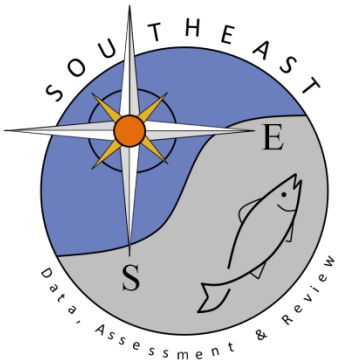
Example Implementation of a Hierarchical Cluster Analysis and Cross-correlations of Selected CPUE Indices for the SEDAR 54 Assessment

Dean Courtney

SEDAR54-WP-06

20 September 2017



Please cite this document as:

Courtney, D. 2017. Example Implementation of a Hierarchical Cluster Analysis and Cross-correlations of Selected CPUE Indices for the SEDAR 54 Assessment. SEDAR54-WP-06. SEDAR, North Charleston, SC. 11 pp.

SEDAR 54 ASSESSMENT DOCUMENT

Example Implementation of a Hierarchical Cluster Analysis and Cross-correlations of Selected CPUE Indices for the SEDAR 54 Assessment

Dean Courtney

NOAA Fisheries
Southeast Fisheries Science Center
Panama City Laboratory
3500 Delwood Beach Drive,
Panama City, FL 32408, USA
E-mail: dean.courtney@noaa.gov

September 2017

Summary

An example implementation of a hierarchical cluster analysis and cross-correlations of selected CPUE indices for the SEDAR 54 assessment was conducted to identify conflicting information among CPUE indices. Hierarchical cluster analysis identified two groupings of time-series. The first group was characterized by time-series which were highly correlated with each other and which had some highly negative correlations with some time-series not included in the group. The second group was characterized by time-series which were less highly correlated with each other or were slightly negatively correlated with each other. Because CPUEs with conflicting information were identified, it may be reasonable to assume that the indices reflect alternative hypotheses about states of nature and to run scenarios for single or sets of indices identified that represent a common hypothesis as alternative sates of nature. Cross-correlations identified strong autocorrelation in some CPUE indices over 2 to 3 years, which could indicate a year-class effect. Cross-correlations also identified strong cross correlation of lagged values of some CPUE indices (at lags between 2 to 10 years) with the current values of other CPUE indices, which could indicate that some CPUE indices represent younger age-classes than others. However, the specific lagged relationships with high correlation were not consistent among the series.

Introduction

An example implementation of a hierarchical cluster analysis and cross-correlations of selected CPUE indices for the SEDAR 54 assessment was conducted to identify conflicting information among CPUE indices. The methods were adapted from those recently implemented in an Atlantic shortfin mako assessment conducted by the International Commission for the Conservation of Atlantic Tunas (ICCAT 2017), and are provided here as an example implementation of the approach for its possible use within SEDAR.

As noted in the Atlantic shortfin mako assessment (ICCAT 2017): "…it is not uncommon for CPUE indices to contain conflicting information. However, when CPUE indices are conflicting, including them in a single assessment (either explicitly or after combining them into a single index) tends to result in parameter estimates intermediate to what would be obtained from the data sets individually. Schnute and Hilborn (1993) showed the most likely parameter values are usually not intermediate but occur at one of the apparent extremes. Including conflicting indices in a stock assessment scenario may also result in residuals not being identically and independently distributed (IID) and so procedures such as the bootstrap cannot be used to estimate parameter uncertainty. Consequently, when CPUEs with conflicting information are identified, an alternative is to assume that indices reflect hypotheses about states of nature and to run scenarios for single or sets of indices that represent a common hypothesis…"

Data Analysis

CPUE indices were evaluated for conflicting information for the combined Gulf of Mexico and South Atlantic (GOMSA) region. The agreed CPUE indices were evaluated for consistency with the average trend of the combined GOMSA indices based on a smoother fitted to year with series as a factor. Time series of residuals from the smooth fit to the agreed indices were evaluated in the combined GOMSA. Pairwise scatter plots for agreed indices were evaluated to identify correlations and high leverage points among indices. A hierarchical cluster analysis (Murtagh and Legendre, 2014) was used to group the agreed indices based on their correlations. Cross-correlations between agreed indices (i.e., the correlations between series

when each series is lagged by up to 10 years) were evaluated to identify lagged correlations (e.g., due to year-class effects).

