Alaska Department of Environmental Conservation Application of EPA's Virtual Beach Modeling Tool for Kenai River Beaches

Prepared by: J. Petitt[1]

Abstract

During summer 2021, the Alaska Department of Environmental Conservation used the U.S. Environmental Protection Agency's indicator bacteria predictive model building software, Virtual Beach V. 3.0.7, to evaluate its ability to predict exceedances on North and South Kenai beaches. The Alaska Beach Grant Program has monitored Kenai beaches for enterococcus and fecal coliform as the fecal indicator bacteria since 2010, and these bacteria results were compiled alongside concurrent beach environmental data and used to calibrate the predictive models. The models were evaluated for ability to predict exceedances (sensitivity), non-exceedances (specificity), and accuracy. Generalized boosted model and multiple linear regression showed the best fits for the datasets of North Kenai and South Kenai beaches respectively. Future sample events are recommended to ground truth the models.

Basic Waterbody Information

Table 1. Basic Waterbody Information

Assessment Unit ID	AK_B_2030218_002 (North Kenai); AK_B_2030218_003 (South Kenai)
Assessment Unit Name	Kenai River North and South Beaches
Location description	Outlet of Kenai River into Cook Inlet
Water Type	Marine
Area sampled	North and South beaches, areas used by personal use fishery
Time of year sampled	May through August

[1] Nonpoint Source Pollution, Water Quality, Division of Water, Department of Environmental Conservation

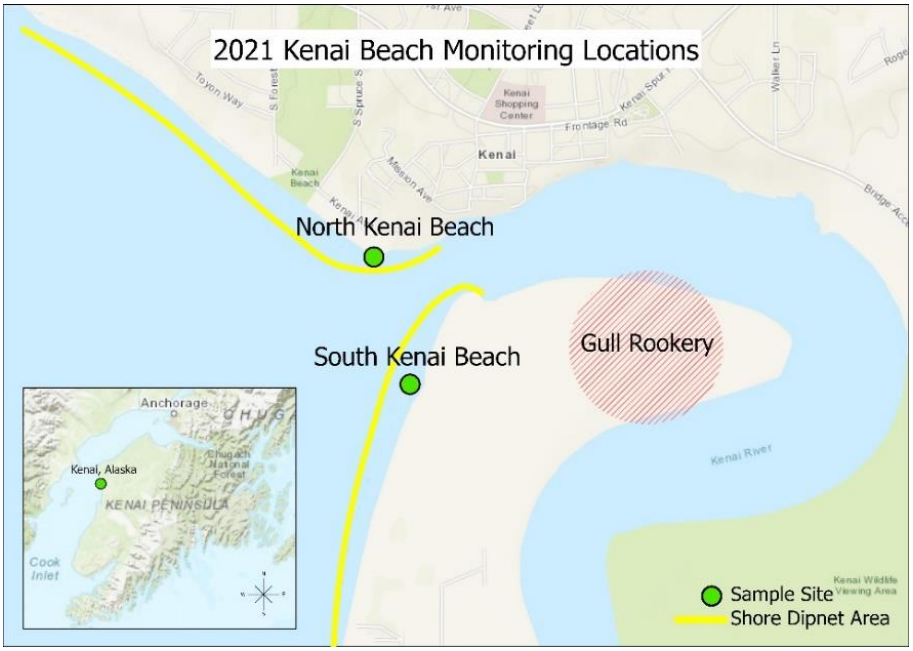

Figure 1. 2021 Kenai Beach Monitoring Sample Sites

Virtual Beach Model Application

Background

The Kenai North and South beaches are positioned at the Kenai River outlet and experience high recreational use during the personal use fishery season which elevates the probability of human contact with water. Kenai Beaches were monitored for bacteria by the Kenai Watershed Forum (KWF) under guidance of the Alaska Department of Environmental Conservation (ADEC) between 2010-2014 and 2018-2020 (KWF 2021). Water quality at both beaches has periodically exceeded the criteria for marine bacteria pathogens (18 AAC 70.020(14)) (ADEC 2021b). The beaches are downstream of a gull rookery, and gulls have been identified as a significant source of bacteria pathogens at the Kenai River beaches through microbial source tracking conducted by KWF in 2019 and 2020. In addition to contributing to elevated bacteria levels, species of gulls are known carriers of anthropogenic antimicrobial resistance contamination (Alstrohm et al. 2019) making monitoring and public outreach necessary for the protection of human health.

