What spatial statistical model is best for predicting fisheries bycatch risk?

BRIAN STOCK
Thank you!

SIO
- Brice Semmens

SWFSC
- Tomo Eguchi

NWFSC
- Eric Ward
- Essential Fish Habitat (Blake Feist)
- West Coast Groundfish Observer Program (Jason Jannot)

PIFSC
- Hawaii Longline Observer Program (Eric Forney)
“Target” vs. “bycatch”

Introduction
“Target” vs. “bycatch”

Introduction

Oceana (2014)
Bycatch is a big (spatial) issue

**Protected species**

**Competing fisheries**

**Recovering species**

**Unmarketable species**

**Introduction**
Difficult when they move so much...

Introduction
Static vs. dynamic management

Introduction
Static vs. dynamic management

Dynamic

1. Effectively protected?
2. Huge loss of fishing area

Static

Introduction

Oceana (2012)
Static vs. dynamic management

**Dynamic**

EXPERIMENTAL PRODUCT

Avoid fishing between solid black 62.6°F and 85.3°F lines to help reduce leatherback and loggerhead sea turtle interactions.

Avoid fishing between solid black 72.3°F and 74.1°F lines or east of 140°W to help reduce leatherback sea turtle interactions.

**Static**

Howell et al. (2008, 2015)
Tools for dynamic management

Need map of bycatch risk
Introduction

Tools for dynamic management

Need map of bycatch risk
Tools for dynamic management

Need map of bycatch risk

- temperature
- depth
- substrate
- spatial field

Introduction
Q1: Which spatial model is best?

• temperature
• depth
• substrate
• spatial field

GLM
GAM
GMRF
RF

Research question
The data (fisheries observers)

West Coast Groundfish Trawl
- 2002-2013
- 55,835 tows

Hawaii Longline
- 1994-2014
- 16,714 sets (swordfish only)

Research question
Q2: Does the answer depend on species?

<table>
<thead>
<tr>
<th>Species</th>
<th>Habitat</th>
<th>Movement</th>
<th>Bycatch Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fish</td>
<td>Bottom</td>
<td>Med</td>
<td>29%</td>
</tr>
<tr>
<td>Habitat</td>
<td>Rocky bottom</td>
<td>Low</td>
<td>18%</td>
</tr>
<tr>
<td>Movement</td>
<td>Rocky bottom</td>
<td>Low</td>
<td>0.3%</td>
</tr>
<tr>
<td>Bycatch Rate</td>
<td>Open ocean</td>
<td>High</td>
<td>96%</td>
</tr>
<tr>
<td>Habitat</td>
<td>Open ocean</td>
<td>High</td>
<td>1.4%</td>
</tr>
<tr>
<td>Movement</td>
<td>Open ocean</td>
<td>High</td>
<td>0.7%</td>
</tr>
</tbody>
</table>

Research question
Q2: Does the answer depend on species?

<table>
<thead>
<tr>
<th>Habitat</th>
<th>Movement</th>
<th>Bycatch Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom</td>
<td>Med</td>
<td>29%</td>
</tr>
<tr>
<td>Rocky bottom</td>
<td>Low</td>
<td>18%</td>
</tr>
<tr>
<td>Rocky bottom</td>
<td>Low</td>
<td>0.3%</td>
</tr>
<tr>
<td>Open ocean</td>
<td>High</td>
<td>96%</td>
</tr>
<tr>
<td>Open ocean</td>
<td>High</td>
<td>1.4%</td>
</tr>
<tr>
<td>Open ocean</td>
<td>High</td>
<td>0.7%</td>
</tr>
</tbody>
</table>

Research question
Q2: Does the answer depend on species?

<table>
<thead>
<tr>
<th>Habitat</th>
<th>Movement</th>
<th>Bycatch Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom</td>
<td>Med</td>
<td>29%</td>
</tr>
<tr>
<td>Rocky bottom</td>
<td>Low</td>
<td>18%</td>
</tr>
<tr>
<td>Rocky bottom</td>
<td>Low</td>
<td>0.3%</td>
</tr>
<tr>
<td>Open ocean</td>
<td>High</td>
<td>96%</td>
</tr>
<tr>
<td>Open ocean</td>
<td>High</td>
<td>1.4%</td>
</tr>
<tr>
<td>Open ocean</td>
<td>High</td>
<td>0.7%</td>
</tr>
</tbody>
</table>
Q2: Does the answer depend on species?

