Use of machine-learning to estimate rare-event bycatch in the CA swordfish drift gillnet fishery

James V. Carretta
Jeffrey E. Moore
Karin A. Forney

Southwest Fisheries Science Center, La Jolla, CA

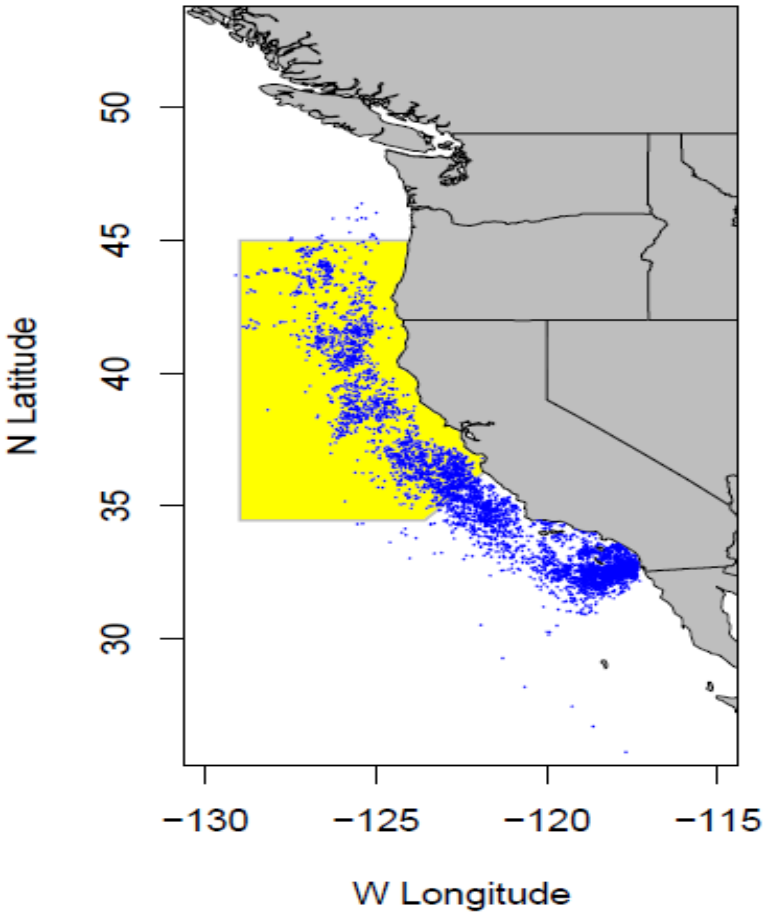
National Oceanic and Atmospheric Administration
U.S. Department of Commerce

Advances in bycatch estimation that serve the following partners:

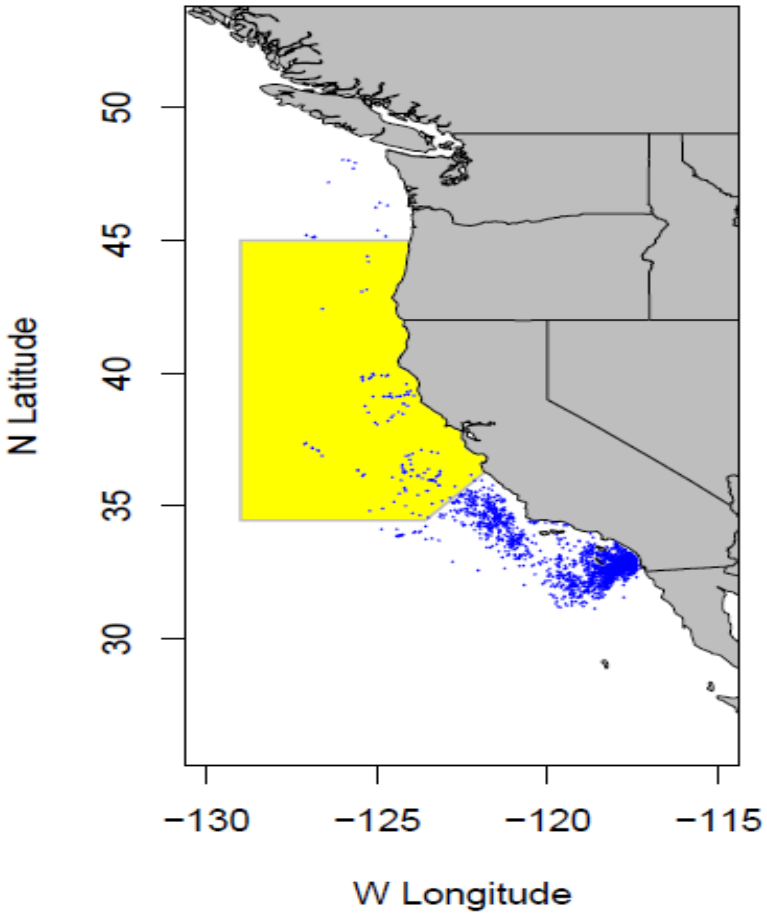
- A **Council** desire to address rare-event bycatch in the context of low observer coverage.
- A **Take Reduction Team** need for reliable bycatch estimates to effectively mitigate the fishery.
- An **NGO** desire to know if bycatch is occurring even in the absence of observations.
- A **DGN fleet** need for realistic bycatch estimates and identification of 'bycatch drivers' so that mitigation and management can be implemented appropriately while maintaining fishing opportunity.

Bycatch estimation: CA drift gillnet fishery

Observed sets 1990-2000 (n=5,973)



Observed sets 2001-2015 (n=2,738)



How we used to estimate DGN bycatch ... using within-year data and ratio estimates

- 20% observer coverage + rare events

Observed	Estimated
0	**0**
1	**5**
2	**10**
3	**15**

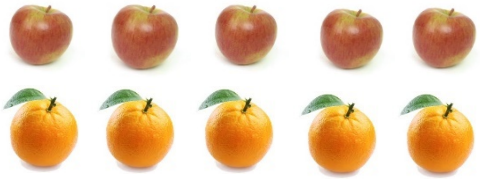
If true bycatch = 2 whales / 1,000 annual sets, then annual ratio estimates are always too low / high.

