SCRS/2023/104 Collect. Vol. Sci. Pap. ICCAT, 80(1): 168-175 (2023)

A HIERARCHICAL CLUSTER ANALYSIS OF SOUTH ATLANTIC SWORDFISH CPUE SERIES

Nathan G. Taylor[1]

SUMMARY

This document presents a method for clustering CPUE series with similar trends and applies it for Southern CPUE series. The method consists of visual examination of the series with Lowess fitting, residual plots, cross-correlation analysis, and hierarchical cluster analysis. The hierarchical cluster analysis uses complete linkage clustering. This computes all pairwise dissimilarities between the correlation coefficients in cluster 1 and the elements in cluster 2. It then considers the largest value of these dissimilarities as the distance between two clusters. I focused on defining two clusters. The analysis shows a highly correlated group of indices (the Uruguayan longline, the Uruguayan historical longline, the South African longline, the late Japanese longline, the western Spanish longline, the Chinese Taipei longline, and the northern Spanish longline), and a second group that have different trajectories (the early Japanese longline, the Brazilian longline, and the late Chinese Taipei longline series). Each cluster can define scenarios or Operating Models in stock assessment or Management Strategy Evaluation, respectively.

RÉSUMÉ

Ce document présente une méthode de regroupement des séries de CPUE présentant des tendances similaires et une application aux séries de CPUE du Sud. La méthode consiste en un examen visuel des séries au moyen de l'ajustement de Lowess, de graphiques des valeurs résiduelles, d'une analyse de corrélation croisée et d'une analyse hiérarchique en grappes. L'analyse hiérarchique en grappes utilise le regroupement par liens complets. Elle calcule toutes les différences par paire entre les coefficients de corrélation de la grappe 1 et les éléments de la grappe 2. Elle considère ensuite la plus grande valeur de ces différences comme la distance entre les deux grappes. Ce document se concentre sur la définition de deux grappes. L'analyse montre un groupe d'indices fortement corrélés (la palangre uruguayenne, la palangre historique uruguayenne, la palangre sud-africaine, la palangre japonaise de la fin de la série, la palangre espagnole occidentale, la palangre du Taipei chinois et la palangre espagnole septentrionale) et un autre groupe qui présente des trajectoires différentes (la palangre japonaise du début de la série, la palangre brésilienne et la palangre du Taipei chinois de la fin de la série). Chaque grappe peut définir des scénarios ou des modèles opérationnels pour l'évaluation des stocks ou l'évaluation des stratégies de gestion, respectivement.

RESUMEN

Este documento presenta un método para agrupar series de CPUE con tendencias similares y lo aplica a las series de CPUE del sur. El método consiste en un examen visual de las series con ajuste de Lowess, gráficos de residuos, análisis de correlación cruzada y análisis jerárquico de conglomerados. El análisis jerárquico de conglomerados utiliza la agrupación completa de las vinculaciones. En él se calculan todas las disimilitudes por pares entre los coeficientes de correlación del conglomerado 1 y los elementos del conglomerado 2. A continuación se considera el mayor valor de estas disimilitudes como la distancia entre dos conglomerados. El análisis se centra en la definición de los conglomerados. El análisis muestra un grupo de índices altamente correlacionados (el palangre uruguayo, el palangre histórico uruguayo, el palangre sudafricano, el palangre japonés tardío, el palangre español occidental, el palangre de Taipei Chino y el palangre español septentrional), y un segundo grupo que tienen trayectorias diferentes (las series del palangre japonés temprano, el palangre brasileño y el palangre del Taipei Chino tardío). Cada grupo puede definir escenarios o modelos operativos en la evaluación de stocks o en la evaluación de estrategias de ordenación, respectivamente

---

[1] ICCAT Secretariat. Nathan.Taylor@iccat.int

KEYWORDS

Cluster analysis, Operating models, Catch statistics, Fishery statistics, Catch/effort, Autocorrelation, Artificial intelligence, Stock assessment, Time series analysis

