Predictive water virology using regularized regression analyses for projecting virus inactivation efficiency in ozone disinfection

Syun-suke Kadoya a, Osamu Nishimura a, Hiroyuki Kato b, Daisuke Sano a, c, *

a Department of Civil and Environmental Engineering, Graduate School of Engineering, Tohoku University, Aoba 6-6-06, Aramaki, Aoba-ku, Sendai, Miyagi, 980-8579, Japan
b New Industry Creation Hatchery Center, Tohoku University, Sendai, Miyagi, Japan
c Department of Frontier Sciences for Advanced Environment, Graduate School of Environmental Studies, Tohoku University, Aoba 6-6-06, Aramaki, Aoba-ku, Sendai, Miyagi, 980-8579, Japan

A R T I C L E   I N F O

Article history:
Received 3 January 2021
Received in revised form 5 February 2021
Accepted 6 February 2021
Available online 12 February 2021

Keywords:
Log reduction value
Ozone disinfection
Waterborne viruses
Regularized regression analyses
Hierarchical bayesian modeling

A B S T R A C T

Wastewater reclamation and reuse have been practically applied to water-stressed regions, but waterborne pathogens remaining in insufficiently treated wastewater are of concern. Sanitation Safety Planning adopts the hazard analysis and critical control point (HACCP) approach to manage human health risks upon exposure to reclaimed wastewater. HACCP requires a predetermined reference value (critical limit: CL) at critical control points (CCPs), in which specific parameters are monitored and recorded in real time. A disinfection reactor of a wastewater treatment plant (WWTP) is regarded as a CCP, and one of the CCP parameters is the disinfection intensity (e.g., initial disinfectant concentration and contact time), which is proportional to the log reduction value (LRV) of waterborne pathogens. However, the achievable LRVs are not always stable because the disinfection intensity is affected by water quality parameters, which vary among WWTPs. In this study, we established models for projecting virus LRVs using ozone, in which water quality and operational parameters were used as explanatory variables. For the model construction, we used five machine learning algorithms and found that automatic relevance determination with interaction terms resulted in better prediction performances for norovirus and rotavirus LRVs. Poliovirus and coxsackievirus LRVs were predicted well by a Bayesian ridge with interaction terms, and lasso with quadratic terms, respectively. The established models were relatively robust to predict LRV using new datasets that were out of the range of the training data used here, but it is important to collect LRV datasets further to make the models more predictable and flexible for newly obtained datasets. The modeling framework proposed here can help WWTP operators and risk assessors determine the appropriate CL to protect human health in wastewater reclamation and reuse.

© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Water shortages are a critical global concern and endanger human lives all over the world. Wastewater reclamation and reuse have significantly contributed to the improvement of the accessibility of safe water, but waterborne viruses such as norovirus, rotavirus, adenovirus and enterovirus have been detected from effluents of wastewater treatment plants (WWTPs) (Subahir et al., 2019; Hoque et al., 2018; Croker et al., 2018; Bauri et al., 2019; Li et al., 2018; Silva-Sales et al., 2020; Katayama et al., 2008; Adeifisove et al., 2016; Shlindwein et al., 2010). The risk of viral infection to humans, including users of reclaimed wastewater, needs to be addressed because waterborne viruses in wastewater have imposed a significant disease burden on human society (Utsumi et al., 2020; Subahir et al., 2019; Hoque et al., 2018; Croker et al., 2018; Bauri et al., 2019; Li et al., 2018).

International and domestic guidelines for wastewater reclamation and reuse employ the multiple-barrier concept to address the waterborne disease risk in the usage of reclaimed wastewater, in which the target log reduction value (LRV) of waterborne pathogens is determined beforehand, and the target LRV is allocated to each step at a wastewater treatment plant (WWTP) (Ito et al., 2017; Sano et al., 2016). The World Health Organization (WHO) has recommended applying Sanitation Safety Planning (SSP) to the safe
Fig. 1. Distribution of operational and water quality parameters. Water quality and operational parameters except for strain types were plotted for each virus species as swarm, box (median, 25 and 75 percentile) and violin plots. “Assay” is classified as A1 (genome or infectivity) or A2. (If classifying as infectivity, the variable is also classified as plaque forming unit (PFU) or most probable number (MPN) method.).

| Parameter | Ref |
|-----------|-----|
| Norovirus | 68  |
| Rotavirus | 65  |
| Poliovirus| 182 |
| Coxsackievirus | 111 |

Table 1
Information about datasets used for the model construction (NLRV: the number of log reduction values, NE: the number of articles, NS: the number of virus types, I: initial ozone concentration, k: decay constant, C: Ct-value, p: pH, T: temperature, A1: infectivity or genome, A2: plaque forming unit or most probable number method, W: water types, V: log10 initial virus concentration).

| Parameter | Ref |
|-----------|-----|
| Norovirus | 68  |
| Rotavirus | 65  |
| Poliovirus| 182 |
| Coxsackievirus | 111 |

| Parameter | Ref |
|-----------|-----|
| Norovirus | 68  |
| Rotavirus | 65  |
| Poliovirus| 182 |
| Coxsackievirus | 111 |

| Parameter | Ref |
|-----------|-----|
| Norovirus | 68  |
| Rotavirus | 65  |
| Poliovirus| 182 |
| Coxsackievirus | 111 |

| Parameter | Ref |
|-----------|-----|
| Norovirus | 68  |
| Rotavirus | 65  |
| Poliovirus| 182 |
| Coxsackievirus | 111 |

| Parameter | Ref |
|-----------|-----|
| Norovirus | 68  |
| Rotavirus | 65  |
| Poliovirus| 182 |
| Coxsackievirus | 111 |

| Parameter | Ref |
|-----------|-----|
| Norovirus | 68  |
| Rotavirus | 65  |
| Poliovirus| 182 |
| Coxsackievirus | 111 |

| Parameter | Ref |
|-----------|-----|
| Norovirus | 68  |
| Rotavirus | 65  |
| Poliovirus| 182 |
| Coxsackievirus | 111 |

| Parameter | Ref |
|-----------|-----|
| Norovirus | 68  |
| Rotavirus | 65  |
| Poliovirus| 182 |
| Coxsackievirus | 111 |
use of excreta, wastewater and greywater, and the hazard analysis and critical control point (HACCP) approach is employed in SSP in order to manage human health risks upon exposure to reclaimed wastewater (WHO, 2006 and 2015). The HACCP approach includes the identification of critical control points (CCPs), which are important operational steps to determine the magnitude of hazardous factors in the final products (U.S. FDA, 1997). The parameters at the CCP are monitored and recorded in real time. In addition, the parameters are compared to reference values called critical limit (CL) to maintain the safety of final products (U.S. FDA, 1997). The corrective action needs to take place when the parameters monitored at CCP deviate from CL. The disinfection intensity, determined by some operational parameters such as Ct-value in a disinfection reactor, is suitable as a CCP parameter at WWTPs since it determines the LRV of waterborne pathogens at a disinfection reactor.