Results

The CPUE time series are plotted in **Figure 1**, along with a smoother fitted to CPUE each year using a general additive model (GAM) to compare trends for Gulf of Mexico and South Atlantic combined, GOMSA. The overall trend for the indices is an initial decrease, a more dramatic decrease beginning in the late 1980s through the 1990s, and an increase in the 2000s continuing through the most recent years.

Residuals from the smoother fits to CPUE are compared in **Figure 2** to look at deviations from the overall trends. This allows conflicts between indices (e.g. highlighted by patterns in the residuals) to be identified. For example, in both the NMFS-LL-SE and NMFS-NE time-series, there is a series of negative residuals followed by a series of positive residuals indicating that these time-series do not follow the overall trend, and provide evidence of a more rapidly increasing trend in the stock trajectory in recent years than the overall trend. Similarly, in the PLL-OP and SEAMAP-LL-SE time-series, there is a series of positive residuals followed by a series of negative residuals indicating that these series also do not follow the overall trend, but that, in contrast, these series provide evidence of a more gradually increasing trend in the stock trajectory in recent years than the overall trend.

Correlations between indices are evaluated in **Figure 3**. The lower triangle shows the pairwise scatter plots between indices with a regression line, the upper triangle provides the correlation coefficients, and the diagonal provides the range of observations. A single influential point may cause a strong spurious correlation, so it is important to look at the plots as well as the correlation coefficients. Also, a strong correlation could be found by chance if two series only overlap for a few years.

A hierarchical cluster analysis implemented for the indices using a set of dissimilarities is provided in **Figure 4**. If indices represent the same stock components, then it is reasonable to expect them to be correlated. If indices are not correlated or are negatively correlated, i.e. they show conflicting trends, then this may result in poor fits to the data and bias in the parameter

estimates obtained within a stock assessment model. Therefore, the correlations can be used to select groups of indices that represent a common hypothesis about the evolution of the stock (ICCAT 2017).

The hierarchical cluster analysis identified two groupings of time-series. The first group included VA-LL, NMFS-LL-SE, BLLOP-2 and NMFS-NE and was characterized by time-series which were highly correlated with each other and which had some highly negative correlations with some time-series not included in the group. The second group included BLLOP-1, COASTSPAN-NE-LL, LPS, PLL-OP, SCDNR-Red-dr, COASTSPAN-SE-LL, and SEAMAP-LL-SE, and was characterized by time-series which were less highly correlated with each other or were slightly negatively correlated with each other.

Cross-correlations are plotted in **Figure 5**. The diagonals show the autocorrelations of an index lagged against itself by -10 to 10 years. The upper and lower triangles show the lagged correlation of the rows (i.e., the row lagged by -10 to 10 years) with the current value of the column.

Strong negative and positive autocorrelations over 2 to 3 years were identified for COASTSPAN SE LL, LPS, NMFS LL SE, PLL OP, and VA LL. Strong positive correlations were identified between lagged values of COASTSPAN SE LL (at lags between 2 to 10 years) and current values of LPS, NMFS LL SE, PLL OP, and VA LL.

Discussion

The hierarchical cluster analysis identified two groupings of time-series. Consequently, CPUEs with conflicting information were identified, and it may be reasonable to assume that the indices reflect alternative hypotheses about states of nature and to run scenarios for single or sets of indices identified that represent a common hypothesis.

Cross-correlations identified strong negative and positive autocorrelation in some indices over 2 to 3 years, which could indicate a year-class effect. Cross-correlations also identified strong cross correlation of lagged values of some indices (at lags between 2 to 10 years) with the current values of other indices, which could indicate that some indices reflect younger age-

classes than others. However, the specific lagged relationships with high correlation were not consistent among the series.

