A New Approach to Generating Spatial Age-Length Keys

Based on Using a Gaussian Field Approximation with Support for Physical Barriers

Jonathan Babyn

Dalhousie University,
Mathematics & Statistics, PhD Student,
supervised under Joanna Mills Flemming

November 6 2019

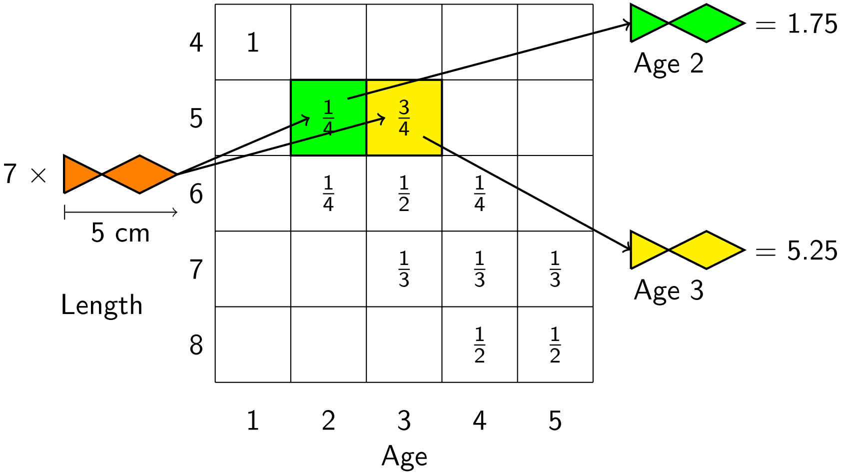
Outline

- Background
- Spatial Age-Length Key Methods
- Simulation Study
- Conclusions

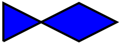
Motivation

- Age structured models common in stock assessments
- Direct aging is expensive
- Accuracy could be improved
- Working towards a spatial model

Age-Length Keys



Age-Length Keys

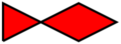
10 × 

7 cm

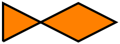
Length

Length \ Age	1	2	3	4	5
4	1				
5		$\frac{1}{4}$	$\frac{3}{4}$		
6		$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	
7	??	??	??	??	??
8				$\frac{1}{2}$	$\frac{1}{2}$

Age

 = ??

Age 3

 = ??

Age 4

Age-Length Keys

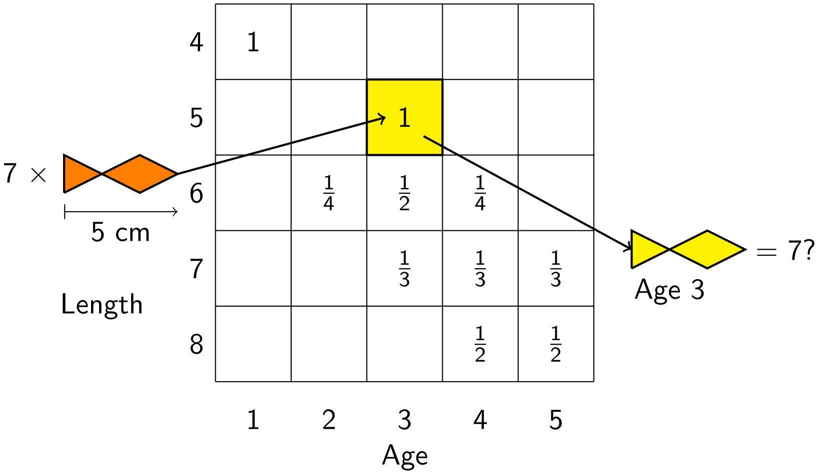


Figure: So can other sampling artifacts.

Model-based ALKs

- Can create ALKs via statistical techniques that can give a probability of being age j given length and other covariates
- Statistical techniques like multinomial regression, ordinal regression, machine learning, etc.
- Smooth over gaps and noise
- Predict where not observed
- Continuous covariates a possibility, including space!

