Predicting Freshwater Ecotoxicological Concentrations for Chemicals in Household Products Using Linear Regression and Random Forest

Maryse Suppiger[1], Susan Oginah[2], Bu Zhao[1], Peter Fantke[2], Justin Colacino[1], Kristin Isaacs[3], Lei Huang[1] , Olivier Jolliet[1]

[1]Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, MI, USA
[2]Quantitative Sustainability Assessment, Technical University of Denmark, Kgs. Lyngby, Denmark
[3]Center for Computational Toxicology and Exposure, U.S. EPA, Durham NC, USA

1. Introduction

- Ecotoxicological data are limited for many chemicals and species.
- Regression and machine learning-based approaches have been utilized to predict ecotoxicological concentrations when data are lacking. Specifically, random forest modelling has been shown to successfully predict aggregated ecotoxicity at a chemical-specific level (Hou et al., 2020).
- This study aims to develop and compare models for predicting chemical concentrations that cause 10% of individuals in 20% of freshwater species (hazardous concentration-20th percentile, or HC20) to experience chronic toxic impacts.

2. Method

- Species sensitivity distribution (SSD) data were curated from work completed by colleagues at National Institute for Public Health and the Environment (RIVM) (Posthuma et al., 2019). This dataset included acute hazardous concentration values (HC50) for 7,439 chemicals, which were used to derive chronic HC20 values.
- Using this SSD-derived data, chronic HC20 values were derived for 4,887 chemicals. Acute HC50 data were used to generate chronic HC20 data due to the high coverage of this value in the dataset.
- Chemical properties data and QSAR-derived toxicity parameters were pulled from the Open (Quantitative) Structure-activity/property Relationship App (OPERA) (Mansouri et al., 2018) and Toxicity Estimation Software Tool (TEST) (Martin, 2016).
- Data for each model was split into a training set (70% of data) and a test dataset (30% of data). Models were built with the training set and performance was evaluated using test data.
- Acute HC50 values, OPERA data, and TEST data were entered as inputs into linear regression and random forest models to predict chronic HC20.
- Models were constructed and compared based on performance (test R^2/Q^2, cross-validation Q^2) and the number of chemicals included in model construction.
- Three models were selected and used to predict chronic HC20 (from chronic HC10) for chemicals in the US EPA Chemical and Products Database (CPDat) (Williams, 2017), with the highest performing models being used in priority.

3. Results

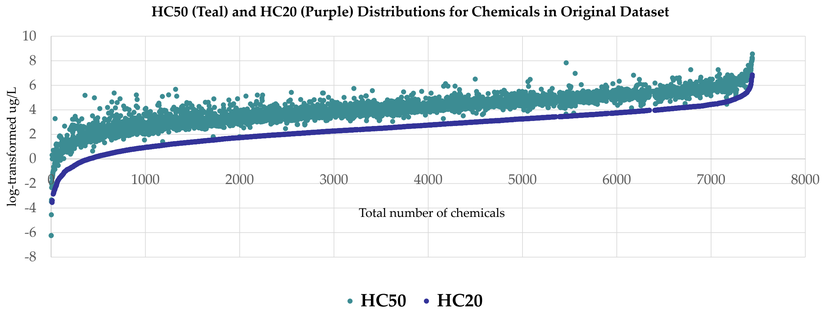

Fig.1. Distributions of HC50 values (teal) and HC20 values (blue) in original datasets.

Model 1: Linear Regression, Acute HC50 as only input

- When acute data were available for a chemical, chronic HC20 were predicted using simple linear regression (N=3,421; test R^2=0.89).
- The regression equation is y=1.05x – 1.81, where y=Chronic HC20 and x=Acute HC50.
- The univariate linear model performed similarly to a random forest model with the same input (test R^2=0.88).
- The distributions of acute HC50 and chronic HC20 values are plotted above (Fig.1.).

Model 2: Random Forest, TEST and OPERA parameters

- To generate HC20 predictions for chemicals lacking acute HC50 values in the original dataset, OPERA and TEST data parameters were input into a random forest model.
- All 29 toxicity and chemical property inputs were ranked by variable importance (increase in mean squared error) and added to the model sequentially, until model performance was optimized (Fig.2).
- Only eight inputs generated from the TEST consensus QSAR prediction method and four OPERA chemical properties were needed to maximize performance (N=1,230; test R^2=0.59) (Fig.3).

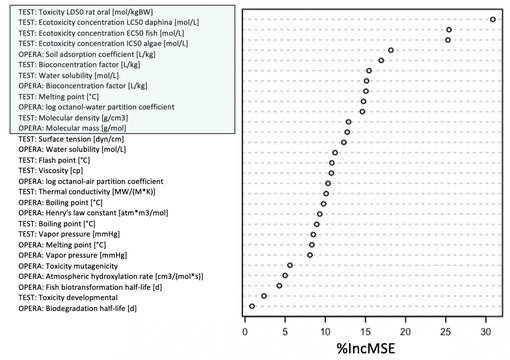

Fig.2. Top 12 most important (highest increase in mean squared error, or MSE, when excluded from model) TEST/OPERA variables for random forest model performance (highlighted in blue).

