Bayesian modeling of virus removal efficiency in wastewater treatment processes

T. Ito, T. Kato, K. Takagishi, S. Okabe and D. Sano

ABSTRACT

Left-censored datasets of virus density in wastewater samples make it difficult to evaluate the virus removal efficiency in wastewater treatment processes. In the present study, we modeled the probabilistic distribution of virus removal efficiency in a wastewater treatment process with a Bayesian approach, and investigated how many detect samples in influent and effluent are necessary for accurate estimation. One hundred left-censored data of virus density in wastewater (influent and effluent) were artificially generated based on assumed log-normal distributions and the posterior predictive distribution of virus density, and the log-ratio distribution were estimated. The estimation accuracy of distributions was quantified by Bhattacharyya coefficient. When it is assumed that the accurate estimation of posterior predictive distributions is possible when a 100% positive rate is obtained for 12 pairs of influent and effluent, 11 out of 144, 60 out of 324, and 201 out of 576 combinations of detect samples gave an accurate estimation at the significant level of 0.01 in a Kruskal-Wallis test when the total sample number was 12, 18, and 24, respectively. The combinations with the minimum number of detect samples were (12, 9), (16, 10), and (21, 8) when the total sample number was 12, 18, and 24, respectively.

Key words | Bayesian model, left-censored data, truncated log-normal distribution, virus removal efficiency, wastewater treatment

T. Ito
S. Okabe
D. Sano (corresponding author)
Division of Environmental Engineering, Faculty of Engineering,
Hokkaido University,
North 13, West 8, Kita-ku, Sapporo,
Hokkaido, 060-8628,
Japan
E-mail: *dsano@eng.hokudai.ac.jp*

T. Kato
Department of Computer Science, Graduate School of Engineering,
Gunma University,
Tenjinmachi 1-5-1, Kiryu,
Gunma, 376-8515,
Japan

K. Takagishi
Division of Electronics and Informatics, Faculty of Science and Technology,
Gunma University,
Tenjinmachi 1-5-1, Kiryu,
Gunma, 376-8515,
Japan

INTRODUCTION

Enteric viruses, including human noroviruses, are causing outbreaks of waterborne diseases worldwide (Bosch *et al.* 2008). Since extremely high numbers of these pathogenic viruses are shed in the feces of infected or asymptomatic individuals (Lee *et al.* 2007; García *et al.* 2006), it is necessary to pay great attention to virus removal efficiency in wastewater treatment processes for preventing viral contamination of water environments receiving effluent from wastewater treatment plants. If the virus removal efficiency of each element of the process of wastewater treatment is assessed prior to the operation, it is possible to estimate the total virus removal efficiency in a wastewater treatment process and to calculate infection risks arising from the usage of treated wastewater for various purposes such as irrigation by quantitative microbial risk assessment (QMRA) (Shuval *et al.* 1997; USEPA 2012).

However, the evaluation of virus removal efficiency in wastewater treatment processes can be a challenging task. Although the virus removal efficiency of wastewater treatment processes is usually evaluated as a ratio of virus density in effluent to that in influent of a wastewater treatment process (Ottoson *et al.* 2006), non-detect samples in which the quantity of enteric viruses is below the analytical quantification limit make it difficult to obtain a precise ratio of the virus density before and after a treatment step. In particular, treated wastewater is prone to yielding left-censored datasets, which have high numbers of non-detects (Helsel 2006).

To address this problem of left-censored data, several statistical approaches have been proposed, in which the density of object substances is expressed by probabilistic density functions (PDF) (Kennedy & Hart 2009; Kennedy 2010; EFSA 2010). Paulo *et al.* (2005) proposed a Bayesian approach adapted for the left-censored data of residual pesticide concentrations in food. In our previous study, the Paulo model was applied to the artificially created enteric virus density in wastewater with a slight modification, in which the occurrence of the real zero of virus density is not assumed (Kato *et al.* 2013). Our previous study

doi: 10.2166/wst.2015.402

concluded that eight or more detect samples in a dataset (up to the total sample number of 48) are required to accurately estimate the posterior predictive distributions of virus density and the log-ratio posterior distributions as a virus removal efficiency, when 100% of untreated wastewater samples (influent) give positive results for the presence of enteric viruses (Kato *et al.* 2013). However, the prerequisite condition of a 100% positive rate in influent samples does not coincide with reality, and non-detects also frequently appear in influent samples. It is necessary to confirm what level of accuracy in the estimation of posterior predictive distribution is obtained for a given sample size and number of non-detects.