<table>
<thead>
<tr>
<th>Habitat</th>
<th>Movement</th>
<th>Bycatch Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom</td>
<td>Med</td>
<td>29%</td>
</tr>
<tr>
<td>Rocky bottom</td>
<td>Low</td>
<td>18%</td>
</tr>
<tr>
<td>Rocky bottom</td>
<td>Low</td>
<td>0.3%</td>
</tr>
<tr>
<td>Open ocean</td>
<td>High</td>
<td>96%</td>
</tr>
<tr>
<td>Open ocean</td>
<td>High</td>
<td>1.4%</td>
</tr>
<tr>
<td>Open ocean</td>
<td>High</td>
<td>0.7%</td>
</tr>
</tbody>
</table>

Research question
Q3: How much bycatch can they prevent?
“Species distribution models”

Fundamental ecological question: *where are they?*

- temperature
- depth
- substrate
- spatial field
“Zero-inflated” data

More zeros than expected

Zuur et al. (2009)
“Zero-inflated” data

Approach 1: Zero-inflated distributions

- ZI-Poisson
- ZI-Neg Binomial
“Zero-inflated” data

Approach 2: Delta (hurdle) model

Methods

Zuur et al. (2009)
“Zero-inflated” data

Approach 2: Delta (hurdle) model

\[ \logit(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = X_i\beta \]

\[ Y_i \sim Bernoulli(p_i) \]
“Zero-inflated” data

Approach 2: Delta (hurdle) model

Pr(some bycatch)

\[ E(\text{bycatch} \mid \text{some bycatch}) \]

\[ \log(\mu_i) = X_i \beta \]

\[ Y_i \sim \text{Gamma}(\mu_i, \nu) \quad \text{for} \quad Y_i > 0 \]
“Zero-inflated” data

Approach 2: Delta (hurdle) model

- Pr(some bycatch)
- E(bycatch | some bycatch)
- E(bycatch)

Methods
Q1: Which spatial model is best?

- temperature
- depth
- substrate
- spatial field

Methods

GLM  GAM  GMRF  RF
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>

Methods: evaluation
Q1: Which spatial model is best?

Goal: prediction
5-fold cross validation repeated 10x

Binomial

Positive

AUC

RMSE, $R^2$ (pred – obs)

$$\sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$
Q3: How much bycatch can they prevent?

Simulate management:

1. Predict bycatch risk at test locations
Q3: How much bycatch can they prevent?

Simulate management:
1. Predict bycatch risk at test locations
2. Remove X% of fishing effort with highest bycatch risk

Methods: evaluation
Q3: How much bycatch can they prevent?

Simulate management:

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)
Q1: Which spatial model is best?

- **GLM**
- **GAM**
- **GMRF**
- **RF**

Observations are modeled as a function of environmental predictors (temp, depth, ...):

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

- \( Y_i \sim Bernoulli(\logit^{-1}[X_i\beta]) \) (Binomial)
- \( Y_i \sim Gamma(e^{X_i\beta}, \nu) \) (Positive)

Dormann (2007)
Q1: Which spatial model is best?

obs \sim \text{environmental predictors (temp, depth, ...)}

Y_i \sim \text{Bernoulli}(\logit^{-1}[X_i \beta])

Y_i \sim \text{Gamma}(e^{X_i \beta}, \nu)

How much variability can we explain?

- with covariates
- without spatial locations
Q1: Which spatial model is best?

- GLM
- GAM
- GMRF
- RF

Environmental predictors (temp, depth, ...)

\[
\begin{align*}
\text{obs} & \sim \text{environmental predictors (temp, depth, ...)} \\
Y_i & \sim \text{Bernoulli}(\logit^{-1}[X_i\beta]) \\
Y_i & \sim \text{Gamma}(e^{X_i\beta}, \nu)
\end{align*}
\]

Problem:
spatial correlation in residuals (Prediction – Observed)

Dormann (2007)
Why does spatial correlation matter?

1. Valid statistical inference
   - Observations not independent
   - Lower effective sample size (i.e. CI should be wider)
Why does spatial correlation matter?

2. Get the temporal trend right

Methods

Shelton et al. (2014)
Why does spatial correlation matter?

2. Get the temporal trend right

Methods

Shelton et al. (2014)
Why does spatial correlation matter?

2. Get the temporal trend right

[Image of fish and map with graph showing density over time with spatial and no space markers]
3. Effect of habitat vs. schooling

Why does spatial correlation matter?

Methods

Agostini et al. (2008)
3. Effect of habitat vs. schooling

Why does spatial correlation matter?

Methods

Agostini et al. (2008)
Why does spatial correlation matter?

3. Effect of habitat vs. schooling

![Graph showing the effect of habitat on hake density with and without space.]
Q1: Which spatial model is best?

Generalized Additive Models

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

- Common, simple approach
- Parameterized by spline basis functions (not spatial correlation)

Methods

Wood (2006)
Q1: Which spatial model is best?

Gaussian Markov random field (GMRF)

Methods

Lindgren, Rue, & Lindstrom (2011)
Q1: Which spatial model is best?

**Gaussian Markov random field**
- Models *covariance* as function of spatial locations
- \( \text{obs} \sim \text{environmental predictors} + \text{MVN}(0, \Sigma) \)
Q1: Which spatial model is best?

Gaussian Markov random field

- Models covariance as function of spatial locations:
  \[
  \text{obs} \sim \text{environmental predictors} + \text{MVN}(0, \Sigma) \]

Problem...

