Fisheries Research 102 (2010) 323–326

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Short communication

# Exploratory analysis of spatial–temporal patterns in length–frequency data: An example of distributional regression trees

Cleridy E. Lennert-Cody[a,*], Mihoko Minami[b], Patrick K. Tomlinson[a], Mark N. Maunder[a]

[a] Inter-American Tropical Tuna Commission, 8604 La Jolla Shores Drive, La Jolla, CA 92037, USA
[b] Department of Mathematics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan

ARTICLE INFO

*Article history:*
Received 15 July 2009
Received in revised form
28 November 2009
Accepted 30 November 2009

*Keywords:*
Multivariate regression tree
Kullback–Leibler divergence
Length–frequency
Yellowfin tuna

ABSTRACT

Understanding the spatial–temporal distributions of fish populations is important for their assessment and management. Given the complex structure often present in fisheries length–frequency samples, there is a need for flexible statistical techniques to explore patterns with these types of data. We present a multivariate regression tree method for binned frequencies that uses the Kullback–Leibler divergence to measure node heterogeneity. To illustrate this approach, we apply the method to length–frequency data for yellowfin tuna caught in the purse-seine fishery of the eastern Pacific Ocean.

## 1. Introduction

Development of spatially structured stock assessment models is complicated by the nature of catch length–frequency data, which can be skewed, or even multimodal. For species such as tunas, which are assessed by catch-at-length models, there is a need for flexible techniques to explore spatial–temporal patterns in the size structure, extending methods based on summary length statistics (Phillips et al., 2005). Multivariate tree-based techniques (*e.g.*, Segal, 1992; Yu and Lambert, 1999; De'ath, 2002; Molinaro et al., 2004; Cariou, 2006; Nerini and Ghattas, 2007) have been developed to extend classical regression trees (CART; Breiman et al., 1984) to complex data. These techniques include distance-based approaches for basis function approximations of sample curves. However, for exploratory analysis of length–frequency data, an impurity-based approach that uses binned frequencies seems preferable, especially as length intervals are often directly interpretable in terms of cohorts and recruitment. In this paper we present a multivariate regression tree method for exploratory analysis of binned length–frequency data, where node heterogeneity is measured by the Kullback–Leibler divergence (KLD; *e.g.*, Wang et al., 2005), and we demonstrate use of this method with length–frequency data for yellowfin tuna (*Thunnus albacares*) collected from the purse-seine fishery of the eastern Pacific Ocean.

## 2. Length–frequency samples

Data used in this example are from the Inter-American Tropical Tuna Commission port-sampling program (Suter, 2008) of purse-seine catches of vessels with fish-carrying capacity greater than 363 t. We use length–frequency data for yellowfin tuna caught in purse-seine sets on tunas associated with dolphins from 2003 to 2007. The data represent forklength (to the nearest mm) from 797 samples (∼50 fish measured per sample). The minimum spatial–temporal resolution of these data is one month by a 5° area. To increase the number of samples per 5° area, and to be consistent with the stock assessment, fish were 'grown' (or 'shrunk') from the month to the quarter according to the Gompertz growth model of Wild (1986). For this example, the length data were binned into 11 intervals: ≤58 cm, 59–69 cm, . . ., 136–146 cm, 147–159 cm, and ≥160 cm. The proportion of fish per length interval was computed by sample, which creates a multivariate response with 11 values per sample. For this example, the length intervals were chosen to be small enough to capture the structure in the data, but large enough to avoid a dominance of zero values in any given sample. As with any frequency-based method, sensitivity of results to the choice of bin size should be explored.

## 3. A regression tree approach for length–frequency data

### 3.1. Description of the method

CART uses recursive partitioning to try to determine a series of binary decision rules ('splits') that divide the data into smaller

[*] Corresponding author. Tel.: +1 858 546 7190; fax: +1 858 546 7133.
*E-mail address:* clennert@iattc.org (C.E. Lennert-Cody).

0165-7836/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.fishres.2009.11.014

more homogeneous subgroups. The split variables and their values are selected to provide the greatest decrease in the heterogeneity ('impurity') of the data, as measured by the sum of squares loss function. To adapt the CART approach to an impurity-based analysis of binned frequencies, we follow the general concept of functional data analysis (Ramsay and Silverman, 2002), which is to view each length–frequency distribution as the data. The KLD is an appropriate measure for quantifying the difference between frequency distributions. For two probability distributions $P$ and $Q$, the KLD of $P$ from $Q$ has the following form:

$$D(P|Q) \equiv \sum_j p(j) \log\left(\frac{p(j)}{q(j)}\right)$$

where $p(j)$ and $q(j)$ are the frequencies associated with the $j$th interval. For a collection of distributions $P_1$ ,. . ., $P_n$, the equally weighted combined distribution is given by

