A Bayesian Approach to Recreational Water Quality Modeling and Cross Validation in the Presence of Measurement Error

Eric Potash[a†]

March 19, 2020

Abstract

Methods for measuring water quality vary in analysis time, precision, availability, and cost. Decision-makers often use predictions from statistical models to compensate for the shortcomings of available measurements. However, these models and analyses of their performance have largely omitted an important source of uncertainty: measurement error. This has led to inefficient models and misestimation of their performance. In this study we show how Bayesian simulations can be used to account for measurement error in cross validation. To that end, we develop a new Bayesian multivariate, multilevel, measurement error model of fecal indicator bacteria at 19 recreational beaches in Chicago. We find that estimates of prediction performance change substantially when accounting for measurement error. We also find that when used to make predictions the Bayesian model is expected to outperform past models with the greatest improvement under limited sampling resources.

Keywords: Bayesian; multilevel; cross validation; recreational water quality; fecal indicator bacteria

Abbreviations: Fecal indicator bacteria (FIB), quantitative polymerase chain reaction (qPCR), cell equivalents (CE), mean squared error (MSE), mean absolute error (MAE), area under receiver operating characteristic curve (AUC), credible interval (CI).

[a] University of Chicago, 1307 E 60th St, Chicago, IL 60637, USA

[†] Corresponding author epotash@uchicago.edu

1 Introduction

Recreational waterways are subject to contamination by bacteria from various sources including stormwater, sewage, and wildlife (Whitman and Nevers 2008). Exposure to contaminated water has been associated with gastrointestinal illness (Prüss 1998). Managers of recreational beaches monitor the presence of fecal indicator bacteria (FIB) as a proxy measure of contamination. To mitigate exposure, managers issue warnings or close sites based on this information.

Traditionally managers relied on culture-based methods which take at least 12-24 hours. Because water quality can change rapidly, decisions based on measurements that are 24 hours delayed are likely to result in unnecessary closures as well as exposure (Kinzelman et al. 2003). More recently, quantitative polymerase chain reaction (qPCR) methods have been employed which can quantify indicator bacteria in less than 2 hours (Noble et al. 2010). However, qPCR testing is substantially more expensive, somewhat less precise, and not as widely available (Whitman, Ge, et al. 2010).

In this study we focus on water quality monitoring in Chicago, which currently manages 19 recreational beaches and has been studied extensively. Statistical models have been used to aid estimation of bacteria levels. For example, when using culture measurements which are 24 hours delayed, covariates such as rainfall, temperature, and sunlight have been shown to improve predictions of current bacteria levels (Shively et al. 2016). In 2017 administrators switched to using the geometric mean of two qPCR measurements per site, which we dub the *empirical model*. Due to the cost of these measurements, the city has proposed sampling half of the sites and using a random forest model to predict levels at the remaining sites (Lucius et al. 2019).

We found that, when evaluating predictive performance of these models, researchers assumed that a measurement (or mean measurement, when multiple measurements were made) represented the true bacteria level at the time the sample was taken (Shively et al. 2016; Lucius et al. 2019). However, it is known that measurement error is substantial (Whitman and Nevers 2004; Whitman, Ge, et al. 2010). The omission of measurement error thus distorts estimates of prediction performance, though the magnitude of this effect is unknown.

In this study we develop a Bayesian multivariate, multilevel, and measurement rror model of FIB levels which we use to make two primary contributions to the literature. First, simulations from the model are used to account for measurement error in cross validation and provide more realistic estimates and uncertainty of the performance of any prediction model. This allows us to re-evaluate prediction models from the literature as well as provide the first estimates of the performance of the currently employed empirical model. Second, we use the Bayesian model to make predictions of FIB levels and we find that its performance is expected to be superior to existing models.

2 Methods

2.1 Bayesian model

We denote the true level of *Enterococci* cell equivalents (CE) per mL on the (natural) log scale by θ_{jt} with $j = 1 \ldots J$ a site index and $t = 1 \ldots T$ a date index.

Next we propose a linear regression (on the log scale) model for these states with coefficients varying by site:

$$\theta_{jt} \sim X_{jt}\beta_j + \epsilon_{jt} \tag{1}$$

where X_{jt} is a vector of $K = 15$ covariates (including an intercept) and β_j is a vector of K coefficient parameters.

The covariates X_{jt} are listed in table 1 and mirror those of (Lucius et al. 2019) with minor changes. First, we excluded forecasts of future meteorological conditions based on a prior belief that, conditional on past conditions, current bacteria levels are independent of future conditions. Second, since our model is linear as opposed to their non-linear random forest, we parameterized day of year as a B-spline with 4 degrees of freedom and a separate wind speed for each cardinal direction. Finally, we reparameterized aggregated covariates to reduce their correlation to speed up model fitting and improve interpretability. For example we included 2-3 day total rainfall instead of their 3 day total rainfall as the former is less correlated with 1 day total rainfall.

Category	**Covariate**
Precipitation	1 day total rainfall 2-3 days total rainfall 1-2 day change in water level
Sunlight	1 day average cloud cover 2-3 day average cloud cover
Wind	1 day average North wind speed 1 day average South wind speed 1 day average East wind speed 1 day average West wind speed
Temporal	Day of year B-spline Weekday indicator

Table 1: Bayesian Model Covariates

While our model is similar to previous regression models of bacteria levels, we add three innovations. First we add a multivariate normal error distribution with covariance matrix Σ to induce correlation in the errors across beaches on a given date t:

$$\epsilon_t \sim \text{Normal}(0, \Sigma) \tag{2}$$

This enables us to combine the measurements at other beaches with those at a given beach in estimating the bacteria level at that beach.