Acknowledgements

Analyses and figures were adapted from R Markdown code kindly provided by Laurie Kell (ICCAT secretariat). All analyses were conducted in R using FLR and the diags package. FLR provides a set of common methods for reading these data into R, plotting and summarizing them (e.g., see: http://www.flr-project.org/).

References

International Commission for the Conservation of Atlantic Tunas (ICCAT). 2017. Report of the 2017 ICCAT Shortfin Mako Data Preparatory Meeting (Madrid, Spain 28-31 March, 2017).

Murtagh F., and P. Legendre. 2014. Wards hierarchical agglomerative clustering method: Which algorithms implement wards criterion? Journal of Classification, 318 31(3): 274–295

Schnute J.T., and R. Hilborn. 1993. Analysis of contradictory data sources in fish stock assessment. Canadian Journal of Fisheries and Aquatic Sciences, 50 (9): 1916-1923.

Table 1. CPUE indices obtained for the SEDAR 54 assessment for the combined Gulf of Mexico and South Atlantic (GOMSA) region.

YEAR	LPS	BLLOP_1	BLLOP_2	VA-LL	NMFS LLSE	COASTSPAN NE LL	NMFS-NE	PLLOP	COASTSPAN SE LL	SCDNR-Red dr	SEAMAP LL SE
1960											
1961											
1962											
1963											
1964											
1965											
1966											
1967											
1968											
1969											
1970											
1971											
1972											
1973											
1974											
1975				2.362							
1976											
1977				1.629							
1978											
1979											
1980				2.106							
1981				2.406							
1982											
1983											
1984											
1985											
1986	1.183										
1987	0.363										
1988	1.184										
1989	1.352										
1990	0.471			0.299							
1991	0.762			0.408							
1992	0.584			0.149				0.593			
1993	0.261			0.755				0.483			
1994	0.175	223.74						0.192			
1995	0.138	188.64		0.606	0.215			0.304			
1996	0.164	178.42		0.626	0.110		0.0005	0.071			
1997	0.198	284.33		0.619	0.199			0.281			
1998	0.051	298.58		0.935			0.0032	0.113		0.140	
1999	0.081	168.69		0.854	0.090			0.300		0.595	
2000	0.085	103.26		0.767	0.137			0.112	0.308	0.058	
2001	0.370	360.60		0.883	0.205	3.529	0.0016	0.085	0.683	0.350	
2002	0.145	189.97		0.422	0.151	1.232		0.007	1.269	0.231	
2003	0.066	308.88		0.425	0.170	3.414		0.006	2.027	0.154	
2004	0.030	223.06		0.519	0.131	3.312	0.0015	0.110	5.876	0.338	
2005	0.156	226.42		0.298	0.049	3.524		0.032	4.275	0.155	
2006	0.046	299.50		0.795	0.083	1.815		0.161	5.078	0.279	
2007	0.104	388.02		0.251	0.214	1.864	0.0075	0.094	4.656		1.681
2008	0.135		535.52	0.834	0.162	0.581		0.109	4.894		1.205
2009	0.201		1370.66	1.188	0.409	4.620	0.0121	0.138	2.512		0.862
2010	0.106		1157.62	1.110	0.478	2.084		0.075	2.522		0.740
2011	0.086		729.47	0.624	0.371	3.351		0.097	2.864		0.346
2012	0.070		1380.63	1.146	0.636	0.862	0.0165	0.081	2.542		0.289
2013	0.275		909.50	0.959	0.443	2.400		0.128	3.015		0.301
2014	0.461		935.61	0.749	0.480	5.697		0.079	3.604		0.417
2015	0.232		1584.08	0.469	0.704	3.485	0.0270	0.126	1.177		0.589

