Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P045-cd

Analysis of Shrinkage Estimators and Bayesian Decision Rules for Bioburden Density Estimation in Planetary Protection Probabilistic Risk Assessment

Andrei Gribok

Idaho National Laboratory, USA. E-mail: andrei.gribok@inl.gov

Michael DiNicola

California Institute of Technology, Jet Propulsion Laboratory, USA. E-mail: michael.dinicola@jpl.nasa.gov

Lisa Guan

California Institute of Technology, Jet Propulsion Laboratory, USA. E-mail: lisa.guan@jpl.nasa.gov

The discipline of forward planetary protection aims to minimize microbial contamination on spacecraft in order to prevent the inadvertent contamination of other planetary bodies. Understanding the number of microorganisms, or bioburden, launched with the spacecraft is fundamental to achieving this outcome and is calculated using estimates of the bioburden density (bioburden per unit area or volume) across the spacecraft.
While extremely simple, the deterministic estimators based on NASA-specified and implied bioburden densities may, under certain conditions, have quadratic risk lower than data-driven estimators, with no data estimators being uniformly better (i.e., the estimators are admissible). By comparing risks of deterministic and data-driven estimators, different sampling schedules and volumes can be analyzed to optimize the performance of these estimators. This paper contrasts two approaches used for bioburden calculations—frequentist and Bayesian—and evaluates their performance using data collected from NASA's InSight mission. Specifically, we calculate quadratic risks of different types of shrinkage estimators and compare the risks with the Bayesian approach. An analysis for different regions of the parameter space found estimators with the lowest risks for bioburden values most frequently occurring in practice.

Keywords: Planetary protection, Bayesian inference, InSight mission, loss function, risk, Gamma-Poisson model.

1. Introduction

The primary objective of forward planetary protection is to minimize the inadvertent microbial contamination of other planetary bodies via hitchhiking microbes on robotic spacecraft sent to these planetary bodies. Planetary Protection (PP) engineers thereby constantly monitor, assess, and mitigate the microbiome of spacecraft surfaces and cleanroom assembly environments to ensure the responsible exploration of the solar system. NASA's InSight mission, a lander delivered to the Martian surface in 2018 and retired by NASA in December 2022, explored the interior structure and processes of Mars. This mission had an at-launch bioburden requirement for the entire spacecraft of 1.50×10^5 spores while the cruise stage had a requirement of 5×10^5 spores. The landed spacecraft bioburden had to remain $<3 \times 10^5$ spores while maintaining a bioburden density of <300 spores/m^2. This paper contrasts two approaches to the bioburden density estimation: frequentist and Bayesian, by comparing their corresponding risks. The merits of the risks for evaluating the uncertainty of different estimators are analyzed and compared.

2. Data Collection and Processing

Although previous work has used a suite of molecular techniques to thoroughly characterize and profile the microbiome of various cleanroom environments and spacecraft, the gold standard remains the physical enumeration of microbes via culturing samples from spacecraft and associated surfaces. These samples go through laboratory processing and result in colony forming unit (CFU) counts that are ultimately represented as bioburden density estimates (CFU/m^2). These estimates were tracked in a PP equipment list, and rollup calculations generated current best estimates

(CBE) of bioburden at higher-level nodes (i.e., subsystem and system) to certify compliance. Data samples were collected using either cotton Puritan (Guilford, ME) 806C swabs or Texwipe (Kernersville, NC) TX3211 polyester wipes. Swabs sampled a 0.0025 m^2 surface area maximum while wipes sampled up to a 1.0 m^2 surface area. Due to this 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. Also, in this paper, we assume that the sampling efficiency, i.e., the ability of sampling devices to remove spores from the surface and recover them in culture, is 100%. However, due to technical, budgetary, and programmatic constraints, only a manageable portion of the entire spacecraft surface is directly sampled. To generate the bioburden CBE for components not directly verifiable, we applied a NASA-defined bioburden estimate based on the components' manufacturing or assembly environment (Hendrickson et al., 2020). This approach utilizes a prespecified bioburden density estimation that applies a maximum value across the total surface area of the specified component. For hardware components that underwent similar assembly processes, an implied bioburden is adopted for all components, based on a direct verification of a representative component within the same lot. Once all components have a CBE, the bioburden estimates are generated. In this paper, we consider statistical risks estimates for all three types of components: sampled, specified, and implied.

