SCRS/1994/132

Col.Vol.Sci.Pap. ICCAT, 44 (2) : 332-336 (1995)

A NUMERICAL EVALUATION OF GLM METHODS FOR ESTIMATING INDICES OF ABUNDANCE FROM WEST ATLANTIC SMALL BLUEFIN TUNA CATCH PER TRIP DATA WHEN THE DATA ARE AGGREGATED TO MINIMIZE ZERO CATCHES

Porch, C.E.

National Marine Fisheries Service, Southeast Fisheries Center, 75 Virginia Beach Drive, Miami, Florida 33149, U.S.A.

SUMMARY

General linear models are often used to estimate indices of abundance from catch per unit effort data. One of the difficulties with this approach stems from the fact that many data sets have a large proportion of zero observations. A number of methods have been developed to deal with this problem, but a commonly used alternative has been to simply eliminate most of the zeroes by aggregating the data across several different observations. This study uses Monte Carlo simulations to examine the relative efficacy of several GLM approaches to predicting inter-annual trends in abundance when the data are aggregated.

RESUME

Les modèles linéaires généraux (GLM) sont fréquemment utilisés pour évaluer les indices d'abondance à partir des données de capture par unité d'effort. L'un des problèmes rencontrés avec cette approche est que de nombreux jeux de données ont une forte proportion d'observations égales à *zéro*. Un certain nombre de méthodes ont été élaborées afin de résoudre ce problème. Toutefois, l'une des possibilités à laquelle on a le plus fréquemment recours est d'éliminer la plupart des *zéro* en concentrant les données sur plusieurs observations différentes. Cette étude reprend les simulations de Monte-Carlo pour examiner l'efficacité relative de plusieurs approches GLM visant à prévoir les tendances interannuelles dans l'abondance lorsque les données sont concentrées.

RESUMEN

Con frecuencia, los modelos lineales generalizados se utilizan para estimar índices de abundancia de datos de captura por unidad de esfuerzo. Una de las dificultades de emplear este enfoque proviene del hecho de que muchos conjuntos de datos tienen una gran proporción de observaciones cero. Se ha desarrollado un número de métodos para tratar este problema, pero una alternativa común ha consistido en simplemente eliminar la mayor parte de los ceros agregando los datos obtenidos a través de varias observaciones diferentes. Este estudio utiliza las simulaciones de Monte-Carlo para examinar la eficacia relativa de varios enfoques GLM para predecir las tendencias interanuales en la abundancia cuando se agregan los datos.

INTRODUCTION

Porch and Scott (1994) used a Monte-Carlo approach to examine the ability of general linear models (GLM's) of catch per trip to index the interannual trends in the abundance of West Atlantic small bluefin tuna. They found that the Delta-lognormal model of Lo et al. (1992) best reproduced the true trends. The index adopted by the SCRS in their 1993 assessments, however, was based on the average catch per line-hour, pooled across five different trips, rather than the catch on a single trip (Brown and Browder 1994). The question arises as to whether the relative performance of the different GLM methods is consistent with Porch and Scott (1994) when the data are pooled.

This study essentially repeats Porch and Scott's analysis, only it uses Brown and Browder's aggregated data set as the template for generating the artificial data sets.

METHODS

Model Evaluation

Three types of GLM models were studied: ln(CPUE + c), CPUE/c, and the delta-lognormal. The mathematical details of these models are explained fully in Porch and Scott (1994). Four constants were examined in the ln(CPUE + c) GLM's: an arbitrary small value-- 0.001; the optimal value according to Berry's (1987) criteria-- 1024 (see Figure 1); the most often used constant in practice-- 1.0; and ten times the highest observed CPUE in any strata-- 3594. The constant used for the GLM on CPUE/c was also 3594.

Each of the above models were applied to 500 different artificially-generated data sets. The estimated indices (standardized CPUE) for each year y in data set k (I_{yk}) were normalized by their mean ($R_{yk} = I_{yk}/I_{.k}$) and compared to the normalized vector of known abundance ($A_y = N_y/N_.$). The error was calculated as $E_{yk} = R_{yk} - A_y$.

Accuracy was evaluated by examining the average error over all data sets $E_{y.}$ and the coefficient of error:

$$CE_y = \frac{\langle E_{yk}^2 \rangle}{A_y} \ ,$$

where the angle brackets denote the ensemble average. (Note that the coefficient of error is essentially the coefficient of variation with the sample mean replaced by the known value.)

