LINEAR POISSON MODELS: A NEW TOOL FOR LUNAR AND PLANETARY SCIENCE

P.D. Tar, N.A. Thacker, J.D. Gilmour, University of Manchester, Oxford Road, Manchester, UK Paul.Tar@manchester.ac.uk

Introduction: Data distributions summarized as sampled histograms are commonly found within planetary science, such as ToF mass spectra. Other data can often be easily converted to histograms, such as images by sampling patterns of pixels. However, making efficient use of histograms isn't always possible using traditional statistical methods, as low parameter descriptions (e.g. based on moments, percentiles, frequency ratios etc.) struggle to express all complexities and variations. This motivates the use of contemporary A.I. and machine learning approaches (e.g. deep learning, decision trees etc.), which are seen as more powerful. However, these contemporary methods often lack statistical rigor; are empirically driven (i.e. follow a try-it-and-see-what-happens philosophy); and overlook the theoretical consequences of different noise assumptions. We therefore introduce our new machine learning method – Linear Poisson Models (LPMs) – designed to bridge the gap between traditional statistical methods and contemporary A.I.

LPMs [1] are capable of supervised, semi-supervised and unsupervised learning of complex data distributions. The method describes histograms as linearly weighted combinations of probability mass functions (PMFs), where weights can be interpreted as quantity measurements and PMFs can be interpreted as components of variation:

$$H_X \approx \sum_k P(X|k) Q_k$$

where X is a histogram bin of $\mathbf{H}$, $P(X|k)$ is the probability of an X event, given component k, and $\mathbf{Q}$ is the quantity of a component present in the data. In addition, an LPM error analysis provides theoretical predictions of statistical and systematic effects of Poisson sampling noise in training (model) and testing data (data), and a chi-square goodness-of-fit verifies that data is being appropriately modeled:

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

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

where error covariance matrix elements consider changes in quantities, $\mathbf{Q}$, with respect to noise in incoming data, $\mathbf{H}_X$, and previously modelled components $\mathbf{H}_{X|k}$ from training data. We have successfully demonstrated LPMs for:

- making terrain area measurements from martian imagery [2];
- estimating false positive crater annotations in citizen science lunar data [3][4];
- and correcting Xenon ToF spectra for blank and contamination [5].

In each case, raw data is preprocessed to best approximate the statistical assumptions made by LPMs, i.e. histograms with independent Poisson bins, generated from linearly additive sources describable as a combination of fixed PMFs.

Martian Terrain Area Measurements: HiRISE imagery from varied martian terrains (dunes, chaotic terrains, CO2 features...) were sampled using a 'Poisson blob' encoding, which captured local variations in light/dark pixel patterns. The encoding, inspired by BRIEF, grouped similar adjacent patterns together, permitting blobs to be treated as spatially occurring Poisson events. Combinations of blob PMFs approximated the varying complex textures found within the terrain images, with groups of PMFs being associated with different classes. Large quantities of ground-truth imagery were generated via a bootstrap re-sampling technique, allowing compositions of test images to be controlled. Histograms of pixel pattern frequencies were used to train and test LPMs, with linear weights being used as area measurements. Area measurements were achieved with accuracies between **0.6%** and **1.6%**, with error predictions matching within a factor of 2, confirming the error theory accounts for the majority of noise effects.

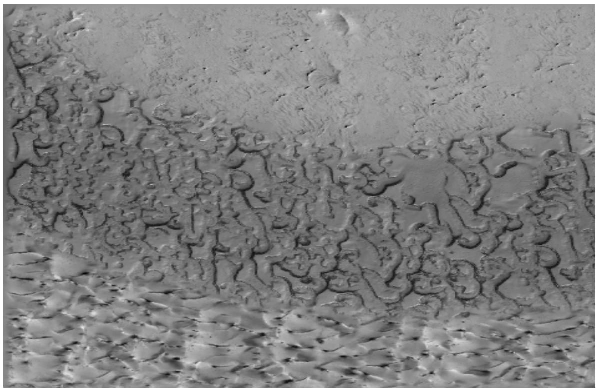

Fig 1. Example of composite martian terrain image used for area quantification tests. This image contains dunes and CO2 features.

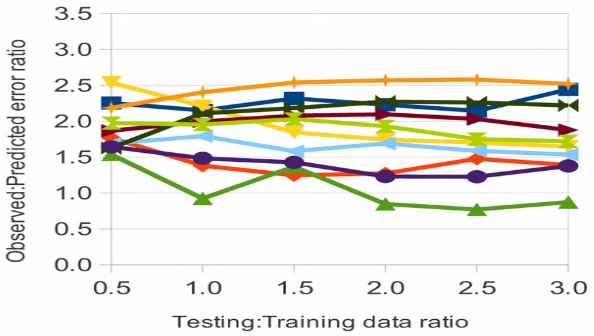

Fig 2. Agreement between predicted and empirical measurement accuracy for varied martian terrains.

