Sign-constrained linear regression for prediction of microbe concentration based on water quality datasets

Tsuyoshi Kato, Ayano Kobayashi, Wakana Oishi, Syun-suke Kadoya, Satoshi Okabe, Naoya Ohta, Mohan Amarasiri and Daisuke Sano

ABSTRACT

This study presents a novel methodology for estimating the concentration of environmental pollutants in water, such as pathogens, based on environmental parameters. The scientific uniqueness of this study is the prevention of excess conformity in the model fitting by applying domain knowledge, which is the accumulated scientific knowledge regarding the correlations between response and explanatory variables. Sign constraints were used to express domain knowledge, and the effect of the sign constraints on the prediction performance using censored datasets was investigated. As a result, we confirmed that sign constraints made prediction more accurate compared to conventional sign-free approaches. The most remarkable technical contribution of this study is the finding that the sign constraints can be incorporated in the estimation of the correlation coefficient in Tobit analysis. We developed effective and numerically stable algorithms for fitting a model to datasets under the sign constraints. This novel algorithm is applicable to a wide variety of the prediction of pollutant contamination level, including the pathogen concentrations in water.

Key words | censored datasets, environmental regression, sign constraints, Tobit analysis, water quality

Tsuyoshi Kato
Naoya Ohta
Division of Electronics and Informatics,
Faculty of Science and Technology,
Gunma University,
Tenjin-cho 1-5-1, Kiryu, Gunma 376-8515, Japan,
and
Center for Research on Adoption of NextGen
Transportation Systems (CRANTS),
Gunma University,
Aramaki-machi 4-2, Maebashi, Gunma, 371-8510,
Japan

Tsuyoshi Kato
Integrated Institute for Regulatory Science,
Waseda University,
Tsurumaki-cho 513, Shinjuku-ku, Tokyo 162-0041,
Japan

Ayano Kobayashi
Wakana Oishi
Satoshi Okabe IWA
Division of Environmental Engineering, Faculty of
Engineering,
Hokkaido University,
North 13, West 8, Kita-ku, Sapporo,
Hokkaido 060-8628, Japan

Syun-suke Kadoya
Mohan Amarasiri IWA
Daisuke Sano IWA (corresponding author)
Department of Civil and Environmental
Engineering, Graduate School of Engineering,
Tohoku University,
Aoba 6-6-06, Aramaki, Aoba-ku, Sendai,
Miyagi 980-8579, Japan
E-mail: *daisuke.sano.e1@tohoku.ac.jp*

Daisuke Sano
Department of Frontier Science for Advanced
Environment, Graduate School of Environmental
Studies,
Tohoku University,
Aoba 6-6-06, Aramaki, Aoba-ku, Sendai, Miyagi
980-8579, Japan

This article has been made Open Access thanks to the generous support of a global network of libraries as part of the Knowledge Unlatched Select initiative.

doi: 10.2166/wh.2019.001

INTRODUCTION

Water safety plans (WSP) and sanitation safety plans (SSP) are international planning schemes for water and wastewater developments (Goodwin *et al.* 2015). Hazard analysis and critical control point (HACCP) is a basic concept of these plans, in which some operational parameters are identified as critical control points (CCPs), and critical limit values are monitored for ensuring safety in water usage (Garner *et al.* 2016). A possible option for monitoring microbial risks in water usage is the use of on-line sensors for pathogens. Sensitive sensors for pathogens in water have been proposed (Kitajima *et al.* 2016), but they are still too expensive to use in daily operation, and the quantification limit is still not low enough for applying to pathogens in environmental water.

This study focuses on an alternative approach for monitoring microbial risks, which is the employment of other water quality parameters that have been proved to have a statistically significant relationship with the concentration of pathogens (Cho *et al.* 2010; Jurzik *et al.* 2010). To exploit other water quality parameters, the prediction model must be fitted in advance to a training sample by regression analysis. An obstacle in the model fitting stage is that pathogen concentration in water is frequently found to be below the quantification limit of an analytical method (Kato *et al.* 2016). Such data are referred to as non-detects in this paper.

To cope with this censoring problem in regression analysis, several approaches have been attempted. One approach is substitution of non-detects (data under a quantification limit) with zero or a value of quantification limit (Antweiler 2015). However, value substitution is not always appropriate in a statistical treatment of a dataset because it can distort the statistical estimation of parameters (Helsel 2006, 2010; Huynh *et al.* 2014, 2016). Another approach is Tobit analysis (Amemiya 1984) that employs a probabilistic model. This approach suffers from overfitting to samples for training when the sample sizes are small.

If abundant training data were available, combining more multiple independent variables could boost the prediction performance of regression models (Harwood *et al.* 2005). When the correlations between pathogen concentration and explanatory variables are weaker, more explanatory variables are needed to obtain high prediction accuracy. However, too many independent variables would result in excess conformity if the training sample size were small. The scientific uniqueness of this study is to alleviate the excess conformity by applying domain knowledge, which is the accumulated scientific knowledge regarding the correlations between pathogen concentration and the other variables. The present study investigated whether the performance of regression analyses using censored datasets is increased by the application of domain knowledge in the form of *sign constraints*. The sign constraints are to fix signs (non-negative or non-positive) of the regression coefficients in regression analysis, when the signs have been shown to be statistically significant in previous studies. For example, we expect a positive correlation between indicator microorganisms (such as coliforms) and pathogenic bacteria (such as enterohemorrhagic *Escherichia coli*) in water, but the sign of the sample correlation may be reversed when the sample size is too small.

The main purpose of this study was to evaluate the effect of the sign constraints on the performance of regression analyses using censored datasets. In our simulation, water quality data acquired from a watershed were used as explanatory variables, and the common logarithmic value of *E. coli* concentration was used as an alternative response variable to the pathogen concentration. The reason why *E. coli* data were used for this simulation is that *E. coli* can be measured with little censoring that could be easily used to generate datasets with various given quantification limits, which reflects different level of censoring. Six left-censored datasets with different qualification limit values were prepared for investigating the power of the sign constraints.