Precision was evaluated by examining the standard deviation of the index

$$s_y = \sqrt{\frac{\sum_k^{500} (I_{yk} - I_{y.})^2}{499}}$$

Data generation

Each of the 500 artificial data sets were created by randomly drawing with replacement from strata-specific negative binomial distributions of catch per line-hour. The strata examined were distinguished by year, boat type, month of the year and region as described in Brown and Browder (1994). The number of draws from each strata was the same as the number of aggregated observations.

The strata-specific negative binomial distributions were determined from the expected CPUE and the variance in CPUE for that strata as described in Porch and Scott (1994). The regressions relating the variance to the expected CPUE,

$$V[CPUE]_{private} = 1.50 E^{1.49}[CPUE]$$

$$V[CPUE]_{charter} = 1.69 E^{1.49}[CPUE] \ ,$$

explained 48% and 80% of the variation in CPUE variance, respectively. The dispersion coefficient for the negative binomial was computed from the formula

$$K = \frac{E^2[CPUE]}{V[CPUE]e^{zs_a} - E[CPUE]} \ ,$$

where $V[CPUE]$ is obtained from the regressions above, s_a is the estimated standard deviation of the multiplicative parameter in the regression (0.86 for private and 0.43 for charter) and z is a standard normal random variable.

The expected CPUE for a given strata was computed from the model

$$E[CPUE] = q_0 B_b M_m R_r V_{mr} N_y$$

q_o = reference catchability
B_b = boat type effect on q
M_m = month effect on q and N
R_r = regional effect on q
V_{mr} = month and region interaction
N_y = abundance in year y.

The reference catchability, q_o = 0.000442, was obtained by regressing the strata-specific observed mean CPUE for private vessels on VPA estimates of stock abundance for ages 2 and 3. The boat type effect B was set equal to 1.0 for private boats and 2.0 for charter boats (= $q_{charter}/q_{private}$). The month and region were approximated by the relative variation in observed catch rates among strata (see Table 1). The stock abundance values for each year were equated with the VPA estimates for two and three year-old fish published in the 1991 SCRS report. The value used for 1992 was 30,000 fish.

Table 1. Month and region 'effects' used to generate CPUE. The row and column marked with an asterisk (*) give the main effects. The remaining entries are the interaction effects.

	REGION						
MONTH	1	2	3	4	5	6	*
June	2.3	0.5	1.7	2.3	1.1	2.6	1.7
July	1.6	0.5	0.4	1.7	1.0	2.1	1.1
August	0.5	0.2	0.0	0.0	0.0	0.0	0.1
*	1.3	0.3	0.8	1.4	0.7	1.6	

RESULTS

Accuracy

The six models estimated very similar patterns that generally reflected the underlying abundance, particularly during the later years of the time series (Figure 2). Overall, the most accurate estimates were obtained using the GLM's on ln(CPUE+3594) and CPUE/3594, but the difference between the six models was negligible (Figures 3 and 4).

Precision

The precision of the index estimates, measured by their standard deviation, is demonstrated in Figure 5. On average, the GLM estimates of the standard deviation matched the observed standard deviations pretty well.

The level of precision differed greatly between methods. The most precise predictions were obtained using the smallest constant (0.001).

CONCLUSIONS

The results of this study indicate that, in terms of predicting the trends accurately, the choice of GLM model is unimportant when the data are aggregated in the manner discussed (at least for this particular fishery). This is not entirely surprising since the problems associated with a high proportion of zero observations have been largely eliminated. Brown and Browder (1994), on the other hand, found that the choice of additive constants had a profound influence on the estimated CPUE indices. This discrepancy between our conclusions is partially explained by the fact that we are talking about two different measures. The emphasis in this paper is on the relative rather than absolute magnitudes of the estimated CPUE. The relative CPUE indices in Brown and Browder's paper are much less sensitive to the choice of additive constant than the absolute values. Moreover, Brown and Browder's conclusion is based on a single realization, whereas the conclusions in this paper are based on 500 realizations. On average there does not appear to be much difference between methods, but for any particular data set the estimates of standardized CPUE might differ substantially with different additive constants.

As found earlier in Porch and Scott (1994), all of the models predicted the standard deviation of the index estimates reasonably well. It must be reiterated, however, that this does not imply that the GLM estimates of the standard deviation of the index are useful for deciding between competing models. There was no correlation between precision and accuracy found in this study.

Interestingly, the GLM using Berry's constant was not as precise as the GLM's using smaller constants. This is in contrast to Porch and Scott's (1994) findings, which agree with Berry's assertion that statistical tests using his constant will be more powerful and robust than those of the other models. Possibly this is related to the fact that very few of the pooled catch rates in this study were zero and to the fact that averaging tends to make the data more normally distributed (central limit theorem). Moreover, Berry's constant was not well-determined for this data set. More work is needed.

