SEDAR-17-AW-06

Methods for combining multiple indices into one, with application to south Atlantic (U.S.) Spanish mackerel

August 23, 2008

Prepared by Paul B. Conn
National Marine Fisheries Service
NOAA Center for Coastal Fisheries & Habitat Research
Beaufort, NC

1 Introduction

Fisheries analysts often want to fit population models to several indices. If the selectivity associated with the gears for each index differ markedly, this may be a reasonable thing to do, especially if the age or length composition of the fishery changes over time. However, this is often difficult or impossible with simpler assessment methods, as with surplus production models (cf. Schaefer, 1954, 1957; Prager, 1994) or stock reduction analysis (Walters et al., 2006). In these cases, it would be useful to perform inference with one common index since the numerical routines used to fit these models and are often sensitive to the quality (and quantity) of datasets available.

For Spanish mackerel on the Atlantic coast of the U.S., a total of nine different indices of abundance had been computed, few of which correlated well with each other (SEDAR, 2008a). Since population indices are supposed to reflect relative abundance, we anticipated problems would arise because each index told a different "story" with regard to whether the population had been increasing or decreasing over the last 22 years.

In this working paper, we introduce a novel method for combining multiple indices into one common index. The method works by assuming that observed relative abundance trends are sampled from a common population trend, but subject to process and sampling error. Inference then focuses on describing trends in the mean index. In particular, we give details on data manipulations needed to perform the analysis on Spanish mackerel in the south Atlantic (U.S.), and provide examples of how the models can be fitted to Spanish mackerel indices.

2 Methods for combining indices

I assume that the investigator has access to I different indices, denoted $\mathbf{U_i} = \{U_{it}\}$, where the i subscript gives the index, and t is a time subscript ($t \in 1, \ldots, T$). Owing to differences in scale between indices, inference focuses on *changes* in relative abundance each year, $\Delta_{it} = U_{i,t+1}/U_{it}$, which are related to the finite rate of population increase (λ) in population ecology. Observed changes in relative abundance for each index $\boldsymbol{\Delta}_i$ are assumed to be sampled from a common underlying population trend, which we denote $\boldsymbol{\lambda}$, where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_{T-1}]$. Observed changes in each index are subject to both process (σ_i^p) and sampling ($\boldsymbol{\sigma_i^s} = [\sigma_{i1}^s, \sigma_{i2}^s, \ldots, \sigma_{i,T-1}^s]$) errors. The latter are easily obtained as long as sampling errors are calculated and reported together with each of the indices (for example, as a result of design-based sampling or a byproduct of model fitting). For instance, assuming that indices are constructed in such a manner that sampling covariance is negligible, a first order Taylor series approximation gives

$$V\hat{a}r(U_{i,t+1}/U_{it}) = \frac{V\hat{a}r(U_{i,t+1})}{U_{it}^2} + \frac{U_{i,t+1}^2}{U_{it}^4} V\hat{a}r(U_{it})$$

(Casella and Berger, 1990). In contrast, process errors are impossible to detect given one index alone. However, they should often be expected in indices due to annual changes in catchability (for example, because of differences in spatial coverage, timing of migration, or due to age or length-based gear selectivity).

To incorporate process and sampling error, I assume that $\boldsymbol{\Delta}_i$ and $\boldsymbol{\lambda}$ are connected through a sequence of latent population changes $\mathbf{L_i} = [L_{i1}, L_{i2}, \ldots, L_{i,T-1}]$, where the latent time series is subject to process error (but not sampling error). I specify a hierarchical model for the observed data as follows:

$$[\Delta_{it}|\lambda_t, \sigma_i^p, \sigma_{it}^s] = [\Delta_{it}|L_{it}, \sigma_{it}^s][L_{it}|\sigma_i^p, \lambda_t][\lambda_t][\sigma_i^p].$$

