A multivariate tree-based method for exploring stock structure in multiple data sets

Cleridy E. Lennert-Cody, Mark N. Maunder, Carolina Minte-Vera, Haikun Xu, Juan Valero, Alex Aires-da-Silva, Jon Lopez

Stock Assessment Program
Inter-American Tropical Tuna Commission

CAPAM

Spatial Stock Assessment Models
La Jolla, October 1- 5 2018

CIAT IATTC

Outline of presentation

- The challenge: defining spatial units for stock assessments.
- The "simultaneous tree" method:
  - Description of the method;
  - Illustration of the method: bigeye tuna from Japanese longline fishery in the eastern Pacific Ocean (EPO);
  - Summary and comments on improving the methodology.

Defining spatial units

- Spatial management requires ‘spatial units’ be defined, typically a limited number of large areas.
- To define spatial units, population spatial structure can be studied with many types of data.
- In fisheries, direct indicators of population structure are not always available, however, monitoring programs can yield large amounts of catch and size composition data.
- With some assumptions, these data may be considered indirect indicators of population structure.
- Inference with indirect measures may be improved by studying the spatial structure in the two data types simultaneously.

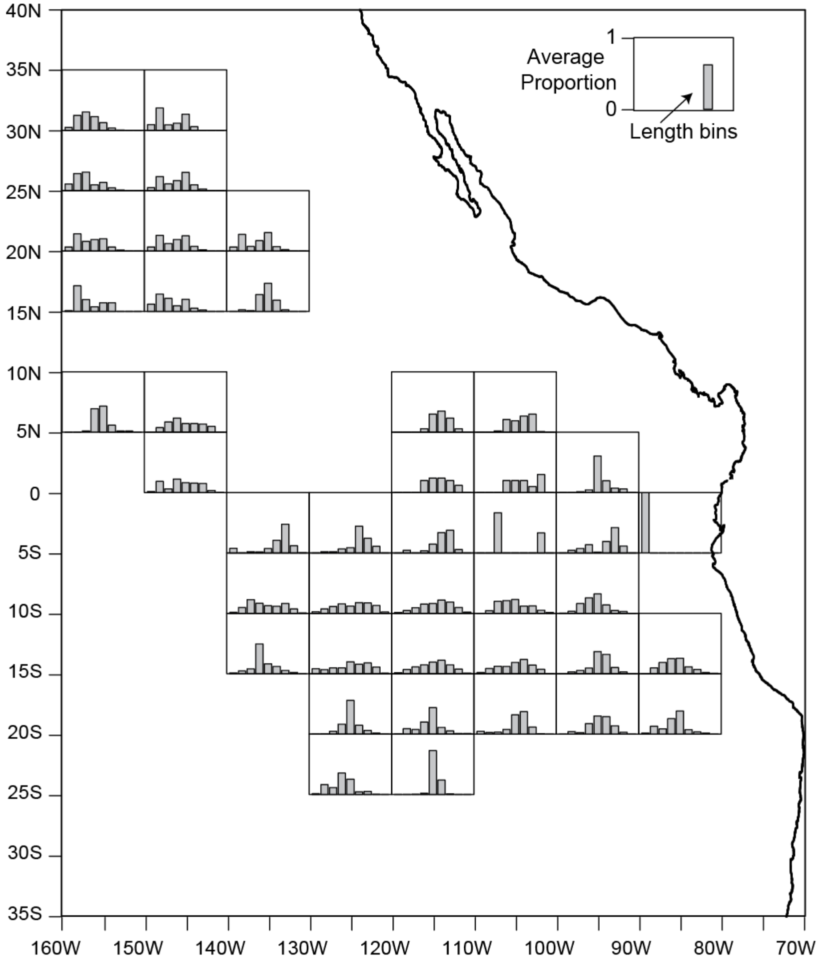
Simultaneous tree method

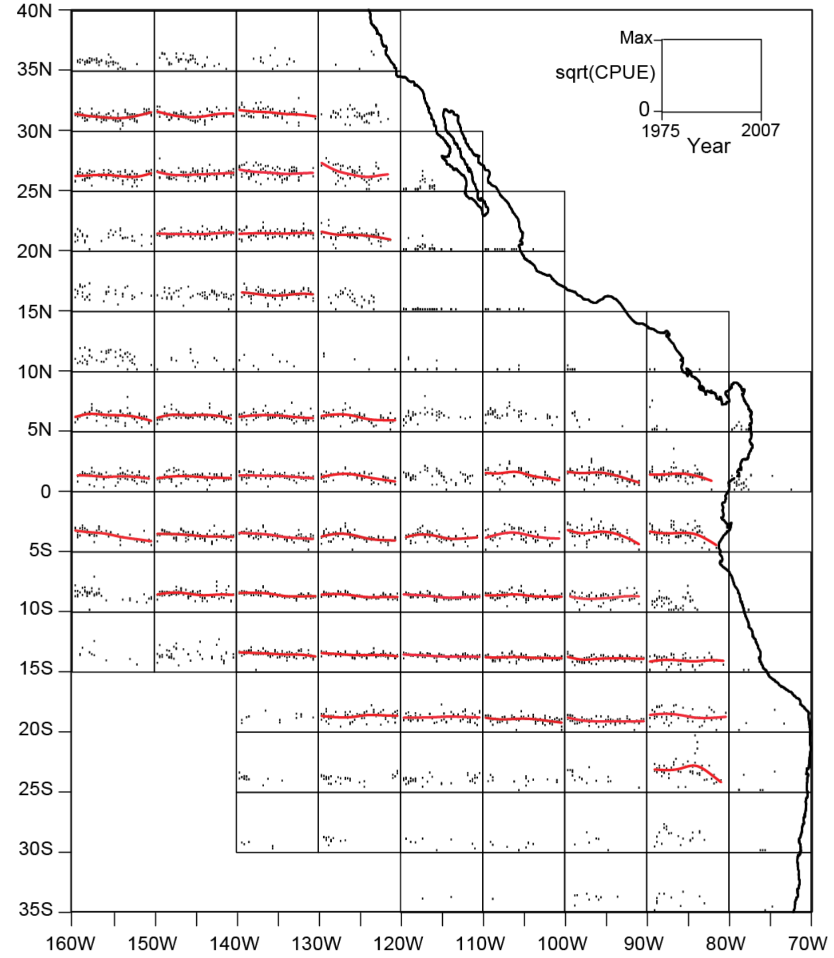
- Task: develop a method for exploratory analysis of large-scale spatial pattern *simultaneously* in different data types.

- Overview of the method:
  1) Construct multivariate response variable and select impurity measure for each data type;