Introduction

In fisheries stock assessments that have several available CPUE series, choosing CPUE series or given sets of CPUE series essentially constitute making different hypotheses about the trajectory of the stock. Consider a simplified example of a stock assessment having two indices with one index going up and another index going down. Let us assume that each of these indices are supposed to represent the same ages and areas of the stock. In such situations, the first problem is logical: both indices cannot be representative of the stock. If they are assumed to be so, then the selection of only the increasing index will likely result in a stock assessment showing an increasing trend, but choosing to fit the model only to the index in decline will probably cause the assessment to produce a negative trend. Fitting both CPUE series is not always a satisfactory solution either: fitting two conflicting CPUE series can result in a nearly flat stock trajectory that results in the model failing to fit either index.

What should be done with conflicting indices in a stock assessment? A possible solution to this dilemma in the simple two index scenario is to present one model fit to one index (a hypothesis that the stock is rising) and a second model fit to the declining index (a hypothesis that the stock is declining. This reformulates the model from a statistical problem to a decision problem where absent any reason to reject one index, the decision maker must choose between behaving like the stock is increasing or like the stock is decreasing. Using a Traditional Approach (Butterworth 2007) or Best Assessment (BA), paradigm this could be illustrated with scenarios considering the effects of total allowable catches given either the increasing or the decreasing stock hypotheses. In the MSE paradigm, these scenarios can be considered with different Operating Models; Management Procedures can be demonstrated to be robust (or not) to both hypotheses about the state of the system.

While having two indices is a relatively simple but challenging scenario, stock assessment models in tuna assessment may be fit to five, ten, or even as many as twenty indices (BFT). When there are multiple indices in a stock assessment, the situation is more complicated. There can be increasing, decreasing, and flat trends. Beyond simple visual analyses, selecting clusters of indices that reflect alternative hypotheses about the stock requires more elaborate statistical techniques to justify these choices.

Delineating indices into similar groups can be done using a technique called Hierarchical Cluster Analysis (Frank *et al.* 2016). Hierarchical Cluster Analysis calculates a measure of dissimilarity between datasets or time series. In most methods of hierarchical clustering, this is achieved by use of an appropriate distance d, such as the Euclidean distance, between single observations of the data set, and a linkage criterion, which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets (Frank *et al.* 2016, Murtagh *et al*. 2014). Such analyses have been conducted for the purpose of grouping CPUE series (Rice *et al.* 2023) in Indian Ocean Albacore Tuna and I broadly follow their analysis here.

Methods

I explore the degree of correlations between the CPUE series using a series of techniques. These are: visual examination of the series with Lowess fitting, residual plots, cross-correlation analysis, and hierarchical cluster analysis. R (R Core Team 2023) has functions to hierarchical clustering analysis relatively simply. In this case, the distance d is calculated in terms of the relative degree similarly in the trends between CPUE series expressed in term of the correlation coefficients between them.

The analysis is summarized in a series of figures. The CPUE time series for swordfish in the South Atlantic are plotted in **Figure 1**, along with a Lowess smoother fitted to CPUE each year using a general additive model (GAM) to compare trends for the submitted CPUEs. Residuals from the smoother fits to CPUE are compared in **Figure 2** to look at deviations from the overall trends. 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 examine plots. Also, a strong correlation could be found by chance if two series only overlap for a few years. A hierarchical cluster analysis evaluated 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, correlations can be used to select groups of indices that represent a common hypothesis about the evolution of the stock (ICCAT 2017).

All analyses were conducted using R and FLR and the diags package which provides a set of common methods for reading these data into R, plotting and summarizing them. The hierarchical cluster analysis was performed using the corrplot package. CPUE indices were clustered using complete linkage clustering. This computes all pairwise dissimilarities between the correlation coefficients in cluster 1 and the elements in cluster 2, and considers the largest value (i.e., maximum value) of these dissimilarities as the distance between the two clusters.