Continuation-Ratio Logits

- Form of Ordinal Regression
- Easy to implement, easy to relax assumption of equal slopes across age classes

Definition:

$$\text{logit}(\pi_a[\boldsymbol{x}_i]) = P(Y = a | Y \geq a)$$

$$\pi_a[\boldsymbol{x}_i] = \frac{p_a[\boldsymbol{x}_i]}{p_a[\boldsymbol{x}_i] + \ldots + p_A[\boldsymbol{x}_i]}$$

$$P(Y = a) = \begin{cases} \pi_a[\boldsymbol{x}_i], \quad a = R \\ \pi_a[\boldsymbol{x}_i] \sum_R^{a-1}(1 - \pi_i[\boldsymbol{x}_i]) \quad R < a < A \\ 1 - \sum_R^{A-1}(1 - \pi_i[\boldsymbol{x}_i]) \quad a = A \end{cases}$$

Age a, R- First age in model, A- Last age/Plus Group, $\boldsymbol{x_i}$- Set of covariates for observation i, p_a proportion at age a

CRL ALK Model

- Simplest CRL model: $\text{logit}(\pi_a[\boldsymbol{x}_i]) = \alpha_a + \beta_a l_i$
- α_a: Intercept for age a
- l_i: Length of observation i
- β_i Slope term for length of age a

ALK Creation:

- Predict model at desired length bins
- Get unconditional probabilities which form the ALK

$$P(Y = a) = \begin{cases} \pi_a[\boldsymbol{x}_i], & a = R \\ \pi_a[\boldsymbol{x}_i] \sum_R^{a-1}(1 - \pi_i[\boldsymbol{x}_i]) & R < a < A \\ 1 - \sum_R^{A-1}(1 - \pi_i[\boldsymbol{x}_i]) & a = A \end{cases}$$

GAM based Spatial ALKs

- Berg & Kristensen (2012) presented a number of ALK models incorporating spatial information in a Generalized Additive Model
- $\text{logit}(\pi_a[\boldsymbol{x}_i]) = \alpha_a + \beta_a l_i + f(\boldsymbol{s})$
- $f()$ being of latitude and longitude using thin-plate regression splines
- They found improved internal and external consistencies for survey indices using spatial versions of the model vs. non-spatial versions.

Gaussian Fields

- GFs are collections of Gaussian Random Variables indexed across space
- Described by mean function $\mu(s)$ and covariance function $\mathrm{Cov}(s,t)$
- Matérn: $c(s,t) = \sigma_u^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{8\nu}\frac{||s-t||}{r}\right) K_\nu \left(\sqrt{8\nu}\frac{||s-t||}{r}\right)$
- Too slow to use directly for larger cases

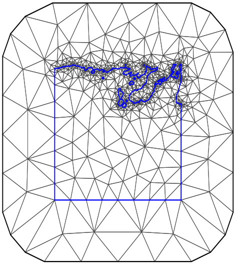
Gaussian Markov Random Fields:

GFs with the Markov Property :

$P(X_n = x_n | X_{n-1}, ..., X_0 = x_0) = P(X_n = x_n | x_{n-1})$

ν-smoothness parameter, r-range parameter, σ_u-standard deviation of GF, K_ν -Bessel function of the second kind, Γ function

GF Approximation



- Lindgren et. al found an explicit link between GFs and GMRFs when using a Matérn covariance function
- A valid semi-positive definite covariance matrix is the solution to a set of Stochastic Partial Differential Equations
- This ensures a sparse structure in the covariance matrix

Barrier Approach

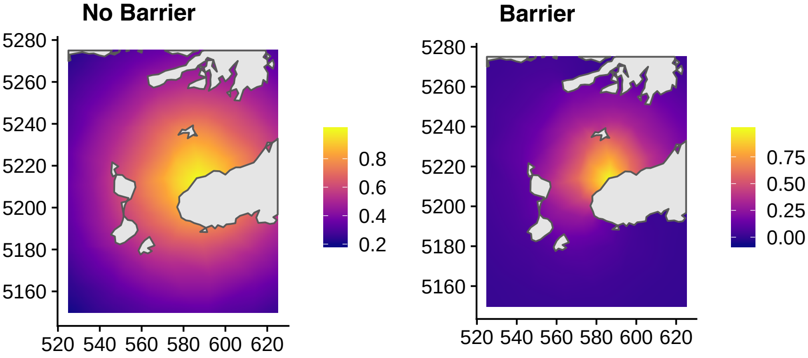
- Bakka et. al extended the SPDE GF approximation to include the ability to handle physical barriers
- They did this by have the range parameter inside the area of barrier be a fraction of the value outside the barrier