METHODS

Water quality parameters

River water sampling was conducted about twice a month from January 2012 to April 2013 in four rivers: Toyohira River (Site A), Nopporo River (Site B), Atsubetsu River (Site C), and Motsukisamu River (Site D) in Sapporo City, Japan (Figure SD.1 in the Supplementary materials, available with the online version of this paper). *E. coli*

concentration, water temperature, pH, electrical conductivity, dissolved oxygen, suspended solids, biological oxygen demand, total nitrogen, and total phosphorus in water samples were measured according to *Standard Methods for the Examination of Water and Wastewater* (APHA 2005). The flow rate of the river water was also measured at each sampling event. We acquired 96 observations of these water quality parameters (Table SD.1 in the Supplementary materials, available online).

In this study, the common logarithmic value of *E. coli* concentration was used as an alternative for the pathogen concentration. Six left-censored datasets were generated by manually setting different quantification limit values (1.5, 2.0, 2.5, 3.0, 3.5, and 4.0-log MPN/100 mL), in which the *E. coli* concentration values less than the quantification limit were regarded as non-detects. These left-censored datasets were used as response variables in the following regression analysis.

The explanatory variables include water temperature, pH, electric conductivity, suspended solids, dissolved oxygen, biological oxygen demand, total nitrogen, total phosphorus and flow rate in the regression analysis. One of these variables, pH, was divided into the following two separate explanatory variables:

$$\mathrm{pH}_{+} := \max(0, \mathrm{pH} - 7.0) \tag{1}$$

$$\mathrm{pH}_{-} := \max(0, 7.0 - \mathrm{pH}) \tag{2}$$

Sign-constrained regression after deletion/substitution

Linear regression analysis is a task for determining the value of d regression coefficients $w_1, \ldots, w_d \in \mathbb{R}$ that approximate the linear relationship $\mathrm{y} \cong w_1 x_1 + \ldots + w_d x_d = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ between d water quality parameters, $x_1, \ldots, x_d \in \mathbb{R}$ and an *E. coli* concentration y, where $\boldsymbol{x} := [x_1, \ldots, x_d]^{\mathrm{T}}$ and $\boldsymbol{w} := [w_1, \ldots, w_d]^{\mathrm{T}}$. Regression coefficients in $\mathbf{w}$ are usually determined by minimizing the mean square error

$$\mathrm{P}(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} (\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - y_i)^2 \tag{3}$$

where $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ are n observations in a given dataset. Typically, in the water quality engineering field, the signs of the true correlations of some explanatory variables to a pathogen concentration can be expected in advance. However, in the case of a small dataset or weak correlations, the signs of the corresponding regression coefficients are often opposite to those of the true correlations. To alleviate this phenomenon, sign constraints were introduced in this study. Let $\mathbf{c} \in \{\pm 1, 0\}^d$ be a constant vector where h-th entry c_h is $+1$ if the h-th explanatory variable x_h is known to have a positive correlation to the *E. coli* concentration y; $c_h := -1$ if the negativity of the correlation is known; 0 otherwise. The feasible region is then expressed as:

$$S := \{\boldsymbol{w} \in R^d | \forall h, \ c_h w_h \geq 0\} \tag{4}$$

The sign-constrained least square estimation is a task to find the regression coefficients that minimize the mean square error in the feasible region. In other words, the sign-constrained least square problem can be expressed as:

$$\min P(\boldsymbol{w}) \ \mathrm{wrt}\, \boldsymbol{w} \in S \tag{5}$$

When some *E. coli* concentration values in a given dataset are not detected due to a quantification limit, the mean square error cannot be assessed. Simple solutions to this issue are substitution or deletion of non-detects before the sign-constrained least square estimation. In the substitution method, a constant value (the quantification limit value or half the quantification limit value) was substituted into a non-detect. Then, the estimation task was reduced to the sign-constrained linear regression problem with dataset size n. The deletion method deletes non-detects to obtain the sign-constrained linear regression problem with dataset size n_v where n_v is the number of detected *E. coli* concentration.

Sign-constrained Tobit

In the Tobit model (Amemiya 1984), *E. coli* detection failure is described with a mass probability, and the regression coefficients $\boldsymbol{w}$ are determined by maximum likelihood estimation. Let us denote the n_v data pairs of water quality parameter vectors and *E. coli* concentration values by $(\boldsymbol{x}_1^v, y_1^v), \ldots, (\boldsymbol{x}_{n_v}^v, y_{n_v}^v) \in \mathbb{R}^d \times \mathbb{R}$. Note that all *E. coli* concentration values y_i^v in the n_v data pairs are over the quantification limit u. Denote $n_h (:= n - n_v)$ data pairs with

non-detects by $(\boldsymbol{x}_1^h, y_1^h), \ldots, (\boldsymbol{x}_{n_h}^h, y_{n_h}^h) \in \mathbb{R}^d \times \mathbb{R}$. When *E. coli* is not detected, the information available is that the true *E. coli* concentration values do not exceed the quantification limit. For such a dataset, the log-likelihood function of the Tobit model is given by:

$$L(\boldsymbol{w}, \beta) := \sum_{i=1}^{n_v} \log N(y_i^v; \langle \boldsymbol{w}, \boldsymbol{x}_i^v \rangle, \beta^{-1}) + \sum_{i'=1}^{n_h} \int_{-\infty}^{u} \log N(y_{i'}^h; \langle \boldsymbol{w}, \boldsymbol{x}_{i'}^h \rangle, \beta^{-1}) dy_{i'}^h \tag{6}$$

where β is referred to as the precision parameter. In the classical Tobit model, the log-likelihood function $L(\boldsymbol{w}, \beta)$ is maximized without any constraints.

In this study, the sign constraints were introduced for maximum likelihood estimation, which is written as:

$$\max L(\boldsymbol{w}, \beta) \quad \text{wrt} \quad \boldsymbol{w} \in S,\ \beta \in \mathbf{R} \tag{7}$$

To solve this maximization problem, a new expectation-maximization (EM) algorithm (MacKay 2003) was developed. In the EM algorithm, a posterior distribution $q_t(y_{i'}^h)$ is introduced for each of the non-detects at each iteration, and the E-step and M-step are repeated alternately until convergence. Denoting by $(\boldsymbol{w}^{(t)}, \beta^{(t)})$ the pair of the regression coefficient vector and the precision parameter at t-th iteration, the posterior distribution is defined in the E-step as:

$$q_t(y_{i'}^h) = \begin{cases} \dfrac{1}{\Phi(\xi_i^{(t)})} N(y_i^h; \mu_i^{(t)}, 1/\beta^{(t)}) & \text{if } y_i^h \le u \\ 0 & \text{if } y_i^h > u \end{cases} \tag{8}$$

where

$$\mu_i^{(t)} := \langle \boldsymbol{w}^{(t)}, \boldsymbol{x}_i^h \rangle, \quad \text{and } \xi_i^{(t)} := (u - \mu_i^{(t)}) \sqrt{\beta^{(t)}} \tag{9}$$