Results and Discussion

The overall trend for all indices is a general declining trend from the beginning of the time series until approximately 2010 (**Figure 1**). The trajectories diverge in this year: whereas some CPUE time series continue to decline after 2010 such as:Brazilian longline (BRA.LL), Chinese Tai Pei longline (CPT.LL2), South African Longline, ZAF.LL). Others longline series show slight increases after 2010 like Japanese Longline (JPN.LL) as well as the Spanish Longline series (W.SPN.LL and N.SPN.LL).

Figure 2 is a plot of the Lowess fit residuals. This allows for any conflicts between indices to be highlighted by similar residual patterns. For example, the indices that show declines, i.e.,Brazilian longline (BRA.LL), Chinese Tai Pei longline (CPT.LL2), South African Longline, ZAF.LL all have negative residuals in the recent years whereas the increasing indices have positive residuals.

Figure 3 is a the pairwise correlation plot of the CPUE series. The upper right triangle of the matrix shows the correlation coefficient between the CPUEs and the degree of significance of the correlation (see **Figure 3** label). There were very strong and significant positive correlations between ZAF.LL and URU.LL.hist, n.SPN.LL and w.SPN.LL, URU.LL.hist and BRA.LL indices. Negative correlations tended not to be significant and smaller in magnitude.

Figure 4 shows the Hierarchical Clustering results. It identifies two groupings of CPUE series. The first cluster includes URU.LL, URU.LL.hist, ZAF.LL, JPM.LL1, JPN.LL2, w.SPN.LL, CTP.LL1, nSPN.LL. This group has CPUE series that are highly correlated with each other, and which have some highly negative correlations with some CPUE series in the second group. The second cluster includes the other indices i.e., JPN.LL1, BRA.LL, CTP.LL2. This cluster has CPUE series that have small correlation coefficients with each other and/or are negatively correlated with one or more of the indices in the first cluster.

So how could this kind of analysis be used in practice? The two clusters in the hierarchical clustering analysis can define different stock assessment or different OM scenarios. Effectively, choosing one cluster over another would replace the choice of the increasing or decreasing index discussed in the Introduction. Accordingly, the BA paradigm, one assessment model would fit to the indices of one cluster and the second scenario would consider the second cluster. In the MSE paradigm each cluster would consist of an OM or a factor to be tested (in combination with other factors) in a set of OMs.

I only considered two clusters for this analysis. There is nothing in the analytical method that requires there be two clusters. Three clusters could be considered as well; this could include for example a set of indices with largely flat trends. In this instance, I did not explore this option very much because the indices in the case are relatively coherent. This coherence between indices is not always the case: the Southern Swordfish stock does not present a challenging set of indices to cluster because barring the last 10-15 years, the indices show very similar (declining trends) but other fish stocks may show pronounced differences in CPUE trends that would results in very different CPUE trends.

The effects of choosing one cluster or another for southern swordfish remains to be seen. A priority would be to consider different clusters in the OMs. This would illustrate the effect of these choices on inferred parameter set that defines each OM. The set of Management Strategy Evaluation simulation would illustrate which MPs would be robust to different trajectories of the stock.

References

Butterworth, D. S. 2007. Why a management procedure approach? Some positives and negatives. ICES Journal of Marine Science, 64: 613–617.

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).

Kell, L. T., I. Mosqueira, P. Grosjean, J-M. Fromentin, D. Garcia, R. Hillary, E. Jardim, S. Mardle, M. A. Pastoors, J. J. Poos, F. Scott, R. D. Scott. 2007. FLR: an open-source framework for the evaluation and development of management strategies. ICES J Mar Sci, 64 (4): 640-646. doi: 10.1093/icesjms/fsm012

Murtagh, Fionn and Legendre, Pierre 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? Journal of Classification, 31, 274–295. doi:10.1007/s00357-014-9161-z.

Nielsen, F. 2016. "8. Hierarchical Clustering". Introduction to HPC with MPI for Data Science. Springer. pp. 195–211. ISBN 978-3-319-21903-5.

