SCRS/1993/075 Col.Vol.Sci.Pap. ICCAT, 42 (1) : 240-245 (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

Porch, C.E., G.P. Scott

Southeast Fisheries Science Center, National Marine, Fisheries Service, 75 Virginia Beach Drive, Miami, Florida 33149, USA

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. Methods that necessitate logarithmic transformations therefore require the addition of a constant to the data to avoid taking the logarithm of zero. The choice of the optimal constant has been addressed by Berry (1987). Lo *et al* (1992) suggested an alternative approach based on delta-lognormal methods. The 1991 SCRS Bluefin Working Group employed an approach based on a Taylor series approximation to ln(CPUE+constant). This study used Monte-Carlo simulations to examine the efficacy of these methods in predicting known inter-annual trends in abundance.

RESUME

Les modèles linéaires généralisés sont souvent utilisés pour estimer les indices d'abondance à partir des données de prise par unité d'effort. L'une des difficultés de cette méthode vient du fait que de nombreux jeux de données comportent une importante portion d'observations nulles. Les méthodes qui entraînent des transformations logarithmiques ont donc besoint d'ajouter une constante aux données pour éviter de prendre le logarithme zéro. Le choix de la constante optimale a été traité par Berry (1987). Lo *et al.* (1992) ont suggéré une méthode alternative se basant sur des modèles delta-lognormaux. Le groupe de travail thon rouge du SCRS a employé en 1991 une méthode se basant sur une approximation à ln (CPUE + constante) des séries de Taylor. Cette étude utilise des simulations Monte-Carlo pour examiner le degré d'efficacité de ces méthodes pour prédire les tendances inter-annuelles connues de l'abondance.

RESUMEN

Con frecuencia se usan los modelos lineales generalizados para estimar los índices de abundancia basándose en datos de captura por unidad de esfuerzo (CPUE). Una de las dificultades de este enfoque estriba en el hecho que muchos conjuntos de datos tienen una gran proporción de observaciones cero. Por lo tanto, los métodos que requieren transformaciones logarítmicas, requieren también la incorporación de una constante a los datos, con el fin de evitar tomar el logaritmo de cero. Berry (1987) ha tratado acerca de la elección de la constante óptima. Lo *et al* (1992) sugerían un enfoque alternativo basado en modelo lognormales-delta. El grupo SCRS que estudió el atún rojo en 1991, aplicó un enfoque basado en una serie Taylor de aproximación al ln (CPUE+constante). Este estudio emplea simulaciones Montecarlo para examinar la eficacia de estos métodos en la predicción de tendencias interanuales de la abundancia.

INTRODUCTION

The average catch per unit effort (CPUE) of a fishing fleet during a given period of time is often assumed to be proportional to the average abundance N of the stock it is fishing-- $CPUE = qN$. The proportionality factor q is called the catchability coefficient. If catchability remains constant, then inter-annual trends in CPUE should parallel inter-annual trends in abundance.

In practice, catchability tends to change through time with changes in factors such as engine horsepower, captain skill, or gear modifications. Moreover, catchability tends to vary from area to area and among different components of a fishery (e.g., private versus charter boats). Accordingly, changes in the distribution of effort or sampling among the different strata will result in a change in the overall nominal CPUE that is independent of the abundance of the stock.

The key to developing an accurate index of abundance lies then in distinguishing the trends in abundance from the trends in catchability. Gulland (1956) first introduced the idea of using general linear models (GLM) to do this. His work was later expanded on by a number of others including Robson (1966), Gavaris (1980), and Kimura (1980). Since then, GLM models have become a standard tool of assessment scientists.