Here the notation $[X|Y]$ specifies the conditional distribution of X given Y. Note that I have implicitly assumed prior independence between the population trend and process errors by assuming that $[\lambda_t, \sigma_i^p] = [\lambda_t][\sigma_i^p]$. Note also that the level of process error is dependent upon the index only (i.e., does not change over time). I also assume the investigator has access to index sampling errors, as output from index construction (through fitting with delta GLM's for example; cf., Lo et al. 1992, Maunder and Punt 2004). Writing a model for the observed indices in this manner is useful because it facilitates calculation of posterior distributions for λ_t and σ_i^p via Bayesian analysis. Posterior prediction may also be used to generate densities for μ_t, which can be interpreted as a "combined" index. This can be done by setting the first value of the index to a constant (e.g., $\mu_1 = 1$), and then applying estimated values of λ_t, so that $\hat{\boldsymbol{\mu}} = [1, \lambda_1, \lambda_1\lambda_2, \ldots, \prod_t \lambda_t]$.

Setting up a Bayesian analysis requires that one specify probability distributions for all likelihood components, as well as prior distributions for $[\lambda_t]$ and $[\sigma_i^p]$. For purposes of this paper, I suggest that making likelihood components normally distributed is a reasonable choice. As it is easier to calculate Bayesian quantities when the model is specified in terms of precision (τ) than standard deviation (σ), I made the substitutions $\tau_{it}^s = 1/(\sigma_{it}^s)^2$

and $\tau_i^p = 1/(\sigma_i^p)^2$), and then specified following probability models:

$$[\Delta_{it}|L_{it}, \tau_{it}^s] : \quad \text{Normal}(\Delta_{it}; L_{it}, 1/\tau_{it}^s)$$

$$[L_{it}|\lambda_t, \tau_i^p] : \quad \text{Normal}(L_{it}; \lambda_t, 1/\tau_t^p).$$

I also specified the priors

$$[\sigma_i^p] : \quad \text{Gamma}(\sigma_i^p; \alpha, \beta), \text{and}$$

$$[\lambda_t] : \quad \text{Uniform}(a, b).$$

The gamma distribution is a logical choice in this case because it is the conjugate prior for the precision parameter under a normal likelihood; thus when sampling the posterior via Markov chain Monte Carlo (MCMC), this parameter may be simulated directly without an accept/reject step. The choice of a uniform prior on the mean index is somewhat arbitrary.

3 Spanish mackerel analysis

We gathered Spanish mackerel indices from the SEDAR 17 Data workshop report (SEDAR, 2008a) for hierarchical analysis. We performed analysis on seven of the nine indices, not electing to use SEAMAP indices because these reflected the relative abundance of new recruits rather than population-level relative abundance. Using the model specified above, we used a hybrid MCMC sampler (see, e.g., Gelman et al., 2004) to approximate the posterior distribution of population growth rates, latent gradients, and process errors. Population growth rates and latent gradients were sampled individually using Metropolis-Hastings steps with normally distributed proposals with mean zero and a standard deviation chosen to give acceptance rates in the 30-40%

range (Gelman et al., 2004). Conditional on the other parameters, process precision parameters (τ_i^p) were available analytically and were simulated as

$$\tau_i^p \sim \text{Gamma}\left(0.5 \sum_t I_{it} + \alpha,\ \ 0.5 \sum_t I_{it}(L_{it} - \lambda_t)^2 + \beta\right).$$

Here, the gamma distribution is referenced with a rate parameter rather than a scale parameter, and I_{it} is an indicator variable that takes on the value 1.0 if λ_{it} is defined in year t and zero otherwise. The values $\alpha = 0.01$ and $\beta = 0.01$ were chosen to give a diffuse distribution for the prior precision. Similarly, a diffuse Uniform(0.1,10.0) was chosen for population growth rates (λ_t). This prior assumes that the population had zero probability of declining by more than 90% or increasing it's size tenfold in any given year. This seemed reasonable given our knowledge of Spanish mackerel life history.