The aim of this study is to clarify the minimum number of detect samples for the estimation of the probabilistic distribution of virus removal efficiency. One hundred paired datasets of virus density in influent and effluent were generated artificially from log-normal distributions for each combination of a sample size (12, 18, or 24) and the number of detected samples (from 1 to 24) in influent and effluent. The left-censored datasets of influent and effluent were created by setting a limit value of analytical quantification to obtain the assumed number of detect samples. Then, the modified Paulo model was applied to the generated left-censored datasets, in order to estimate posterior predictive distribution of virus density in influent and effluent. A log-ratio posterior distribution, regarded as virus removal efficiency, was also obtained by dividing two posterior predictive distributions of virus density in influent and effluent. Finally, the difference between the true distribution and log-ratio posterior distribution was evaluated by the Bhattacharyya coefficient.

MATERIAL AND METHODS

Estimation of the predictive distribution of enteric virus density and the log-ratio distribution

In this study, a parametric probabilistic model called the truncated log-normal distribution is employed to represent the underlying distribution of enteric virus concentrations. Each of the observations in a given dataset is written by a tuple (x, y), where x is the numerical value of the observed enteric virus concentration and y is a Bernoulli variable indicating the presence of the detect; $y = 1$ if the observation is detected, and $y = 0$, otherwise. The value of x is undefined if $y = 0$. Each tuple has an additional variable θ representing the limit of quantification. In the truncated log-normal model, the concentration x is assumed to be drawn from the following PDF:

$$\mathrm{TLN}(x; \mu, \beta^{-1}, \theta) := \frac{1}{Z(\mu, \beta^{-1}, \theta)x} \exp\left(-\frac{\beta}{2}(\mu - \log_{10}x)^2\right)$$

where $Z(\mu, \beta^{-1}, \theta) := \sqrt{2\pi}\ln(10) \cdot (1 - \varphi(\sqrt{\beta}(\theta - \mu)))\beta^{-1}$ and φ is the cumulative density function of the standard normal distribution. The probability of the Bernoulli variable is given by $(1 - \varphi(\sqrt{\beta}(\theta - \mu))$ for $y = 1$, and by $\varphi(\sqrt{\beta}(\theta - \mu))$ for $y = 0$. This probabilistic model thus contains two model parameters, μ and β, to be inferred.

The likelihood function is required to infer the model parameters. Let us denote a dataset either of influent or of effluent by $X = \{(x_i, y_i)\}_{i=1}^{n}$, where the total number of samples is n in the dataset. Let θ_i be the quantification limit for i-th sample (x_i, y_i). The likelihood function can then be written as $p(X|\mu, \beta) = \prod_{i=1}^{n} (\varphi(\sqrt{\beta}(\theta_i - \mu)))^{1-y_i}$ $((1 - \varphi(\sqrt{\beta}(\theta_i - \mu)))TLN(x_i; \mu, \beta^{-1}, \theta_i)^{y_i}$. To obtain the inference of the model parameters in the form of a posterior distribution, say $p(\mu, \beta|X)$, this study adopted Paulo *et al.*'s prior $\mu \sim N(0, 100)$ and $\beta \sim \mathrm{Gam}(0.01,\ 0.01)$, where $N(m, v)$ and $\mathrm{Gam}(a, b)$ (Paulo *et al.* 2005), respectively, denote the normal distribution with mean m and variance v and the Gamma distribution with shape parameter a and rate parameter b.

The posterior predictive distributions of an unknown concentration can be obtained by applying Bayesian inference to the posterior distribution of the model parameters. Given a dataset X, the probabilistic density at the common logarithm of a concentration $x_{\log}$ is written as $p_{\mathrm{pred}}(x_{\log}|X) = N(x_{\log}; \mu, \beta^{-1})p(\mu, \beta|X)d\mu d\beta$. It would be ideal if $p_{\mathrm{pred}}(x_{\log}|X)$ was close to $N(x_{\log}; \mu_*, \beta_*^{-1})$ where (μ_*, β_*^{-1}) is the true value of the model parameter. When both a dataset of influent X_{inf} and a dataset of effluent X_{eff} are given, the probabilistic distribution of the log-ratio between two respective concentrations can be computed from the corresponding posterior predictive distributions. The probabilistic distribution of the log ratio is simply referred to as the log-ratio distribution, and denoted by $p(r|X_{\mathrm{inf}},\ X_{\mathrm{eff}})$.

