Dynamic Factor Analysis of pre-recruit data

Introduction

This is a quick summary of (1) a PCA, (2) a Dynamic Factor Analysis (DFA), and (3) a Bayesian DFA on timeseries pelagic juvenile (pre-recruit) rockfishes. There are a potentially a few things to finalize before producing a write up, but you could probably steal most of the text from this document.

As noted above, I have done three sets of analyses for comparison. I reproduced the original PCA just for my own edification and to have as a reference in this "document". Originally, I had planned on using a regular, MARSS-style DFA, so those results are here. I kept them for comparison, again mostly for my own reference. The primary results are the results from the Bayesian DFA from 'bayesdfa' in R.

Note: this is just a quick update adding in AR1 structure to the Bayesian DFA model.

Methods

Because DFA allows us to include time series with missing data, I recalculated the indices of abundance from the raw data for each region by calculating the mean log(x+1) abundance for each of the 12 regions from 2001 - 2019. For the DFA analyses, zeros were converted to 'NAs' after calculating the mean. Including a zero in the DFA analyses would assert that there were no rockfish pre-recruits in the water when in reality there were probably very few but sampling did not encounter them. Including the NA allows the model to 'interpolate' (sort of) these values based on a Kalman Filters. In some cases, the DFA models do not run well if there are zeros. Prior to analysis the time series were standardized (z-scored).

Note, the MS text at present reads that the indices were calculated using delta-glms but then notes that the PCA analysis is preliminary and uses regional means. If you choose to use delta-glms, I can always rerun the analysis pretty quickly.

Results

Below is a plot of the mean log(x+1) pre-recruits indices by region for 2001 to 2019 for all regions.

For the PCA analysis reproduced here, I used the same data as in the orignal analysis with limited regions and only those years with positive data. The following plot shows just the indices used in the orginal PCA analysis (repeated here for my own edification).

Principal components analysis

As I noted above, I re-ran the PCA mostly to have the results in one section.

Below is a biplot of the first two PCs connected by year. These are handy for showing a change in the overal state of the system. For example, you can see that 2013-2017 are quite different from the rest of the years and that 2014-2016 were also some what different to just before and just after. Also note that the middle blob years (2014-2017) are separate from just before (2013) and just after (2017) in terms of spatial variation in recruitment.

I have not reproduced the loadings here because we mostlikely won't use this output.

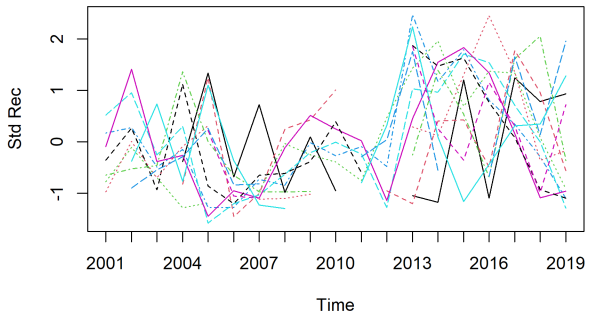


Figure 1: Plot of mean logx+1 pre-recruits by region for 2001 to 2019 for all regions. Data were standardiZed by substracting the mean and dividing by the standard deviation.

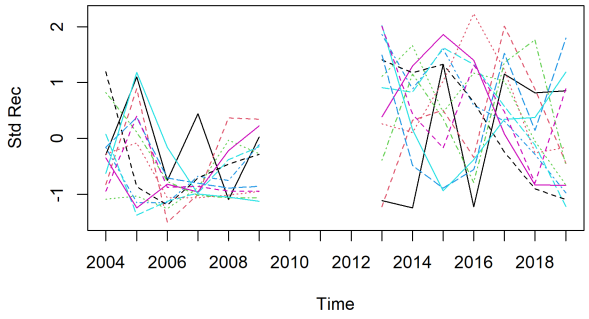


Figure 2: Time series used in the PCA analysis.