- This random forest model could predict roughly 20% more of the variance in HC20 values, compared to a multilinear regression with the same inputs (test R^2=0.39).

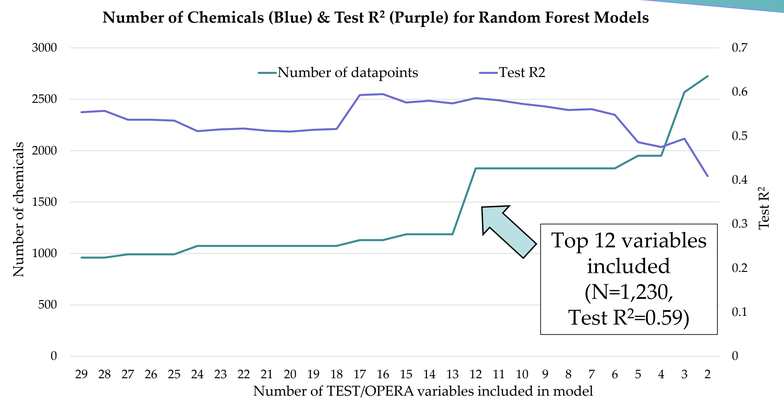

Fig.3. Number of chemicals and test R^2 of 28 random forest models. OPERA/TEST variables were sequentially added to a random forest model in order of increasing importance to model performance. Model performance and applicability was optimized with the top 12 variables.

Model 3: Random Forest, OPERA parameters only

- Finally, when TEST model predictions were missing for a chemical, thirteen chemical properties from only the OPERA model were entered in a separate random forest model (N=2,976; test R^2=0.39).
- Similarly to Model 2, random forest significantly outperformed linear regression for this set of inputs (Table 1).
- Once the models were built and trained, predictions could be generated for 1,342 out of the 1,363 chemicals in the CPDat dataset.
- When all three models were applied to the CPDat dataset, chronic HC20 coverage increased from 58% to 98.5%.

Table 1. Performance metrics for the three recommended prediction models, listed/highlighted in recommended hierarchy. Linear regression outperformed random forest when acute HC50 was the sole model input, but random forest outperformed regression modeling when only TEST/OPERA variables were used.

Model	# of data points	Test R^2, Linear Regression (Cross validation Q^2)	Test R^2 Random forest (Cross-validation Q^2)	Net increase in CPDat chemical coverage (N=1,363) with addition of each model
1. Acute HC-50 alone	4,887	**0.89 (0.89)**	0.86 (0.88)	8.9% (121 chemicals)
2. Top 12 TEST/OPERA variables	1,828	0.42 (0.39)	**0.56 (0.59)**	14.5 % (197 chemicals)
3. OPERA only	4,251	0.22 (0.22)	**0.43 (0.39)**	17.6% (240 chemicals)

4. Conclusions

Quantifying ecotoxicological concentrations is a necessary step towards successful ecosystem conservation efforts and thus, the applicability of predictive modeling should continue to be a focus of ecotoxicological research. In silico methods, such as machine learning algorithms, allow for an expansion of toxicological data without the cost, time, and ethical complications of in vivo studies.

5. References

1. Hou, P., Jolliet, O., Zhu, J., & Xu, M. (2020). Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. *Environment International*, *135*, 105393. https://doi.org/10.1016/j.envint.2019.105393
2. Posthuma, L., van Gils, J., Zijp, M. C., van de Meent, D., & de Zwart, D. (2019). Species sensitivity distributions for use in environmental protection, assessment, and management of aquatic ecosystems for 12 386 chemicals. Environmental Toxicology and Chemistry, 38(4), 905–917. https://doi.org/10.1002/etc.4373
3. Mansouri, K., Grulke, C. M., Judson, R. S., & Williams, A. J. (2018). OPERA models for predicting physicochemical properties and environmental fate endpoints. Journal of Cheminformatics, 10(1), 10. https://doi.org/10.1186/s13321-018-0263-1
4. Martin, T. User's Guide for T.E.S.T. (version 4.2) (Toxicity Estimation Software Tool) A Program to Estimate Toxicity from Molecular Structure. U.S. EPA Office of Research and Development, Washington, DC, EPA/600/R-16/058, 2016.
5. Williams, Antony (2017): The Chemical and Products Database (CPDat) MySQL Data File. The United States Environmental Protection Agency's Center for Computational Toxicology and Exposure. Dataset. https://doi.org/10.23645/epacomptox.5352997

Society of Toxicology
Creating a Safer and Healthier World by Advancing the Science and Increasing the Impact of Toxicology

The views expressed in this poster are solely those of the authors and do not represent the policies of the U.S. EPA. Mention of trade names of commercial products should not be interpreted as an endorsement by the U.S. EPA.