Second we add a multilevel structure on the coefficients, that is we have the second-level model:

$$\beta_{jk} \sim \text{Normal}(\mu_{\beta_k}, \sigma^2_{\beta_k}) \tag{3}$$

This allows us to partially pool information across beaches to more efficiently estimate the coefficients at a given beach (Stow et al. 2009; Cha et al. 2010).

Our final innovation is to incldue an additive and normally distributed measurement error with standard deviation τ (Gronewold et al. 2009). That is, if y_{ijt} is an observation at beach j on date t then it is normally distributed with mean θ_{jt} and variance τ^2:

$$y_{ijt} \sim \text{Normal}(\theta_{jt}, \tau^2) \tag{4}$$

We put the following uninformative priors on these parameters (Gelman et al. 2013). Decomposing Σ into a correlation matrix Ω and a vector of coefficient scales σ

$$\Sigma = \text{diag}(\sigma) \cdot \Omega \cdot \text{diag}(\sigma) \tag{5}$$

we put a uniform prior over Ω and a $\text{Cauchy}_+(0, 1)$ prior on the components of σ. The mean and variance hyperparameters μ_{β_k} and $\sigma^2_{\beta_k}$ are given uninformative $\text{Cauchy}(0, 1)$ and $\text{Cauchy}_+(0, 1)$ priors, respectively. All priors are defined after standardizing all predictors and the outcome.

We denote by Ψ the collection of model parameters (Σ, β, τ). We fit the model using the Markov Chain Monte Carlo software Stan (Carpenter et al. 2017), which uses No-U-Turn sampling (Hoffman and Gelman 2014), an extension Hamiltonian Monte Carlo (Duane et al. 1987). We generated 4 chains with 1000 iterations each, saving the last 500 to produce 2000 draws from the joint posterior parameter distribution. We used the criteria $\hat{R} < 1.05$ and $n_{eff}/N > .001$ to assess mixing (Gelman et al. 2013).

2.2 Estimators

Various estimates of the *Enterococci* CE θ_{jt} at site j on date t have been proposed. Here we review several of these as well as describe our proposed Bayesian estimator. In all cases the input to the estimator on day t is a vector Y_t observations of I observations. In considering various estimators, we include the sampling design, i.e. the number of samples taken at each site each day, as part of the estimator. Thus the same type of estimator (e.g. empirical) under different sampling designs are considered different estimators.

Formally we write a sampling design as a matrix F of dimension $N \times J$ where N is the total number of samples per day. If the i^{th} sample is at site j then $(n, j) = 1$ and the remaining entries are zero. Thus if Y_t is a vector of observations with design F then

$$\mathbb{E}[Y_t] = F\theta_t. \tag{6}$$

Here we consider three sampling designs: one sample per site, two samples per site, and the targeted design of Lucius et al. (2019).

2.2.1 Bayesian estimator

Our Bayesian estimator of θ_t is the posterior mean of θ given measurements Y_t and covariates X_t:

$$\hat{\theta}_t^{\text{bayes}} := \mathbb{E}[\theta_t | Y_t, X_t]. \tag{7}$$

Note that given the design matrix F we can rewrite equation 4 as

$$Y_t \sim \text{Normal}(F\theta_t, \mathbb{1}_I \tau^2) \tag{8}$$

where $\mathbb{1}_I$ is the identity matrix of dimension I. For each posterior sample of parameters Ψ, the posterior $\theta_t | Y_t, X_t, \Psi$ is normal with a closed form mean and variance Eaton, Giovagnoli, and Sebastiani 1996, lemma 3.1.

2.2.2 Empirical estimator

The simplest estimator, which is currently being used in Chicago, is what we call the empirical estimator. When there is one sample y_{jt} per site, the empirical estimate is equal to the observation: $\hat{\theta}_{jt}^{\text{emp}} = y_{jt}$. When the sampling design F includes more than one sample at a site, the empirical estimate is the site-specific (geometric) mean, which can be written in matrix notation as $\hat{\theta}^{\text{emp}} = (FF')^{-1}F'Y$.

2.2.3 Random forest estimator

A random forest model and sampling design was proposed in Lucius et al. (2019). The design uses two measurements at each of 10 beaches and then uses their results together with covariates to predict levels at the remaining beaches. Thus this estimator depends on 20 samples, though in our analysis we include it among the Bayesian and empirical estimators using one sample per site, i.e. 19 samples.

2.2.4 Exceedance predictions

We have presented estimators of continuous FIB levels but Environmental Protection Agency guidance suggests making management decisions based on the binary event of exceeding 1000 CE (United States Environmental Protection Agency 2012). For this an estimator's continuous predictions must be transformed into binary predictions. This is done using a threshold decision rule, i.e. $D(\hat{\theta}) = \hat{\theta} > C$ where the threshold C may depend on the estimator.

Regarding the choice of threshold, decisions currently made in Chicago using the empirical estimator with two samples per site simply use the allowance of 1000 CE as the threshold.

The Bayesian model is richer than the others and so when predicting exceedance we replace the posterior expectation estimator above with the posterior probability of exceedance:

$$\hat{\theta}_t^{\text{bayes}} = \mathbb{E}[\theta_t > \log(1000) | Y_t, X_t] \tag{9}$$

For their random forest, (Lucius et al. 2019) calibrated the threshold to match the specificity of a reference model (Shively et al. 2016). We follow this approach, taking the current empirical estimator using two samples per site as our reference and calibrating thresholds for all other estimators to match its specificity.

2.3 Cross validation

In cross validation we evaluate the fidelity of estimated states $\hat{\theta}$ to the true state θ by a function $L(\theta, \hat{\theta})$, where L is one of various performance measures such as mean squared error (MSE) described below. We include dates t in the *test period* which we chose to be the most recent beach season, 2019. For the random forest and Bayesian estimators, model parameters were fit using data from the *training period*, i.e. prior to 2019.

