Tina Memo No. 2012-006
Internal

A Connected Blob Image Representation for Poisson Linear Models.

Paul Tar and Neil.A.Thacker.

Last updated
11 / 9 / 2012

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

A Connected Blob Image Representation for Poisson Linear Models

abstract

Tina Memos 2011-003, 2011-004 and 2012-003 describe statistical techniques for determining the composition of linearly combined histogram data assuming Poisson generation processes. Tina Memos 2011-002, 2011-005 and 2012-004 suggest the use of a BRIEF-like representation for translating images into linear histograms so these statistical techniques can be applied to make measurements of surface textures in planetary terrains. The simple BRIEF-like representation has proven ineffective and unstable. We believe this is due to spatial correlations which are not correctly accounted for if images are considered on a location-by-location basis assuming independence between each pixel. An alternative representation is presented here which groups image data points together which share common BRIEF-like patterns. It is believed that the resulting connected blobs share a common spatial property and appear as individual Poisson events (or are approximated as Poisson events), thus when applied to image data Linear Poisson Models should be blob-based. As blobs can be irregular in shape and size the new encoding requires additional interpretation to convert model quantities into meaningful area measurements. This conversion also requires additions to be made to error estimates. The method of area measurement and two alternative approaches to area covariance calculation are presented. Details and results using the new representation and error models are presented from Monte-Carlo simulated histograms and simulated Martian data.

1 Blobs and Areas

Previously, histograms were constructed with bins over the BRIEF pattern space, with each bin (i.e. pattern) represented by an X. In the new blob representation the meaning of each X has been updated to reflect connected image points sharing a common BRIEF pattern such that an entry is made in a histogram bin for individual blobs, rather than multiple entries being made for each constituent connected image location. As blobs encompass multiple image locations the X now encodes two pieces of information: π, the BRIEF descriptor common to all image locations within the blob (formally, this was the entire representation); and γ, a size band indicating the size of the blob. The size bands are organised into discrete logarithmic size bins to span a large dynamic range of possible blob sizes whilst keeping the pattern space reasonably small. For example, a base-2 log binning would have size bands of $\gamma_0 \in \{1\}$, $\gamma_1 \in \{2, 3\}$, $\gamma_2 \in \{4, 5, 6, 7\}$, etc.

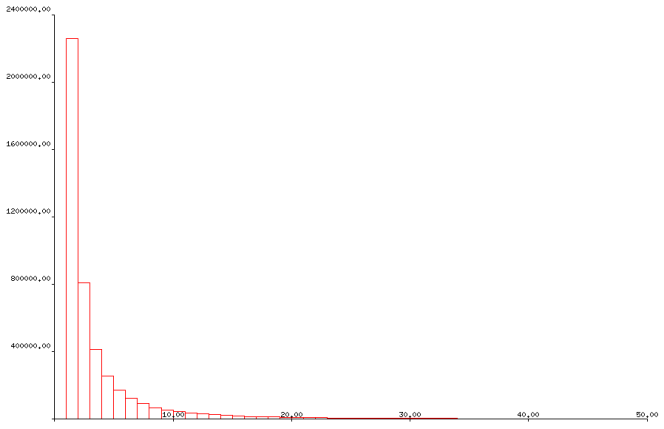

Figure 1: Distribution of blob sizes from typical Martian terrain image

In addition to the X encoding used in the Linear Poisson Model, the location and precise size (in pixels) of each blob is recorded separately for use in area calculations.

The range of blob sizes in real data follows an exponential distribution with the majority of blobs being small in size, with larger blobs becoming less common. Figure 1 shows this distribution for a selected Martian terrain image. The logarithmic size banding used by the new representation allows the within-size-band variation to be approximated by a uniform distribution between the lower and upper size band boundaries.