However, an application of SSP to the WWTP operation is difficult at present since a universal method for CL determination has not yet been developed. In addition, the difference in water quality among WWTPs makes it difficult to establish a universal approach that determines the CLs appropriate for respective WWTPs, because chemical disinfectants are consumed by contaminants in wastewater, and the decay of disinfectant concentration results in deviation from the expected disinfection intensity. Thus, the disinfection intensity that achieves the target LRV needs to be determined by taking water quality and operational information at each WWTP into account. Previously, we proposed the concept of predictive water virology, in which the models predicted virus LRV by using water quality and operational parameters as explanatory variables (Kadoya et al., 2019). The predictive water virology enables us to derive the appropriate CL to attain the target LRV under site-specific water quality. Kadoya et al. (2019 and 2020) previously proposed the concept of predictive water virology and reported that the predictive inactivation models, based on hierarchical Bayesian modeling (HBM) and regularized regressions, flexibly correspond to the changes and variety in water quality parameters among WWTPs (e.g., seasonal variation). Predictive

Fig. 2. Mean squared errors (MSEs) in Trial 1. Total datasets are randomly divided into test and training datasets in Trial 1. MSEs are compared between test (solid) and training (dashed) datasets (left), and the absolute values of difference between test and training data are also displayed (right).
water virology can help us understand the required operational conditions to achieve the target LRV, in which there is no gap between the predicted and target LRVs. However, these previous studies were limited in their predictions of newly obtained datasets that are not used for model construction, in other words, we need to know the predictable range of present models and identify which algorithms are more robust to predict LRVs using new datasets.

Basic regularized regression analyses are ridge, lasso and elastic net regressions, in which the multicollinearity problem is solved, and thus all explanatory variables can be used. All these regression approaches are extended versions of the least square method by adding a regularization term corresponding to each algorithm. The ridge regression shrinks absolute values of the coefficients of model variables (Hoerl and Kennard, 1970), whereas the lasso regression makes the coefficients of inessential variables zero with a decrease in the magnitude of model variables (Tibshirani, 1996). The elastic net regression combines regularized terms used in both the ridge and lasso regressions (Zou and Hastie, 2006). The lasso and elastic net regressions are classified into the sparse estimation, which selects the plausibly essential variables. In Bayesian regularized regression (Bayesian ridge and automatic relevance determination (ARD)), an objective variable is assumed to be expressed as Gaussian functions, and regularization terms are then estimated as hyperparameters by Bayes rule. A Bayesian ridge employs all explanatory variables as well as basic ridge regression, whereas ARD is also one of the sparse estimations (Tipping, 2001). Previous studies have applied several machine learning algorithms to establish predictive models for water quality and rainfall runoff (Park et al., 2018; Lu and Ma, 2020; Safari et al., 2020), but the predictive inactivation models for waterborne viruses, which are based on Bayesian regularized regression analyses using water
quality and operational parameters, have not been reported.

In this study, we introduced the Bayesian regularized regression analysis into the predictive water virology to explore appropriate algorithms, which indicated the best prediction performance and were capable of predicting newly obtained datasets and avoiding overfitting to training datasets. We here have focused on inactivation datasets related to ozone, which is one of disinfectants effective for inactivation of waterborne viruses (Roy et al., 1981; Kim et al., 1980) and is utilized at many WWTPs (Hembach et al., 2019; Kharel et al., 2020; Wang et al., 2018). We first collected peer-reviewed articles following a systematic review method and extracted LRV datasets with water quality and operational information from water disinfection experiments using ozone. We applied three regularized regressions (ridge, lasso, and elastic net) and two Bayesian regularized regressions (Bayesian ridge and ARD) to construct the predictive inactivation models, and then verified the prediction performance and overfitting problem based on two validation approaches: (1) the models constructed by training datasets (80% of total dataset) predicted test datasets (20% of total dataset) for evaluation of the prediction performance, and (2) one article was selected and its dataset was used as a test dataset, which had no relation to the training datasets, to confirm whether the model overfitted. We then tried to combine the best regularized regression-based models with HBM, which took experimental errors and virus type-dependent sensitivity into account (Kadoya et al., 2019). Finally, we provided the applicable range of explanatory variables to help readers obtain more precise prediction values and the required disinfection conditions under the water quality level to achieve 4 LRV, which has been recommended by the USEPA (USEPA, 1999).

2. Methods

2.1. Systematic Review and data collection

Peer-reviewed articles relating to virus inactivation using ozone were collected using Google Scholar from April 2019 to June 2020, following PRISMA guidelines (Moher et al., 2009). The keywords input to Google Scholar were "disinfection" and "ozone" combined with keywords of each virus species ("norovirus", "rotavirus", "poliovirus", "adenovirus", "hepatitis A virus", "coxsackievirus" and "echovirus"), and we checked all records published from 1950 to 2020. Dissertations, book chapters, reviews, conference reports and non-English articles were eliminated from the collected articles (first screening). We read whole texts in the first-screened records, and the articles relevant to this study aim, which included the information about the virus LRV and ozone disinfection conditions such as contact time, initial concentration and Ct-value, were then selected from the first-screened records. To establish models using the highly qualified datasets, we assumed that LRVs must be differentiated under varied conditions of disinfection in well-performed experiments, in which water quality and operational parameters were strictly controlled. If the LRVs were same in spite of distinct disinfection conditions, the datasets were excluded. The process of systematic review was fulfilled by three persons.