The basic assumption behind most GLM approaches is that the error structure and effects are multiplicative, e.g.,

$$CPUE = U \prod Y_y B_{ij}^{x_{ij}} e^{\epsilon} \quad .$$

In the above model U is an unknown constant, Y_y is the year-effect on N in year y, B_{ij} is the effect on q of category j in category type i, x_{ij} is equal to 1 when category j occurs and zero otherwise, and ϵ is the error. The corresponding logarithmic transform,

$$\ln(CPUE) = \ln(U) + \ln(Y_y) + \sum \ln(B_{ij}) x_{ij} + \epsilon \quad , \qquad (1)$$

falls quite naturally into the framework established for GLM's. Provided the catchability coefficient is parameterized properly, a standardized measure of CPUE (and an unbiased index of abundance) during year j is

$$I_y = e^{\ln(U) + \ln(Y_y) + \frac{1}{2}\sigma^2} \quad . \qquad (2)$$

In similar fashion measures of the strata-specific catchabilities can be obtained (see Kimura 1980).

The GLM estimates of the parameters can be substituted into Equation (2), but the resulting index is no longer unbiased. Bradu and Mundlak (1970) develop the unbiased formulation, which is asymptotically equivalent to

$$I_y = e^{\ln(U) + \ln(Y_y) + \frac{m+1}{2m}(s^2_R - s^2_{\ln(UY)})}$$

(Ebbeler 1973). The terms s^2_R, m, and $s^2_{\ln(UY)}$ are the residual variance, the degrees of freedom for the residual variance, and the estimated variance of $\ln(U)+\ln(Y_y)$, respectively. In practice the asymptotic formulae are sufficient if the number of degrees of freedom in the regression exceed 50 (Gavaris 1980).

Bradu and Mundlak's formula is the minimum variance estimator provided CPUE is lognormally-distributed. The CPUE frequency distributions for many commercial fisheries are in fact approximately lognormal (Gulland 1956), so the GLM approach appears tailored for the task. The CPUE frequency distributions for recreational fisheries or fisheries directed at relatively rare resources, however, typically have a large proportion of zero catches. Thus the lognormal distribution is inappropriate and, more importantly, Equation (1) is undefined.

The solution of many researchers has been to add a constant c to each datum so that observations of zero catches can be incorporated into the analysis. Thus a new model is defined:

$$\ln(CPUE+c) = \ln(U') + \ln(Y'_y) + \sum \ln(B'_{ij}) x_{ij} + \epsilon,$$

where the unbiased index of abundance is

$$I_y = e^{\ln(U') + \ln(Y'_y) + \frac{m+1}{2m}(s^2_R - s^2_{\ln(U'Y')})} - c \quad .$$

The choice of the additive constant c greatly affects the conclusions of statistical tests of hypotheses because it affects the shape of the frequency distribution of the residuals. Berry (1987) has developed criteria for identifying the constant that makes the residuals closest to a sample from a normal distribution. This in principle should make the associated statistical tests more powerful and more robust to outliers. In practice, however, one may be more interested in producing an unbiased and precise index than in testing whether the model effects are statistically significant. This being the case, it is interesting to know how the choice of c effects the behavior of the estimator.

An alternative to the $\ln(CPUE+c)$ GLM was used by the 1991 bluefin tuna working group of the International Commission for the Conservation of Atlantic Tunas. This approach employs a first order Taylor series approximation to $\ln(CPUE+c)$, which for large c reduces to $CPUE/c + \ln(c)$. The suggested value of c was the largest observed CPUE value multiplied by 10. The assumed error structure is therefore additive and the GLM would then be performed on CPUE/(10*MAX[CPUE]). The corresponding index of abundance is

$$I_y = c(U'' + Y''_y) \quad .$$

The impetus behind the working group's application of this model was to avoid the roundoff error potentially accompanying the summing of small CPUE's with the large constant.

A second alternative to ln(CPUE+c) GLM's was advanced by Lo et al. (1992). It employs a delta-lognormal model wherein separate GLM's are developed for the logarithms of the positive CPUE data and proportion of observations for which there was a positive catch (PPOS). In this study the GLM on ln(PPOS) was weighted by the number of observations determining PPOS in each strata. The corresponding index of abundance is simply the multiple of the separate GLM estimators (Equation 3).

The purpose of this paper is to evaluate the accuracy and precision of these methods when they are applied to West Atlantic small (< 150 cm) bluefin tuna.