In the M-step, the value of $(\boldsymbol{w}, \beta)$ is updated so that the Q-function is increased, where the Q-function is defined as:

$$\mathrm{Q}(\boldsymbol{w}, \beta; q_t) := \frac{n}{2} \log(\beta) - \frac{\beta}{2} \left(\left\| \boldsymbol{X}^{\mathrm{T}} \boldsymbol{w} - \bar{\boldsymbol{y}}^{(t)} \right\|^2 + v^{(t)} \right) \tag{10}$$

where

$$\boldsymbol{X} := \left[\boldsymbol{x}_1^v, \ldots, \boldsymbol{x}_{n_v}^v, \boldsymbol{x}_1^h, \ldots, \boldsymbol{x}_{n_h}^h \right]$$

$$\mathbf{y}^{\mathrm{v}} := \left[y_1^v, \ldots, y_{n_v}^v \right]^{\mathrm{T}}, \qquad \mathbf{y}^{\mathrm{h}} := \left[y_1^h, \ldots, y_{n_h}^h \right]^{\mathrm{T}}$$

$$\bar{\boldsymbol{y}}^t := \begin{bmatrix} \mathbf{y}^{\mathrm{v}} \\ \mathbb{E}_{q_t}[\mathbf{y}^{\mathrm{h}}] \end{bmatrix}, \qquad \mathrm{v}^{(\mathrm{t})} := \mathbb{E}_{q_t}\left[\|\boldsymbol{y}^h\|^2 \right] + \left\| \mathbb{E}_{\boldsymbol{q}_t}\left[\boldsymbol{y}^h \right] \right\|^2$$

Therein, $\mathbb{E}_{q_t}$ is the operator taking the expectation over the posterior distribution defined in the E-step of the t-th iteration. With this Q-function, $\boldsymbol{w}$ and β are updated as:

$$\boldsymbol{w}^{(t+1)} := \mathrm{argmax}_{\boldsymbol{w} \in S} \mathrm{Q}(\boldsymbol{w}, \beta^{(t)}; q_t) \tag{11}$$

and

$$\beta^{(t+1)} := \mathrm{argmax}_{\beta \in S} \mathrm{Q}(\boldsymbol{w}^{(t+1)}, \beta; q_t) = \frac{n}{\left\| \boldsymbol{X}^{\mathrm{T}} \boldsymbol{w} - \bar{\boldsymbol{y}}^{(t)} \right\|^2 + v^{(t)}} \tag{12}$$

The definition of the Q-function implies that the update rule of $\boldsymbol{w}$ is reduced to the sign-constrained least square problem, which can be solved efficiently. More detailed information of sign-constrained Tobit, including EM steps is indicated in Appendix E of the Supplementary materials (available online).

Evaluation process

In order to evaluate the generalized prediction performance for unseen data, 96 data pairs were divided into training and evaluation datasets. The size of the training datasets, say n, was between 10 and 58. There were four approaches to deal with the censored dataset: Tobit analysis (Tobit), the substitution of non-detect with the quantification limit value (DL), the substitution of non-detect with half the quantification limit value (DL/2), and the deletion of non-detects (Del). For each approach, the generalization performance of the conventional sign-free regression was compared with that of the sign-constrained regression. Sign-constrained Tobit, DL, DL/2, and Del were expressed as SC-Tobit, SC-DL, SC-DL/2, and SC-Del, respectively, whereas those without sign constraints were SF-Tobit, SF-DL, SF-DL/2, and SF-Del, respectively. In total, eight methods were examined.

In the evaluation, the regression coefficients in $\boldsymbol{w}$ were determined using a training dataset of the eight approaches (SC-Tobit, SF-Tobit, SC-DL, SF-DL, SC-DL/2, SF-DL/2, SC-Del, and FC-Del). Then, the root mean square deviation (RMSD) was calculated using an evaluation dataset as follows:

$$\mathrm{RMSD} = \sqrt{\frac{1}{n_{\mathrm{tst}}} \sum_{i=1}^{n_{\mathrm{tst}}} \left(y_i^{\mathrm{tst}} - \langle \boldsymbol{w}, \boldsymbol{x}_i^{\mathrm{tst}} \rangle\right)^2} \tag{13}$$

where $(\boldsymbol{x}_1^{\mathrm{tst}}, y_1^{\mathrm{tst}}), \ldots, (\boldsymbol{x}_{n_{\mathrm{tst}}}^{\mathrm{tst}}, y_{n_{\mathrm{tst}}}^{\mathrm{tst}}) \in \mathbb{R}^d \times \mathbb{R}$ are n_{tst} data pairs in the evaluation dataset. The size of every evaluation dataset is $n_{\mathrm{tst}} = 25$, and the points are chosen from (96 − n) data points. This process of training and evaluation was repeated 100 times for each approach, and the average and standard deviation of RMSD were calculated.

To make the evaluation process reproducible, a step-by-step description of the evaluation process is given as follows. Two subsets with size n and $n_{\mathrm{tst}} = 25$ were chosen from 96 data points, to obtain the training and evaluation datasets, respectively. The intersection of the two subsets were empty. This step is repeated 100 times, and 100 different training/evaluation datasets were generated for each $n \in \{10, 13, 16, \ldots, 58\}$. For each of 100 different training/evaluation datasets and each of eight methods, the regression coefficients were estimated with the training dataset and RMSD was evaluated with the evaluation dataset.

RESULTS

Before examining the effects of the sign constraints, the dataset that was used for examining the sign constraints was overviewed. The minimum and maximum values of the common logarithm of *E. coli* concentration were 0.89 and 5.38, respectively, and the median was 3.20. To visualize the relationship between the response variable and each explanatory variable, the non-parametric density estimation was performed using the Parzen window (Figure 1(a)). Weak but statistically significant correlations were observed between the response variable and each explanatory variable except for pH (pH values less than 7.0). The Pearson's correlation coefficients and p values in t-test with sample size 96, shown in Table 1, indicate that there were statistically significant positive correlations ($p < 0.05$) with water temperature, electric conductivity, suspended solids, biological oxygen demand, total nitrogen, and total phosphorus, but statistically significant negative correlations ($p < 0.05$) with pH_+ (pH values higher than 7.0), dissolved oxygen, and flow rate.

