Sample size to evaluate ballast water standards: a Bayesian nonparametric approach

Eliardo G. Costa, Carlos Daniel Paulino & Julio M. Singer

July 1, 2019

Abstract

We employ a nonparametric Bayesian approach to compute sample sizes for estimating the organism concentration in ballast water. As a criterion to obtain the sample size we use the total cost minimization, which is the sum of the Bayes risk and a sampling cost function, under a Dirichlet process mixture based on a Poisson model for the concentration of organisms in aliquots of ballast water taken from the ship tank. This semiparametric model provides greater flexibility in modeling the organism distribution and robustness against the misspecification than allowed by parametric models. Credible intervals obtained via the proposed model may be used to verify compliance with international standards.

1 Introduction

The determination of the sample volume to verify the compliance of a ship with the D-2 standard of the International Maritime Organization requires

careful statistical analysis. An important feature of this problem is the inherent heterogeneous nature of the organism concentration in the ballast tank (Murphy *et al.*, 2002). Recently, Costa *et al.* (2015, 2016) proposed methodologies to compute the sample size for evaluate ballast water standards controlling the probabilities of Type I and II errors and the estimation error, respectively, in a frequentist approach using a negative binomial model. On the other hand, Costa *et al.* (2019a,b) also use a negative binomial model but in a Bayesian approach to compute the sample size with criteria controlling summaries of the credible intervals and the Bayes risk. The advantage of the latter approach is that we may incorporate (if available) prior knowledge acquired over the time.

Assume that we collect n aliquots of ballast water and that the i-th aliquot has a λ_i organism concentration with a correspondent number of organisms, which we denote by X_i, $i = 1, \ldots, n$. Then, in the i-th aliquot we expect to find $w\lambda_i$ organisms, *i.e.* $\mathbb{E}\left[X_i \middle| \lambda_i\right] = w\lambda_i$. For $i = 1, \ldots, n$, suppose that, given λ_i, X_i follows a Poisson distribution with mean $w\lambda_i$ and that given a probability measure F, yet unknown wholly or partly, λ_i follows such a distribution. In Costa *et al.* (2019a,b) the authors suppose that F is the probability measure of a gamma distribution, but instead of specifying a known form for F, we may establish a set of possible distributions in which F may vary, which allows greater flexibility in the modeling and robustness against the misspecification. A way to avoid the specification of a parametric form for F is using random probability measures (RPM), which are probability distributions under a space of probability measures (here in $\mathbb{R}_+$), and a example of a RPM is the Dirichlet process. Ferguson

(1973) introduced the Dirichlet process as a possible solution for the problem of prior specification in a nonparametric Bayesian approach, where the prior space is a set of probability distributions under a given sample space. See Phadia (2016), for example.

2 The semiparametric Bayesian model

Suppose that F follows a Dirichlet process with parameters α and F_0, symbolically $F \sim \mathrm{DP}(\alpha, F_0)$. Under this setting we have $\mathbb{E}[F(A)] = F_0(A)$ and $\mathrm{Var}[F(A)] = F_0(A)[1 - F_0(A)]/(\alpha + 1)$, where A is an element of the σ-field of Λ (parameter space of λ_i), F_0 is called base-distribution and α a precision parameter. In our problem we consider F_0 to be a gamma distribution function with mean λ_0 and shape parameter θ_0, both known.

Its properties and the fact that the posterior distribution is easily obtained show why the Dirichlet process is so attractive when analyzed under a Bayesian perspective. Note that in our case the Dirichlet process is the prior assigned to the distribution of the concentrations associated with the conditional Poissonian observations. In this sense, we may write the model hierarchically as follows

$$X_i|\lambda_i \overset{\text{ind}}{\sim} \text{Poisson}(w\lambda_i), \quad i = 1, 2, \ldots, n; \tag{1}$$

$$\lambda_i|F \overset{\text{iid}}{\sim} F, \quad i = 1, 2, \ldots, n; \tag{2}$$

$$F \sim \mathrm{DP}(\alpha, F_0). \tag{3}$$

The parameter of interest is the mean, say $\overline{\lambda}_{\mathrm{R}}$, of the (unknown) real