The vector of model weights, $\mathbf{Q} = \{Q(k=1), Q(k=2), \ldots, Q(k=n)\}$, derived from the Linear Poisson Model are now interpreted as blob quantities

$$Q(k) = \sum_X P(k|X)H(X)$$

where $Q(k)$ is the estimated quantity of blobs associated with model component k; $P(k|X)$ is the probability that component k was the source of a blob of type X; and $H(X)$ is the histogram frequency of blobs of type X. The estimated area covered by component k is then

$$A(k) = \sum_d P(k|X_d)a_d = \sum_X P(k|X)a_X$$

$$a_X = \sum_{d,\delta(X_d=X)} a_d$$

where $A(k)$ is the area estimate; the sum over d is a sum over each individual blob within the data; and a_d is the specific area of blob d. Areas can be expressed as a sum over blob types, X, with specific blob sizes being summed separately

for all blobs which share a common X. The blob sizes, a_d, are drawn from the exponential distribution (or approximated by the local X size band uniform distribution) and the Central Limit Theorem can be applied to the accumulation of each a_X, making them approximately Gaussian distributed.

2 Area Errors

The Linear Poisson Model provides a n by n covariance matrix, $\mathbf{C}_Q$, for the n histogram component weights (blob quantities) from which an area covariance matrix can be computed. We consider two methods: dividing the covariance matrix across individual X bins then accumulating an area covariance by scaling the individual X contributions by the corresponding blob areas; and applying conventional error propagation to the sources of uncertainty within the area estimation calculation.

2.1 Per X bin covariance scaling approach

The covariance for blob quantities is composed of statistical and systematic errors, $C_{\mathbf{Q},data}$ and $C_{\mathbf{Q},model}$, where the statistical errors come from the fluctuation in the data sample and the systematic errors are due to the errors arising from the estimated model components. As a consequence statistical data errors are independent over the set of blobs, while systematic model errors are identical (i.e. not independent) between blobs of the same X. As the model is based upon a non-parametric (histogram based) density estimate the systematic model errors can instead be considered independent over the set of X.

For systematic errors, logically we must be able to write the total covariance as a sum over X's.

$$C_{\mathbf{Q},model} \;=\; \sum_X C_{\mathbf{Q},model}(X)$$

then $C_{\mathbf{Q},model}(X)$ can be interpreted as the contribution to the measurement covariances arising due to **all equivalent** X patterns and giving rise to the contribution $P(k|X)H(X)$ to total $Q(k)$. We can write this as a measurement and covariance for each contribution to $\mathbf{Q}$

$$\mathbf{Q}_X \;=\; H(X)\mathbf{P}(X) \pm C_{\mathbf{Q},model}(X)$$

where $\mathbf{P}(X)$ is the vector of posterior probabilities $[P(0|X), P(1|X), P(2|X), ...P(K|X)]^T$. By simple scaling, the contributions to the total area $P(k|X)a_X$ must therefore have an associated covariance of $C_{\mathbf{Q},model}(X)(a_X/H(X))^2$.

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

$$C_{\mathbf{A},model} \;=\; \sum_X \frac{a_X^2 C_{\mathbf{Q},model}(X)}{H(X)^2}$$

While for the statistical error we have instead

$$C_{\mathbf{Q},data} \;=\; \sum_X C_{\mathbf{Q},data}(X) \;=\; \sum_d M_{\mathbf{Q},data}(X_d)$$

as each measured blob is equivalent we know that $M_{\mathbf{Q},data}(X_d) \;=\; C_{\mathbf{Q},data}(X_d)/H(X_d)$ which is the uncertainty in the final area estimate associated with each individual blob, i.e.

$$\mathbf{Q}_d \;=\; \mathbf{P}(\mathbf{X}) \pm C_{\mathbf{Q},data}/H(X_d)$$

and so continuing as above

$$C_{\mathbf{A},data} \;=\; \sum_d \frac{a_d^2 C_{\mathbf{Q},data}(X_d)}{H(X_d)}$$