What happens when the sample size is small is demonstrated preliminary to reporting the effects of sign constraints. The positive correlations between the response variable and six explanatory variables (water temperature, electric conductivity, suspended solids, biological oxygen demand, total nitrogen, and total phosphorus) were statistically significant when all 96 data were used for the computation. However, these correlations would not be significant when the sample size was too small. Similarly, significant negative correlations with pH_+, dissolved oxygen, and flow rate would not be detected when the sample size was too small. In order to confirm this intuition, a Pearson's sample correlation coefficient was computed using five randomly selected data out of 96 for each explanatory variable. This simulation was repeated 10,000 times, and 10,000 values of the sample correlation coefficient were obtained for each explanatory variable. Then, the histograms of the sample correlation coefficients were plotted (Figure 1(b)), and the percentage values of positive and negative sign of the sample correlation coefficient were calculated (Table 2). As shown in Figure 1(b), the correlation coefficients for all explanatory variables were scattered in a broad interval, which is due to the small size of the sample. The strongest positive correlation was obtained with total nitrogen (Table 1), but 3.9% of the sample correlation coefficient values were negative (Table 2). The strongest negative correlation was obtained with dissolved oxygen (Table 1), but 5.5% of the sample correlation coefficient values were positive (Table 2). Suspended solids had a significant positive correlation (Table 1), but 22.0% of the sample correlation coefficient values were negative (Table 2). The flow rate had a significant negative correlation (Table 1), but 28.0% of the sample correlation coefficient values were positive (Table 2). These results clearly indicate that the sign reversal between sample and population correlation coefficient values occurs when the sample size is small, which exacerbates the performance

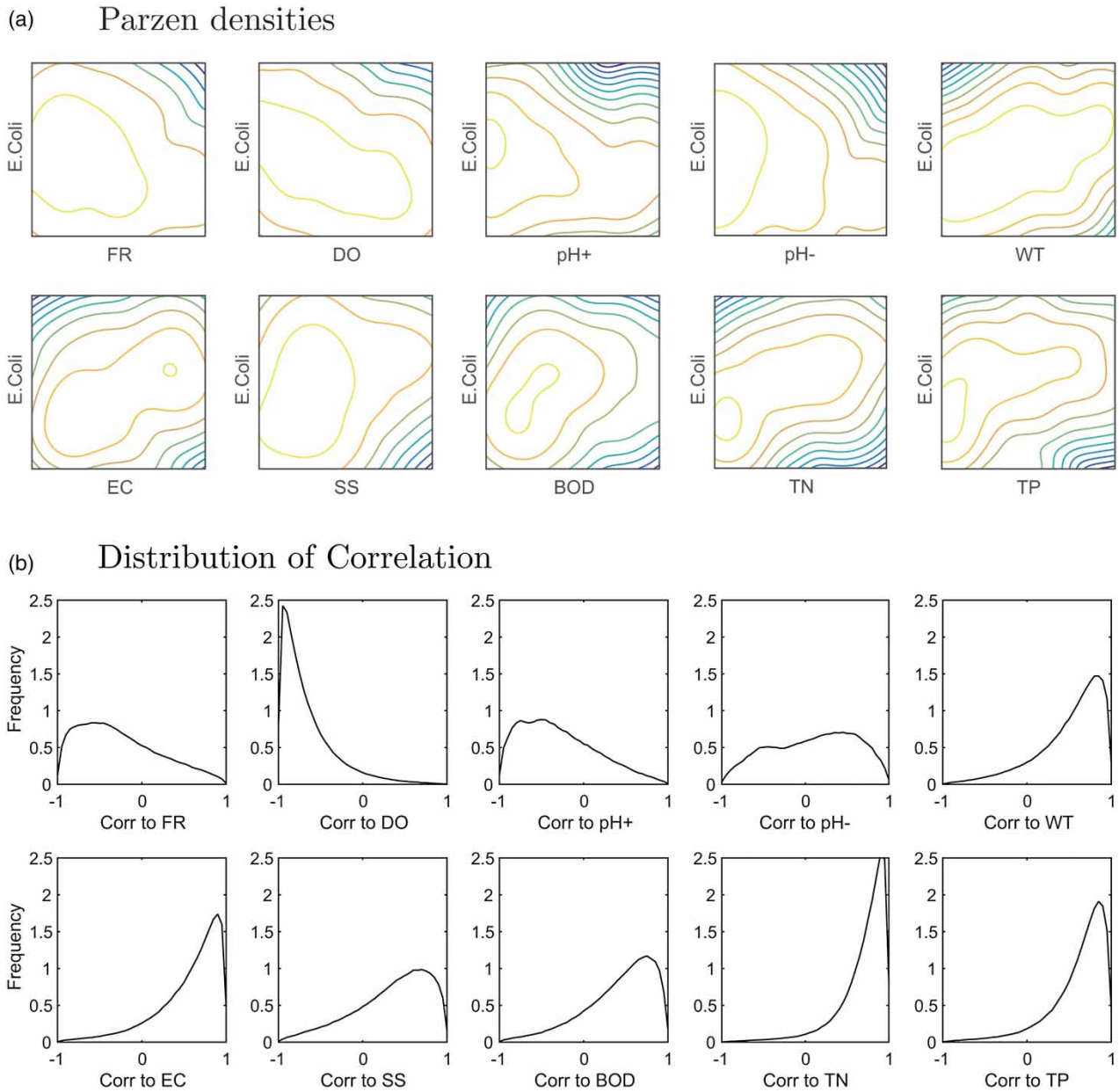

Figure 1 | Observations of the water quality monitoring data and *E. coli* concentration values. (a) Distributions over an explanatory variable and the response variable, where the explanatory variable is one of the water quality monitoring data, and the response variable is the *E. coli* concentration values. (b) Distributions of the Pearson correlation coefficient between an explanatory variable and the response variable when the sample size is limited to five. FR: flow rate, DO: dissolved oxygen, WT: water temperature, EC: electric conductivity, SS: suspended solids, BOD: biological oxygen demand, TN: total nitrogen, TP: total phosphorus.

of conventional regression analysis based on the standard least square estimation.

