Deciphering factors controlling groundwater arsenic spatial variability in Bangladesh

ZHEN TAN[1,2], QIANG YANG[3],
CHUNMIAO ZHENG[1,2], AND YAN ZHENG[2,3]

[1] Peking University, Institute of Water Sciences, Beijing, China (tanzhen@pku.edu.cn)
[2] Southern University of Science and Technology, School of Env. Sci. & Engr., Shenzhen, China
[3] Columbia University of New York, Lamont-Doherty Earth Observatory, Palisades, NY, United States

Introduction and Method

Factors controlling groundwater arsenic spatial variability in Bangladesh are investigated using two weak-learner ensemble models (random forest and boosted regression trees, RF and BRT) and a traditional linear model (backward logistic regression, BLR). All models considered hydrochemical, soil, geomorphological and hydrologeological spatial parameters and hydrology spatial parameters. Three major rivers of Bangladesh, i.e. Brahmatrupa, Ganges, and Meghna, along with uplifted terraces, demarcated Bangladesh to four regions as modeling domain: the South, Northeast, Northwest and the Tract. Models were bulit for the entire nation as well as the four regions to identify and to compare the controlling factors for the probability of detecting arsenic in groundwater in exceedance of any of the three threshold values: 5 μg/L to approximate median value of the dataset, 10 μg/L to represent the WHO guideline value, and 50 μg/L to represent Bangladesh drinking water standard. Oversampling is used to reduce the model bias caused by imbalanced data.

Results

Iron and phosphorus are the important parameters of different regional models and national models with all parameters, where are depth, distance to three major river, elevation, irrigation or Indo-Burman Ranges in differents models only using spatial parameters. Ensemble models outperformed BLR at three thresholds in all predictive performance measures. BRT model have similar performence with RF, and BRT yielded better robustness prediction than RF. Pre-oversampling reduced difference between sensitivity and specificity by 51.6% on average facing imbalanced data without decline of accuracy. The probabilities of arsenic occurrence in groundwater is best mapped via the BRT model.