$$\bar{P} = \frac{1}{n}\sum\nolimits_{i=1}^{n} P_i$$

Using the above definitions, the KLD-based definition of impurity of a group of $n$ distributions can be expressed as

$$I_{\text{KLD}} = \sum_{i=1}^{n} D(P_i|\bar{P}) \tag{1}$$

The improvement ($Imp$) toward reducing the within-node impurity achieved by one binary split of a CART regression tree is defined as the impurity of the parent node less that of each of its two children (Breiman et al., 1984). From Eq. (1), it can be shown that the improvement, $Imp(G_L, G_R)$, achieved by one binary split of the parent node collection of distributions $G$ into smaller collections $G_L$ and $G_R$ is given by:

$$\begin{aligned} Imp(G_L, G_R) &= n_{all}H(\bar{P}_{all}) - n_L H(\bar{P}_L) - n_R H(\bar{P}_R) \\ &= n_L D\left(\bar{P}_L|\bar{P}_{all}\right) + n_R D\left(\bar{P}_R|\bar{P}_{all}\right) \end{aligned} \tag{2}$$

where $n_{all}$ is the number of distributions in the parent collection, and $n_L$ and $n_R$ are the number of distributions in the left and right child node collections, respectively, and $H(P) = -\sum_j p(j) \log p(j)$ is the entropy of distribution P.

In our example, for $m$ length intervals and $n$ length–frequency samples, Eq. (1) becomes

$$I_{\text{KLD}} = \sum_{i=1}^{n}\sum_{j=1}^{m} p_i(j) \log\left(\frac{p_i(j)}{\bar{p}(j)}\right) \tag{3}$$

where $p_i(j)$ is the proportion of fish in the $j$th length interval of the $i$th sample, and $\bar{p}(j)$ is the average proportion of fish in the $j$th length interval (average computed over samples in the collection). Eq. (2) becomes

$$Imp(G_L, G_R) = n_L \sum_{j=1}^{m} \bar{p}_L(j) \log\left(\frac{\bar{p}_L(j)}{\bar{p}(j)}\right) + n_R \sum_{j=1}^{m} \bar{p}_R(j) \log\left(\frac{\bar{p}_R(j)}{\bar{p}(j)}\right) \tag{4}$$

Eqs. (3)–(4) ($m$ = 11) were implemented by modifying the *mrt.c* routine of the library *mvpart* (mvpart, 2007) of the statistical freeware *R* (R Development Core Team, 2009). The initial tree grown was pruned using cross-validation to the '1-se' tree (Breiman et al., 1984), which is a commonly used rule for establishing the size (number of terminal nodes) of the final tree. Predicted frequencies at the terminal nodes are given by the $\bar{p}(j)$.

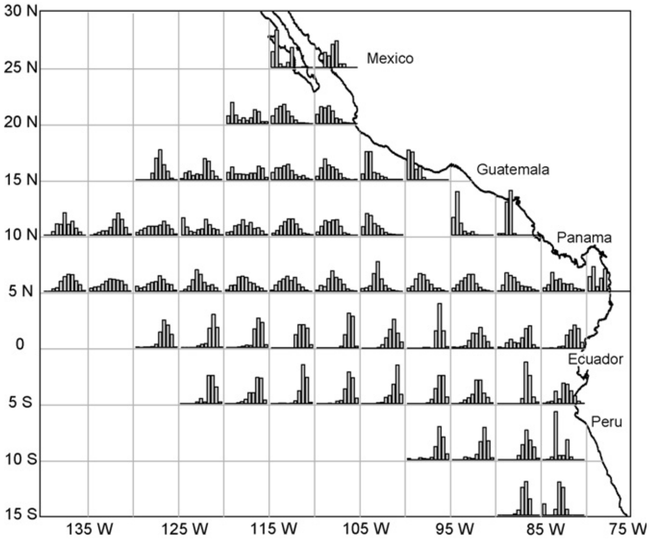

**Fig. 1.** Mean proportion (over samples) of yellowfin tuna by length interval within each 5° area (data pooled over quarters and years).

### 3.2. Application to length–frequency data

Spatial structure in the binned yellowfin tuna length-frequencies is clearly evident (Fig. 1). To explore the spatial–temporal pattern in these data, the method outlined above was applied to the binned frequencies of each year, and to the data pooled over the five-year period, using three predictors: the 5° degree latitude and longitude, and the quarter. (These predictors were treated as numeric, which imparts some 'smoothness' to the splits.) Given the geometry of the large-scale oceanographic circulation patterns of the eastern Pacific Ocean (Kessler, 2006), these spatial predictors, and the types of partitions produced by CART-like binary splits, are a sensible starting point for analyzing the yellowfin length–frequency data. For simplicity, in what follows, we focus on the spatial aspects of the results.