The MCMC analysis was coded in R (R Development Core Team, 2007), and was run for 110,000 iterations with the the first 10,000 iterations thrown out as a burn-in (Gelman et al., 2004). Standard MCMC diagnostics (Figure 1) suggested that the chain mixed well and had indeed converged to the posterior distribution. Examination of posterior distributions for process error (Figure 2) indicated that the best estimate of process error for most indices was in the 0.1-0.3 range; an exception was the handline index compiled from logbook entries; this index included enough noise that it was virtually useless. Posterior estimates of λ indicated that the population increased more than it decreased over the course of the study (Figure 3). The "overall" index derived from λ (Figure 4, Table 1) confirmed this visually, as relative abundance was predicted to increase over the course of the study. In general, estimates of population level trends followed trends in indices with lower estimated levels of process error. Also available for comparison were latent, index-specific estimates of λ, which included process error with sampling error removed (Figure 5). The former are analogous to shrinkage estimators in random effects models.

References

Casella, G., and R. L. Berger. 1990. Statistical Inference. Duxbury Press, Belmont, CA.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian Data Analysis, 2nd Edition. Chapman and Hall, Boca Ration.

Lo, N. C., L. D. Jacobson, and J. L. Squire. 1992. Indices of relative abundance from fish spotter data based on delta-lognormal models. Canadian Journal of Fisheries and Aquatic Sciences **49**:2515–2526.

Maunder, M. N., and A. E. Punt. 2004. Standardizing catch and effort data: a review of recent approaches. Fisheries Research **70**:141–159.

Prager, M. H. 1994. A suite of extensions to a nonequilibrium surplus-production model. Fisheries Bulletin **92**:374–389.

R Development Core Team, 2007. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

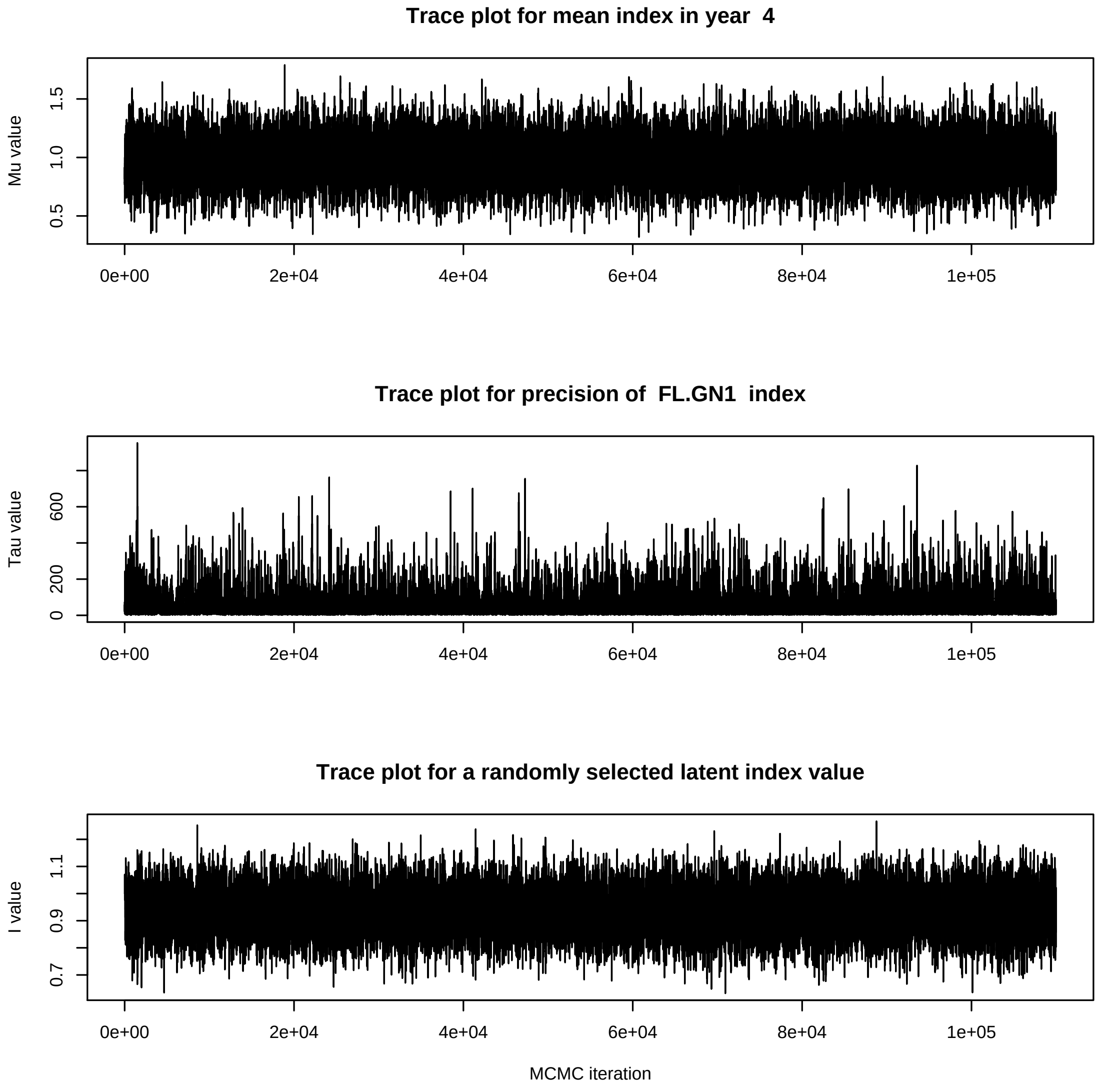
Schaefer, M. B. 1954. Some aspects of the dynamics of populations important to the management of the commercial marine fisheries. Bulletin of the Inter-American Tropical Tuna Commission **1**:27–56.