3. Gamma-Poisson Model

In this paper, we use the Gamma-Poisson compound distribution model to estimate bioburden density and risks associated with the estimate (Gribok et al., 2022). The Gamma-Poisson model assumes a Poisson distribution as a data generating model and a Gamma distribution as a prior distribution for parameter of the Poisson distribution. The Gamma and Poisson distributions are a conjugate pair, allowing for analytical calculations for Bayesian inference. On the other hand, the Gamma distribution is flexible enough to model a variety of prior assumptions about the Poisson parameter's distribution. For the i^{th} component, the model can be represented schematically as:

$$X_i = x|\lambda^i_{true} \sim Poisson(\lambda^i_{true} \cdot E_i)$$
$$\lambda^i_{true}|\alpha,\beta \sim Gamma(\alpha,\beta) \quad (1)$$

for i=1…N where N is the number of components, X_i is the random variable describing CFU counts, x is the actual number of CFUs found on the total exposure area E_i calculated as the area covered with a swab or wipe multiplied by the corresponding pour fraction, and λ^i_{true} is the bioburden density or expected number of CFUs per unit of exposure, which is unknown and the subject of the statistical inference. The unknown λ^i_{true} are drawn from the Gamma distribution with shape parameter α and rate parameter β. The mean and variance of the Gamma distribution under this parametrization can be expressed as $\frac{\alpha}{\beta}$ and $\frac{\alpha}{\beta^2}$, respectively. The Gamma distribution parameters can either be set to reflect the noninformative nature of the prior information about λ^i_{true} or be inferred from previously collected data to implement an empirical Bayes estimation.

If the observed CFU count on the i^{th} component is x_i for a given exposure E_i, λ_i can be estimated as:

$$\widehat{\lambda_\iota} = \frac{x_i}{E_i}, i = 1, \dots N, \quad (2)$$

where N is the number of components. This estimate is the maximum likelihood estimate (MLE) (Atwood et al., 2003) currently used by NASA to evaluate the bioburden density and total CFU counts for biologically sensitive missions (Beaudet, 2013). The MLE allows the bioburden density for each sample to be examined separately, and it has a number of desirable statistical properties. For example, MLE is unbiased in a frequentist sense. However, it also has a number of shortcomings, such as large variance, and most importantly, for a small number of observed CFUs, it can overfit the data. For example, if the number of CFUs registered on an exposure surface is zero, MLE will produce a bioburden density estimate of zero, which is highly unlikely, as achieving absolute cleanliness is practically impossible, considering the presence of humans during spacecraft assembly. These shortcomings motivated the search for other estimators to calculate bioburden density,

such as Bayesian estimators. Bayesian inference using the Gamma-Poisson model will produce an estimator through (Martz et al., 1991):

$$\widehat{\lambda_i^{Bayes}} = \frac{x_i+\alpha}{E_i+\beta}, \quad (3)$$

which is the mean of the posterior Gamma distribution. While a biased estimator in a frequentist's sense, the Bayes estimator has a number of advantages, such as not producing zero bioburden estimates for components with zero CFU counts and also often having a lower error with respect to the true bioburden density values. The lower error is achieved by balancing an estimator bias against its variance.

Due to budgetary and time constraints, on average, only 10% of the spacecraft is sampled. The rest of the spacecraft's components are either specified or implied. Since no data are collected from such components, the specified values are used as-is, and are examples of no-data or deterministic estimators. Additionally, for hardware components that underwent similar assembly processes, an implied bioburden is adopted for all components, based on the direct verification of a representative component within the same lot. Such components are called implied components and their bioburden density is also estimated with a no-data estimate, which is applied from other components. In this paper, we analyze risks associated with data-driven and no-data estimators to evaluate the uncertainty associated with these estimators and for sampling design.