R Core Team 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rice, J. 2022. Stock assessment of albacore tuna (Thunnus alalunga) in the Indian Ocean using Stock Synthesis. IOTC-2022-WPTmT_8 (Virtual meeting July-11-22 2022)

Taylor, N. G. , McAllister, M. K., Lawson, G. L., Carruthers, T., & Block, B. A. 2011. Atlantic bluefin tuna: A novel multistock spatial model for assessing population biomass. PLoS ONE, 6(12). https://doi.org/10.1371/journal.pone.0027693

Taylor, N.G., Murato, B., and Parker, D., 2022. Preliminary close-loop simulation of Management Procedure performance for southern swordfish. Collect. Vol. Sci. Pap. ICCAT, 79(2): 705-714

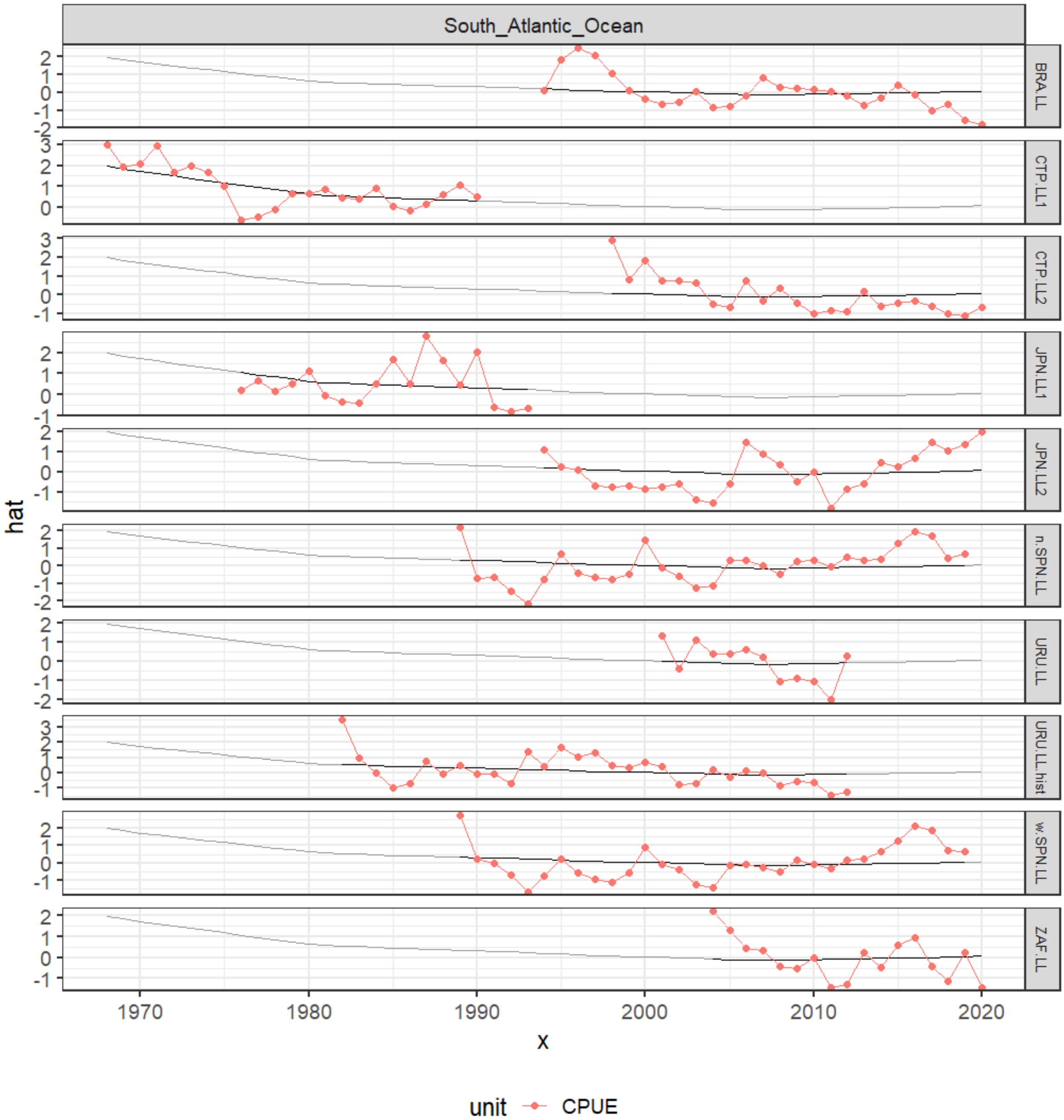

Figure 1. South Atlantic time series of agreed CPUE indices. Points are the standardized values. Continuous black lines are a lowess smoother showing the average trend by area (i.e., fitted to year for each area with series as a factor). X-axis is time, Y-axis are the scaled indices.