REFERENCES

Berry, D. A. 1987. Logarithmic transformations in ANOVA. Biometrics 43:439-456.

Brown, C. A., and J. A. Browder. 1994. Standardized catch rates of small bluefin tuna in the Virginia - Rhode Island (U.S.) rod and reel fishery. ICCAT Collective Volume of Scientific Papers 42(1): 181-191.

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

Porch, C. E. and G. P. Scott. 1994. A numerical evaluation of GLM methods for estimating indices of abundance from West Atlantic Bluefin tuna catch per trip data when a high proportion of the trips are unsuccessful. ICCAT Collective Volume of Scientific Papers 42(1): 240-245.

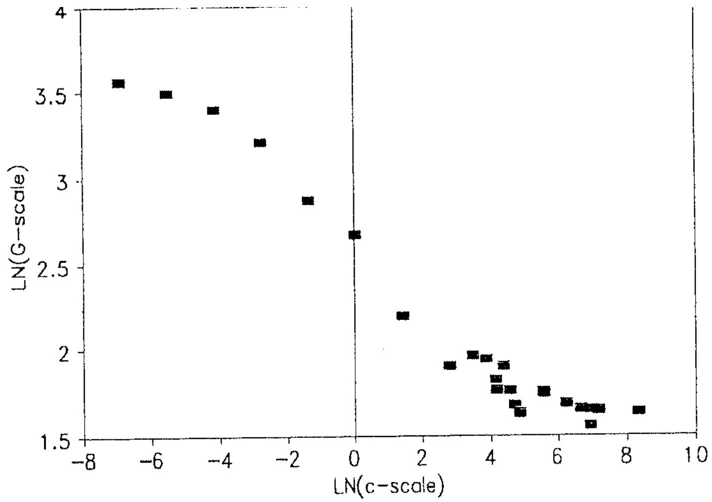

Figure 1. Berry's (1987) 'G' function measuring the skewness and kurtosis of the distribution of the GLM residuals. The value of *c* that minimizes G is optimal according to Berry's criteria. The points are averages of ten independent runs using different data sets. The 'minimum' selected here was the lowest observed value--1024.

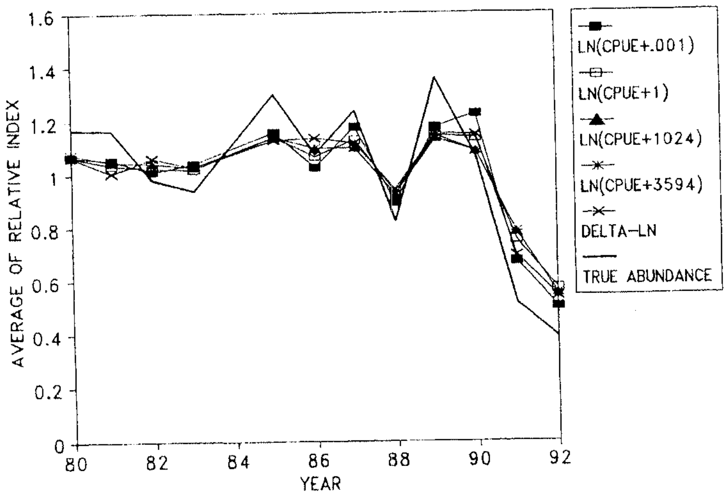

Figure 2. Average values of the estimated indices of relative abundance versus the known trends of [simulated] relative abundance.

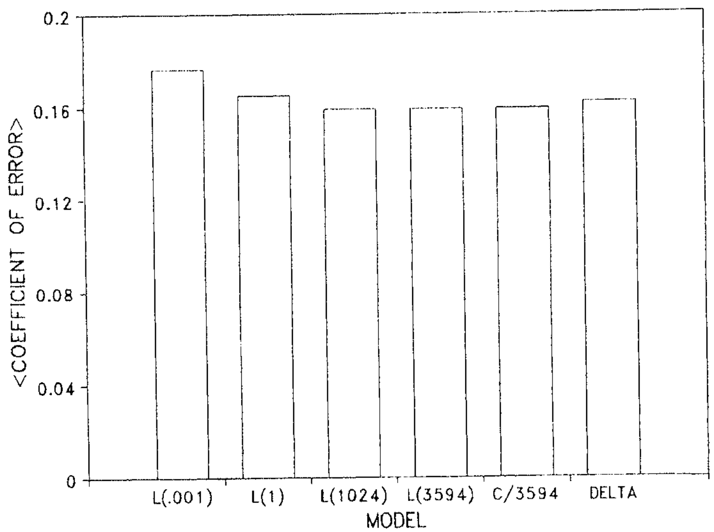

Figure 3. Average coefficient of error associated with each method.