concentration distribution in the tank, F. If we consider the prior Dirichlet process for F, the corresponding prior (random) mean

$$\overline{\lambda} = \int_{\Lambda} uF(du),$$

is known as a functional of the Dirichlet process. For example, see Cifarelli & Regazzini (1990), Cifarelli & Melilli (2000), James *et al.* (2008), Regazzini *et al.* (2002), among others, for how to obtain the probability distribution function of functionals of the Dirichlet process and its properties. In our case we do not need to know the probability distribution function of the functional, it is sufficient to know how to draw samples of $\overline{\lambda}$ (given an observed sample) and we may use the result in Hjort & Ongaro (2005, Proposition 2) whose obtained a stochastic representation for functionals of the Dirichlet process from the stick-break representation. In our case, for $\overline{\lambda}$ we have

$$\overline{\lambda} =_d B\xi + (1 - B)\overline{\lambda},$$

if $\mathbb{E}\left[\log(1 + |\xi|)\right] < \infty$, where $\xi \sim F_0$ and $B \sim \text{Beta}(1, \alpha)$, a beta distribution with mean $1/(1 + \alpha)$. In the right side of the equation the terms B, ξ and $\overline{\lambda}$ are independents. The notation '$=_d$' means the same distribution. Since in our case F_0 is the gamma distribution function then $\mathbb{E}\left[|\xi|\right]$ is finite and using Jensen's inequality we conclude that $\mathbb{E}\left[\log(1 + |\xi|)\right]$ is finite and we may use the stochastic representation. The simulation strategy for $\overline{\lambda}$ is exploiting a Markov chain of the form

$$\overline{\lambda}_t = B_t\xi_t + (1 - B_t)\overline{\lambda}_{t-1}, \quad t \geq 2.$$

In our case we have that $\overline{\lambda}_t \to \overline{\lambda}$ in distribution as $t \to \infty$, in addition $\overline{\lambda}_t$ is geometrically ergodic (Guglielmi & Tweedie, 2001, Theorem 1). Using this chain and the algorithm in Guglielmi *et al.* (2002) we may draw samples of $\overline{\lambda}$. Given a random sample $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ consider the posterior random mean

$$\overline{\lambda}^{(n)} = \int_{\Lambda} u F^{(n)}(du),$$

with

$$F^{(n)} = (F|\boldsymbol{x}_n) = \int_{\Lambda^n} \mathrm{DP}(\alpha + n, G_n)\nu(d\boldsymbol{\lambda}_n|\boldsymbol{x}_n),$$

where $\boldsymbol{\lambda}_n = (\lambda_1, \ldots, \lambda_n)$ and $G_n = (\alpha F_0 + \sum_{i=1}^n \delta_{\lambda_i})/(\alpha + n)$ with $\delta_{\lambda_i}(A) = 1$, if $\lambda_i \in A$, and $\delta_{\lambda_i}(A) = 0$, otherwise. In addition, we have

$$\nu(d\boldsymbol{\lambda}_n|\boldsymbol{x}_n) \propto \prod_{i=1}^{n} g(x_i|\lambda_i) \left[\alpha F_0(d\lambda_i) + \sum_{j=1}^{i-1} \delta_{\lambda_i}(d\lambda_j)\right],$$

where $g(\cdot|\lambda)$ is the probability function of a Poisson distribution with mean $w\lambda$.