The challenge in cross validation here is that we never observe θ_t. In the literature, θ_t is often assumed to be exactly equal to the empirical estimate $\hat{\theta}_t^{\text{emp}}$ using all available samples (Shively et al. 2016; Lucius et al. 2019). However, this does not account for uncertainty. For example, this would imply that predictions of the empirical estimator using all available samples are perfect. We term this method of cross validation *naive*, and propose two additional methods: *non-parametric* and *simulated*. The methods are desribed below and summarized in figure 1.

All of the cross validation methods involve producing for each estimator $\hat{\theta}$ an observation vector Y_t according to the estimator's sampling design. In naive and non-parametric validation, when the estimator's design uses fewer measurements than are available in the data, we will produce the vector Y_t by *subsampling*. For example, if an estimator uses one observation per site and there are two in the data, one sample at each site will be selected at random to create Y_t. Simulated validation, on the other hand, simulates these observations from the Bayesian model.

2.3.1 Naive validation

In naive validation we simply assume that the empirical estimate is true: $\theta_t = \hat{\theta}_t^{\text{emp}}$. When the estimator $\hat{\theta}$ uses all available samples, we let $Y = Y^{\text{obs}}$ and for any performance measure L we have a single number $L(\theta, \hat{\theta}(Y))$.

As mentioned above this naive validation method is flawed as can be seen from the fact that when it is used to evaluate the empirical estimator that uses all available samples, the predictions are exactly equal to the "truth" ($\theta = \hat{\theta}$) and so are considered perfect. We expect naive validation to over-estimate the performance of the empirical estimator using fewer samples as well.[1]

When evaluating the empirical and Bayesian estimators using one sample per site we need to subsample the observed measurements, as described above.[2] In this case there are many possible subsamples so we sample among them, resulting in a distribution for $L(\theta, \hat{\theta})$.

Thus in this case we have (by necessity) augmented the usual naive validation procedure (Shively et al. 2016; Lucius et al. 2019) and incorporated uncertainty in the measurements Y. However this does not apply to estimators using all available samples (i.e. two samples per site) and in no case does it account for uncertainty in the true state θ.

2.3.2 Non-parametric validation

Consider the special case of MSE of the empirical estimator with one sample per site. Then, assuming only the measurement error model in equation 4, the expected MSE is simply τ^2. So to estimate the MSE of the empirical estimator under this assumption is equivalent to estimating τ^2, i.e. the sample variance for each beach-day, which can be done using the standard sample variance estimate and then averaged across beaches and dates.

Can this result be extended to include other estimators? For simplicity let us assume that there are two observations per site in the data and the estimator $\hat{\theta}$ uses only one observation per site.[3] Next sample Y_t accordingly and denote the remaining or *hold-out* observations by Y_t^{hold}. In appendix A we derive the following estimate of the MSE:

$$\frac{1}{T}\sum_t |\hat{\theta}(Y_t) - Y_t^{\text{hold}}|^2 - \frac{1}{2}|Y_t - Y_t^{\text{hold}}|^2 \tag{10}$$

which we can bootstrap across t to estimate the sampling distribution.

Note that this result does not make any distributional assumptions. We only assumed that the observations are equal to the true state plus independent measurement error. However, this approach only applies to estimators using strictly fewer observations of each state θ_{jt} than are available. Thus we only use non-parametric validation to estimate the empirical and Bayesian estimators with one sample per site. It is also limited to MSE.[4]

2.3.3 Simulated validation

An alternative approach is a kind of multiple imputation using simulations from our Bayesian model (Rubin 2004). There are two simulation steps. First we simulate the true states θ_t from the posterior distribution $\theta_t | Y_t^{\text{obs}}, X_t$. Next we simulate measurements $Y_t | \theta_t$ according to the design of whatever estimator we are evaluating. Note that whereas parameters for the prediction model are fit using only training data, parameters used for simulation of the true states and measurements are fit using all of the data, including the test set.

The simulated cross validation method assumes that the Bayesian model is correct. Thus one may suspect that using this method to validate predictions from the same model will lead to (optimistically) biased

[1] We do not expect naive validation to be biased for estimators based on prior day *E. Coli* measurements because in this case the measurement error in the predictions is independent of that in the estimated true state. However, naive validation still has the disadvantage here of not reporting uncertainty.

[2] This does not apply to the random forest estimator because, while it uses an average of one sample per site it in fact uses two samples at half the sites.

[3] In fact these results can be generalized to whenever there are more observations available than used by the estimator.

[4] One might try using the bootstrap to estimate the distribution of θ_{jt} which would provide a non-parametric approach to cross validation for any metric or estimator. But with only two samples of each θ_{jt} this bootstrap approach is not viable here.

performance estimates. However note that the simulations incorporate uncertainty in the Bayesian model parameters. And the simulations are made with the Bayesian model fit to all data while predictions are made by a model fit only to past data. Thus the simulation and prediction models are not strictly the same.

The simulated validation method has two advantages. First, unlike the non-parametric validation above which can only estimate MSE, simulation can be used to estimate any prediction performance measure (Table 2). Second, it can be used in scenarios where there are not enough remaining observations to form the hold-out vector as in the non-parametric approach, or even when there are fewer observations in the data than used by the estimator.

We used MSE and mean absolute error (MAE) to evaluate predictions of the continuous FIB level while other metrics evaluate predictions of binary exceedance (section 2.2.4). The area under the receiver operating curve (AUC) evaluates these exceedance predictions as continuous scores. The remaining metrics use binary classifications which are obtained from scores using a threshold as described in section 2.2.4: precision measures the proportion of exceedance predictions which are correct; sensitivity measures the proportion of exceedances which are correctly predicted; accuracy measures overall performance as the proportion of all predictions which are correct.