Virtual Beach V. 3.0.7 (VB$_3$) is a statistical modelling program that uses site specific environmental conditions to predict fecal indicator bacteria exceedances. Walter Frick and Zhongfu Ge of the Environmental Protection Agency (EPA) developed the first release version of Virtual Beach (VB$_1$), which allowed users to build linear regression models with the support of manual analysis of data sets via visual inspection of data plots and an iterative process of testing, comparing, and evaluating the models (EPA 2020). EPA enhanced the functionality of the subsequent release of VB$_2$, including a map module that provided a method for defining beach orientation allowing for the calculation of wind, wave, and current components into

predictor variables. Additionally, later versions released by EPA added several comparative criteria for the ranking of Multiple Linear Regression (MLR) models: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and sensitivity. EPA improved VB further with the addition of Generalized Boosted Model (GBM) and Partial Least Squares (PLS) modules, allowing for more flexibility in modeling datasets.

Virtual Beach has been successfully used to produce statistical models of bacteria levels in both fresh and saltwater beaches, including Huntington Beach, OH; various Great Lakes' beaches; Miami, FL; and La Monserrate, Puerto Rico (EPA 2020). Predictive modeling is effective because it produces results instantaneously. This may provide a more accurate description of current beach bacteria conditions rather than results that are from samples taken days in advance (Cyterski et al. 2019), and because of the variability in bacteria concentrations that can occur over the incubation period of samples (Brooks et al. 2015). The use of regression models can produce bacteria predictions more frequently while requiring less sampling events, making it a more cost effective method compared to traditional field sampling. ADEC evaluated the applicability of VB$_3$ to Alaskan Beaches during the 2021 recreation season.

Objectives

The objectives for the 2021 VB$_3$ model building project were to:

- Organize historic Kenai beaches bacteria pathogen data alongside concurrent environmental conditions to create a complete dataset on which to calibrate predictive VB$_3$ models.
- Evaluate the fitness of VB$_3$ statistical models for Kenai Beaches, and select the best fit model for each beach for use predicting bacteria conditions for the 2022 recreation season.

Methods

Bacteria pathogen data from North Kenai (2011-2021) and South Kenai (2014-2021) beaches were used to develop the predictive models in VB$_3$. Samples were collected following established protocols under respective Quality Assurance Program Plans (ADEC 2021a). Bacteria pathogen data was organized in an Excel spreadsheet along with corresponding environmental data prior to importing to VB$_3$.

Environmental data obtained from online sources was converted to a standard format in Excel. Historic airport METAR (meteorological aerodrome report) data was retrieved from the Iowa Environmental Mesonet, Kenai River discharge data retrieved from the USGS National Water Information System: Web Interface, and local tide levels retrieved online from NOAA Tide Predictions. Lists of variables retrieved and links for these websites are included in appendix A. Environmental variables with a significant number of missing values as well as dates with missing observations were excluded from the model building exercise. Duplicate rows and blank values were flagged for removal after import to VB$_3$.

Further cleaning and manipulation of data occurred within the VB interface. The orientations of the beaches were used to create alongshore and offshore (A/O) components of wind speed and wind direction, which were added to the list of variables available for modeling. A logarithm with base 10 transformation was applied to the response variable (enterococci bacteria) prior to modeling in VB$_3$ to normalize the distribution of the data.

The response variable was analyzed against the explanatory environmental variables using three different model builders within VB$_3$: MLR, GBM, and PLS. See Appendix B for a description of these modules. One model of each type was built for both Kenai beaches resulting in six total models.

The regulatory standard of each model was manually adjusted lower than the state water quality criteria for marine bacteria pathogens of 130 MPN/100 ml (18 AAC 70.020(14)) to balance the exceedances and non-exceedances in the model training dataset and to improve the predictive performance while attempting to not adversely affect the sensitivity of the model. This practice trains the model to understand how environmental conditions can increase or decrease bacteria concentrations. The adjustment is impermanent and is not used for beach regulatory management decisions. When the predicted values are generated by the model, they will be compared to the state water quality criterion of 130 MPN/100 ml. North Kenai beach favored a lower regulatory standard ranging from 30-60 because fewer exceedances were observed in the training data. South Kenai beach historically has a greater frequency of exceedances, so a regulatory standard ranging from 95-100 was low enough to balance the exceedances and non-exceedances.

Bayesian Information Criterion was used as the criterion for MLR model selection, and the models were cross validated to determine predictive ability. Out of the available criterion, BIC gives the largest penalty for the number of parameters which reduces the tendency MLR has of overfitting to the training dataset. For cross validation, a random sample size of novel observations were set aside, and the models were refit to the remaining training data. The models were then used to make predictions based on the environmental parameters of the novel observations, which were compared to the actual bacteria pathogen observations. The cross validated models were listed by mean squared predicted error (MSEP). The models with the lowest BIC and MSEP were considered the best fit. The decision criterion was manually adjusted at 100 and lowered if necessary, to maximize sensitivity and specificity values.