Schaefer, M. B. 1957. A study of the dynamics of the fishery for yellowfin tuna in the eastern tropical Pacific Ocean. Bulletin of the Inter-American Tropical Tuna Commission **2**:247–268.

SEDAR, 2008a. SEDAR 17 Data Workshop Report.

Walters, C. J., S. J. D. Martell, and J. Korman. 2006. A stochastic approach to stock reduction analysis. Canadian Journal of Fisheries and Aquatic Sciences **63**:212–223.

Figure 1: Trace plots of three randomly selected parameters from the MCMC analysis. The plot indicates good mixing and no indication of alternative minima or maxima.


Trace plot for mean index in year 4

Trace plot for precision of FL.GN1 index

Trace plot for a randomly selected latent index value

Figure 2: Estimated posterior distributions for the process errors, σ_i^p. Shown are values for CPUE indices derived from different gears, locations, and times: Florida gillnet preceding net ban (FL.GN1), Florida gillnet after the net ban (FL.GN2), Florida castnet (FL.CN), Florida handlines (FL.HL), the Marine Recreational Fisheries Statistics Survey (MRFSS), logbook records from gillnet fisheries north of Florida (LB.GN), and logbook records from handline fisheries north of Florida (LB.HL). Large process errors indicate that what is measured by a given index is not necessarily indicative of what is going on at the population level.