METHODS

Model Evaluation

Three types of GLM models were studied: ln(CPUE+c), CPUE/c, and the delta-lognormal approach of Lo et al. (1992). 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-- 0.064 (see Figure 1); the most often used constant in practice-- 1.0; and ten times the highest observed CPUE in any strata-- 110. The constant used for the GLM on CPUE/c was also 110.

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 ($A_{yk}=N_{yk}/N_{.k}$). The error was calculated as $E_{yk}=R_{yk}-A_{yk}$.

Accuracy was evaluated by examining the average error over all data sets $E_{y.}$ and the root-mean-square error:

$$RMSE_y = \sqrt{\frac{\sum_{k}^{500}(E_{yk})^2}{500}} \quad .$$

Precision was evaluated by examining the standard deviation of the index

$$s_y = \sqrt{\frac{\sum_{k}^{500}(I_{yk}-I_{y.})^2}{499}} \quad . \qquad (3)$$

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 trip. The strata examined were distinguished by year (1980-1992), boat type (private or charter), month of the year (June-September), and region (Virginia, Maryland and Delaware combined, New Jersey, New York, and Connecticut and Rhode Island combined). The number of draws from each strata was the same as the number of trips targeting bluefin tuna in the actual data base. Not all strata are represented in the data collected during the years 1980-1984, 1986, and 1988. There were no data for private boats in 1981. Data from 1984 were excluded because boat type and fishing area were not recorded (Brown and Browder 1993).

The strata-specific negative binomial distributions were determined from the expected CPUE and the variance in CPUE for that strata. Following the approach of Small and Downham 1985, Porch 1990, and Lester et al. 1991, the expected variance in CPUE was related to the expected CPUE by regressing the logarithm of observed variance in catch per trip on the logarithm of mean catch per trip. The negative binomial distribution could then be determined solely from the expected CPUE.

Separate regressions were performed for charter and private vessels (Figure 2). The results,

$$V[CPUE]_{private} = 3.75\,E^{1.2}[CPUE]$$

$$V[CPUE]_{charter} = 4.05\,E^{1.3}[CPUE] \quad ,$$

accounted for 90% and 91% of the variation in CPUE variance, respectively.

The dispersion parameter of the negative binomial distribution was computed as

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

where $V[CPUE]$ is obtained from the regressions above, s_e is the estimated standard deviation of the multiplicative parameter (0.69 for charter, 0.68 for private) and z is a normally-distributed random variable with mean 0 and variance 1. The error term was included to mimic natural variability within strata.

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_0 = 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 specific values selected for each of these parameters were patterned after the U.S. recreational fishery targeting small bluefin tuna. The reference catchability, q_0 = 0.0000069, 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 2.3 for charter boats ($= q_{charter}/q_{private}$). The month effect considered two factors-- the change in stock abundance among months due to mortality and a 10% per month decrease in overall catchability. The resulting effects were $M_{June}=1.05$, $M_{July}=1.02$, $M_{August}=0.98$, and $M_{September}=0.95$. The regional effects on catchability were equated with the relative magnitudes of the observed mean catch rates: R_r = 1.0, 0.3, 1.2, 1.0, 1.6. The month/region interaction effect, V_{mr}, reflects the migration of fish through the various regions. This effect was approximated by the relative variation in observed catch rates among month/region 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/Region interaction 'effects' used to generate CPUE.

	REGION				
MONTH	Va	Md-De	NJ	NY	Ct-RI
June	2.4	0.0	0.0	0.0	0.0
July	1.1	1.7	0.8	0.1	0.0
August	0.5	2.3	0.9	0.1	0.5
September	0.0	0.0	0.5	3.8	3.5

RESULTS

Accuracy

The six models estimated similar patterns that generally reflected the underlying abundance, particularly during the later years of the time series (Figure 3). The relative rmse (rmse divided by the true value of normalized abundance) is averaged across years in Figure 4. Overall, the most accurate estimates were obtained using the delta-lognormal model. The average relative rmse was 14% using the delta-lognormal approach compared to 17% for the next most accurate method-- the GLM on ln(CPUE+110). The delta-lognormal indices were not necessarily the most accurate in any given year however, as Figure 5 clearly demonstrates.