The effects of sign constraints in Tobit, DL, DL/2, and Del were compared with sign-free approaches when the quantification limit value was 1.5, 2.0, or 2.5-log MPN/100 mL (Figure 2) and 3.0, 3.5, or 4.0-log MPN/100 mL (Figure 3). RMSD between sign-constraint and sign-free approaches and *p*-values at each number of training examples is indicated in Tables SA.1–SA.24 of the Supplementary materials (available with the online version of this paper). The Tobit model is one of the statistically sophisticated models that include a truncated probabilistic

Table 1 | Pearson's correlation coefficient and *p* value when all 96 data pairs are used

Explanatory variable	Correlation coefficient	*p* value
Flow rate	−0.287	5.23×10^{-5}
Dissolved oxygen	−0.683	5.67×10^{-26}
pH_+^a	−0.293	3.88×10^{-5}
pH_-^b	−0.019	4.02×10^{-1}
Water temperature	0.548	1.41×10^{-15}
Electrical conductivity	0.562	1.89×10^{-16}
Suspended solids	0.369	2.17×10^{-7}
Biological oxygen demand	0.381	8.14×10^{-6}
Total nitrogen	0.691	9.17×10^{-27}

[a]pH_+: max (0, pH − 7.0).
[b]pH_-: max (0, 7.0 − pH).

distribution, in which true values in non-detects are expressed using mass probabilities and estimated using maximum likelihood estimation in the marginal distributions. It was remarkable that the sign-constraint approaches (triangles) always gave smaller RMSD values than the sign-free approaches (squares), particularly when the training data size was small. This means that the accuracy of the regression is improved by the employment of sign-constraint approaches compared to conventional sign-free approaches.

The sign-free approaches were compared among Tobit, DL, DL/2, and Del (Figure 4(a)–4(f)). When the quantification limit was 3.0-log MPN/100 mL, the sign-free Tobit (square) performed best in terms of the regression accuracy if the training sample size was larger than 40 (Figure 4(d)). The sign-constraint approaches were also compared among Tobit, DL, DL/2, and Del (Figure 4(g)–4(l)). The performance of SC-DL/2 was the best when the quantification limit value was smaller than 2.5-log MPN/100 mL, but the difference among SC-DL, SC-DL/2, and SC-Del was almost negligible. When the training sample size was 21, the RMSD of SC-Tobit was larger than those of SC-DL, SC-DL/2, and SC-Del. The difference of RMSD among SC-Tobit, SC-DL, SC-DL/2, and SC-Del was very small when the training sample size was large. SC-Tobit performed best when the training sample size was large and the quantification limit value was higher than 3.0-log MPN/100 mL, whereas SC-DL/2 was the best when the quantification limit value was less than 2.5-log MPN/100 mL. These results indicate that the best approach (the smallest RMSD) is determined by the sample size when the quantification limit value is larger than 3.0-log MPN/100 mL.

Table 2 | Percentages of positive and negative Pearson correlation coefficients when the sample size is limited to five

Explanatory variable	Positive correlation %	Negative correlation %
Flow rate	28.0	72.0
Dissolved oxygen	5.4	94.6
pH_+^a	25.0	73.5
pH_-^b	51.2	36.6
Water temperature	88.4	11.6
Electrical conductivity	90.3	9.7
Suspended solids	78.1	21.9
Biological oxygen demand	83.6	16.4
Total nitrogen	96.1	3.9

[a]pH_+: max (0, pH − 7.0).
[b]pH_-: max (0, 7.0 − pH).

DISCUSSION

When there is a significant correlation between pathogen concentration in water and explanatory variables, prediction accuracy is improved by the linear combination of multiple explanatory variables, if the training sample size is large enough (Chatterjee & Price 1977). However, the sign of the sample correlation coefficient may become opposite when the training sample size is not large enough, as shown in Table 2, which results in performance deterioration of the predictor. In this study, we propose the following simple assignment on the sign of correlation coefficient: (1) non-negative regression coefficient ($w_h \geq 0$) of an explanatory variable if it is common to have a positive correlation coefficient; and (2) non-positive regression coefficient ($w_h \leq 0$) of an explanatory variable if it is common to have a negative correlation coefficient. The effect of the sign constraints was clear, as can be seen in Figures 2 and 3: the averages of RMSD values were always lower (the prediction performance is always higher) in the sign-constraint approaches compared with the sign-free approaches.

The sign constraints play a role in removing the explanatory variables in a given dataset that violates the domain knowledge