Comparison of the 1-se tree of the pooled data to those individual years highlights consistent spatial structure over the five years, as well as regions of inter-annual variability (Fig. 2). The overall spatial structure identified relates to a presence of large fish in the catch south of 5°N, moderate-sized fish offshore of 115°W and north of 5°N, and small to moderate-sized fish inshore of 115°W north of 5°N. Noteworthy is the relatively homogeneity south of 5°N, compared to the area north of 5°N and east of 115°W. The simplest annual structure was that of years 2003 and 2006, with a north-south split at 5°N and an east-west split at 115°W north of 5°N (Fig. 2). In the other three years, the area north of 5°N and east of 115°W was further divided by both latitude and longitude, although the boundaries were not always coincident across years (not shown). The region south of 5°N showed some differences between years (not shown), but appears relatively homogeneous compared to the area to the north.

The manner in which these types of results might play a role in defining stratification for stock assessment depends on the species. Assuming sampling coverage is adequate, comparison of pooled and individual-year tree results can identify spatial structure that is strongly indicated in every year, and that which is only present in select years, perhaps as a result of strong recruitment or changes in catchability, or if sampling coverage is not adequate, sampling variability. For example, the results above indicate a clear north-south division at 5°N, and an inshore–offshore division north of 5°N at 115°W, with possible inter-annual variability in recruitment within this northern inshore region. In a stock assessment, this northern

