Bayesian Crater-Counting Analysis Accounting for Observation Error

Xiangmei Zhang

May 20, 2017

1 Introduction

Crater counting is a tool used by astrophysicists to determine the relative ages of two celestial surfaces by comparing the numbers of craters that can be seen in areas defined on the two surfaces. Most of such studies [1–3] have relied on analysts to mark craters, recording their locations and sizes on high-resolution photographs of the surfaces. Crater counts can vary due to individual differences between analysts and even for a single analyst because of inevitable observation error. We have applied a Bayesian method to analyze a large real dataset consisting of crater observations by several analysts on an image of the moon's surface and estimate the real crater number on the area covered by that image. We demonstrate the effectiveness of the method using a simulated data set (where "ground truth" is known and can be used to judge effectiveness).

2 Pretreatment Process

The real crater measurement dataset that motivates our work contains over $25,000$ crater observations made by 13 analysts who mapped and identified sizes of craters in the same region of a *Lunar Reconnaissance Orbiter* Camera (LROC) image. Suppose

the total number of observations is S. For $i = 1, 2, \ldots, S$, each observation s_i consists of crater location (x_i, y_i), a crater diameter d_i, the observer name, and the camera type employed (WAC or NAC). Variation in identification and measurement of craters by different analysts could be produced by individual biases and differences among physical interfaces and methods. A graphic of reported crater locations and diameters represented as circles shows inconsistent but overlapping circles from different analysts.

Therefore, before we analyze real crater data, pretreatment needs to be done to deal with small real variations in crater locations and sizes recorded by different observers for the same crater. We apply a clustering algorithm to attempt to group observations that represent the same crater. For each pair of crater observations s_1 and s_2 (s_i with location (x_i, y_i) and diameter d_i), a dissimilarity between these two crater circles can be defined as

$$D = \frac{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + \alpha(d_1 - d_2)^2}}{\min(d_1, d_2)} \ ,$$

where α is a tuning parameter that adjusts scaling between location and diameter differences.

The resulting set of dissimilarities is used in a hierarchical cluster analysis using the `R` function `hclust()`. A variety of criteria for quantifying dissimilarities of clusters, including average linkage, single linkage, complete linkage, Ward's minimum variance and the unweighted pair group method with arithmetic mean (UPGMA) will tend to allow (for a given number of clusters) single observations to be ungrouped, and are thus attractive for our purposes. We have found that choice between these criteria is best done on a case by case basis. We plot the cluster tree and cut the tree at a height so that it is plausible to treat each branch below the cut as multiple observations of the same individual real crater and every leaf of such a branch as observations from different observers. If the total number of observers is J, the number of crater observations for a single real crater is less than or equal to J. So the cut criteria should ensure that any branch below the cut has a number of leaves no more than J. Within each branch below the cut, we replace the location and diameter records

with the means of leaves in that branch and obtain a cleaned dataset that we use for subsequent analysis.

2.1 Pretreatment Example

We picked a 1000 × 500-pixel sub-image from the LROC image as an example to illustrate the pretreatment process. There are 1356 crater observations in this rectangle produced by 13 observers. Figure 1 shows inconsistency and overlapping among circles representing different observations for single craters.

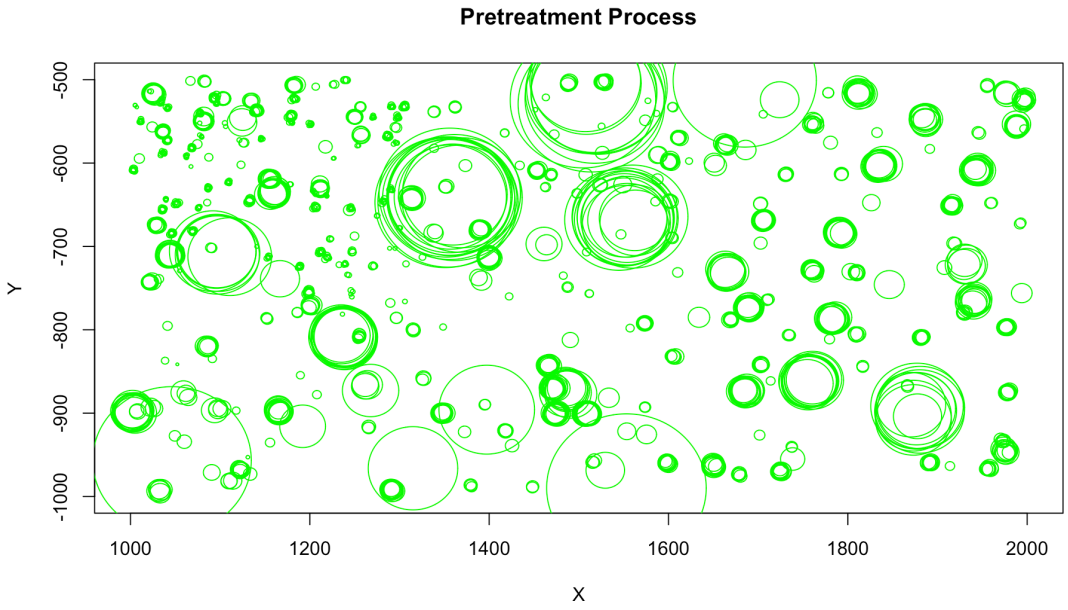

Figure 1: Crater Observations Before Pretreatment

We set α to be 0.25 so the dissimilarity between two crater observations is

$$D = \frac{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + 0.25(d_1 - d_2)^2}}{\min(d_1, d_2)} .$$

We performed a hierarchical cluster analysis using the dissimilarities for the 1356 objects in `R` with the function `hclust()` (using the agglomeration method and `"average"` (UPGMA) criterion). Transforming the clustering structure to the class `"dendrogram"` and plotting consistent with the resulting tree, we got Figure 2. This shows the first 270 objects in the cluster tree. In view of the number of observers, we cut the tree at height= 1, and each branch below the cut has no more than 13 leaves.

The crater circles in the same branch are clustered as observations of one real crater and those locations and diameters were replaced with the average of locations and diameters for circles in that branch. Figure 3 shows in red the crater markings after the pretreatment/data cleaning process.