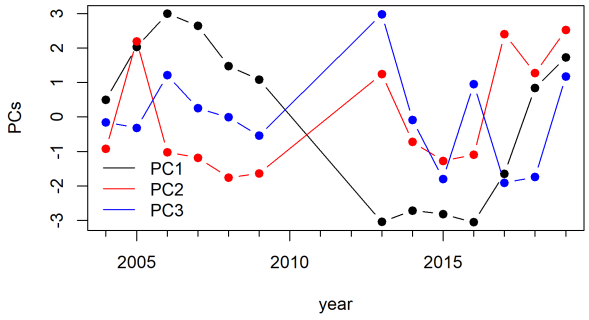
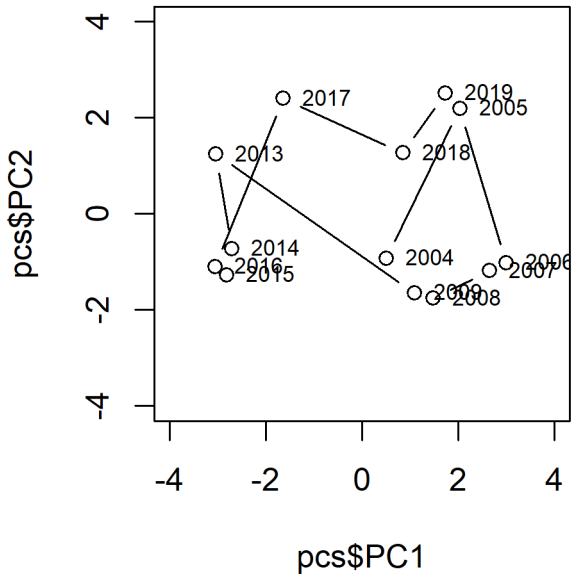


Figure 3: Principal components



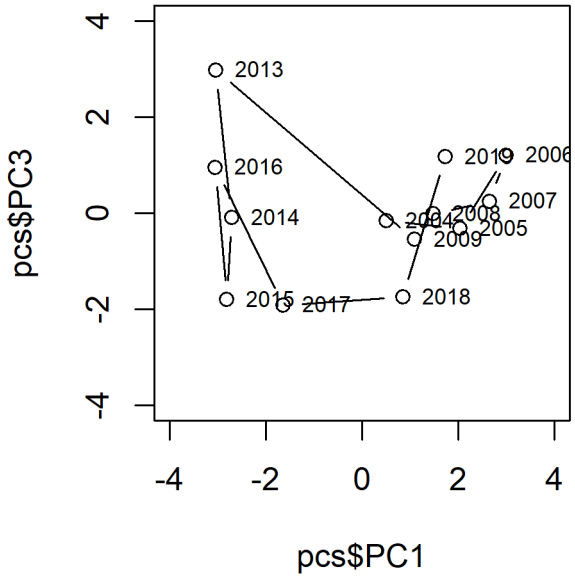


Figure 4: Biplot of PC1 vs PC2

Regular MARSS DFA

I ran a regular DFA (MARSS package) primarily becauase I started this before emailing a bit with Rick B. I switched to a Bayesian DFA below, but I kept these results for comparison with the Bayesian DFA.

The MARSS-based DFA here uses all time series (12) and all years (19). In part, it is important to include the missing years so that the model knows how far apart observations are given that X_t is a function of X_{t-1}. The model can then estimate those missing values.

There were 2 models with delta AICc values The best fit model (lowest delta AICc) had 2 dynamic factors and a diagonal and unequal R matrix.Model fits are shown below. We might chose the m = 1 model becasue it has the fewest parameters. However, I've taken the m = 2 model as the one with the lowest AIcc. It does not really matter because we will use the Bayesian DFA.

```
##                         model m      AIC     AICc n_par converge     delta
## 2 diagonal and unequal m = 2 2 479.9119 495.5640    35      yes  0.000000
## 1 diagonal and unequal m = 1 1 490.1143 497.0910    24      yes  1.526993
## 6   diagonal and equal m = 2 2 496.0095 502.9862    24      yes  7.422173
## 7   diagonal and equal m = 3 3 490.1908 504.8822    34      yes  9.318127
## 5   diagonal and equal m = 1 1 505.6869 507.6760    13      yes 12.111913
## 3 diagonal and unequal m = 3 3 480.4419 507.8591    45      yes 12.295102
## 8   diagonal and equal m = 4 4 504.1358 528.8678    43      yes 33.303747
## 4 diagonal and unequal m = 4 4 492.1163 533.9473    54      yes 38.383260
```

Below are the model fits to the data for each region including the correlation between the model fit and the observed data. Overall, the model does pretty well, but does not predict the two souther regions (SCI and NCI) particularly well.

DFA Loadings

The DFA loadings are shown here. They are qualitatively similar to the PCA results with a north & south grouping and a central grouping of regions.

DFA time series

Below are the two dynamic factors trough time (with 95% CLs) and a biplot showing the state of the system through time.

Overall, the DFA gives similar byt slightly different results from the PCA. We can still see the division along the first DF puttingl the blob years off to the left. You can see essentially two states of nature contrasting 2013-2017 vs other years on DF1 with DF2 explaining some of the variation within those two big regimes. Again, 2013 (just before) and 2017 (just after) are different from the blob years and other years.

BAYESIAN Dynamic factor analysis

I expect this is the analysis we would use in the MS. I've included methods text regarding the calculation of the indices for completeness.