The LRV was expressed as below:

\[
\text{LRV} = \log_{10}(N_0 / N_t)
\]  

where \(N_0\) is virus concentration at time 0 and \(N_t\) is that at time \(t\). ImageJ software was used to extract numerical datasets of the LRV from figures, such as an inactivation curve (Shneider et al., 2012). Additionally, we obtained water quality and operational parameters such as initial ozone concentration (I) [mg/L], decay rate of
ozone concentration \((k)\) [min\(^{-1}\)], contact time \((t)\) [min], Ct-value \((C)\) [min mg/L], pH \((p)\) [-], temperature \((T)\) [\(^{\circ}\C\)]], assays for measuring virus concentration \((A)\): qPCR or infectivity, \(A2\): plaque assay or MPN method when \(A1\) is infectivity), types of water used for the disinfection experiment \((W)\): puriﬁed or environmental water) and log\(_{10}\) initial virus concentration \((V\) [log\(_{10}\) (genome copy number, PFU or MPN/mL)]). Virus types \((S)\) were also extracted from the collected articles (Table S1), of which rotavirus and coxsackievirus were used as categorical variables because they are the only two types that were tested in the articles. When datasets did not include Ct-values, we calculated them by an integration of the disinfectant decay formula (Eq. (2)):

\[
C(t) = I \exp(-kt) \tag{2}
\]

where \(C(t)\) is ozone concentration at time \(t\) [min], \(I\) is initial concentration of ozone and \(k\) is the decay constant of ozone concentration in water.

2.2. Basic regularized regression analysis

Explanatory variables in the regular regression analyses were \(I, k, t, C, p, T, A1, A2, W, V\) and \(S\), indexed above. The variables \(A1\), \(A2\), \(W\) and \(S\) were binary ones (0: genome copy (in \(A1\)), PFU (in \(A2\)), puriﬁed water and one virus type, 1: infectivity (in \(A1\)), MPN (in \(A2\)), environmental water, another type). Interaction and interaction-quadratic terms were added in a stepwise procedure using the “PolynomialFeatures” function of the scikit-learn library (Pedregosa et al., 2011) in Python 3.7. Regularized regression analysis requires a standardization of explanatory variables \((\text{mean} = 0 \text{ and standard deviation} = 1)\), which was performed by the function “StandardScaler” in scikit-learn library.

The response variable LRVs \((y)\) were expressed as a linear function:

\[
y = X\beta + \epsilon \tag{3}
\]

where \(y\) is a matrix of response variables \((y_1, ..., y_r)\), \(\beta\) is a coeﬃcient matrix \((\beta_1, ..., \beta_p)\), \(\epsilon\) is the matrix of observation errors of \((\epsilon_1, ..., \epsilon_r)\), \(X\) is the design matrix of model variables of \((x_{11}, ..., x_{1p})\).
$x_{(i)}$ is a matrix of $(x_{1p}, ..., x_{nq})$, $p$ is the number of datasets, and $n$ is the number of model variables). Minimizing the sum of squares error $(S_i)$ derives the estimates of coefficients:

$$\min p S_i = \min p \left\{ \frac{1}{2n} \| y - X \beta \|^2 + \lambda R(\beta) \right\} \quad (4)$$

where $\min p$ is a minimization function for $\beta$, $\lambda$ is the regularization parameter and a regularization term $R(\beta)$ determines which regularization methods are used for constructing models. $R(\beta)$ for ridge regression is expressed as $1/2||\beta||^2$ ($||\beta||^2 = \sum_{i=1}^{n} \beta_i^2$) (Hoerl and Kennard, 1970) while for lasso regression $||\beta||$ ($||\beta|| = \sum_{i=1}^{n} |\beta_i|$) is applied (Tibshirani, 1996). Elastic net regression combines the $R(\beta)$ of ridge and lasso regressions $(\alpha||\beta||^2 + (1 - \alpha)||\beta||_1$, where $\alpha$ determines the proportion of the ridge to the lasso regression) (Zou and Hastie, 2006). Suitable values of two parameters ($\lambda$ and $\alpha$) were concurrently searched from $10^{-6}$ to $10^2$ and from 0 to 1, respectively, using the grid search method that enumerated all combinations of those parameters and identify an appropriate combination. The standardization of variables and the addition of regularization term can ensure the homoscedasticity of errors.

### 2.3. Bayesian regularized regression analysis

Regularization parameters are determined not by fitting to data but by the random procedure in the grid search method as described above. Bayesian regularized regression analysis enables us to estimate regularization parameters with coefficients of regression models by fitting to data, and can avoid overfitting by making the model less complex (Tipping, 2001).

The coefficient $\beta$ follows a spherical Gaussian distribution in Bayesian ridge regression analysis, and two regularization parameters ($f$ and $\sigma^2$) are given by gamma distributions as prior distributions (Tipping, 2001; Bishop, 2006). In this case, the likelihood of $y$ is expressed as:

$$p(y | \beta, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \| y - X \beta \|^2 \right\} \quad (5)$$

The prior for $\beta$ is described as:

$$p(\beta | f) = \prod_{i=0}^{p} f(\beta_i | 0, f_i^{-1}) \quad (6)$$

and the prior distributions for $f$ and $\sigma^2$ are:

$$p(f) = \prod_{i=0}^{p} \text{Gamma}(f_i | a, b) \quad (7)$$

$$p(\sigma^2) = \text{Gamma}(\sigma^2 | c, d) \quad (8)$$

where $a$ and $c$ are shape parameters, and $b$ and $d$ are rate parameters of each gamma distribution. These parameters ($a, b, c$ and $d$) were determined from non-informative distributions. Maximizing the logarithmic posterior distribution $(\ln p(\beta | y))$ is equal to the minimization of a regularization term. Automatic relevance determination (ARD) employs the elliptical Gaussian distribution for a prior distribution of $\beta$, and follows the same procedure of Bayesian
ridge regression (Tipping, 2001; Bishop, 2006). These Bayesian regularized regression analyses were also conducted by the scikit-learn library in Python 3.7; codes and source data are available in the supplementary materials.