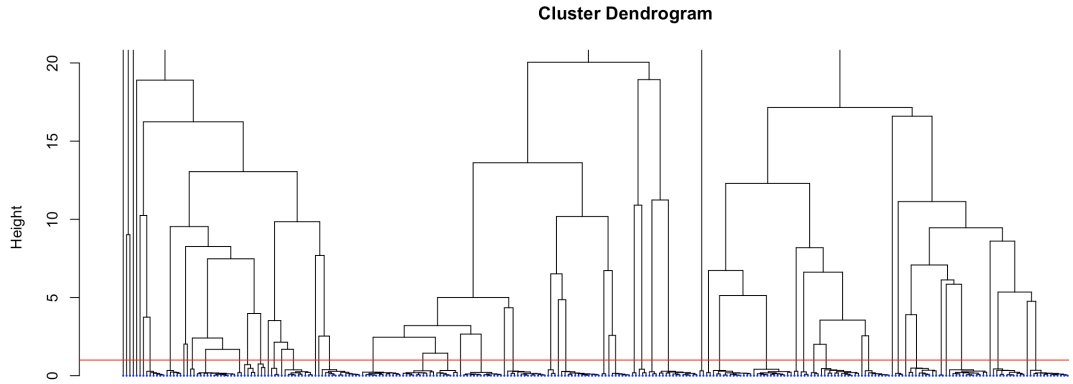

Figure 2: Cluster Tree Cut at Height=1

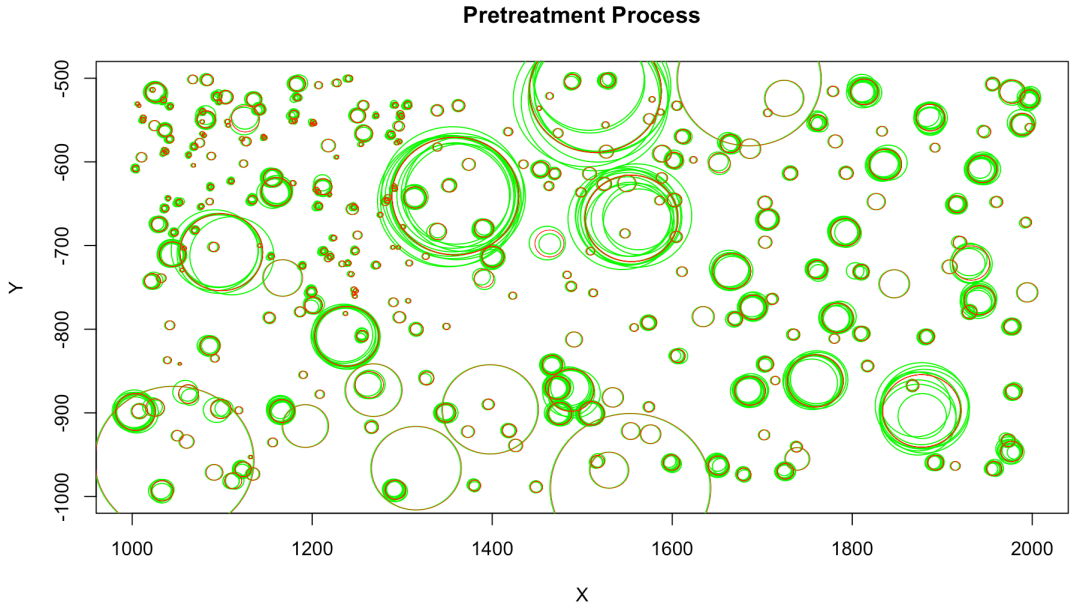

Figure 3: Crater Observations Before Pretreatment and The Observations Used for Analysis (in Red)

3 Modeling and Method

3.1 General Model for Observations and Latent Variables

We assume that observers identify craters with centers in an area A and can not only report real crater locations but also record errant observations that do not represent actual caters. We will call these latter observations "phantoms."

We apply independent Homogeneous Poisson Process (HPP) models for locations of real craters and for phantoms generated by each observer with real crater intensity ρ and phantom intensity ρ_j from observer j for $j = 1, 2, \ldots, J$. The crater measurement data (after pretreatment) will consist of

- N_{2+} craters whose locations and diameters are recorded by 2 or more of J analysts;
- For $j = 1, 2, \ldots, J$, a number S_j of possible craters counted only by analyst j, each of which may be either a real crater (missed by all other analysts) or a phantom non-crater seen only by analyst j. We'll call the number of actual craters in this group N_j and write $F_j = S_j - N_j$ for the number of phantoms seen by analyst j.

We'll further allow that it is possible for analysts to miss craters in their counting and that there are N_0 craters missed by all in the counting. And ultimately the number of real craters in a region of area A is

$$N = N_0 + N_{2+} + \sum_{j=1}^{J} N_j$$

that we will assume to follow a Poisson distribution with mean ρA.

Suppose that analyst j misses a real crater of diameter d in his/her counting with probability $p(d|\boldsymbol{\gamma}_j)$ for a parameter vector $\boldsymbol{\gamma}_j$. (In more complex modeling, this could be a function of other information like local conditions around the crater center or the overall density of real craters.) This could be a simple step function as

$$p(d|\gamma_1, \gamma_2) = I[d < \gamma_1] + \gamma_2 I[d \geqslant \gamma_1]$$

that takes only the value 1 near 0 and some positive value γ_2 to the right of a critical diameter $d = \gamma_1$. Another possible form is

$$p(d|\gamma_1, \gamma_2, \gamma_3) = I[d < \gamma_1] + I[d \geqslant \gamma_1] \left(\gamma_2 + (1 - \gamma_2) \exp\left(-\frac{d - \gamma_1}{\gamma_3} \right) \right) .$$

This function is 1 at and below the threshold $d = \gamma_1$, decreases in exponential fashion to the right of the threshold with a rate parameter γ_3, and has limit γ_2 as $s \to \infty$.

Next, suppose that real craters have diameters that are iid with marginal density $f(d|\boldsymbol{\beta})$ for a parameter vector $\boldsymbol{\beta}$, and phantom craters for analyst j have recorded diameters that are iid with marginal density $h(d|\boldsymbol{\eta}_j)$ for a parameter vector $\boldsymbol{\eta}_j$. A possible form for the real crater diameter distribution is gamma with shape parameter β_1 and rate parameter β_2. A plausible parametric form for the distribution of phantom diameters is up for discussion. It could be as simple as uniform or gamma.

In any event, the basic parameters of the modeling are

$$\boldsymbol{\beta}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_J, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J, \rho, \rho_1, \rho_2, \ldots, \rho_J \, .$$

Various simplified versions of what follows can be had by assuming that some parameter(s) is (are) fixed across j, or are known, etc.