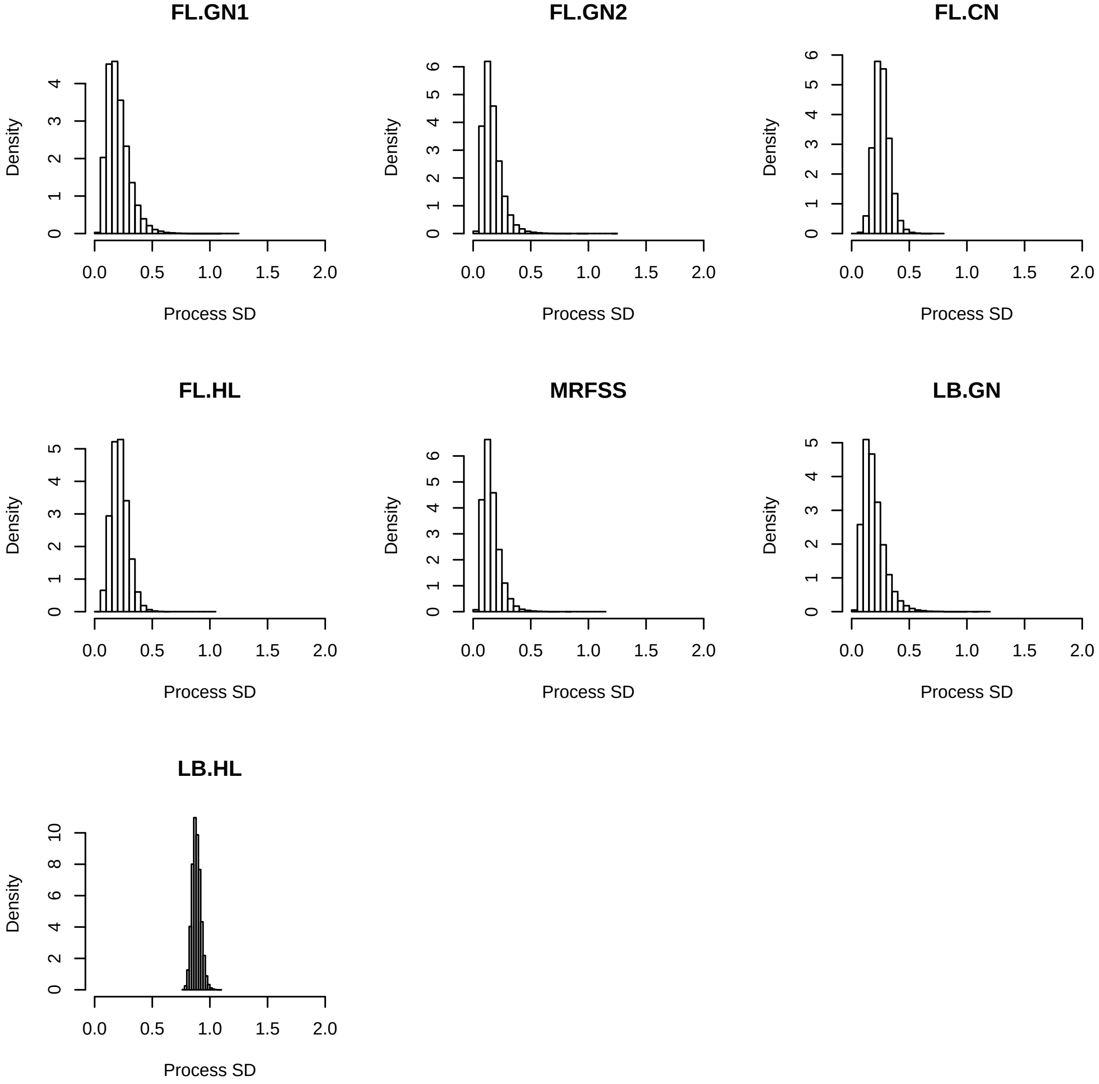


Figure 3: Estimates (posterior means) for the rate of population change, 1985-2006.

