Smoothed factor analysis for multivariate time series

Eric J. Ward[1] and …

[1]Conservation Biology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Blvd E, Seattle WA, 98112, USA

Abstract

Introduction

Ecological data can be characterized by multiple sources of variability, including stochastic natural variation, and errors associated with data collection (observation, sampling, and measurement errors). Disentangling these sources of variability is often challenging, and necessitates the use of complex statistical methods, including state space models. Such approaches have become ubiquitous in ecology, particularly for time series data (Auger-Méthé et al. 2020) – in part because these models allow researchers to make inferences about ecological processes that aren't directly observable via corrupted observations. Applications of these models include estimating population change over time (Clark and Bjørnstad 2004), understanding movement dynamics (Patterson et al. 2008), and understanding spatiotemporal variation (Anderson and Ward 2019).

Estimating multiple sources of variation in state space models is numerically complex, and can be constrained explicitly or implicitly in ecological models via model assumptions. For example, discrete time state-space models of population trajectories generally assume latent population size can be approximated by an autoregressive process in log-space $x_{t+1} = f(x_t) + \epsilon_t$, where $x_t = log(n_t)$ and ϵ_t are normally distributed process deviations representing stochastic variability of the natural system (Dennis et al. 2006). The autoregressive assumption is critical here; without such a constraint, the variance of the stochastic noise ϵ_t is not estimable in the presence of an observation or data model. If model inference is not dependent on parameters of ecological interest (e.g. growth rates, density dependence), a wide range of alternative semi-parametric approaches exist that can be used to model the trajectory of x_t, including generalized additive

models (GAMs, (Wood 2011)) and Gaussian process models (Roberts et al. 2013). Because these models are not autoregressive with a constant time step, the 'wiggliness' of the model can be adjusted as part of the model fitting. In addition to their flexibility, these alternative models of x_t may be better suited for situations when data are patchily distributed in time or unequally spaced, making estimation of process and observation errors more difficult.

Challenges posed by univariate time series models also apply to multivariate time series models, though an additional complexity in the multivariate setting is that the number of latent time series may be variable, $k = 1, ..., m$, where m is the number of time series observed. At one extreme, $k = m$, and each time series can be thought of as corresponding to a unique latent state. Motivating questions for involved in analyzing these kinds of data include estimating correlated latent processes or trends, or estimating effects of environmental covariates (Hovel et al. 2016). At the other extreme, $k = 1$, where each time series can be thought of representing multiple measurements of the same trajectory of states, or trend, with optional offsets or coefficients included for each time series (offsets allowing for differing detectability). Applications focused on estimating a single trend from multivariate data include the development of ecological indicators. Models with intermediate numbers of latent states 1<k<m require mapping of time series to latent trends. These may be specified a priori (Ward et al. 2010) or estimated within the modeling framework using dimension reduction techniques.

Many statistical approaches have been proposed in recent years for clustering or estimating common signals in multivariate time series (Liao 2005). Examples include clustering based on similarities among time series features (Sardá-Espinosa 2019), identifying common patterns in the frequency domain (Holan and Ravishanker 2018), and clustering based on neural networks (Cherif et al. 2011). Application of these methods to ecological data has been limited, however, in part because many of these approaches identify clusters from raw data and don't explicitly account for observation error. An alternative approach that has been used in ecology to map collections of multivariate time series to latent states, while accounting for observation error, is dynamic factor analysis (DFA) (Zuur et al. 2003b, 2003a). DFA is an extension of factor analysis for time series data, and estimates a small number of common trends that can describe observed data. Mapping time series to trends is done via an estimated matrix of factor loadings – these allow each time series to be modeled as a mixture of the estimated latent trends, rather than assigning each time series to a single trend.

The objective of this analysis is to introduce a new class of DFA models for multivariate time series. Just as the univariate autoregressive model described above can be approximated with smooth functions, DFA models may be extended to use smooth functions in lieu of autoregressive processes. Recent work has

highlighted the application of hierarchical GAMs for multiple data sources (Pedersen et al. 2019). These approaches are flexible and likely to provide similar inference to DFA for single latent trend, however these methods have not been extended to include more than one process. We illustrate two options for modeling smooth functions for latent trends: basis splines ('b-splines') and Gaussian process models. Both approaches are compared to conventional autoregressive DFA models for two datasets on marine fishes from the west coast of the USA. All data and code are for replicating our analysis are on Github, and in our existing R package 'bayesdfa' (Ward et al. 2019).

