Machine Learning in the Environmental Discipline: Statistical Paradigm Shifts, Combining Data Mining with Geographic Information Systems (GIS), and Educating the Next Generation of Environmental Researchers

Falk Huettemann
-EWHALE lab-
419 Irving 1
Inst. of Arctic Biology
Biology & Wildlife Dept
Fairbanks, Alaska 99775 USA
fhuettmann@alaska.edu
What’s at stake?

- a better pre-screening of data (data mining)
- a better inference (better model fits)
- a better hypothesis (more relevant)
- a faster and more accurate analysis
- a better prediction (pro-active decisions)

=> progress in science & society
=> a new culture
This presentation

- Environmental Sciences (Sustainability)
- Paradigm Shift in the Sciences
- Geographic Information Systems (GIS)
- Teaching Machine Learning at Public Institutions
The earth and its atmosphere ...we only have one ...
The atmosphere
Rio Convention 1992 and its goals
...people are everywhere these days, mostly at the coast, affecting Wildlife, Habitats and Ecological Services...
An Earth Sampler: The Three Poles

1.

2.

3.
Three poles: Ice, glaciers and snow

=> A cool (or a hot?) place to be...

Sources: NASA & Natural Resources Defense Council
No data ...and “nothing is known” ...?
Climbing and Understanding Data Mountains with Data Mining:
A free copy of GBIF data... [www.gbif.org],
seamap.env.duke.edu/ , mountainbiodiversity.org

Megascience
Examples of people using and publishing on Machine Learning in the Biodiversity, Wildlife and Sustainability Sciences...
Classify, pre-screen, explain and predict data

A heap of data,
e.g. online and merged,
free of research design

Response $\sim$ Predictor1 + Predictor2 + Predictor3 + Predictor4 + ....
(e.g. clustering & multiple regression)
Multiple Regressions: Are they any good?! 

Response ~ Predictor1 + Predictor2 + Predictor3 + Predictor4 +....

The issues are for instance:
- Many predictors, e.g. 30
- Correlations
- Interactions (2-way, 3-way and higher)
- Coefficients + Offset
- (Multi-) hypothesis testing
- Model Selection
- Parsimony

=> A single formula vs a software code as a result ?!
What statistical philosophies and schools exist?

- Frequency Statistics
  - R. A. Fisher

- Bayesian Statistics
  - Bayes

- Machine Learning
  - L. Breiman

Occam (Parsimony)
What statistical philosophies and schools exist?

*e.g.* Frequency Statistics
- Normal distribution
- Independence
- Linear functions
- Variance explained
- Parsimony
What statistical philosophies and schools exist?

- Bayesian Statistics
  - Priors
  - Fit a function
  - Variance explained
  - (alternative) confidence intervals
  - Linear functions (usually)
What statistical philosophies and schools exist?

E.g. Machine Learning
- Inherently non-linear
- No prior assumptions (non-parametric)
- Fast
- Computationally intense
What linear models and frequency statistics cannot achieve

Linear
(~unrealistic)

Non-Linear
(driven by data)

‘Mean’
SD
=> potentially low r2

‘Mean’?
SD?

Machine Learning, e.g.
CART, TreeNet &
RandomForest
(there are many
other algorithms !)

Virtually not Pro-Active

Pro-active Potential
~Pre-cautionary
What machine learning is

A numerical pattern

Machine Learning algorithm

The mimick’ed pattern

Virtually NOT parsimonious (no AIC, no R2 etc etc)
What is a non-linear algorithm ...?

Through the origin vs.

Not through the origin

Single formula vs.

Many formulas
What is a non-linear algorithm ...?

Through the origin vs. Not through the origin
What is a non-linear algorithm ...?

Single formula vs. Many formulas
What is a non-linear algorithm ...?

Through the origin vs.
Not through the origin

Single formula vs.
Many formulas

A binary software algorithm that captures the mathematical relationship vs.
A single formula with (a) coefficient
Linear regression as an artefact, *aka* statistical & institutional terror?!
Predictions as the prime goal, and linked with Adaptive Management (done in Random Forest)

Best data used and available?

Pelagic King Penguin prediction using RF (french telemetry data from SCAR-MARBIN)
Outliers and variance ...

Root Mean Square Error (RMSE)
Outliers and variance ...

Confusion Matrix (=> ROC curve)

<table>
<thead>
<tr>
<th>Reality</th>
<th>Predicted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>80%</td>
<td>No</td>
</tr>
<tr>
<td>20%</td>
<td>20%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Confusion Matrix (=> ROC curve)
Moving into Data Mining
(non-parametric, rel. high data error)

A heap of data,
e.g. online and merged,
free of research design

Your classic Frequentist Statistician
might get very scared and horrified
with such data...
Describe Patterns, Outliers

Informed Traditional Analysis

Screening

Data Quantification (~prediction)

Moving into Data Mining

A heap of data, e.g. online and merged

Craig, E., and F. Huettmann. (2008). Using “blackbox” algorithms such as TreeNet and Random Forests for data-mining and for finding meaningful patterns, relationships and outliers in complex ecological data: an overview, an example using golden eagle satellite data and an outlook for a promising future. Chapter IV in Intelligent Data Analysis: Developing New Methodologies through Pattern Discovery and Recovery (Hsiao-fan Wang, Ed.). IGI Global, Hershey, PA, USA.
Craig, E., and F. Huettmann. (2008). Using "blackbox" algorithms such as TreeNet and Random Forests for data-mining and for finding meaningful patterns, relationships and outliers in complex ecological data; an overview, an example using golden eagle satellite data and an outlook for a promising future. Chapter IV in Intelligent Data Analysis: Developing New Methodologies through Pattern Discovery and Recovery (Hsiao-fan Wang, Ed.). IGI Global, Hershey, PA, USA.