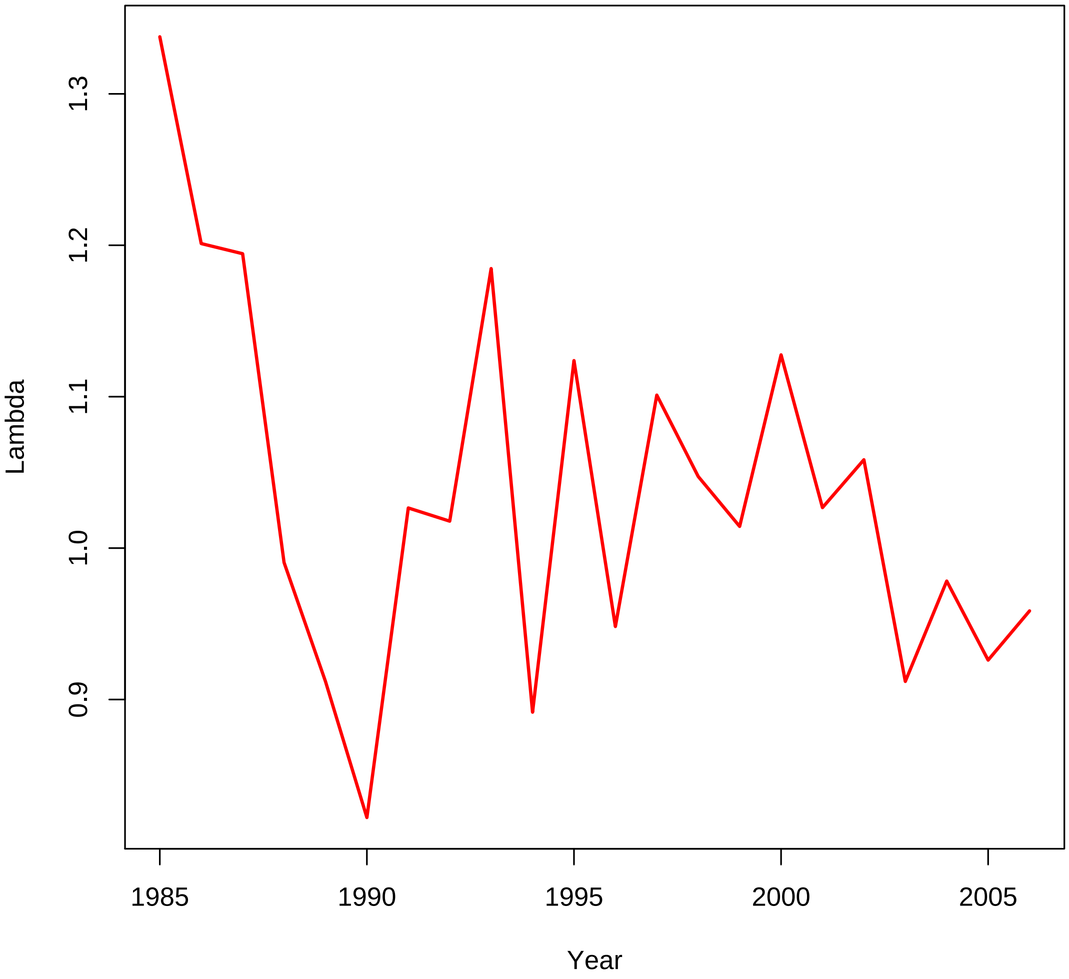


Figure 4: Estimates of a single combined relative abundance index for Spanish mackerel based on posterior prediction.