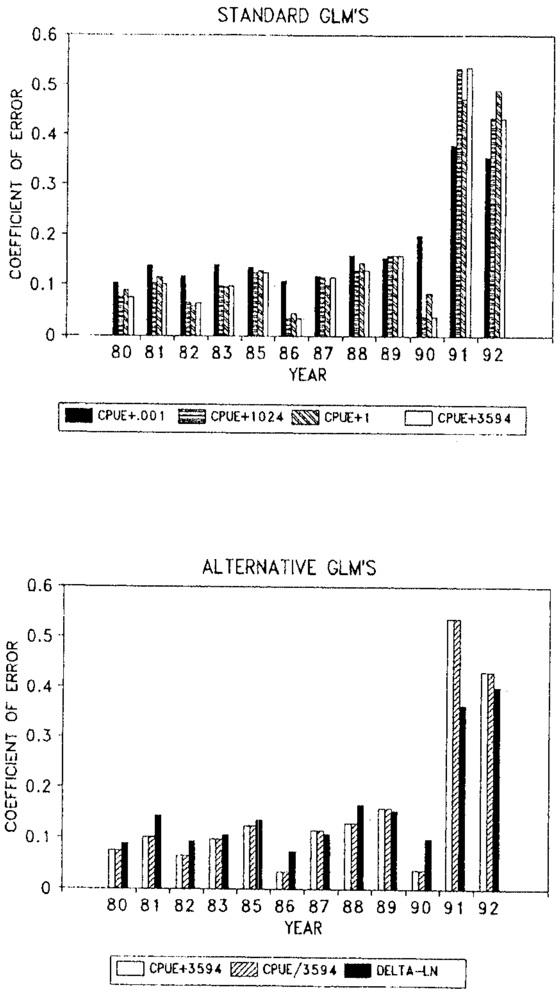

Figure 4. Coefficients of error by year.

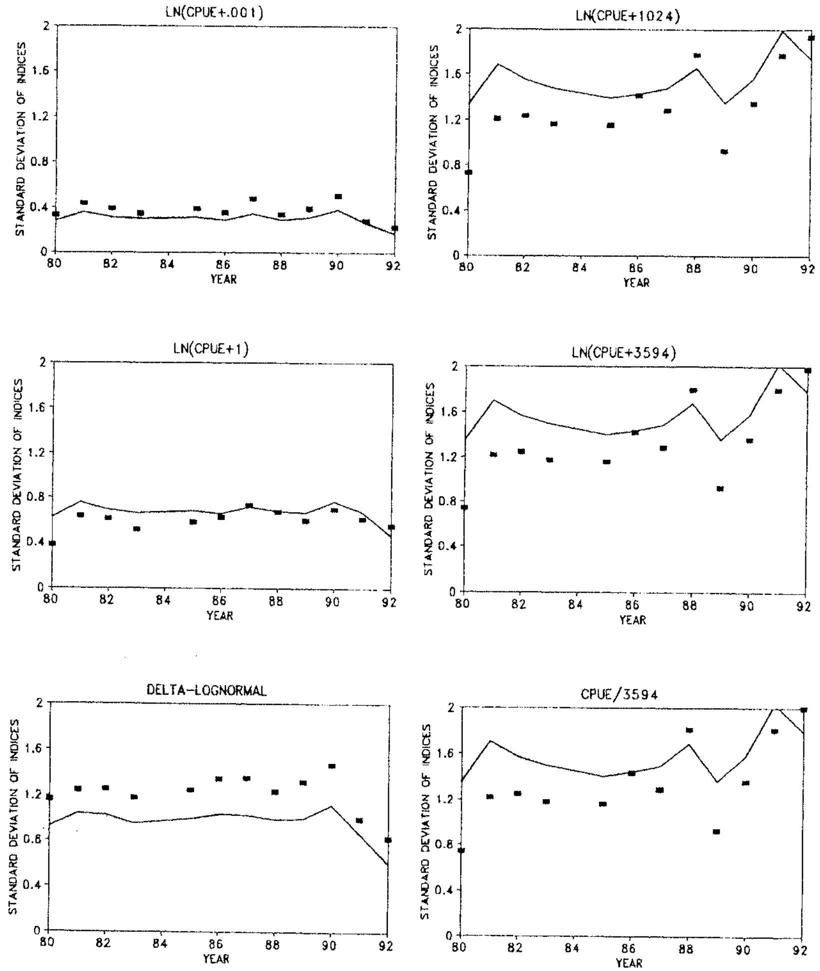

Figure 6. Standard deviation of the estimates of standardized CPUE by year. The lines track the average of the GLM estimates and the points represent the observed standard deviation of the 500 simulations.