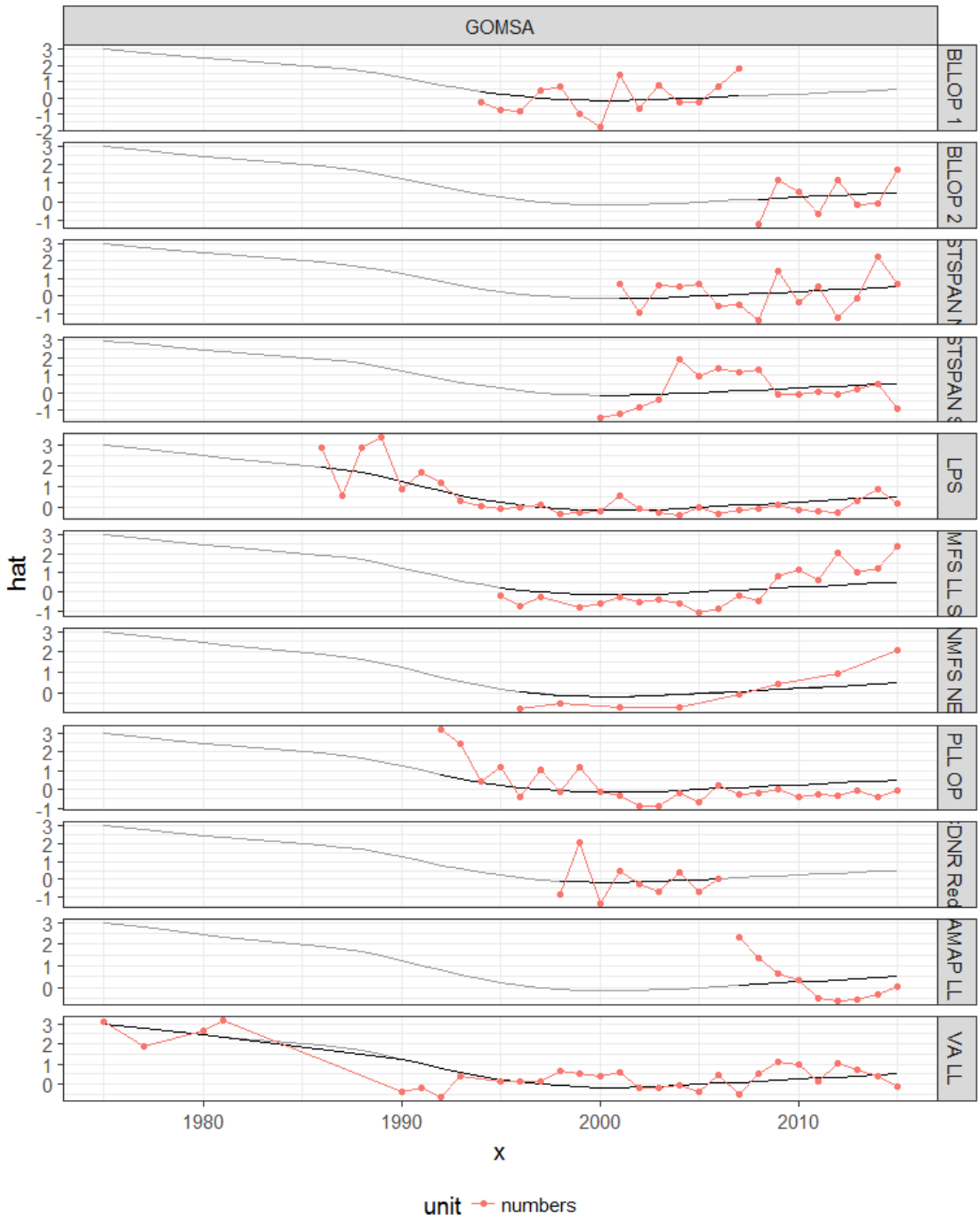


Figure 1. Smooth fit to CPUE indices obtained for the SEDAR 54 assessment for the combined Gulf of Mexico and South Atlantic (GOMSA) region. Points are the CPUE indices, continuous black lines are the smoother showing the average trend for the combined GOMSA region (i.e. GAM fitted to year with series as a factor). X-axis is time, Y-axes are the scaled indices.