Taking into account the inherent clustering of the λ_i's in the Dirichlet process, we concentrate the conditioning quantities λ_j's on $n^* \leq n-1$ distinct values λ_j^*, with n_j quantities taking this common value. Then, we may use the following full conditional probability (Escobar & West, 1998, Section 1.3.1)

$$\nu(d\lambda_i|\boldsymbol{\lambda}_{(-i)}, \boldsymbol{x}_n) \propto q_0 g(x_i|\lambda_i) F_0(d\lambda_i) + \sum_{j=1}^{n^*} n_j q_j^* \delta_{\lambda_j^*}(d\lambda_i), \tag{4}$$

where $\boldsymbol{\lambda}_{(-i)} = \{\lambda_j | j \neq i, j = 1, \ldots, n\}$ with

$$q_0 \propto \alpha \int_\Lambda g(x_i|\lambda_i) F_0(d\lambda_i) \quad \text{and} \quad q_j^* \propto g(x_i|\lambda_j^*),$$

such that $q_0 + \sum_j n_j q_j^* = 1$. In our problem for q_0 we have the mixture of a Poisson distribution and a gamma distribution, which arises a negative binomial distribution. Hence, we may use a Gibbs sampling to draw samples of $\nu(d\boldsymbol{\lambda}_n|\boldsymbol{x}_n)$. Escobar & West (1998) comment that when we use the above conditional distribution in a Markov chain Monte Carlo algorithm, there may occur problems if the sum of the q_j^*'s becomes very large relative to q_0 on any iteration. In order to prevent this problem it is helpful to "remix" the λ_j^*'s after every step. Conditioning on n^*, consider $s_i = j$ if $\lambda_i = \lambda_j^*$ so that, given $s_i = j$ and λ_j^*, $X_i \sim \text{Poisson}(w\lambda_j^*)$. The cluster structure is defined by the set $\boldsymbol{s} = \{s_1, \ldots, s_n\}$, the $n_j = \#\{s_i = j\}$ observations in cluster j share the common value λ_j^*. Define J_j as the set of indexes of observations in cluster j, *i.e.*, $J_j = \{i | s_i = j\}$. Let $x_{(j)} = \{x_i | s_i = j\}$ be the corresponding cluster of observations. Then, we use the following posterior distribution to "remix" the λ_j^*'s in the Gibbs sampling

$$h(\lambda_j^*|\boldsymbol{x}_n, \boldsymbol{s}, n^*) = h(\lambda_j^*|x_{(j)}, \boldsymbol{s}, n^*) = \prod_{i \in J_j} g(x_i|\lambda_j^*) F_0(d\lambda_j^*),$$

for $j = 1, \ldots, n^*$. In our problem we have

$$h(\lambda_j^*|\boldsymbol{x}_n, \boldsymbol{s}, n^*) \propto (\lambda_j^*)^{\theta_0 - 1 + \sum_{i \in J_j} x_i} e^{-\left(n_j + \frac{\theta_0}{\lambda_0}\right)\lambda_j^*}, \tag{5}$$

which is a gamma distribution. To draw samples of $\nu(d\boldsymbol{\lambda}_n|\boldsymbol{x}_n)$ we use (4) in

a Gibbs sampling and (5) to "remix" the λ_j^*'s.

For the posterior random mean we have the following representation (Hjort & Ongaro, 2005, equation 5.3)

$$\overline{\lambda}^{(n)} =_d B\overline{\lambda} + (1-B)\sum_{i=1}^{n} D_i Z_i,$$

where $B \sim \text{Beta}(\alpha, n)$ is a beta distribution with mean $\alpha/(\alpha+n)$, D_i, $i = 1, \ldots, n$ are the elements of a vector with multivariate uniform distribution and $(Z_1, \ldots, Z_n) \sim \nu(d\boldsymbol{\lambda}_n|\boldsymbol{x}_n)$. Taking all these features into account we are able to simulate samples of $\overline{\lambda}^{(n)}$. Similar strategies and results may be used in the case that the functional of the Dirichlet process is $\int_\Lambda k(u)F(du)$ where $k(u)$ is a measurable function (Guglielmi *et al.*, 2002; Hjort & Ongaro, 2005).

