Early look: Performance of machine learning to classify age from age-length data

Caleb A. Aldridge[1*], Michael E. Colvin[1], and Leandro E. Miranda[2]

[1]Department of Wildlife, Fisheries & Aquaculture, Mississippi State University, Mississippi State, MS 39762-9690
[2]Mississippi Cooperative Fish & Wildlife Research Unit, Mississippi State, MS 39762-9691
[*]Corresponding author: caa134@misstate.edu

Introduction

- Fish length and age are common measurements used in growth, mortality, and recruitment
- Estimating age of individual fish is costly, therefore age structure is estimated
  - Done using a probability matrix of age and length-class to partition individuals (age-length key [ALK])
  - Aging error can occur at multiple stages (e.g., aging mistakes, many ages represented in a length-class)
- Covariates can help more accurately assign age to fish
  - Additional data collection can be resource limited
  - Likely that only length and age data will be used

Objective

To evaluate if naïve Bayes classifier (NB) is more sensitivity (% of correct classifications) than ALK.

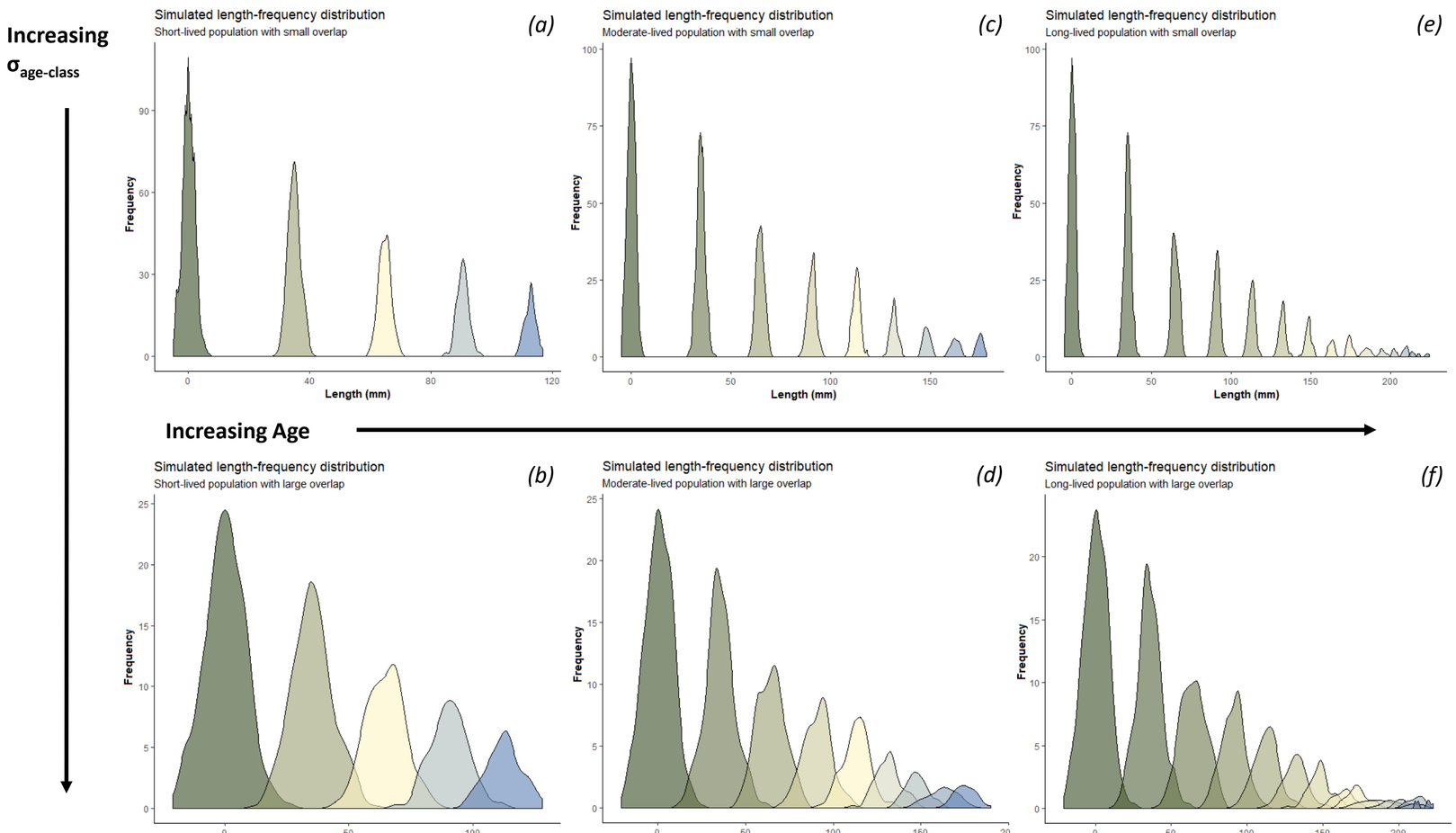
Methods

- Simulated age-length data sets (**Fig. 1, *a–f***).

Constant:
- Initial cohort (N_0 = 500)
- Mortality (A_{annual} = 0.3)
- Growth coefficient (K = 0.1)
- Time coefficient (t_0 = 0)
- Asymptotic length (L_∞)

Varied:
- Maximum age (t_{max})
  - 4, 8, and 16
- Age-class variance ($\sigma_{age\text{-}class}$)
  - 2 and 8

- 10 × repeated 5-fold cross-validation
  - Generating 600 observations
    - 50 per classifier per data set
- Compared mean sensitivity of ALK and NB
  - Observed difference to resampling differences
- Compared sensitivity of ALK and NB—empirical data (**Fig. 1, *g***)
- Constructed von Bertalanffy growth curves—empirical data
  - Compared curves built from data assigned by each method



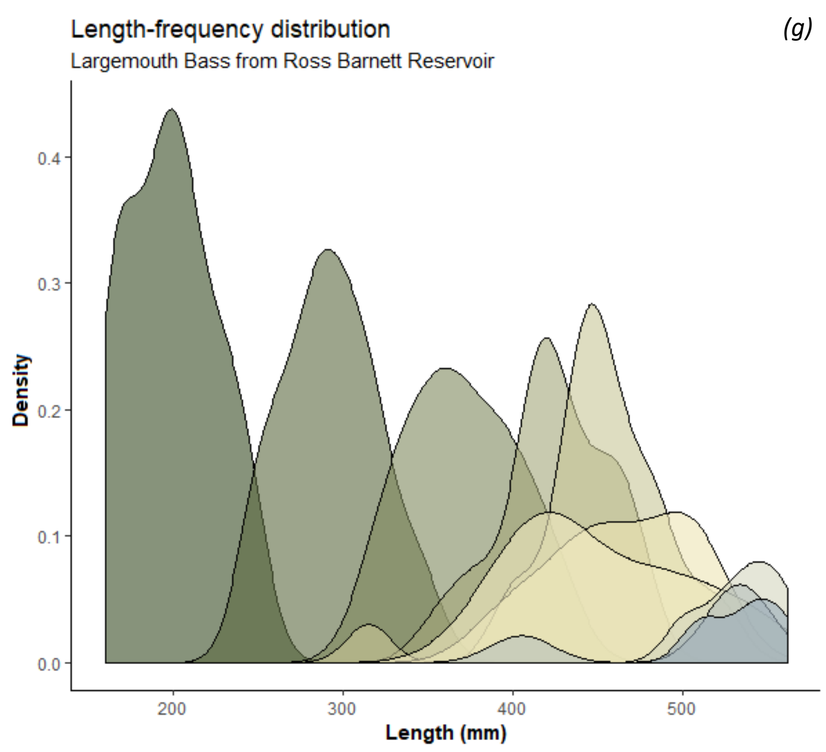


Figure 1. *Simulated (a–f) and empirical (g) data sets used in our analyses.*

Results

- NB classifier had higher sensitivity
  - Across all data sets (all P lower than 0.005)
  - On average 9% higher
- Differences less pronounced with lower $\sigma_{age\text{-}class}$
- Differences less pronounced with lower t_{max} (**Fig. 2**)
- Differences negligible for the empirical data (ALK = 52% vs. NB = 53%)
- Graphically, von Bertalanffy growth curves from ALK classified ages had less bias from complete empirical curve (**Fig. 3**)