Methods

Dynamic Factor Model

The basic DFA model can be written as a multivariate state space model, consisting of a latent process model and observation or data model. In its simplest form, the process model is expressed as a random walk, $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{w}_t$, where $\mathbf{w}_t \sim MVN(0, \mathbf{Q})$. For identifiability constraints (Zuur et al. 2003b, Holmes et al. 2012), the covariance matrix $\mathbf{Q}$ is generally constrained to be an identity matrix. Additional features may be incorporated into the process model, including autoregressive or moving average coefficients, covariates, or deviations that are more extreme than that of the normal distribution (Ward et al. 2019). The observation model in a DFA is expressed as a linear combination of trends $\mathbf{x}_t$ and matrix of loadings coefficients $\mathbf{Z}$, $\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{B}\mathbf{d}_t + \mathbf{e}_t$. In addition to the trends and loadings, time varying covariates $\mathbf{d}_t$ may be optionally included and linked to the observations through estimated coefficients $\mathbf{B}$. The vector $\mathbf{e}_t$ represents residual observation error. These typically are modeled as a diagonal matrix, $\mathbf{e}_t \sim MVN(0, \mathbf{R})$ but off-diagonal elements may be estimated (Holmes et al. 2020). Further details of the Bayesian implementation of the DFA model are provided in (Ward et al. 2019).

B-splines

The DFA model is extremely flexible, but a downside is that the latent process x_t

Gaussian predictive process

The discrete time Gaussian process model of trends treats the vector representing each trend $\mathbf{x}_k$ as being drawn from a multivariate normal distribution. As data in a DFA are generally standardized, we can assume the mean of each trend to be 0, and all inference about the Gaussian process occurs the covariance matrix,

$\mathbf{x}_k \sim MVN(0, \Sigma)$. Rather than estimate each element of Σ independently, smooth covariance functions or 'kernels' are chosen to represent the covariance between points in time (typical choices include the exponential, Gaussian, and Matern functions). For the purpose of our DFA modeling, we adopt a Gaussian kernel. For a hypothetical DFA model with more than one trend, the covariance between points i and j at times t_i and t_j on trend k can be expressed as $cov(x_{i,k}, x_{j,k}) = \sigma_k^2 exp\left(\frac{-(t_i - t_j)^2}{2\theta_k^2}\right)$, where σ_k controls the magnitude of variation, and θ_k controls how smoothly correlation decreases as time points become further apart. We allow each trend to have its own covariance parameters, allowing each to have differing degrees of smoothness. Because of potential computation issues in high dimensionality problems, we also allow this Gaussian process model to be expressed as a Gaussian predictive process model. The difference between the predictive process approach and the full Gaussian process model is that instead of modeling the $\mathbf{x}_t$ themselves as random variables, random variables are modeled at a subset of locations $\mathbf{x}^*_k$ and projected to the locations of the data $\mathbf{x}_k$. If we assume $\mathbf{x}^*_k \sim MVN(0, \Sigma^*)$, then this projection can be done as $x_k = \Sigma'_{k,k^*}\Sigma^{*-1}x^*_k$, where the matrix Σ'_{k,k^*} is the transpose of the matrix describing the covariance between x_k and x^*_k. The location of k^* can be spaced equally or depend on data; for the purposes of our DFA modeling, we assume that the k^* are equally spaced within each time series (with the endpoints also acting as knots).

Application: larval fish survey

As a first application of smooth factor analysis models, we apply DFA to a long term time series of larval fishes collected off California. The California Cooperative Oceanic Fisheries Investigations (CalCOFI) survey has been collecting physical and biological samples since 1949, to monitor annual, seasonal, and spatial changes to the California Current Ecosystem (Bograd et al. 2003). The CalCOFI data has been incorporated into models used to assess population status (MacCall 2003), and numerous publications have used time series of larval fishes from the CalCOFI survey as indicators of ecosystem state (Mcclatchie et al. 2008). These types of motivating questions also present an opportunity to apply DFA with both conventional and smoothed trends to summarize ecosystem state indices. For this application, we focus on the dynamics of three species of juvenile rockfishes: Aurora rockfish (*Sebastes aurora*), Shortbelly rockfish (*S. jordani*), and Bocaccio rockfish (*S. paucispinis*). For the purposes of this analysis, we restrict the time series to data collected since 1985, when sampling has been consistent in space and time (Moser et al. 2001). Though CalCOFI cruises are done throughout the year, we are primarily interested in estimating interannual trends, and thus restrict our analysis to considering spring cruises from 1-April to 22-May when densities of most rockfish species are highest (Mosek et al. 2000). All data were retrieved using R software (R Core Team 2020) and the rerddap package (Chamberlain 2020).