The software implementing the algorithm developed for inferring posterior predictive distribution of virus removal efficiency is available upon request to the corresponding author.

Bhattacharyya coefficient

Provided that the two datasets, X_{inf} and X_{eff}, are artificially generated, the true distribution of the log ratio is known, allowing us to assess the accuracy of the inference by comparing the inferred distribution with the true distribution. In this study, the Bhattacharyya coefficient (Bhattacharyya 1943; Derpanis 2008) is employed for assessment of inferred distributions. Denoting by $p_{\mathrm{true}}(r)$ the true distribution, the Bhattacharyya coefficient is defined by $\mathrm{BC} = \int p(r|X_{\mathrm{inf}},\ X_{\mathrm{eff}})\, p_{\mathrm{true}}(r)\mathrm{dr}$. The coefficient takes a value between zero and one. Better inference gets a higher Bhattacharyya coefficient; $\mathrm{BC} = 1$ implies the exact inference.

Statistical treatment

We assumed that the accurate estimation of posterior predictive distributions is possible when a 100% positive rate is obtained for 12 sample pairs of influent and effluent. This assumption can be translated so that the estimation is regarded to be accurate when Bhattacharyya coefficient values of a certain combination of detect samples are not statistically different from those at a 100% positive rate for 12 pairs of influent and effluent. For this purpose, the Kruskal-Wallis test at a significant level of 0.01 was performed under the null hypothesis that the median value of the Bhattacharyya coefficient is identical with that at a 100% positive rate for 12 pairs of influent and effluent. Before performing the Kruskal-Wallis test, outliers in the Bhattacharyya coefficient values were detected by using an interquartile range (IQR) between the first (25%tile) and third (75%tile) quartiles, in which any Bhattacharyya coefficient values at a greater distance from first or third quartiles than 1.5 times IQR were regarded as outliers. The H statistic in Kruskal-Wallis test is defined by $H = \dfrac{12}{N(N+1)}\sum_{i=1}^{C}\dfrac{R_i^2}{n_i} - 3(N+1)$, where N is the total number of data and C is the number of detected sample combination. When the median values of three combinations of detect samples (ex., (12, 12), (12, 9), and (9, 12)) that have 100 Bhattacharyya coefficient each were compared in a Kruskal-Wallis test, the number of detected sample combination a $C = 3$ and the total number of data $N = 300$. The mean of ranks in each group of the Bhattacharyya coefficient values is found by dividing the sum of ranked scores $\left(\sum^{Ri}\right)$ by n_i observations in the i-th group. In the matter of the correction factor for ties, it should be denoted by $H_0 = \dfrac{H}{1 - \sum (t_i^3 - t_i)/(N^3 - N)}$, where t_i is the number of tied value in the i-th rank. The H_0 statistic (chi-squared value) has been shown to be distributed approximately as a chi-squared distribution ($df = \mathrm{C}$-1).

RESULTS AND DISCUSSION

Accuracy of the estimated distribution of virus removal efficiency when the total sample number is 12 each for influent and effluent

When the total sample number is 12 each for influent and effluent, there are 144 combinations of detect samples from (1, 1) to (12, 12). One hundred pairs of virus density data in influent and effluent were generated for each combination of detect samples, based on two assumed (true) log-normal distributions ((μ, σ^2) = (4, 1) and (1, 1)), respectively. Then, 100 each of posterior predictive distributions of virus density in influent and effluent were estimated, and 100 log-ratio posterior distributions for each combination of detect samples were subsequently obtained. These 100 log-ratio posterior distributions were compared with the log-ratio distribution derived from the true log-normal distributions which were used to generate simulated datasets for evaluating the estimation accuracy, by using the Bhattacharyya coefficient. One Bhattacharyya coefficient value was obtained for one inferred log-ratio posterior distribution, which means that 100 values of Bhattacharyya coefficients were calculated for each combination of detected sample.

