of the techniques proposed by Clauset et al. (2009) for the analysis of size–frequency distribution data. We find these techniques applicable to the analysis of blocks or craters in the planetary sciences and apply them for the first time for these purposes. As a test of the approach we reevaluate the block size–frequency distribution for Itokawa and compare the results with a previous analysis (Mazrouei et al., 2014). Additionally, we evaluate a variety of block sizing approaches and evaluate their impact on the goodness-of-fit to a power-law.

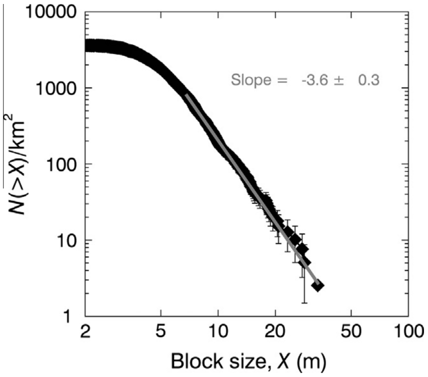

Fig. 1. The global cumulative size–frequency distribution of blocks on Itokawa as a function of block size using the equivalent spherical radius sizing method.

3. Revised method

Clauset et al. (2009) proposed a three-step procedure as an alternative to linear regression methods; we use that method and notation in this work. The three steps are:

1. Estimate x_{min} and the scaling parameter α.
2. Calculate the goodness-of-fit between the data and the power law in order to provide an assessment of the power-law hypothesis.
3. Compare the power-law with alternative hypotheses via a likelihood ratio test.

Only steps 1 and 2 will be covered here as we have found good agreement with our datasets and power-law behavior.

3.1. Estimation of the power-law parameters

The general form of the PDF is a probability distribution function that integrates to 1 over the interval $x > x_{min}$:

$$p(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\alpha}.$$

To apply this model PDF to a dataset, it must be scaled by the number of blocks with $x > x_{min}$. Power-law fitting results are expressed as the estimate, $\hat{\alpha}$, of the true scaling parameter α and the standard error σ of $\hat{\alpha}$, along with the minimum block size x_{min} and its standard error $\sigma_{x_{min}}$. To estimate α, the method of maximum likelihood is used. Assuming that the distribution follows a power law then the estimate of the scaling parameter, $\hat{\alpha}$, can be obtained as

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1},$$

with standard error given as

$$\sigma_{\hat{\alpha}} = \frac{\hat{\alpha} - 1}{\sqrt{n}} + \text{terms of order } (1/n).$$

Here n is the number of objects (the sample size) used to calculate $\hat{\alpha}$. The value of n is known a priori but x_{min} is not. Estimating α and x_{min} is done by calculating $\hat{\alpha}$ for every possible x_{min} and evaluating the goodness of the fit to the model data. Clauset et al. (2009) tested a variety of methods and suggest the best method to be a Kolmogorov–Smirnov (KS) statistic that calculates the difference between the model and the dataset. Minimization of the KS statistic then provides the best estimates of α and x_{min}.

Synthetic datasets for values of $x > x_{min}$ are easily generated using the calculated power-law model for the data as a probability distribution function and randomly calculating n, where n is the number of values in the dataset where $x > x_{min}$. For $x < x_{min}$, the approach is not as straightforward. The power-law estimate depends on x_{min}, therefore for the synthetic data to generate rational and independent results, the non power law distribution must also be modeled. Here we follow Clauset et al. (2009), where they use the actual data set to randomly select values below x_{min} to guarantee a similar roll-over below x_{min}. Other approaches could be used below x_{min}, for example, the data could be fitted with a simple functional form that is continuous with the power-law model at x_{min} and whose derivative is also continuous at x_{min}; samples could be randomly selected using this distribution and then scattered with Poisson statistics.

The generation of synthetic datasets can be used to estimate the confidence interval for x_{min}. The standard deviation of the estimates for x_{min} for each synthetic dataset provides this assessment.

3.2. Assessment of the fit to a power law

Having calculated the best estimate of the power law from the dataset, is the power law fit a good model? The recommended approach is to use a power law hypothesis based on the estimates of x_{min} and α to generate random power law (synthetic) datasets and assess the goodness-of-fit to the power law based on comparing the synthetic dataset to its estimated power law. If the ratio of synthetic data that have a KS distance, D, greater than the D of the estimated power law is greater than 0.1, as suggested by Clauset et al. (2009), then the estimated power-law model would be considered a good fit to the dataset. If the departures are significantly larger for the dataset, then the model should be questioned. A large probability ($p > 0.1$) suggests a good fit while a small probability suggests that random fluctuations from the power law are not adequate to explain the dataset and alternative forms may be more appropriate.

The number of synthetic datasets tested impacts the uncertainty in the calculated probability. An empirically derived "rule-of-thumb" suggested by Clauset et al. (2009) for the uncertainty (ϵ) in the probability (p) is