With only three time series, we focus on identifying models estimating a single shared trend. Other types of models, including hierarchical GAMs (Pedersen et al. 2019) or models allowing estimated offsets may also be useful in this type of application. Where the DFA model differs is that unlike models with random intercepts or additive offset terms, the DFA factor loadings $\mathbf{Z}$ are multiplicative and may be close to zero. These cases may arise when a particular time series has a low signal to noise ratio, or if there is low correspondence with the latent trends estimated among all other time series. In addition to estimating a conventional 1-trend DFA model with a latent autoregressive process, we evaluate 1-trend b-spline and GP models. Because we have no a priori hypotheses about the complexity of these smoothed factor models, we evaluated 5 models for each, using 6, 12, 18, 24, and 30 equally spaced knots.

Application: commercial fisheries landings

As a slightly more complex example of the smooth factor analysis model, we examine the performance of 2-trend models, using a dataset of commercial fisheries catches from the west coast of the USA. This dataset consists of commercial landings by dominant species, and is reported annually to the Pacific Fishery Management Council (cite SAFE tables). This dataset consists of 13 species or groups reported over a 39 year period (1981 – 2019). Landings on the west coast are dominated by Pacific hake (also Pacific whiting, Merluccius productus), but also include substantial catches of rockfishes (*Sebastes spp.*) and flatfishes (e.g. Dover sole, *Solea solea*). Over the course of the last 4 decades, these species have experienced variability associated with population dynamics and the environment, but the patterns of landings also reflects a dynamic fisheries management process. Examples of changes include temporarily closing areas to fishing to protect species of conservation concern, and implementing catch share programs. These processes, combined with environmental conditions that have been positive for many species have resulted in many increasing populations (Warlick et al. 2018). Given these various management and ecological changes, it is important to summarize patterns of landings, and identify common trends as indicators for management and ecosystem status (Harvey et al. 2018).

As with our previous example, we compared conventional DFA models to those modeling the trends with smooth functions. Preliminary model fitting suggested that 2-trend models were more supported by the data, and thus will be the focus of our analysis. In addition to modeling the 2-trend model with conventional DFA, we evaluated B-spline and Gaussian process models with 6 to 30 equally spaced knots.

Estimation and model selection

We developed our DFA model in a Bayesian framework, using Stan and the package rstan (Stan Development Team 2016), which implements Markov chain Monte Carlo (MCMC) using the No-U Turn Sampling (NUTS) algorithm (Hoffman and Gelman 2014, Carpenter et al. 2017). For each model considered, we ran 3 parallel MCMC chains for 3000 iterations each, discarding the first 50% of the samples. Convergence diagnostics were assessed using summary statistics (R-hat, (Gelman and Rubin 1992)). Previous approaches have used the Leave One Out Information Criterion (LOOIC) as a model selection tool (Vehtari et al. 2017, 2020). Preliminary model checks using LOOIC for the models included in our analysis indicated that many models had 1-4 data points that had high Pareto-k statistics (possibly because of model-misspecification or model flexibity; Vehtari et al. (2017)). As a slower but potentially more robust model selection approach, we implemented k-fold cross validation. There are many possible ways to assign 'folds' in cross-validation, and because of our focus on the temporal aspect of these DFA models, we assigned each year of data to a unique fold. Implementing these models in a Bayesian framework requires significant computational time; instead of fitting each model to a dataset once, a single model is re-fit to a dataset for as many years as there are data (35 years for our application to CalCOFI data; 39 years for our application to commercial landings data).

Results

In our comparison of 1-trend DFA models applied to the three timeseries of juvenile rockfishes, we found that models with smooth trends were better supported over conventional random walk models. In calculating LOOIC, all models had a few (1-3) data points with Pareto-k statistic values between 0.7 - 1, likely because of the flexibility of these models. Across models considered, we found that the B-spline trend model with 24 knots had slightly lower LOOIC values than alternatives, and appeared to capture the dynamics of the three *Sebastes* time series @ref(fig:fig1). All three species were estimated to have positive loadings on the estimated trend, with shortbelly rockfishes had the highest loading (0.99, 95% posterior interval = 0.35-2.38) followed by aurora (0.82, 95% posterior interval = 0.24-2.06) and bocaccio rokfishes (0.65, 95% posterior interval = 0.12-1.72).