	Naive	Non-parametric	Simulated
Empirical^1	✓	✓	✓
Random Forest^1	✓		✓
Bayesian^1	✓	✓	✓
Empirical^2	✓		✓
Bayesian^2	✓		✓

(a)

	Naive	Non-parametric	Simulated
Mean squared error	✓	✓	✓
Mean absolute error	✓		✓
AUC	✓		✓
Accuracy	✓		✓
Precision	✓		✓
Sensitivity	✓		✓

(b)

Table 2: Applicability of cross validation methods to estimators (a) and prediction performance measures (b). Estimator superscripts indicate number of samples per site used. Abbreviations: area under receiver operating characteristic curve (AUC).

```
Input: Test set observations $Y^{\text{obs}}$, covariates $X$
       Estimator $\hat{\theta}$ with sampling design $F$
       Performance measure $L$
Result: Estimate of $L(\hat{\theta}, \theta)$
Estimate $\theta$ using the empirical estimator $\theta := \hat{\theta}^{\text{emp}}(Y^{\text{obs}})$
Subsample $Y$ from $Y^{\text{obs}}$ according to $F$
Estimate $\hat{\theta} := \hat{\theta}(Y, X)$
Calculate $L(\hat{\theta}, \theta)$
```

(a) Naive cross validation

```
Input: Test set observations $Y^{\text{obs}}$, covariates $X$
       Estimator $\hat{\theta}$ with sampling design $F$
Result: Estimate of $\text{MSE}(\hat{\theta}, \theta)$
Bootstrap resample $\tilde{Y}^{\text{obs}}$ across $t$ of $Y^{\text{obs}}$
Subsample $\tilde{Y}^{\text{obs}}$ to $Y$ and $Y^{\text{hold}}$ according to $F$
Estimate $\hat{\theta} := \hat{\theta}(Y, X)$
Calculate $\frac{1}{T} \sum_t |\hat{\theta}(Y_t) - Y_t^{\text{hold}}|^2 - \frac{1}{2}|Y_t - Y_t^{\text{hold}}|^2$
```

(b) Non-parametric cross validation

```
Input: Test set observations $Y^{\text{obs}}$, covariates $X$
       Estimator $\hat{\theta}$ with sampling design $F$
       Performance measure $L$
Result: Estimate of $L(\hat{\theta}, \theta)$
Sample Bayesian model parameters $\Psi$ from posterior
Sample $\theta | Y^{\text{obs}}, X, \Psi$
Sample Observations $Y | \theta$ according to $F$
Estimate $\hat{\theta} := \hat{\theta}(Y, X)$
Calculate $L(\hat{\theta}, \theta)$
```

(c) Simulated cross validation

Figure 1: Cross validation methods

3 Results and discussion

3.1 Bayesian model Fit

The Bayesian model was fit using qPCR measurements from the 2015 to 2018 seasons. There were 9329 such observations made at the 19 beaches on 328 dates. A subset of these measurements at a cluster of beaches are shown in figure 2.

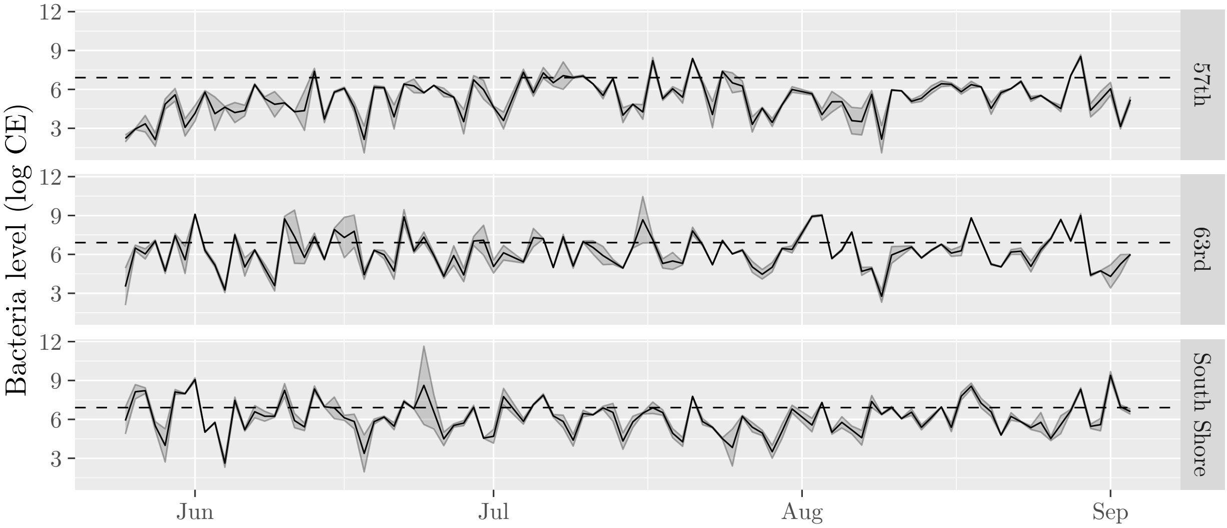


Figure 2: Fecal indicator bacteria measurements at a cluster of 3 of the 19 Chicago beaches during the 2018 beach season. Gray region spans minimum and maximum measurements, solid line connects daily means, i.e. empirical estimates currently used for decision making. Dashed lines indicate action threshold of 1000 cell equivalents (CE).