Suppose that for a real crater the recording pattern for all J analysts is a vector of 0's and 1's

$$\boldsymbol{I} = (I_1, I_2, \ldots, I_J) \, ,$$

where $I_j = 1$ indicates that observer j records the crater. Then define the probability that a crater of diameter size d is seen by exactly those analysts j with $I_j = 1$ as

$$q^{\boldsymbol{I}}(d|\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) = \prod_{j=1}^{J} \left(1 - p(d|\boldsymbol{\gamma}_j)\right)^{I_j} \left(p(d|\boldsymbol{\gamma}_j)\right)^{1-I_j} .$$

Then the probability that a crater of diameter size d is seen by none of J analysts (the probability that $\boldsymbol{I} = \mathbf{0}$) is

$$q^{\mathbf{0}}(d|\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) = \prod_{j=1}^{J} p(d|\boldsymbol{\gamma}_j) \, .$$

The probability that a crater of diameter size d is seen only by analyst j is

$$q^{j}(d|\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) = \left(1 - p(d|\boldsymbol{\gamma}_j)\right) \prod_{j' = j} p(d|\boldsymbol{\gamma}'_j) \, .$$

And the probability that a crater of diameter size d is seen by at least 2 analysts is

$$q_{2+}(d|\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) = 1 - q^{\mathbf{0}}(d|\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) - \sum_{j=1}^{J} q^j(d|\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J)\,.$$

Versions of all the above quantities q averaged across diameter sizes of craters can be defined and depend upon both the parameters $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J$ and upon the parameter $\boldsymbol{\beta}$ that describes the diameter distribution. Then we have the probability that a crater is seen by exactly those analysts j with $I_j = 1$ is

$$q^{\boldsymbol{I}}(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) = \int q^{\boldsymbol{I}}(d|\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) f(d|\boldsymbol{\beta})\,\mathrm{d}d\,.$$

And the probability that a crater is seen by none of J analysts is

$$q^{\mathbf{0}}(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) = \int q^{\mathbf{0}}(d|\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) f(d|\boldsymbol{\beta})\,\mathrm{d}d\,.$$

The probability that a crater is seen only by analyst j is

$$q^{j}(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) = \int q^{j}(d|\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) f(d|\boldsymbol{\beta})\,\mathrm{d}d\,.$$

The probability that a crater is seen by at least 2 analysts is

$$q_{2+}(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) = 1 - q^{\mathbf{0}}(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J) - \sum_{j=1}^{J} q^j(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J)\,.$$

Conditional on the parameters $\boldsymbol{\beta}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_J, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J, \rho, \rho_1, \rho_2, \ldots, \rho_J$, we suppose that N_0, N_{2+}, and $S_1, \ldots, S_J$ are independent random variables with

$$N_0 \sim \text{Poisson}\left(q^{\mathbf{0}}(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J)\rho A\right),$$

$$N_{2+} \sim \text{Poisson}\left(q_{2+}(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J)\rho A\right),$$

and

$$S_j \sim \text{Poisson}\left(q^{j}(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J)\rho A + \rho_j A\right) \text{ for } j = 1, 2, \ldots, J.$$

For each $j = 1, 2, \ldots, J$ suppose that there are latent variables $T_1^j, T_2^j, \ldots, T_{S_j}^j$ that are (conditional on all before) independent Bernoulli variables with success probabilities ("success" meaning that a possible crater is real and not a phantom)

$$\frac{q^{j}(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J)\rho}{q^{j}(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J)\rho + \rho_j}\,.$$

Then $N_j = \sum_{l=1}^{S_j} T_l^j$ and $F_j = S_j - N_j$ are independent Poisson variables with mean $q^j(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J)\rho A$ and $\rho_j A$ respectively. So in an area A the number of real craters

$$N = N_0 + N_{2+} + \sum_{j=1}^{J} N_j$$

is Poisson with mean

$$q^{\mathbf{0}}(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J)\rho A + q_{2+}(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J)\rho A + \sum_{j=1}^{J} q^j(\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_J)\rho A = \rho A \, .$$

So we have defined a probability structure for the counts that is consistent with the Poisson process assumptions, the crater size generating assumptions, and the analyst crater-detection assumptions.

3.2 Simplified Modeling and Bayesian Analysis

In this study, we applied a simplified version of the general model for crater data by considering phantom intensities ρ_j's from analyst $j = 1, 2, \ldots, J$ and parameters $\boldsymbol{\gamma}_j$'s in the "missing-probability" and $\boldsymbol{\eta}_j$'s in the phantom size marginal distributions to be consistent across j. We will call these common values $\rho^*, \boldsymbol{\gamma}$ and $\boldsymbol{\eta}$. We further assumed that diameter distributions for real craters and phantoms are gamma. For real craters the size density function used was

$$f(d|\beta_1, \beta_2) = \frac{\beta_2^{\beta_1}}{\Gamma(\beta_1)} d^{\beta_1 - 1} e^{-\beta_2 d}.$$

And for phantoms the size density function used for all analysts was

$$h(d|\eta_1, \eta_2) = \frac{\eta_2^{\eta_1}}{\Gamma(\eta_1)} d^{\eta_1 - 1} e^{-\eta_2 d}.$$

We used a step function for the missing-probability for all analysts, namely

$$p(d|\gamma_1, \gamma_2) = I[d < \gamma_1] + \gamma_2 I[d \geqslant \gamma_1] \, .$$

The parameters in the simplified model are ρ, ρ^*, $\boldsymbol{\beta} = (\beta_1, \beta_2)$, $\boldsymbol{\eta} = (\eta_1, \eta_2)$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$, and we next derive the quantities q for this simplified model. We will abbreviate following $q^{\boldsymbol{I}}(\beta_1, \beta_2, \gamma_1, \gamma_2)$, $q^{\mathbf{0}}(\beta_1, \beta_2, \gamma_1, \gamma_2)$, $q^j(\beta_1, \beta_2, \gamma_1, \gamma_2)$, and

$q_{2+}(\beta_1, \beta_2, \gamma_1, \gamma_2)$ as $q^{\boldsymbol{I}}$, $q^{\mathbf{0}}$, q^j, and q_{2+} respectively. Let n be the number of 1's in observation pattern $\boldsymbol{I}$ for a crater (which means n analysts among J marked that crater). Then using the notation F for cumulative distribution functions,

$$\begin{aligned}
q^{\boldsymbol{I}} &= \int q^{\boldsymbol{I}}(d|\gamma_1, \gamma_2) f(d|\beta_1, \beta_2)\, \mathrm{d}d \\
&= \int (1 - p(d|\gamma_1, \gamma_2))^n \, (p(d|\gamma_1, \gamma_2))^{J-n} \, f(d|\beta_1, \beta_2)\, \mathrm{d}d \\
&= \int_{\gamma_1}^{\infty} (1 - \gamma_2)^n \, (\gamma_2)^{J-n} \, f(d|\beta_1, \beta_2)\, \mathrm{d}d \\
&= (1 - \gamma_2)^n \, (\gamma_2)^{J-n} \, (1 - F(\gamma_1|\beta_1, \beta_2)) \,,
\end{aligned}$$

$$\begin{aligned}
q^{\mathbf{0}} &= \int q^{\mathbf{0}}(d|\gamma_1, \gamma_2) f(d|\beta_1, \beta_2)\, \mathrm{d}d \\
&= \int (p(d|\gamma_1, \gamma_2))^J \, f(d|\beta_1, \beta_2)\, \mathrm{d}d \\
&= \int_0^{\gamma_1} f(d|\beta_1, \beta_2)\, \mathrm{d}d + \int_{\gamma_1}^{\infty} \gamma_2^J f(d|\beta_1, \beta_2)\, \mathrm{d}d \\
&= F(\gamma_1|\beta_1, \beta_2) + \gamma_2^J (1 - F(\gamma_1|\beta_1, \beta_2)) \,,
\end{aligned}$$