Percentiles (5, 25, 50, or 75%tile) of Bhattacharyya coefficient values when the total sample number is 12 each for influent and effluent are indicated in Figure 1. The x-axis and y-axis are the number of detect samples of influent and effluent, respectively, and the z-axis is the value of Bhattacharyya coefficient ranging from 0 to 1. The 25 and 75%-tile values of the Bhattacharyya coefficient were used for the rejection of outliers using IQR. Intuitively, the Bhattacharyya coefficient values are increased with the number of detect samples, although how large a Bhattacharyya coefficient value is enough for assuring estimation accuracy is not clear. This is why we investigated the relative accuracy of the estimation by comparing with Bhattacharyya coefficient values at a 100% positive rate for 12 pairs of influent and effluent.

Bhattacharyya coefficient values without outliers are shown in Figure 2(a). The result of the Kruskal-Wallis test indicated that 11 out of 144 combinations of the number of detect samples gave a comparative estimation accuracy to that at the 100% positive rate for 12 pairs of influent and effluent (Figure 2(b)). The minimum number of detect

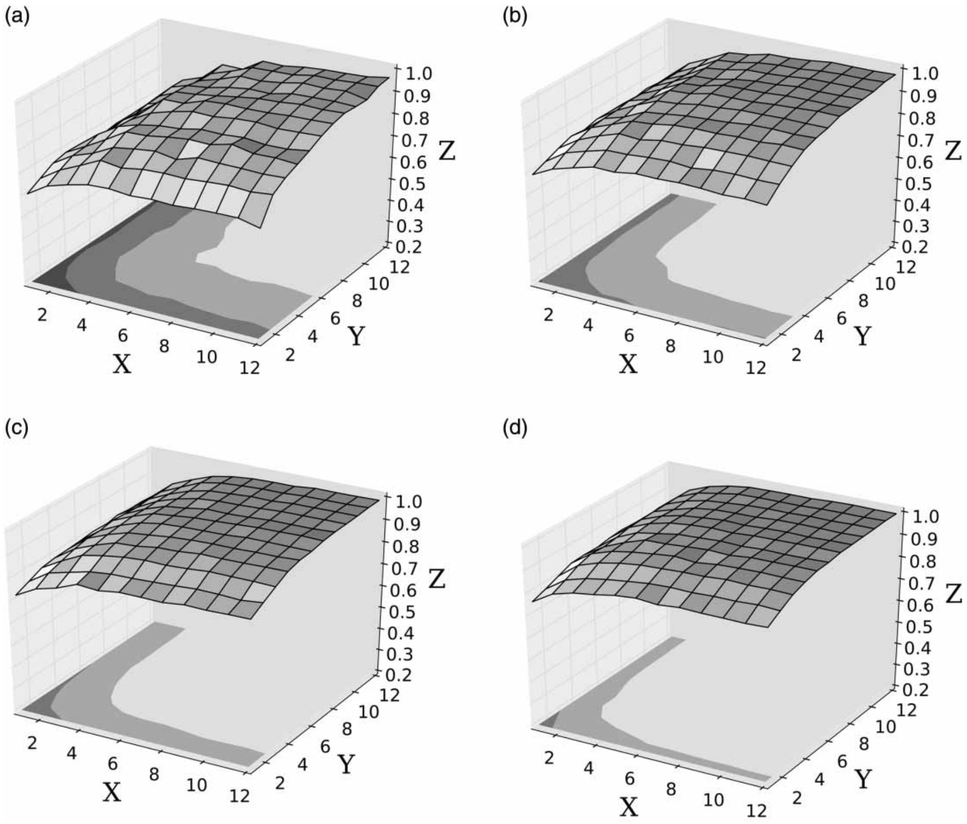
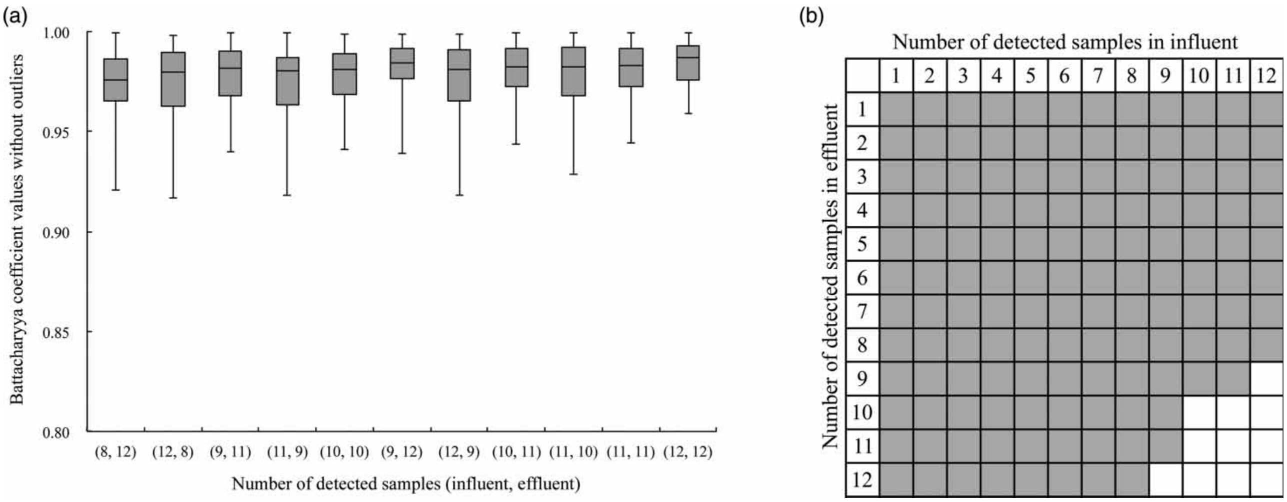


