Can we use random forests for spatiotemporal CPUE modeling?

BRIAN STOCK, ERIC WARD, BRICE SEMMENS
What we want (from Rick Methot)

- **Fast** (coding vs. runtime vs. interpretation)
- **Replicable** (method well-defined, get same answer)
- **Robust** (insensitive to distributional assumptions, outliers)
- **Predictive ability** (minimal errors, fill in space/time gaps)
- **Covariate effects** (nonlinear, interactions)
- **Uncertainty estimates** (with known properties)
- **Specifiable structure** (e.g. correlation through time, biology)
- **Unbiased** (relative vs. absolute abundance)

Introduction
What we want (from Rick Methot)

- Fast
- Replicable
- Robust
- Predictive ability
- Covariate effects
- Uncertainty estimates
- Specifiable structure
- Unbiased

Introduction

Story 1: Bycatch hotspots
What we want (from Rick Methot)

- Fast
- Replicable
- Robust
- Predictive ability
- Covariate effects
- Uncertainty estimates
- Specifiable structure
- Unbiased

Story 2: Total bycatch estimation

\[ \sum \]
What we want (from Rick Methot)

- Fast
- Replicable
- Robust
- Predictive ability
- Covariate effects
  - Uncertainty estimates
- Specifiable structure
- Unbiased

Introduction

Story 3: CPUE standardization
Tools for dynamic management

Need map of bycatch “risk”

1. Introduction
Tools for dynamic management

Need map of bycatch "risk"

1. Introduction
Tools for dynamic management

Need map of bycatch "risk"

• temperature
• depth
• substrate
• spatial field
Q: Which spatial model is best?

1. Research question

- temperature
- depth
- substrate
- spatial field

- GLM
- GAM
- GMRF
- RF
Q: Which spatial model is best?

- temperature
- depth
- substrate
- spatial field

1. Research question
What are these models exactly?

1. Methods

obs ~ environmental predictors (temp, depth, ...)

\[ Y_i \sim Binomial\left(\logit^{-1}[X_i\beta]\right) \]  
\[ Y_i \sim Gamma\left(e^{X_i\beta}, \nu\right) \]
What are these models exactly?

1. Methods

- **GLM**
  \[ \text{obs} \sim \text{environmental predictors (temp, depth, ...)} \]

- **GAM**
  \[ \text{obs} \sim \text{environmental predictors} + s(\text{lat,lon}) \]
What are these models exactly?

- **GLM**
  \[ \text{obs} \sim \text{environmental predictors (temp, depth, ...)} \]

- **GAM**
  \[ \text{obs} \sim \text{environmental predictors} + s(\text{lat, lon}) \]

- **GMRF**
  \[ \text{obs} \sim \text{environmental predictors} + MVN(0, \Sigma) \]

1. Methods
What are these models exactly?

1. Methods

- **GLM**
  \[ \text{obs} \sim \text{environmental predictors (temp, depth, \ldots)} \]

- **GAM**
  \[ \text{obs} \sim \text{environmental predictors} + s(\text{lat, lon}) \]

- **GMRF**
  \[ \text{obs} \sim \text{environmental predictors} + \text{MVN}(0, \Sigma) \]

- **RF**
  \[ \text{obs} \sim \text{environmental predictors} + \text{lat} + \text{lon} \]
Fisheries observer data

1. Methods

U.S. West Coast
Groundfish
Trawl

Hawaii
Swordfish
Longline
Generally:

- GLM
- GAM
- GMRF
- RF

1. Results

- Binomial
Generally: \( \text{GLM} < \text{GAM} < \text{GMRF} < \text{RF} \)

Less clear for rarer species

1. Results

- For common species:
  - \( N^+ = 7,660 \)
  - 18%

- For rare species:
  - \( N^+ = 143 \)
  - 0.3%

Model AUC (test data):
- Common species: [Boxplot]
- Rare species: [Boxplot]

Binomial
Generally: GLM < GAM < GMRF < RF

1. Results Positive
Q: How much bycatch can they prevent?

Crude management simulation:

1. Predict bycatch risk at test locations
Q: How much bycatch can they prevent?

Crude management simulation:

1. Predict bycatch risk at test locations
2. Remove X% of fishing effort with highest bycatch risk
Q: How much bycatch can they prevent?