	Trend.model	Knots	ELPD	SE
8	Gaussian process	12	62.87196	36.66923
10	Gaussian process	24	57.07978	41.76271
11	Gaussian process	30	56.99396	42.13878

	Trend.model	Knots	ELPD	SE
4	B-spline	18	56.94945	43.33349
5	B-spline	24	55.88544	42.09490

	Trend.model	Knots	ELPD	SE
1	Random walk	NA	-12686.72	1961.7972
2	B-spline	6	-25250.76	294.5618
5	B-spline	24	-25375.16	315.0257
6	B-spline	30	-25548.36	286.1946
4	B-spline	18	-25556.58	305.8288

Acknowledgments

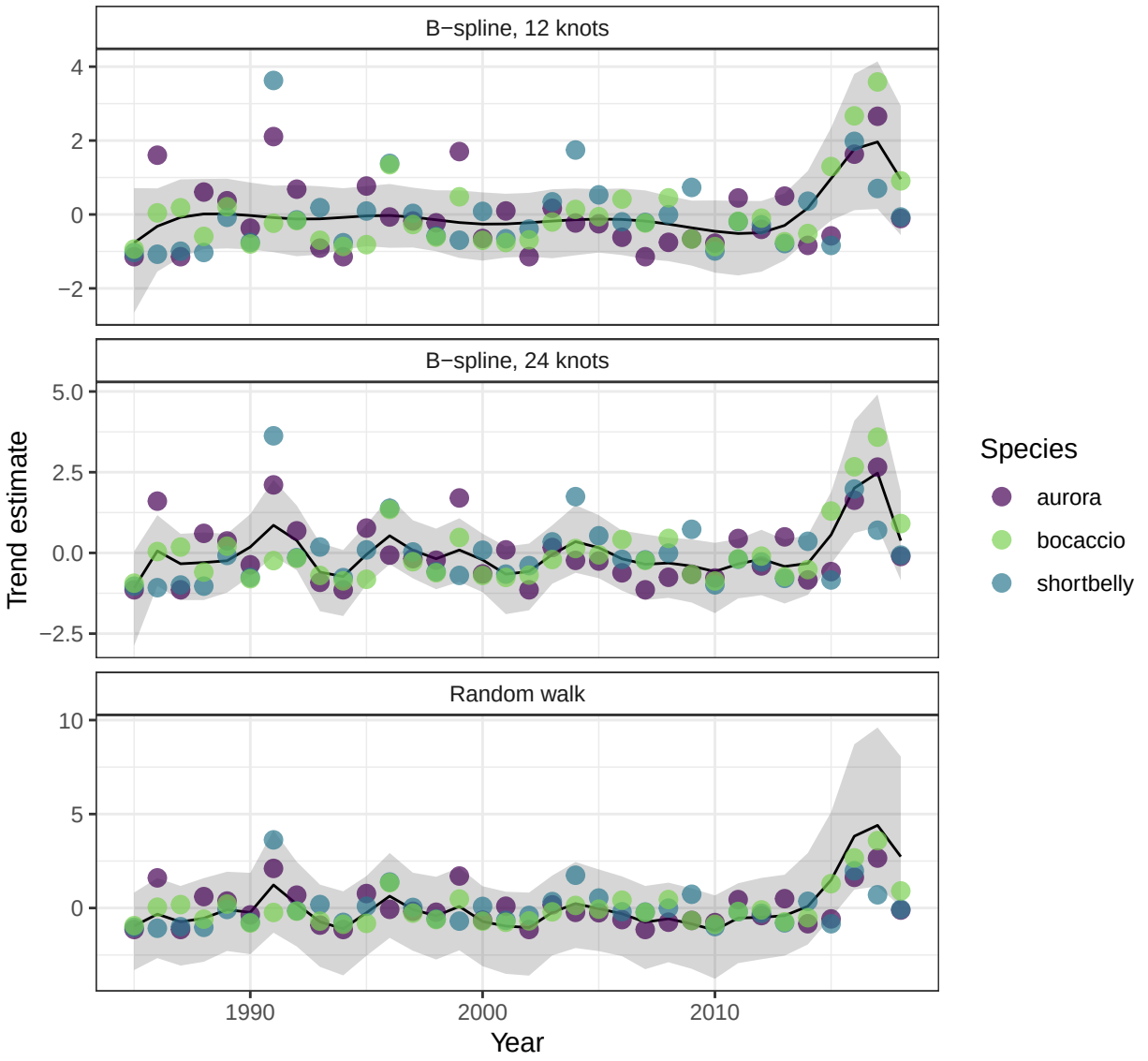

Figure 1: Standardized densities of juvenile rockfish species collected in the CalCOFI survey, and estimates of latent trends for three candidate models. A B-spline model with 24 knots had slighty lower LOOIC values than other B-spline models (such as the more coarse 12 knot model) and the conventional DFA, modeled with a random walk. The posterior mean from each model is shown as a solid black line, and 95% credible intervals are shown in the grey region.