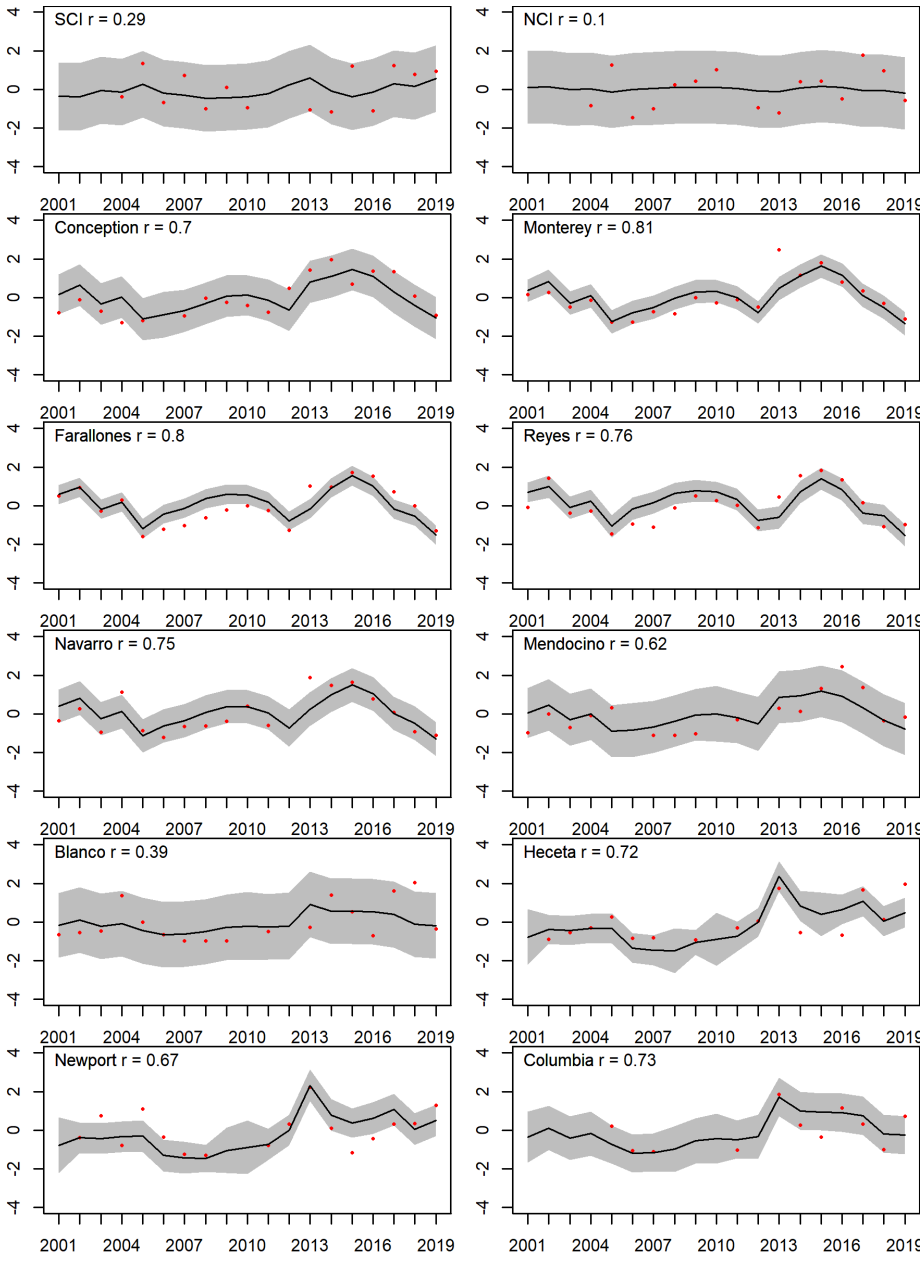


Figure 5: Model fits for the DFAs to observed data. r is the correlations between the DFA model and the original data.

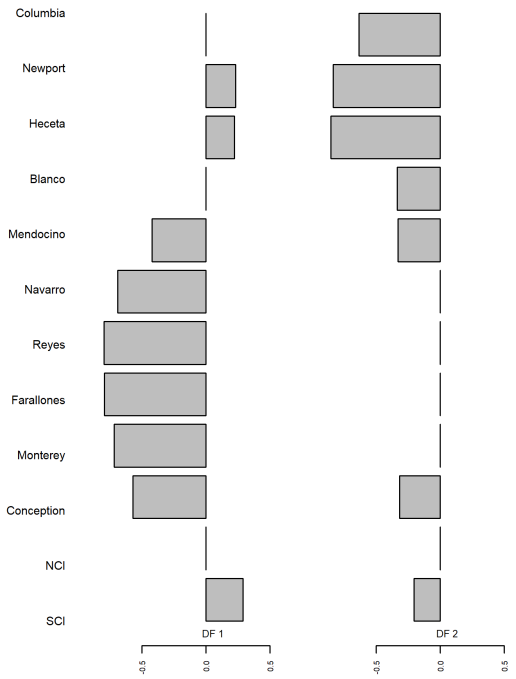


Figure 6: Dynamic factor loadings. Loadings less than 0.2 (absolute value) are not shown.