4. Estimator Loss Functions and Risks

In statistics, there are two approaches to determining the uncertainty associated with an estimator: frequentist and Bayesian (Martz et al., 1991). Both approaches start the analysis of estimators' performance by defining a loss function; for example, the measure of deviation of the estimate from the true value of the estimand (i.e., the parameter of interest, in our case, the bioburden density). The most commonly used loss function is the squared error loss (SEL) function, which in its general form can be written as:

$$L_k\left(\widehat{\lambda_i}, \lambda_{true}^i\right) = \frac{\left(\widehat{\lambda_i}-\lambda_{true}^i\right)^2}{(\lambda_{true}^i)^k}, \quad (4)$$

where k is a positive integer and $\widehat{\lambda_i}$ is an estimator. For k=0, Eq. (4) is reduced to the familiar least squares error function, L_0. We use L_0 as our loss function due to its analytical tractability and the fact that it is the currently assumed function for performing bioburden density estimation, which makes the estimators in this paper directly comparable with currently utilized statistical approaches.

Having defined the loss function, the next step is to select estimators that will be used to obtain estimates of bioburden density. For this paper, three estimators are an obvious choice: deterministic estimators as the ones used for implied and specified components; MLE, Eq. (2), as the estimator currently used by NASA; and the Bayes estimator, Eq. (3), as a competing estimator with MLE. Since the loss function in Eq. (4) is a random variable, it cannot be used directly to evaluate estimator performance. In order to derive efficient comparison criteria, the loss function is averaged either over the assumed data distribution (frequentist) or over assumed λ_{true} distribution (Bayesian). Unfortunately, these two approaches are incompatible as they are making fundamentally different assumptions. The frequentist approach postulates that the randomness in the loss function in Eq. (4) originates exclusively from the data, and hence, the estimator $\widehat{\lambda_i}$ is the only source of randomness as it is a function of the data. The unknown parameter λ_{true} is assumed to be a fixed constant. In contrast, the Bayesian approach considers the collected data to be fixed, with λ_{true} is a random variable and averaging is performed over its distribution. The averaged loss is called risk in both paradigms; however, to make a distinction, the loss averaged over the data is called frequentist risk and the risk averaged over the posterior distribution of the parameter is called posterior expected loss (PEL) or posterior risk (Berger, 1995).

Formally, the two risks for the Gamma-Poisson model can be expressed as:

$$R\left(\lambda_{true}^i, \widehat{\lambda_i}\right) = \sum_{x=0}^{\infty}\left(\widehat{\lambda_i(x)} - \lambda_{true}^i\right)^2 \cdot \frac{\left(\lambda_{true}^i \cdot E_i\right)^x}{x!} \cdot e^{-\lambda_{true}^i \cdot E_i}, \quad (5)$$

where R is the frequentist risk. An important feature of the frequentist risk is its dependence on the unknown parameter λ_{true}^{i}, which makes it a function and not a single number. This fact may complicate a comparison and ranking of different estimators as it is more difficult to compare two functions than two numbers. One of the frequentist approaches to deal with this situation is the minimax technique. Using the minimax approach, for the whole range of the parameter, the estimator with lowest worst-case risk is selected (minimum of the maximum risk). The estimator's risk needs to be calculated for the whole range of the parameter since different estimators may have lower risks for different λ_{true}^{i} ranges (Berger, 1995). Moreover, for a given estimator, the minimax value could in the parameter range irrelevant for the problem in hand.