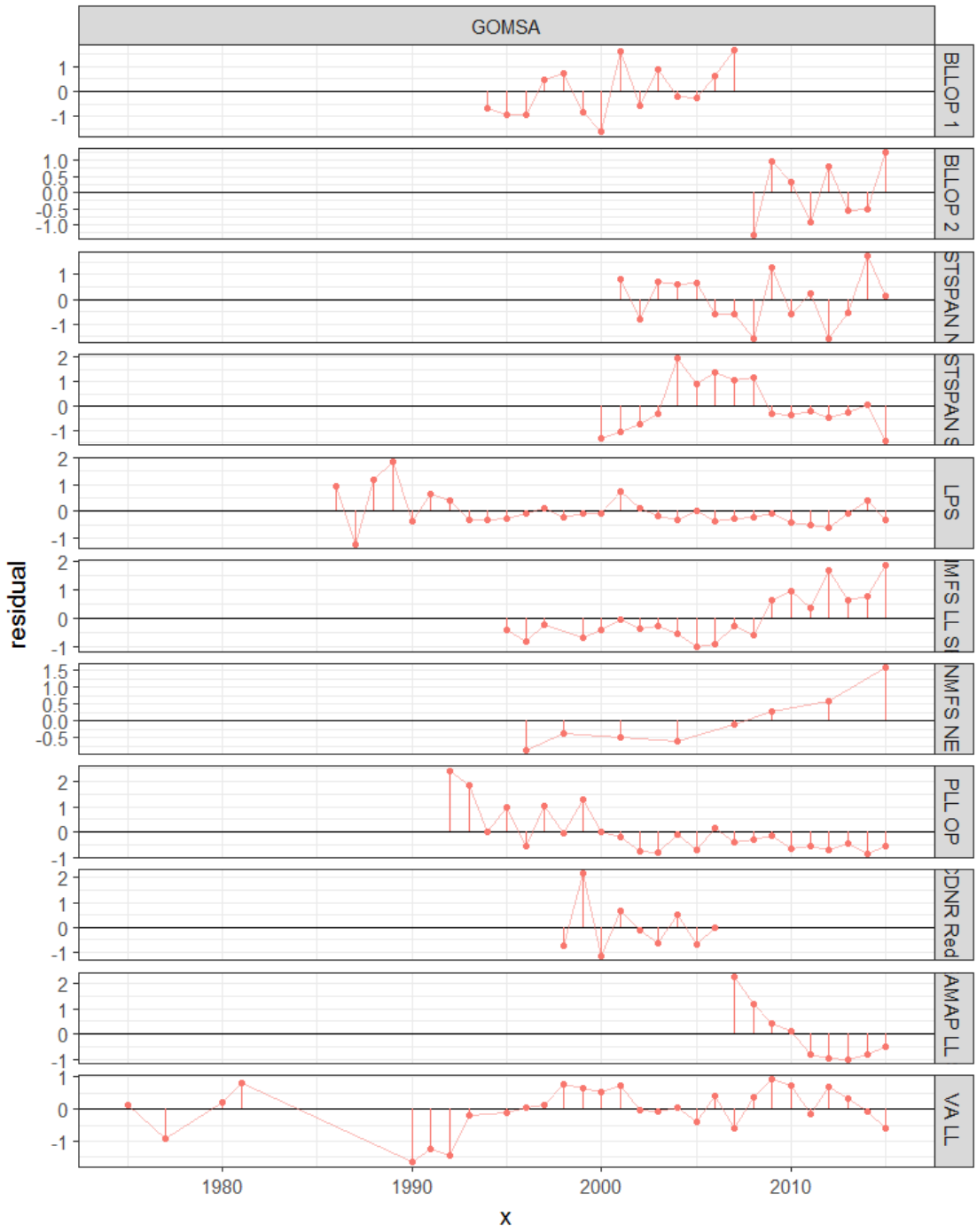


Figure 2. Residuals of the smooth fit to CPUE indices obtained for the SEDAR 54 assessment for the combined Gulf of Mexico and South Atlantic (GOMSA) region. Points are residuals of the scaled CPUE indices to the average trend for the combined GOMSA region (**Figure 1**). X-axis is time, Y-axes are residuals of the scaled indices.