The response is to pool >=5 yrs of annual estimates to derive better mean estimates.

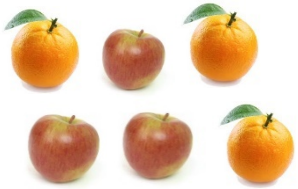
5 yrs of pooling isn't enough to reduce estimation bias (Carretta and Moore 2014).

Rare Event Bycatch = Needle in a haystack problem

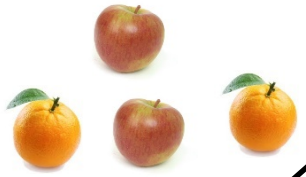
Classification Trees



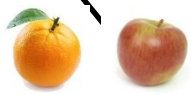
weight < 150 g | weight >= 150 g

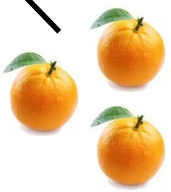


Calcium < 10 mg | Calcium >= 10 mg



Asymmetry < 0.5 | Asymmetry >= 0.5





Find variables that best reduce the variance of the response when used to split data.

Step 1: Evaluate variable importance from simulated bycatch data

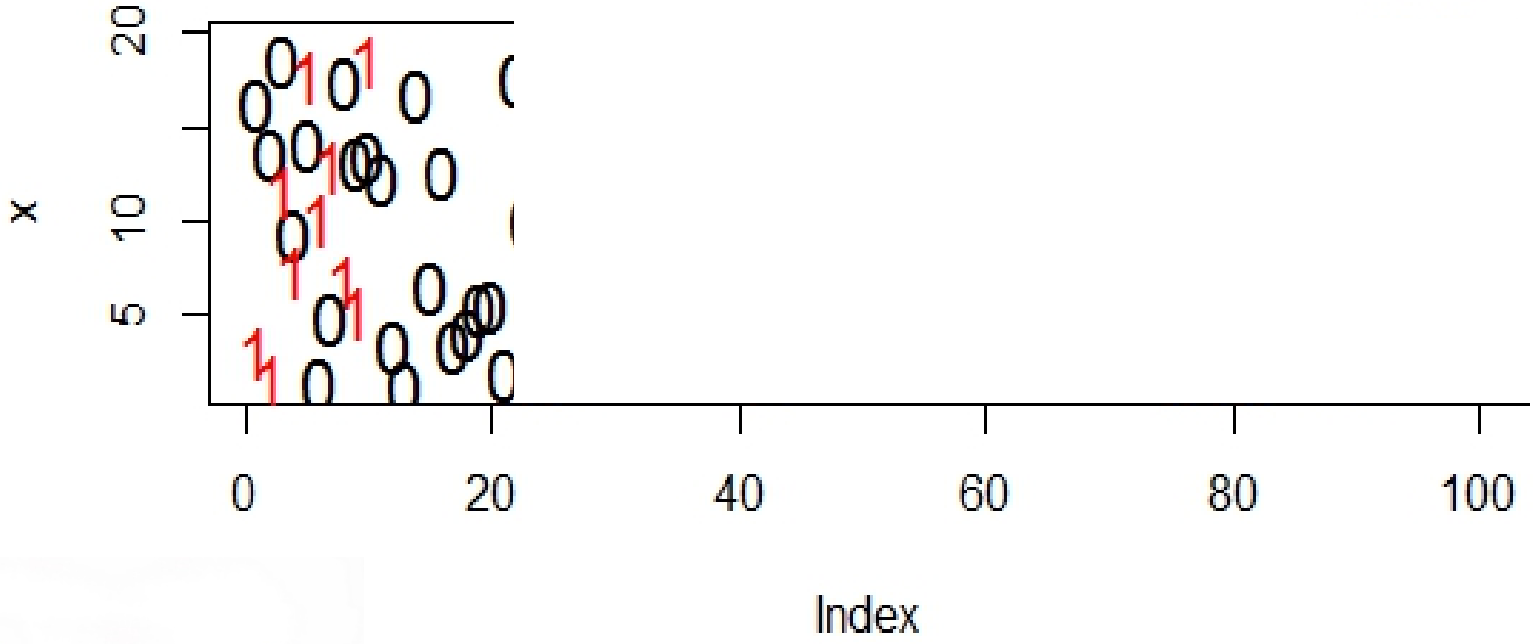
Simulations: 4 to 9 bycatch events from ~ 8,500 sets (30 simulated data sets).

Logistic model : increasing probability of bycatch as a function of a secret variable.

Can you ID the variable linked to bycatch and how often?

- Convert response data into classes (0 or 1).
- Create classification tree RF. If variable has no value, data are split randomly between resulting nodes.
- Variables with predictive value will 'purify' your response data. This is quantifiable through the *Gini Index*.
- Permute response data *n* times and measure Gini Index scores each time = null distribution of Gini metrics from which p-values of variables are derived.

Deal with zero-inflation by increasing contrast in data.