The Bayesian approach to quantifying estimator risk is to average over the posterior distribution of λ_{true}^{i} while regarding the collected data x_i as fixed. This leads to the following formula for the PEL for the Gamma-Poisson model:

$$\rho\left(\alpha,\beta,x_i,\widehat{\lambda_l(x_i)}\right)=\int_0^{\infty}\left(\widehat{\lambda_l(x)}-\lambda_{true}\right)^2\cdot Gamma(\lambda_{true}|x_i,\alpha,\beta)d\lambda_{true}. \quad (6)$$

The PEL is a function of the parameters of the posterior distribution of λ_{true}, collected data, and estimator. Notice the Gamma distribution dependence on collected data x_i, which makes it a posterior distribution. In contrast to the frequentist risk, however, all these values are known quantities. While the PEL is a function of collected data x_i, its value is known and considered fixed. The uncertainty in PEL comes entirely from the posterior distribution of λ_{true}. On the other hand, the uncertainty in frequentist risk comes entirely from the data.

While the two risks are not compatible and produce different results for the same estimator, they can be reconciled if the frequentist risk R is integrated over a prior distribution of λ_{true} or PEL is integrated or summed over all possible realizations of the data, X. In this case, the resulting risk is called r-integrated risk and can be written as:

$$r\left(\alpha,\beta,\widehat{\lambda_l}\right)=\int_0^{\infty}R\left(\lambda_{true},\widehat{\lambda_l}\right)Gamma(\lambda_{true}|x_i,\alpha,\beta)d\lambda_{true}=\sum_{x=0}^{\infty}\rho\left(\alpha,\beta,x_i,\widehat{\lambda_l(x_l)}\right)\cdot NB(\alpha,\tfrac{\beta}{\beta+E_i}). \quad (7),$$

where $NB(\alpha,\frac{\beta}{\beta+E_i})$ is negative binomial distribution. The r-integrated risk is no longer dependent on λ_{true} or on data X. It represents a single number that can be used to rank estimators or design sampling strategies. For the Gamma-Poisson model and L_0 loss function, the three risks for three different estimators considered in this paper are shown in Table 1. The first estimator in Table 1 is d-estimator, the deterministic, no-data estimator used for

Table 1. Risks for three different estimators.

Estimator	R-Frequentist Risk	ρ-Posterior Expected Loss	r- Integrated Risk
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)	$\frac{\lambda_{true}}{E}$	$\frac{x+\alpha}{(E+\beta)^2}+\left(\frac{x}{e}-\frac{x+\alpha}{e+\beta}\right)^2$	$\frac{\alpha}{E\cdot\beta}$
$\frac{x+\alpha}{E+\beta}$ (Bayes)	$\frac{\lambda_{true}}{E}\cdot\left(1-\frac{\beta}{E+\beta}\right)^2+\left(\cdot\frac{\beta}{E+\beta}\left(\lambda_{true}-\frac{\alpha}{\beta}\right)\right)^2$	$\frac{x+\alpha}{(E+\beta)^2}$	$\frac{\alpha}{\beta\cdot(E+\beta)}$

specified and implied components; the second estimator is MLE currently used by NASA for bioburden density estimation; and the third one is the Bayes estimator produced by the Gamma-Poisson model. Since the deterministic estimator does not depend on data, its frequentist risk associated with the L_0 cost function is just a squared difference between the estimator value and true bioburden density value, which is the squared frequentist bias. Due to its indifference to the data, the deterministic estimator has no frequentist variance. Furthermore, its PEL and integrated risk are the same, as the estimator does not depend on data. The PEL and integrated risk is a sum of two terms: the variance of prior distribution Gamma (α,β) and the squared Bayesian bias. The Bayesian bias is different from traditional frequentist bias as it measures the difference between an estimator and the prior's mean. If the deterministic estimator happens to coincide exactly with the prior mean, both PEL and integrated risks are just the variance of the prior distribution.

The MLE in the second row of Table 1 is a data-driven estimator, and its frequentist risk is a function of λ_{true} as well as exposure E. The frequentist risk is directly proportional to λ_{true} and inversely proportional to exposure. Larger λ_{true} values as well as data with a small exposure will produce estimates with larger uncertainty. The largest uncertainty will be for large λ_{true} and small exposure values. The risk also coincides with the variance of the MLE estimator, and it is obvious that the MLE estimator is unbiased in the frequentist sense. The posterior expected MLE loss is similar to the deterministic estimator; however, since the MLE uses collected data, the loss is decomposed into the variance of the posterior distribution and square of the Bayesian bias. The Bayesian bias in this case is the difference between the MLE estimate and posterior mean. In contrast to the frequentist risk, the PEL does not depend on λ_{true} but on collected data x, which is a known value. The PEL should however be considered as a function of the random variable X and hence it is a random variable. The integrated risk is not a random variable, as both random variables have been integrated out. The MLE's integrated risk is a function of prior mean value and exposure. The deterministic and Bayes estimators are shrinkage estimators as they reduce the MLE variance.