The GBM calibration began with all possible variables. After the first run, the four variables with the least influence were dropped. The process of rerunning remaining variables and dropping those with the least influence was repeated until the number of variables equaled five to ten percent of the observations for the data set was left, five for North Kenai Beach and four for South Kenai Beach. The decision criterion was automatically set by the program during model calibration to the best ratio of correct exceedances to non-exceedances and was not manually adjusted.

Partial least squares regression also ran all possible variables and listed them by influence. The decision criterion for PLS regression was automatically determined during model calibration and was not manually adjusted.

All models were compared and evaluated by sensitivity (exceedances correctly predicted), specificity (non-exceedances correctly predicted), accuracy, and visual inspection of diagnostic plots. The number of variables ultimately selected for each model was based on a rule of thumb that independent variables should equal five to ten percent of each site's total observations.

Results

Table 4. Results of VB$_3$ model evaluation for North and South Kenai beaches.

Beach Location	Model Evaluation	MLR	GBM	PLS
Kenai North Beach	Specificity	80%	**83%**	83%
	Sensitivity	63%	**53%**	26%
	Accuracy	76%	**77%**	70%
	Most influential variables per model	Tide low/high Relative humidity >24h Air temp	**Relative humidity** **River discharge** **>24h wind speed** **Tide range** **>24h Air temp**	River discharge >24h River-discharge
Kenai South Beach	Specificity	**85%**	79%	90%
	Sensitivity	**68%**	48%	42%
	Accuracy	**80%**	68%	75%
	Most influential variables per model	**Wind direction** **Low tide level** **River discharge** **>24 Air temp** **>24 Wind speed**	River discharge Low tide level Tide range >24h Air temp	River discharge

Specificity is percent non-exceedances correctly predicted; sensitivity is percent exceedances correctly predicted; accuracy is percent total correct predictions.
Bolded columns show best fit model for Kenai beaches.
">24 hour" signifies conditions observed the 24 hours prior to the sample date.

Evaluative conditions for all models calibrated in VB$_3$ shown in Table 4. The North Kenai beach model with the highest accuracy of 77% total correct predictions was built with GBM regression. The South Kenai beach model with the highest accuracy of 85% was built with MLR.

Of the variables used, river discharge, past 24-hour air temperature, and past 24-hour wind speed showed the highest influence on the models. PLS model calibration produced only river discharge and past 24-hour river discharge as influential variables.

Conclusion

The models selected for use during the 2022 recreation season are MLR for South Kenai beach and GBM for North Kenai beach. The sensitivity and specificity of both models are in a usable and normal range for beach bacteria models, comparable to a California bacteria model usability study (Searcy et al. 2018). The Kenai predictive models have accuracies of 75-76% when validated against the training data, so predicted bacteria values for the beaches will also have a 75% chance of being true exceedances/ non-exceedances.

The sample results of Kenai beaches for the 2021 season arrived no sooner than 48 hours after the time of collection, likely causing variance between the result of the sample and the actual bacteria conditions at the beach at the time of issuing public notices. The models were calibrated with environmental conditions that occurred at the time of sampling; therefore, the bacteria values were predicted based on environmental conditions that occurred just before public notices are issued. This could negate the 25% chance of incorrectly predicted exceedances or non-exceedances (with model accuracies of 75%) when compared to the accuracy of beach water sample results provided days after collection.

Kenai River discharge was an influential variable in five of the six models calibrated with the training data. For the PLS models, river discharge was the only influential variable. The high influence that river discharge has on the models could be a result of the beaches proximity to the outlet of the river, which could affect several environmental characteristics like water temperature.

The implementation of Virtual Beach models will allow for more frequent and up to date bacteria predictions at two high use recreational beaches in Kenai, Alaska. The dataset used to calibrate the models should be improved by collecting more bacteria samples in the future. Each variable used for regression requires a minimum of 5-10 unique samples. Because of the variability of environmental conditions, more observations are recommended to reduce risk of fitting the model to outliers. Although we have met the minimum number of observations, more samples would increase the confidence level of predictions. Collecting bacteria samples will also allow for comparison of results to predictions to ground truth the model. With predictive modeling, the reliance on beach monitoring can still be reduced while continuing to supply beach goers with information on the bacteria condition of local beaches.

Recommended Next Steps

The recommended next steps of the Alaska BEACH Program's VB$_3$ project are to:

- Continue to calibrate models during the 2022 recreation season by collecting water samples at Kenai North and South beaches four times during the summer.
- The Alaska BEACH Program will be working with ADEC's webpage programmers to design a format to present bacteria exceedance predictions on ADEC's public webpage beaches.alaska.gov.

