Northwest Atlantic Fisheries Organization

Serial No. N1693 SCIENTIFIC COUNCIL - 1989 NAFO SCR Doc. 89/93

WORKING GROUP ON SHRIMP AGEING - OCTOBER 1989

Some Observations on Modal Analysis of Shrimp Length Frequency Distributions

by

D. G. Parsons

Science Branch, Department of Fisheries and Oceans,
P. O. Box 5667, St. John's, Newfoundland, Canada A1C 5X1

and

L. Savard

Fisheries Research Division, Dept. of Fisheries and Oceans, Maurice-Lamontagne Institute
P. O. Box 1000, Mont-Joli, Quebec, Canada G5H 3Z4

Introduction

Age determination of northern shrimp (*Pandalus borealis*) is based on the assumption that regularly occurring modes in the length frequency distributions represent different cohorts. It is also assumed that maturation is synchronous within cohorts (i.e. males reach maturity at the same age and change sex at the same age) and that primiparous and multiparous females can be distinguished by the presence or absence of sternal spines (McCrary 1971). The verity of the first two assumptions is always debateable, as they are very difficult to test either in the natural habitat or under laboratory conditions. Differentiating males and females separates the younger and older animals while the sternal spines distinguish the young and old females. Males also can be separated into mature and immature categories which comprise different modal groups. At this point, however, one is left with a group which is made up of a number of ages with overlapping length distributions and then some numerical method is required to separate the components.

Numerous programs exist to separate components in a mixture of normal (or non-normal) distributions (e.g. NORMSEP, ENORMSEP, MIX, MULTIFAN). Most are based of methods of maximum likelihood with indicators of goodness of fit. The commonly used routines require input parameters on the number of components and their shape. Output generally includes estimates of proportions, means and standard deviations (with standard errors) along with expected frequencies from the "best" model. A good fit is one with a low chi-square and low standard errors, the latter being the more important indicator.

All of the available routines require practice to gain familiarity with the procedures and an understanding of how the output is generated. With experience, the outputs can be evaluated as to which best explains the data in terms of age composition. This paper draws attention to the sensitivities of these types of analyses from three perspectives; the selection of the number of components, the choice of starting parameters and the preparation of data. The purpose of the paper is to provide the inexperienced user with some idea of how easily the outputs can be affected and to generate some discussion at this meeting on how best to deal with the problems. It is <u>not</u> the purpose of this paper to question the integrity of any of the methods, either generally or specifically, practically or theoretically.

Methods

Samples of shrimp from West Greenland research surveys were analysed by Carlsson et al.(1988) for age composition. Several of these samples were selected (with the first author's consent) to investigate the sensitivity of modal analysis. The Macdonald and Pitcher (1979) MIX program (release 2.3) was used to separate the length frequency distributions into normal components. One sample was analysed assuming a different number of components on each run. A second was analysed under two different perceptions of how the components were overlapped (different standard deviations) and a third to demonstrate the affects of leading and/or trailing zeros.

Results and Discussion

1. Number of Components

The length frequency distribution analysed under different interpretations of the number of components is given in Fig. 1. The interpretation of the number of modes in this figure is highly subjective and depends on whether or not the peaks are interpreted as modes or as noise. In the first instance, six components were interpreted. All input and output of the model are given in Table 1. The procedure ran freely with no constraints on the parameters in the final run. The observed and expected data were in very good agreement (P = 0.97) and the standard errors of the parameters were generally low, except in some problem areas in the middle of the distribution.

If some of the peaks are seen as noise, especially in the middle, then it might be interpreted that perhaps only three components are present with modes at roughly 14, 17 and 20 mm. Under this assumption, and the inputs given in Table 1, again reasonable results were obtained. A low chi-square was obtained (P = 0.64) as were the standard errors of the parameters for the first two components. The errors for the last were high because of the overlap. Nevertheless, this would be considered as a good fit, assuming three components.

This example can be carried farther by obtaining good and statistically acceptable results in assuming two and, in the extreme, one component in the data (Table 1). In terms of goodness of fit, the six component model is far ahead of the others but it is quite possible that we are modelling the noise in this case. It also is possible that the "correct" answer is four or five components but the procedure would not run freely under either of those scenarios. Parameters can be contstrained at certain values if there are some ancillary biological data to suggest that such constraints are appropriate. However, other than separating the samples by sex and maturity stages, there are no other pieces of information known at present which can be used to aid the modal analysis.

2. Choice of Starting Values for Standard Deviations

The results of any modal analysis can be easily affected by the interpretation of the relative strength of the components, even though there may be no confusion as to how many modes are present. The most critical starting parameters to estimate in the Macdonald and Pitcher method are the standard deviations of each normal component. The length distribution to demonstrate the problem associated in determining the shape of the curves is given in Fig. 2. In the first run, it was interpreted that all components were strongly represented in the data and starting values of the standard deviations were given as 0.7, 0.7 and 0.8. In the second, the middle mode was considered to be much stronger than the adjacent modes and values were estimated at 0.6, 1.3, and 0.6. All remaining starting parameters were identical in both runs (Table 2).

The goodness of fit was the same in both cases (P = 0.65) and the standard errors for most of the estimated parameters were low.