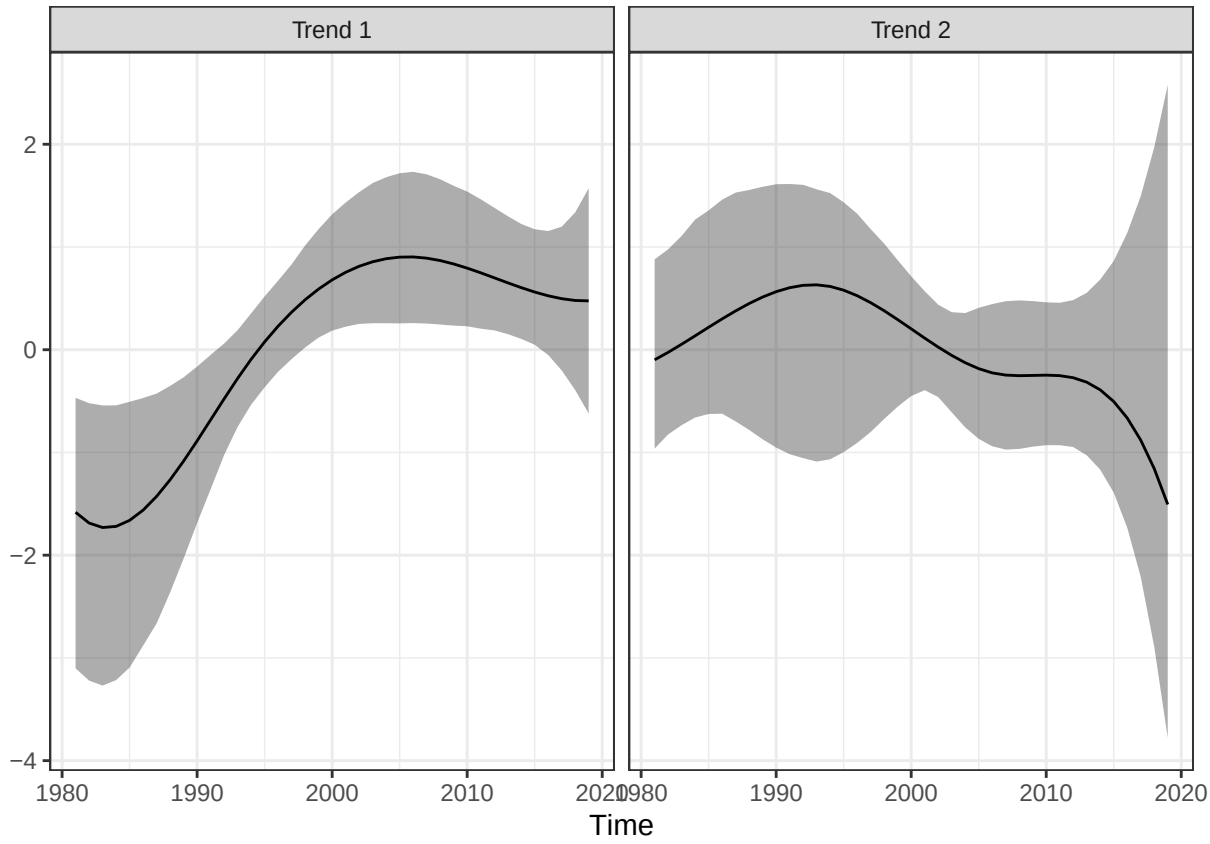

Figure 2: Estimated trends from the 2-trend DFA model applied to commercial groundfish landings off the west coast of the United States. The model results with lowest LOOIC is shown, a model that allows trends to be approximated with B-spines (6 knots). The posterior mean for each trend is shown, with ribbons representing 95% credible intervals.