3 Sample size determination

We use the total cost minimization approach used by Costa *et al.* (2019a) to determine the sample size. Under this approach it is necessary to specify a loss function $L(\overline{\lambda}_{\text{R}}, d_n)$ based on a sample $X_1, \ldots, X_n$ and a decision d_n, when the parameter is $\overline{\lambda}_{\text{R}}$. In the problem of interval inference, a decision corresponds to the determination of two quantities, the lower [say, $a = a(\boldsymbol{x}_n)$] and upper [say, $b = b(\boldsymbol{x}_n)$] limits which form a credible interval for the parameter of interest $\overline{\lambda}_{\text{R}}$. For simplicity of notation, we drop the argument $\boldsymbol{x}_n$. In this context, the posterior Bayes risk may be written as

$$r(F^{(n)}, d_n) = \int_{\mathcal{X}^n} \mathbb{E}[L(\overline{\lambda}_{\text{R}}, d_n)|\boldsymbol{x}_n] g(\boldsymbol{x}_n) d\boldsymbol{x}_n. \tag{6}$$

The decision d_n^* which minimizes $r(F^{(n)}, d_n)$ among all the possible decisions d_n is the so-called Bayes rule. Then, the sample size desired is the one which minimizes the total cost defined as

$$\mathrm{TC}(n) = r(F^{(n)}, d_n^*) + cn,$$

where c is the cost of sampling one aliquot. It is not always possible to compute $r(F^{(n)}, d_n^*)$ analytically. We use Monte Carlo simulations to estimate $r(F^{(n)}, d_n^*)$ for a set of n's by drawing samples of $\boldsymbol{x}_n$, computing the expected value in (6) applied in d_n^* and taking the mean of these values. Details are presented in the Supplementary Material. With the estimates of $r(F^{(n)}, d_n^*)$ for each n we fit the following curve which may be linearized and viewed as a linear regression equation (Costa *et al.*, 2019a)

$$\mathrm{TC}(n) = \frac{E}{(1+n)^H} + cn,$$

which leads to the required sample size is the largest integer next to

$$\left(\frac{\widehat{E}\ \widehat{H}}{c}\right)^{1/(\widehat{H}+1)} - 1, \tag{7}$$

where $\widehat{E}$ and $\widehat{H}$ are the estimates obtained by the linear regression fitting (least squares, for example) of E and H, respectively. We use the same loss functions used by Costa *et al.* (2019a) defined as follows.

3.1 Loss function 1

The first loss function we use is

$$L(\overline{\lambda}_{\mathrm{R}}, d_n) = \rho\tau + (a - \overline{\lambda}_{\mathrm{R}})^+ + (\overline{\lambda}_{\mathrm{R}} - b)^+, \tag{8}$$

where $0 < \rho < 1$ is a weight, $\tau = (b-a)/2$ is the radius of the interval, the function x^+ is equals to x if $x > 0$ and equals to zero, otherwise and a decision $d_n = d_n(a, b)$ corresponds to determine the bounds of an credible interval. The correspondent Bayes rule is the quantiles of probabilities $\rho/2$ and $1 - \rho/2$ of the distribution of $\overline{\lambda}^{(n)}$ (Rice *et al.*, 2008). For this loss function we have

$$\mathbb{E}\left[L(\overline{\lambda}^{(n)}, d_n^*)\right] = \mathbb{E}\left[\overline{\lambda}^{(n)}\delta_{\overline{\lambda}^{(n)}}(A_{b^*})\right] - \mathbb{E}\left[\overline{\lambda}^{(n)}\delta_{\overline{\lambda}^{(n)}}(A_{a^*})\right], \tag{9}$$

where $A_{b^*} = [b^*, \infty)$, $A_{a^*} = (0, a^*]$, a^* and b^* are the correspondent bounds of the Bayes rule d_n^*. In Table 1 we present sample sizes computed using the total cost minimization criterion and loss function 1.

3.2 Loss function 2

The second loss function is

$$L(\overline{\lambda}_{\mathrm{R}}, d_n) = \gamma\tau + (\overline{\lambda}_{\mathrm{R}} - m)^2/\tau,$$

where $\gamma > 0$ is a fixed constant and $m = (a+b)/2$ is the center of the credible interval. In this case, the Bayes rule correspond to the quantities which form the interval $[a^*, b^*] = [m - \mathrm{SD}_\gamma, m + \mathrm{SD}_\gamma]$, where $(m, \mathrm{SD}_\gamma) =$

Table 1: Sample size (n) computed with $\rho = 0.05$ and under the Poisson/Dirichlet process (1)-(3) model with F_0 a gamma distribution function with mean $\lambda_0 = 10$ and shape parameter θ_0, and using the loss function 1.