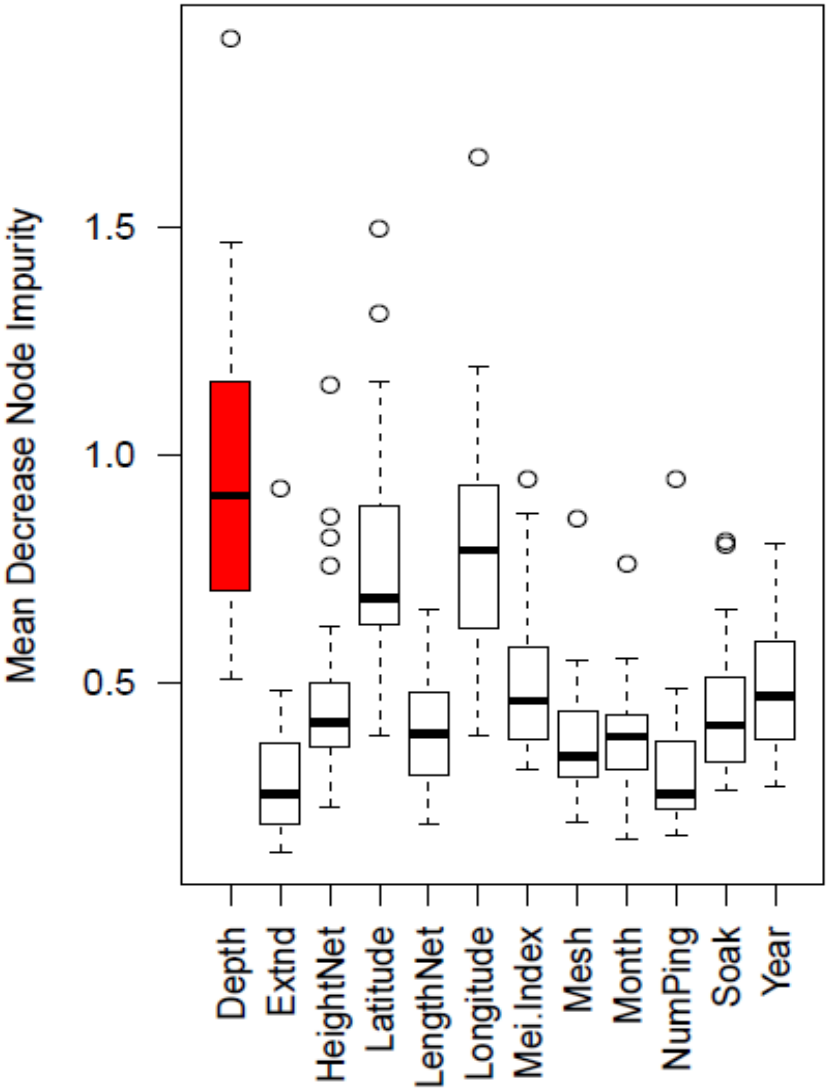


Balance sample sizes (via bootstrap) to increase signal-to-noise ratio.

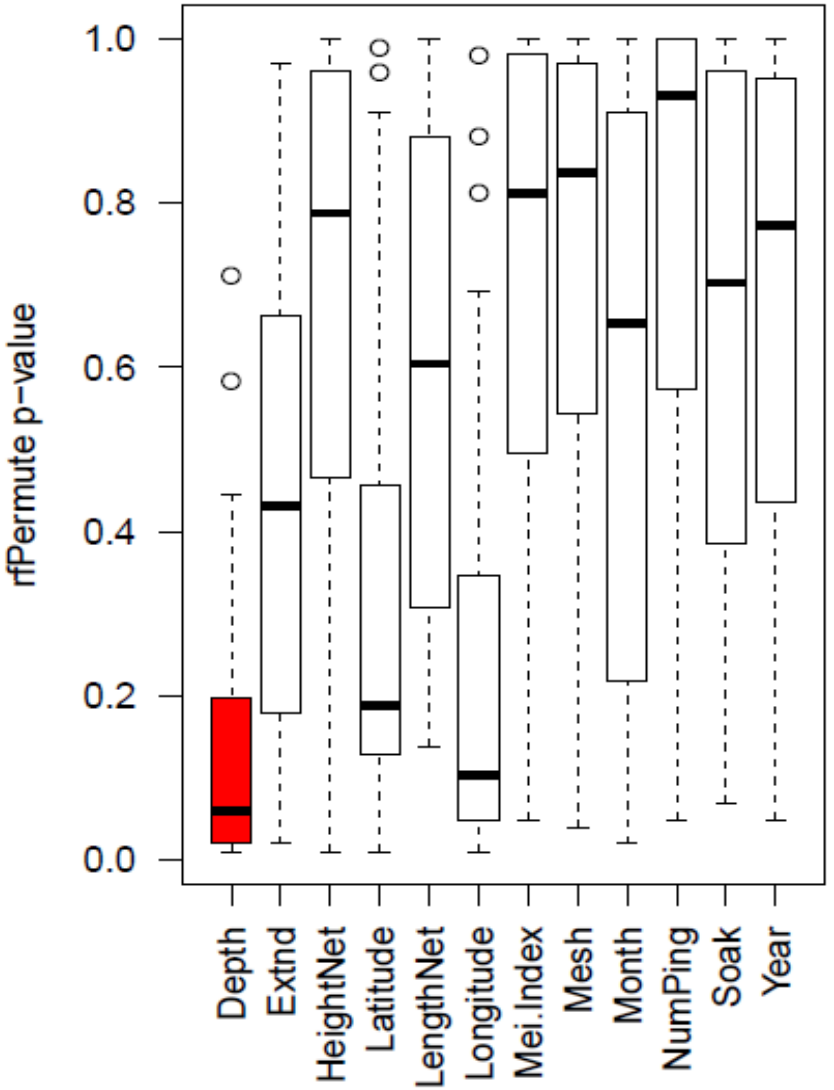
Xie, Y., Li X., Ngai E.W.T. and Ying, W. 2009. Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3), pp.5445-5449.

Variable selection: Simulated rare-event bycatch

Gini index metric



rfPermute p-value



R-package *rfPermute* (Archer 2016)

Model validation and variable selection

- **Performance testing on simulated bycatch data.**

  4 to 9 simulated bycatch events from ~ 8,500 sets.
  30 realizations of this bycatch process.