The last row in Table 1 is the Gamma-Poisson Bayesian estimator. For an SEL cost function, its frequentist risk can be decomposed into variance and squared bias. The variance is represented by the first term, and the squared bias is represented by the second term. The trade-off between the two is regulated with parameter B, which is between zero and one and is a function of prior parameter β and exposure E. For large exposures, the parameter is converging to zero and the Bayesian estimator is converging to MLE, reflecting the fact that, for a large amount of collected data, the data will outweigh any prior information. Also, if B differs from one, the Bayes estimator variance is guaranteed to be lower than the MLE variance. On the other hand, a very small exposure will set B nearly to one and the frequentist risk will be dominated by the bias term. Notice that, if the mean of the prior $\frac{\alpha}{\beta}$ is equal to λ_{true}, the Bayes estimator is unbiased in a frequentist sense. Parameter β controls the prior distribution variance, and for its large values, parameter B is set to one, meaning that in this case the estimator relies entirely on prior information. On the contrary, for a small β, parameter B is near zero and the estimator relies on data because the prior variance is large, and consequently, the prior information is vague. The PEL of Bayes estimator is just the posterior distribution variance. Finally, the integrated risk for the Bayes estimator depends on the prior distribution and exposure. For zero exposure, the risk becomes the variance of the prior distribution as no data are collected.

5. Risks Behavior of Estimators

The deterministic or no-data estimators are estimators that do not use data from the estimated component but rather rely either on NASA-specified values or the bioburden density estimated using the data collected from similar components. For specified components, the only information provided is the bioburden density value for that component. For the InSight mission, the NASA-specified values range from 10 to 10000 CFUs/m^2. For implied components, the

bioburden information is provided by adopting posterior distributions from similar sampled components. Figure 1 shows frequentist risks for the deterministic estimator and its comparison with frequentist risks of the MLE and Bayes estimator. The frequentist risk of the deterministic estimator does not depend on a component's exposure or the data. The frequentist risks for the MLE and Bayes estimator, on the other hand, depend on exposure. In Fig. 1, the assumed exposure is 0.002 m^2, corresponding to swab exposure, and for the Bayes estimator, the prior is assumed to be constrained noninformative (CNI) (Atwood, 2003), which constrains the mean value of the prior distribution but is otherwise vague. For large mean values, this prior is converging to Jeffreys noninformative prior (Jeffreys, 1946). The risks are plotted as a function of the difference between the estimator and λ_{true} values. Parameter B for the Bayes estimator is set to 0.5.

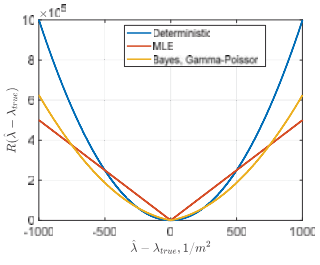

Fig. 1. Frequentist risks of three estimators.

