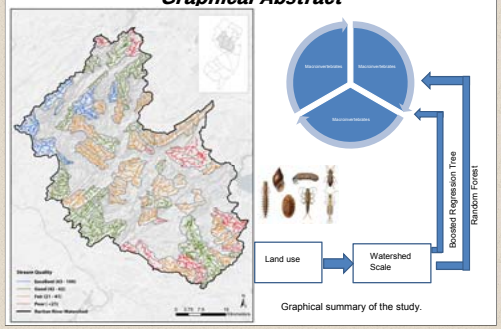
Evaluating the Impact of Land Uses on Stream Integrity Using Machine Learning Algorithms

Subhasis Giri, Richard Lathrop, Daryl Krasnuk
School of Environmental and Biological Sciences

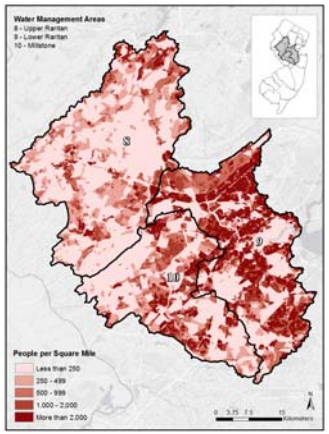
Graphical Abstract



Graphical summary of the study.

Introduction

- The population in the Raritan River Watershed(RRW) has an increasing trend. During 1990 to 2010, an increase in 25.6 percent of population was observed.
- Similar to population trend, an increasing housing density was found in the RRW. As a result; upland forest, agricultural lands, and wetlands are converted into low, medium, and high density residential, as well as new urban centers.



Population density distribution in the RRW (source: Giri et al., 2017. Sustainable Raritan report).

- Resulting increases in impervious surface (road, parking lots, and residential areas) are expected to have negative consequences on water quality and watershed health.

Rationale

- The composition of the macroinvertebrate community is often used as an indicator of the biological integrity of aquatic ecosystems. The intensity and type of human land uses may affect the macroinvertebrate community through multifaceted ways: water chemistry/quality, runoff patterns, habitat composition.

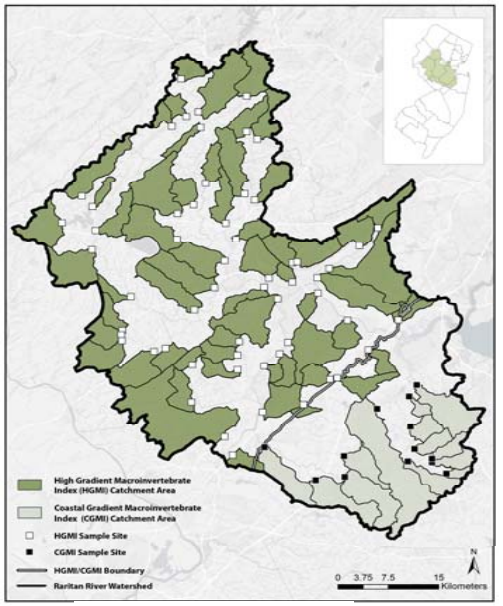
Objectives

- To determine the form and sensitivity of the relationship between land use intensity and stream health as characterized by macroinvertebrate integrity index data.
- To compare the utility of two different data mining techniques, random forests and boosted regression trees to address the above

Materials and Methods

Study Area

- The study area includes the entire Raritan River basin in central New Jersey including parts of Hunterdon, Mercer, Middlesex, Monmouth, Morris, Somerset, and Union Counties.
- Out of 71 watersheds, 58 are distributed randomly in high gradient macroinvertebrates index boundary and 13 are located in the coastal gradient macroinvertebrates index boundary.



Location of 71 study sub-basins in the study area.

- A total of 139 sub-basins were delineated using Streamstats based on macroinvertebrates data from New Jersey Department of Environmental Protection (NJDEP). Out of which, half of the watersheds were eliminated using following two criteria, i) no nestedness and spatial autocorrelation among watersheds, ii) threshold watershed area (>1,000 acre).

Land Use Matrix

- 2012,2007,2002, 1995- land use/cover data obtained from NJDEP was used in this study. Level III NJDEP plus Hasse-Lathrop reclassification systems were used to get the final land use matrix.

Stream Integrity Index

Photo: Jacqui Banfield-Taylor

Photo: Tasman District Council

Photo: National Forest Foundation

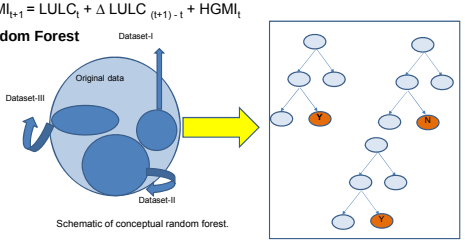
- Stream integrity index is calculated in the form of a multimetric index known as New Jersey Impairment Score (NJIS) using benthic macroinvertebrates samples collected at each station by NJDEP
- High Gradient Macroinvertebrate Index (HGMI) consists of seven distinct metrics including total number of genera, percent of non-insect genera, percent of sensitive EPT (Ephemeroptera-Plecoptera-Trichoptera), percent of scraper genera, Hilsenhoff Biotic Index, total number of Attribute 2 genera, and total number of Attribute 3 genera used for this study.
- Four rounds of HGMI data from round-2 to round-5 collected during the years 1998-1999, 2004, 2009, 2014 were used.
- Habitat data were used as additional explanatory variables to land uses to explain the processes occurring closer to the sampling location.

