Predicting Human Alteration of River and Stream Salinity Using Random Forest Models

Franco Sanchez[1], John Olson[2], Steven Kim[1]

[1]Department of Mathematics and Statistics, California State University, Monterey Bay

[2]School of Natural Sciences, California State University, Monterey Bay

Introduction

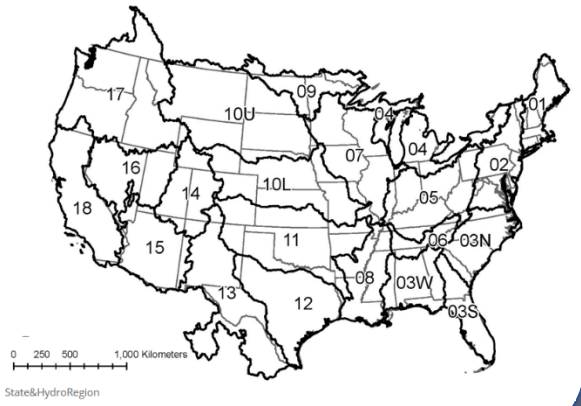
Salt concentration measured as specific conductivity (SC) in µS/cm, is an important measure in assessing water quality in streams, rivers, and watersheds. Human activities are known to cause changes in stream SC [1]. In particular,

- High levels of SC have been associated with negative effects on aquatic life [2].
- SC levels above 1000 µS/cm may be unsuitable for industrial and agricultural purposes [3].
- SC varies naturally based on temporal changes in the atmosphere [4], and so modeling SC variability is challenging due to the complex interactions between dynamic and static climate factors [5].
- Olson and Hawkins (2012) found that a non-parametric approach using random forest was superior to multiple linear regression when predicting the natural background SC of streams throughout the contiguous U.S. In this study, we follow a similar approach provided by Olson and Hawkins.

Models that can predict natural stream SC may serve as an ancillary reference when predicating the impact of human activity on steam SC. Modern climate data also makes it possible to model SC changes caused by natural and human factors based on spatial and temporal variations [2].

Research Objectives

- Develop a spatial/temporal model that relates human activity and natural environmental predictors to stream SC.
- Predict SC alteration throughout the contiguous U.S.



Methods

Data Collection, Processing and Formulating the Response as SC Alteration

- Monthly SC data (Jan 2001 -- Dec 2015) was gathered across the contiguous U.S. (total 2.6 million) (see **Table A** for data sources).
- 101 static predictors were obtained from StreamCat database [6].
- Monthly temporal predictors were matched with SC observations, following the methods from Olson and Cormier (2018).
- To limit outliers, only SC observations between 1 and 10,000 µS/cm were included in the model (n = 685,148).
- The SC observations were transformed by natural log, and the SC alteration was defined as the observed SC minus the natural background SC in the log-scale.

Developing a Random Forest Model

- Models were built using the "randomForests" package in R (Liaw and Wiener 2002) using all default settings.
- Model was built using 10% of the data (n = 68,513) determined by stratified sample on Level II Ecoregion.
- First iteration was constructed using all predictors (p = 129). Final model used only 45 predictors.

Variable Selection and Validation

- A PCA approach identified uncorrelated predictors with the strongest association to SC alteration (following a method suggest by R. A. Hill and E. W. Fox, personal communication).
- Further variable selection was determined by variable importance and partial dependence plots (**Fig B**).
- Stratified sample on Level II Ecoregion was used to validate, using another 10% of the data not already used in training.

STORET (USEPA 2016b)
USGS National Water Information Systems (USGS 2016)
State natural resources agencies, provide by Tetra Tech, Inc.
Predicted natural background SC from Olson & Hawkins (2012) Model

Table A. All sources used to obtain SC data

Results

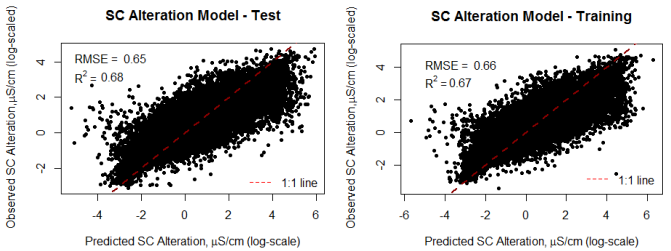


Figure A. Plots for predicted and observed SC alteration (log-scaled) in both training and validation subsets.

