THEORETICALLY INFORMED MACHINE LEARNING (ML) MODELS FOR PREDICTING PESTICIDE LEACHING IN NEW YORK (August 2025, NABEC)

Isaiah Guenther, Tammo S. Steenhuis, and others of Soil & Water Lab,
Biological and Environmental Engineering Dept. - Cornell University
With and for New York State Department of Environmental Conservation (NYSDEC), Bureau of Pesticide Management

BACKGROUND

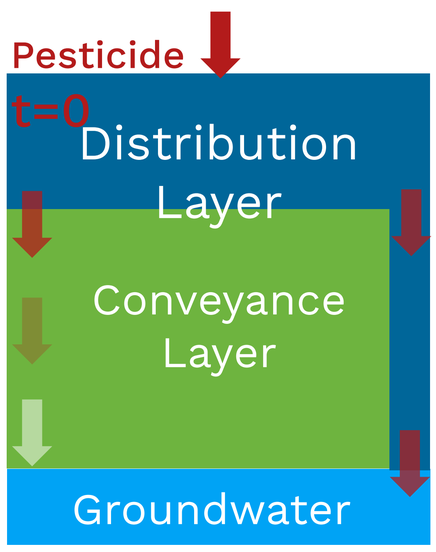
- Eastern Long Island 1970s insecticides in groundwater.
- Upstate diversely different from Long Island.
- Many eclectic pesticides and uses. Complex to forecast groundwater occurrence.
- Many non-detect results in monitoring.

Objectives

- Develop ML model that predicts leaching.
- Compare ML to theoretical model, same inputs.

MODELS AND DATA

Theoretical Groundwater Ubiquity Score (TGUS) [1]



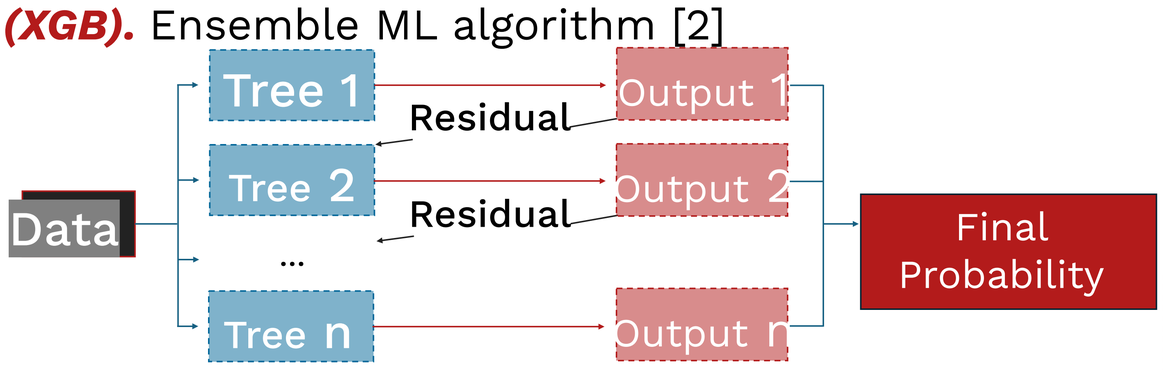
$$t_{LRP} = \frac{t_{\frac{1}{2}}}{0.69} ln\left(\frac{\lambda\Phi}{K_{oc} f_{oc} \rho}\right)$$

$t_{\frac{1}{2}} = soil\ halflife [days],$
$\lambda = preferential\ flow\ fraction$

$$\Phi, = \frac{application\ rate}{detection\ limit\ * distribution\ layer\ thickness}$$

K_{oc}
$= organic\ carbon\ sorption\ partition\ coef. \left[\frac{ml}{g}\right],$
$f_{oc} = soil\ organic\ carbon\ fraction,$
$\rho = soil\ bulk\ density\ [\frac{Mg}{m^3}]$

Machine Learning – Extreme Gradient Boosting (XGB). Ensemble ML algorithm [2]



XGB classification flowchart. Final class probability is the sigmoid function applied to the sum of N log-odds outputs from each tree.

Data sources and selection

- Long Island groundwater analytical data (USGS [3]), 1832 cases and NYSDEC lab upstate, 761 cases.
- USDA NRCS soil surveys, NYSDEC Pesticide Sales and Use Reporting (PSUR), pesticide product labels, and pesticide properties databases and literature.
- Cases omitted for missing chem properties and Long Island analytes with pesticide use+sales <50 lb over 6 years in zip code.