Aliquot volume (w)	Aliquot cost (c)	α	Shape parameter (θ_0) 1.0	2.5	5.0	7.5	10.0
0.5	0.005	0.5	20	16	14	13	10
		1.5	22	18	15	13	12
		2.5	23	17	14	12	11
		5.0	21	15	12	10	9
		10.0	17	12	9	7	6
	0.010	0.5	12	10	9	8	8
		1.5	14	11	9	8	7
		2.5	14	10	8	7	6
		5.0	13	9	7	6	5
		10.0	10	7	5	4	3
1.0	0.005	0.5	19	15	13	12	11
		1.5	22	17	14	12	11
		2.5	22	17	13	12	11
		5.0	20	15	12	10	9
		10.0	17	12	9	7	7
	0.010	0.5	12	10	8	7	7
		1.5	14	10	9	8	7
		2.5	14	10	8	7	6
		5.0	13	9	7	6	5
		10.0	10	7	5	4	4

$\left(\mathbb{E}\left[\overline{\lambda}^{(n)}\right], \gamma^{-1/2}\sqrt{\text{Var}\left[\overline{\lambda}^{(n)}\right]}\right)$. For more details see Rice *et al.* (2008). For this loss function we have

$$\mathbb{E}\left[L(\overline{\lambda}^{(n)}, d_n^*)\right] = 2\gamma^{1/2}\sqrt{\text{Var}\left[\overline{\lambda}^{(n)}\right]}. \tag{10}$$

In Table 2 we present sample sizes computed using the total cost minimization criterion and loss function 2.

4 Discussion

Sample size is directly affected when we vary θ_0 with α fixed, or vary α with θ_0 fixed (Tables 1 and 2), Costa *et al.* (2019a) also observed this behavior with another model. This change in n is more evident in loss function 2. In general, the sample sizes obtained via loss function 1 are much smaller than those obtained via loss function 2 (see Tables 1 and 2), which is also observed in Costa *et al.* (2019a). The authors justify this by the fact that loss function 2 seems to provide more conservative intervals.

We also observed that for a fixed θ_0 the sample size increases with α until some value and then decreases, which is more evident in loss function 2. This may be explained by two facts: (i) as $\alpha \to \infty$ the Dirichlet process tends to concentrate around F_0, which in our problem is a gamma distribution; (ii) Sethuraman & Tiwari (1981) showed that $\text{DP}(\alpha, F_0) \to \delta_\lambda(\lambda')$ in distribution as $\alpha \to 0$, where $\lambda' \sim F_0$. Also note that for θ_0, α and c fixed, the value of the aliquot volume w does not affect much the sample size n, suggesting the choice of smaller w in order to decrease the total volume and the cost of sampling. On the other hand, when the aliquot cost c increases the n decreases, which is more evident in loss function 2.

Table 2: Sample size (n) computed under the Poisson/Dirichlet process (1)-(3) model with F_0 a gamma distribution function with mean $\lambda_0 = 10$ and shape parameter θ_0, and using the loss function 2.