- Evaluate use of VB_3 in Southeast Alaska for Ketchikan beaches.
- Informing recreational beach users of best practices to maintain cleanliness of Alaskan beaches and to protect their health remains a top priority. Proper disposal of fish waste and litter reduces attraction of gulls to beaches. Washing after exposure to water and rinsing and cooking fish to an internal temperature of 145^{o}F reduces likelihood of foodborne illness.

Appendix A

Data Retreival Websites

Website Name	Variables Retreived	Link
Iowa State University Environmental Mesonet	Air temperature, dew point, relative humidity, wind speed, wind direction	https://mesonet.agron.iastate.edu/ASOS/
USGS National Water Information System	River discharge	https://waterdata.usgs.gov/nwis
NOAA Tides and Currents	Tide levels, lows, and highs	https://tidesandcurrents.noaa.gov/

Appendix B

Virtual Beach Model Types

VB$_3$ provides users with three model options for fitting explanatory variables to a dependent variable: MLR, GBM, and PLS.

MLR models the relationship between explanatory variables by fitting the observed data to a linear equation. VB$_3$ gives a list of MLR models that are calibrated to a dataset of imported observations. In VB$_3$, the models are ordered by fitness according to a selected criterion, including Akaike information criterion, R squared, and Bayesian information criterion (BIC).

GBM is an example of a random forest model that uses a group of up to 10,000 regression trees, or sets of binary decisions, instead of linear equations (Cyterski et al. 2019). VB$_3$ GBM runs all possible variables and lists them by percent influence.

PLS regression can build models with many variables and decomposes the explanatory variables into mutually orthogonal, or statistically independent, components which are then used as covariates in the regression model (Brooks et al. 2015). Like GBM, VB$_3$ PLS runs all possible variables and lists them by influence.

Both GBM and PLS options automatically conduct 5-fold cross validation meaning the data are split randomly into five sections, and five models are built to predict exceedances on each of the five sections (Cyterski et al. 2019). This way the accuracy of the predictions is tested with novel observations rather than fitting to past observations which could lead to overfitting.

The fitness of predictive models in VB$_3$ are evaluated by the comparison of observed values and predicted values:

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

When a model's output gives more false positives than true positives, the sensitivity will be less than 50%. Adjusting the decision threshold can increase the number of true positives predicted, but doing so may also increase the number of false negatives predicted thereby lowering the specificity. Specificity is the inverse of sensitivity describing the number of true negatives correctly predicted. Virtual Beach gives the user many options to calibrate models and assess statistical parameters.

References

ADEC (Alaska Department of Environmental Conservation). 2020. Water Quality Standards: 18 AAC 70.

ADEC (Alaska Department of Environmental Conservation). 2021a. Kenai BEACH Water Quality Monitoring and Bacteria pathogen Detection: Quality Assurance Project Plan, V. 5.

ADEC (Alaska Department of Environmental Conservation). 2021b. Waterbody Field Report, Kenai River North and South Beaches, Kenai, Alaska.

Ahlstrom C, Ramey A, Woksepp H, Bonnedahl J. 2019. Repeated Detection of Carbapenemase-Producing *Escherichia coli* in Gulls Inhabiting Alaska. Antimicrobial Agents and Chemotherapy 63: 1-4. http://doi.org/10.1128/AAC.00758-19

Brooks W, Corsi S, Fienen M, Carvin M. 2016. Predicting recreational water quality advisories: A comparison of statistical methods. Environmental Modeling & Software 76: 81-94. https://dx.doi.org/10.1016/j.envsoft.2015.10.012

Cyterski M, Brooks W, Galvin M, Wolfe K, Carvin R, Roddick T, Fienen M, Corsi S. 2019. Virtual Beach 3.0.7: User's Guide. Athens, GA: National Exposure Research Laboratory USEPA. https://www.epa.gov/ceam/virtual-beach-307-user-guide

EPA (Environmental Protection Agency). 2020. Virtual Beach (VB). EPA. Retrieved November 1, 2021, from https://www.epa.gov/ceam/virtual-beach-vb.

KWF (Kenai Watershed Forum). 2021. 2020 Kenai Beach Bacteria Monitoring Report. Prepared by Kenai Watershed Forum for Alaska Department of Environmental Conservation, Division of Water, Anchorage.

Searcy R, Taggart M, Gold M, Boehm A. 2018. Implementation of an automated beach water quality nowcast system at ten California oceanic beaches. Journal of Environmental Management 223:633-643. http://doi.org/10.1016/j.jenvman.2018.06.058