INL/CON-24-78825-Revision-0

Sampling Size Optimization for Bioburden Density Estimation in Planetary Protection

July 2024

Lisa Guan, Andrei Vasilyevich Gribok, Michael DiNicola

Changing the World's Energy Future

Idaho National Laboratory

INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC

INL/CON-24-78825-Revision-0

Sampling Size Optimization for Bioburden Density Estimation in Planetary Protection

Lisa Guan, Andrei Vasilyevich Gribok, Michael DiNicola

July 2024

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Jet Propulsion Laboratory
California Institute of Technology

Sampling Size Optimization for Bioburden Density Estimation in Planetary Protection

Andrei Gribok[a], Mike DiNicola[b] ,and Lisa Guan[b]

[a] Idaho National Laboratory, Idaho Falls, ID, USA
[b] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

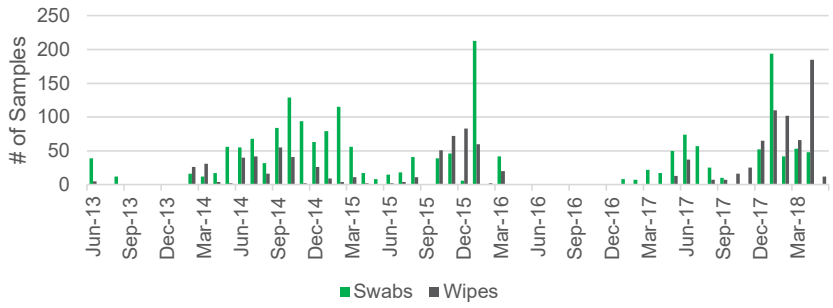
45th Scientific Assembly of the Committee on Space Research (COSPAR), July 13 - July 21, 2024, BEXCO, Busan, Republic of Korea.

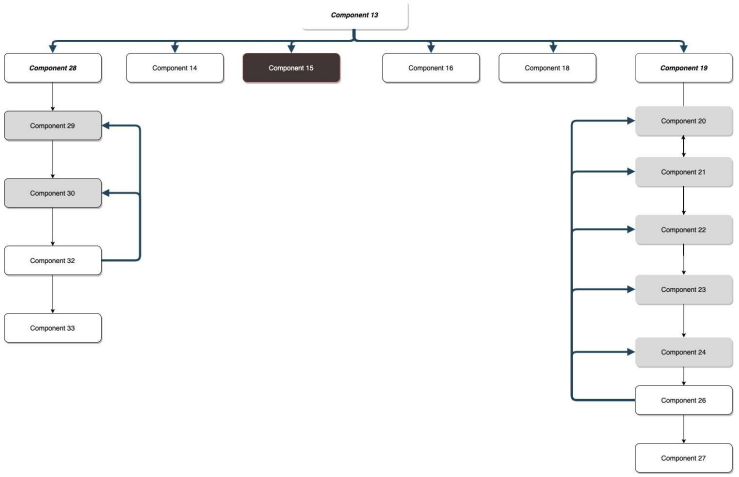
Battelle Energy Alliance manages INL for the U.S. Department of Energy's Office of Nuclear Energy
Idaho National Laboratory

Jet Propulsion Laboratory
California Institute of Technology

Planetary Protection Dataset Overview – Interior Exploration using Seismic Investigations, Geodesy and Heat Transport (InSight) Mission

- ~10% of spacecraft surface areas were verified for biological cleanliness.
  - 2,031 swabs, 1,266 wipes, and 39,379 petri dishes were processed between June 2013 and May 2018.
  - 118,137 data points were recorded as part of the biological cleanliness verification campaign.
- A majority of the samples yielded clean results.
  - 93% of the swabs and 63% of the wipes had 0 colony forming units (CFUs), resulting in ~85% of the 39,379 petri dishes yielding 0 CFUs (Hendrickson et al., Astrobiology 2019).





