EGU23-8832, updated on 29 Mar 2024
https://doi.org/10.5194/egusphere-egu23-8832
EGU General Assembly 2023

Non-parametric Bayesian modeling for risk-based management of Bathing Water Quality

Wolfgang Seis[1,2], Pascale Rouault[1], David Steffelbauer[1], Marie-Claire Ten Veldhius[2], and Gertjan Medema[2]

[1]Kompetenzzentrum Wasser Berlin gGmbH, Hydroinformatics, Berlin, Germany (wolfgang.seis@kompetenz-wasser.de)
[2]Delft University of Technology, Water Management Department, Faculty of Civil Engineering and Geosciences, Stevinweg 1, 2628 CN Delft, the Netherlands

Bayesian non-parametric models are rarely used for predictive modeling of recreational waters. In the present study, we use a Dirichlet Process Gaussian Mixture Model (DPMM) for model-based clustering of hydrologic data collected at three river bathing sites (3 rivers, N = 256, N = 281, N = 1170). The three sites differ in their climatic conditions. *Rivers 1 and 3* are continentally influenced (highly unbalanced dataset with few but severe contamination episodes); *River 2* is more maritime-influenced (regular rainfall leads to balanced data set with regularly occurring pollution episodes); DP models can be used for model-based clustering, where the number of clusters does not have to be pre-defined but is inferred from the dataset itself. For each new observation x_l, the probability of belonging to an already existing cluster as well as the probability of belonging to a new cluster is calculated. We used this property to identify unknown, i.e. high-risk situations, at the individual river sites.

We first applied the DPMM to the available hydraulic training data for model training before conditionally updating a predefined lognormal prior for each cluster, representing the *E.coli* concentration in the river. For prediction, we first evaluated whether a new observation belongs to an existing cluster or whether it constitutes a new cluster. Based on this evaluation, we used either the posterior predictive distribution or the prior predictive distribution for cases where a new cluster was identified. The water quality assessment was subsequently based on the 90th and 95th percentiles of the individual predictive distribution. Model performance was evaluated by means of calculating four criteria: (i) the root mean squared error (RMSE), (ii) the percentage coverage of predictive intervals in relation to the test data (80%), (iii) the detection rate of confirmed contaminations (*E.coli* > 1800 MPN/100 mL), and (iv) the number of predicted bathing days in the test data. The ratio between training and test data was incrementally altered from 10-70%. We compared the DPMM model with four alternative data-driven algorithms: (i) an intercept-only model (zero model), (ii) a multiple linear regression based on stepwise variable selection (stepwise), (iii) a quantile random forest (QRM) and (iv) a Bayesian updating approach, where individual clusters were predetermined manually based on hydrologic characteristics instead of being inferred by the DPMM. The results show that especially for River 1 and 3, only the Bayesian models could predict over 90% of observed contaminations. Through its ability to identify

unknown hydraulic situations and its combination with a prior predictive distribution, the DPMM algorithm can predict high-risk periods without the need to be trained on a dataset that includes this specific contamination information. This is achieved as it identified new hydrologic information as anomalies related to the training set. Thereby, the approach is especially suitable as a precautionary approach for recreational waters, where information-rich datasets are often missing.