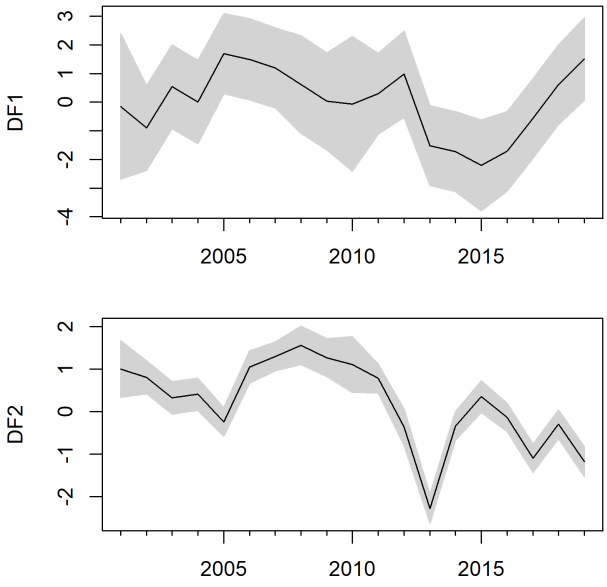


Figure 7: Time series of the dynamic factors. Error envelopes are 95 percent CLs. Note the difference in scale on the axes.

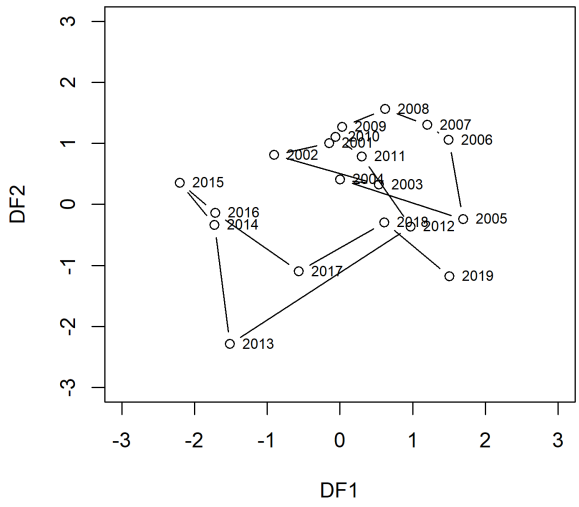


Figure 8: Biplot of dynamic factors

Methods

We used dynamic factor analysis (DFA; Zuur etal 2003a,b) in a Bayesian framework (Ward et al 2018) to look for common trends among regions in the abundance of pre-recruits through time from 2001-2019. DFA is analagous to pricipal components analysis in that it is a data reduction technique that looks for common patterns among a group of variables, but it is relevant for time-series analysis because x_t is function of x_{t-1} (Zuur etal 2003a,b). More specifically, time series are modeled as a linear combination of latent (hidden) trends. These trends reflect the shared temporal variation among the time series (here regions) and error terms, which are specific to populations: $y_{i,t}i, = Zx_{i,t} + v_{i,t}$. The latent trends ($x_t$) are a function of x_{t-1} with noise component (w): $x_{i,t} = \phi x_{i,t-1} + w_{i,t}$ and $w_{i,t} \sim$ MVN(0,1). When ϕ approaches 1.0 the the trend behaves as a random walk. When ϕ approaches zero the trend behaves as white noise. The Z matrix contains factor loadings, and the residual error is $v_{i,t} \sim$ MVN(0,R), where R is the variance-covariance matrix (). An important additional feature of DFA is that it can handle time series with missing data through the use of Kalman Filters (Zuur etal 2003a,b).

We recalculated the indices of total abundance for pre-recruits from the raw data for each region by calculating the mean log(x+1) abundance for each of the 12 regions from 2001 - 2019. Zeros were converted to 'NAs' after calculating the mean. Including a zero in the DFA analysis would assert that there were no rockfish pre-recruits in the water when in reality there were probably very few but sampling did not encounter them. Including the NA allows the model to predict these missing values based on the use of Kalman Filters (Zurr 2003a,b). Prior to analysis the time series were standardized (z-scored).

We fit 12 Bayesian DFA models using the time series for all 12 sites from 2001 - 2019. Models were fit using the 'bayesdfa' package in R v4.0.0 (R Core Team 2020, Ward et al 2019). Model fitting included 5000 iterations and 4 chains. We considered models with up the three dynamic factors and with normal or studentized errors. We also considered models with diagonal and equal or diangonal and unequal observation variance-covariance matrices (R matrix).