We can now recombine these covariances to estimate the uncertainty on $\mathbf{A}$

$$C_{\mathbf{A}} \;=\; C_{\mathbf{A},model} \;+\; C_{\mathbf{A},data}$$

We can test the above expressions for consistency by checking the covariance estimates for the case when blobs are pixels. In this case $a_X \;=\; H(X)$ and

$$C_{\mathbf{A},model} \;=\; \sum_X \frac{H(X)^2 C_{\mathbf{Q},model}(X)}{H(X)^2} \;=\; \sum_X C_{\mathbf{Q},model}(X) \;=\; C_{\mathbf{Q},model}$$

also $a_d \;=\; 1$ and

$$C_{\mathbf{A},data} \;=\; \sum_d \frac{C_{\mathbf{Q},data}(X_d)}{H(X_d)} \;=\; \sum_X C_{\mathbf{Q},data}(X) \;=\; C_{\mathbf{Q},data}$$

as expected.

2.2 Direct error propagation approach

To apply error propagation the area measurement calculation can be rewritten using Bayes Theorem in terms of the two sources of uncertainty: the correlated $Q(k)$ weights and the independent blob sizes.

$$A(k) = \sum_X \left(\frac{P(X|k)Q(k)}{M(X)} \right) a_X = Q(k) \sum_X \frac{P(X|k)a_X}{M(X)}$$

where $P(X|k)$ is the probability of X given histogram component k; and $M(X)$ is the modelled frequency of X. Then error propagation can be applied giving an n by n area covariance

$$\mathbf{C}_A = \nabla_Q \mathbf{C}_Q \nabla_Q^T + \sum_X [\nabla_{a_X} \otimes \nabla_{a_X}^T] \sigma_{a_X}^2$$

where ∇_Q is the matrix of partial derivatives

$$\nabla_{Q,ij} = \frac{\partial A(i)}{\partial Q(j)}$$

and ∇_{a_X} is the vector of derivatives

$$\nabla_{a_X,k} = \frac{\partial A(k)}{\partial a_X}$$

For the case when $i = j$ is

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

$$= \sum_X \frac{P(X|j) Q(j) a_X}{M(X) Q(j)} = \sum_X \frac{P(j|X) a_X}{Q(j)}$$

$$= \frac{A(j)}{Q(j)}$$

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

$$\nabla_{a_X,k} = Q(k) \frac{P(X|k)}{M(X)}$$
$$= P(k|X)$$

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

$$\sigma_{a_X}^2 = H(X) \sigma_{a_{Xd}}^2 = \sum_{d,\delta(X_d=X)} (a_d - < a_{Xd} >)^2$$

which can be done as above using sample variances. Alternatively, assuming a uniform size-band distribution the blob variances can be computed using

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

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

3 Monte-Carlo Simulation Results

Simulated histograms composed of 9 subcomponents spread over 64 X pattern bins were generated with different quantities of training and testing data to test the systematic and statistical contributions to area measurements. Both the per X covariance scaling and the error propagation error estimation methods were applied to predict the accuracy of results over 1,000 trials per experiment. Predicted accuracies were compared to observed accuracies with results plotted in figure 2. Tests were done using blobs of fixed unit size, fixed finite size, random sizes, and random sizes as a function of the X pattern. Both methods achieved good error predictions in most conditions, however the per X covariance scaling method failed in the most realistic tests, that of random blob sizes as a function of X. This excluded the use of this method in further tests.

The effect of the random spread of individual blob sizes within any given X bin is shown to be negligible in figure 3. Here, the effect of omitting this additional variance from the area error estimates ($\sum_X [\nabla_{a_X} \otimes \nabla^T_{a_X}] \sigma^2_{a_X}$) 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). The range of quantities tested was kept small to test the error theory as it approached zero. In this range both quantity and area errors are overestimated, as the true variation in measurements is truncated at zero.