2.4. Hierarchical bayesian modeling

We adopted probability distributions specific to the objective variables’ LRV based on Akaike’s information criterion (AIC) using the “fitdistplus” packages of R software (Delignette-Muller and Dutang, 2015). The population parameter shaping the probability distribution ($Y$) was described as a logarithmic link function because the LRV datasets followed a Weibull or gamma distribution (Table S2). The logarithmic link function is expressed as below:

$$\log_e Y = \beta_0 + \beta_E X$$

(9)

where $\beta_0$ is an intercept, $\beta_E$ is a matrix of coefficients and $X$ is a design matrix. We assumed that LRVs involved an experimental error, and the matrix of coefficients was indexed “$E$”. The $\beta_E$ was assumed to follow a normal distribution because of experimental error:

$$\beta_E \sim \text{Normal}(\mu_{\text{ST},E}, \sigma_E)$$

(10)

where $\sigma_E$ is a standard deviation generating variations among

---

**Table 3**

| Algorithm      | Polynomial terms |
|----------------|------------------|
| Norovirus      | Automatic relevance determination | Interaction |
| Rotavirus      | Automatic relevance determination | Interaction |
| Poliovirus     | Bayesian ridge   | Interaction  |
| Coxsackievirus | Lasso            | Quadratic    |

---

Fig. 7. Simulation of poliovirus and coxsackievirus LRVs. Heatmaps indicate the LRVs by changing specific explanatory variables, and the white dotted lines are the boundary of 4LRV. The gray cell means the out-of-predictable-range value (negative value) in the models. Temperature $T$ and $\log_{10}$ virus concentration $V$ correspond to summer conditions. ($I$: initial ozone concentration, $k$: decay constant, $t$: contact time, $C$: Ct-value, $A$: assays to measure virus concentration, $W$: water types, $p$: pH, $T$: temperature, $V$: initial virus concentration).
inactivation experiments, and $\mu_{ST[i]}$ is a mean of virus type $i$, which implies the type-dependent sensitivity to ozone. The virus type-dependent parameter $\mu_{ST[i]}$ also follows a normal distribution:

$$
\mu_{ST[i]} \sim \text{Normal}(\mu_0, \sigma_{ST})
$$

where $\mu_0$ is the averaged mean among all virus types and $\sigma_{ST}$ is the standard deviation bearing the type-dependent sensitivity. The hierarchical Bayesian models were constructed by two approaches that employed explanatory variables selected by the best regularized regression models or variance inflation factor (VIF). A multicollinearity problem possibly remained in the HBM, so the models that removed an explanatory variable having more than 10 VIF values were prepared. HBM was conducted on R software version 3.5.0 (https://www.r-project.org/) and Stan (https://mc-stan.org/).

2.5. Model validation

We conducted two trials to determine which regularized regression algorithms exhibited the best prediction performance and escaped overfitting to training datasets. In Trial 1, we randomly separated total datasets into training (80%) and test datasets (20%) to evaluate the prediction powers for interpolation. In Trial 2, the datasets derived from an article were used as test datasets, which resulted in no relationship between training and test datasets and enabled us to understand the predictability for extrapolation. Both trials adopted leave-one-out (LOO) cross-validation, in which temporal models were repeatedly generated using N-1 training datasets ($N$ is the number of training datasets), and a remaining dataset was used for model validations. The coefficients were the averaged values of the generated $N$ models. To evaluate the prediction performance, we calculated mean squared values (MSEs). We also generated a number of the predicted LRVs by changing, averaged values of the generated models. To evaluate the predictability for extrapolation. Both kinds of test datasets were used as test datasets. In Trial 1, the number of explanatory variables was regarded as an explanatory variable in the regularized regression analyses. Taken together, the number of explanatory variables was 9 (norovirus), 8 (rotavirus), 7 (poliovirus) and 7 (coxsackievirus) (Table 1).

3. Results

3.1. Article collection and data characteristics

We first identified 10102 records according to a systematic review method. Dissertations, book chapters, reviews and non-English articles were eliminated from the list of the first-collected articles, which resulted in a decrease to 4058 articles. After excluding the articles not relevant to our study (e.g., no information about the LRV and no water disinfection), 21 articles remained. In detail, the numbers of screened articles for norovirus, rotavirus, poliovirus, coxsackievirus, echovirus and adenovirus were 6, 4, 12, 4, 3 and 2, respectively. The numbers of LRV data point were 68 (norovirus), 65 (rotavirus), 182 (poliovirus), 111 (coxsackievirus), 29 (echovirus) and 29 (adenovirus), respectively (Lim et al., 2010; Brie et al., 2018; Hinneisen et al., 2011; Tondera et al., 2015; Shin and Sobsey, 2003; Thurston-Enriquez et al., 2005; Harakeh and Bulter, 1984a; Harakeh and Bulter, 1984b; Meunier et al., 2006; Vaughn et al., 1987; Finch and Fairbairn, 1991; Farooq and Akhlaque, 1983; Katzenelson and Biedermann, 1976; Moore and Magolin, 1994; Roy et al., 1981; Roy et al., 1982; Engelbrecht et al., 1980; Katzenelson et al., 1979; Emerson et al., 1982; Wang et al., 2018; Wolf et al., 2018). The Scikit-learn library (https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html) and previous studies suggest that more than 50 datasets should be analyzed (Cui and Gong, 2018; Leonard et al., 2017), so we excluded echovirus and adenovirus data from our analyses.