We compared models using a combination of metrics: (1) leave one one information critera (looic) using Pareto-smoothed importance sampling (Vehtari et al 2017), (2) expected log predictive density (elpd) among models using the 'loo' package in R (Vehtari et al 2019), (3) model weights, (4) model fits to the original data, and (5) number of model parameters. Convergence was evaluated using the effective sample size, and

Rhat.

Results

Based on looic values, the best-fit model had three DFs, studentized errors, and diagonal and unequal R (m3-s-un). However, the elpd +/- SE for a similar but more simple model (m3-s-eq) overlapped that of the former and had similar weight. The 3m-s-eq model was more simlpe because the diagonal and equal R matrix had fewer estimated variances (1) than did the R matrix for m3_s_un with separate variance estated for each time series (12). The model with equal variance also produced better fits (higher correlations) to the original data. Thus we chose the best-fit model as the one with three dynamic factors (3DFs), studentized errors, and diagonal and equal R (3m_s_eq).

I added in AR1 structure to this 3DF model to have a look at the resulting values. Once John finalizes the recruit indices, I'll re-run things and go through the model selection again

Note For reference in the following tables * m3 = 3 dynamic factors * s = student errors, Student errors are useful when there are extermen values * n = normal errors * un = diagonal and unequal observation error matrix * eq = diagonal and equal observation error matrix * ar1 includes first order autocorrlation * ma1 includes multivatriate autocorrelation

Model weights

```
## Method: stacking
## ------
##             weight
## m1_s_eq     0.000
## m1_n_eq     0.000
## m2_s_eq     0.000
## m2_n_eq     0.000
## m3_s_eq     0.409
## m3_n_eq     0.000
## m1_s_un     0.000
## m1_n_un     0.000
## m2_s_un     0.000
## m2_n_un     0.000
## m3_s_un     0.311
## m3_n_un     0.278
## m3_s_eq_ar1 0.002
```

Autocorrelation

Remember that: $x_{i,t} = \phi x_{i,t-1} + w_{i,t}$. When ϕ approaches 1.0 the the trend behaves as a random walk. When ϕ approaches zero the trend behaves as white noise. Violin plots of ϕ are shown below. They are a bit ugly. I'll clean them up later.

Trend 1 tends a bit more to the white noise side of things, while Trend 3 follows more of a random walk. Trend 2 is somewhere in the middle. So, Trend 3 (coast-wide recruitment here) depends a more on the previous year that does the latitudinal variation among regions.

In the initial round of DFAs, I did not include AR1. Functionally that just sets $\phi = 1.0$. I think we can just choose to include AR1 structure because it is interesting.

```
## Using  as id variables
```

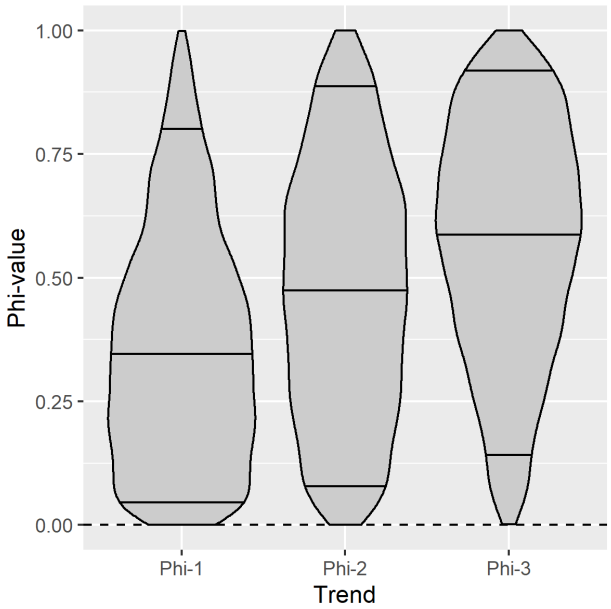


Figure 9: Plot of phi parameters. Lines within the violin plot are the 5%, 50% and 95% percent quantiles.