Aliquot volume (w)	Aliquot cost (c)	γ	α	Shape parameter (θ_0)				
				1.0	2.5	5.0	7.5	10.0
0.5	0.005	1	0.5	108	92	83	78	74
			1.5	133	106	92	84	79
			2.5	138	109	92	83	77
			5.0	138	106	85	76	69
			10.0	126	92	72	61	54
		1/4	0.5	69	59	53	49	47
			1.5	83	67	58	52	49
			2.5	86	68	57	50	47
			5.0	84	64	52	45	41
			10.0	75	55	42	35	31
	0.010	1	0.5	69	59	53	49	47
			1.5	84	67	57	52	49
			2.5	86	68	57	51	47
			5.0	85	64	51	45	41
			10.0	75	54	42	36	31
		1/4	0.5	45	37	33	31	30
			1.5	53	42	36	32	30
			2.5	54	42	35	31	28
			5.0	52	39	31	27	24
			10.0	45	32	24	20	18
1.0	0.005	1	0.5	103	85	75	70	66
			1.5	128	100	85	77	72
			2.5	135	104	87	77	72
			5.0	135	102	82	73	66
			10.0	125	91	71	61	55
		1/4	0.5	67	54	48	44	42
			1.5	82	64	54	49	45
			2.5	85	65	54	48	44
			5.0	84	62	50	44	40
			10.0	74	53	42	36	32
	0.010	1	0.5	66	54	47	44	42
			1.5	81	64	53	48	45
			2.5	85	65	53	48	44
			5.0	84	63	50	44	40
			10.0	75	53	41	36	32
		1/4	0.5	43	34	30	28	27
			1.5	52	40	33	30	28
			2.5	53	41	33	30	27
			5.0	52	38	30	26	24
			10.0	44	32	25	21	19

References

Cifarelli, D. M. & Melilli, E. (2000). Some new results for dirichlet priors. *The Annals of Statistics* **28**, 1390–1413.

Cifarelli, D. M. & Regazzini, E. (1990). Distribution functions of means of a Dirichlet process. *The Annals of Statistics* **18**, 429–442. Correction in *The Annals of Statistics* **22**, 1633-1634.

Costa, E. G., Lopes, R. M. & Singer, J. M. (2015). Implications of heterogeneous distributions of organisms on ballast water sampling. *Marine Pollution Bulletin* **91**, 280–287.

Costa, E. G., Lopes, R. M. & Singer, J. M. (2016). Sample size for estimating the mean concentration of organisms in ballast water. *Journal of Environmental Management* **180**, 433–438.

Costa, E. G., Paulino, C. D. & Singer, J. M. (2019a). Sample size determination to evaluate ballast water standards: a decision-theoretic approach. *Submitted* **-**, –.

Costa, E. G., Paulino, C. D. & Singer, J. M. (2019b). Sample size for estimating organism concentration in ballast water: a bayesian approach. *Submitted* **-**, –.

Escobar, M. D. & West, M. (1998). Computing nonparametric hierarchical models. Em *Practical nonparametric and semiparametric Bayesian statistics*. Springer, 1–22.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Guglielmi, A., Holmes, C. C. & Walker, S. G. (2002). Perfect simulation involving functionals of a Dirichlet process. *Journal of Computational and Graphical Statistics* **11**, 306–310.

Guglielmi, A. & Tweedie, R. L. (2001). Markov chain Monte Carlo estimation of the law of the mean of a Dirichlet process. *Bernoulli* **7**, 573–592.

Hjort, N. L. & Ongaro, A. (2005). Exact inference for random Dirichlet means. *Statistical Inference for Stochastic Processes* **8**, 227–254.

James, L. F., Lijoi, A. & Prünster, I. (2008). Distributions of linear functionals of two parameter poisson: Dirichlet random measures. *The Annals of Applied Probability* **18**, 521–551.

Murphy, K. R., Ritz, D. & Hewitt, C. L. (2002). Heterogeneous zooplankton distribution in a ship's ballast tanks. *Journal of Plankton Research* **24**, 729–734.

Phadia, E. G. (2016). *Prior processes and their applications*, 2 ed. Springer.

Regazzini, E., Guglielmi, A. & Nunno, G. D. (2002). Theory and numerical analysis for exact distributions of functionals of a Dirichlet process. *The Annals of Statistics* **30**, 1376–1411.

Rice, K. M., Lumley, T. & Szpiro, A. A. (2008). Trading bias for precision: decision theory for intervals and sets.

http://www.bepress.com/uwbiostat/paper336. Working Paper 336, UW Biostatistics.

Sethuraman, J. & Tiwari, R. C. (1981). Convergence of Dirichlet measures and the interpretation of their parameter. Relat tico M583, Florida State University. DTIC Document.