Water quality and operational parameters were also extracted from the screened articles. Initial concentrations of ozone ($I$) are mostly less than 5 mg/L (Fig. 1, I), and the decay constant ($k$) is described only in articles for norovirus (Fig. 1, k). Data points of the contact time ($t$) gather around lower values, but a part of the data for rotavirus and poliovirus was distributed around 10 min (Fig. 1, t). Ct-values ($C$) were not listed in articles about rotavirus and poliovirus (Fig. 1, C). The articles about coxsackievirus listed $C$ but included no descriptions of $I$. The values of pH ($p$) are centered around 7 ± 1 (Fig. 1, p), and almost all disinfection experiments are conducted at room temperature (around 20 °C) or 4 °C (Fig. 1, T). Purified water, like a buffer, was mainly used for disinfection tests (Fig. 1, W). Infectivity assay (plaque assay or MPN method) is one of the main methods to measure virus concentration (Fig. 1, A1 and A2). Initial log virus concentration ranges from approximately four to eight (Fig. 1, V). Since the collected studies used only two types of rotavirus and coxsackievirus (Table 1 and Table S1), the types were regarded as an explanatory variable in the regularized regression analyses. Taken together, the number of explanatory variables was 9 (norovirus), 8 (rotavirus), 7 (poliovirus) and 7 (coxsackievirus) (Table 1).

3.2. The prediction performance of regularized regression models in trial 1

Based on training datasets randomly split from the total dataset, predictive inactivation models for four virus species were constructed. MSEs for both training and test datasets were calculated to identify the best algorithms for the LRV prediction and the required polynomial terms (Fig. 2). The addition of higher terms to the norovirus inactivation model made prediction performances for both datasets better than only linear terms and decreased the differences of MSEs between training and test data (ΔMSE = [MSEtrain - MSEtest]) except for ridge regression. Also, MSEs decreased by adding interaction terms to the rotavirus inactivation models, whereas ΔMSEs in the models with only linear terms were lower. Among models for predicting poliovirus LRV, the ridge regression with polynomial terms indicated the best prediction performance, and Bayesian regularized regression-based models with higher terms also showed smaller MSEs. The coxsackievirus inactivation models were improved very slightly by the addition of polynomial terms, and the ridge-based model with linear terms had the smallest ΔMSE. The algorithms providing better prediction performance were ARD with interaction terms (norovirus and rotavirus) and ridge with interaction terms (poliovirus and coxsackievirus) at this time.

3.3. The prediction performance for new datasets (trial 2)

In Trial 1, there could be relationships between training and test datasets, which made it difficult to confirm the ability to predict newly obtained datasets. Therefore, we selected datasets extracted from an article as test datasets in Trial 2 to validate the models appropriate for predicting new datasets not related to training datasets. The model indicating smaller MSE and ΔMSE is preferable for the predictive inactivation model. In norovirus models, lasso and elastic net with linear terms and ARD with interaction terms indicated smaller MSE and ΔMSE (Fig. 3). Only ARD with interaction terms in rotavirus models seemed to avoid the overfitting problem. All poliovirus models had very small ΔMSE and smaller MSEs, and Bayesian regularized regressions with polynomial terms slightly improved prediction performances. The coxsackievirus model based on lasso regression with quadratic terms had lower MSEs and ΔMSE.
3.4. The best predictive inactivation models and their characteristics

Taken together with the results from Trial 1 and 2, the best predictive inactivation models were ARD with interaction terms for norovirus and rotavirus, Bayesian ridge with interaction terms for poliovirus and lasso with quadratic terms for coxsackievirus (Table 3). Comparative results for observed and predicted LRVs indicated that not only training datasets but also many of the test datasets were predicted well by the best models, but some predicted values (two test data of norovirus in Trial 2, both training and test data of coxsackievirus in Trial 1 and 2) largely deviated from the observed values (Fig. 4). Many of prediction values tended to be on the y = x line and some were randomly deviated from the line (i.e., residuals are distributed randomly), which indicates the independence of residuals. The estimated coefficients of explanatory variables in the best prediction models are displayed in Fig. 5. Except for the poliovirus model, the number of explanatory variables decreased because of the sparse estimations. All models had the variables directly related to the disinfection condition, such as I, C and t, indicating that they are essential to explain LRVs. Other water quality parameters are also included in the best models although some of them are interaction terms.

3.5. No improvement of the prediction performance by HBM

We designed VIF-based and regularized-regression-based hierarchical Bayesian models (Table S3) and compared MSE values among them (Table 2). The hierarchical Bayesian models based on regularized regression analyses showed higher prediction performances than those based on VIF. However, the prediction performances in HBM for test datasets of norovirus, poliovirus and coxsackievirus were lower than the best regularized regression models, indicating that the HBM did not contribute to an improvement of the predictive inactivation models using ozone (Figs. S1 and 2).

3.6. Identification of disinfection conditions achieving 4LRV and unpredictable ranges

The LRVs were generated using the major variables indicated in Fig. 5 to identify the predictable ranges in the models constructed and the disinfection conditions that achieved 4LRV. For all simulations, we set fixed values based on a practical WWTP in Sendai City, Japan, according to the seasonal prevalence of the viruses (summer: poliovirus and coxsackievirus, winter: norovirus and rotavirus). The p slightly affected the 4LRV achievement (Fig. 6a), but as the k increased, the 4LRV was achieved easily (Fig. 6b), which does not seem to be intuitive since a high value of k reduced the effectiveness of ozone disinfection. The rotavirus model clearly demonstrated the out-of-predictable ranges in certain l values (Fig. 6c, d). As well as norovirus, the p had a small effect on the LRV in the poliovirus model (Fig. 7a), but the higher T made it difficult to achieve the 4LRV in a disinfection reactor (Fig. 7b). The linear relationship between I and t was likely to be found (Fig. 7c). The coxsackievirus model also indicated unpredictable regions, and higher p and T made it easy to achieve 4LRV (Fig. 7d, e).

4. Discussion

In this study, we constructed ozone inactivation models predicting virus LRV by regularized regression analyses. The best performances for predicting norovirus and rotavirus LRVs were found in ARD with interaction terms, in which overfitting to training datasets could be avoided (Figs. 2–4 and Table 3). Bayesian ridge with interaction terms and lasso with quadratic terms indicated the best prediction performance for poliovirus and coxsackievirus, respectively, and had the ability to predict new datasets not relevant to the training dataset (Figs. 2–4 and Table 3). However, some test and training datasets have still not been predicted well (Fig. 4), and the HBM did not improve the prediction performances (Table 2). We also provided information about the unpredictable range of some variables and the desirable disinfection condition for the achievement of 4LRV (Figs. 6 and 7).