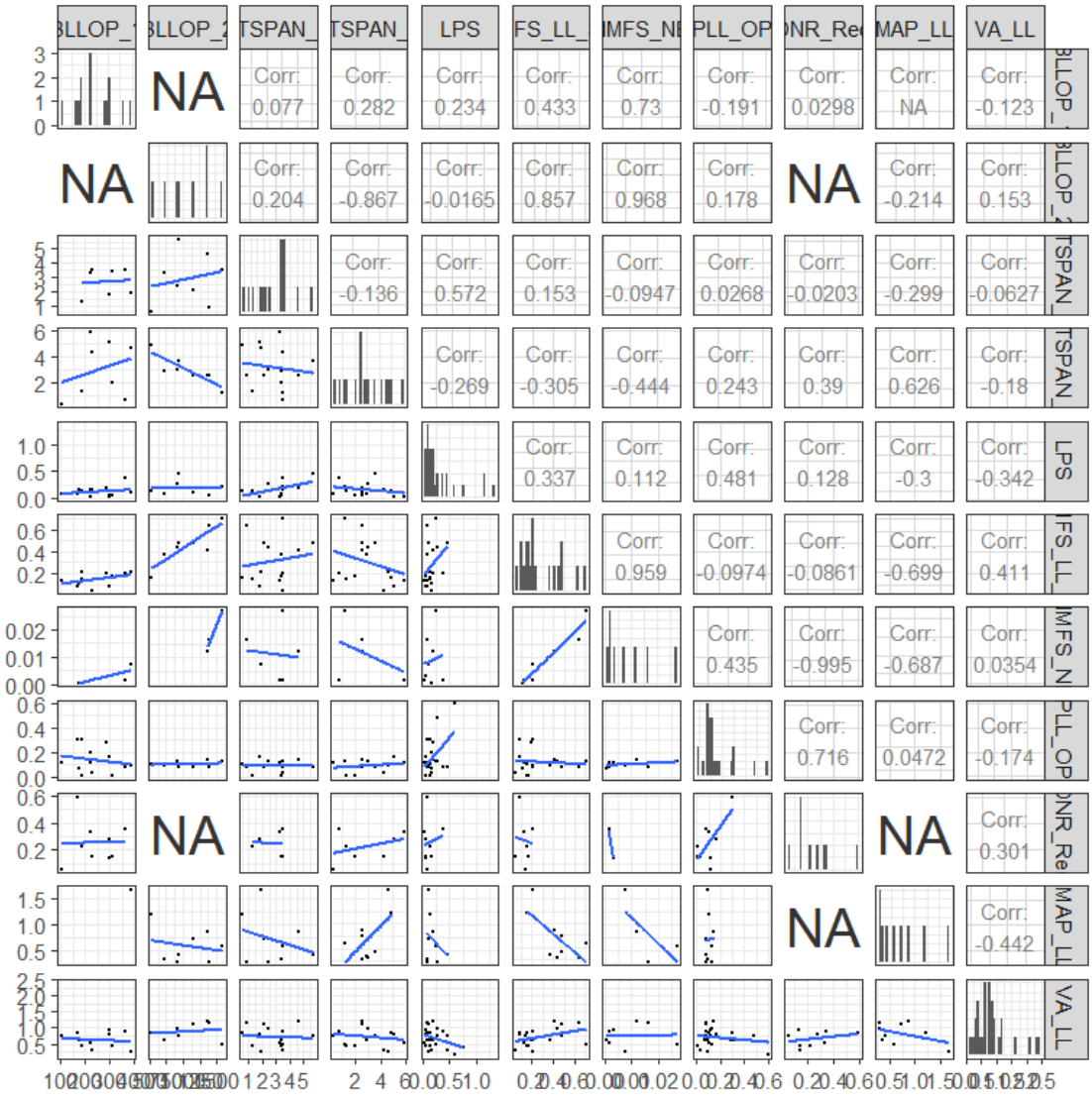


Figure 3. Pairwise scatter plots of CPUE indices obtained for the SEDAR 54 assessment for the combined Gulf of Mexico and South Atlantic (GOMSA) region. X- and Y-axis are indices.