  2) Grow a small tree, with an combined split criterion that is based on the impurity measures of (1) (do not prune);
  3) From tree structure, identify candidate spatial units.

- Although many types of data could be considered, we focus on length-frequency distributions and relative abundance trends.

Length-frequency distributions and relative abundance trends: an example





Response variable and impurity: frequency distributions

- Starting point
  - Raw length-frequency distributions

- Multivariate response
  - Proportion of individuals in each binned length interval, $\{p_l(j)$, j = 1, …, # intervals for sample or data unit $l\}$.

- Impurity measure for a collection of units $\{l\}$
  - Based on the Kullback-Leibler divergence ('KLD')
  - $I_{KLD} = \sum_l \sum_j p_l(j) log\left(\frac{p_l(j)}{\bar{p}_.(j)}\right)$

Response variable and impurity: trends

- Starting point
  - Nominal CPUE times series.
- Multivariate response
  - First generate vector of annual estimates of relative abundance, $\hat{C}_l$, for a complete time series of m years in grid cell or unit l.
  - Response is then vector of first-differenced annual relative abundance estimates:

$$\Delta\hat{C}_l = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ & & \dots & & \\ 0 & \dots & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \hat{C}_{l1} \\ \vdots \\ \hat{C}_{lm} \end{bmatrix}$$

  - Interested in trends, not absolute magnitude.
- Impurity measure for a collection of units $\{l\}$
  - Sum of squares-based measure:

$$I_{SS} = \sum_{l} \sum_{y=1}^{m-1} \left(\left(\Delta\hat{C}_l\right)_y - \left(\Delta\tilde{C}\right)_y \right)^2$$

    where $\tilde{C}$ is relative abundance estimated from the pooled data.

