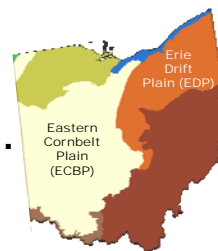




# Data exploration of eco-epidemiological relationships for a large-scale multi-stressor database using statistical classification techniques.

Leslie Faggiano<sup>1</sup>, Katherine Kapo<sup>2</sup>, Scott Dyer<sup>3</sup>, Dick DeZwart<sup>4</sup> and Leo Posthuma<sup>4</sup>



<sup>1</sup>University of Girona, Spain.

<sup>2</sup>Montani Run LLC, Kenna, West Virginia, USA

<sup>3</sup>The Procter & Gamble Company, Cincinnati, Ohio, USA

<sup>4</sup>RIVM, Bilthoven, The Netherlands.

Email contact: leslie.faggiano@udg.edu

## INTRODUCTION:

- Screening-level data exploration approaches can be useful for identifying variables which are influential in the structure of the raw biological data. These relatively simple types of analyses applied to large databases can narrow data selections and provide guidance for subsequent more complex analysis approaches which require more time and effort.
- A dataset of biological, water chemistry, habitat, toxicity, and various other environmental parameters were collected for approximately 2000 catchments across the state of Ohio. Of the >100 available environmental parameters, some initial expert judgment was utilized for variable selection, as well as PCA analysis to reduce numerous water chemistry variables to representative factors. Correlation matrices (corrgram, R version 2.10.1, see Figure in poster WEPC4-7) were also constructed to examine data relationships between environmental parameters.
- The Index of Biotic Integrity (IBI), compiled by the Ohio Environmental Protection Agency, OEPA) was evaluated using decision tree analysis using both regression trees and random forest approaches to identify important environmental parameters at the ecoregion level to investigate region-specific data relationships.
- A biological trait (fish size) was evaluated using Artificial Neural Network to identify major drivers explaining the distribution of each functional group (Small, Medium and Big). Actually, this approach could potentially quantify "ecosystem health" through the functional biodiversity of communities.
- Analyses using biological metric (IBI) and biological trait (size) for the Eastern Corn belt Plains and the Erie Drift Plain ecoregions are presented in this poster.

## Material and Method:

- Regression Tree Analysis** (Rpart, R version 2.10.1) was applied to the IBI and the environmental parameters shown in Figure 2 (with exception of land use parameters) as an initial variable selection tool. Variables with the greatest influence on IBI are located closest to the "roots" (origin) of the tree, with the end "nodes" of the tree representing groups of sites having particular environmental conditions that result in their average biological value.
- Random Forest Analysis** (RandomForest, R version 2.10.1) was subsequently applied to the IBI and environmental parameters following regression tree analysis. Variable importance is provided as a ranked list, based on the reduction in model accuracy (fit) given the removal of the particular variable from its order in the list.
- The mean body mass of 60 fish species collected in Ohio was used to split them in three categories: Small (n= 24; mean body mass: 3.20g), Medium (n=17; mean body mass: 18.03g) and Big (n= 19; mean body mass: 407.81g).

Fish assemblage data collected at each sampling sites were then used to weight the occurrence of this biological trait.

A **multilayer perceptron with a backpropagation learning algorithm (BPN)** was applied in order to make quantitative prediction of these relative frequencies according to 16 environmental variables. The BP algorithm is a supervised learning algorithm designed to minimize the mean square error between the results or predictions computed by the network and the observations. Model performance was determined using the correlation coefficient (r) between observed and estimated values of the 3 output variables (Tab 1). Finally, a sensitivity analysis (the 'PaD' for Partial Derivatives) was applied in order to determine the contribution of each predictor for explaining the distribution of each functional group.

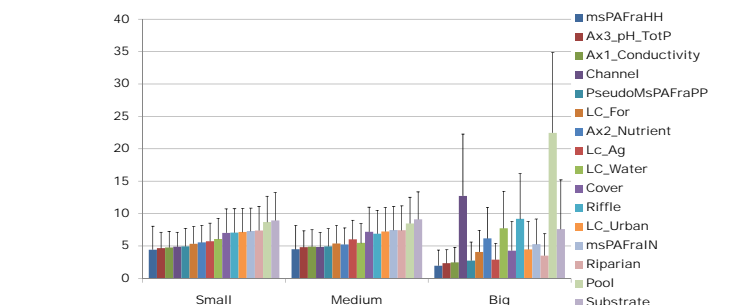
## RESULTS:

Figure 1 and 2 show the regression tree results for IBI in the Eastern Cornbelt Plains (Fig. 1) and in the Erie Drift Plain (Fig. 2) ecoregions.