Tree structure of hardware component 13. The sampled subcomponents are shown as white rectangles, implied components are denoted in grey, and the specified component is shown in black. The names of rollup components 13, 19, and 28 are italicized on the diagram.

IDAHO NATIONAL LABORATORY

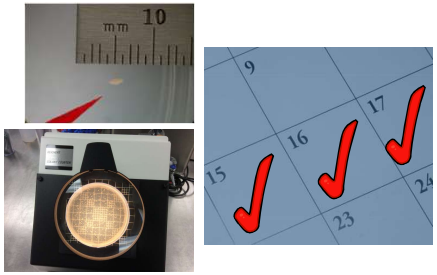
Jet Propulsion Laboratory
California Institute of Technology

Planetary Protection Biological Cleanliness Verification Sampling

Clean hardware (and table or bag) with a solvent wipe prior to sampling or installation

Sample with swab or wipe (water is used as the solvent)

Process swabs and wipes, up to 24 hours required post-assay



Count plates – 24 h, 48 h, and 72 h

- Swabs sampled a 0.0025 m^2 surface area maximum.
- Wipes sampled up to a 1.0 m^2 surface area.
- Due to the experimental procedure, the swabs assume a pour fraction of 0.8 and the wipes 0.25, representing the portion of the total sample solution plated and analyzed for CFU counts.
- Exposure is the sampled area multiplied by the pour fraction (i.e., for swabs it would be E= 0.0025 m^2 ·0.8=0.002 m^2).

Spore/CFU, a heat-tolerant reproductive cell capable of developing into a new individual without fusion with another reproductive cell.

Jet Propulsion Laboratory
California Institute of Technology

Data Generating Model and Parameter Estimators

$$X_i = x|\lambda^i_{true} \sim Poisson(\lambda^i_{true} \cdot E_i)$$

$$\lambda^i_{true}|\alpha,\beta \sim Gamma(\alpha,\beta), i = 1,\dots N$$

$\widehat{\lambda_i}(x_i) = \frac{x_i}{E_i}$ - Maximum Likelihood Estimator (MLE)

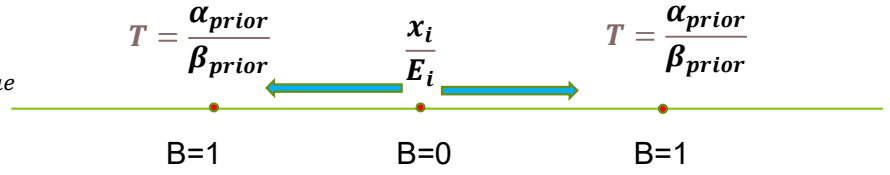
$\widehat{\lambda_i}(x_i) = \frac{x_i+\alpha}{E_i+\beta}$ - Conjugate prior Bayes Estimator

$\widehat{\lambda_i} = d$ - Deterministic Estimator

- Bayesian and Deterministic estimators are shrinkage estimators with respect to MLE
- Shrinkage estimators pulls MLE towards a prespecified target value
- Shrinkage reduces the variance of the estimate at the expense of introducing bias
- Variance reduction can outweigh increase in bias thus improving mean squared error of the estimator

$$\widehat{\lambda_i} = \frac{x_i}{E_i} - B \cdot \left(\frac{x_i}{E_i} - T\right) \quad B = \frac{\beta_{prior}}{E_i + \beta_{prior}} \leq 1$$

X_i – random variable describing the number of CFUs on i-th sampled component
x_i – observed number of CFUs in a sample from i-th sampled component
E_i – exposure for a sample, determined by multiplying the sampled area by pour fraction
λ^i_{true} – true bioburden density for a sample
$Gamma(\lambda/\ \alpha,\beta)$ – Gamma prior distribution of bioburden density λ^i_{true}
α,β – parameters of the Gamma prior distribution
$Poisson(\lambda^i_{true} \cdot E_i)$ – Poisson likelihood of having x CFU counts at exposure E_i given λ^i_{true}
N-number of components