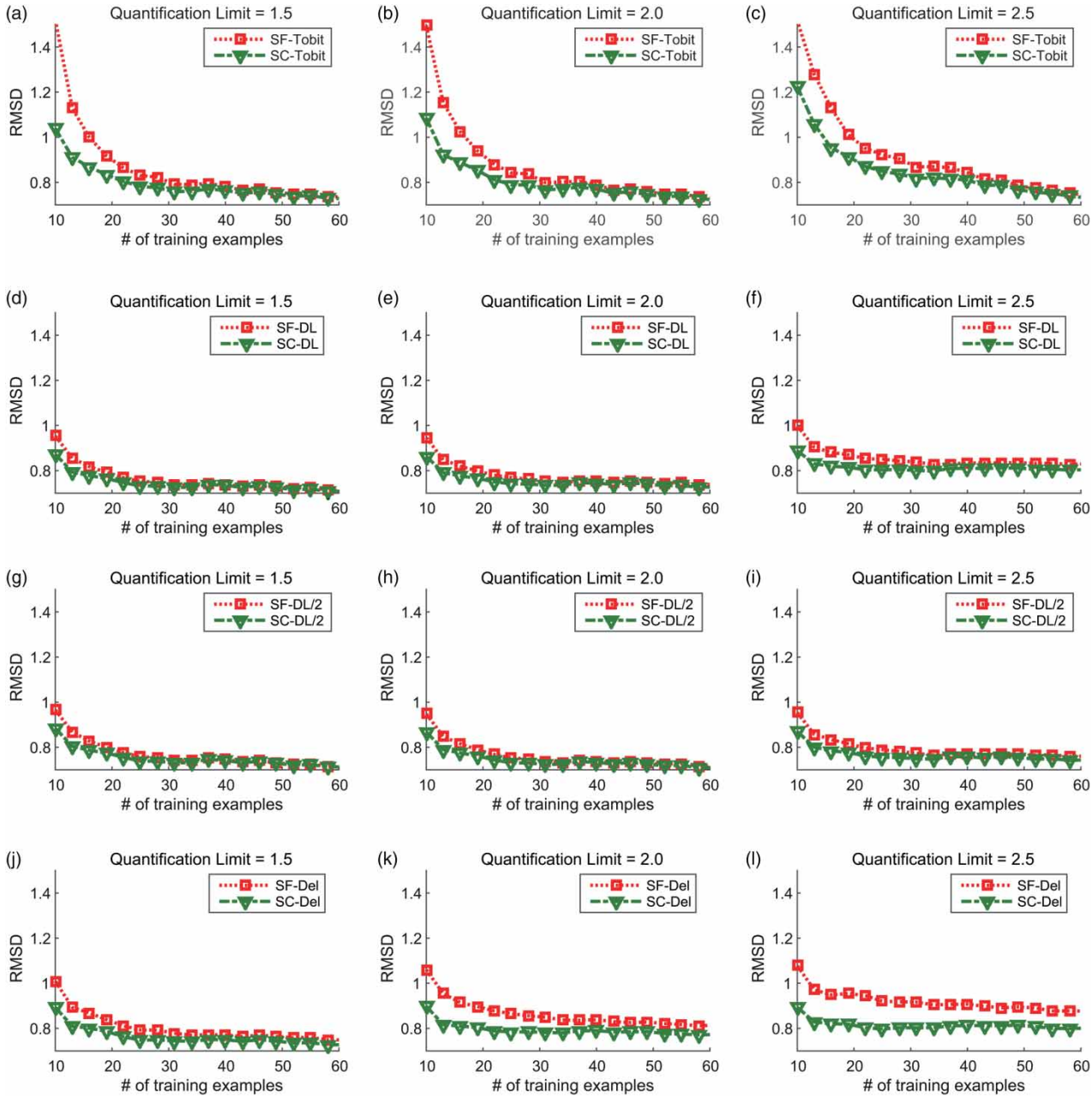

Figure 2 | The average of RMSD between the sign-constrained and the sign-free fittings when the quantification limit is between 1.5-log and 2.5-log MPN/100 mL. SC: sign-constrained, SF: sign-free, DL: the substitution of non-detect with the quantification limit value, DL/2: the substitution of non-detect with half the quantification limit value, Del: the deletion of non-detect.

by zeroing the corresponding regression coefficients. The box plots in Figures SB.1–SB.24 of the Supplementary materials (available with the online version of this paper) show the distributions of the regression coefficient values obtained in regression analysis. The coefficient values were distributed widely in the feasible region associated with the sign constraints for each explanatory variable. The coefficient values frequently vanished, especially when the training dataset size is smaller. Note that for a coefficient value that did not vanish, the estimated coefficient values would not change even if the corresponding sign constraints were excluded from the optimization problem for model fitting. On the other hand, the zero

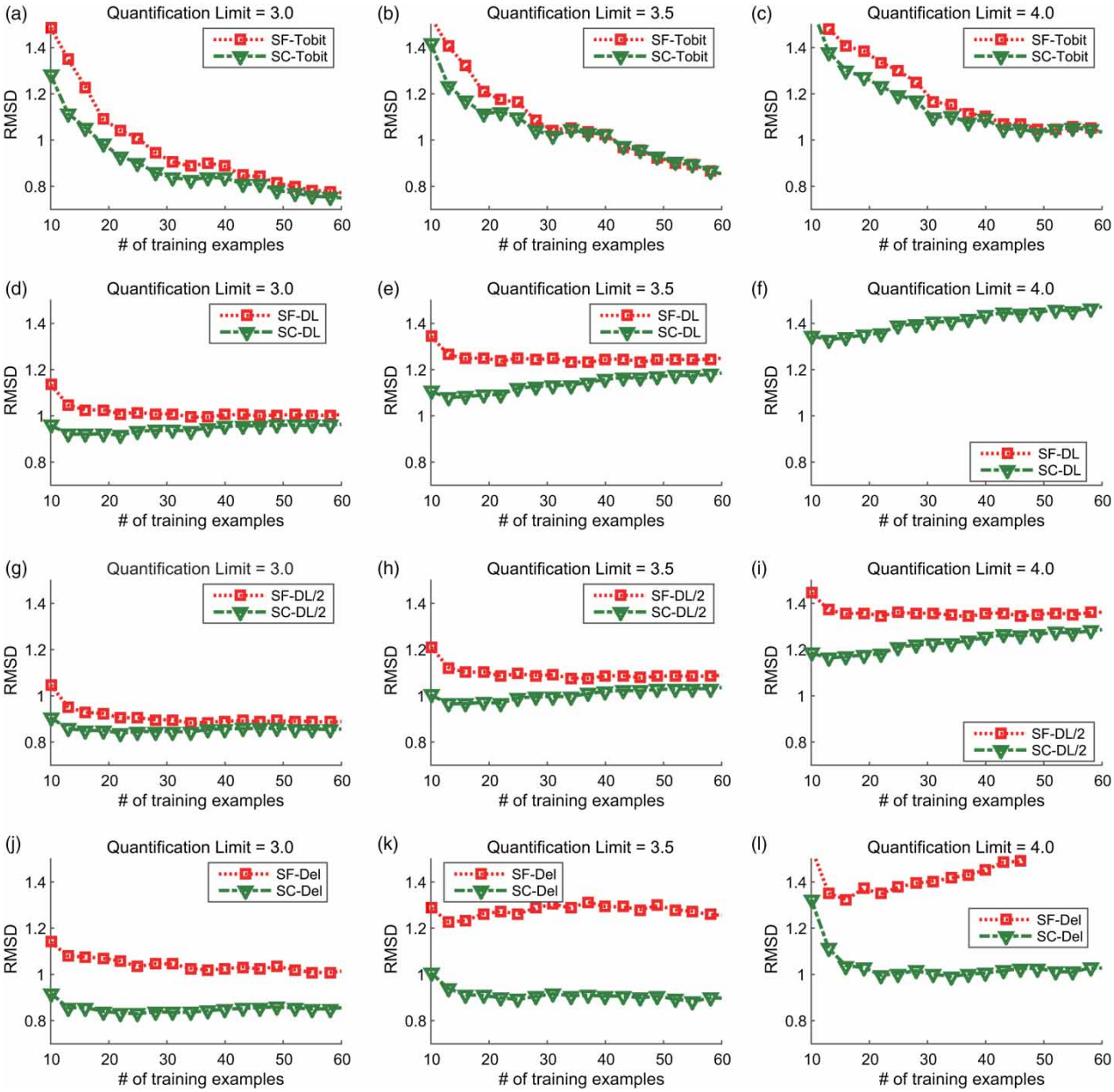

Figure 3 | The average of RMSD between the sign-constrained and the sign-free fittings when the quantification limit is between 3.0-log and 4.0-log MPN/100 mL. SF: sign-constrained, SF: sign-free, DL: the substitution of non-detect with the quantification limit value, DL/2: the substitution of non-detect with half the quantification limit value, Del: the deletion of non-detect.