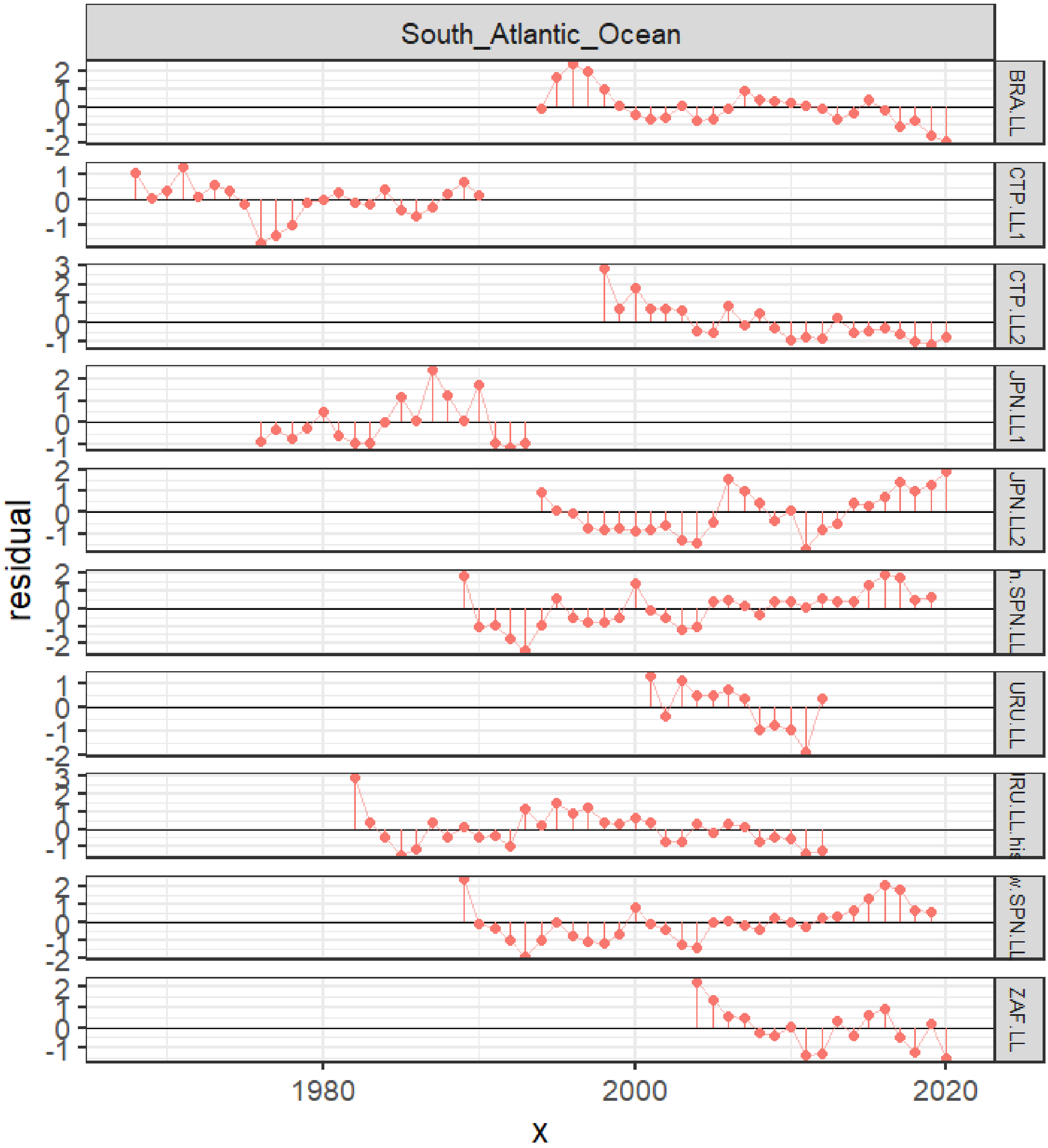

Figure 2. Time series of residuals from the smooth fit to CPUE indices. X-axis is time, Y-axis are the scaled indices.

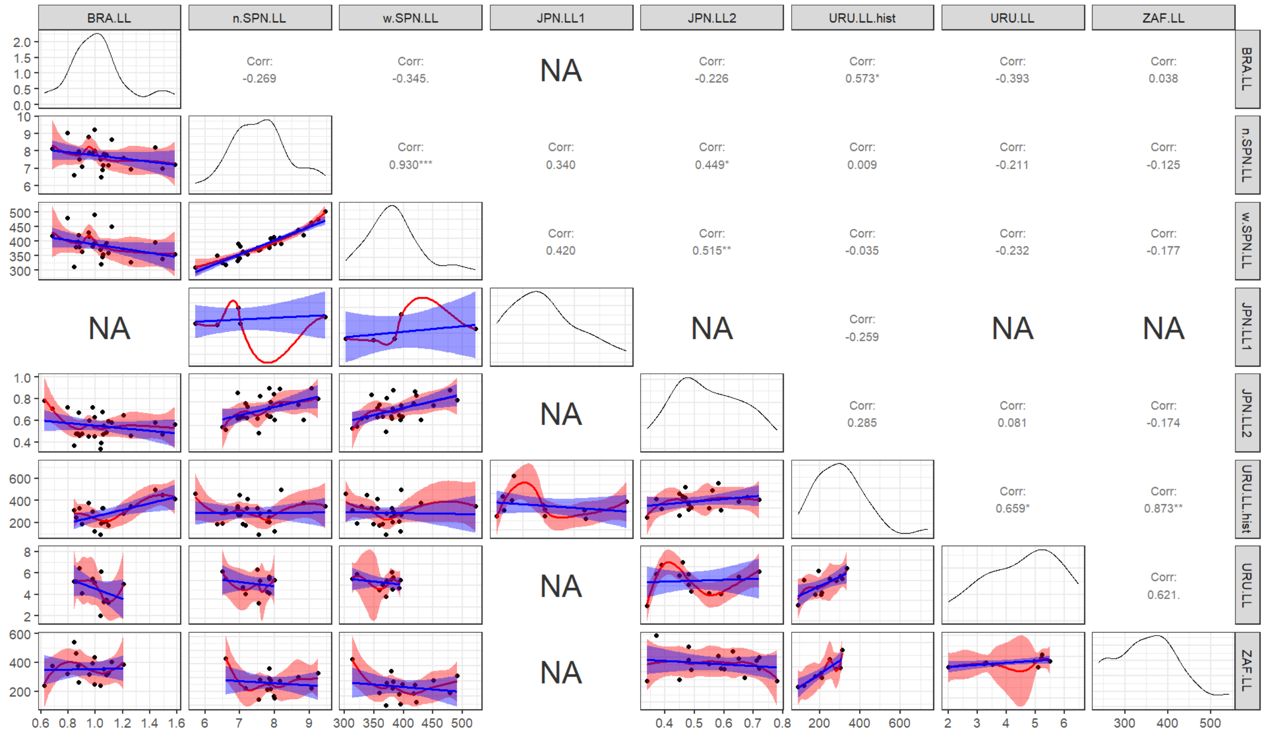

Figure 3. Pairwise scatter plots for CPUE indices. X- and Y-axis are the unscaled indices. The upper right triangle represents the correlation coefficient. They are labeled as follows: *** if the p-value is < 0.00; ** if the p-value is < 0.01; * if the p-value is < 0.05; and ". " if the p-value is < 0.10.