The frequentists risks cannot be calculated in practice due to their dependency on λ_{true}; however, its comparative analysis is still meaningful and insightful. For example, Fig. 1 shows that, for a single component, the deterministic estimator has a smaller risk than MLE as long as the difference between λ_{true} and the estimator is smaller than 500 CFU's/m^2. It also has a smaller risk than the Bayes estimator for differences under 250 CFU/m^2. Also, notice that the Bayes estimator has a smaller risk than MLE up to 750 CFU/m^2. While in practice the differences are not available, if the deterministic estimator is accompanied by a measure of uncertainty, it can be decided whether it is worth using any data-driven estimators instead of the deterministic estimator. For example, if deterministic estimator has value of 400 CFU/m^2 $\pm$ 50 CFU/m^2, it will outperform MLE in terms of frequentist risk but will be dominated by the Bayes estimator as the Bayes estimator has a smaller frequentist risk starting with 250 CFU/m^2. The PEL for the three estimators is shown in Column 3 of Table 1. Since only a single bioburden density value is provided for specified components, its PEL can only be estimated under certain assumptions. The first assumption is that the provided bioburden value is the mean value of the Gamma distribution with some parameters α and β. The second assumption is that, in the absence of any other information, it is reasonable to place CNI on the bioburden density of a specified component. The CNI can be parametrized as Gamma (0.5, 1/(2·μ)), where μ is the specified component's bioburden value. Since, by assumption, the specified estimator is the mean value of the CNI distribution, its PEL is the CNI's variance because the bias term is zero. When reporting risk values, it is convenient to report the risk's square root as, in this case, the risk's units are the same as the units of bioburden density. For example, for specified components with bioburden values of 10, 300, and 1000 CFU/m^2, the corresponding square roots of PEL and integrated risk will be 14.2, 424.3, and 1414.2 CFU/m^2 calculated as standard deviations of the corresponding CNI distributions. The PEL for MLE and Bayes estimators can be calculated for the situation when zero CFUs are collected from the component. Assuming CNI with a mean value fixed at 300 CFU/m^2, the corresponding posterior risks for MLE and Bayes estimators are 236.2 and 192.8 CFU/m^2. Notice that the deterministic estimator risk for the same prior is 424.3 CFU/m^2. In general, the posterior expected losses for the three estimators can be ranked as follow:

$$\rho_{Bayes} \leq \rho_{MLE} \leq \rho_d \tag{8}$$

Summarizing, the deterministic estimator may have the smallest risks of the three estimators provided its value is close to the true value of the parameter. This is achieved by the virtue of zero variance of the deterministic estimator. If its bias is small, the frequentist risk of the deterministic estimator will be small. In case, when the deterministic estimator is exactly equal to the value of the true parameter its frequentist risk is unassailable as it is guaranteed to be at least as good as any other estimator.

6. Integrated Risk and Sampling Size Determination

The integrated risk is sometimes called the pre-posterior risk since it can be considered as risk prior to any data collection. The integrated risk does not depend on λ_{true} or collected data. As can be seen from Table 1, for the deterministic estimator only depends on parameters of the prior distribution and, for data-driven estimators, it also depends on exposure. It is this dependence on exposure that makes it possible to use this risk for sampling size determination (SSD). Since sampling incurs a cost, which quantifies a financial risk, the total risk of determining the bioburden density of a component using the Bayes estimator can be represented as:

$$TR(E) = \frac{\alpha}{\beta \cdot (E+\beta)} + C_0 + C \cdot E \qquad (9)$$

where the first term is the integrated risk of Bayes estimator, C_0 the initial cost of sampling setup, and C the recurring cost per subsequent sample. As can be seen from Eq. (9), the sampling cost is represented as a linear function of exposure. For swabs used for the InSight mission, the cost was \$9.11 for the first swab and \$1.45 for each additional swab. The optimal exposure can be found by differentiating Eq. (9) with respect to E and equating the derivative to zero:

$$E_{opt} = max\left\{0, \sqrt{\frac{\alpha}{c \cdot \beta}} - \beta\right\}. \qquad (10)$$

Equation 10 can be used, for example, to make a decision about the benefits of sampling implied components. Table 2 shows SSD results for five implied components. The first two columns are component numbers: the number of the implied component and its implicant.

Table 2. SSD for implied components using integrated risk.