coefficients suggest that the existence of the sign constraints might change the estimated solution from the sign-free approach. The frequencies of trials of 100-time repeated experiments in which specified numbers of regression coefficients vanished were plotted in a stacked bar format in Figures SC.1–SC.24 of the Supplementary materials (available online). Each bar is a stack of three sections at most. The lowest sections indicated with '<' are the frequencies with which RMSD was reduced by imposing the sign constraints; the middle sections indicated with '=' are the frequencies with which no difference in RMSD between the sign-constrained and the sign-free approaches was observed; and the highest sections indicated with '>' are the frequencies with which RMSD was increased with the sign constraints.

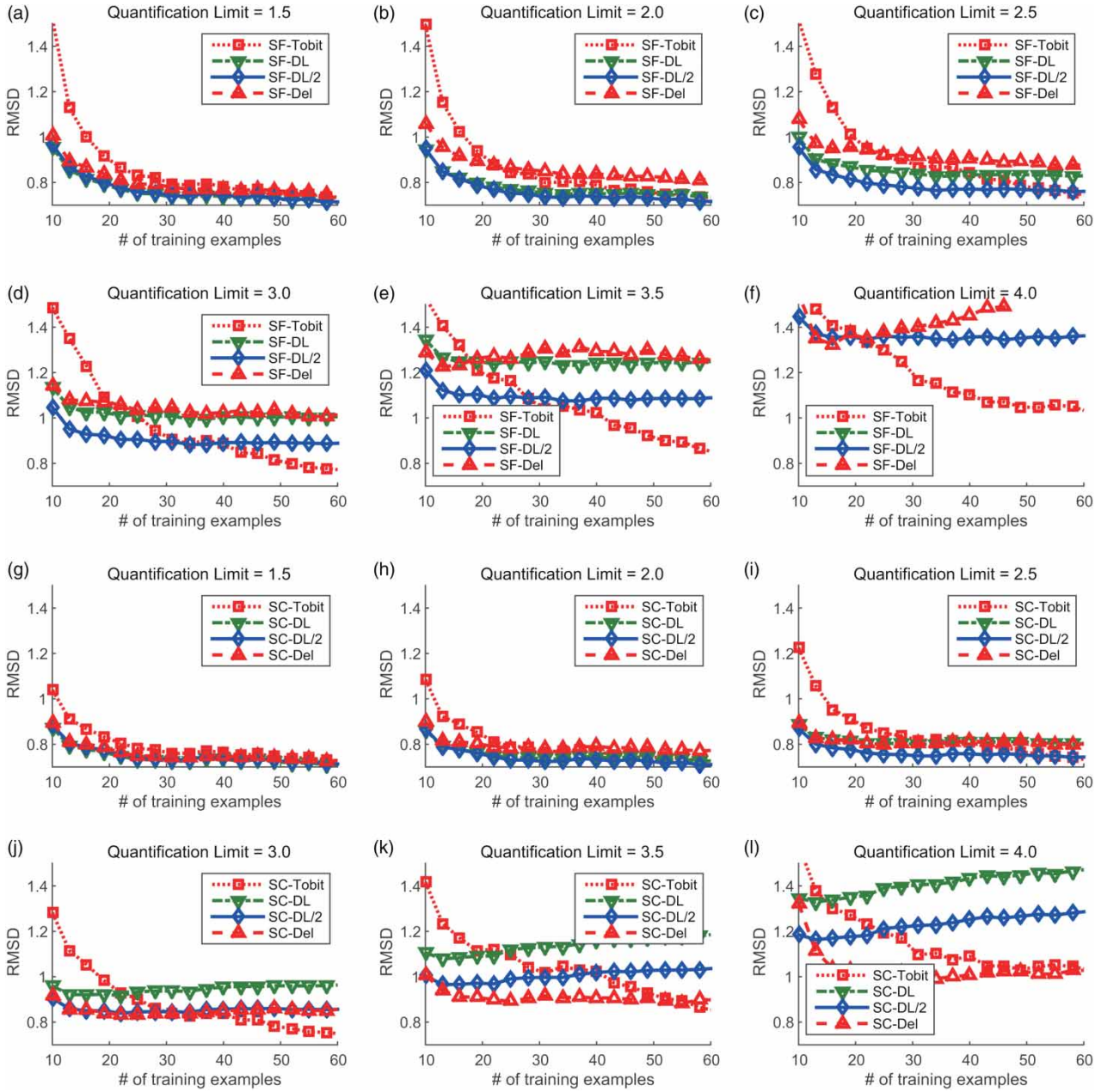

Figure 4 | Comparison of different methods for the sign-constrained and the sign-free fittings. SC: sign-constrained, SF: sign-free, DL: the substitution of non-detect with the quantification limit value, DL/2: the substitution of non-detect with half the quantification limit value, Del: the deletion of non-detect.

For Tobit methods, RMSD was often reduced when the training sample size was small and the quantification limit was low. For the other three approaches, the sign constraints tended to decrease RMSD. The averaged numbers of vanished coefficients against the training dataset size are plotted in Figure SC.25 of the Supplementary materials and these averaged numbers and the standard deviations are shown in Tables SC.1–SC.4 of the Supplementary materials (available online). For Tobit, DL, and DL/2 methods, more vanished coefficients were observed when the quantification limit is lower, reflecting the fact that the domain knowledge disagreed less frequently in training datasets containing more censored data. For Del method, higher quantification limit reduced the data fed to model fitting (recall that Del method uses only uncensored

data), which produced more explanatory variables contradicting domain knowledge and thereby increased the number of zeroed coefficients. These results also suggest that the explanatory variables are discarded if the corresponding sign constraints are contradicted with a given sample, and the sign constraints are effective even with a small number of wrong constraints given.