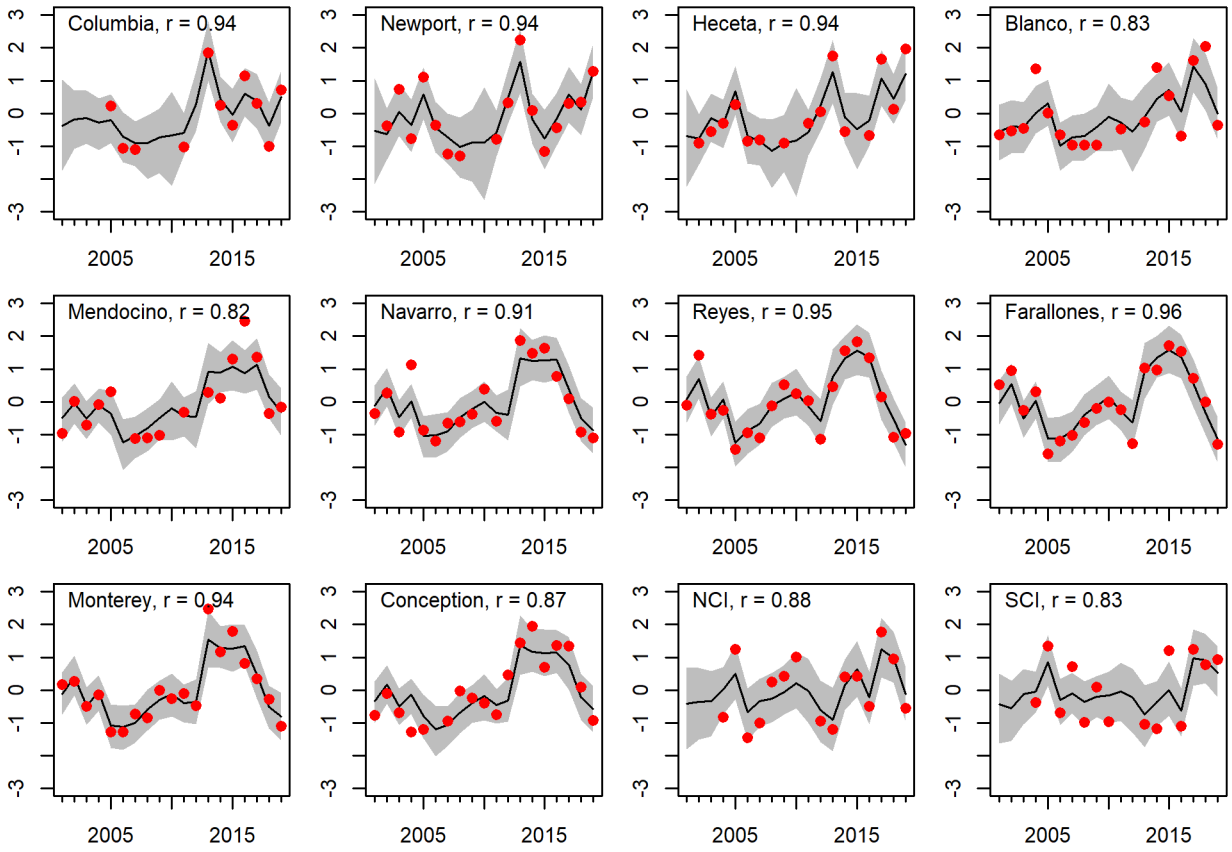


Figure 10: Fit of the DFA model to the original data, including the correlations between the model fit and the observed data. Points are the observed data; lines are the fitted model prediction. Grey envelopes indicate 95% CL. Regions are orders north-south moving left to right across each row and then down rows.

The DFA model fits the observed data well with correlations between the DFA and the original data all above r = 0.8. While still strong, model fits were lowest from Point Conception to the south.

Trend 1 captures coherence is pre-recruit abunance among loctions from Cape Mendocino to Heceta back and those south of Point Conception, which all had positive loadings. Mid-latitide sites had negative or zero loadings, while the Columbia River region, the farthest north, also had negative loadings.

Trend 2 also captured latitudinal differneces in pre-recruit trends. The far northern sites, Heceta to Columbia, had negative loadings. Regions from Cape Blanco to had more positive loadings, but loadings for the southern most region (SCI) were negative indicating that it trended similarly to the northern sites. Thus, both trend 1 and trend 2 indicate some coherence in pre-recruit abundance between northern and southern regions, with the central regions showing the opposite trends.

Trend 3 (DF3) captured coherent variation in recruitment along the entire coast with most regions having positive loadings (although some overlap zero).

Note that the order if the dynamic factors is not meaningful as is the order of compenents in a PCA.