The error propagation method was used when testing with simulated Martian terrains. 30 real Martian images were used to generate synthetic Martian surfaces by tiling and merging randomly selected regions together. Tiles were stretched in both the X and Y direction by up to 10% and additional grey level noise added to each image so that each tile could be considered an independent sample. These images were grouped into sets of 3. Error agreement is shown in figure 4 and model fits and percentage errors in 5. Both quantity and area measurement error predictions achieved agreement with observed errors within a factor of 2 to 3.

4 Conclusion

The method of converting blob quantity estimates into area measurements is both straightforward and effective involving a simple scaling of the former by an average blob-to-area ratio. The conversion of blob quantity errors to area errors requires additional thought, as the conversion ratio itself is subject to additional perturbations due to the random fluctuations in individual blob areas. Attempting to account for these fluctuations on an X bin by X bin basis by distributing the contribution of the quantity error covariance across the pattern space did not achieve the desired results despite the logical arguments for the approach. It is believed this failure is due to additional hidden terms in the $C_{\mathbf{Q}}(X)$ per X covariance contributions which might be revealed by deriving them directly rather than taking the short-cut of dividing the full covariance by the number of X patterns. The alternative error propagation approach achieved good error predictions and also highlighted the negligible effect individual blob size variations had on final area measurements. For practical use the error model could be simplified to $\mathbf{C}_A = \nabla_Q \mathbf{C}_Q \nabla_Q^T$ without loss of useful predicative capabilities.

The final results now demonstrate an overall under-prediction of error by a factor of $1.5 - 2.0$ for the new approach. This is clearly much better than the order of magnitude errors we were seeing when treating each pixel as in dependant, and at this level the error estimates are potentially useful. However, we would like to understand this problem better, and improve the level of agreement further. It is our belief that this error is due to residual correlations between X patterns (i.e. blobs) for specific textures. It should therefore be possible to estimate this effect and correct for this as an effective degree of freedom change during error estimation. This remains an area for future work.