The GLM on ln(CPUE+1) was nearly as accurate as the GLM on ln(CPUE+110), but the GLM's on ln(CPUE+.001) and ln(CPUE+.064) [Berry's constant] were not competitive.

The results from the GLM on CPUE/110 were almost identical to those from the GLM on ln(CPUE+110), the latter being slightly more accurate and precise. The similarity is not surprising inasmuch as the former is simply an approximation of the latter. The suspected roundoff error does not appear to be an important factor when 16 bit or greater machines are employed.

Precision

The precision of the index estimates, measured by their standard deviation, is demonstrated in Figure 6. On average, the GLM estimates of the standard deviation closely match the observed standard deviation computed from the 500 simulation runs using Equation 3.

The level of precision differed greatly between methods. The most precise predictions were obtained using the constant determined by Berry's criteria (0.064).

CONCLUSIONS

The results of this study indicate that, in terms of accuracy, the best GLM model for this particular fishery is the delta-lognormal model followed closely by the GLM on ln(CPUE+c) with c equal to ten times the maximum observed CPUE. The GLM using Berry's constant, however, yielded substantially more precise index estimates than any of the other models. This suggests that the statistical tests using Berry's constant will also be more powerful and robust than those of the other models. Accordingly, one should consider using Berry's constant when GLM statistics are used to identify variables that do not contribute significantly to the regression. The delta-lognormal or ln(CPUE+10*MAX[CPUE]) models could then be used to obtain the final estimates of the indices from the selected variables.

It was also found that all of the models predicted the standard deviation of the index estimates reasonably well. This does not imply, however, that GLM estimates of the standard deviation of the index are useful for deciding between competing models. This would be true only if the estimators were unbiased in each individual year, which from Figure 1 is clearly not the case. Thus, it is possible for a more precise estimator (e.g., Berry's GLM) to be less accurate on average than a less precise estimator (e.g., the delta-lognormal model). There was no correlation between precision and accuracy found in this study.

Further studies should be conducted to elucidate the relationship between index precision and sample size, and to examine the generality of the conclusions to data sets with different underlying CPUE distributions.

REFERENCES

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

Bradu, D., and Y. Mundlak. 1970. Estimation in log-normal linear models. Journal of the American Statistical Association 65:198-211.

Brown, C. A., and J. A. Browder. 1993. Standardized catch rates of small bluefin tuna in the Virginia - Rhode Island (U.S.) rod and reel fishery. ICCAT Working Document SCRS/93/67.

Ebbeler, D. H. 1973. A note on estimation in log-normal linear models. Journal of Statistics and Computer Simulation 2: 225-231.

Gavaris, S. 1980. Use of a multiplicative model to estimate catch rate and effort from commercial data. Canadian Journal of Fisheries and Aquatic Sciences 37:2272-2275.

Kimura, D. K. 1980. Standardized measures of relative abundance based on modelling log(c.p.u.e.), and their application to Pacific ocean perch (*Sebastes alutus*). Journal du Conseil 39:211-218.

Lester, N. P., M. M. Petzold, and W. I. Dunlop. 1991. Sample size determination in roving creel surveys. American Fisheries Society Symposium 12:25-39.

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. 1990. Simulating the dynamic trends of fisheries regulated by small daily bag limits. Transactions of the American Fisheries Society 119:836-849.

Robson, D. S. 1966. Estimation of the relative fishing power of individual ships. ICNAF Res. Bull. 3:5-14.

Small, I., and D. Y. Downham. 1985. The interpretation of angler's records (trout and seatrout, *Salmo trutta* L., and *Salmo salar* L.). Aquaculture and Fisheries Management 16:151 169.