The conventional kinetics of disinfection for viruses, such as the Chick-Watson, Hom and efficiency factor Hom (EFH) models, use only disinfection parameters (disinfectant concentration, contact time and decay constant) as model parameters (Chick, 1908; Watson, 1908; Hom, 1972; Haas and Joffe, 1994). Boehm et al. established the predictive model for the natural decay of enteric viruses, in which several water quality parameters were used as explanatory variables, and found the larger effects of water quality on natural decay rates (Boehm et al., 2019). The disinfection intensity of ozone is also affected by water quality parameters as the decay rate of ozone increases at higher temperatures (Sohn et al., 2004). Classical models are likely to be inadequate to predict LRVs under varied water quality status, but predictive water virology provides more practical models for LRV prediction using disinfectants because some operational and water quality information can be incorporated into the model (Fig. 5).

The selection of explanatory variables is critical for model construction (Abdul-Wabab et al., 2005). In fact, sparse estimation, which is applied for the predictive inactivation models except for poliovirus, implies a mitigation of overfitting to training datasets (Fig. 3), although an addition of polynomial terms is required to improve prediction performances. Ozone is generally consumed by organic contaminants in water (Park et al., 2001), and the established models lack variables related to disinfectant-consuming substances (e.g., DOC and SS) (Sohn et al., 2004; Du et al., 2020), so the polynomial terms might compensate for the missing but essential variables. However, in this study, only the small number of variables was available and thus an addition of polynomial terms does not always support the prediction, which might attribute to the lack of fit to some datasets. To predict LRVs that are not predicted well with current models (Fig. 4), other types of water quality information needs to be incorporated into our models. The collection of datasets that completely involve water quality information is required to improve the prediction performance, which may be achievable by establishing a widely accessible database that includes datasets provided from laboratories and WWTPs, such as Global Water Pathogen Project (GWPP; https://www.waterpathogens.org/). If water quality information is difficult to prepare, the Ct-value that is calculated using a decay constant can be a substitution of disinfectant consuming matters (Kadoya et al., 2019). Additionally, we are able to construct the robust regularized regression-based models that exclude outliers from an estimation of model coefficients by using other machine learning algorithms (e.g., Random Sample Consensus, Theil-Sen Estimator) (Fischler and Bolles, 1981; Dang et al., 2008).

Predictive water virology helps us identify the water quality and operational conditions that are required for 4LRV, but the models established here have to be updated continuously because the out-of-predictable range has remained large (Figs. 6 and 7). The LRVs for norovirus, rotavirus and poliovirus are relatively stable under near-neutral pH levels (Figs. 6 and 7), which is consistent with previous reports (Lin et al., 2014; Meng et al., 1987; Ward and Ashley, 1979). On the other hand, coxsackievirus LRVs are positively proportional to the pH-value since coxsackievirus is easily inactivated by higher pH (McCready et al., 1979). Although waterborne viruses are easily inactivated with higher temperatures (Moresco et al., 2016;
Biziagos et al., 1988), the model implies that poliovirus is inactivated more easily with lower temperatures, which may be attributable to a data bias (e.g., small datasets using low temperature, in which higher LRVs are observed). Also, as shown in Fig. 6b, an unintuitive result that the LRV became high as decay rate k increased is likely to be explained by the data bias owing to the small number of low k values (Fig. 1). Thus, virus LRV datasets obtained under various water quality conditions need to be collected to eliminate the data bias and expand the predictable range of models.

The predictive inactivation models established here are mostly based on datasets obtained in bench-scale experiments, so the models need to be carefully applied to LRV prediction at real-scale plants. In practice, there can be uncertainty related to hydraulics. For example, heterogeneity of DOC concentration, which is caused by the degree of incomplete mixing, could affect the decay of ozone concentration and the Ct-value. Also, under the mixing effect at a WWTP, water quality parameters need to be expressed using probability distributions (Sin et al., 2009 and 2011). Again, the model limitation must be solved by preparing a high-quality database of virus LRV, in which water quality parameters at WWTP are deposited. Alternatively, the prediction performance is developed by making the regularized regression-based models with explanatory variables that are expressed using probability distributions.

Several descriptions about the CL determination at WWTPs are available in the literature, but we do not have a common and flexible approach for such a determination (Hallowell, 2014). The current models for the LRV prediction help us determine the CL at the disinfection step at any WWTPs. The accepted concentration of virus in effluent is determined so as to correspond to the reference values of tolerable infectious risk and disease burden (e.g., DALY loss per person per year) (USEPA, 1990; WHO, 2017). When a probability distribution of virus concentration in influent is known, the accepted virus concentration in effluent derives the target LRV (Ito et al., 2017). The operational and water quality parameters at WWTPs are monitored, and then we can apply the predictive inactivation models established in this study to determine the CL, which retains the target LRV.

5. Conclusions

Under the concept of predictive water virology, we provided the ozone inactivation models for waterborne viruses, which enable operators at a WWTP and risk assessors to determine the CL at the disinfection reactor. The proposed models are useful for predicting LRVs that are affected by some water quality parameters within the predictable range, and to make the models predictable, we need to continue to collect water quality information with LRV data and then update the models.

Acknowledgement

This study was supported by Gesuido Academic Incubation to Advanced Project, Japan Ministry of Land, Infrastructure, Transport and Tourism, and “The Sanitation Value Chain: Designing Sanitation Systems as Eco-Community Value System” Project, Research Institute for Humanity and Nature (RIHN, Project No. 14200107).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.watres.2021.100093.

References

Abdul-Wahab, S., Balhecht, C.S., Al-Alawi, M., 2005. Principal component and multiple regression analysis in modeling of ground-level ozone and factors affecting its concentrations. Environ. Model. Software 20, 1261–1271.

Adefiose, M.A., NWodo, U.U., Green, E., Okoh, A.I., 2016. Quantitative PCR detection and characterization of human adenovirus, rotavirus and hepatitis A virus in discharged effluents of two wastewater treatment facilities in the Eastern Cape, South Africa. Food. Environ. Virol. 8, 262–274.