Estimators' Loss Functions and Risks

$$L\left(\widehat{\lambda_i}(x), \lambda^i_{true}\right) = \left(\widehat{\lambda_i}(x) - \lambda^i_{true}\right)^2 - Loss\ function\ for\ a\ data-driven\ estimators,\ (\text{MLE and Bayes})$$

$$L\left(\widehat{\lambda_i}, \lambda^i_{true}\right) = \left(\widehat{\lambda_i} - \lambda^i_{true}\right)^2 - Loss\ function\ for\ a\ deterministic\ estimator$$

$$R_{Frequentist}\left(\lambda^i_{true}, \widehat{\lambda_i}(x)\right) = \sum_x \left(\lambda^i_{true} - \widehat{\lambda_i}(x)\right)^2 \cdot p\left(x; \lambda^i_{true}\right) = \sum_{x=0}^{\infty} \left(\lambda^i_{true} - \widehat{\lambda_i}(x)\right)^2 \cdot \frac{\left(\lambda^i_{true} \cdot E_i\right)^x}{x!} \cdot e^{-\lambda^i_{true} \cdot E_i}$$

$$\rho_{Posterior\ expected\ loss}\left(\alpha, \beta, x_i, \widehat{\lambda_i}(x)\right) = \int_0^{\infty} \left(\lambda^i_{true} - \widehat{\lambda_i}(x_i)\right)^2 \cdot Gamma(\lambda^i_{true} | x_i, \alpha, \beta) d\lambda^i_{true}$$

$$r_{Integrated}\left(\alpha, \beta, \widehat{\lambda_i}\right) = \int_0^{\infty} R\left(\lambda^i_{true}, \widehat{\lambda_i}(x)\right) \cdot Gamma(\lambda^i_{true} | x_i, \alpha, \beta) d\lambda^i_{true} = \sum_{x=0}^{\infty} \rho\left(\alpha, \beta, x_i, \widehat{\lambda_i}(x)\right) \cdot NB\left(x | \alpha, \frac{\beta}{\beta + E_i}\right)$$

Risks for Three Different Estimators

Estimator	R-Frequentist Risk (FR)	ρ-Posterior Expected Loss (PEL)	r- Integrated Risk (IR)
d (deterministic)	$(d-\lambda_{true})^2$	$\frac{\alpha}{\beta^2}+\left(d-\frac{\alpha}{\beta}\right)^2$	$\frac{\alpha}{\beta^2}+\left(d-\frac{\alpha}{\beta}\right)^2$
$\frac{x}{E}$ (MLE)	$\frac{\lambda_{true}}{E}$	$\frac{x+\alpha}{(E+\beta)^2}+\left(\frac{x}{E}-\frac{\alpha}{\beta}\right)^2$	$\frac{\alpha}{E\cdot\beta}$
$\frac{x+\alpha}{E+\beta}$ (Bayes)	$\frac{\lambda_{true}}{E}\cdot(1-B)^2+\left(B\cdot\left(\lambda_{true}-\frac{\alpha}{\beta}\right)\right)^2$ $B=\frac{\beta}{\mathrm{E}+\beta}\leq 1$	$\frac{x+\alpha}{(E+\beta)^2}$	$\frac{\alpha}{\beta\cdot(E+\beta)}$
Properties of the Risks	• The FR is a function of unknown parameter λ_{true} • Functions are harder to compare than numbers • Still useful, for example, $\lambda_{true}<2\cdot\lambda_{true}$	• PEL depends only on known values (data x) • PEL is a number • PEL may not exist for some types of prior	• IR is a number • IR depends neither on data nor λ_{true} • IR can be used for sampling optimization

* From a dimensional analysis perspective, risk and loss functions above have units [CFU/Exposure]^2. This can be seen to be consistent with the above formulas via approrprate transformation; e.g., for the R-Frequentist Risk of the MLE estimator, $\frac{\lambda_{true}}{E}=\frac{\lambda_{true}E}{E^2}$.