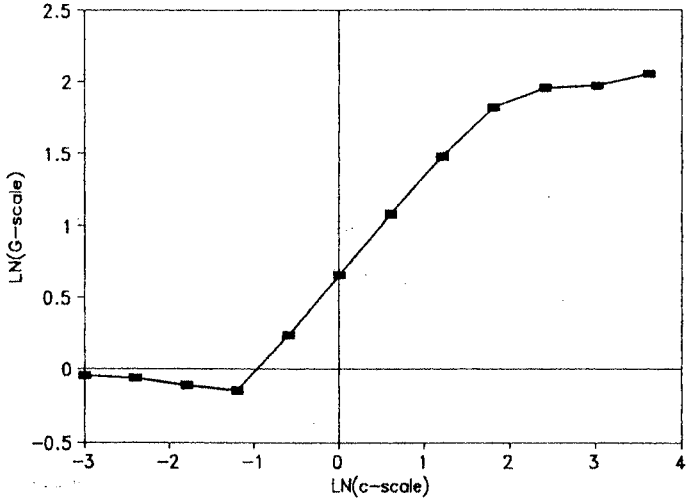

Figure 1. Berry's (1987) 'G' function measuring the skewness and kurtosis of the distribution of the GLM residuals. The value of *c* that minimizes G is optimal according to Berry's criteria.

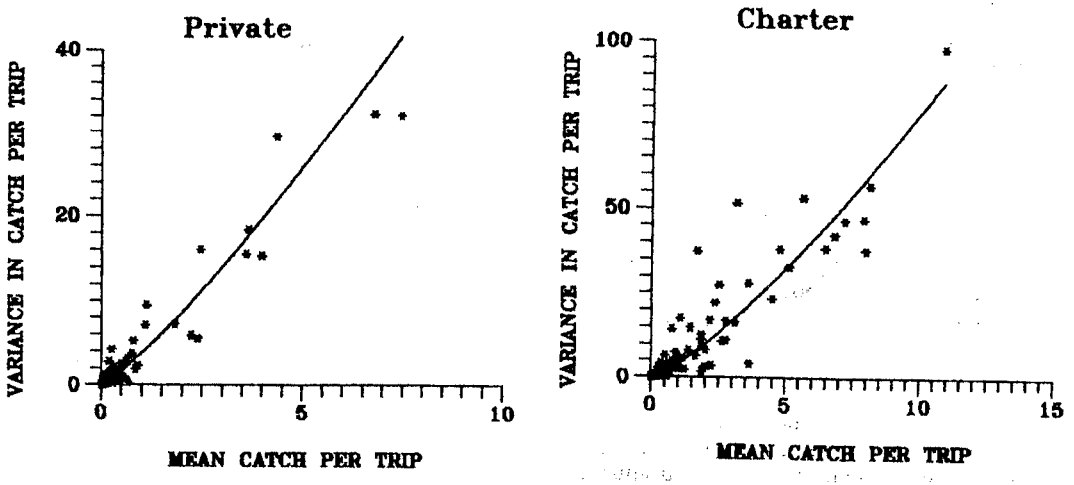

Figure 2. Regressions of variance in catch per trip against average catch per trip for private and charter boats.

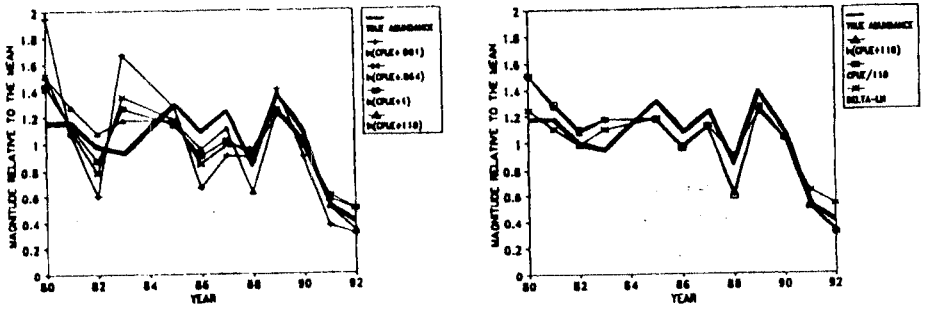

Figure 3. Average values of the estimated indices of relative abundance versus the known trends of [simulated] relative abundance.

