

Title: Predicting the Spatial Distribution of Organic Contaminants in an Estuarine System using a Random Forest Approach

Authors: Walsh, E.S.¹, Cantwell, M.G.¹, Kreakie, B.J.¹, Nacci, D.¹.

Abstract: Modeling the magnitude and distribution of estuarine sediment contamination by pollutants of historic (e.g. PCB) and emerging concern (e.g., personal care products, PCP) is often limited by incomplete site knowledge and inadequate sediment contamination sampling. We tested a modeling process using Random Forest to predict low, medium or high sediment concentration of a representative PCP (triclosan). The PCP was primarily sourced from wastewater treatment plant (WWTP) and combined sewer overflow discharges, in subestuaries of Narragansett Bay. We built the models using a limited number and non-random distribution of sediment triclosan measurements. The explanatory variables were accessible and commonly measured data on site features such as bathymetry, sediment composition, and distance to a point source. The sample coordinates were used as a proxy for the urbanization gradient along Narragansett Bay. We found that our models were sensitive to class binning. The model fit based on the model's comparison between predicted and actual class designations improved from 66% to 88% agreement as the data extremes were marginalized, e.g., predictions of very high and medium concentrations improved. The importance ranks of the explanatory variables as measured by the Random Forest algorithm differed among the three models tested. The best-fit model identified the expected variables of top importance which were distance to WWTP and CSO, and sediment composition. They did not include longitude, which was a poor proxy for Narragansett Bay's urbanization gradient compared to latitude. Our results suggested the small number of contaminant measurements limited our models predictions, while better coverage for sediment characteristics could extend the prediction area. Overall, the process we tested appears to be a promising option to extrapolate distribution information from limited spatial concentration data. Further studies are necessary to evaluate the robustness of the predictions and test the transferability to other estuarine systems.

¹US Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, Narragansett, RI, USA