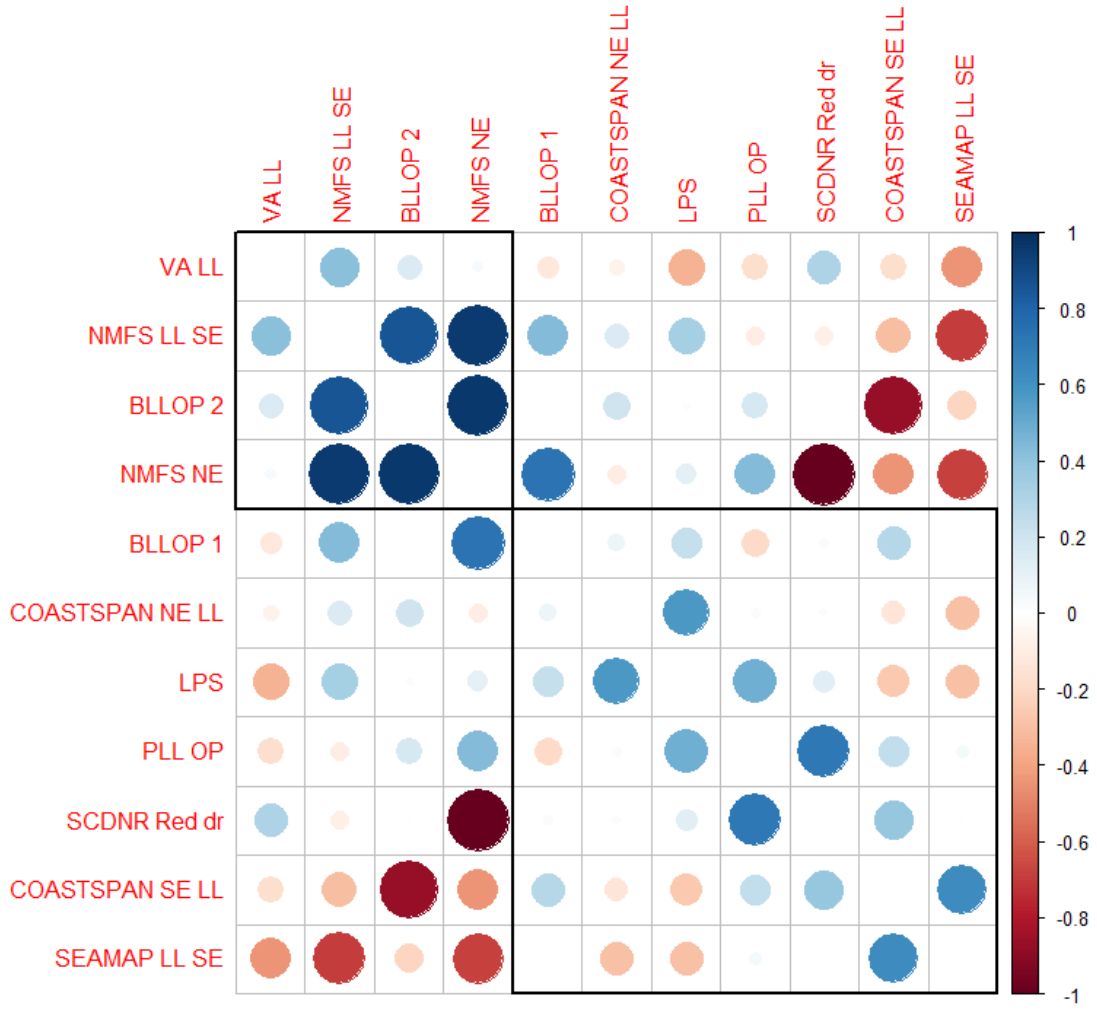


Figure 4. Correlation matrix for CPUE indices obtained for the SEDAR 54 assessment for the combined Gulf of Mexico and South Atlantic (GOMSA) region. Blue indicates positive and red negative correlations. The order of the indices and the rectangular boxes are chosen based on a hierarchical cluster analysis using a set of dissimilarities.