- Can modify I_{SS} by weighting by inverse of variance of $\Delta\hat{C}_l$.

Combined split criterion

- Each binary partition of the data sets is evaluated with the following combined criterion:

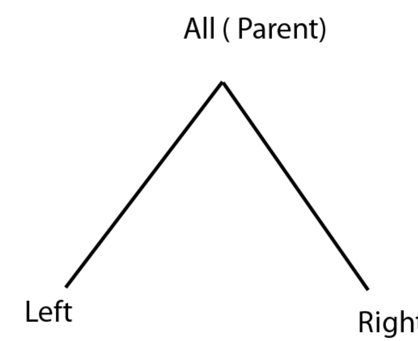
$$\gamma\left[\frac{Imp\_KLD}{max_{candiate\ splits}(Imp\_KLD)}\right] + (1-\gamma)\left[\frac{Imp\_SS}{max_{candidate\ splits}(Imp\_SS)}\right]$$

where

γ (0<γ<1) is a user-specified weight

$$Imp\_KLD = n_{left}\sum_j \bar{p}_{left}(j)\ log\left(\frac{\bar{p}_{left}(j)}{\bar{p}_{.}(j)}\right) + n_{right}\sum_j \bar{p}_{right}(j)\ log\left(\frac{\bar{p}_{right}(j)}{\bar{p}_{.}(j)}\right)$$

$$Imp\_SS = I_{SS;\,all} - \left(I_{SS;\,left} + I_{SS;\,right}\right)$$



- *In principle*, the criterion takes on values between 0 and 1.
- The best split choice maximizes this criterion.

Growing the tree

- A small tree of is built by binary recursive partitioning, using the combined split criterion.

- The tree size can be based on, for example, the number of strata (*e.g.*, areas) that can be handled in the population assessment model.

Example: bigeye tuna in the EPO Japanese longline fishery

- Available data are aggregated (catch, effort at 5° x 5° x month; lengths at 5° x 10° x month).

- Spatial and temporal resolution used for the analysis:
  5° latitude x 10° longitude, quarter.

- Why?
  - Minimum common spatial resolution is 5° latitude x 10° longitude.
  - Assessment model has quarterly time step.
  - Interested in knowing if large-scale spatial pattern varies quarterly.

- In the interest of time, skipping description of data processing.

- Predictors (all numeric): 5° latitude, 10° longitude, quarter, cyclic quarter.

Input data: length-frequency distributions

- Raw data:
  - fish counts by 2 cm interval, years 1986-1991.

- Multivariate response:
  - proportion of fish per sample in each of 9 binned length intervals (*i*.*e*., binned length-frequency distribution).

Input data: relative abundance trends

- Raw data: nominal *cpue* = # fish/# hooks, for 1975-1991.

- Trends were estimated by fitting a simple cubic spline model to data in each grid cell l:

$$\text{sqrt}(cpue_{l,y,n_y}) = f\left(year_{l,y,n_y}\right) + \varepsilon_{l,y,n_y}$$

f a smooth function;
ε error;
y indexes year, n_y data points of year y;
sqrt is square root, used to stabilize variance;
basis dimension, knots, smoothing parameter fixed for all l.

- Multivariate response: first-differenced times series of predicted annual sqrt(*cpue*).