$$\begin{aligned}
q^j &= \int q^j(d|\gamma_1, \gamma_2) f(d|\beta_1, \beta_2)\, \mathrm{d}d \\
&= \int (1 - p(d|\gamma_1, \gamma_2)) \, (p(d|\gamma_1, \gamma_2))^{J-1} \, f(d|\beta_1, \beta_2)\, \mathrm{d}d \\
&= \int_{\gamma_1}^{\infty} (1 - \gamma_2) \, (\gamma_2)^{J-1} \, f(d|\beta_1, \beta_2)\, \mathrm{d}d \\
&= (1 - \gamma_2) \, (\gamma_2)^{J-1} \, (1 - F(\gamma_1|\beta_1, \beta_2)) \,,
\end{aligned}$$

and

$$\begin{aligned}
q_{2+} &= \int q_{2+}(d|\gamma_1, \gamma_2) f(d|\beta_1, \beta_2)\, \mathrm{d}d \\
&= 1 - q^{\mathbf{0}} - \sum_{j=1}^{J} q^j \\
&= \left(1 - J \cdot \gamma_2^{J-1} + (J-1)\gamma_2^J\right) (1 - F(\gamma_1|\beta_1, \beta_2)) \,.
\end{aligned}$$

Now consider partitioning the crater measurement data into three subsets according

to observation pattern of each potential crater:

- N_{2+} ($\sim$ Poisson$(q_{2+}\rho A)$) craters seen by at least 2 analysts. For $l = 1, 2, \ldots, N_{2+}$ each potential crater, indexed with l, in this part has a size d_l and detection pattern $\boldsymbol{I}_l$, the latter having $n_l \geqslant 2$ entries of 1. The joint density for size d_l and detection pattern is a function of d_l and n_l

$$\begin{aligned} f(d_l, n_l) &= \frac{q^{\boldsymbol{I}}(d_l|\gamma_1, \gamma_2)}{q_{2+}} f(d_l|\beta_1, \beta_2) \\ &= \frac{(1 - p(d_l|\gamma_1, \gamma_2))^{n_l} (p(d_l|\gamma_1, \gamma_2))^{J-n_l}}{q_{2+}} f(d_l|\beta_1, \beta_2) \\ &= P(n_l|\gamma_1, \gamma_2) \cdot f(d_l|\beta_1, \beta_2) \, . \end{aligned}$$

  So we have independent variables n_l and d_l, where $n_l \sim$ Binomial$(J, 1 - p(d_l|\gamma_1, \gamma_2))$ truncated to the range $[2, J]$ and $d_l \sim$ Gamma(β_1, β_2).

- N_0 ($\sim$ Poisson$(q^{\mathbf{0}}\rho A)$) craters not seen by any analyst. We treat N_0 as a missing or latent value.

- For $j = 1, 2, \ldots, J$, S_j ($\sim$ Poisson$(q^j \rho A + \rho^* A)$) possible craters counted by only analyst j. For $m = 1, 2, \ldots, S_j$ each crater m has a diameter and latent variable (an indicator of the possible crater being real and not a phantom) pair (d_m, T_m). Conditional on it being a real crater,

$$f(d_m|T_m = 1) = \frac{q^j(d_m|\gamma_1, \gamma_2)}{q^j} f(d_m|\beta_1, \beta_2) \, .$$

  Conditional on it being a phantom,

$$f(d_m|T_m = 0) = h(d_m|\eta_1, \eta_2)$$

  and $T_m \sim$ Bernoulli $\left(\dfrac{q^j \rho}{q^j \rho + \rho^*} \right)$. Then the joint density for size d_m and latent variable T_m is

$$\begin{aligned} f(d_m, T_m) &= f(d_m|T_m = 1)P(T_m = 1) + f(d_m|T_m = 0)P(T_m = 0) \\ &= \frac{q^j(d_m|\gamma_1, \gamma_2) f(d_m|\beta_1, \beta_2)\rho + h(d_m|\eta_1, \eta_2)\rho^*}{q^j \rho + \rho^*} \\ &= \frac{(1 - p(d|\gamma_1, \gamma_2)) (p(d|\gamma_1, \gamma_2))^{J-1} f(d_m|\beta_1, \beta_2)\rho + h(d_m|\eta_1, \eta_2)\rho^*}{q^j \rho + \rho^*} \, . \end{aligned}$$

Therefore

$$f(d_m, T_m) = \begin{cases} \dfrac{h(d_m|\eta_1, \eta_2)\rho^*}{q^j\rho + \rho^*} & \text{for } d_m \leqslant \gamma_1 \\ \\ \dfrac{(1-\gamma_2)\gamma_2^{J-1} f(d_m|\beta_1, \beta_2)\rho + h(d_m|\eta_1, \eta_2)\rho^*}{q^j\rho + \rho^*} & \text{for } d_m \geqslant \gamma_1 \end{cases} .$$

Now we have a data model for all crater observations and latent variables that depends on a set of parameters $\boldsymbol{\theta} = (\beta_1, \beta_2, \eta_1, \eta_2, \gamma_1, \gamma_2, \rho, \rho^*)$. What we need next for a Bayes analysis is a prior $\pi(\boldsymbol{\theta})$, a joint distribution for all parameters in our data model that represents our beliefs or *a priori* knowledge concerning values of $\boldsymbol{\theta}$. It could be a complex distribution that depends on some additional unknown parameters, or a product of univariate distributions for each parameter in the data model.

We have no empirical or domain knowledge external to our example dataset to suggest a good choice of a prior distribution, thus in this study we will construct a prior of independence

$$\pi(\boldsymbol{\theta}) = \pi(\beta_1)\pi(\beta_2)\pi(\eta_1)\pi(\eta_2)\pi(\gamma_1)\pi(\gamma_2)\pi(\rho)\pi(\rho^*)$$

for mathematical convenience, trying to make choices of factors that do not too much dictate the form of our empirical conclusions.

4 Bayesian Analysis Examples

We have implemented an MCMC (Markov chain Monte Carlo)-based Bayes analysis based on the modeling of Section 3.2 in the `Rstan` system. In this section we apply it to a simulated dataset and then to the motivating real example.

4.1 Analysis of simulated data

We simulated a crater data set on a region of area $A = 2000 \times 2000$ with the expected number of real craters $\rho A = 2000$, for $J = 10$ analysts. For $j = 1, \ldots, J$ the expected

number of phantoms for each analyst is $\rho^* A = 50$. The values of parameters in the missing-probability function are $\gamma_1 = 3$ and $\gamma_2 = 0.2$, so the step function for the missing-probability for every analyst was

$$p(d|\gamma_1 = 3, \gamma_2 = 0.2) = I[d < 3] + 0.2 \cdot I[d \geqslant 3] \, .$$

And the parameter values in diameter densities were $(\beta_1, \beta_2) = (2.3, 0.1)$ for real craters and $(\eta_1, \eta_2) = (3, 0.1)$ for phantoms, *i.e.* the diameter distributions for real craters and phantoms were Gamma$(2.3, 0.1)$ and Gamma$(3, 0.1)$ respectively.

