Identification of the Most Influential Watershed Characteristics for Bifenthrin Contamination in Stream Sediments in California

*Dan Wang, *Robert Budd, *Christopher DeMars, ^Bryn Philips, ^Brian Anderson, *Nan Singhasemanon, *Kean S. Goh

*California Department of Pesticide Regulation,
^Department of Environmental Toxicology, University of California, Davis, Granite Canyon Laboratory

INTRODUCTION

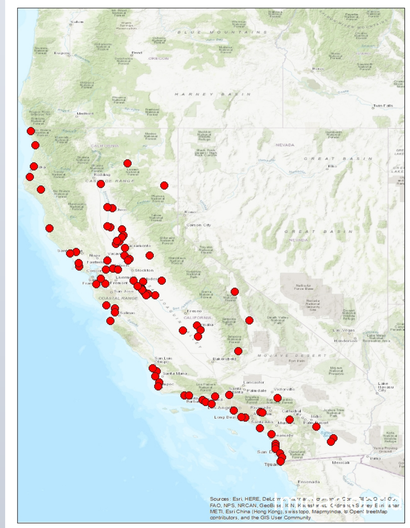
Bifenthrin is an insecticide under intensive investigation in California because of its high use and high toxicity to aquatic invertebrates. This pesticide has high affinity for soil and sediment particulate matter. Its fate and transport in the aquatic system are largely associated with the movement of sediments and appear to be greatly influenced by the hydrological, geomorphological and anthropogenic characteristics of the contributing watershed. This study uses a tree-based statistical learning method, random forests model, to identify the most influential watershed characteristics for bifenthrin concentration in stream sediments. Random forests models were chosen because: (1) they work when there are complex interactions among those watershed characteristics as well as possible non-linear relationships between the characteristics and the bifenthrin concentrations, and (2) they tend not to overfit the tree, i.e., corresponds too closely to training data and may fail to predict.

MATERIALS AND METHODS

Monitoring data

California State Water Resource Control Board's Surface Water Ambient Monitoring Program (SWAMP) - Stream Pollution Trends (SPoT) Monitoring[1]

- 87 sites (see map to the right), 1 dry season sample per year, 4–8 years, 2008–2015
- 564 samples, 327 in range of quantification (ROQ), 237 left-censored by reporting or detection limits
- 36 counties in 9 regions: Central Coast, Central Valley, Colorado River Basin, Lahontan, Los Angeles, North Coast, San Diego, San Francisco Bay, Santa Ana Region



Watershed characteristics

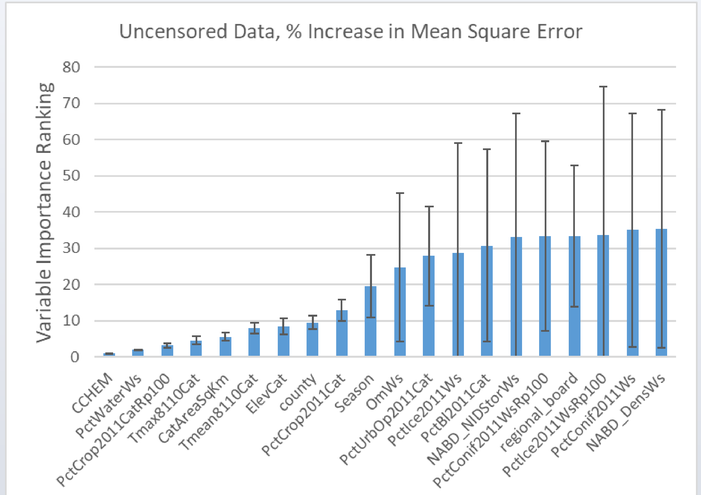
The characteristics of the immediate catchment as well as entire contributing watershed to each monitoring site are extracted from the StreamCat database[2] after projecting the coordinates of the sites to the NHDPlus HUC14 catchments[3].

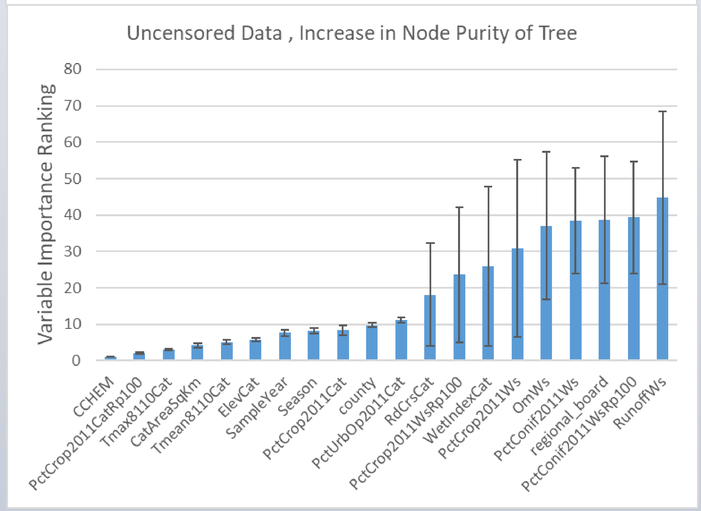
- Selected 197 variables: potential to impact bifenthrin fate and transport, do not have missing data; NLCD 2011 land use data used for average condition in 2008–2015