- \(\Sigma\) has \(O(N^2)\) elements
- Computations scale as \(O(N^3)\) from \(|\Sigma|\) and \(\Sigma^{-1}\)

Methods

Lindgren, Rue, & Lindstrom (2011)
Q1: Which spatial model is best?

Gaussian Markov random field

- Models covariance as function of spatial locations
- \( \text{obs} \sim \text{environmental predictors} + \text{MVN}(0, \Sigma) \)

Solution:

- correlation = 0 for “far away” points \( \rightarrow \) sparse matrix

Methods

Lindgren, Rue, & Lindstrom (2011)
Q1: Which spatial model is best?

Gaussian Markov random field

- Models covariance as function of spatial locations

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

- Discrete approximation of continuous space

Methods

Lindgren, Rue, & Lindstrom (2011)
Q1: Which spatial model is best?

Gaussian Markov random field
- Models *covariance* as function of spatial locations
  \[ \text{obs} \sim \text{environmental predictors} + \text{MVN}(0, \Sigma) \]
- Increasing adoption in fisheries

Methods

Shelton et al. (2014)
Q1: Which spatial model is best?

Gaussian Markov random field

- Models covariance as function of spatial locations
  \[ \text{obs} \sim \text{environmental predictors} + \text{MVN}(0, \Sigma) \]

- Increasing adoption in fisheries

---

Thorson et al. (2015)
Q1: Which spatial model is best?

Random Forest

Pr = 0.12

Depth >= 250 fm

Pr = 0.15
Temp >= 1

Pr = 0.11
Temp < 1

Pr = 0.21
Depth < 250 fm

Pr = 0.18

Methods

Breiman (2001)
Q1: Which spatial model is best?

- Machine learning, designed for prediction
- “Black box”
- Predictor-bycatch relationships not modeled
- No spatial field (add LAT, LON)

Breiman (2001)
Bycatch risk maps

Results

Binomial
Generally:

- GLM < GAM < GMRF < RF

AUC (test data) vs. Model

Results: Binomial
Generally:

- GLM < GAM < GMRF < RF

Less clear for rarer species

N+ = 7,660
18%

N+ = 143
0.3%
Generally:

```
GLM < GAM < GMRF < RF
```
Q3: 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>

Results
**Q1: Which spatial model best predicts bycatch?**

GLM < GAM < GMRF < RF

**Q2: Does the answer depend on species?**

No, RF had consistent advantage

(larger for species with higher bycatch rates)

**Q3: How much bycatch can they prevent?**

Enough to consider using them in management
Discussion

If the goal is purely *prediction*:

GLM < GAM < GMRF < RF

...but if we care about *inference on processes* affecting bycatch:
Covariate effects

Discussion

Palczewksa (2013), Welling (2016)
Covariate effects

Are random forests really “black boxes”? 

GMRF

RF

Survey

Depth

In RCA

Discussion

Palczewksa (2013), Welling (2016)
Can random forests do better?

Identifying covariate interactions

Discussion
Thank you!

SIO
  ◦ Brice Semmens

SWFSC
  ◦ Tomo Eguchi

NWFSC
  ◦ Eric Ward
  ◦ Essential Fish Habitat (Blake Feist)
  ◦ West Coast Groundfish Observer Program (Jason Jannot)

PIFSC
  ◦ Hawaii Longline Observer Program (Eric Forney)
Research opportunities in applied math/statistics and fisheries science
We can easily harvest too many fish

McCauley et al. 2015
We use models in management

1. Sustainable harvest → need to assess populations
1. Sustainable harvest $\rightarrow$ need to assess populations

2. Primarily, *how many* and *where*
Build & test population models

1. Stock assessment
2. Simulate alternative harvest strategies
What are the effects on fish of:

1. Ocean productivity?
Work with physics/climate modelers

What are the effects on fish of:
1. Ocean productivity?
2. Dispersal of eggs and larvae?
Work with physics/climate modelers

What are the effects on fish of:

1. Ocean productivity?
2. Dispersal of eggs and larvae?
3. Range shifts?
How to gauge model performance?

Goal: prediction
5-fold cross validation repeated 10x

- Binomial
- ROC curve (AUC)
- Positive
- RMSE
West Coast Groundfish covariates

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

Chapter 2: Bycatch prediction

Shelton et al. (2014)
Hawaii Longline covariates

\[ \text{Binomial} \sim \text{sst} + \text{sst}^2 + \]
\[ \text{Positive} \quad \text{day of year} + \]
\[ \text{spatial field} \]

Shelton et al. (2014)
RF

+ Better at prediction
+ More complex covariate relationships (incl. interactions)
+ Much quicker to set up and run (~2 min vs. 5-15 hours)
+ Not just a “black box”?  

GMRF

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

Discussion
What is a “feature contribution”??

Covariate effects

Depth >= 250 fm

Depth < 250 fm

Temp >= 1

Temp < 1

Pr = 0.18

Pr = 0.12

Pr = 0.15

Pr = 0.11

Pr = 0.21

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

Prediction \( i \) = 0.11 = 0.18 – 0.06 (Depth) – 0.01 (Temp)