Implied Component #	Implied from Component # (implicant)	Implied Bioburden Density, $\hat{\lambda}$, CFU/m^2	Implied Risk, CFUs/m^2	Total Surface Area of the Implied Component, m^2	Optimal Sampling Area, m^2	Optimal Risk, CFUs/m^2	Optimal Cost, \$
2	10	12.05	17.03	0.256	0.162	7.68	126.80
106	108	13.51	19.11	0.533	0.178	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	0.166	0.166	23.41	129.46

Since all implicant statistical characteristics are transferred to the implied component, its bioburden density and risk is identical to the implicant. The risk in this case is the PEL of the implicant. The PEL for the implicant is always available since all implicants are sampled components. For the results in Table 2, the PEL was calculated using sampled data and Jeffreys noninformative prior (Jeffreys,1946). The implicant PEL is shown in Column 4 of Table 2. The goal of SSD for implied components is to determine whether their additional sampling can be justified in terms of risk reduction and cost. The optimal sampling area obtained according to Eq. (10) is shown in Column 6, while the optimal risk and cost are in Columns 7 and 8, respectively. For Implied Components 2 and 106, the optimal sampling area is much smaller than the total surface area of the components and a significant risk reduction can be achieved with optimal sampling. For Implied Components 133, 36, and 71, no additional sampling is required as they all have low implied bioburden densities and small risks implied from their corresponding implicants. On the other hand, Implied Component 29 has a high value of implied

bioburden density and implied risk, so according to Eq. (10), its entire surface needs to be sampled to reduce risk significantly.

7. Conclusions

We analyzed the different risks for three different estimators. All risk frameworks have their merits and limitations. The frequentist risk averages over the data distribution and is a function of the unknown parameter. Nevertheless, a relative comparison of different estimators is possible, and valuable insights about their corresponding regions of domination can be obtained. If the deterministic estimator is accompanied by an uncertainty estimate, the frequentist risk can be used to gain insights into the benefits of data-driven estimators for specified or implied components. The posterior expected loss is only conditioned on known values and is most appropriate for sampled components as they have available collected data and assumed prior distribution of λ_{true}. For the Bayes estimator and squared error loss function, the risk is the posterior distribution variance, which is readily available as a result of Bayesian inference. This risk should also be used as an uncertainty estimate for implied components since implied components directly inherit all their values from corresponding sampled components. Finally, the integrated or pre-posterior risk is a very valuable tool for sampling size determination. Since the risk only depends on prior distribution and exposure parameters, it can be optimized to find a trade-off between the estimator's uncertainty and cost. Future work will include a risk optimization for other types of cost functions to account for different emphases on bioburden density over- and underestimation and sampling efficiency.

Acknowledgement

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). This work was supported by U.S. DOE-NASA Strategic Partnership Project (SPP) #19701. This manuscript has been authored by Battelle Energy Alliance, LLC under Contract No. DE-AC07-05ID14517 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes. The authors are grateful to Dr. J. Nick Benardini, Dr. Elaine E. Seasly, and Arman Seuylemezian for their contributions to this project.

References

Atwood, C. L., J. L. LaChance, H. F. Martz, D. J. Anderson, M. Englehardt, D. Whitehead, and T. Wheeler. (2003). Handbook of Parameter Estimation for Probabilistic Risk Assessment. NUREG/CR-6823, SAND2003-3348P, U.S. Nuclear Regulatory Commission.

Beaudet, R. A. (2013). The Statistical Treatment Implemented to Obtain the Planetary Protection Bioburdens for the Mars Science Laboratory Mission. *Advances in Space Research 51*, 2261–2268.

Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis. Springer, New York.

Gribok, A. and A. Seuylemezian. (2022). Performance of Shrinkage Estimators for Bioburden Density Calculations in Planetary Protection Probabilistic Risk Assessment. PSAM 16, June 26–July 1, 2022. Honolulu, Hawaii.

Hendrickson, R., G. Kazarians, and J. N. Benardini. (2020). Planetary Protection Implementation on the Interior Exploration Using Seismic Investigations, Geodesy and Heat Transport Mission. *Astrobiology 20*, 1151–1157.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences 186*, 453–461.

Martz, H. F. and R. A. Waller. (1991). *Bayesian Reliability Analysis*. Reprinted with corrections. Krieger Publishing Co.