To create a simulated data set, we first randomly drew a number N from the Poisson$(\rho A = 2000)$ distribution as the number of real craters. Then we randomly generated locations and diameters for N real craters using the `R` functions `runif()` and `rgamma()`.

Second, we simulated observation behavior for the real craters in the region of area A. All craters with diameter smaller than 3 were removed from the observation data according to the missing-probability function $p(d|\gamma_1 = 3, \gamma_2 = 0.2)$ (because these craters were small enough that they will be missed by all analysts). The remaining craters were missed by each analyst with probability 0.2. We implemented this using the `R` function `rbinom()` by randomly generating a set of latent variables to indicate "missing" or "not" for every combination of crater with diameter larger than 3 and analyst. Considering the possible dispersion of observations for a real crater, we multiplied crater diameters and locations by a set of small random normal errors generated by the `R` function `rnorm()` with SD = 0.001 for the diameters and SD = 0.0001 for the locations.

Next we allowed each analyst to observe phantoms with expected number $\rho^* A = 50$. We randomly drew a number F_j from the Poisson$(\rho^* A = 50)$ distribution as the number of phantoms for analyst $j = 1, \ldots, J$, and then used the `R` functions `runif()` and `rgamma()` to generate the locations and diameters for the F_j phantoms from analyst $j = 1, \ldots, J$.

Finally, we obtained a simulated observation data set by combining the data for

craters that are not missed and phantoms for all analysts. The simulated observation data contains 16356 rows, each row denotes a crater record (including location and diameter) identified by a particular analyst.

Before fitting the model, we needed to pretreat the simulated data set as described in Section 2. We set the tuning parameter to 0.25 in the dissimilarity function, and performed a hierarchical cluster analysis using `hclust()` in `R` with the agglomeration method `"single."` We plotted the cluster tree and cut at height= 0.5 with the intent that similar observations for a real crater be grouped into a subset. Then the locations and diameters were replaced with group means in each subset, and we separated the pretreated data into two parts as discussed in Section 3 according to the number of repeated records for each crater. There are 2475 unique craters in the pretreated simulated data, 1969 craters were recorded by more than two analysts and 506 craters were recorded only by a single analyst.

We applied Bayesian inference in `Rstan` based on the `R` program. We declared variables in 3 blocks in `Stan` programs corresponding to their use: the data block, the parameter block and the transformed parameter block. Then we defined the Bayesian model including priors and data model in the model block. Because in the HPP models the area A is a constant, we abbreviate ρA and $\rho^* A$ to ρ and ρ^* respectively in the following discussion.

The data block consists of the variables for the N_{2+} craters seen by multiple observers and the $\sum_j S_j$ craters and phantoms seen by a single observer. The parameters block contains all parameters in the data model $\beta_1, \beta_2, \eta_1, \eta_2, \gamma_1, \gamma_2, \rho, \rho^*$ as well as the latent number of craters missed by all analysts N_0. The transformed parameters block consists of all "q" quantities $q_{2+}, q^j, q^{\mathbf{0}}$. In the model block we set improper priors for ρ and ρ^* of the form $\pi(\rho) = 1$ and $\pi(\rho^*) = 1$ on the non-negative real line. The prior for γ_2 is a beta distribution with support of $(0, 1)$, and the prior for each of $\beta_1, \beta_2, \eta_1, \eta_2, \gamma_1$ is chosen as a gamma distribution with a large variance. We also include a statement

to add a log probability for N_0 at the end the model block. Because N_0 follows a Poisson distribution with mean $q^{\mathbf{0}}\rho$, we add $-q^{\mathbf{0}}\rho + N_0 \cdot \log(q^{\mathbf{0}}\rho) - \log(N_0!)$. We use the gamma function property $\Gamma(N_0 + 1) = N_0!$ and treat N_0 as a continuous gamma variable in the computations for convenience and (presumably) without real loss of precision.

We performed Markov chain Monte Carlo (MCMC) sampling for the parameters, q quantities, and missing value N_0. Four Markov chains were generated with 1000 iterations. Treating the first 500 iterations for each chain as burn-in, the total number of draws employed in analysis was 2000. The program run time was 9 minutes in total on a laptop with 2 processor cores. The results are shown in Table 1.

Table 1: Posterior Summary Statistics in Simulated Data Analysis

Parameter	True value	mean	SE(mean)	SD	95% lower	95% upper	$\hat{R}$
ρ	2000	1998.51	1.24	46.00	1910.50	2090.54	1
ρ^*	50	506.68	0.49	21.95	465.09	549.42	1
γ_1	3.0	2.99	0.03	0.81	1.61	4.75	1
γ_2	0.2	0.19	0	0	0.19	0.20	1
β_1	2.3	2.52	0	0.08	2.37	2.67	1
β_2	0.1	0.10	0	0	0.10	0.11	1
η_1	3.0	3.28	0.01	0.19	2.92	3.68	1
η_2	0.1	0.11	0	0.01	0.10	0.12	1
N_0		28.02	0.57	18.33	4.57	74.31	1
q^j		0	0	0	0	0	1
q_{2+}		0.98	0	0.01	0.96	1	1
$q^{\mathbf{0}}$		0.02	0	0.01	0	0.04	1

The potential scale reduction factor $\hat{R}$ for all variables are less than 1.1, which indicates the chains converged. 95% credible intervals cover the true values for all parameters in $(\beta_2, \eta_1, \eta_2, \gamma_1, \gamma_2, \rho)$. The posterior estimate and 95% credible interval for β_1 (in the size distribution) are above the true value but not much so.

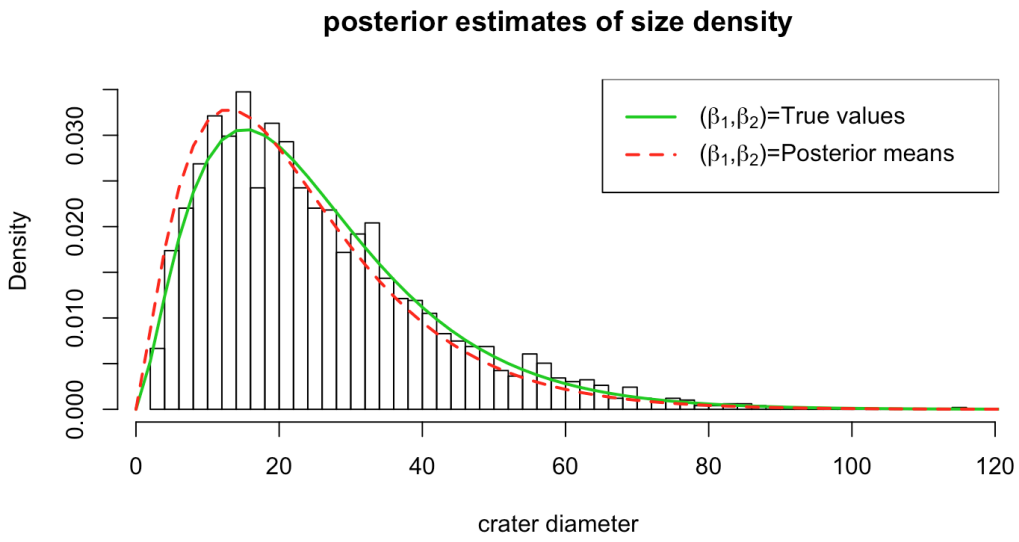

Figure 4: Posterior Analysis of Diameter Density

Figure 4 shows a small shift between the true gamma distribution of size and one with posterior means plugged in as parameters, but the latter may actually be a better fit to the simulated crater sizes. So the inconsistency between the true value and posterior estimate for β_1 may be attributable to sampling errors for crater sizes.