MODEL APPLICATION

TGUS	Predict leaching with t_{LRP} : $pred. = \begin{cases} detect, & t_{LRP} \geq 100\ days \\ nondetect, & t_{LRP} < 100\ days \end{cases}$
ML: XGB	1. CV train on 80% stratified data split. tune hyperparameters/threshold. 2. Select "fold" with max F1. 3. Test prediction performance on 20% held-out data set. 4. Repeat 50 iterations steps 2-3, record all predictions and metrics

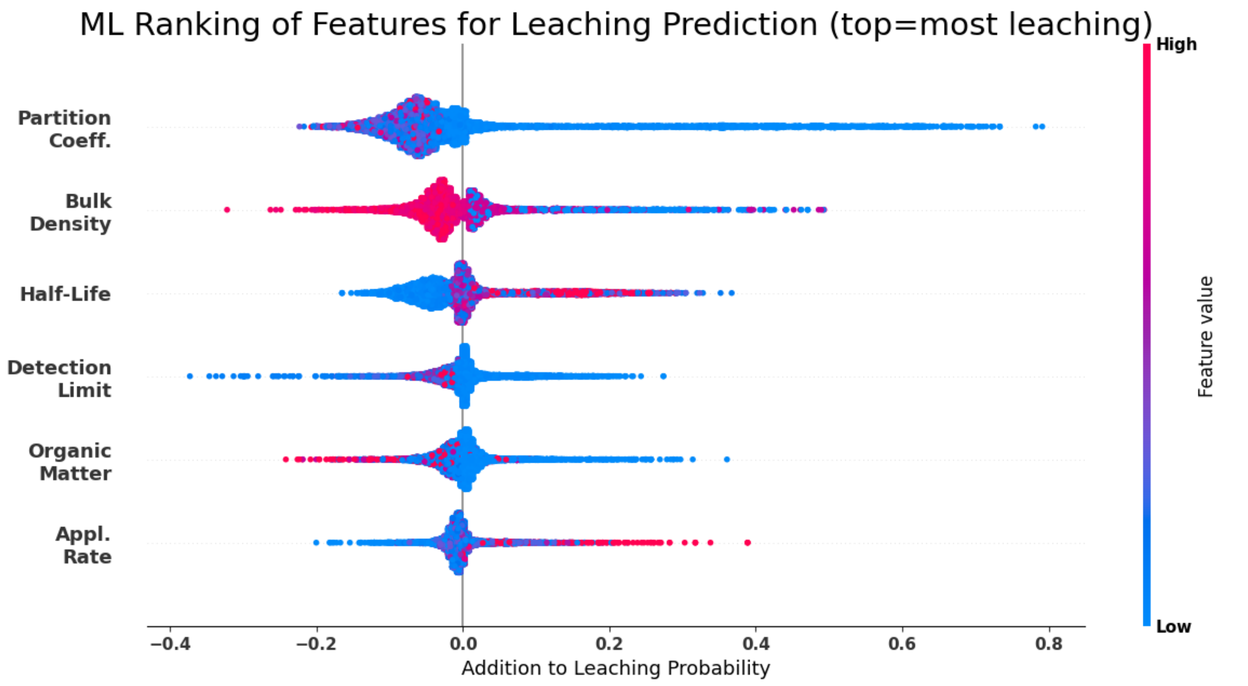
RESULTS

Model	Peformance Metrics (train/test)		
	Precision [%]	Recall [%]	F_β [%]
TGUS Upstate (untrained)	-/20	-/99	-/56
XGB, upstate 6 vars	77/83	80/95	79/92
TGUS Long Isl. (untrained)	-/21	-/74	-/49
XGB, Long Island 6 vars	52/58	79/89	71/80

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \qquad F_\beta = (1 + \beta^2) * \frac{Precision\ * Recall}{(\beta^2 * Precision) + Recall} \qquad \beta = 2$$

Performance metrics from model outcomes: true positive (TP), false positive (FP), true negative (TN), false negative (FN).

Shapley beeswarm plot, six variable XGB model, upstate 20% testing dataset



left=lower effect, right=higher, blue=low variable, red=high variable

Conclusions

- XGB outperformed TGUS in most metrics, both locales.
- XGB learned patterns consistently with TGUS theory. This may corroborate TGUS method.

Considerations

- Limited datasets and empirical scope of ML tools.
- No analysis of feature interactions (yet).
- Two of eight TGUS params not readily available.

REFERENCES AND CREDITS

[1] Steenhuis, T. S., et. al. 2024.. J. of Hydrology and Hydromechanics. DOI: 10.2478/johh-2024-0016
[2] Chen, T. & Guestrin, C. 2016. DOI: 10.48550/arXiv.1603.02754
[3] Fisher, I. J., et. al. 2021. DOI: 10.1016/j.scitotenv.2020.141895

Algorithm and ML coding: Isaiah Guenther, Steven Pacenka
Database:: Steven Pacenka, Naaran Brindt, Xin Shen, Xingliang Cao, Xiyue Luo, Yuchen Tao
TGUS theoretical formula: Tammo Steenhuis
Principal investigators: Tammo Steenhuis, Brian Richards
Poster: Isaiah Guenther, Steven Pacenka, Yuchen Tao.
Special thanks to Scott Steinchneider for ML advice.

Soilandwaterlab@cornell.edu https://soilandwaterlab.cornell.edu/