SPDE is solution to these equations:

$$u(\boldsymbol{s}) - \nabla \cdot \frac{r^2}{8} \nabla u(\boldsymbol{s}) = r \sqrt{\frac{\pi}{2}} \sigma_u \mathcal{W}(\boldsymbol{s}) \quad \text{for } s \in \Omega_n \tag{1}$$

$$u(\boldsymbol{s}) - \nabla \cdot \frac{r_b^2}{8} \nabla u(\boldsymbol{s}) = r_b \sqrt{\frac{\pi}{2}} \sigma_u \mathcal{W}(\boldsymbol{s}) \quad \text{for } s \in \Omega_b \tag{2}$$

Barrier Approach – Correlation at a Point



Spatial ALK with GF

The Spatial ALK model using a GF is

$$\text{logit}(\pi_a[\boldsymbol{x}_i]) = \alpha_a + \beta_a l_i + \xi_{a,\boldsymbol{s}} \tag{3}$$

$$\xi_{a,\boldsymbol{s}} = \begin{cases} \text{MVN}\left(\mathbf{0}, \frac{\sigma_u^2}{(1-\varphi_a^2)} c(\boldsymbol{s})\right) & a = 1 \\ \text{MVN}\left(\varphi_a \xi_{a-1,s}, \sigma_u^2 c(\boldsymbol{s})\right) & a > 1. \end{cases} \tag{4}$$

Model was implemented using Template Model Builder and Maximum Likelihood (ML) estimation and optimized with `nlminb`

Penalization

- With large numbers of categories in ordinal regression models, using ML estimation can cause an optimizer to easily fall into a local minimum
- Penalization can help avoid that and improve prediction accuracy
- $\log L - \frac{1}{2}\lambda\boldsymbol{\beta}'\boldsymbol{P}\boldsymbol{\beta}$

SimSurvey

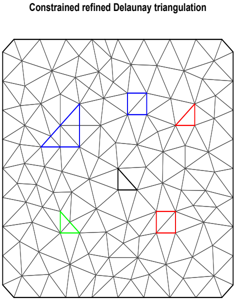
- SimSurvey is an R package created by Dr. Paul Regular at Fisheries and Ocean Canada
- Simulates Bottom Trawl Research Vessel-like survey data from a stratified random sample design and an estimate of abundance at age calculated using the stratified mean method
- Simulates a population with known abundance at age and distribute it spatially among an area
- Offers control over strata design, tow distance, and other survey settings
- Publicly available on GitHub

Simulation Models

Four different ALKs methods were compared:

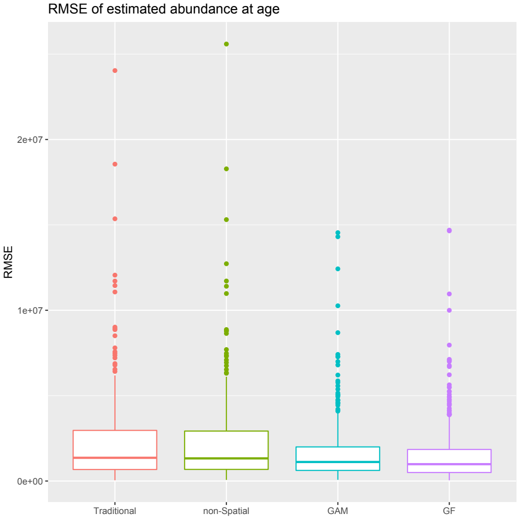
- Traditional ALK
- Non-spatial CRL model: $\text{logit}(\pi_a[\boldsymbol{x}_i]) = \alpha_a + \beta_a l_i$
- GAM model: $\text{logit}(\pi_a[\boldsymbol{x}_i]) = \alpha_a + \beta_a l_i + f(\boldsymbol{s})$
- GF model: $\text{logit}(\pi_a[\boldsymbol{x}_i]) = \alpha_a + \beta_a l_i + \xi_{a,\boldsymbol{s}}$