Figure 1 | Percentile values of Bhattacharyya coefficient between the true distribution of virus removal efficiency and estimated log-ratio posterior distribution when the total sample number is 12 each for influent and effluent: (a) 5%tile, (b) 25%tile, (c) 50%tile, and (d) 75%tile. The *x*-axis and *y*-axis are the number of detect samples in influent and effluent, respectively. The *z*-axis is the value of the Bhattacharyya coefficient (ranging from 0 to 1). Level lines on the bottom panel are indicating Bhattacharyya coefficient values of 0.9, 0.8 and 0.7.



	1	2	3	4	5	6	7	8	9	10	11	12
1	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
2	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
3	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
4	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
5	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
6	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
7	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
8	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
9	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	□
10	▨	▨	▨	▨	▨	▨	▨	▨	▨	□	□	□
11	▨	▨	▨	▨	▨	▨	▨	▨	▨	□	□	□
12	▨	▨	▨	▨	▨	▨	▨	▨	□	□	□	□

Figure 2 | Estimation accuracy of log-ratio distribution depending on combinations of detect samples in influent and effluent. (a) Box plot of Bhattacharyya coefficient values without outliers when the total sample number is 12 each for influent and effluent. Maximum, minimum and quartile values are indicated. (b) White cells indicate that the combination of detect samples gives accurate estimation at the significant level of 0.01 in the Kruskal-Wallis test, while grey cells do not.