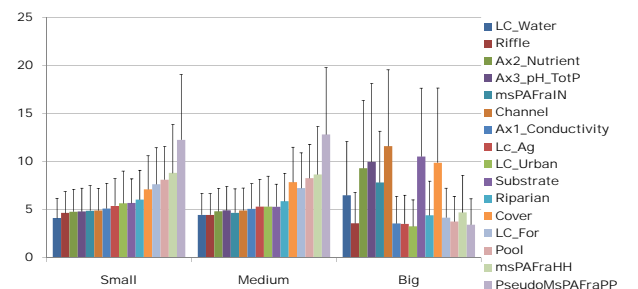
Table 1 gives the mean and standard deviation of correlation coefficients values between observed and predicted values (MLP-BP) for each trait modality and for each ecoregion.

		Small	Medium	Big
Eastern Cornbelt Plain	Mean	0.15	0.21	0.22
	Sd	0.07	0.07	0.08
Erie Drift Plain	Mean	0.20	0.20	0.27
	Sd	0.09	0.09	0.13

For the Eastern Cornbelt Plain ecoregion (below) the quality of substrate, pool and cover were highly influential for Small and Medium fish while the quality of pool, channel and riffle were the most important factor explaining the distribution of Big fishes. For Small and Medium organisms, industrial toxicity as well as the percentage of urban area (LC\_Urban) had strong contribution in the hierarchical ANN.



For the Erie Drift Plain ecoregion (below), variables representing pesticide and pharmaceutical toxicity were the most contributing factors for Small and Medium Fish while habitat variables (channel, substrate and cover) as well as chemical variables (Conductivity, pH and Total Phosphorus: PCA factor) were highlighted for Big ones.



### Conclusions:

- Habitat quality were consistently highlighted as being highly influential variables in prediction of both IBI and trait response showing that fish communities are primarily structured by large-scale differences in habitat. Other important variables were more dependent on the specific type of biological response or geographic area (ecoregion) being evaluated. We demonstrate a significant link between fish traits and toxicity (pesticide and pharmaceutical) for Small and Medium size organism in the Erie Drift Plain ecoregion highlighting the usefulness of a functional approach.
- In this study, the combined use of a taxonomical approach (IBI) and a functional one (size) as well as the use of different advanced modeling method provide preliminary hypotheses and guidance for more detailed quantitative analyses.

Figure 1

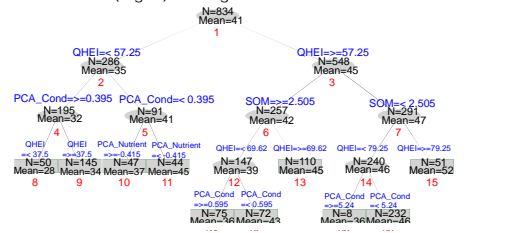
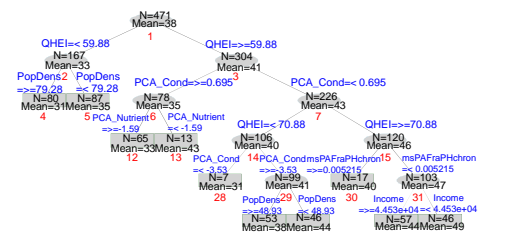
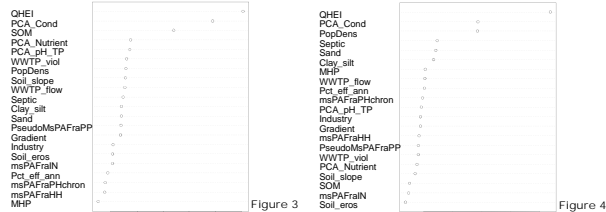


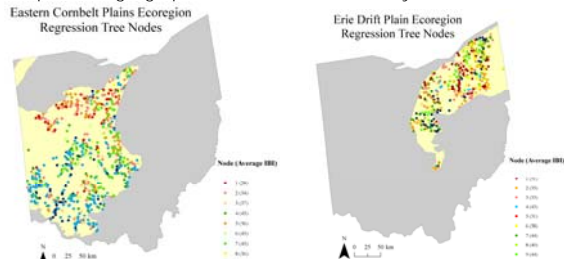
Figure 2



For the Eastern Cornbelt Plain ecoregion (Fig. 3; a highly agricultural landscape) soil organic matter and nutrients were highly ranked in addition to habitat and conductivity (PCA). For the Erie Drift Plain ecoregion (Fig. 4; a comparatively more urban landscape), a variable representing pharmaceutical toxicity was highlighted by regression tree analysis and more highly ranked by random forest analysis compared to other ecoregions.



Regression tree nodes representing sites having similar environmental conditions can be plotted to provide a geographic visualization of the analysis results



### Acknowledgement:

The authors thank the expertise and cooperation of the Ohio EPA, whom without collecting biological data throughout the state would not make it possible to test the relationship of prognostics and diagnostics at diverse geographical scales. A PhD funding was provided by the EU project keybioeffects (MRTN-CT-2006-035695).