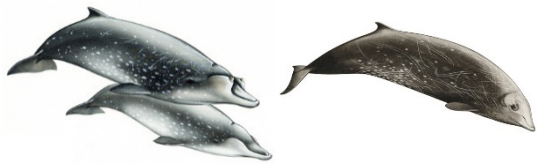
- **How often did randomForest correctly identify the 'secret variable' from a suite of 12 variables?**

  23/30 cases *depth* was ranked 1st or 2nd in predictive power. In 25/30 cases either *depth* or *longitude* was ranked as most important.

Cross-validated error rates with 10,000 balanced-sample classification trees

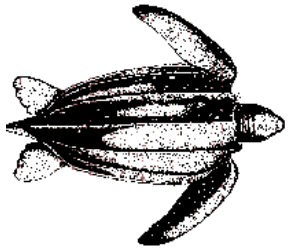
Error Rates for positive events: Observed (and expected by chance)

50% (~100%)



18% (99%)

61% (96%)

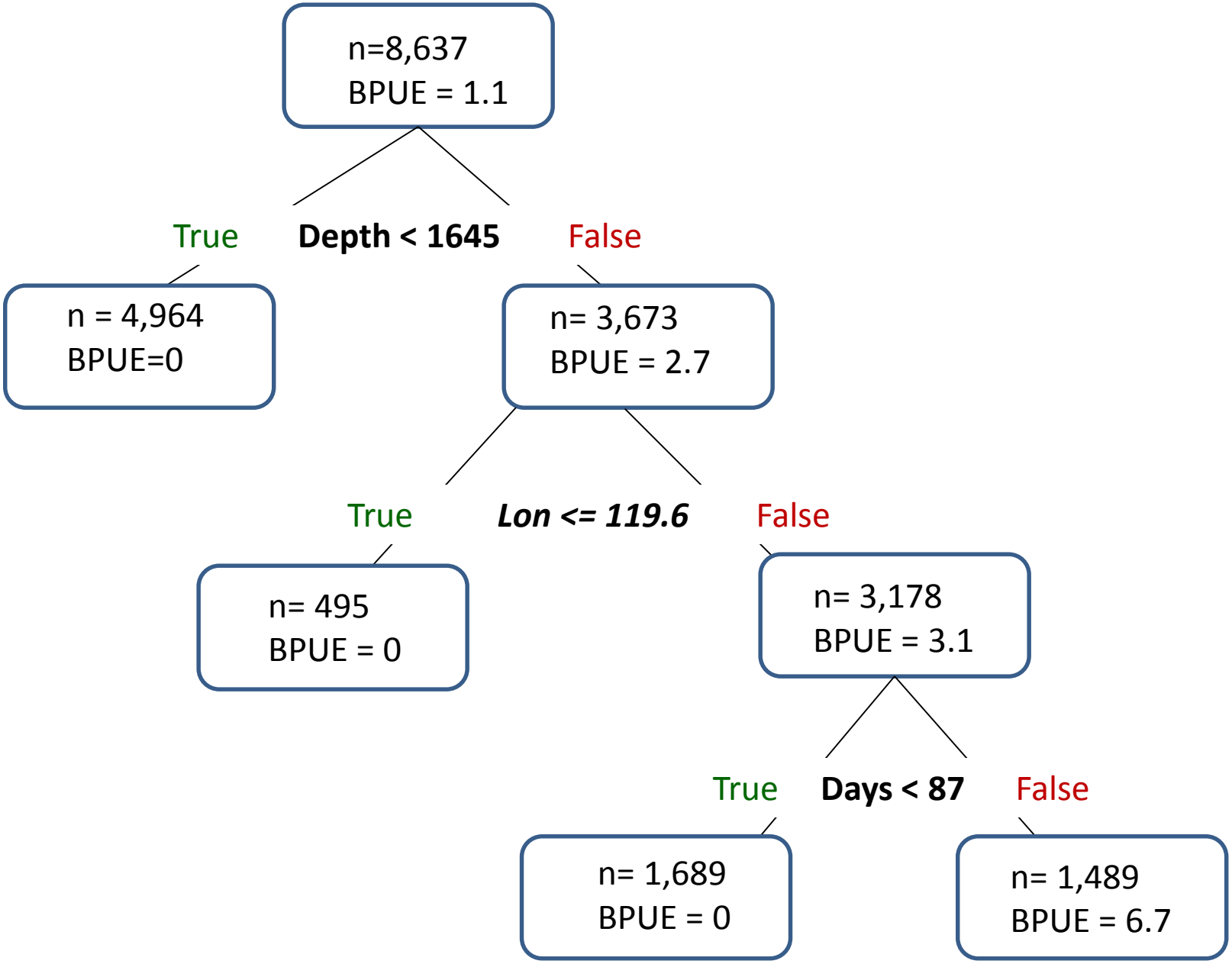


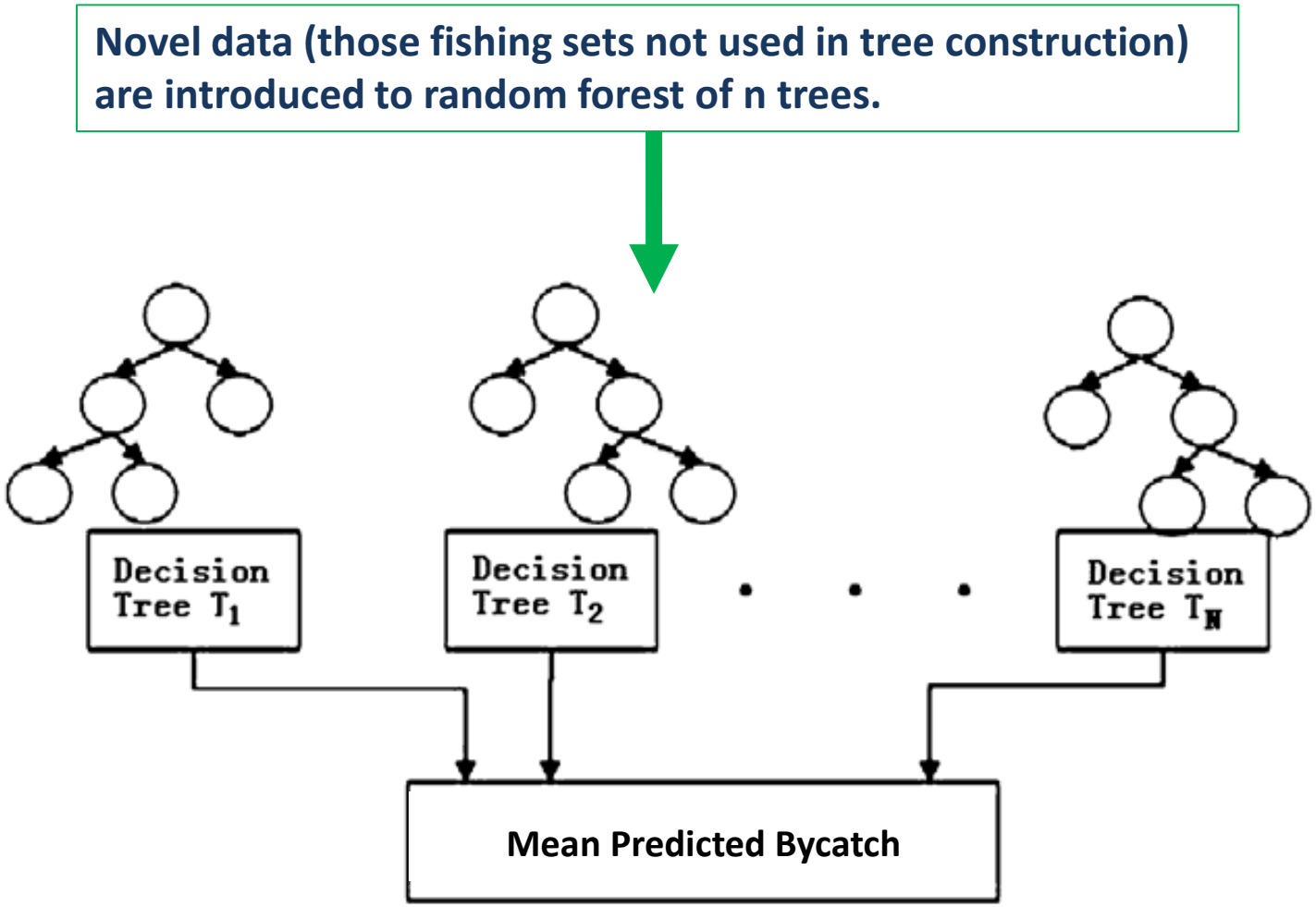
32% (99%)

40% (99%)

Step 2: Estimation

BPUE per 1,000 sets