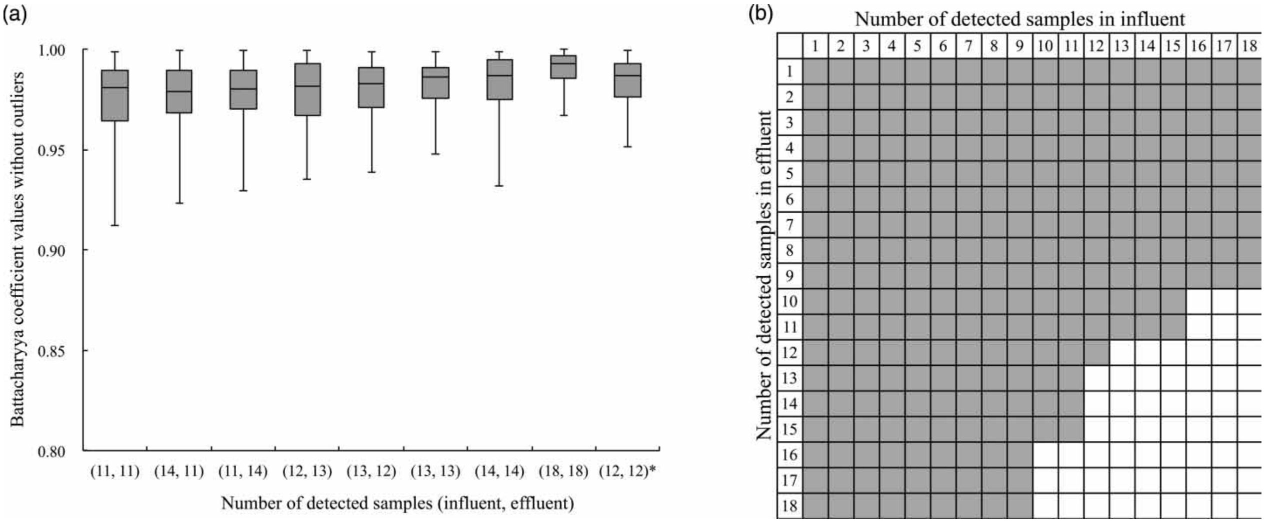
samples of nine was acceptable, when 12 of the coupled samples were all positive in the virus quantification. In our previous study, eight or more detect samples in treated wastewater (effluent) were required for accurate estimation, when the positive rate in influent was 100% and the total sample number was 12 each for influent and effluent (Kato *et al.* 2013), while the present study indicated that more than 11 detect samples under the same condition (100% of influent samples) are positive. This difference is attributable to the employment of different parameters for evaluating estimation accuracy. Kullback-Leibier divergence was used in the previous study, and the Bhattacharyya coefficient was used in the present study. Although these two parameters are able to give us information on estimation accuracy, we had better employ the larger value of nine detect samples out of a total sample number of 12 from a conservative viewpoint.

Accuracy of the estimated distribution of virus removal efficiency when the total sample number is 18 each for influent and effluent

When the total sample number is 18 each for influent and effluent, there are 324 combinations of detect samples from (1, 1) to (18, 18). One hundred paired datasets of virus density in influent and effluent were generated for each combination of detect samples as well as the total sample number, which was 12 each. One hundred Bhattacharyya coefficient values between the true distribution and the log-ratio posterior distribution were computed, and outliers are excluded as well. Bhattacharyya coefficient values without outliers are shown in Figure 3(a). As a result, 60 out of 324 combinations of the number of detect samples gave comparative estimation accuracy to that at a 100% positive rate for 12 pairs of influent and effluent (Figure 3(b)). The minimum number of detect samples was 10, when the companion number of detect samples is larger than 15 (Figure 3(b)).

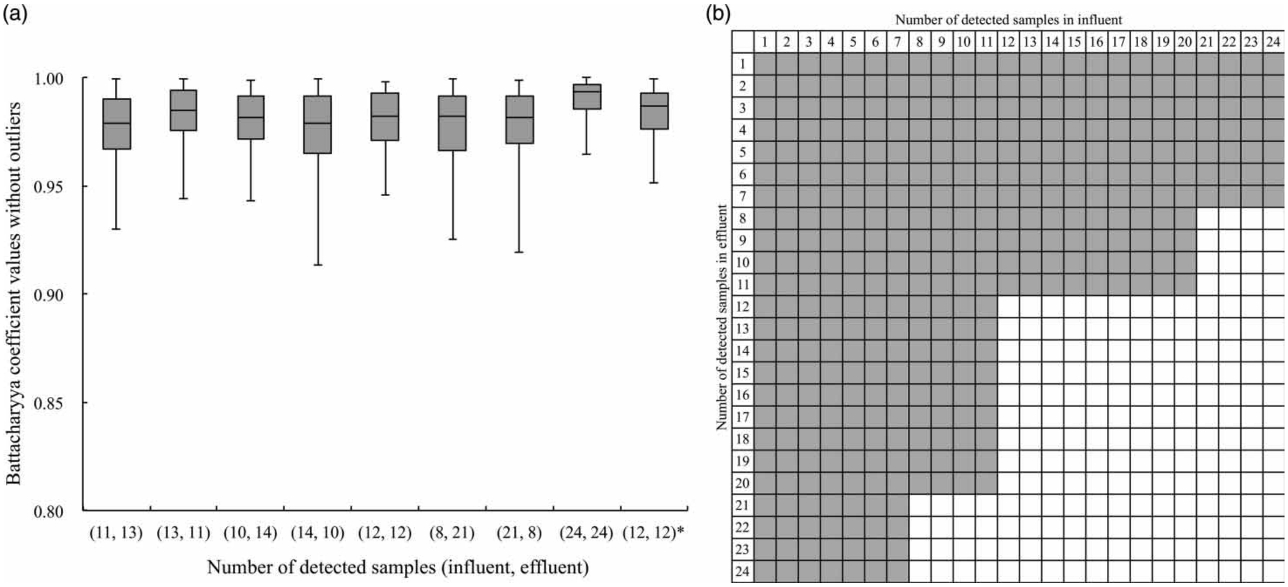
Accuracy of the estimated distribution of virus removal efficiency when the total sample number is 24 each for influent and effluent