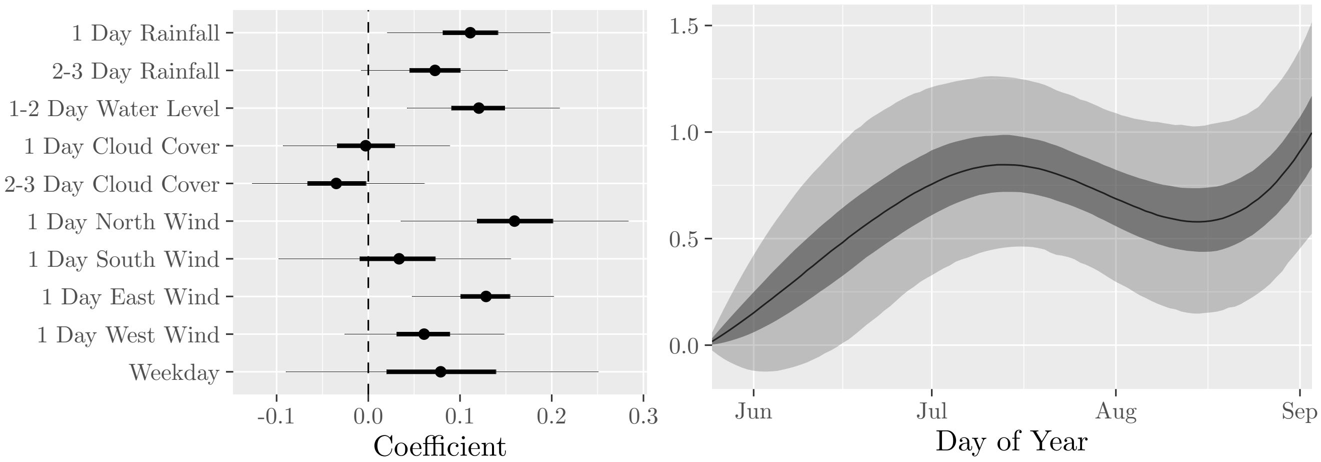
Our MCMC diagnostic criteria were satisfied and there were no divergent transitions. The posterior estimates of the multilevel regression coefficients μ_{β_k} are summarized in figure 3. We note that it makes sense that the precipitation coefficients (rainfall and change in lake level) are positive since stormwater causes combined sewage overflows that discharge into the lake Olyphant and Whitman 2004. However, many of the covariates (precipitation, cloud cover, wind speed, etc.) are correlated so our interpretation of their coefficients is limited.

The posterior distribution of measurement error τ^2 had median 0.77 (95% CI, 0.74 to 0.8). We separately fit the model on E Coli culture measurements and estimated that measurement error to be 0.37 (95% CI, 0.34 to 0.4), which is consistent with the estimate .401 of Whitman and Nevers (2004), Table 2. We conclude that, as previously suggested by Whitman, Ge, et al. (2010), measurement error is greater for qPCR than culture tests, albeit with respect to different units.

3.2 Cross validation

During the 2019 season 3780 qPCR measurements were made over 102 days. We restricted the test period to those days with two samples at each of the 19 beaches so that all estimators could be evaluated. There were 67 such days.

According to the empirical estimates, the median level of indicator bacteria was 92 CE and 4.9% of these beach-days were in exceedance of the 1000 CE threshold. Using the Bayesian estimator the median level was estimated to be 93 CE (95% CI, 36 to 239 CE) and 4.0% (95% CI, 3.6% to 4.2%) of beach days exceeded the threshold.



(a)

(b)

Figure 3: Bayesian model estimated (a) standardized coefficients $\mu_{\beta_k}/\text{sd}(x_k)$ and (b) day of year trend.

3.2.1 Comparing cross validation methods

We started by examining the three cross validation methods on the estimators and metrics where they could all be compared, that is the empirical and Bayesian models with one sample per site and the MSE measure (figure 4).

There are two findings here. First, while naive cross validation shows the empirical estimator performs best,[5] non-parametric cross validation shows the Bayesian estimator performs best. We anticipated above that naive validation would overestimate the performance of the empirical estimator since the same measurement errors are unaccounted for in both the true state in naive validation and the predictions of the empirical estimator. Because non-parametric validation accounts for measurement error, we are inclined to trust its results and dismiss naive validation which is a priori flawed.

Our second finding is a remarkable agreement between non-parametric and simulated MSE estimates. Because non-parametric validation makes few assumptions, this agreement provides evidence to support our use of the simulated validation method to further explore the performance of estimators and metrics for which we do not have a non-parametric method.

3.2.2 Comparing estimators

Next we used the simulated and naive cross validation methods to evaluate all estimators and metrics. The results are shown in figure 5. While naive validation gives point estimates for these measures, simulated validation provides uncertainty estimates which we summarize with 50% and 95% intervals.[6]

The simulated results show that we expect the Bayesian estimator to outperform the others at each number of samples per site, with the greatest improvement in continuous predictions as measured by MSE and MAE. The degree of uncertainty varies across the performance measures with estimates of MAE the most certain and sensitivity the least. The quantification of this uncertainty is an asset of the simulated validation approach.

The discrepancy between simulated and naive validation, first documented in section 3.2.1 above, continues here across more estimators and measures. In most cases the naive estimates are more optimistic (e.g. lower MSE, higher precision, etc.) as we anticipated above. This is explained by the fact that the empirically

[5] Note that in naive validation of the empirical estimator with one observation per site, uncertainty collapses to zero due to symmetry between the observation and hold-out.

[6] The empirical estimator with two samples per site, which uses a binary classification threshold of 1000 CE, had an expected specificity of 97.2%. Classification thresholds for all other estimators were calibrated to match this (see section 2.2.4).