The most remarkable technological contribution of this study is the finding that it is possible to incorporate the sign constraints in the estimation of the correlation coefficient in the EM algorithm for Tobit analysis (Amemiya 1984). The sign-constrained regression exists and the Tobit model exists, although no algorithm for the combination has been developed so far. The new discovery was that the M-step in the EM algorithm boiled down to a mathematical problem of non-negative least square estimation when the maximum likelihood estimation is conducted under sign-constraint. Algorithms for the non-negative least square estimation have been studied extensively and an efficient algorithm exists (Lawson & Hanson 1987). By introducing the efficient algorithm for non-negative least square estimation to the M-step of the EM algorithm, an effective and numerically stable algorithm for fitting a model to datasets under the non-negative constraint was obtained. The novel algorithm is applicable to environmental correlation issues, including the prediction of pathogen concentration in water.

The EM approach employed in this study was not a unique choice for finding the maximum likelihood estimator. Under an invertible variable transformation $(\mathbf{w}, \beta) \mapsto (\beta^{0.5}\boldsymbol{w}, \beta^{0.5})$, the Hessian matrix of the negative log-likelihood function is guaranteed to be positive definite, which implies the global concavity (Amemiya 1984). This fact suggests another approach such as gradient-based methods for maximum likelihood estimation under sign constraints. An advantage of the EM algorithm is the ease of extensions to more complicated models and Bayesian approaches. If performing Bayesian inference instead of maximum likelihood estimation, the prior distribution is designed so that the prior probabilistic densities outside the feasible region of the sign constraints are zero, and the variational approximation (MacKay 2003) which is a natural extension of the EM algorithm may be needed to make the Bayesian algorithm tractable.

CONCLUSIONS

In this study, new approaches introducing sign constraints to express such domain knowledge were attempted. Effective and numerically stable algorithms for fitting a model to left-censored datasets under the sign constraints were developed, which must be applicable to a wide variety of environmental prediction problems, including the real-time monitoring of pathogen concentration in water. It was confirmed that the prediction performance of the regression was improved by the employment of sign-constraint approaches compared to conventional sign-free approaches. In particular, more significant improvements were observed when the training sample is small, implying that the sign-constraint techniques are a powerful option for practical analysts when they choose a statistical tool. Another contribution of this paper is to present a novel algorithm for fitting of Tobit model under sign constraints. The presented algorithm was an implementation of the algorithm for the fitting problem.

ACKNOWLEDGEMENTS

This study was supported by JSPS KAKENHI Grant Number JP15K00591. No conflict of interest is declared.

REFERENCES

Amemiya, T. 1984 Tobit models: a survey. *J. Econom.* **24** (1–2), 3–61.

Antweiler, R. C. 2015 Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets. II. Group comparison. *Environ. Sci. Tech.* **49**, 13439–13446.

APHA 2005 *Standard Methods for the Examination of Water and Wastewater*, 21st edn. American Public Health Association/American Water Works Association/Water Environment Federation, Washington, DC, USA.

Chatterjee, S. & Price, B. 1977 *Regression Analysis by Example*. Wiley, New York, USA.

Cho, K. H., Cha, S. M., Kang, J.-H., Lee, S. W., Park, Y., Kim, J.-W. & Kim, J. H. 2010 Meteorological effects on the levels of fecal indicator bacteria in an urban stream: a modeling approach. *Water Res.* **44**, 2189–2202.

Garner, E., Zhu, N., Strom, L., Edwards, M. & Pruden, A. 2016 A human exposome framework for guiding risk management

and holistic assessment of recycled water quality. *Environ. Sci.: Water Res. Technol.* **2**, 580–598.

Goodwin, D., Raffin, M., Jeffrey, P. & Smith, H. M. 2015 Applying the water safety plan to water reuse: towards a conceptual risk management framework. *Environ. Sci.: Water Res. Technol.* **1**, 709–722.

Harwood, V. J., Levine, A. D., Scott, T. M., Chivukula, V., Lukasik, J., Farrah, S. R. & Rose, J. B. 2005 Validity of the indicator organism paradigm for pathogen reduction in reclaimed water and public health protection. *Appl. Environ. Microbiol.* **71**, 3163–3170.

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

Helsel, D. R. 2010 Summing nondetects: incorporating low-level contaminants in risk assessment. *Integr. Environ. Assess. Manag.* **6** (3), 361–366.

Huynh, T., Ramachandran, G., Banerjee, S., Monteiro, J., Stenzel, M., Sandler, D. P., Engel, L. S., Kwok, R. K., Blair, A. & Stewart, P. A. 2014 Comparison of methods for analyzing left-censored occupational exposure data. *Ann. Occup. Hyg.* **58** (9), 1126–1142.

Huynh, T., Quick, H., Ramachandran, G., Banerjee, S., Stenzel, M., Sandler, D. P., Engel, L. S., Kwok, R. K., Blair, A. & Stewart, P. A. 2016 A comparison of the beta-substitution method and a Bayesian method for analyzing left-censored data. *Ann. Occup. Hyg.* **60** (1), 56–73.

Jurzik, L., Hamza, I. A., Puchert, W., Überla, K. & Wilhelm, M. 2010 Chemical and microbiological parameters as possible indicators for human enteric viruses in surface water. *Int. J. Hyg. Environ Health* **213**, 210–216.

Kato, T., Kobayashi, A., Ito, T., Miura, T., Ishii, S., Okabe, S. & Sano, D. 2016 Estimation of concentration ratio of indicator to pathogen-related gene in environmental water based on left-censored data. *J. Water Health* **14** (1), 14–25.

Kitajima, M., Wang, N., Tay, M. Q. X., Miao, J. & Whittle, A. J. 2016 Development of a MEMS-based electrochemical aptasensor for norovirus detection. *Micro Nano Lett.* **11** (10), 582–585.

Lawson, C. L. & Hanson, R. J. 1987 *Solving Least Squares Problems* (Classics in Applied Mathematics). Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

MacKay, D. 2003 *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK.

First received 16 December 2018; accepted in revised form 6 March 2019. Available online 3 April 2019