When the total sample number is 24 each for influent and effluent, there are 576 combinations of detect samples from (1, 1) to (24, 24). One hundred paired datasets of virus density in influent and effluent were generated for each combination of detect samples as well as the total sample number, which was 12 or 18 each. One hundred Bhattacharyya coefficient values between the true distribution and the log-ratio posterior distribution were computed, and outliers are excluded as well. Bhattacharyya coefficient values without outliers are shown in Figure 4(a). As a result, 201 out of 576 combinations of the number of detect samples gave a comparative estimation accuracy to that at a 100% positive rate for 12 pairs of influent and effluent (Figure 4(b)). The minimum number of detect samples was eight, when the companion number of detect samples was larger than 20 (Figure 4(b)).



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
2	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
3	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
4	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
5	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
6	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
7	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
8	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
9	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
10	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨			
11	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨			
12	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨						
13	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨							
14	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨							
15	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨							
16	▨	▨	▨	▨	▨	▨	▨	▨	▨									
17	▨	▨	▨	▨	▨	▨	▨	▨	▨									
18	▨	▨	▨	▨	▨	▨	▨	▨	▨									

Figure 3 | Estimation accuracy of log-ratio distribution depending on combinations of detect samples in influent and effluent. (a) Box plot of Bhattacharyya coefficient values without outliers when the total sample number is 18 each for influent and effluent. Maximum, minimum and quartile values are indicated. The asterisk indicated the Bhattacharyya coefficient values without outliers at a 100% positive rate for 12 pairs of influent and effluent. (b) The white cells indicate that the combination of detect samples gives an accurate estimation at the significant level of 0.01 in the Kruskal-Wallis test, while the grey cells do not.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
2	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
3	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
4	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
5	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
6	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
7	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
8	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
9	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
10	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
11	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
12	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨													
13	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨													
14	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨													
15	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨													
16	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨													
17	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨													
18	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨													
19	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨													
20	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨													
21	▨	▨	▨	▨	▨	▨	▨																	
22	▨	▨	▨	▨	▨	▨	▨																	
23	▨	▨	▨	▨	▨	▨	▨																	
24	▨	▨	▨	▨	▨	▨	▨																	

Figure 4 | Estimation accuracy of log-ratio distribution depending on combinations of detect samples in influent and effluent. (a) Box plot of Bhattacharyya coefficient values without outliers when the total sample number is 24 each for influent and effluent. Maximum, minimum and quartile values are indicated. The asterisk indicated the Bhattacharyya coefficient values without outliers at a 100% positive rate for 12 pairs of influent and effluent. (b) The white cells indicate that the combination of detect samples gives an accurate estimation at the significant level of 0.01 in the Kruskal-Wallis test, while the grey cells do not.

These results indicate that estimation accuracy depends on the number of detect samples, but the dependency was not clear for the positive rate of virus quantification. In the case of other occasions, such as the total sample number larger than 24 and unequal sample numbers between influent and effluent, it is better to estimate estimation accuracy using artificially generated data as performed in this study. Furthermore, we assumed a log-normal distribution of enteric virus density in wastewater samples, but the other distributions such as gamma distribution should be also tested in further studies.

The virus removal mechanism is complex, and its efficiency depends on types of wastewater treatment and species of enteric viruses (Sano 2006; Miura *et al.* 2015). The difference in virus removal tendency is reflected by different estimation results in the proposed approach, and theoretically, the number of positive samples needed for the estimation is not affected by the particularities of wastewater treatment and enteric virus species. The methodology for virus quantification also does not affect estimation results, but algorithm users have to use identical methodology for both influent and effluent, although it is not necessary for quantification limit values to be identical between influent and effluent samples.