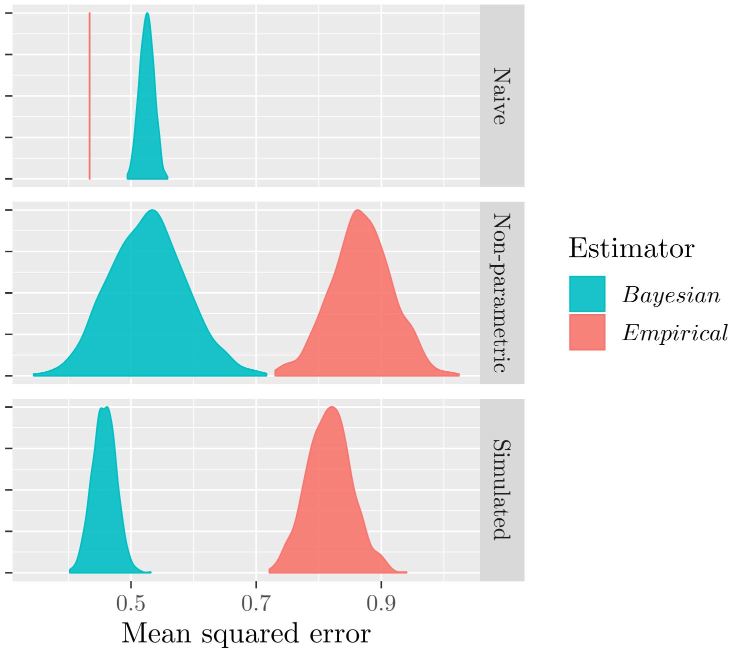


Figure 4: Mean squared error distributions estimated using naive, non-parametric, and simulted cross validation methods for empirical and Bayesian estimators each using one sample per site.

estimated state θ in naive validation includes some of the same measurement error which is used by the estimators to make their predictions.

We used simulated cross validation to estimate the difference in each performance measure between the Bayesian model with 1 sample per site and the currently used empirical model with two samples per site. Note that estimates of prediction performance using simulated validation for different estimators are jointly distributed. This is taken into account to estimate differences in performance. The results are displayed in figure 6.

The Bayesian estimator with one sample per site performs almost as well as the empirical estimator with two samples per site across all metrics. This suggests that using the Bayesian estimator in 2019, beach administrators would have been able to cut sampling resources by half with minor changes in the fidelity of their estimates, and in turn the utility of their decisions based on those estimates.

3.3 Limitations and future work

One natural way to extend the Bayesian model is through autoregression. These models were considered but initial testing (not shown) confirmed previous findings that system dynamics here are too fast for memory effects of daily samples to provide substantial explanatory power (Dorevitch et al. 2017). The parameters in our model are also stationary in time (though they may be updated by periodically refitting the model). If the true parameters are changing, the performance we estimated here may not be representative of future performance. This could be overcome to some extent by modeling the parameters as varying in time (Petris, Petrone, and Campagnoli 2009).

There is some evidence suggesting that measurement error may vary with the bacteria level (Whitman, Ge, et al. 2010). The model could be extended with a heterogeneous measurement error which varies with the bacteria level, as well as other factors such as turbidity.

The data for this study came from Chicago's existing water quality monitoring program. This program relies on two samples per beach per day with each pair collected in the same place and time. However it has

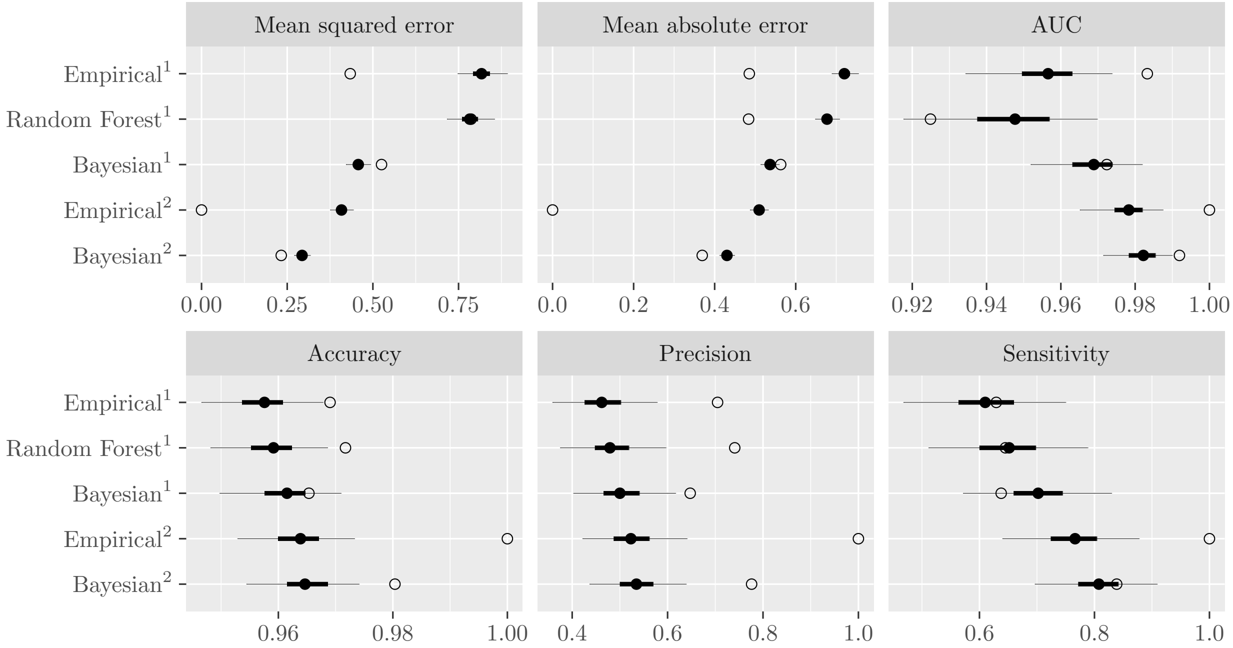


Figure 5: Prediction performance measures. Solid dots and intervals show median and 50% and 95% credible intervals using simulation to account for measurement error. Open dots show naive estimates without accounting for measurement error. Model superscripts indicate the number of measurements per site for the model. Abbreviations: area under receiver operating characteristic curve (AUC).

