Interpolating missing data and comparing performance of common interpolation techniques from a 30-year water quality dataset

Authors: Wako Bungula[1], Danelle M. Larson[2], Killian Davis[3], Richard A. Erickson[4], Amber Lee[3], Casey McKean[3], Frederick Miller[3], Alaina Stockdill[3], Enrika Hlavacek[5]

[1] University of Wisconsin La-Crosse, Department of Mathematics and Statistics, 1027 Cowley Hall, La Crosse, WI 54603

[2] U.S. Geological Survey, 2630 Fanta Reed Road, La Crosse, WI 54603; dmlarson@usgs.gov; no fax; ORCID: 0000-0001-6349-6267; [i] Corresponding Author: dmlarson@usgs.gov

[3] University of Wisconsin La-Crosse, Research Experience for Undergraduates Program, 1027 Cowley Hall, La Crosse, WI 54603

[4] U.S. Geological Survey, 2630 Fanta Reed Road, La Crosse, WI 54603; ORCID: 0000-0003-4649-482X

[5] U.S. Geological Survey, 2630 Fanta Reed Road, La Crosse, WI 54603; ORCID: 0000-0002-9872-2305

Missing data is a common issue, especially in long-term and extensive ecological data sets. The Upper Mississippi River Restoration Program collects long-term water quality data to monitor environmental conditions and to take restoration actions. Given the extent of this dataset (200,000 sampling sites over 30 years), missingness is an issue for three variables: total nitrogen, total phosphorus, and water velocity. Data interpolation (i.e., estimating missing values using known values within the data set) can relieve issues, and so we compared the performance of five interpolation methods using several error metrics. For all three water quality variables, the interpolation method ‘random forests’ outperformed the methods of kriging, polynomial regression, regression trees, and inverse distance weighting. Total phosphorus had very high prediction accuracy (percent error of 2% and mean absolute error of 0.03 mg/L-TP). No interpolation method accurately predicted the water velocity variable (percent errors ranged from 100–175%), indicating that more frequent river sampling and hydraulic models may be better suited for improving velocity predictions. In the current era of ‘big data’, interpolation becomes an imperative step prior to ecological analyses. Our research provides an approach and data analysis scripts that allow intercomparisons of interpolation methods for many applications and contexts.