The algorithm developed in this study for inferring posterior predictive distribution of virus removal efficiency has distinguished advantages compared to already published ones in terms of the treatment of left-censored data, but its availability should be compared with other approaches in further studies.

CONCLUSIONS

The estimation accuracy of log-ratio distributions as the probabilistic distributions of virus removal efficiency in wastewater treatment processes was dependent on the number of detect samples, rather than the positive rate of virus quantification. When it is assumed that the accurate estimation of posterior predictive distributions is possible when a 100% positive rate is obtained for 12 pairs of influent and effluent, 11 out of 144, 60 out of 324, and 201 out of 576 combinations of detect samples gave the accurate estimation when the total sample number was 12, 18, and 24, respectively. The combinations with the minimum number of detect samples were (12, 9), (16, 10), and (21, 8) when the total sample number was 12, 18, and 24, respectively.

REFERENCES

Bhattacharyya, A. 1943 On a measure of divergence between two statistical populations defined by their probability distribution. *Bulletin of the Calcutta Mathematical Society* **35**, 99–110.

Bosch, A., Guix, S., Sano, D. & Pinto, R. 2008 New tools for the study and direct surveillance of viral pathogens in water. *Current Opinion in Biotechnology* **19**, 295–301.

Derpanis, K. G. 2008 The Bhattacharyya measure. *Mendeley Computer* **1** (4), 1990–1992.

European Food Safety Authority 2010 Management of left-censored data in dietary exposure assessment of chemical substances. *EFSA Journal* **8** (3), 1557.

García, C., DuPont, H. L., Long, K. Z., Santos, J. I. & Ko, G. 2006 Asymptomatic norovirus infection in Mexican children. *Journal of Clinical Microbiology* **44** (8), 2997–3000.

Helsel, D. R. 2006 Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere* **65** (11), 2434–2439.

Kato, T., Miura, T., Okabe, S. & Sano, D. 2013 Bayesian modeling of enteric virus density in wastewater using left-censored data. *Food and Environmental Virology* **5** (4), 185–193.

Kennedy, M. C. 2010 Bayesian modeling of long-term dietary intakes from multiple sources. *Food and Chemical Toxicology* **48**, 250–263.

Kennedy, M. & Hart, A. 2009 Bayesian modeling of measurement errors and pesticide concentration in dietary risk assessments. *Risk Analysis* **29** (10), 1427–1442.

Lee, N., Chan, M. C. W., Wong, B., Choi, K. W., Sin, W., Lui, G., Chan, P. K. S., Lai, R. W. M., Cockram, C. S., Sung, J. J. Y. & Leung, W. K. 2007 Fecal viral concentration and diarrhea in norovirus gastroenteritis. *Emerging Infectious Diseases* **13** (9), 1399–1401.

Miura, T., Okabe, S., Nakahara, Y. & Sano, D. 2015 Removal properties of human enteric viruses in a pilot-scale membrane bioreactor (MBR) process. *Water Research* **76**, 33–42.

Ottoson, J., Hansen, A., Bjorlenius, B., Norder, H. & Stenstrom, T. A. 2006 Removal of viruses, parasitic protozoa and microbial indicators in conventional and membrane processes in a wastewater pilot plant. *Water Research* **40**, 1449–1457.

Paulo, M. J., van der Voet, H., Jansen, M. J. W., ter Braak, C. J. F. & van Klaveren, J. D. 2005 Risk assessment of dietary exposure to pesticides using a Bayesian method. *Pest Management Science* **61** (8), 759–766.

Sano, D., Ueki, Y., Watanabe, T. & Omura, T. 2006 Membrane separation of indigenous noroviruses from sewage sludge and treated wastewater. *Water Science and Technology* **54** (3), 77–82.

Shuval, H., Lampert, Y. & Fattal, B. 1997 Development of a risk assessment approach for evaluating wastewater reuse standards for agriculture. *Water Science and Technology* **35** (11), 15–20.

USEPA 2012 *Guidelines for water reuse. USEPA and US Agency for International Development*, Washington, DC.

First received 24 December 2014; accepted in revised form 20 July 2015. Available online 1 August 2015