Diversity of random forest predictions is a direct measure of estimate uncertainty.

Response is mean predicted bycatch per set.

Total Bycatch in year *y:*

$$T_y = \bar{b}_{s,y} * u_{s,y} + \sum o_{s,y}$$

mean predicted bycatch per set

unobserved sets

observed bycatch

Assumption that annual observer data is representative of unobserved fishing effort.

Bycatch rate is based on gear and environment characteristics of observed fishing sets in year *y* and a model that is informed by all 26 years of observer data.

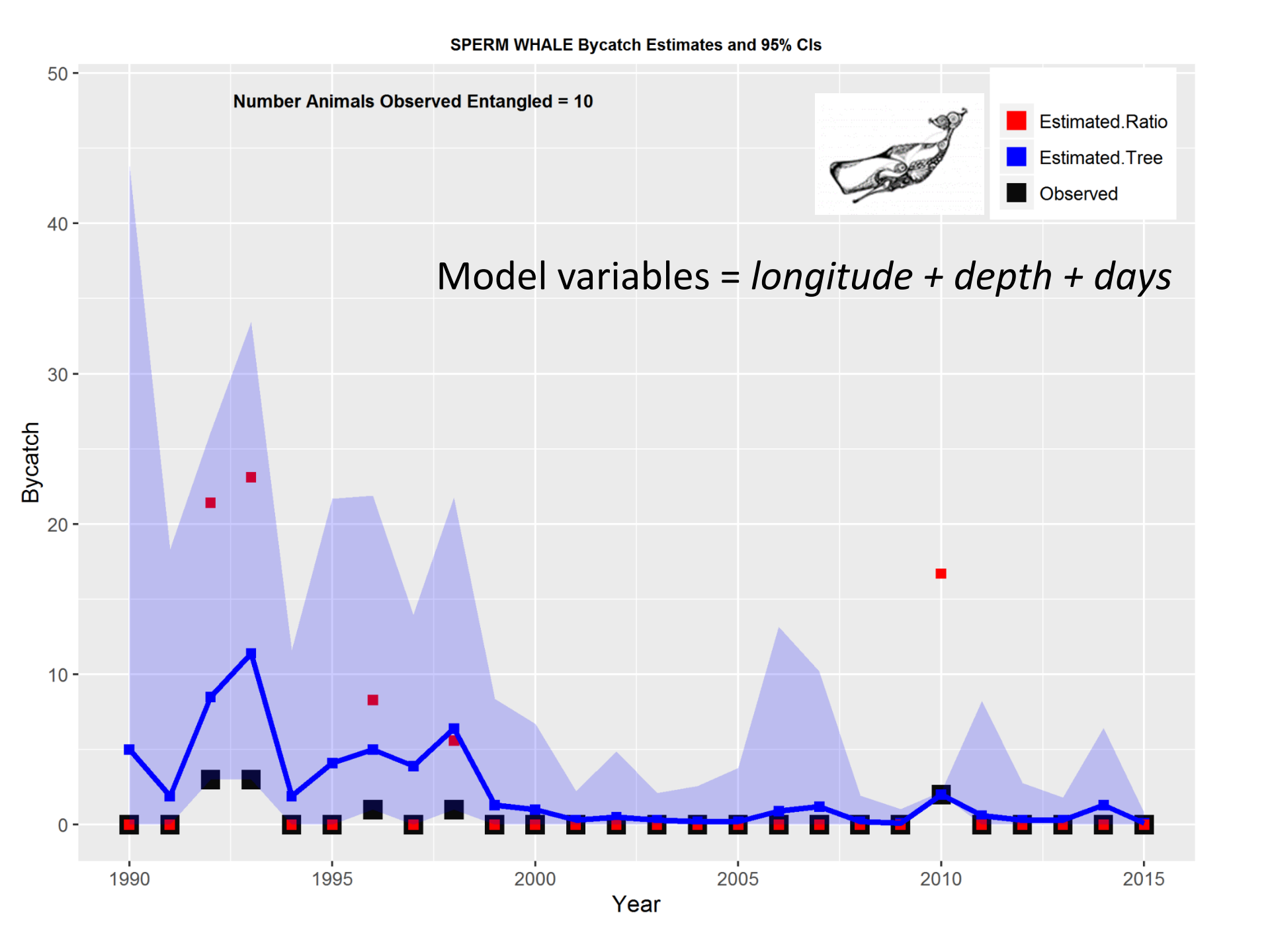
It is no longer based on within-year observations, which previously suffered from small sample size biases.