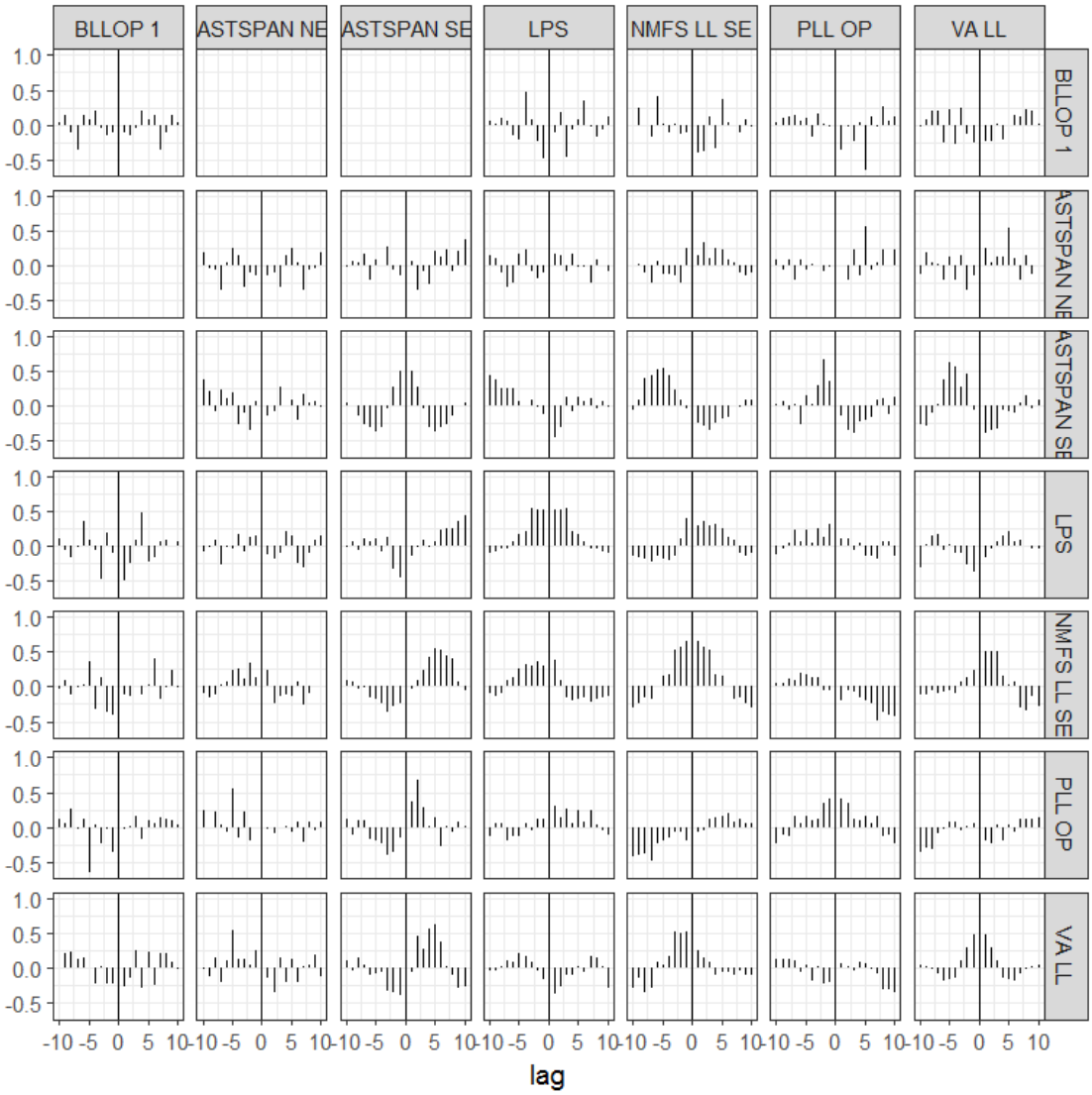


Figure 5. Cross-correlations between CPUE indices obtained for the SEDAR 54 assessment for the combined Gulf of Mexico and South Atlantic (GOMSA) region. X-axis is the cross-correlation at each lag, and Y-axis is cross-correlation lag number.