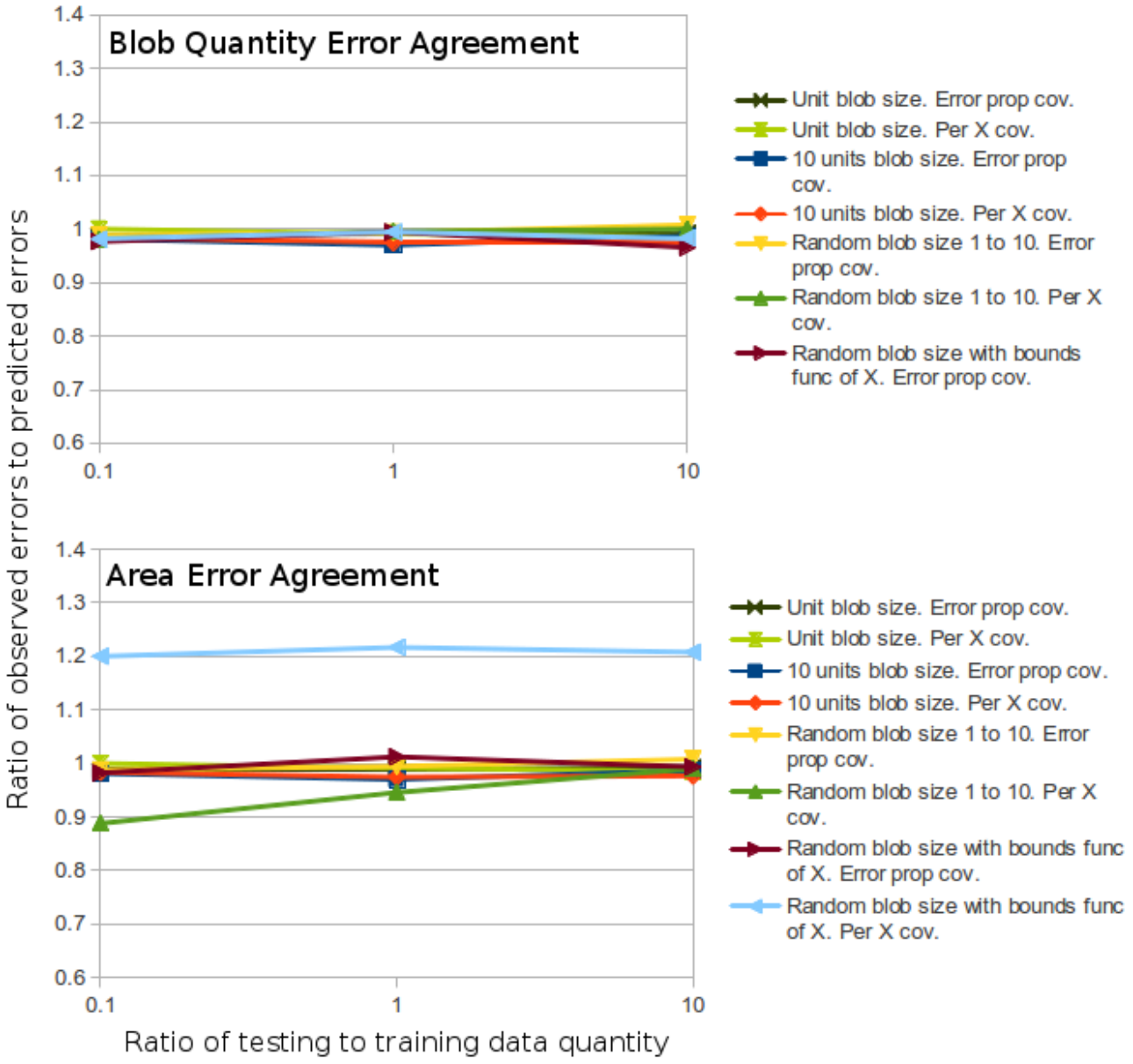

Figure 2: Comparison of error estimation methods using fixed and random blob sizes. The error propagation method produces consistent predictions of error (ratio of observed to predicted errors of unity) over all tests and relative quantities of testing to training data. The per-X covariance scaling method fails to produce consistent error estimates on areas when the random blob sizes are selected with bounds which are a function of X.