The posterior estimate of ρ^* is much larger than its true value 50, which indicates that for a crater that is only recorded by a single analyst, the probability that it is a real crater $\dfrac{q^j\rho}{q^j\rho + \rho^*}$ is much smaller than it should be. And the posterior estimate for q^j is nearly 0 with a very narrow 95% credible interval. That is, a crater in "single observer" part of the observation data is too likely to be labeled as a phantom.

The 95% credible interval for N_0 is quite wide and uninformative. We have no direct information for the number of missed real craters from the observation data, so this is perhaps to be expected. Figure 5 shows there is a strong positive relationship between N_0 and the critical-size-to-be-seen γ_1, because all craters with a size smaller than γ_1 absolutely will be missed and counted towards N_0.

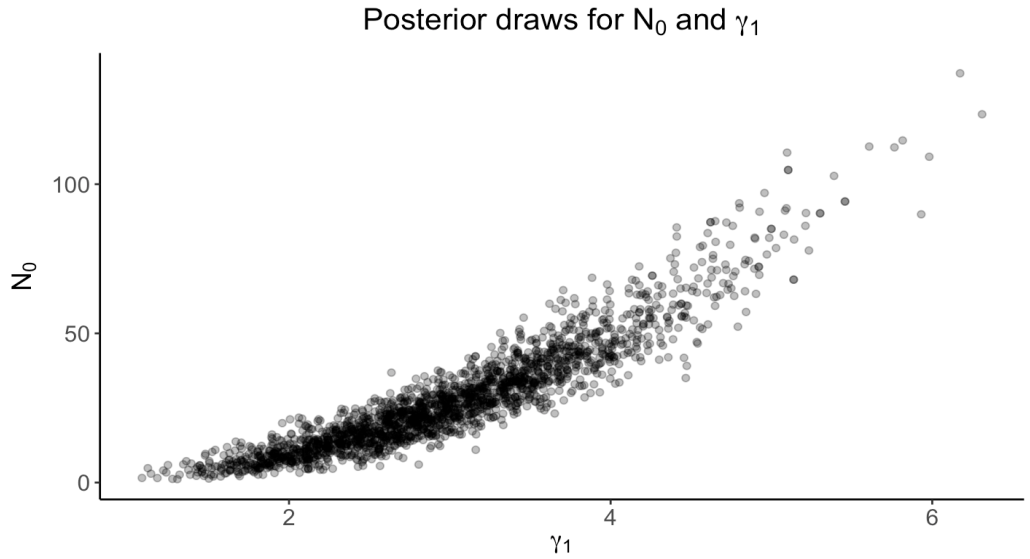

Figure 5: Pairs of N_0 and γ_1 Simulated from the Posterior

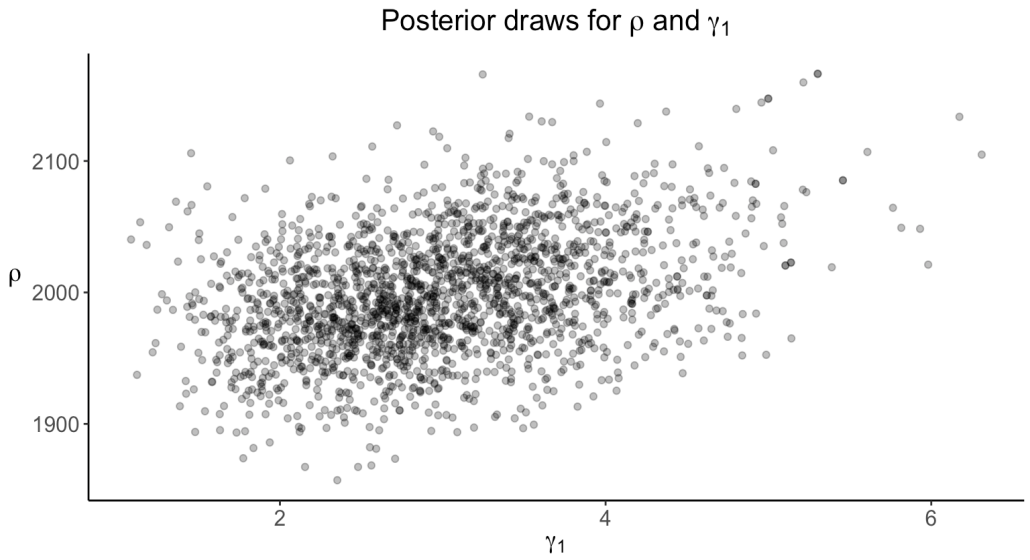

Figure 6: Pairs of ρ and γ_1 Simulated from the Posterior

The 95% credible interval for γ_1 is also wide. Figure 6 shows there is a moderately strong positive relationship between the number of real craters ρ and the critical-size-to-be-seen γ_1. So it is hard to tell from finite data whether there are more real craters with a large γ_1 or there are fewer real craters with a small γ_1, without more definitive prior information about the missing-probability function.

Overall, the `Rstan` results give us clues to understanding the parameters in our model. In this Bayesian analysis of simulated data, the posterior inferences for all

parameters $\beta_1, \beta_2, \eta_1, \eta_2, \gamma_1, \gamma_2, \rho, \rho^*$ are reasonable. From the discussion above we can see that there is a high posterior dependence between the parameters, especially for γ_1, ρ and ρ^*. To produce a more precise estimate of the number of real craters in an area A, we may need more information about the missing-probability and phantom generation pattern for a fixed analyst.

4.2 Analysis of real data

The real crater data set includes two image types "WAC" and "NAC," the "WAC" type covers surface area $\{(x, y)|x \in (0, 1300), y \in (-2800, 0)\}$ and the "NAC" type covers surface area $\{(x, y)|x \in (0, 4100), y \in (-2200, 0)\}$. We treated them separately because they represent two different images captured by different cameras, and in this study we focused on the NAC image.