Mean Absolute Error (MAE)	0.45
Mean Squared Error (MSE)	0.44
Nash-Sutcliffe Efficiency (NSE)	0.46
Percent Bias (PBIAS%)	-0.1

Table B. Model performance on training data

Mean Absolute Error (MAE)	0.45
Mean Squared Error (MSE)	0.43
Nash-Sutcliffe Efficiency (NSE)	0.47
Percent Bias (PBIAS%)	0.1

Table B. Model performance on validation data

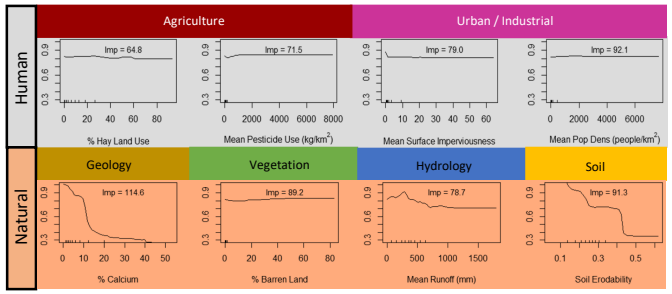


Figure B. Partial dependence plots for top predictors in human and natural classes. Variable importance is noted.

Discussion

Regarding random forest model,

- Based on validation, the final model explains 68% of variation in the data. The model makes well enough predictions and would be useful in understanding the effects of humans on salinity.
- The top natural features explain the variation in SC alteration more so than the top human features.
- Final random forest model (n = 68513, p = 45) had a computation time of 277 mins.

Regarding partial dependence plots,

- High calcium areas are less vulnerable to alteration.
- Soil that is more erodible is less vulnerable to alteration
- Regions with mean runoff < 700mm are less vulnerable to alteration. Relationship seems weak greater than 700mm.
- Barren land coverage has a weak relationship with alteration.
- Human features seem to have a weaker association with SC when the natural features are adjusted in the model.

Future Work

- Computation and time limitations only allowed us to use 20% of the data; we plan to increase data usage.
- Find optimal methods to decrease computational time in developing random forest models on the data.
- Determine how SC varies from normal conditions during prolonged droughts.
- Consider the usage of mixed effects linear models to make further assessment of human impacts on SC.
- Investigate the interaction between human activities and drought.

References:

1. Kaushal, S.S., Likens, G.E., Pace, M.L. Utz, R.M., Haq, S., Gorman, J., Grese, M., M., 2018. Freshwater salinization syndrome on a continental scale. Proc. Natl. Acad. Sci. U. S. A., doi: 10.1073/pnas.1711234115.
2. Cañedo-Argüelles, M., Hawkins, C.P., Kefford, B.J., Schäfer, R.B., Dyack, B.J., Brucet, S., Buchwalter, D., Dunlop, J., Frör, O., Lazorchak, J., Coring, E., 2016. Saving freshwater from salts. Science, 351, 914-916.
3. Olson, John R., Cormier, Susan M., , *in prep*. Modeling spatial and temporal variation in natural background specific conductivity.
4. Interlandi, S.J., Crockett, C.S., 2003. Recent water quality trends in the Schuylkill River, Pennsylvania, USA: a preliminary assessment of the relative influences of climate, river discharge and suburban development. Water Res., 37, 1737-1748.
5. Kundzewicz, Z.W., Krysanova, V., 2010. Climate change and stream water quality in the multi-factor context. Climatic Change, 103, 353-362
6. McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., Rea, A., 2012. NHDPlus Version 2: user guide. National Operational Hydrologic Remote Sensing Center, Washington, DC.
7. Fox, E.W., Hill, R.A., Leibowitz, S.G., Olsen, A.R., Thornbrugh, D.J., Weber, M.H., 2017. Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environ. Monit. Assess, 189, p.316.
8. Olson, J.R., Hawkins, C.P., 2012. Predicting natural base-flow stream water chemistry in the western United States. Water Resour. Res., 48.
9. Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), 18-22.