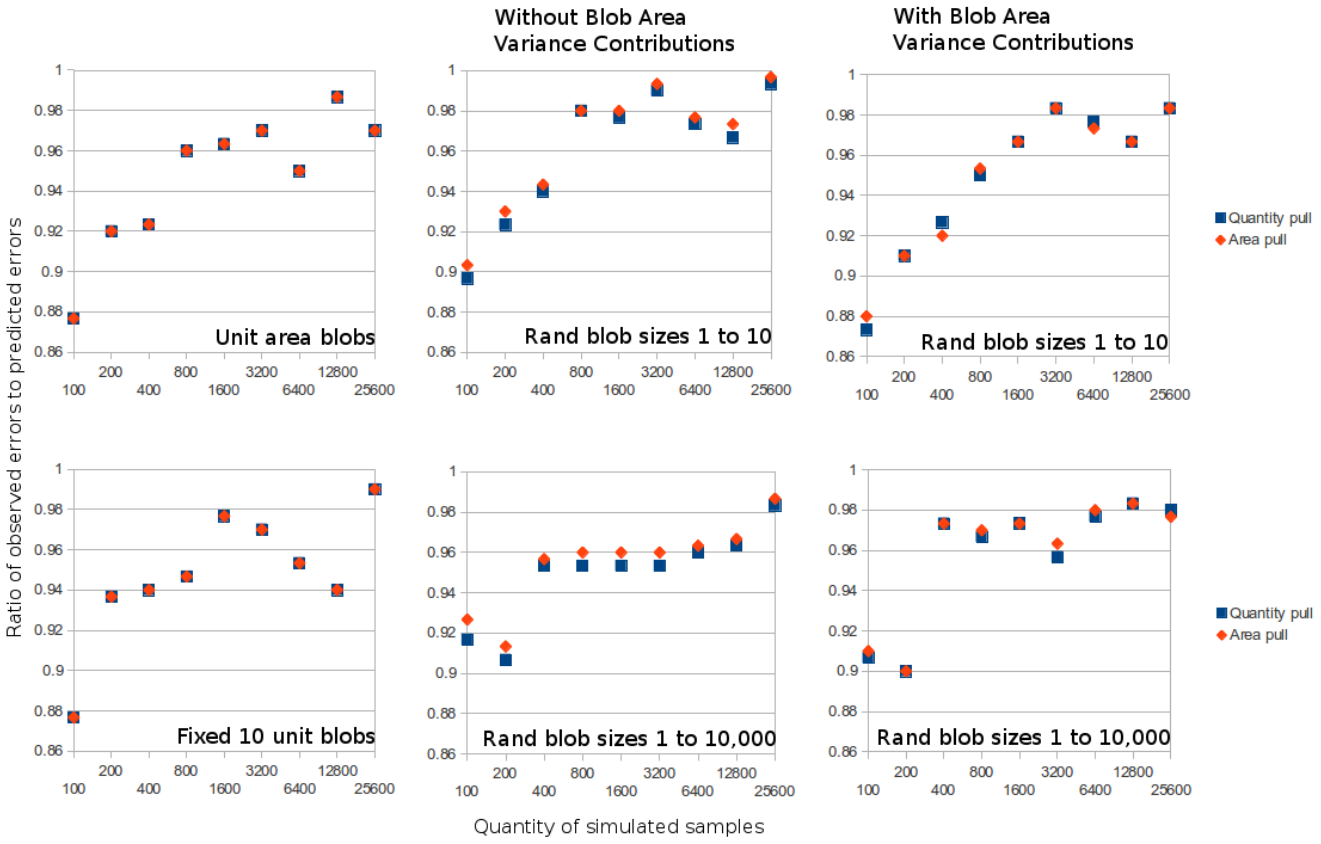

Figure 3: Illustration of the negligible effect variation in blob sizes has upon final area covariances when using the error propagation method. Right: fixed blob sizes, i.e. no variance on blob sizes; Centre: random blob sizes without variance contribution to area errors; Left: random blob sizes with variance contribution to areas. As can be seen, there is a systematic underestimate of area errors unless the additional terms from blob sizes is included.