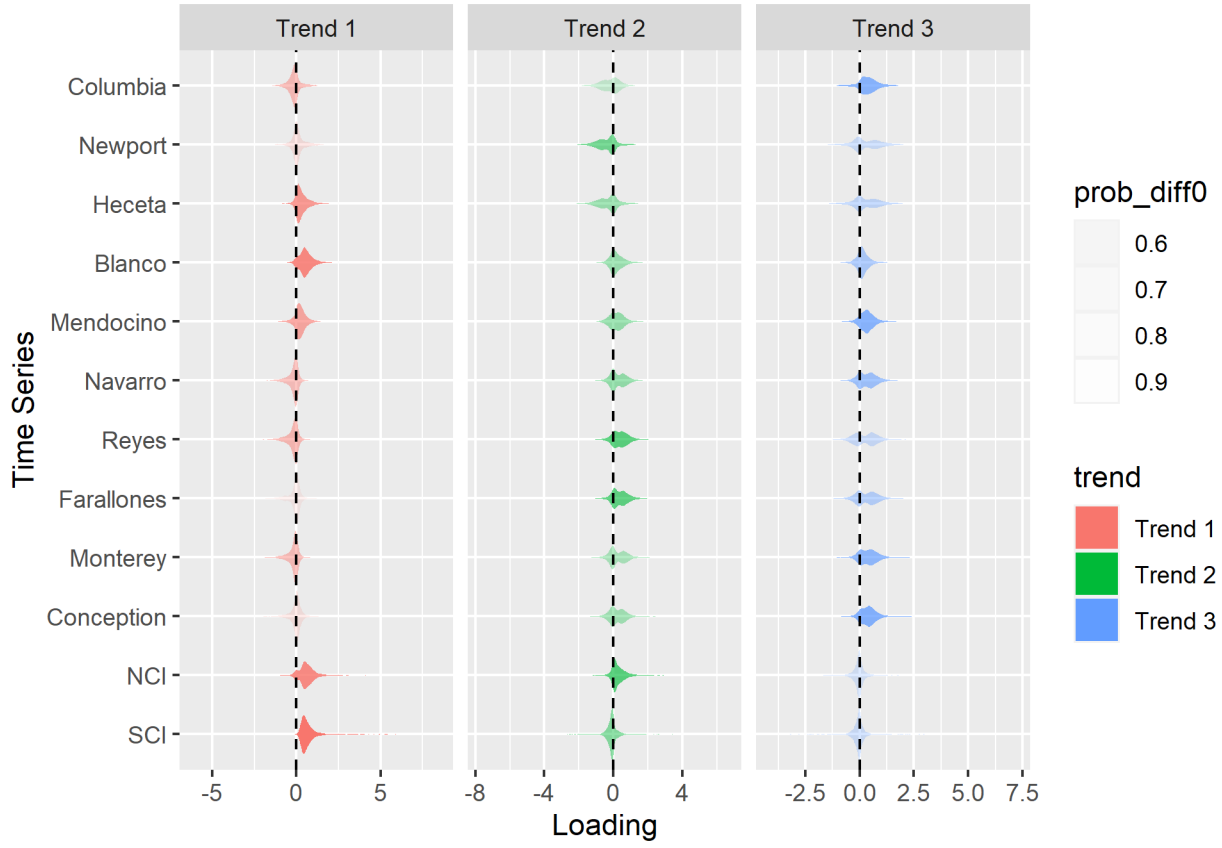


Figure 11: Loadings for the DFA. Regions are ordered north-south.

Table of loadings

```
##             [,1]  [,2]  [,3]
## Columbia   -0.13 -0.17  0.40
## Newport     0.09 -0.37  0.30
## Heceta      0.32 -0.27  0.30
## Blanco      0.56  0.19  0.15
## Mendocino   0.22  0.21  0.35
```

```
## Navarro    -0.17  0.27  0.36
## Reyes      -0.20  0.40  0.27
## Farallones -0.09  0.39  0.33
## Monterey   -0.18  0.25  0.41
## Conception -0.02  0.23  0.42
## NCI         0.62  0.29 -0.03
## SCI         0.65 -0.08 -0.04
```

Overall, the three trends follow generally similar patterns, although the timing of variation seems to differ. For example, most DFs increase in the latter portion of the times series around the years of the marine heatwave (REFERENCE) in northeast Paficic waters. However, the increse is seen most sharply and earliest in DF3.

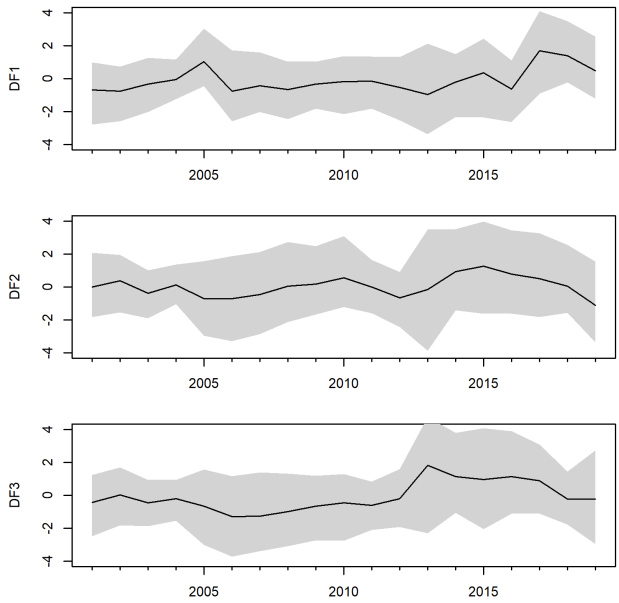


Figure 12: Time series of three dynamic factors from 2001-2019. Grey envelope indicates 95% CL.