Integrated Risk and Sampling Size Determination for Implied Components

$$TR_{Bayes}\,(\alpha,\beta,E) = \underbrace{R(\alpha,\beta,E)_{Bayes}}_{Uncertainty} + \underbrace{C_0 + E\cdot C}_{Cost} = \frac{\mathrm{k}\cdot\alpha}{\beta\cdot(\mathrm{E}+\beta)} + C_0 + C\cdot E$$

For swabs used for the InSight mission, the cost C was \$9.11 for the first swab and \$1.45 for each additional swab
C_0-sampling setup cost
k-scaling factor that dollarizes the r-integrated risk

$$E_{opt} = max\left\{0, \sqrt{\frac{k\cdot\alpha}{C\cdot\beta}} - \beta\right\}$$

Implied Component #	Implied from Component # (implicant)	Implied Bioburden Density, $\hat{\lambda}$, CFUs/m^2	$\sqrt{\text{Implied Risk}}$, CFUs/m^2	Total Surface Area of the Implied Component, m^2	Optimal Sampling Area, m^2	$\sqrt{Optimal\ Risk}$, CFUs/m^2	Optimal Cost, \$
2	10	12.05	17.03	0.26	0.16	7.68	126.80
106	108	13.51	19.11	0.53	0.17	7.91	138.78
133	131	5.10	4.16	0.013	0	4.16	0
36	38	15.50	4.38	2.0	0	4.38	0
71	70	2.47	2.02	7.0	0	2.02	0
29	32	104.16	65.88	0.16	0.16	23.41	129.46

Jet Propulsion Laboratory
California Institute of Technology

Analysis of InSight's Sampled Components

Component	Bioburden density, $\hat{\lambda}$, CFUs/m^2	$\sqrt{Risk}$, CFUs/m^2	Total surface area m^2	Area sampled, m^2	Optimal additional sampling area , m^2	$\sqrt{Optimal\ Risk}$, CFUs/m^2	Optimal Cost, Units
Component 233	31.25	44.19	0.0285	0.02	0.0285	26.49	2.44
Component 236	173.07	81.58	0.5	0.0325	0.17	29.48	13.86
Component 237	93.75	76.54	0.5	0.02	0.14	24.23	11.65
Component 238	125.00	176.77	0.0179	0.005	0.0179	75.54	1.54
Component 242	15.62	22.09	0.298	0.04	0.051	13.71	4.32
Component 243	48.24	20.57	0.298	0.28	0.003	20.29	0.27
Component 245	35.71	50.50	0.282	0.0175	0.098	17.80	8.16
Component 246	62.50	88.38	0.237	0.01	0.130	21.25	10.63
Component 26	27.55	14.73	0.454	0.255	6.61E-05	14.72	0.0057
Component 46	83.33	117.85	0.024	0.007	0.024	52.70	2.05
Component 49	46.87	38.27	0.26	0.04	0.0907	19.54	7.54

Summary

Jet Propulsion Laboratory
California Institute of Technology

- The quality of an estimator is measured by its risks
- The risk can be defined as a mathematical expectation of the difference between the estimate and the true value of the parameter
- Three different risks are commonly used in statistics: frequentist, Bayesian , and integrated
- The integrated risk depends on exposure and parameters of prior distribution
- The integrated risk is a valuable tool to optimize sampling for implied, specified, and sampled components
- The future work will include multi-dimensional risk-cost optimization for multiple components to obtain optimal bioburden estimation strategies

Acknowledgements

This work was supported by U.S. DOE-NASA Strategic Partnership Project (SPP) #19701.

The authors are grateful to Dr. J. Nick Benardini, Dr. Elaine E. Seasly for their contribution to this project.

Jet Propulsion Laboratory
California Institute of Technology

IDAHO NATIONAL LABORATORY