Bauri, M., Wilkinson, A.L., Ropa, B., Feldon, K., Snider, C.J., Anand, A., Tallis, G., Boualam, L., Grabovic, V., Avayag, T., Reze, S.M., Mekonnen, D. Zhang, Z., Thorley, B.R., Shimizu, H., Aposto, L.N.G., Takashima, Y. 2019. Circulating vaccine-derived poliovirus and outbreak response – Papua New Guinea, 2018. MMWR 69 (5), 119–120.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer-Verlag, New York, LLC.

Biziagos, E., Passagot, J., Crance, J.-M., Delonc, R. 1988. Long-term survival of hepatitis A virus and poliovirus type 1 in mineral water. Appl. Environ. Microbiol. 54 (11), 2705–2710.

Boehm, A.B., Silverman, A.I., Schriewer, A., Goodwin, K. 2019. Systematic review and meta-analysis of decay rates of waterborne mammalian viruses and collophages in surface waters. Wat. Res. 164 (1), 114989.

Brie, A., Bouadou, N., Missihid, A., Loutreul, J., Bertrand, L., Ganzter, C., 2018. Inactivation of murine norovirus and hepatitis A virus on fresh raspberries by gaseous ozone treatment. Food Microbiol. 70, 1–6.

Chick, H., 1908. An investigation of the laws of disinfection. J. Hyg. 8 (1), 92–158.

Croker, C., Hathaway, S., Marutani, S., Hernandez, M., Cadavid, C., Rajagopalan, S., 2018. A virus infection among adult patients of a mental hospital – Los Angeles County, 2017. Infect. Control Hosp. Epidemiol. 39 (7), 881–882.

Cui, Z., Gong, C., 2018. The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. Neuroimage 178, 622–637.

Dang, X., Peng, H., Wang, X., Zhang, H., 2008. Theil-Sen Estimators in a Multiple Linear Regression Model. Olemiss Edu, 2008. http://home.olemiss.edu/~xddang/papers/MTSE.pdf.

Delignette-Muller, M.L., Dutang, C., 2015. fitteRdist: an R package for fitting probability distributions. J. Stat. Software 64 (4), 1–34.

Du, H., Wu, Y., Wu, H., Li, F. 2020. Effect of ozone pretreatment on characteristics of dissolved organic matter formed in aerobic and anaerobic digestion of waste-activated sludge. Environ. Sci. Pollut. Res.

Emerson, M.A., Sproul, O.J., Buck, C.E., 1982. Ozone inactivation of cell-associated viruses. Appl. Environ. Microbiol. 43 (3), 603–608.

Englebrecht, R.S., Weber, M.J., Saltier, B.L., Schmidt, C.A., 1980. Comparative inactivation of viruses by chlorine. Appl. Environ. Microbiol. 40 (2), 249–256.

Farooq, S., Akhlaque, S., 1983. Comparative Response of Mixed Cultures of Bacteria and Virology. ACM 24 (6), 381–395.

Haas, C.N., Joffe, J., 1994. Disinfection under dynamic conditions: modification of Hom’s model for decay. Environ. Sci. Technol. 28, 1367–1369.

Hallowell, D., 2014. Utilization of Hazard Analysis and Critical Control Points Approach for Evaluating Integrity of Treatment Barriers for Reuse. WaterReuse Foundation. WRRF-09-03 PProject Synopsis.

Harahe, M.S., Bulter, M., 1984a. Factors increasing the ozone inactivation of enteric viruses in effluent. Ozone Sci. Eng. 6 (4), 235–243.

Harahe, M.S., Bulter, M., 1984b. Inactivation of human rotavirus, SA11 and other enteric viruses in effluent by disinfectants. J. Hyg. Camb. 93, 157–163.

Hernbach, N., Alexander, J., Hiller, C., Wieland, A., Schwartz, T., 2019. Dissemination prevention of antibiotic resistant and facultative pathogenic bacteria ultrafiltration and ozone treatment at an urban wastewater treatment plant. Sci. Rep. 9, 12843.

Hirniesen, K.A., Markland, S.M., Knell, K.E., 2011. Ozone inactivation of norovirus surrogates on fresh produce. J. Food Protect. 74 (5), 836–839.
Hoeri, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for non-orthogonal problems. Technometrics 12 (1), 55–67.

Horn, I.W., 1972. Kinetics of chlorine disinfection in an ecosystem. J. Stat. Eng. Dev. 98 (1), 183–193.

Hoque, S.A., Kobayashi, S., Takanashi, S., Anwar, K.S., Watanabe, T., Khamrin, P., Oktisu, S., Hayakawa, S., Ushijima, H., 2018. Role of rotavirus vaccination on an emerging CRPV[8] rotavirus strain causing an outbreak in central Japan. Vaccine 36, 43–49.

Ito, T., Kitajima, M., Kato, T., Ishii, S., Segawa, T., Okabe, S., Sano, D., 2017. Target virus log10 reduction values determined for two reclaimed wastewater irrigation scenarios in Japan based on tolerable annual disease burden. Wat. Res. 125 (15), 438–448.

Kadoya, S., Nishimura, O., Kato, H., Sano, D., 2019. Predictive water virology: hierarchical Bayesian modeling for estimating virus inactivation curve. Water 11 (10), 2187.

Kadoya, S., Nishimura, O., Kato, H., Sano, D., 2020. Regularized regression analyses for the prediction of virus inactivation efficiency by chloramine disinfection. Environ. Sci. Wat. Res. Technol. 11, 218.

Katayama, H., Haramoto, E., Oguma, K., Yamashita, H., Tajima, H., Nakajima, H., Oghaki, S., 2008. One-year monthly quantitative survey of noroviruses, enteroviruses, and adenoviruses in wastewater collected from six plants in Japan. Wat. Res. 42, 1441–1447.

Katzenelson, E., Bierdemann, N., 1976. Disinfection of viruses in sewage by ozone. Wat. Res. 10 (7), 629–631.

Katzenelson, E., Koorner, G., Bierdemann, N., Peleg, M., Shavul, H.L., 1979. Measurement of the inactivation kinetics of poliovirus by ozone in a fast-flow mixer. Appl. Environ. Microbiol. 37 (4), 715–718.