Bioplots of the DFs give some indication of the system state in terms of recruitment. When plotting DF3 versus either of the other dynamic factorse, we can see that the years of the marine heat wave (approximately 2013-2017) were quite different from other years. Thus,the third DF seems to index changes in pre-recruit abundance related to the marine heatwave. Interestingly 2013 and 2017 (at the beginning and end of the marine heatwave), ordinate differently (DF1 vs DF3) from the core heat wave years (2014-2016) the non-heatwave years, and each other indicating transitional period to and from the anomalous marine heatwave conditions and more typical conditions. Thus, there apperas to have been an overall, coast-wide increase in pre-recruit abudance in 2013 (DF3), but mid-northern and southern regions responded differently than did the central regions during from 2013-2017 (DF1).

Note: I can't remember off-hand the specific blob years, but I think the overal interpretation holds. There were sort of two 'steps' in the state of the pre-recruit environement. Adjust the text to your liking. We also need general reference for the blob

Check for regimes

There appear to be two regimes in the data.

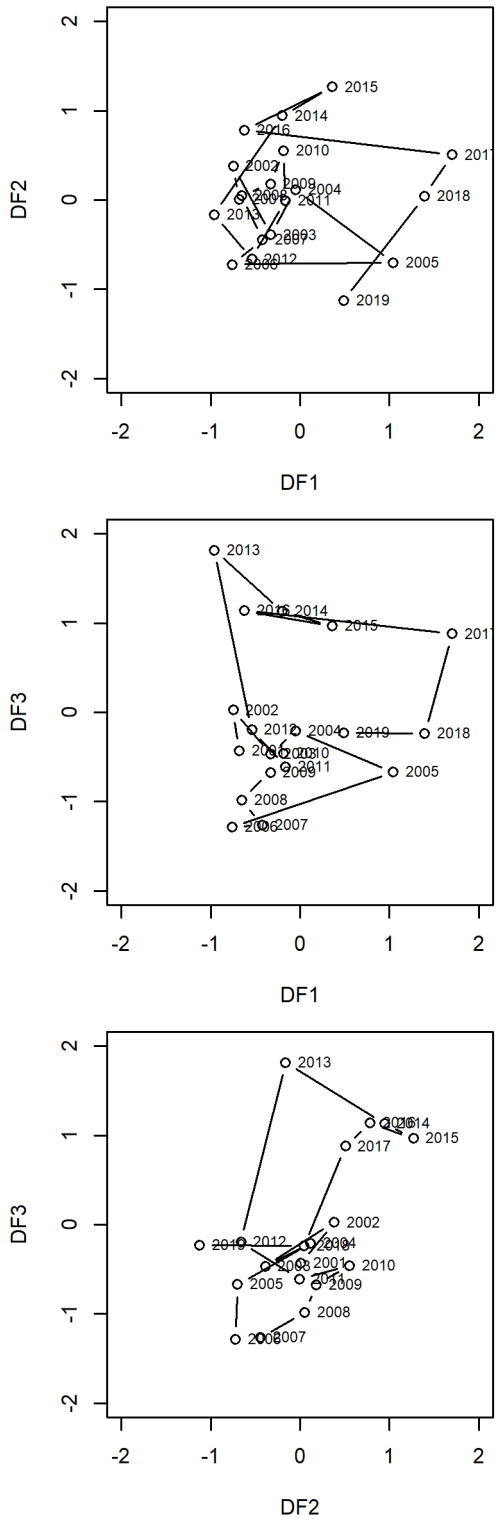


Figure 13: Biplots of DF1, DF2 and DF3 showing the state of the environement (in terms of pre-recruit abundance) through time.

References

Eric J. Ward, Sean C. Anderson and Luis A. Damiano (2019) bayesdfa: Bayesian Dynamic Factor Analysis (DFA) with 'Stan'. R package version 0.1.3. https://CRAN.R-project.org/package=bayesdfa

R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, 27, 1413-1432. doi: 10.1007/s11222-016-9696-4

Vehtari A, Gabry J, Magnusson M, Yao Y, Gelman A (2019) 'loo': Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.2.0.

Ward EJ, Scheuerell MD, Holmes EE 2018. 'atsar': Applied Time Series Analysis in R: An introduction to time series analysis for ecological and fisheries data with Stan. doi.org/10.5281/zenodo.1158021

Zuur AF, ID Tuck, Bailey N. 2003. Dynamic factor analysis to estimate common trends in fisheries time series. Canadian Journal of Fisheries and Aquatic Sciences 60:542-552.

Zuur AF, Fryer RJ, Jolliffe IT, Dekker R, Beukema JJ. 2003. Estimating common trends in multivariate time series using dynamic factor analysis. Environmetrics 14:665-685.