K-nearest neighbors method for prediction of natural mortality rates

Amy Y. Then\textsuperscript{1}, John M. Hoenig\textsuperscript{2}

\textsuperscript{1}Institute of Biological Sciences, Faculty of Science, University of Malaya, Malaysia
\textsuperscript{2}Virginia Institute of Marine Science, Gloucester Pt., VA
Introduction

- Natural mortality $M$ – important & highly influential stock parameter
- True $M$ unknown – direct $M$ estimates are the best information; a ‘constant’ $M$ still useful
- Data-poor species – ‘borrow’ strength from other species (indirect/ empirical methods)
- Common predictor variables: Von Bertalanffy growth parameters ($L_\infty, K$), maximum age ($t_{max}$), mean water temperature ($Temp$)

rougheyne rockfish

Goldman's Goby
Introduction


Original Article

Evaluating the predictive performance of empirical estimators of natural mortality rate using information on over 200 fish species

Amy Y. Then¹,²*, John M. Hoenig¹, Norman G. Hall³,⁴, and David A. Hewitt⁵

- Alverson & Carney (1975) $K, t_{\text{max}}$
- Pauly (1980) $K, L_\infty, \text{Temp}$
- Hoenig (1983) $t_{\text{max}}$
- Jensen (1996) $K$
- one-parameter $t_{\text{max}}$ ($M = c/ t_{\text{max}}$)
**Introduction**

**Original Article**

Evaluating the predictive performance of empirical estimators of natural mortality rate using information on over 200 fish species

Amy Y. Then, John M. Hoenig, Norman G. Hall, and David A. Hewitt

**Rank:** one-parameter $t_{max} \approx$ Hoenig

Alverson & Carney

Jensen $\approx$ Pauly

**Recommendation:**

\[ M = 4.899 \left( t_{max} \right)^{-0.916} \]

\[ M = 4.118 K^{0.73} L_{\infty}^{-0.33} \]
Introduction

- Commonly used approach: fit parametric models, e.g. linear regression (REG) – may pose issues with assumptions
- Non-parametric approaches – may offer more efficient or robust estimators & less bias

Then et al. (2015)
Objectives

Can we do better with empirical estimation of $M$?

- Improve prediction error & residuals
- ‘Benchmark’ regression (REG) models:
  - $t_{max}$-based: $M = a \ t_{max}^b$
    
    $a = 4.899, b = -0.916, n = 226$
  
  - Growth-based: $M = a \ K^b L_\infty^c$
    
    $a = 4.118, b = 0.73, c = -0.33, n = 218$
KNN method

- KNN: k-nearest neighbor
- A type of machine learning (supervised) technique
- Works well for low-dimensionality problems
- Close neighbors are used to predict a given point
- No of neighbors \((k)\) – key parameter
- Distance metric & kernel function

k=4
Method

Training dataset (n = 215)

Test datasets (n = 44, 35)

KNN candidate models

Model tuning & selection

Best KNN models
- $t_{max}$
- growth

Model evaluation & validation

KNN versus REG models

**kknn** package in R

- 10-fold cross validation prediction error (CVPE)
- $k$: 2 to 20 neighbors
- Distance: Euclidean, Manhattan
- Kernel: rectangular, triangular, Gaussian
- Residual patterns
- Stocks with literature $t_{max} > 7$ yr (78%)
Datasets description

Training (n = 215) versus test datasets

\[ t_{max} \] (n = 44)

- \( M \): 0.014 to 5.07 yr\(^{-1} \)
- \( t_{max} \): 0.88 to 205 yr

\[ K \] (log-scale)