Results: EPO

	Variable value	L-F Improvement			CPUE UNWTD Improvement	Simultaneous tree scaled improvement				CPUE WTD Improvement	Simultaneous tree scaled improvement	
Latitude	27.5S	1.13			< 0.001	0.053				-48.49		
	22.5S	8.29			0.001	0.284				-47.12		
	17.5S	13.76	2nd		0.005	0.603	3rd			76.08	0.507	
	12.5S	8.91			0.006	0.494			3rd	328.54	0.603	3rd
	7.5S	4.39			-0.004					309.31	0.443	
	2.5S	2.39			-0.002					259.14	0.330	
	2.5N	2.63			0.004	0.218				110.51	0.191	
	7.5N	3.52		Best	0.014	0.611	2nd		2nd	483.60	0.587	
	12.5N	4.57										
	22.5N			2nd	0.008					323.36		
	27.5N				0.006					272.87		
Longitude	145W	0.70			0.001	0.045				-105.15		
	135W	3.35			0.004	0.247				73.26	0.177	
	125W	8.20		3rd	0.007	0.486				304.36	0.557	
	115W	13.00	3rd	4th	0.006	0.629	Best	**110W**	Best	507.28	0.909	Best
	105W	15.91	Best		0.002	0.580				208.70	0.705	2nd
	95W	12.44			0.005	0.571				134.55	0.523	
	85W	1.01										
Quarter	1	2.15			0.001	0.114				7.44	0.075	
	2	5.09			0.005	0.334				-34.47		
	3	3.37			0.004	0.245				42.82	0.148	
Cyclic quarter	1,4;2,3	5.11			0.002	0.236				17.18	0.178	
	124;3	8.79			0.001	0.317				-19.45		
	134;2	2.29			0.004	0.219				89.94	0.160	

Results: EPO west of 110°W

	Variable value	L-F Improvement			CPUE UNWTD Improvement	Simultaneous tree scaled improvement		
Latitude	12.5S	2.39			5.17E-04	0.38		
	7.5S	3.09	3rd		-1.63E-03			
	2.5S	2.04			9.30E-05	0.31		
	2.5N	2.18			3.71E-03	0.46		
	7.5N	3.32	2nd	Best	1.39E-02	1.00	Best	**10N**
	12.5N	4.35	Best					
	22.5N			2nd	7.47E-03			
	27.5N			3rd	5.30E-03			
Longitude	-145	0.68			3.76E-05	0.10		
	-135	1.17			2.41E-03	0.26		
	-125	1.01			3.63E-03	0.28		
Quarter	1	0.95			1.85E-03	0.21		
	2	2.40			9.38E-04	0.39		
	3	1.75			1.52E-03	0.32		
Cyclic quarter	1,4;2,3	2.17			2.85E-03	0.43		
	124;3	1.64			1.56E-03	0.30		
	134;2	2.53			1.55E-03	0.44		

Summary and comments on improving the methodology

- Useful for exploring similarities in large-scale pattern among several multivariate data types.
- Amenable to other data types and loss functions.
- More complex trend models could be used.

- Variance weighting: is it a good thing?
- Sensitivity to data subsets: implement “bagging”?
- Allow for non-rectangular spatial partitions.
- Negative SS improvements: is model for the pooled data the best choice?

$$I_{SS} = \sum_{l} \sum_{y=1}^{m-1} \left(\left(\Delta \hat{C}_l\right)_y - \left(\Delta \tilde{C}\right)_y \right)^2$$

$$Imp\_SS = I_{SS;\,all} - \left(I_{SS;\,left} + I_{SS;\,right}\right)$$

Thank you!

Special thanks to the National Research Institute of Far Seas Fisheries for the longline fishery data.

All analyses were programmed in R. The spline trend models were fitted with the *mgcv* package.

References

Lennert-Cody, C.E., Minami, M., Tomlinson, P.K., Maunder, M.N. 2010. Exploratory analysis of spatial-temporal patterns in length-frequency data: An example of distributional regression trees. Fisheries Research 102: 323-326.

Lennert-Cody, C.E., Maunder, M.N., Aires-da-Silva, A., Minami, M. 2013. Defining population spatial units: simultaneous analysis of frequency distributions and time series. Fisheries Research 139: 85-92.