Describe Patterns, Outliers

Screening
Craig, E., and F. Huettmann. (2008). *Using “blackbox” algorithms such as TreeNet and Random Forests for data-mining and for finding meaningful patterns, relationships and outliers in complex ecological data: an overview, an example using golden eagle satellite data and an outlook for a promising future.* Chapter IV in *Intelligent Data Analysis: Developing New Methodologies through Pattern Discovery and Recovery* (Hsiao-fan Wang, Ed.). IGI Global, Hershey, PA, USA.

(Adaptive/Science-based) Wildlife Management: How it works...

- Identify a problem
- Do the science
- Provide a management suggestion
- Implement science-based management
- Assess and fine-tune

=> Pro-active (before trouble occurs)
=> Predictive in Space and Time
Paradigm Shifts

- Data Mining as a means to inform an analysis

- Predictions as a goal in itself

- Linear regressions and algorithms as a poor performer for modern problems of relevance to mankind

- Metrics like $r^2$, $p$-values and AIC as being suboptimal

- Statistics based on (very advanced) software

- Decision-support systems, Pro-Active Management and Operations Research
RandomForest as a scientific discipline ?!

-a big and new research field per se (e.g. machine learning, ensemble models)

-comes naturally with computing power and the internet

-changes science (and fieldwork, e.g. design-based)

-changes education (e.g. universities & stats depts)

-changes pro-active/adaptive management, multi-species

-changes society

-affects land- and seascapes
Adding Spatial Complexity: What is a GIS (Geographic Information System)?

“Computerized Mapping”

Consists of Hardware & Software (Monitor, PC and Database)

ArcGIS by ESRI as a major software suite

=> Can be linked with Machine Learning, e.g. for predictions, classifications and data mining
What is a GIS (Geographic Information System)?
RandomForest & GIS:
Mapping Supervised & Unsupervised Classification

**Supervised Classification:**
- Multiple Regression (classification or continuous)
- Multiple Response
  e.g. YAIMPUTE

**Unsupervised Classification:**
1. Proximity Matrix via Bagging/Voting (RF)
2. Similarity Matrix
3. e.g. Regular Clustering (mclust, PAM)
3. Visualize Result
11 Cliome Clusters (RF)
Climate Cluster Data, Canada & Alaska

Credit: M. Lindgren et al.
Teaching Machine Learning

What’s the issue?

- the current teaching concept re. math & stats

- wider introduction of machine learning

- parsimony vs. holistic analysis and approach

- better predictions
Teaching math, statistics and machine learning...

Usually, machine learning starts where maximum-likelihood theory ends...

=> maximum-likelihood starts at college...

But it does not have to be that way!
Teaching math, statistics and machine learning...

<table>
<thead>
<tr>
<th>Student’s age</th>
<th>Topic learned</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>Basic Math</td>
</tr>
<tr>
<td>20</td>
<td>Frequency Statistics (Maximum Likelihood)</td>
</tr>
<tr>
<td>23 - 30</td>
<td>Machine Learning</td>
</tr>
</tbody>
</table>
Teaching math, statistics and machine learning...the better way

<table>
<thead>
<tr>
<th>Student’s age</th>
<th>Topic learned</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>Basic Math</td>
</tr>
<tr>
<td>16</td>
<td>Frequency Statistics</td>
</tr>
<tr>
<td>18</td>
<td>(Maximum Likelihood)</td>
</tr>
<tr>
<td>19</td>
<td>Machine Learning</td>
</tr>
</tbody>
</table>

Why & How done?

via Greybox

Blackbox
Teaching math, statistics and machine learning...

Parsimony (complexity vs. variance explained)

<table>
<thead>
<tr>
<th>Model name</th>
<th>AIC etc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>1000</td>
</tr>
<tr>
<td>Model 2</td>
<td>900</td>
</tr>
<tr>
<td>Model 3</td>
<td>600</td>
</tr>
<tr>
<td>Model 4</td>
<td>500</td>
</tr>
<tr>
<td>Model 5</td>
<td>1200</td>
</tr>
<tr>
<td>Model 6</td>
<td>1100</td>
</tr>
</tbody>
</table>

Breaking down a complex real-world problem into a single linear term...and for a mechanistic understanding...??
Teaching math, statistics and machine learning…

Predictions (knowing things ahead of time)

For spatial predictions
For temporal predictions
For a pro-active management
To...Salford Systems, Dan Steinberg and his team, our students, L. Strecker, S. Linke, and many many project and research collaborators world-wide over the years!!

Questions ?