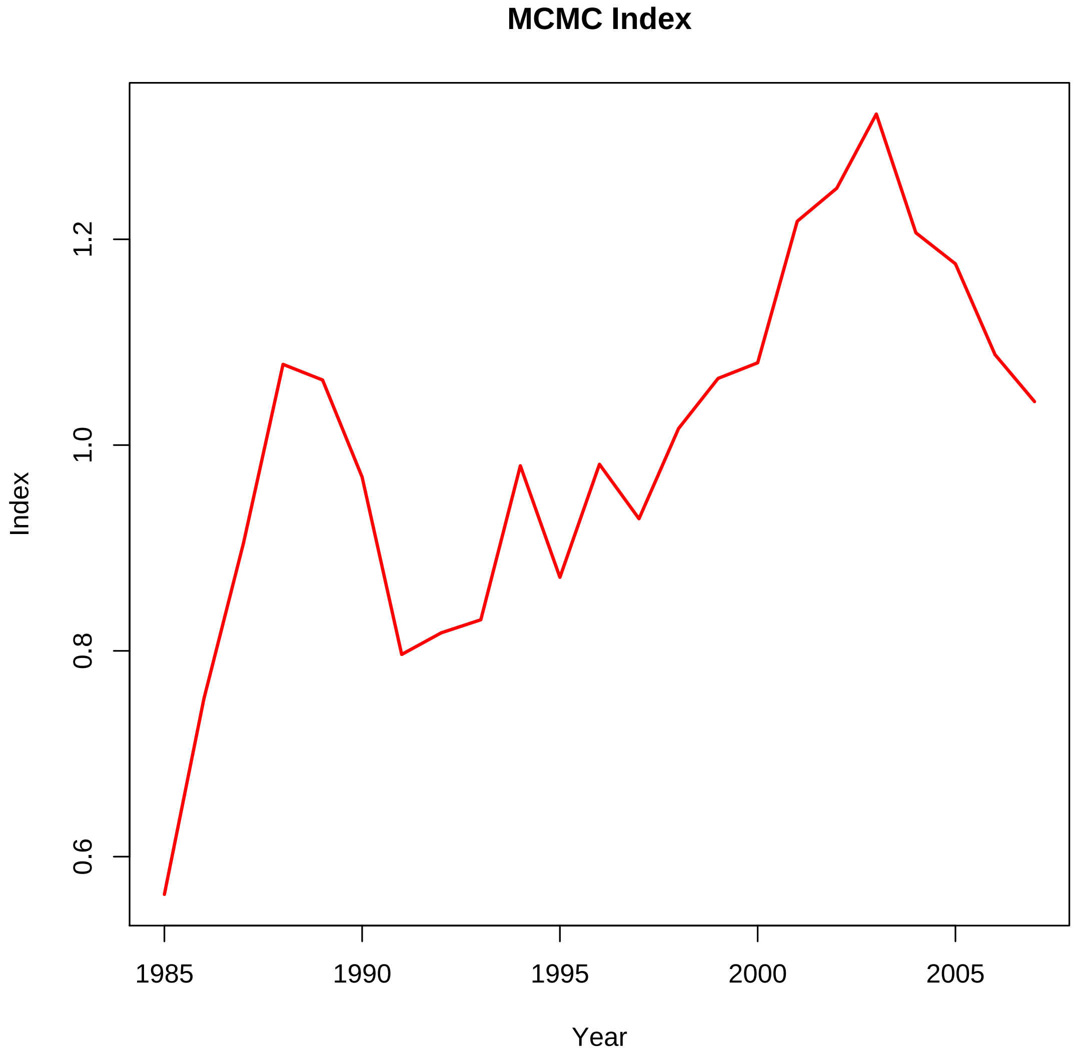


Figure 5: Posterior distributions for latent gradients (solid lines) versus those calculated directly from the indices produced by the SEDAR 17 data workshop (dashed lines). Latent time series include process error, but have sampling error removed.