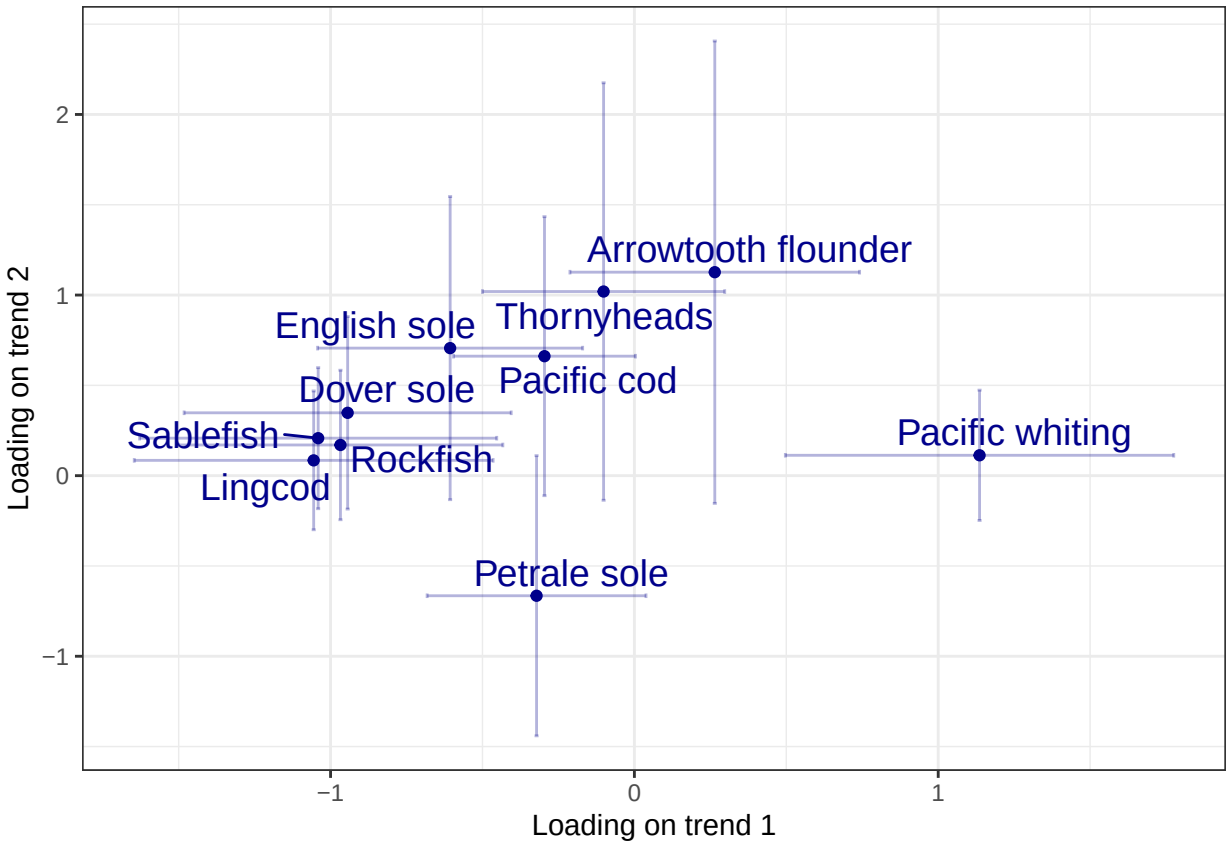

Figure 3: Estimated loadings for each species or group from a 2-trend DFA model with latent trends modeled as B-splines. The posterior mean for each species is shown as a point, with lines representing 95% credible intervals.