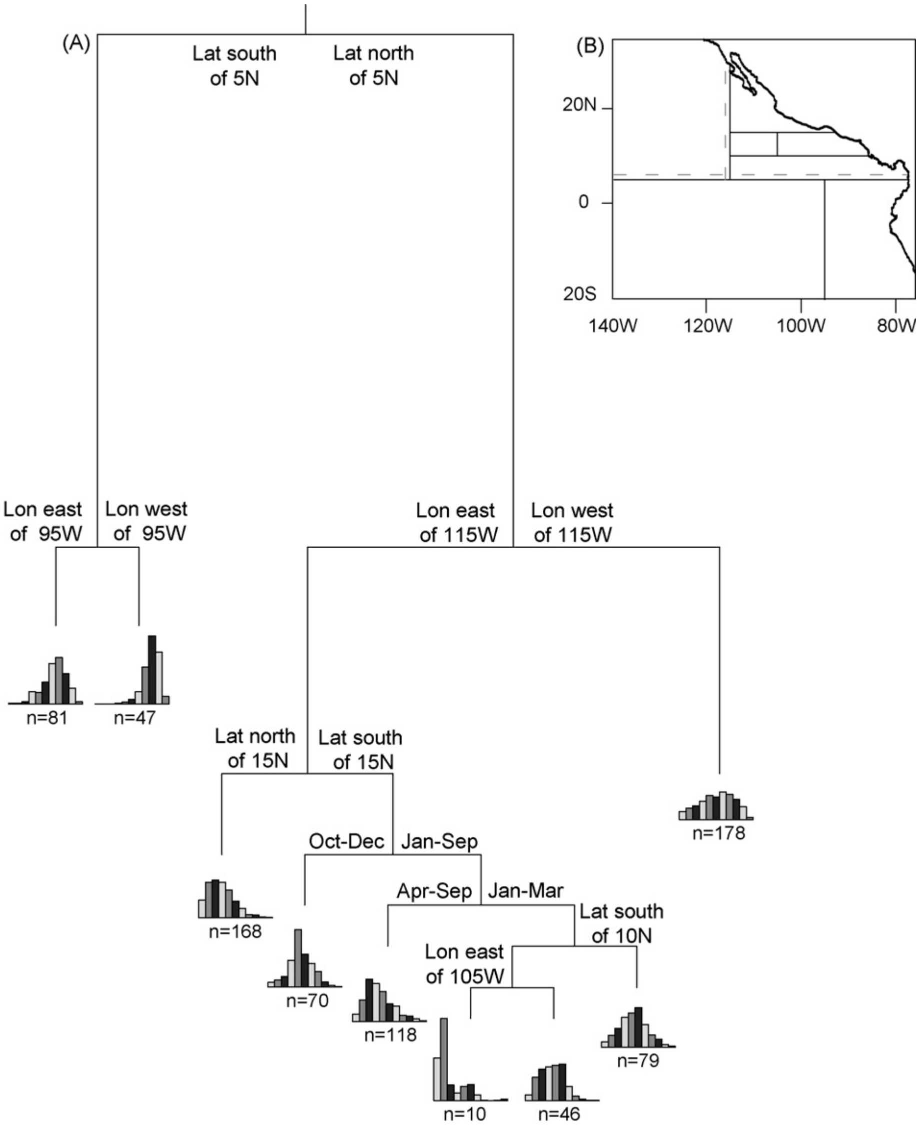

**Fig. 2.** (A) The 1-se tree for 2003–2007. The $\bar{p}(j)$ are shown at the terminal nodes. 'Lat': latitude; 'Lon': longitude; 'Jan': January; 'Mar': March; 'Apr': April'; 'Sep': September; 'Oct': October; 'Dec': December. (B) Maps of spatial splits of the 1-se trees for 2003–2007 (solid black lines), and 2003 and 2006 (dashed gray lines).

inshore area might be modeled with variable selectivity, while the other areas each with constant selectivity. For populations with low recruitment variability, an analysis of just the pooled data might be appropriate.

## 4. Concluding remarks

Although the 1-se rule for determining tree size generally yields sensible results, other criteria could be developed. For example, the final tree size might be that which is smaller than the size of the minimum-error tree but provides the smallest estimated error of the total catch-at-length. Alternatively, tree size might be set by the maximum feasible number of management areas, or by statistically meaningful differences in stock assessment results. Tree instability could be investigated by comparing trees built on bootstrap samples of the original data (Breiman, 1996; Nerini and Ghattas, 2007). In addition, for long time-series, it could be of interest to group years based on their environmental characteristics (*e.g.*, El Niño, La Niña years) to study environmental effects. The influence of strong cohorts could be further investigated with pooled data by modifying Eq. (4) to represent sums of annual impurities (*i.e.*, combined distributions computed within each year) but require the selected splits to be the same across all years. Finally, employing more complex density estimators (*e.g.*, kernel estimators) instead of binned frequencies may provide a useful refinement of the method.

## Acknowledgements

We are grateful to G. De'ath for the *mvpart* library in *R*. We thank William Bayliff, Simon Hoyle and two anonymous referees for comments that improved this manuscript, and Nickolas Vogel and Joydelee Marrow for data base assistance.

## References

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Chapman and Hall/CRC, p. 358.

Breiman, L., 1996. Bagging predictors. Machine Learning 26, 123–140.

Cariou, V., 2006. Extension of multivariate regression trees to interval data: application to electric load profiling. Computational Statistics 21, 325–341.

De'ath, G., 2002. Multivariate regression trees: a new technique for modeling species–environment relationships. Ecology 83, 1105–1117.

Kessler, W.S., 2006. The circulation of the eastern tropical Pacific: a review. Progress in Oceanography 69, 181–217.

Molinaro, A.M., Dudoit, S., van def Laan, M.J., 2004. Tree-based multivariate regression and density estimation with right-censored data. Journal of Multivariate Analysis 90, 154–177.

mvpart, 2007. rpart by Terry M. Therneau, Beth Atkinson. Rport of rpart by Brian Ripley <ripley@stats.ox.ac.uk>. Some routines from vegan–Jari Oksanen <jari.oksanen@oulu.fi>. Extensions and adaptations of rpart to mvpart by Glenn De'ath (2007). mvpart: multivariate partitioning. R package version 1.2-6.

Nerini, D., Ghattas, B., 2007. Classifying densities using functional regression trees: applications in oceanology. Computational Statistics and Data Analysis 51, 4984–4993.

Phillips, N.L., Dunn, A., Hancet, S.M., 2005. Stratification of catch-at-length data using tree based regression: an example using Antarctic toothfish (*Dissostichus mawsoni*) in the Ross Sea. WG-FSA-SAM-05/8, CCAMLR.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org.

Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis: Methods and Case Studies. Springer, p. 190.

Segal, M.R., 1992. Tree-structured methods for longitudinal data. Journal of the American Statistical Association 87, 407–418.

Suter, J.M., 2008. An evaluation of the area stratification used for sampling tunas in the eastern Pacific Ocean and implications for estimating total annual catches. Thesis for Master of Science in Statistics, San Diego State University, San Diego, CA, U.S.A.

Wang, Q., Kulkarnia, S.R., Verdú, S., 2005. Divergence estimation of continuous distributions based on data-dependent partitions. IEEE Transactions on Information Theory 51, 3064–3074.

Wild, A., 1986. Growth of yellowfin tuna, *Thunnus albacares*, in the eastern Pacific Ocean based on otolith increments. Inter-American Tropical Tuna Commission Bulletin 18, 421–482.

Yu, Y., Lambert, D., 1999. Fitting trees to functional data, with an application to time-of-day patterns. Journal of Computational and Graphical Statistics 8, 749–762.