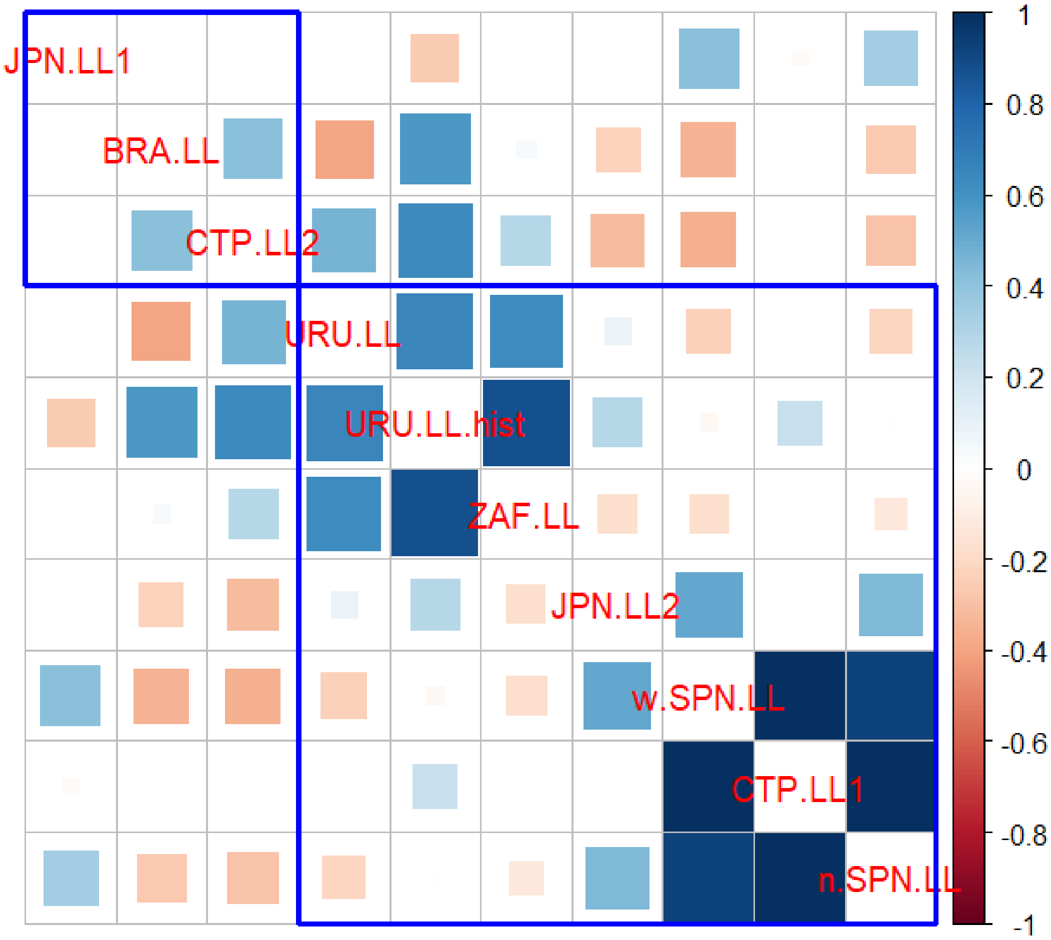

Figure 4. Correlation matrix for CPUE indices. Blue indicates positive and red negative correlations. The order of the indices and the rectangular boxes are chosen based on a hierarchical cluster analysis using a set of correlation dissimilarities.