Model Development Using Machine Learning Algorithms

- Model-1: Composite of Rounds2-5 data of HGMI and corresponding land uses.
- Model-2: Inclusion of HGMI as dependent variable to Model-1 using the following equation:

$$HGMI_{t+1} = LULC_t + \Delta\, LULC_{(t+1)-t} + HGMI_t$$

Random Forest



Schematic of conceptual random forest.

- Random forest is a non parametric method applied in variety of environmental research and it uses multiple learning algorithms to obtain better predictive performance.
- Decision trees are developed based on random selection of data and variables.

Boosted Regression Tree

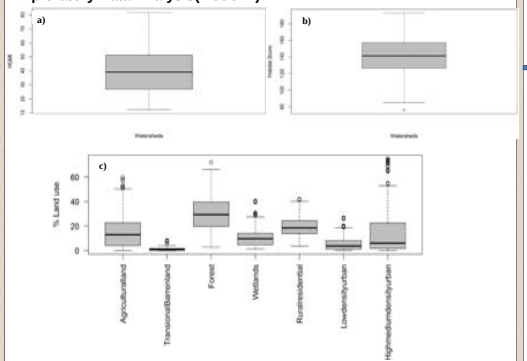
Schematic of conceptual Boosted Regression Tree

Photo: mariajesusbigml

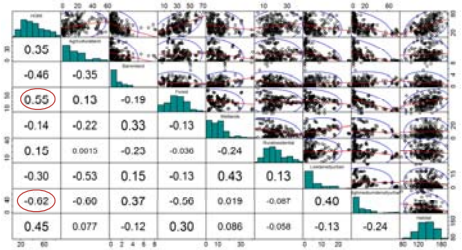
- Boosted regression tree is an advanced form of regression which combines large number of single models to improve the prediction.
- It uses boosting technique(forward stagewise procedure) using the information of residuals from previous tree development.

Results and Discussion

Exploratory Data Analysis(Model-1):



Box and whisker plots illustrating range of (a) HGMI stream integrity, (b) habitat scores, and (c) land use types in the study.

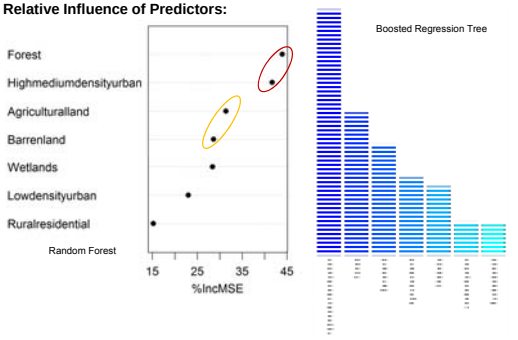


Spearman Correlation Matrix for HGMI stream integrity vs explanatory variables.

Model Performance(Model 1):

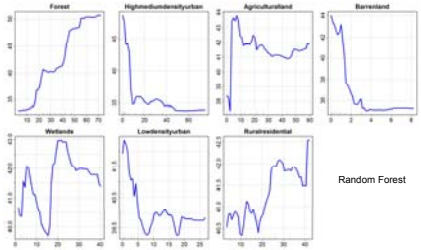
Algorithms		NSE	RMSE	PBIAS
Random Forest	5-fold CV	0.56	12.16	-0.79
Boosted Regression Tree	5-fold CV	0.49	13.04	-0.71

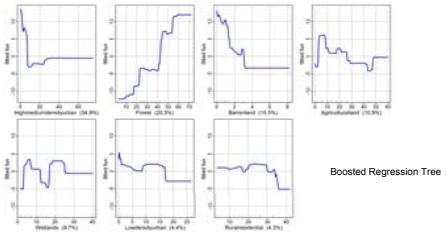
Relative Influence of Predictors:



- Both models agree about the top four most important predictors of stream integrity (forest, high-medium density urban, agricultural land, and barren land), however, the sequence of predictors is slightly different.

Partial Dependence Plot:





- Among the most important predictors, forest is positively related while high-medium density urban is negatively related to stream integrity.
- High-medium density urban and barren land show steep drop at around 10% and 2%, respectively followed by a levelling of the response.
- In case of low density urban, the stream integrity drops rapidly at around 8% beyond which the value shows somewhat constant.

Conclusions

- Forest and high medium density urban land are most important for stream health.
- Identification of thresholds for land uses will enable to craft land use zoning regulation and design restoration program .
- Both machine learning algorithms were able to explain at least 50 percent of HGMI.
- Model-2 performed slightly better over Model-1 due to higher 5-fold CV NSE.

Acknowledgements: Authors would like to thank Johnson Chair in Water Resources for providing funding for this study and the staff of the Bureau of Freshwater and Biological Monitoring, New Jersey Department of Environmental Protection for collecting the macroinvertebrate data and developing into HGMI