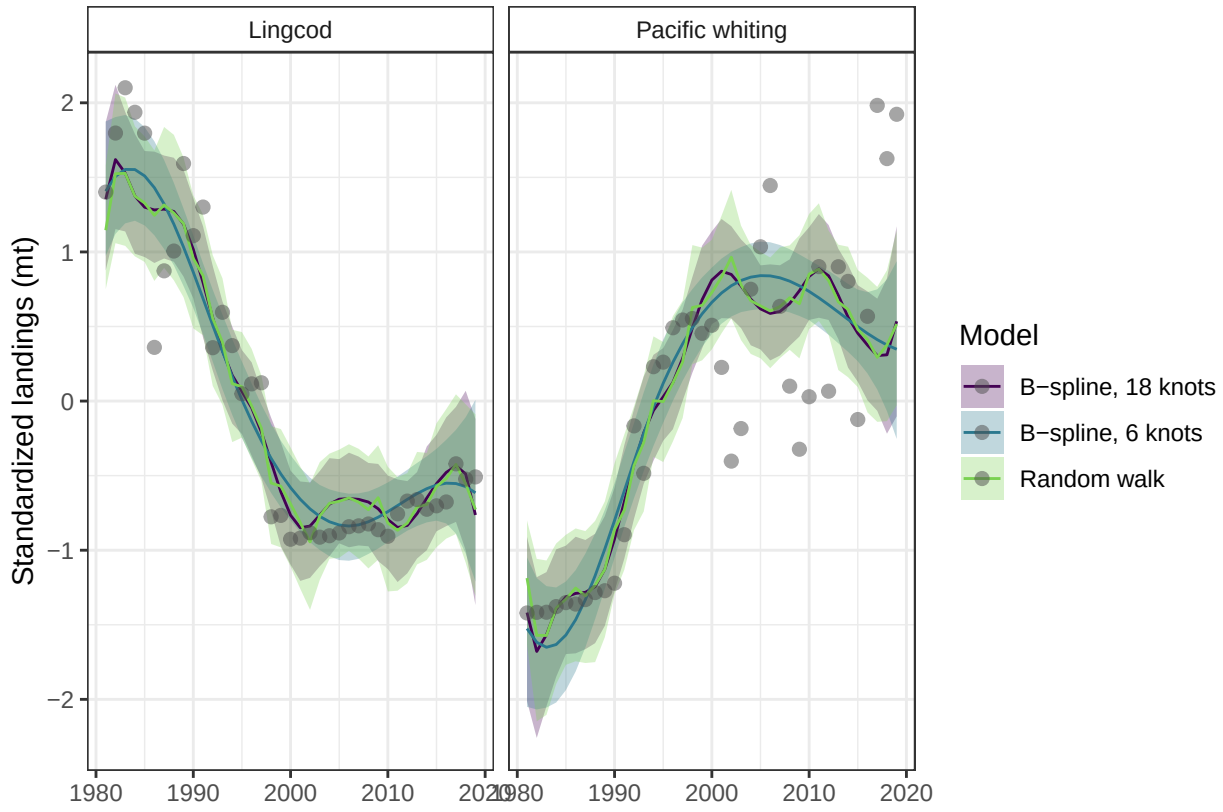

Figure 4: Estimated landings for 2 species included in our analysis, with contrasting trends (lingcod, Pacific whiting). Posterior means and 95% credible intervals (ribbons) for three candidate models are shown: b-spline trend models with 6 and 18 knots, respectively, and a random walk model representing the conventional DFA model.

References

Anderson, S. C., and E. J. Ward. 2019. Black swans in space: Modeling spatiotemporal processes with extremes. Ecology 100:e02403.

Auger-Méthé, M., K. Newman, D. Cole, F. Empacher, R. Gryba, A. A. King, V. Leos-Barajas, J. M. Flemming, A. Nielsen, G. Petris, and L. Thomas. 2020. A guide to state-space modeling of ecological time series. arXiv:2002.02001 [q-bio, stat].

Bograd, S. J., D. A. Checkley, and W. S. Wooster. 2003. CalCOFI: A half century of physical, chemical, and biological research in the California Current System. Deep Sea Research Part II: Topical Studies in Oceanography 50:2349–2353.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. 2017. *Stan* : A Probabilistic Programming Language. Journal of Statistical Software 76.

Chamberlain, S. 2020. Rerddap: General purpose client for 'ERDDAP' servers. manual.

Cherif, A., H. Cardot, and R. Boné. 2011. SOM time series clustering and prediction with recurrent neural networks. Neurocomputing 74:1936–1944.

Clark, J. S., and O. N. Bjørnstad. 2004. Population Time Series: Process Variability, Observation Errors, Missing Values, Lags, and Hidden States. Ecology 85:3140–3150.

Dennis, B., J. M. Ponciano, S. R. Lele, M. L. Taper, and D. F. Staples. 2006. Estimating Density Dependence, Process Noise, and Observation Error. Ecological Monographs 76:323–341.

Gelman, A., and D. B. Rubin. 1992. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science 7:457–472.

Harvey, C., N. Garfield, G. Williams, N. Tolimieri, I. Schroeder, E. Hazen, K. Andrews, K. Barnas, S. Bograd, R. Brodeur, B. Burke, J. Cope, L. deWitt, J. Field, J. Fisher, T. Good, C. Greene, D. Holland, M. Hunsicker, and S. Zador. 2018. Ecosystem Status Report of the California Current for 2018: A Summary of Ecosystem Indicators Compiled by the California Current Integrated Ecosystem Assessment Team (CCIEA).

Hoffman, M. D., and A. Gelman. 2014. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research 15:1593–1623.

Holan, S. H., and N. Ravishanker. 2018. Time series clustering and classification via frequency domain

methods. WIREs Computational Statistics 10:e1444.

