IDENTIFICATION OF CONTAMINANT SOURCE USING TRUNCATED BREAKTHROUGH CURVES IN RIVERS

Siyoon Kwon[1], Il Won Seo[2*], Hyoseob Noh[3]

[1,2,3]*Dept. of Civil & Environmental Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, South Korea*

*Correspondence YP: seoilwon@snu.ac.kr

Abstract When accidental spills of the contaminant occur in natural rivers, it is necessary to identify the contaminant source to minimize the damage from the contamination. Thus, we proposed new data-driven models to identify the pollution source location and spill mass in the river system. The models used the Breakthrough Curve (BTC) obtained from the downstream section as input data. Besides, both Whole Breakthrough Curve (WBTC) and Trunced Breakthrough Curve (TBTC) are capable of input data, which enhances the efficiency of accident response. For developing the models, a large number of numerical simulations under many spill cases with various hydraulic conditions were implemented by the Transient Storage zone Model (TSM). Finally, the performances of six Machine Learning (ML) models for WBTC and TBTC were compared.

Keywords: Contaminant Source identification; Contaminant transport; Transient storage zone; Machine learning.

Methods

1. Contaminant spill scenarios
- Numerical model: Transient Storage zone Model (TSM)
- Scheme: finite difference method and the Crank-Nicolson (approximation of the spatial and temporal derivatives)
- TSM parameters: empirical equations from Principle Components Regression (PCR)
- Validation result with field tracer data: R2 0.9
- Model domain: 39 km with 30 potential spill locations in Gam creek, South Korea

2. Machien Learning (ML) model
- Comparision of 6 ML models
- Input variables: 21 WBTC features vs 8 TBTC features (Ex. Std, Skewness… / rising limb slope, area of rising limb derivative…)

Type	Model	Feature importance	Description
Tree based model	Decision tree	Reduction of Gini index	Bagging
	Random Forest		Bagging, ensemble
	XGBoost		Boosting
Linear model	SVM-linear	Weight coefficient	Hyperplane by linear function
	Ridge regression		L2 regularization
Kernel model (nonlinear)	SVM-RBF		Hyperplane by RBF function

Table 1: Types of ML models

Results

1. Spill location
- WBTC: all non-linear model showed accuracy over 0.9 R^2
- TBTC: only SVM-RBF showed 0.71 R^2

2. Spill mass
- WBTC, TBTC: all non-linear model showed accuracy over 0.9 R^2

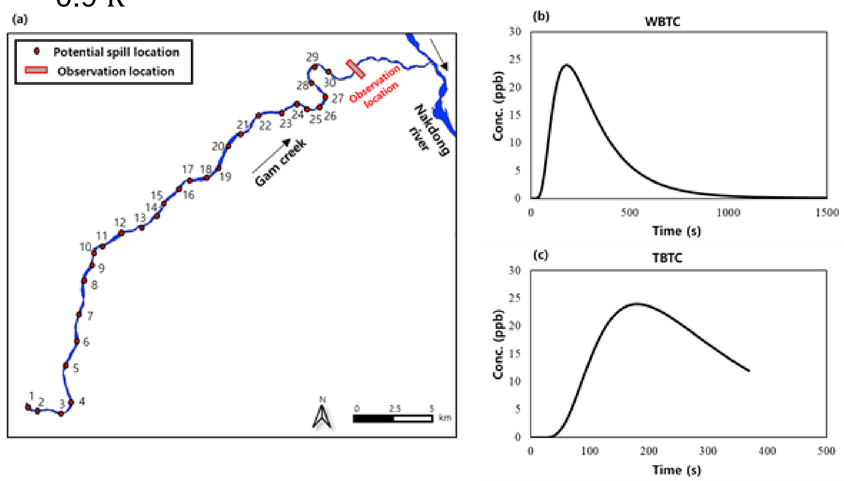

Figure 1: (a) Model domain, (b) WBTC, and (c) TBTC

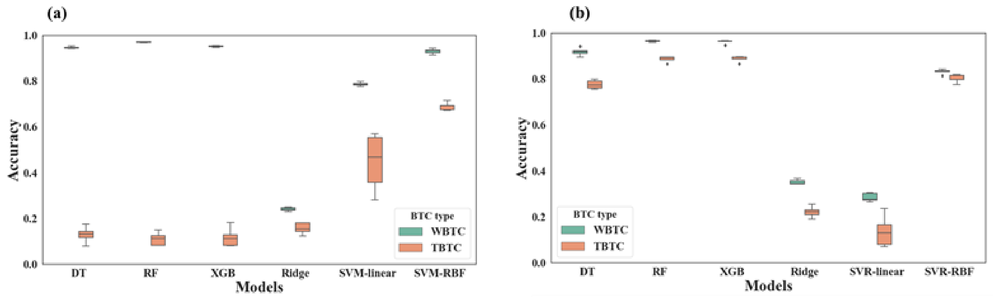

Figure 2: 5-folds cross validation results of (a) spill location and (b) spill mass

Conclusions

DT based models outperformed other models when WBTC is used to predict the spill location and mass. However the tree-based models failed to predict spill location using TBTC features. Only SVM-RBF showed accuracy over 0.7. The reason can be explained by the relative importance of each ML algorithms. Since the tree-based models highly depend on the slope of the tail calculated only in WBTC, these models are difficult to predict well with less contributed TBTC features. In the case of SVM, the overall feature importance is more distributed than the tree-based model. Thus, using only the TBTC features significantly contributes to the prediction. Consequently, nonlinear-SVM is the most optimal model for both spill location and mass.

Bibliography

Singh, R.M., Datta, B. & Jain, A. (2004). Identification of unknown groundwater pollution sources using artificial neural networks. J Water Res PL Asce 130: 506–514.

Srivastava, D., Singh, R.M. (2014). Breakthrough curves characterization and identification of an unknown pollution source in groundwater system using an artificial neural network (ANN). Environ Forensic 15(2):175–189.

Zhang, S. & Xin, X. (2017). Pollutant source identification model for water pollution incidents in small straight rivers based on genetic algorithm. Applied Water Science 7(4): 1955–1963.