Crude management simulation:
1. Predict bycatch risk at test locations
2. Remove X% of fishing effort with highest bycatch risk
3. Calculate “prevented” bycatch and target catch (bycatch:target ratio)
Q: How much bycatch can they prevent?

<table>
<thead>
<tr>
<th>Fishing removed</th>
<th>Bycatch:target reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>8%</td>
</tr>
<tr>
<td>5%</td>
<td>34%</td>
</tr>
<tr>
<td>10%</td>
<td>50%</td>
</tr>
</tbody>
</table>

1. Results
Covariate effects

1. Results

PredOccSurvey  Depth  In/near RCA

Palczewksa (2013), Welling (2016)
Covariate effects

1. Results

Palczewksa (2013), Welling (2016)
How do random forests work?

Single decision tree:
Low bias, high variance model (overfit)

1. Discussion
How do random forests work?

Idea: average across many, uncorrelated trees

\[ E[MSE] = Model\ Bias^2 + Model\ Variance + noise \]

1. **Bagging**: fit each tree on a Bootstrap sample (~63%) of the data, then *Aggregate* across trees (~1000+)

2. Only consider a *random subset* (~p/3) of covariates at each split

---

**1. Discussion**
Covariate effects with RF

What is a “feature contribution”??

Pr = 0.18

Depth >= 250 fm

Pr = 0.12

Depth < 250 fm

Pr = 0.21

Temp >= 1

Pr = 0.15

Temp < 1

Pr = 0.11

1. Discussion

Palczewksa (2013), Welling (2016)
What is a “feature contribution”??

Prediction\textsubscript{i} = 0.11 = 0.18 − 0.06 (Depth) − 0.01 (Temp)
Covariate interactions with RF

Catchability varies by Julian Day

1. Discussion
Need estimates of total bycatch / discards

- Rarely observe 100% of fishing
- Often observe ~20%
#2: Total bycatch estimates

“Ratio estimator”:

\[ B_{unobs} = T_{unobs} \frac{B_{obs}}{T_{obs}} \]

⚠️ Assumes bycatch prop. to target catch / effort
Use a spatial model instead

Cross-validation using dataset with 100% coverage:

1. Choose 20% observed trips
2. Fit spatial model
3. Predict at 80% unobserved
4. Compare sum(predictions) to ratio estimator
Spatial models = lower error

2. Results
2. Results

... bias in spatial model estimates
Why are random forests biased?

1. Extreme values modeled using average of less-extreme points → Regression to the mean

2. Bycatch distribution is right-skewed
Thoughts on RF bias

Bias correction methods:
- Fit linear model in nodes instead of mean (‘Cubist’)
- Fit second model on RF residuals (Xu 2013)

Bycatch estimates (absolute abundance) vs. CPUE standardization (relative abundance)
#3: CPUE data

Eastern Pacific Ocean yellowfin tuna
- 2000-2009 catch + effort
- 1-deg lat/lon bins

Model:
- 2000-2009 catch + effort
- 1-deg lat/lon bins

'\textit{ranger}': \texttt{ranger(cpue \sim \text{lat} + \text{lon} + \text{year}, \ldots)}

'\textit{grf}': \texttt{regression\_forest(dat[,covar], Y=dat$cpue, \ldots)}
3. Methods

CPUE data
3. Methods

Create prediction grid

Area with at least 1 year of data

‘alphahull’ R package

Areas with no data
3. Methods

CPUE data
3. Results

Predicted mean (CPUE)
3. Results

Predicted Var(CPUE)
3. Results

Relative abundance trend
3. Results

Predicted CV(CPUE)

Reasonable scale ★★★
Reasonable pattern ★★★
3. Diagnostics

log(CV) vs. log(Effort)
Standardized residuals

3. Diagnostics
3. Diagnostics

Bias (regression to the mean)
Uncertainty estimates

Need \textit{covariance} between spatiotemporal predictions
\textit{Rapidly evolving...} 34,336 citations Breiman (2001)