First we followed the pretreatment procedure described in Section 2 to cluster crater markings in the NAC image. The dissimilarity tuning parameter was set to 0.25 and a hierarchical cluster analysis was performed using `hclust()` in `R` with the `"average"` agglomeration method. We plotted the cluster tree and cut at height= 0.7 so that similar observations were grouped together. Then the locations and diameters for each group were all replaced with group means. The pretreated data were separated into parts for observations made by at least two analysts and those made by single analysts as discussed in Section 3. $N_{2+} = 2138$ craters were seen by at least two observers and $\sum_j S_j = 1112$ craters were seen by only one observer. Figure 7 shows the pretreatment results of NAC image and there are overall 3250 unique location-size pairs (indicated in red) in the NAC image.

Figure 8 below shows the pattern of counts of observers recording ones of the N_{2+} craters observed by at least two analysts in the NAC image. The smoothed blue line indicates a sharp jump in the counts and a rough flat after the jump, suggesting that it is plausible to use a step function to describe the missing-probability.

We performed Markov chain Monte Carlo (MCMC) sampling in `Rstan` using codes similar to those applied to the simulated data in Section 4.1 and adjusted the

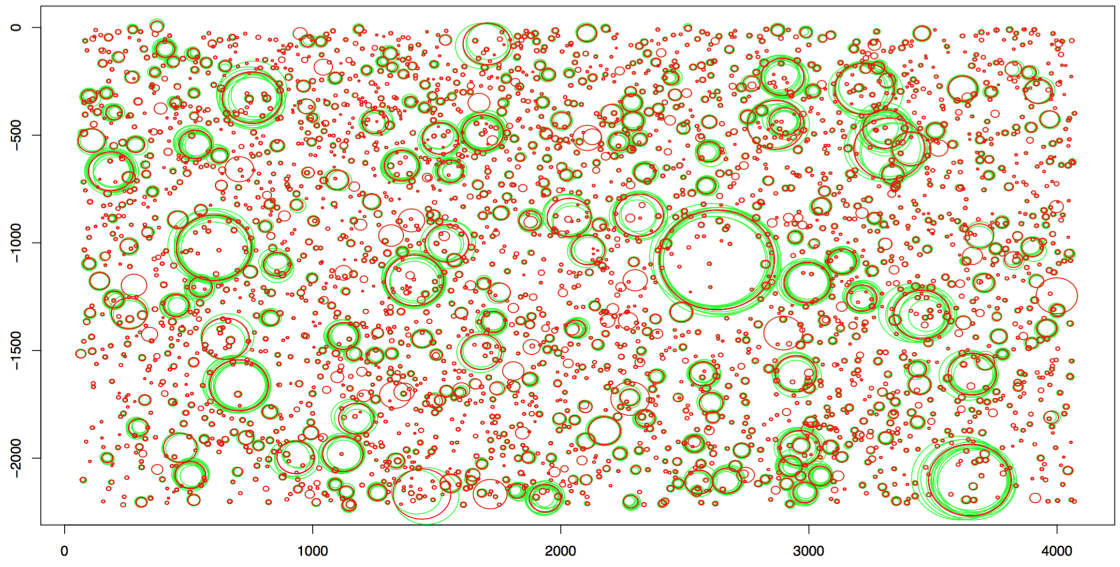

Figure 7: NAC Image Data and Observations Used for Analysis (in Red) After Pretreatment

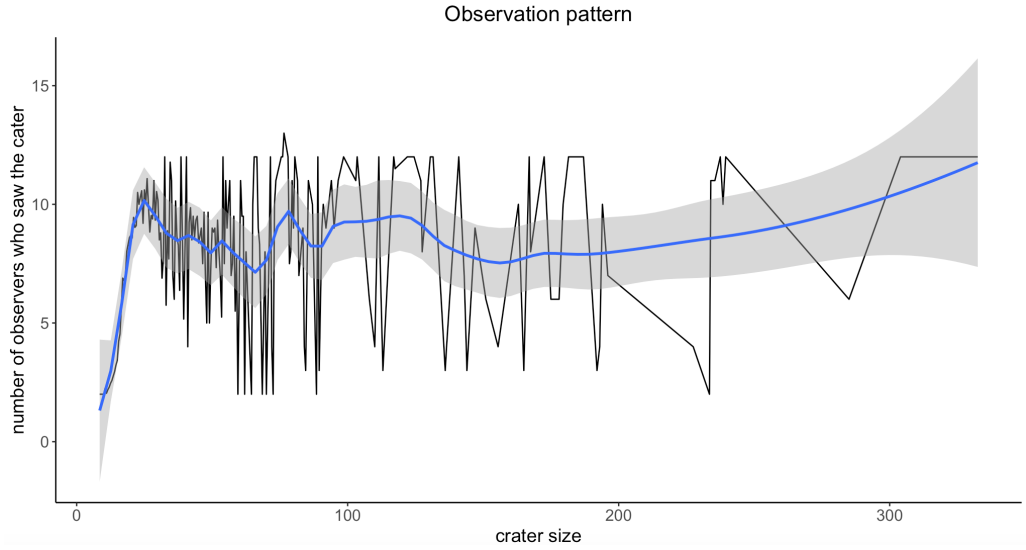

Figure 8: Observation Counts of Craters Recorded Multiple Times in the NAC Image Versus Diameter

parameter values of priors to get good convergence behavior. Four Markov chains were generated through 2000 iterations and the first 1000 iterations were used as burn-in for each chain, giving a total of 4000 simulated parameter vectors. We submitted the job on the `Smaster` cluster server at Iowa State University and the run time of the

`Rstan` program was 25 minutes in total. Summaries of the results are shown in Table 2.

Table 2: Posterior Summary Statistics in Real Data Analysis

Parameter	mean	SE(mean)	SD	95% lower	95% upper	$\hat{R}$
ρ	2177.44	1.79	56.27	2076.31	2299.00	1.00
ρ^*	1106.41	0.89	33.69	1041.58	1177.23	1.00
γ_1	2.90	0.07	1.50	0.63	6.28	1.01
γ_2	0.45	0.00	0.00	0.44	0.45	1.00
β_1	2.39	0.00	0.07	2.26	2.54	1.00
β_2	0.09	0.00	0.00	0.08	0.10	1.00
η_1	3.18	0.01	0.21	2.79	3.61	1.00
η_2	0.18	0.00	0.01	0.16	0.21	1.00
N_0	33.13	1.99	35.54	0.55	134.84	1.01
q^j	0	0	0	0	0	1.00
q_{2+}	0.98	0	0.02	0.94	1	1.01
$q^{\mathbf{0}}$	0.02	0	0.02	0	0.06	1.01

The potential scale reduction factors $\hat{R}$ for all variables are less than 1.1, which indicates the chains converged. The 95% credible intervals for $\beta_1, \beta_2, \eta_1, \eta_2, \gamma_2, \rho, \rho^*$ and q quantities are narrow while the 95% credible intervals for γ_1 and N_0 are relatively wide.

Figure 9 shows there is a strong positive relationship between N_0 and the critical-size-to-be-seen, γ_1. N_0 increases with increasing γ_1 because all craters with a size smaller than γ_1 will be absolutely missed and counted towards N_0.