Moon Zoo False Positive Corrections: The citizen science project, Moon Zoo, enlisted thousands of volunteers to annotate lunar craters. A combination of crater ambiguity and malicious misuse lead to approximately 25% of annotations being false positives. LPMs were applied to learn the distribution of true and false craters, based upon histograms of template match scores. A template crater (the average appearance of a collection of confirmed craters) was compared to each candidate crater to create distributions of scores. Two types of templates (grey levels and derivatives of grey levels) were tested, along with two match scores (mean squared error, and normalised dot product). Bootstrap re-sampling was applied to generate large quantities of ground-truth crater annotations. We found that derivative templates and dot product match scores produced the most separable distributions, allowing corrected crater counts to be achieved with close to limiting case Poisson uncertainty.

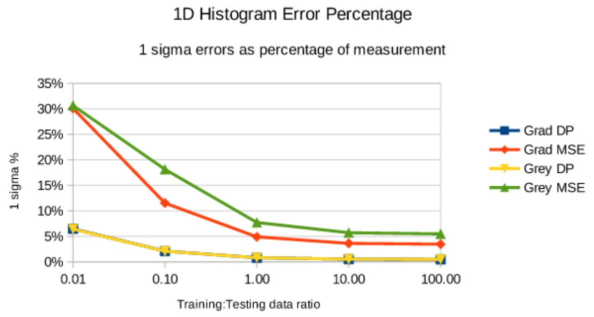

Fig 3. Errors on corrected crater counts for different templates and match scores. 20,000 original annotations were used during experiments.

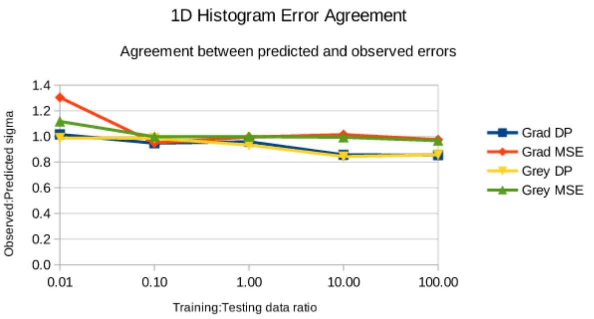

Fig 4. Agreement between predicted and empirical measurement accuracy for different templates and match scores.

Xenon ToF spectra corrections: Xenon isotope ratios are used to identify the origins and histories of solar system samples. Raw spectra can be contaminated and also suffer from significant amounts of variable `blank'. LPMs were used to demonstrate that improved isotope ratio measurements could be made by modeling contamination and blank in a semi-supervised manner. Raw spectra were first aligned using a Fourier interpolation method, then baseline corrected using a locally weighted averaged background before peaks were finally integrated into individual histogram bins. Example of blank spectra were used to train a LPM, which was then fitted with an additional component to model signal. Experiments using air calibration samples showed that measurement precision could be doubled in comparison to a conventional blank subtraction approach – equivalent to having 4 times as much data.

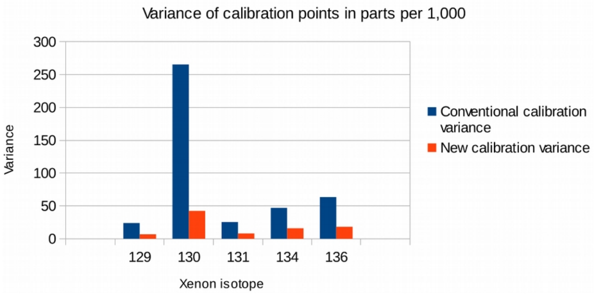

Fig 5. Accuracy comparison between a conventional air calibration using blank subtraction vs LPM analysis.

Acknowledgments: We thank STFC (ST/M001253/1), Leverhulme Trust (RPG-2014-019) and the Moon Zoo team.

References: [1] P. D. Tar, N. A. Thacker, (2014) Annals of the BMVA, vol 2014:1, 1-22 [2] P. D. Tar, N. A. Thacker, J.D. Gilmour, M.A. Jones, (2015) Advances in Space Research, vol 56:1, 92-105 [3] R. Bugiolacchi, S. Bamford, P.D. Tar, N.A. Thacker, I.A. Crawford, K.H. Joy, P. M. Grindrod, C. J. Lintott, (2016), Icarus, 1, 0019-1035 [4] P.D. Tar, N.A. Thacker, J.D. Gilmour, (2016) Earth Moon & Planets [5] A.P. Seepujak, N.A. Thacker, P.D. Tar, J.D. Gilmour, (2016) submitted to JAAS