$$T_y = \bar{b}_{s,y} * u_{s,y} + \sum o_{s,y}$$

mean predicted bycatch per set

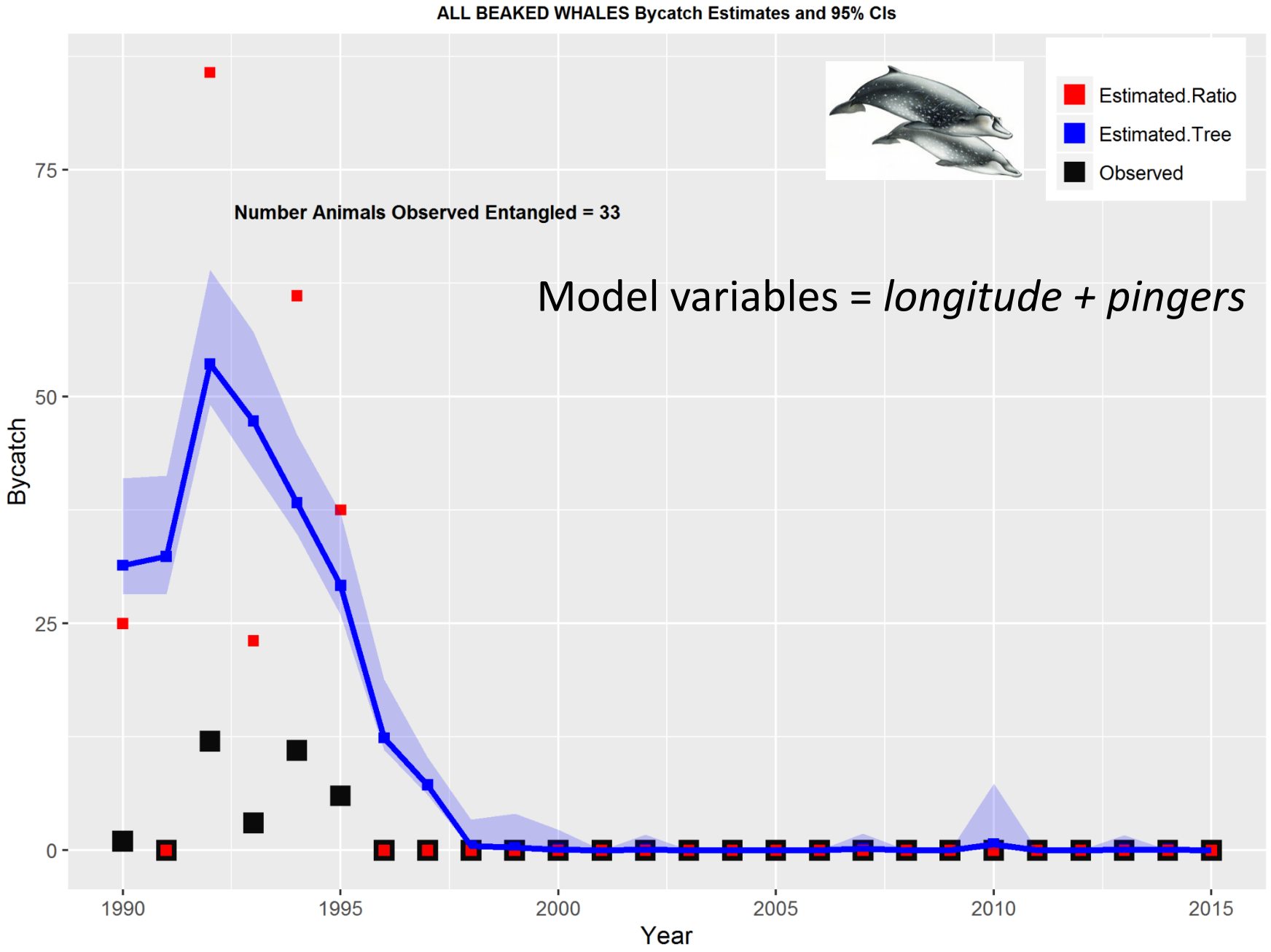
unobserved sets

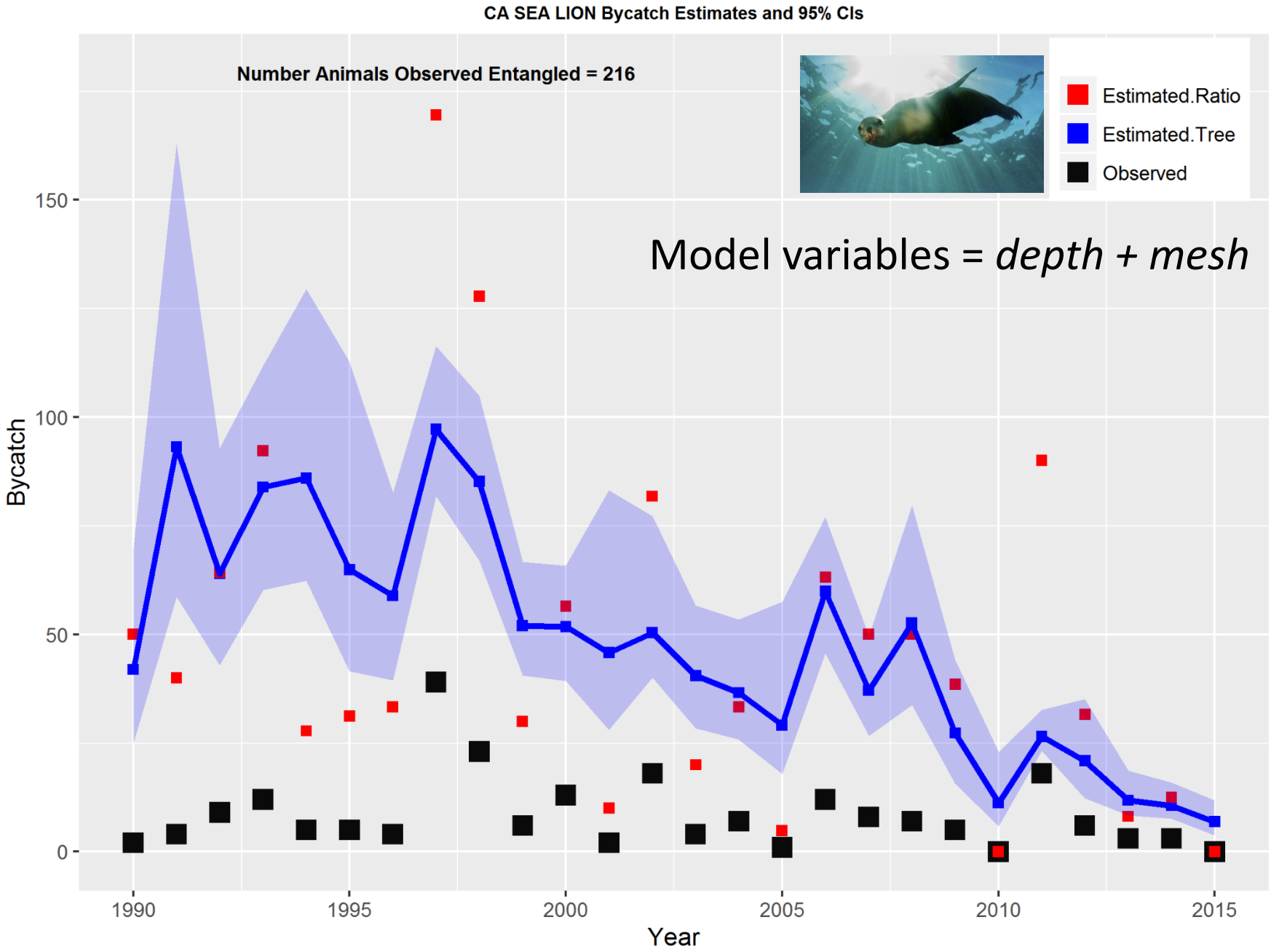
observed bycatch

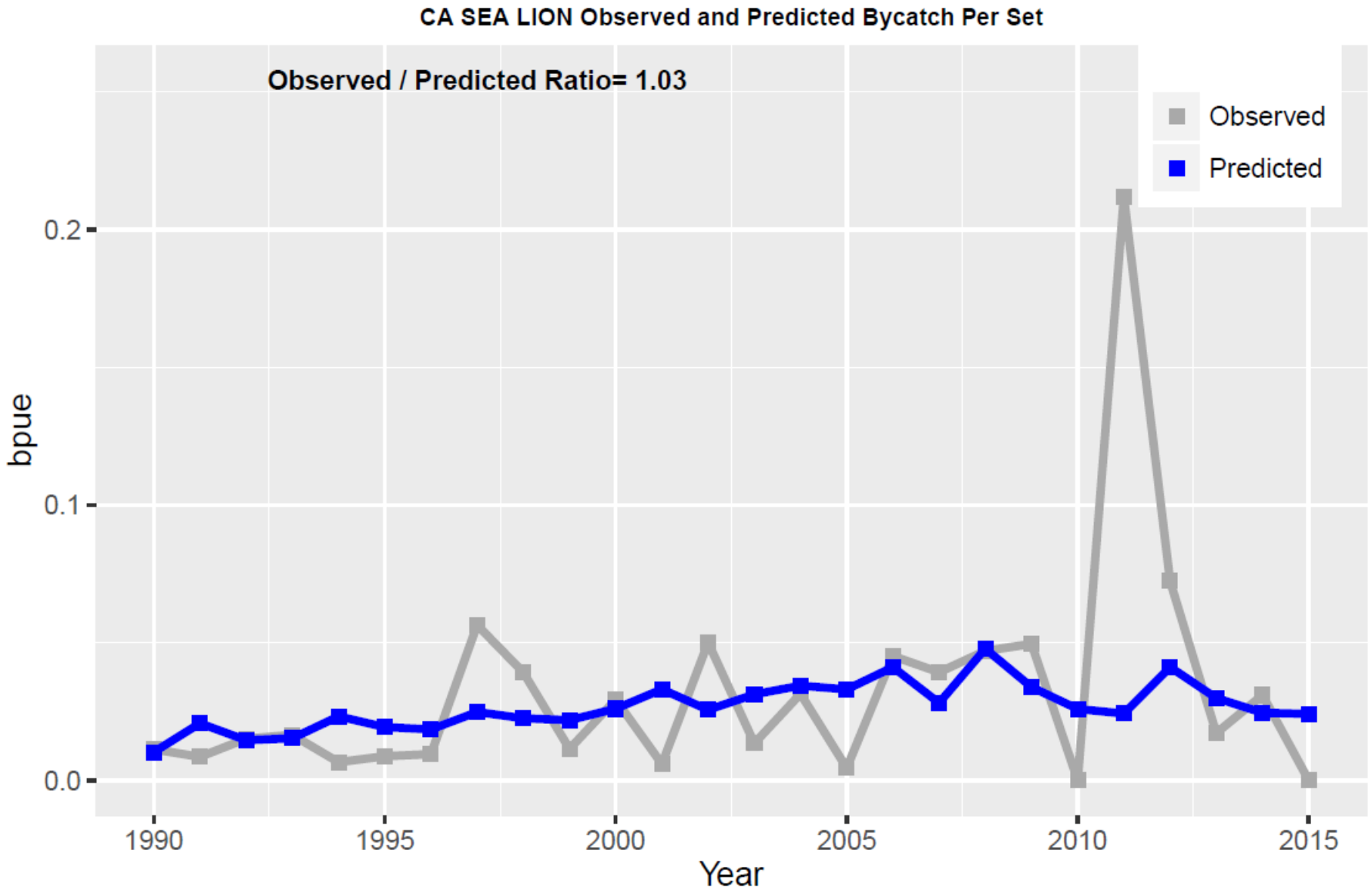

SPERM WHALE Bycatch Estimates and 95% CIs
Number Animals Observed Entangled = 10
Estimated.Ratio
Estimated.Tree
Observed
Model variables = longitude + depth + days
Bycatch
Year
50
40
30
20
10
0
1990
1995
2000
2005
2010
2015

How have estimates changed? Sperm whale example:

- 1990 – 1991: zero sperm whale entanglements observed from 648 fishing sets, during which approx. 9,000 total sets were fished.
- Resulting ratio estimate of bycatch was zero for both years.
- New tree-based estimate is 6.9 entanglements.
- Why? We have a better idea of the long-term bycatch rate for sperm whales. We could not have known this just from the 1st two years of observer data.