Figure 10 shows there is a moderately strong positive relationship between the mean number of real craters ρ and the critical-size-to-be-seen γ_1, and it is hard to tell whether there are more real craters with a larger γ_1 or there are fewer real craters with a smaller γ_1 without more information about the missing-probability function.

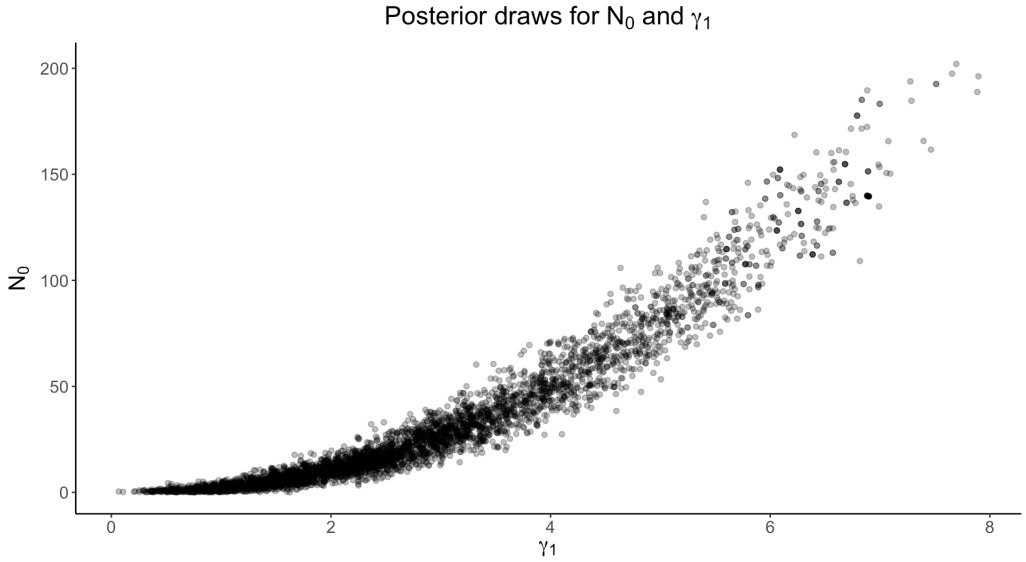

Figure 9: Pairs of N_0 and γ_1 from the Posterior in the Real Data Analysis

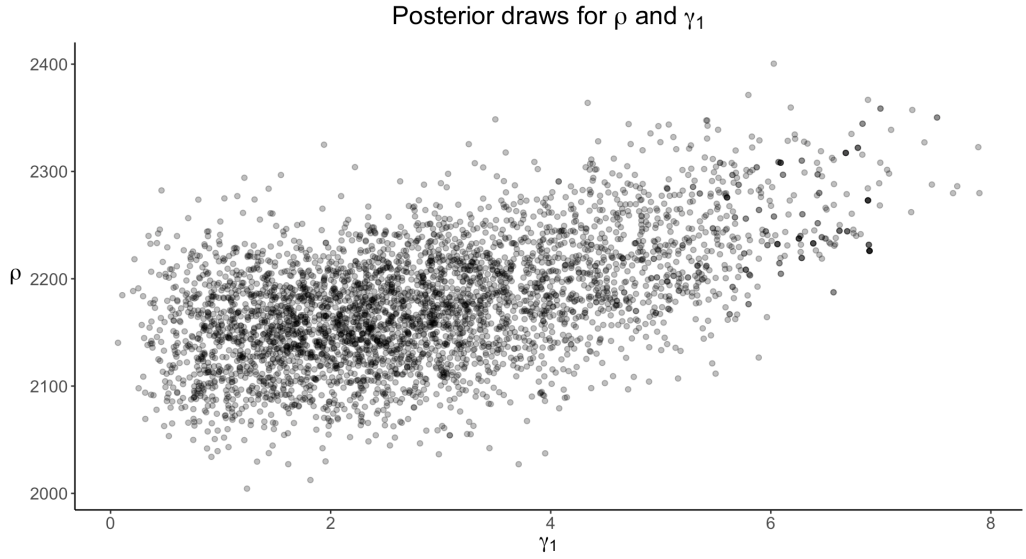

Figure 10: ρ and γ_1 Pairs from the Posterior in the Real Data Analysis

According to the results above, the 95% credible interval for the real crater number in NAC image is $(2076, 2299)$ and the 95% credible interval for the number of phantoms is $(1042, 1177)$.

5 Discussion

In this study, we showed that Bayesian analysis of crater measurement data based on the HPP model is workable. But there are still some issues to be worked out if we want to go further.

The `Stan` program took about 9 minutes on MacBook Pro with 2 processor cores to do a Bayesian analysis of simulated data set. 25 minutes was required for the real data set on the `Smaster` server with 16 processor cores, which is acceptable at this point. But when we need to analysis for larger real datasets or employ more detailed modeling (using more complex missing-probability functions or allowing for observer differences), a super computer may be needed to achieve a shorter run time. When assigning priors in HPP models, we found that the prior parameter values of the crater size that can be seen by analysts (γ_1) is critical in program performance. A prior for γ_1 with large variance will significantly increase the program run time. This indicates that we should obtain subject matter information about the critical-crater-size-to-be-observed (*e.g.* the camera resolution and crater analysis software and hardware used by analysts) and use an informative prior with a small variance for γ_1.

We did a lot of simplification of the general HPP model in the simulated data analysis example by assuming the missing patterns and phantom generation patterns are the same for all analysts. And the missing-probability was treated as a simple step function for mathematical convenience. Actually, a missing-probability with an exponential decreasing shape may be more plausible. However when the missing-probability is a function of diameter size, deriving the q quantities through integration becomes much more difficult to handle in `Rstan` program. How to fit a more complex and flexible model in `Stan` program is an obstacle to be overcome.

The present methodology for analyzing crater measurement data might be applied to other fields. For example it might be applied to flaw detection on a metal casting surface in industry. Of course one will need to adjust parameters in pretreatment process and priors in our modeling to fit each new application.

References

[1] Robbins S J, Antonenko I, Kirchoff M R, et al. *The variability of crater identification among expert and community crater analysts*[J]. Icarus, 2014, 234: 109-131.

[2] Kirchoff M, Sherman K, Chapman C. *Examining lunar impactor population and evolution: Additional results from crater distributions on diverse terrains*[C]. EPSC-DPS Joint Meeting 2011. 2011: 1587.

[3] Kneissl T, van Gasselt S, Neukum G. *Map-projection-independent crater size-frequency determination in GIS environmentsNew software tool for ArcGIS*[J]. Planetary and Space Science, 2011, 59(11): 1243-1254.

6 Appendix

6.1 Codes

Raw data set and `Rstan` codes are available at: `https://github.com/Xiangmei21/Crater-Counting`

6.2 Supplement

The original description of our crater HPP modeling by Dr. Stephen B. Vardeman is available at: `https://github.com/Xiangmei21/Crater-Counting/blob/master/Crater_Counting_Modeling.pdf`