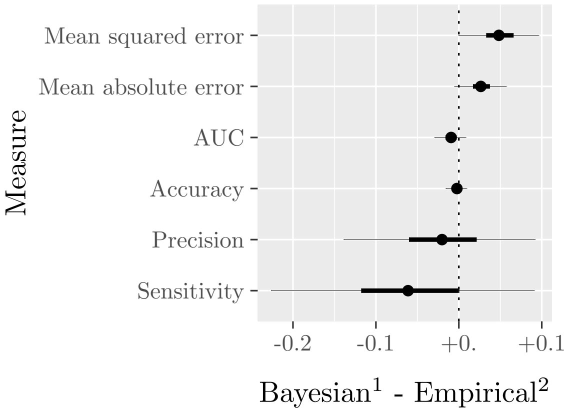


Figure 6: Relative prediction performance. Solid dots and intervals show median and 50% and 95% credible intervals for difference between the Bayesian model with one sample per site and the empirical model with two samples per site using simulation to account for measurement error. Abbreviations: area under receiver operating characteristic curve (AUC).

been shown that there is substantial variation spatially within each beach and temporally within each day (Whitman and Nevers 2004). With the relevant data, our model could be extended to these finer scales.

One innovation of the proposal by Lucius et al. (2019) was to target the allocation of water samples. Our Bayesian model may also be used with a targeted sampling design. Moreover, compared to their random forest model, the Bayesian model is more flexible: it can be used with any sampling design, or combination of designs, without refitting. However, we evaluated the Bayesian model under the targeted design of (Lucius et al. 2019) and found that it underperformed the uniform sampling design (appendix figure B.1). It is possible that a different targeted design may be superior to the uniform one. On the other hand, uniform sampling ensures that measurements are made at every beach which would prove useful if and when a given model is invalidated by non-stationarity.

We found that with Chicago's current sampling design of two samples per beach per day, modeling does not offer substantial improvement over the empirical estimator. However, with one sample per site we demonstrated that statistical models provided substantial improvements and that the Bayesian model likely performs best. There are many other locations where sampling resources are scarce or non-existent. Future work should apply our model to those locations.

Finally, while our model may give better estimates of bacteria levels, to make decisions based on such estimates beach managers will need to incorporate additional information on the consequences of their decisions. This includes information on the effects of human exposure to elevated bacteria levels as well the effects of mitigating actions such as swimming advisories and beach closures.

4 Conclusions

In this study we developed a Bayesian model of FIB levels at recreational beaches in Chicago with three innovations. First, we used a state space approach to explicitly model measurement error. Second, we used a multilevel structure to efficiently estimate regression coefficients that varied across beaches. Finally, the error component of the model had a multivariate normal distribution whose estimated covariance structure enabled incorporating measurements at other beaches in estimating FIB levels at a given beach.

Using this model we made two contributions to the study of water quality monitoring and decision making. First, we showed that the current approach for validating FIB prediction models was naive in taking the observed levels as truth and not accounting for measurement error. We proposed two ways of correcting this: a non-parametric approach for MSE and a parametric approach for any performance measure using simulation from the Bayesian model. We used the non-parametric approach to validate the parametric approach. We found that, across several estimators and measures, estimates of performance using the naive method differed substantially from estimates using simulation to account for measurement error.

Our second contribution was to use the Bayesian model itself to make predictions of FIB levels. For any number of samples per site, we expect the Bayesian model to outperform the empirical model as well as a random forest model from the literature. Moreover, we expect the Bayesian model using one sample per site to have prediction performance on par with the currently deployed empirical model that uses two samples per site.

Acknowledgements

Thanks to Scott Steinschneider for comments on the manuscript as well as Dan Black, Steven Durlauf, Jeff Johnston, Andrew Gelman, Jim Savage, Jackie Shadlen, and Rob Trangucci for useful conversations. Thanks also to Lucius et al. (2019) for transparency about their model and data and assistance in replicating their results.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

A Proofs

Lemma 1. *Let $Y_t = \theta_t + \epsilon$ and $Y_t = \theta_t + \epsilon^{hold}$ where $\epsilon, \epsilon^{hold}$ are independent vectors of length J with mean 0 and variance τ^2 in each component. Then*

$$\mathbb{E}\left[|\hat{\theta}(Y_t) - Y_t^{hold}|^2\right] = \mathbb{E}\left[|\hat{\theta}(Y_t) - \theta|^2\right] + J\tau^2 \tag{11}$$

Proof. We have

$$\begin{aligned}|\hat{\theta}(Y_t) - Y_t^{\text{hold}}|^2 &= |\hat{\theta}(Y_t) - \theta_t - \epsilon^{\text{hold}}|^2 \\ &= |\hat{\theta}(Y_t) - \theta_t|^2 + |\epsilon|^2 - 2\epsilon \cdot (\hat{\theta}(Y_t) - \theta_t).\end{aligned}$$

Next we take expectation over $\epsilon, \epsilon^{\text{hold}}$. The second term becomes $\sum_j \text{Var}[\epsilon_j^{\text{hold}}] = J\tau^2$ and independence of $\epsilon, \epsilon^{\text{hold}}$ implies the third term is zero. □