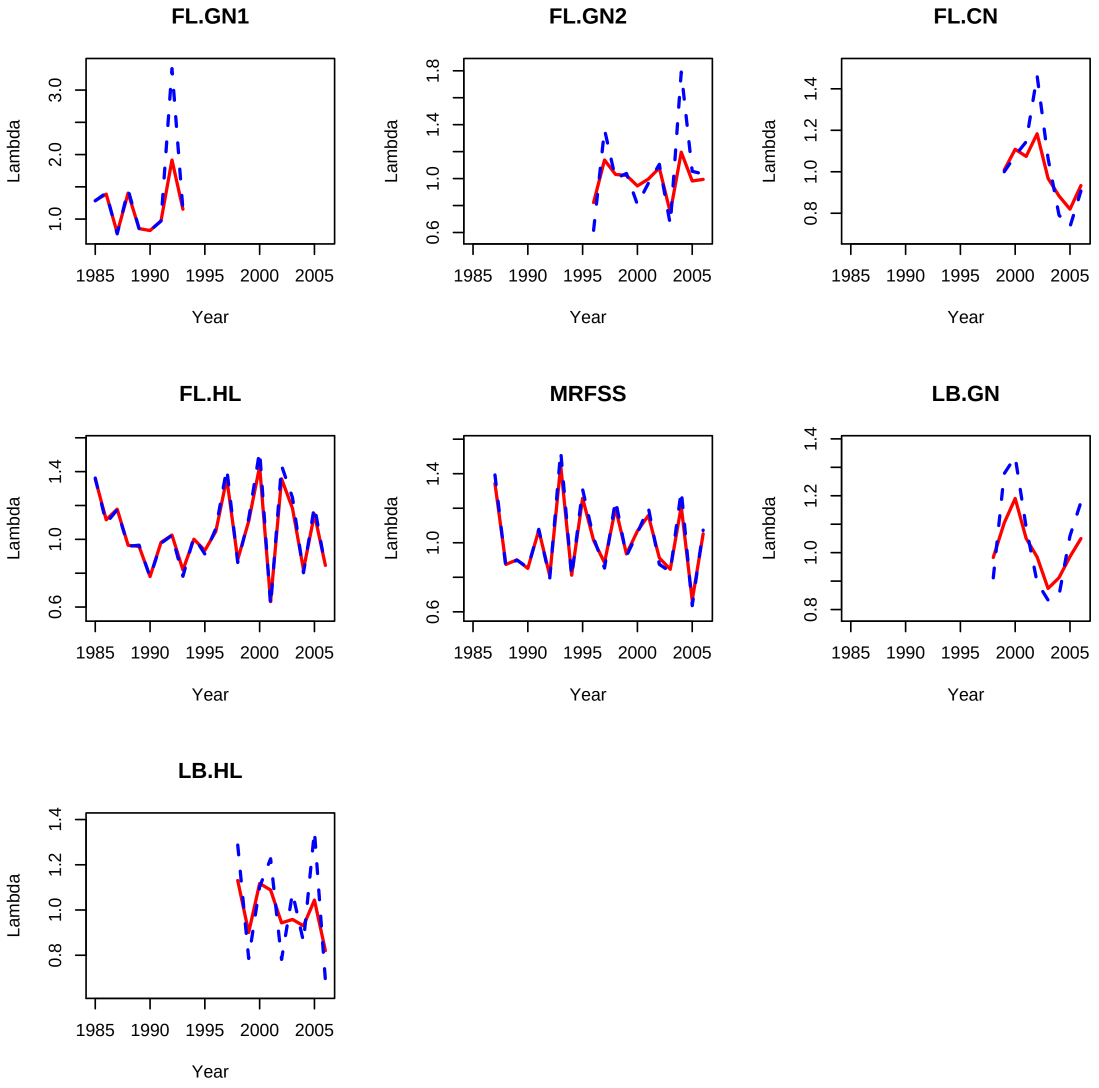


Table 1: Estimates of the combined index for U.S. Atlantic Spanish mackerel (as given by the mean of posterior prediction), together with standard errors. Also presented are posterior means and standard errors for λ_t, which is an estimator for the change in annual abundance of exploitable biomass ($\lambda_t = E_{t+1}/E_t$). This latter quantity was used to fit a stock reduction model to Spanish mackerel (see accompanying working document).

Year	μ_t	CV(μ_t)	λ_t	SE(λ_t)
1985	0.571	0.000	1.338	0.061
1986	0.763	0.184	1.201	0.062
1987	0.915	0.279	1.194	0.036
1988	1.079	0.326	0.991	0.031
1989	1.071	0.356	0.912	0.024
1990	0.979	0.397	0.822	0.022
1991	0.807	0.445	1.027	0.023
1992	0.826	0.473	1.018	0.035
1993	0.851	0.506	1.185	0.040
1994	0.985	0.536	0.892	0.031
1995	0.882	0.580	1.124	0.037
1996	0.982	0.620	0.948	0.024
1997	0.933	0.646	1.101	0.037
1998	1.034	0.669	1.047	0.015
1999	1.089	0.685	1.014	0.011
2000	1.064	0.702	1.128	0.013
2001	1.188	0.711	1.027	0.013
2002	1.196	0.729	1.058	0.013
2003	1.292	0.742	0.912	0.012
2004	1.243	0.758	0.978	0.012
2005	1.161	0.756	0.926	0.013
2006	1.074	0.774	0.959	0.010
2007	1.005	0.785	N/A	N/A