- \( K \): 0.01 to 2.6 yr\(^{-1} \)
- \( L_\infty \): 48.5 to 3164 mm
## Results

### Distance = Manhattan

<table>
<thead>
<tr>
<th>KNN models</th>
<th>Rectangular</th>
<th>Triangular</th>
<th>Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k</td>
<td>CVPE</td>
<td>k</td>
</tr>
<tr>
<td>&quot;M ~ K + Linf + tmax&quot;</td>
<td>4</td>
<td>0.470</td>
<td>4</td>
</tr>
<tr>
<td>&quot;M ~ K + tmax&quot;</td>
<td>3</td>
<td>0.486</td>
<td>5</td>
</tr>
<tr>
<td>&quot;M ~ Linf + tmax&quot;</td>
<td>7</td>
<td>0.328</td>
<td>8</td>
</tr>
<tr>
<td>&quot;M ~ tmax&quot;</td>
<td>6</td>
<td>0.301</td>
<td>4</td>
</tr>
<tr>
<td>&quot;M ~ K + Linf&quot;</td>
<td>12</td>
<td>0.597</td>
<td>8</td>
</tr>
<tr>
<td>&quot;M ~ Linf&quot;</td>
<td>15</td>
<td>0.630</td>
<td>15</td>
</tr>
<tr>
<td>&quot;M ~ K&quot;</td>
<td>12</td>
<td>0.639</td>
<td>10</td>
</tr>
<tr>
<td>&quot;M ~ tmax + Temp&quot;</td>
<td>2</td>
<td>0.400</td>
<td>4</td>
</tr>
<tr>
<td>&quot;M ~ K + Linf + Temp&quot;</td>
<td>6</td>
<td>0.559</td>
<td>11</td>
</tr>
<tr>
<td>&quot;log(M) ~ log(K) + log(Linf)&quot;</td>
<td>2</td>
<td>0.533</td>
<td>2</td>
</tr>
<tr>
<td>&quot;log(M) ~ log(Linf) + log(tmax)&quot;</td>
<td>6</td>
<td>0.323</td>
<td>7</td>
</tr>
<tr>
<td>&quot;log(M) ~ log(K) + log(tmax)&quot;</td>
<td>4</td>
<td>0.313</td>
<td>5</td>
</tr>
<tr>
<td>&quot;log(M) ~ log(Linf)&quot;</td>
<td>7</td>
<td>0.666</td>
<td>15</td>
</tr>
<tr>
<td>&quot;log(M) ~ log(K)&quot;</td>
<td>2</td>
<td>0.687</td>
<td>2</td>
</tr>
<tr>
<td>&quot;log(M) ~ log(tmax)&quot;</td>
<td>4</td>
<td>0.326</td>
<td>4</td>
</tr>
<tr>
<td>&quot;log(M) ~ log(K) + log(Linf) + log(Temp)&quot;</td>
<td>2</td>
<td>0.565</td>
<td>4</td>
</tr>
</tbody>
</table>
Results

Best KNN models (minimum CVPE)

- $t_{\text{max}}$-based:
  \[ M \sim t_{\text{max}} \quad (k = 6) \]

- Growth-based:
  \[ \log(M) \sim \log(K) + \log(L_{\infty}) \quad (k = 2) \]

Distance: Manhattan

Kernel function: Rectangular
Results

CVPE as a function of k

Distance = Manhattan, kernel = rectangular

KNN-$t_{max}$

KNN-growth

$k = 6$

$k = 2$
## Results

<table>
<thead>
<tr>
<th>Models</th>
<th>Training (CVPE)</th>
<th>Test (RMSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{\text{max}}$-based</td>
<td>(n = 215)</td>
<td>(n = 44)</td>
</tr>
<tr>
<td>KNN - $t_{\text{max}}$</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>REG - $t_{\text{max}}$</td>
<td><strong>0.28</strong></td>
<td>0.31</td>
</tr>
<tr>
<td>growth-based</td>
<td>(n = 215)</td>
<td>(n = 35)</td>
</tr>
<tr>
<td>KNN - growth</td>
<td>0.53</td>
<td>0.81</td>
</tr>
<tr>
<td>REG - growth</td>
<td>0.58</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Results: $t_{\text{max}}$-based

Residual plots (training)

REG-$t_{\text{max}}$

KNN-$t_{\text{max}}$

RMSE = 0.281

RMSE = 0.27
Results: $t_{max}$-based

Residual plots ($t_{max} > 7, n = 167$)

REG-$t_{max}$

KNN-$t_{max}$
Results: $t_{\text{max}}$-based

Predictions
Results: growth-based

Residual plots (training)

REG-growth

KNN-growth

RMSE = 0.703

RMSE = 0.357
Results: growth-based

Residual plots \((t_{max} > 7, n = 167)\)

REG-growth

KNN-growth

RMSE = 0.349

RMSE = 0.185
Going back to the objectives

Can we do better with empirical estimation of $M$... with KNN?
Conclusions & Discussion

• KNN method improved prediction of $M$ using von Bertalanffy growth parameters (\& possibly $t_{\text{max}}$)
• Residuals of KNN methods are better-behaved than REG methods
• Temp not important predictor variable
• Computationally feasible
• KNN may perform poorly for stocks in ‘neighbor-poor’ regions
Acknowledgments

Funding

- NOAA Stock Assessment Improvement Award
- Malaysian Ministry of Higher Education Scholarship
- Alaska Sea Grant Travel fund
- University Malaya Conference Travel fund

Data Assistance

- Brooke Lowman, So-Jung Youn, Johnathan D. Maxey, Robert C. Harris
$M$ estimates

‘All models are wrong
BUT some are useful’

~ George E.P. Box ~

Thank you