Simulation Study

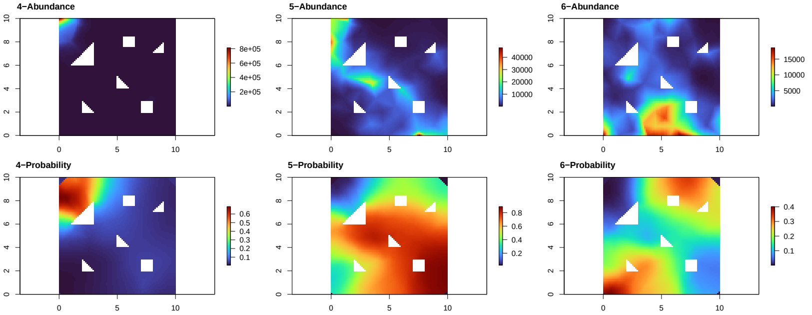


- New population every simulation run, 450 simulated surveys
- Growth from Von Bertalanffy growth curve
- 48 strata based on depth, 96 total sets, an average of 403 fish aged per survey
- Length-Stratified sampling

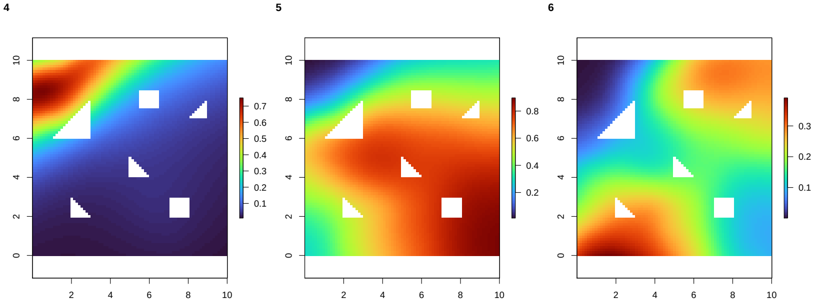
Root Mean Squared Error



Sim - Prob. of being Age 4,5 or 6 with length of 40cm vs. true abundance-GF



Sim - Probability of being Age 4,5 or 6 with length of 40cm-GAM



Conclusions

- If there is a difference in where ages are distributed across space incorporating spatial information can improve estimates of age structure, potentially improving assessment

Future Work: Application to real data set

Acknowledgements

Thank you to Joanna Mills Flemming, Divya Varkey and Paul Regular

References I

- Alan Agresti. *Categorical data analysis*. Vol. 482. John Wiley & Sons, 2003.
- Haakon Bakka. "How to solve the stochastic partial differential equation that gives a Mat ern random field using the finite element method". Mar. 2018.
- Haakon Bakka et al. "Non-stationary Gaussian models with physical barriers". *Spatial Statistics* (2019).
- Haakon Bakka et al. "Spatial modelling with R-INLA: A review". *Wiley Interdisciplinary Reviews: Computational Statistics* 10.6 (2018).
- Casper W Berg and Kasper Kristensen. "Spatial age-length key modelling using continuation ratio logits". *Fisheries Research* 129 (2012), pp. 119–126.

References II

- Frank E Harrell. "Regression modeling strategies". *as implemented in R package 'rms' version* 3.3 (2014).
- Kasper Kristensen et al. "TMB: Automatic differentiation and laplace approximation". *Journal of Statistical Software* 70.5 (2016), pp. 1–21.
- Trine Kvist, Henrik Gislason, and Poul Thyregod. "Using continuation-ratio logits to analyze the variation of the age composition of fish catches". *Journal of applied statistics* 27.3 (2000), pp. 303–319.
- Finn Lindgren and Håvard Rue. "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". *Journal of the Royal Statistical Society. Series B* (2011).

References III

- Anna Rindorf and Peter Lewy. "Analyses of length and age distributions using continuation-ratio logits". *Canadian Journal of Fisheries and Aquatic Sciences* 58.6 (2001), pp. 1141–1152.
- Traiani Stari et al. "Smooth age length keys: observations and implications for data collection on North Sea haddock". *Fisheries Research* 105.1 (2010), pp. 2–12.