1. Quantile regression forests – prediction quantiles
   (‘ranger’, ‘grf’, Meinshausen 2006)

2. Jackknife & infinitesimal jackknife – standard error
   (‘ranger’, Wager et al. 2014)

3. U-statistics – asymptotically normal variance estimate
   (‘surfin’, Mentch & Hooker 2016)

   (‘grf’, Athey et al. 2017)

5. Bayesian additive regression trees – full posterior inference
Other thoughts

Multivariate response:

- Include `model.matrix` as covariates:

```r
levels(Data_Geostat$spp) <- c("A_stomias", "G_chalcogrammus","H_lassodan")
sp.mat <- data.frame(model.matrix(~ spp - 1, Data_Geostat))
mv.dat <- cbind(Data_Geostat, sp.mat)
rfmv = ranger(Catch_KG ~ Lat + Lon + Year + sppA_stomias + sppG_chalcogrammus + sppH_lassodan, data=mv.dat, num.trees=1000, mtry=2, keep.inbag=T, write.forest=T)
```

Buffer distances to smooth predictions:

3. Discussion

https://github.com/thengl/GeoMLA
What we want (from Rick Methot)

- Fast (coding vs. runtime vs. interpretation)
- Replicable (method well-defined, get same answer)
- Robust (insensitive to distributional assumptions, outliers)
- Predictive ability (minimal errors, fill in space/time gaps)
- Covariate effects (nonlinear, interactions)
- Uncertainty estimates (with known properties)
- Specifiable structure (e.g. correlation through time, biology)
- Unbiased (relative vs. absolute abundance)

Discussion
Thank you!

SIO
- Brice Semmens

SWFSC
- Tomo Eguchi

NWFSC
- Eric Ward
- Jim Thorson
- Essential Fish Habitat (Blake Feist)
- West Coast Groundfish Observer Program (Jason Jannot)
Bias-variance tradeoff by species...

2. Results
More worthwhile for rarer species

2. Results
Q1: Which spatial model is best?

Goal: *prediction*

5-fold cross validation repeated 10x

- **Binomial**

**ROC curve (AUC)**

<table>
<thead>
<tr>
<th>ROC curve</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worthless</td>
<td>0.5</td>
</tr>
<tr>
<td>Ok</td>
<td>0.7</td>
</tr>
<tr>
<td>Good</td>
<td>0.8</td>
</tr>
<tr>
<td>Awesome</td>
<td>0.9+</td>
</tr>
</tbody>
</table>
Q1: Which spatial model is best?

Goal: *prediction*
- 5-fold cross validation repeated 10x

- **Binomial**
- **Positive**

**Evaluation Metrics**
- AUC
- RMSE, $R^2$ (pred – obs)

$$\sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$
West Coast Groundfish covariates

---

**Binomial**

\[ \sim \text{positive} \]

- \( \text{sst} + \text{sst}^2 + \)
- \( \text{depth} + \text{depth}^2 + \)
- \( \text{distance to rocky substrate} + \)
- \( \text{size of rocky patch} + \)
- \( \text{in Rockfish Conservation Area} + \)
- \( \text{predicted occurrence (survey)} + \)
- \( \text{day of year} + \)
- \( \text{spatial field} \)

Shelton et al. (2014)
Hawaii Longline covariates

- Binomial
- ~ sst + sst² +
- Positive
- day of year +
- spatial field

Chapter 2: Bycatch prediction

Shelton et al. (2014)
RF

+ Better at prediction
+ More complex covariate relationships (incl. interactions)
+ Easier to set up and run
+ Not just a “black box”?  

GMRF

+ Statistical inference, marginal posteriors for covariate effects
+ Ability to include observation error
Variance of predictions

Discussion

Wager et al. (2014)
Variance of predictions

Discussion

Wager et al. (2014)
Variance of predictions

Non-parametric delta method / “infinitesimal jackknife”