Kharel, S., Stapf, M., Miehe, U., Ekblad, M., Cimbritz, M., Falas, P., Nilsson, J., Ekblad, M., 2018. Thermal and temporal stability of human adenovirus A1–A11, simian rotaviruses and adenoviruses in wastewater collected from six plants in Japan. Appl. Environ. Microbiol. 84 (9), 2614–2619.

Kim, C.K., Gentile, D.M., Sproul, O.J., 1980. Mechanism of ozone inactivation of bacteriophage f2. Appl. Environ. Microbiol. 39 (1), 210–218.

Leonard, J., Flournoy, J., Los angeles, C.P.L., Whitaker, K., 2017. How much motion is too much motion? Determining motion thresholds by sample size for reproducibility in developmental resting-state MRI. Res. Ideas. Outcomes. 3, e12569.

Li, J., Lu, X., Sun, Y., Lin, C., Li, F., Yang, Y., Liang, Z., Jia, L., Chen, L., Jiang, B., Wang, Q., 2018. A swimming pool-associated outbreak of pharyngoconjunctival fever caused by human adenovirus type 4 in Beijing, China. Int. J. Infect. Dis. 75, 89–91.

Lim, A., Kim, J.-M., Lee, J.E., Ko, G.P., 2010. Characterization of ozone disinfection of murine norovirus. Appl. Environ. Microbiol. 76 (4), 1120–1124.

Lin, Y., Fengling, Li, Lianzhu, W., Yuxiu, Z., Yanhua, J., 2014. Function of VP2 protein of murine norovirus. Appl. Environ. Microbiol. 80 (7), 2121–2124.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred reporting items for systematic reviews and meta-analysis: the PRISMA statement. PLoS Med. 6, e1000092.

Park, H.-S., Hwang, T.-M., Kang, J.-W., Choi, H., Oh, H.-J., 2001. Characterization of raw water for the ozone application measuring ozone consumption rate. Wat. Res. 35 (11), 2607–2614.

Park, Y., Kim, M., Pachepsky, Y., Choi, S.-H., Cho, J.-G., Jeon, J., Cho, K.H., 2018. Development of a nowcasting system using machine learning approaches to predict fecal contamination levels as recreational beaches in Korea. J. Environ. Sci. 47, 1094–1102.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, O., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Roy, D., Dong, P.K.Y., Engelbrecht, R.S., Chian, E.S.K., 1981. Mechanisms of enteroviral inactivation by ozone. Appl. Environ. Microbiol. 41 (3), 718–723.

Safar, M.J.S., Arashio, S.R., Mehr, A.D., 2020. Rainfall-runoff modeling through regression in the reproducing kernel Hilbert space algorithm. J. Hydro. 587, 125014.

Sano, D., Amarsi, M., Hata, A., Watanabe, T., Katayama, H., 2016. Risk management of viral infectious diseases in wastewater reclamation and reuse: Review. Environ. Int. 91, 220–229.

Shin, G.-A., Sobsey, M.D., 2003. Reduction of Norwalk virus, poliovirus 1, and bacteriophage MS2 by ozone disinfection of water. Appl. Environ. Microbiol. 69 (7), 3975–3978.

Shindwein, A.D., Rigotto, C., Simoes, C.M.O., Baradi, C.R.M., 2010. Detection of enteric viruses in sewage sludge and treated wastewater effluent. Water Sci. Technol. 61 (2), 537–544.

Shneider, C.A., Rashid, W.S., Eliecer, K.W., 2012. NIH image to ImageJ: 25 years of image analysis. Nat. Methods 9 (7), 655–656.

Silva-Sales, M., Martinez-Puchol, S., Gonzales-Gustavsson, E., Hunadena, A., Girones, R., 2020. High prevalence of rotavirus A in raw sewage from northeast Spain. Viruses 12 (3), 318.

Sin, G., Gernaey, K.V., Neumann, M.B., van Loosdrecht, M.C.M., Gijber, W., 2009. Uncertainty analysis in WWTP model applications: a critical discussion using an example from design. Wat. Res. 43, 2984–2986.

Sin, G., Gernaey, K.V., Neumann, M.B., van Loosdrecht, M.C.M., Gijber, W., 2011. Global sensitivity analysis in wastewater treatment plant model applications: prioritizing sources of uncertainty. Wat. Res. 45, 633–651.

Sohn, J., Amy, G., Cho, J., Lee, Y., Yoon, Y., 2004. Disinfectant decay and disinfection by-products formation model development: chlorination and ozonation by-products. Wat. Res. 38, 2461–2478.

Subahir, M.N., Jeffre, M.S., Hassan, M., Moomad, S.M.G., Suh, F., Ahmed, K., 2019. Norovirus outbreak among students of a boarding school in Kluang, Johor, Malaysia. J. Infect. Def. Cities 13 (4), 274–277.

Thurston-Enriquez, J.A., Haas, C.N., Jacaeglo, J., Gerba, C.P., 2005. Inactivation of enteric adenovirus and feline calicivirus by ozone. Wat. Res. 39, 3650–3656.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Royal. Stat. Soc. Series B. 58, 117–147.

Topping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244.

Tondera, K., Klær, K., Gebhardt, J., Wingender, J., Koch, C., Horstkott, M., Stratham, M., Jurisik, L., Hanza, I.A., Pinekamp, J., 2015. Reducing pathogens in combined sewer overflows using ozonation or UV irradiation. Int. J. Hyg. Environ. Health 218 (8), 731–741.

U. S. Environmental Protection Agency, 1999. Disinfection Principles and Application Guide. U.S. Food & Drug Administration, 1997. HACCP principles and application guidelines. U.S. Environmental Protection Agency, 2006. Guidelines for the Use of Wastewater, Excreta and Greywater in Agriculture and Aquaculture, third ed. World Health Organization, Geneva. WHO, 2006. Guidelines for the Use of Wastewater, Excreta and Greywater in Agriculture and Aquaculture, third ed. World Health Organization, Geneva.

WHO, 2005. Sanitation Safety Planning Manual: For Safe Use and Disposal of Wastewater, Greywater and Excreta. World Health Organization, Geneva.

Zou, H., Hastie, T., 2006. Regularization and variable selection via the elastic net. J. Royal. Stat. Soc. Series B. 67, 301–320.