Random Forests Method, two R packages

- Two R packages: 'randomForest' fits a model for uncensored data and optimizes by concentration. 'randomForestSRC' fits a model for right-censored data, the left-censored data were first transformed by taking the reciprocal; optimizes by cumulative hazard function
- Variable importance ranked according to (1) change in regression or prediction error, and (2) effect on tree structure, node purity or minimal depth[4,5] for uncensored and censored data, respectively
- Method reproducibility: run the model 100 times, lower ranking indicates higher importance, top 20 variables plotted with standard deviation as error bar, lower standard deviation indicates higher reproducibility

VARIABLE RANKING RESULTS: uncensored data in ROQ





RESULTS

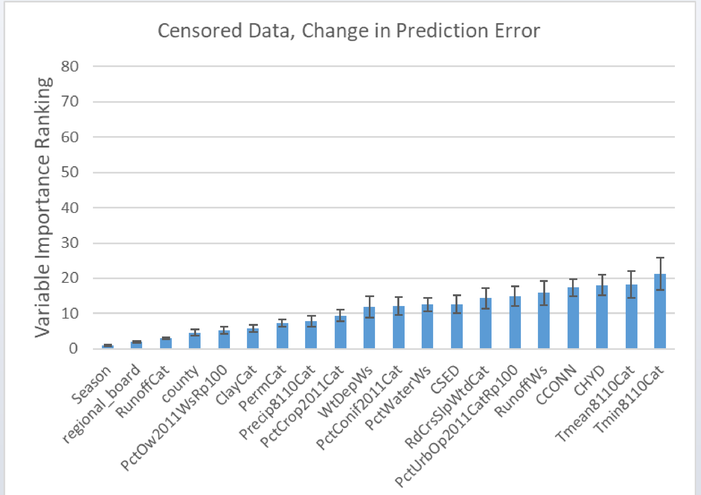
Reproducibility:

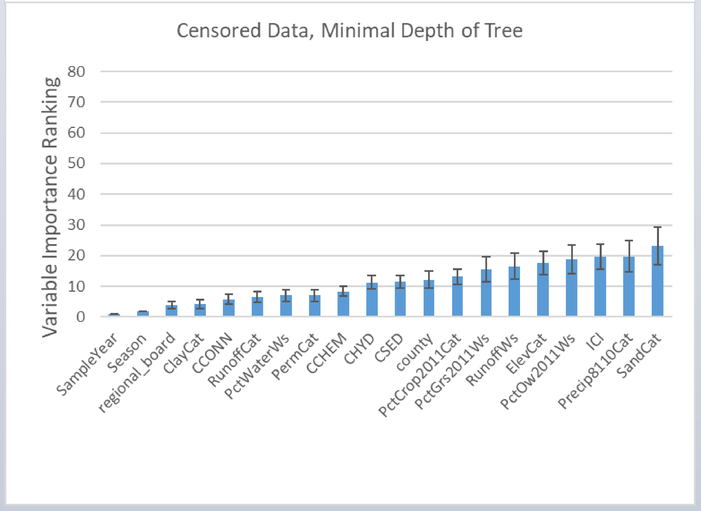
- censored data method had higher reproducibility than uncensored method
- ~ top 10 variables of uncensored method have distinctively higher reproducibility than following variables

Top 20 most important variables selected:

- by 4 methods: county, PctCrop2011Cat, regional, season
- by 2 censored and 1 uncensored methods: PctWaterWs, RunoffWs
- by 1 censored and 2 uncensored methods: Tmean8110Cat, CCHEM ElevCat
- by 2 censored methods: CCONN, CHYD, ClayCat, CSED, PermCat, Precip8110Cat, RunoffCat
- by 2 uncensored methods: CatAreaSqKm, OmWs, PctConif2011Ws/WsRp100, PctCrop2011CatRp100, PctUrbOp2011Cat, Tmax8110Cat
- 10 variables selected by one censored method while 9 variables by one uncensored method

VARIABLE RANKING RESULTS: all data with censoring





DISCUSSION

- Disagreement in the four sets of ranking results is due to differences in input data, model optimization goal, and variable importance matrix
- Next steps: (1) test sensitivity to input data, (2) understand effect of optimization goal: concentration vs. distribution, (3) compare ranking methods, and (4) reveal the driving forces behind the 4 variables selected by all 4 methods (%crop land, county/region, season) by expanding the model and including pesticide use, flow, and additional weather data, etc.

REFERENCES

1. Stream Pollution Trends Monitoring Program (SPoT), https://www.waterboards.ca.gov/water_issues/programs/swamp/spot/
2. Hill, R. A., Weber, M. H., Leibowitz, S. G., Olsen, A. R., & Thornbrugh, D. J. (2016). The Stream-Catchment (StreamCat) Dataset: A Database of Watershed Metrics for the Conterminous United States. *JAWRA Journal of the American Water Resources Association*, *52*(1), 120-128.
3. NHDPlus Version 2, http://www.horizon-systems.com/NHDPlus/index.php
4. Leo Breiman (2001). "Random Forests". Machine Learning 45 (1), 5-32.
5. Ishwaran et al. (2010). High-dimensional variable selection for survival data. JASA 105 (489): 205-217

Corresponding author: Dan.Wang@cdpr.ca.gov

SETAC North America 39th Annual Meeting, Sacramento, CA, USA, November 4-8, 2018