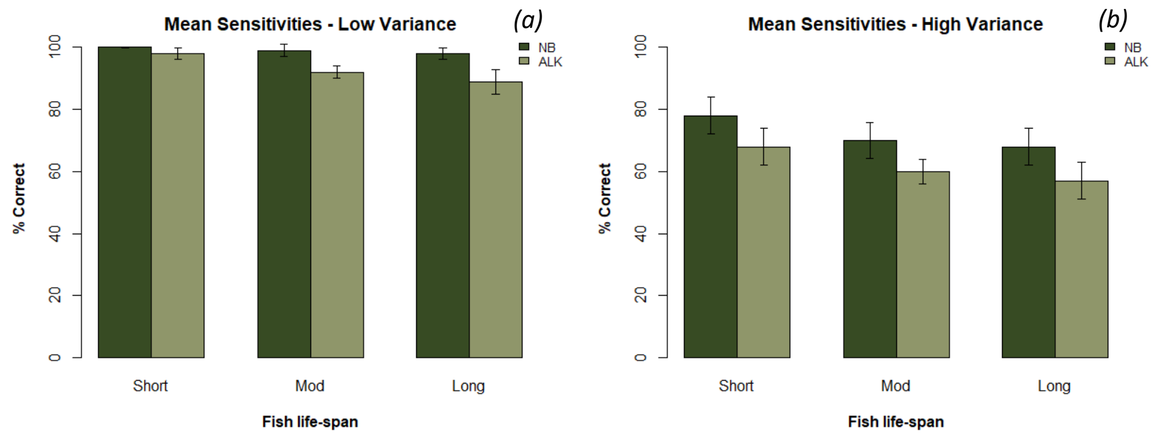


Figure 2. *Average sensitivity (% correct) of 10 × repeated 5-fold cross validation: age-length key (ALK) vs. naïve Bayes (NB) age classification. Simulated data sets with three life-spans (x-axis) and age-class variance (a and b). Error bars represent upper and lower 95% confidence intervals.*

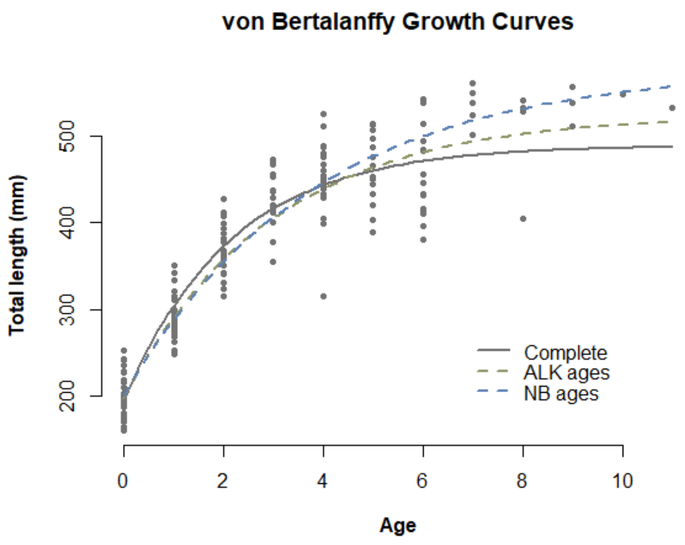


Figure 3. *Comparison of von Bertalanffy growth curves constructed using iterative nonlinear weighted least squares on complete empirical aged data (Complete), age-length key age classified data (ALK ages), and naïve Bayes age classified data (NB ages).*

Discussion

- Naïve Bayes, higher sensitivity but perhaps only for younger age-classes
- Further inspection, NB truncated older (single observations) into lower age-class (multiple observations)
- Additionally, NB assumes Gaussian distribution and misclassified some outliers
- Limitations in interpretation due to one empirical data set
- More analyses are needed to compare the robustness classifiers when dealing with low-observation-at-old-age data—an attribute of most age-length data sets

Future Directions

- Classifiers might be combined
  - Class observation size to decide assigning classifier for each class
  - Build many classification models and "average" (i.e., bootstrap aggregating)

Closing Remarks

- Length and age will continue to be recorded and used in assessing population parameters
- Decisions will continue to be made from graphs and statistics built on data that may inherit aging mistakes
- Improving age classification is one way for improving fisheries management

Your Thoughts

Do you think the growth curves build from the two different classifiers differ? Let us know by scanning the QR code to the left or going to the URL below and answering five short questions.

https://goo.gl/forms/iT5UaNnrCIq3vkT73