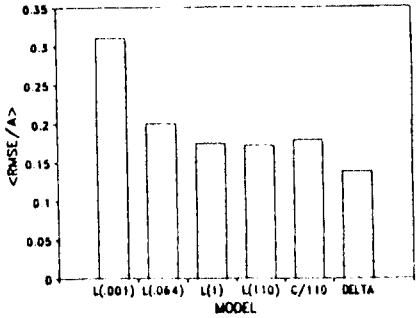

Figure 4. Relative root-mean-square errors averaged across years for each method.

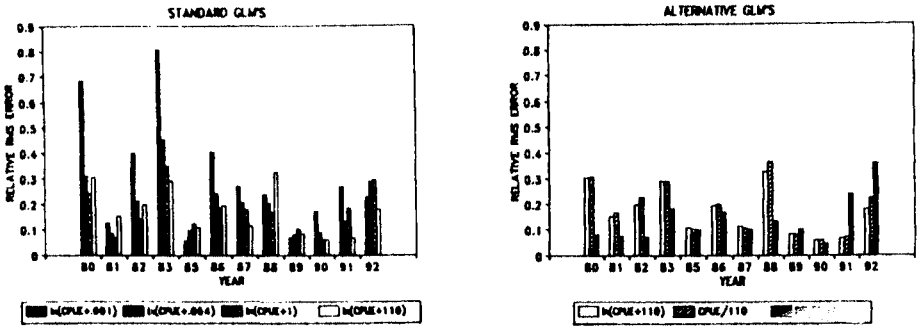

Figure 5. Root-mean-square errors relative to the true value (RMSE/A) by year.

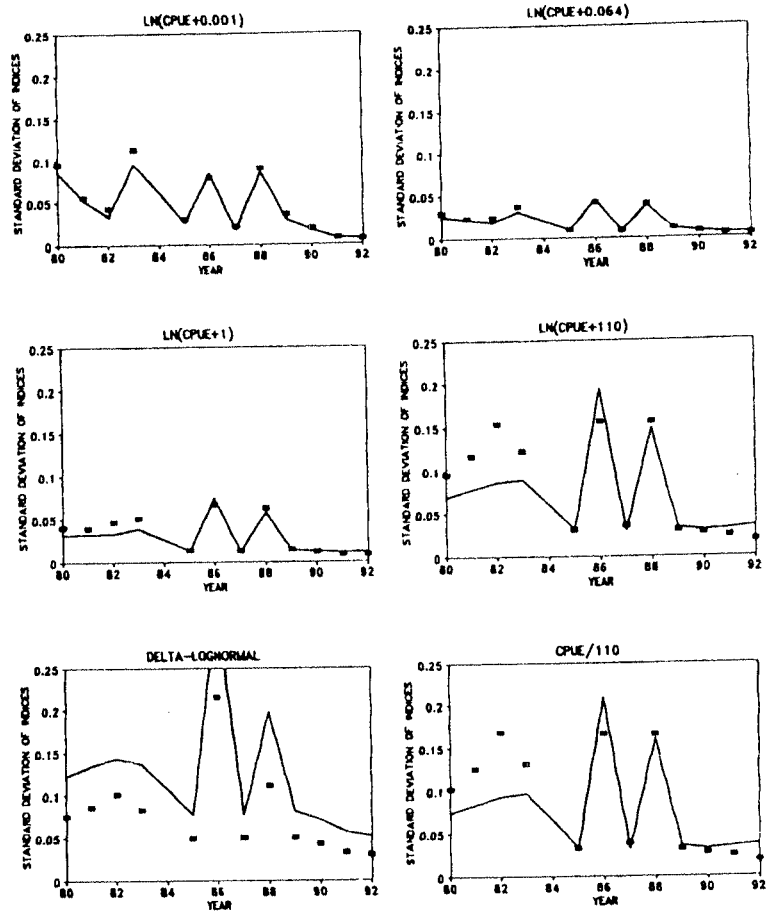

Figure 6. The standard deviation of the estimates of standardized CPUE by year. The lines track the average of the GLM estimates of standard deviation and the points represent the observed standard deviation of the 500 simulations (Equation 3).