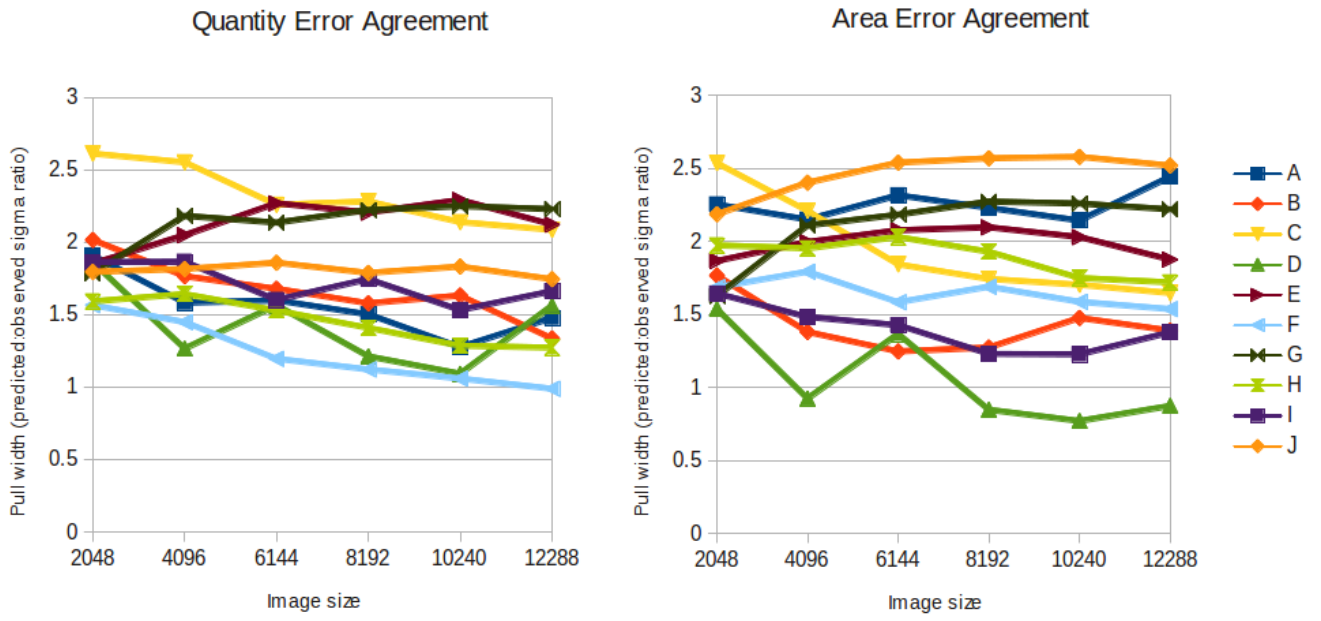

Figure 4: Error agreement when tested with 10 triplets of simulated Martian terrain.

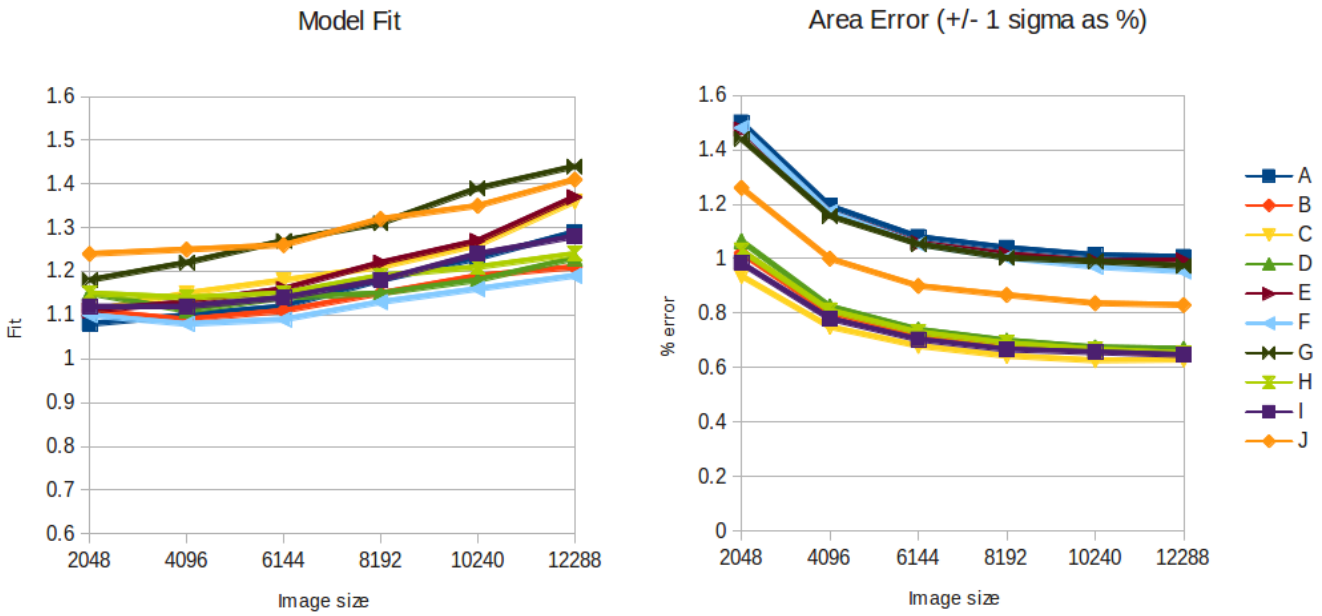

Figure 5: Model fit and predicted errors as percentage of measured areas when tested with simulated Martian terrain.