- Previous 2010 ratio estimate of 16 entanglements (2 observed w/ 12% observer coverage) is now 2 whales (2 observed + zero estimated in 433 unobserved fishing sets).
- Long-term, both estimate types converge on approximately 60 entanglements in 26 years, with 50 entanglements between 1990-1999.


ALL BEAKED WHALES Bycatch Estimates and 95% CIs
Estimated.Ratio
Estimated.Tree
Observed
Number Animals Observed Entangled = 33
Model variables = longitude + pingers
Bycatch
75
50
25
0
1990
1995
2000
2005
2010
2015
Year

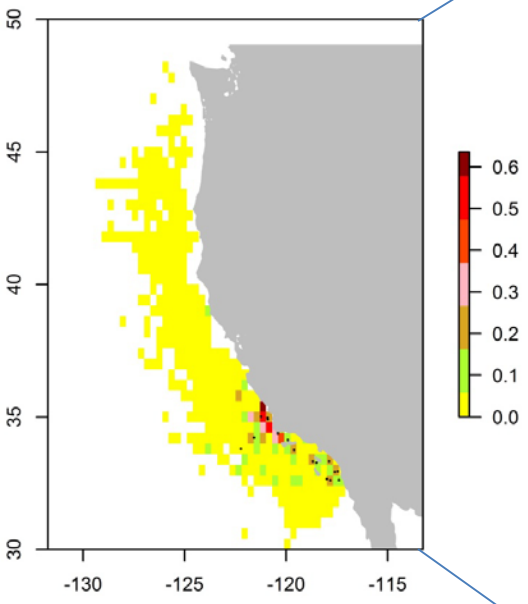

CA SEA LION Bycatch Estimates and 95% CIs
Number Animals Observed Entangled = 216
Estimated.Ratio
Estimated.Tree
Observed
Model variables = depth + mesh
Bycatch
150
100
50
0
1990
1995
2000
2005
2010
2015
Year



Increasing BPUE, reflecting sea lion population growth.

Spatial model validation

Delphinus capensis



Observed (●) and estimated bycatch

Model variables = *latitude + depth*

Caveats + Conclusions + References

You may not identify any significant predictors.

That's ok.

You can use a null model, which merely returns the mean BPUE.

There is nothing special about $p < 0.05$. An ensemble of 'weak' predictors can be powerful.

Archer, E., 2016. rfPermute: Estimate permutation p-values for Random Forest importance metrics. https://github.com/EricArcher/rfPermute

CARRETTA, JAMES V., JAY BARLOW, and LYLE ENRIQUEZ. Acoustic pingers eliminate beaked whale bycatch in a gill net fishery. MARINE MAMMAL SCIENCE 24, no. 4 (2008): 956-961.

Carretta, J.V. and J.E. Moore. 2014. Recommendations for pooling bycatch estimates when events are rare. NOAA Technical Memorandum, NOAA-NMFS-TM-SWFSC-528.

Xie, Y., Li, X., Ngai, E.W.T. and Ying, W., 2009. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, *36*(3), pp.5445-5449.