Holmes, E. E., E. J. Ward, and M. D. Scheuerell. 2020. Analysis of multivariate time-series using the MARSS package. NOAA Fisheries, Northwest Fisheries Science Center, 2725 Montlake Blvd E., Seattle, WA 98112.

Holmes, E. E., E. J. Ward, and K. Wills. 2012. MARSS: Multivariate autoregressive state-space models for analyzing time-series data. R Journal 4:11–19.

Hovel, R., S. Carlson, and T. Quinn. 2016. Climate change alters the reproductive phenology and investment of a lacustrine fish, the three-spine stickleback. Global Change Biology 23.

Liao, T. W. 2005. Clustering of time series data—a survey. Pattern Recognition 38:1857–1874.

MacCall, A. D. 2003. Status of Bocaccio off California in 2003. In Appendix to the status of the Pacific coast groundfish fishery through 2003: Stock assessment and fishery evaluation. Pacific Fishery Management Council, Portland, OR.

Mcclatchie, S., R. Goericke, J. Koslow, F. Schwing, S. Bograd, R. Charter, W. Watson, N. Lo, K. Hill, J. Gottschalck, M. L'Heureux, Y. Xue, W. Peterson, R. Emmett, C. Collins, G. Gaxiola-Castro, R. Durazo, M. Kahru, B. Mitchell, and E. Bjorkstedt. 2008. The state of the California Current, 2007-2008: La Niña conditions and their effects on the ecosystem. California Cooperative Oceanic Fisheries Investigations Reports 49:39–76.

Mosek, H., L. Charter, W. Watson, I. Ambrose, N. Shakon, K. Charter, E. Saniiknoi, S. Fisheiies, S. Center, M. Fi, and H. Service. 2000. Abundance and distribution of rockfish (Sebastes) larvae in the Southern California Bight in relation to environmental conditions and fishery exploitation. ABUNDANCE AND DISTRIBUTION OF ROCKFISH LARVAE CalCOFl Rep 41.

Moser, H. G., R. L. Charter, W. Watson, A. Amurose, P. E. Smith, E. M. Sani, and S. R. Charter. 2001. The CalCOFI Ichthyoplankton time series: Potential contributions to the management of rocky-shore fishes 42:17.

Patterson, T. A., L. Thomas, C. Wilcox, O. Ovaskainen, and J. Matthiopoulos. 2008. State–space models of individual animal movement. Trends in Ecology & Evolution 23:87–94.

Pedersen, E. J., D. L. Miller, G. L. Simpson, and N. Ross. 2019. Hierarchical generalized additive models in ecology: An introduction with mgcv. PeerJ 7:e6876.

R Core Team. 2020. R: A language and environment for statistical computing. manual, Vienna,

Austria.

Roberts, S., M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain. 2013. Gaussian processes for time-series modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371:20110550.

Sardá-Espinosa, A. 2019. Time-Series Clustering in R Using the dtwclust Package. The R Journal 11:22–43.

Stan Development Team. 2016. RStan: The R interface to Stan.

Vehtari, A., J. Gabry, M. Magnusson, Y. Yao, P.-C. Bürkner, T. Paananen, and A. Gelman. 2020. Loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.

Vehtari, A., A. Gelman, and J. Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 27:1413–1432.

Ward, E., J., S. Anderson C., L. Damiano A., M. Hunsicker E., and M. Litzow A. 2019. Modeling regimes with extremes: The bayesdfa package for identifying and forecasting common trends and anomalies in multivariate time-series data. The R Journal 11:46.

Ward, E. J., H. Chirakkal, M. González-Suárez, D. Aurioles-Gamboa, E. E. Holmes, and L. Gerber. 2010. Inferring spatial structure from time-series data: Using multivariate state-space models to detect metapopulation structure of California sea lions in the Gulf of California, Mexico. Journal of Applied Ecology 47:47–56.

Warlick, A., E. Steiner, and M. Guldin. 2018. History of the West Coast groundfish trawl fishery Tracking socioeconomic characteristics across different management policies in a multispecies fishery. Marine Policy 93:9–21.

Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73:3–36.

Zuur, A. F., R. J. Fryer, I. T. Jolliffe, R. Dekker, and J. J. Beukema. 2003a. Estimating common trends in multivariate time series using dynamic factor analysis. Environmetrics 14:665–685.

Zuur, A. F., I. D. Tuck, and N. Bailey. 2003b. Dynamic factor analysis to estimate common trends in fisheries time series. Canadian Journal of Fisheries and Aquatic Sciences 60:542–552.