In particular when using the empirical estimator, $\hat{\theta}^{\text{emp}}(Y_t) = Y_t$ we have
Corollary 2.

$$\mathbb{E}\left[|Y_t - Y_t^{hold}|^2\right] = 2J\tau^2 \tag{12}$$

Combining Lemma 1 and Corollary 2 we have
Corollary 3.

$$MSE(\hat{\theta}(Y_t), \theta_t) = \mathbb{E}\left[|\hat{\theta}(Y_t) - Y_t^{hold}|^2 - \frac{1}{2}|Y_t - Y_t^{hold}|^2\right] \tag{13}$$

Taking the expectation over t, we have
Corollary 4.

$$MSE(\hat{\theta}(Y), \theta) = \mathbb{E}\left[\frac{1}{T}\sum_t |\hat{\theta}(Y_t) - Y_t^{hold}|^2 - \frac{1}{2}|Y_t - Y_t^{hold}|^2\right] \tag{14}$$

B Figures

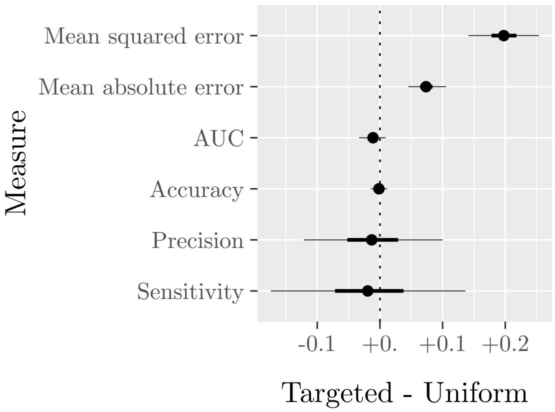


Figure B.1: Prediction performance of Bayesian model with one sample per site, comparison of targeted and uniform sampling designs. Solid dots and intervals show median and 50% and 95% credible intervals for difference between the targeted design and the uniform design using simulation to account for measurement error.

References

Whitman, Richard L and Meredith B Nevers (2008). “Summer E. coli patterns and responses along 23 Chicago beaches”. In: *Environmental science & technology* 42.24, pp. 9217–9224.

Prüss, Annette (1998). “Review of epidemiological studies on health effects from exposure to recreational water”. In: *International journal of epidemiology* 27.1, pp. 1–9.

Kinzelman, Julie et al. (2003). “Enterococci as indicators of Lake Michigan recreational water quality: comparison of two methodologies and their impacts on public health regulatory events”. In: *Appl. Environ. Microbiol.* 69.1, pp. 92–96.

Noble, Rachel T et al. (2010). “Comparison of rapid quantitative PCR-based and conventional culture-based methods for enumeration of Enterococcus spp. and Escherichia coli in recreational waters”. In: *Appl. Environ. Microbiol.* 76.22, pp. 7437–7443.

Whitman, Richard L, Zhongfu Ge, et al. (2010). “Relationship and variation of qPCR and culturable enterococci estimates in ambient surface waters are predictable”. In: *Environmental science & technology* 44.13, pp. 5049–5054.

Shively, Dawn A et al. (2016). “Prototypic automated continuous recreational water quality monitoring of nine Chicago beaches”. In: *Journal of environmental management* 166, pp. 285–293.

Lucius, Nick et al. (2019). “Predicting E. coli concentrations using limited qPCR deployments at Chicago beaches”. In: *Water research X* 2, p. 100016.

Whitman, Richard L and Meredith B Nevers (2004). *Escherichia coli sampling reliability at a frequently closed Chicago beach: monitoring and management implications.*

Stow, Craig A et al. (2009). “Bayesian hierarchical/multilevel models for inference and prediction using cross-system lake data”. In: *Real World Ecology.* Springer, pp. 111–136.

Cha, YoonKyung et al. (2010). “Phosphorus load estimation in the Saginaw River, MI using a Bayesian hierarchical/multilevel model”. In: *Water research* 44.10, pp. 3270–3282.

Gronewold, Andrew D et al. (2009). “Calibrating and validating bacterial water quality models: A Bayesian approach”. In: *Water research* 43.10, pp. 2688–2698.

Gelman, Andrew et al. (2013). *Bayesian data analysis.* Chapman and Hall/CRC.

Carpenter, Bob et al. (2017). “Stan: A probabilistic programming language”. In: *Journal of statistical software* 76.1.

Hoffman, Matthew D and Andrew Gelman (2014). “The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.” In: *Journal of Machine Learning Research* 15.1, pp. 1593–1623.

Duane, Simon et al. (1987). “Hybrid monte carlo”. In: *Physics letters B* 195.2, pp. 216–222.

Eaton, Morris L, Alessandra Giovagnoli, and Paola Sebastiani (1996). “A predictive approach to the Bayesian design problem with application to normal regression models”. In: *Biometrika* 83.1, pp. 111–125.

United States Environmental Protection Agency (2012). *Recreational water quality criteria.*

Rubin, Donald B (2004). *Multiple imputation for nonresponse in surveys.* Vol. 81. John Wiley & Sons.

Olyphant, Greg A and Richard L Whitman (2004). “Elements of a predictive model for determining beach closures on a real time basis: the case of 63rd Street Beach Chicago”. In: *Environmental monitoring and assessment* 98.1-3, pp. 175–190.

Dorevitch, Samuel et al. (2017). “Monitoring urban beaches with qPCR vs. culture measures of fecal indicator bacteria: Implications for public notification”. In: *Environmental Health* 16.1, p. 45.

Petris, Giovanni, Sonia Petrone, and Patrizia Campagnoli (2009). “Dynamic linear models”. In: Springer.