In terms of the latter, it is difficult to tell which run might be better. The estimated means were similar for the two analyses. Proportions, however, were substantially different, especially for the two last elements. In the model in which the middle mode was considered to be strong, only 5% of the animals were estimated to be in the last group compared to 41% under the alternate assumption. There are various constraints within the program to address this problem but there is no biological basis to say, for example, that all standard deviations should be equal or that all coefficients of variation should be equal. At present, results of some analyses reflect the first impression the investigator gets when viewing the length frequency data. The effect is greatest on the proportions and less on the estimated means. The difficulty in estimating relative (or absolute) cohort strength from such analyses is obvious.

Despite the sensitivity to input parameters, it is at times apparent in these analyses that there is only one solution to the data and that these results are obtained over a wide range of starting parameter estimates. This can be looked upon as objectivity in the analysis but the outputs from any run should be reviewed relative to the inputs just in case the results are nonsensical. In many cases, however, the results are quite acceptable and provide the user with an alternative interpretation to the first impression.

3. Preparation of Data

The number of observations within a component is often critical to reaching a solution. In cases where the numbers are small, it is advisable to delete data on the extremes of the distribution rather than have them affect the analysis of the majority of the data. If these "tails" are separated from the main part of the frequency by zeros, it is advisable to delete these as well. Macdonald and Green (1988) suggest avoiding zeros because they increase computation time and render the chi-square invalid. Aside from that, their inclusion can influence the results obtained from the procedure. The data to demonstrate the affect of zeros are the same as shown in Fig. 2. The program was run with a leading, a trailing and finally with both leading and trailing zeros. The estimated parameters from each are given in Table 3. The P - values for the chi-square were high (0.65 - 0.72) and standard errors were not excessively large. Similar to the problem with starting parameters (above), the greatest affect was on the proportions. The contribution of the first component did not change substantially, ranging from 22 to 24%. The second ranged from 38 to 50% and the third from 28 to 38%. The estimated mean lengths did not change to such a degree. From Table 2 (run A) it can be seen that the results obtained with no zeros were different from all the runs in the present example.

It is clear that the preparation of data is very important when using these techniques and care should be taken to exclude low counts and zeros from the data set. On the other hand, the inclusion of zeros on either or both ends of the data has enabled an unconstrained solution to be reached in other exercises. In situations such as these where there is no bioloical or statistical justification for the use of constraints, this becomes very important.

Conclusions

Current methodology for ageing northern shrimp is tenuous at best, given the assumptions of the ontogeny of the animal and the inherent problems of separating length frequency data where the overlap between modes is often severe. Biologists working on the problem are constantly challenged with applying the results to classical fish population models (e.g. yield per recruit, cohort analysis). Based on experience with the technique, it appears that representative length at age data can be obtained for descriptive or comparative purposes and the resulting parameters can be applied to yield per recruit calculations. To attempt a cohort analysis is quite another issue. Typically, the most contentious issues in the

latter are the estimates of natural mortality (M), fishing moratlity in the terminal year (F_t) and partial recruitment rates. Seldom is the integrity of the catch at age matrix challenged. Based on our results with modal analysis, it appears that the basic data required by the model are highly suspect. That is to say, given the sensitivity of the estimated proportions to the input parameters, it would be very difficult to produce reliable catch at age data.

Length-based methods provide some alternatives but are themselves problematical (Lai and Gallucci, 1988). Nevertheless, their potential application to northern shrimp should be thoroughly investigated. In an attempt to provide more reliable estimates of catch at age, it would be advisable to determine which constraints might be appropriate in using the modal analyses. Is it reasonable, for example to constrain all C.V.'s to be equal, all standard deviations to be equal or constant? With some empirical basis for these assumptions, the reliability in the quantitative results might be greatly improved.

The most important conclusion is that the user needs to be very careful in analysing length frequency data and even more careful when applying the results. It is a highly subjective exercise in the first place which can be further confounded by the intricacies of the statistical analyses used.

References

Carlsson, D.M., D.G. Parsons and L. Savard. 1988. Modal Analysis for Davis Strait Shrimp Samples. NAFO SCR Doc. 88/67. Ser. No. N1510. 5p.

Lai, H.L. and V.F. Gallucci. 1988. Effects of parameter variability on length cohort analysis. J. Cons. int. Explor. Mer. 45: 82 - 92.

Macdonald, P.D.M. and T.J. Pitcher. 1979. Age-groups from size-frequency data: A versatile and efficient method of analysing distribution mixtures. J. Fish. Res. Board Can. 36: 987 - 1011.

Macdonald, P.D.M. and P.E.J. Green. 1988. User's guide to program MIX: An interactive program for fitting mixtures of distributions. Ichthus Data Systems. Hamilton, Ontario, Canada. 60 p.

McCrary, J.A. 1971. Sternal spines as a characteristic for differentiating between females of some Pandalidae. J. Fish. Res. Board Can. 28: 98 - 100.

Table 1. Starting values[1] and estimated parameters of normal components from the Macdonald and Pitcher analysis (k = 6, 3, 2, and 1).

Components	1	2	3	4	5	6
Proportions	0.186	0.180	0.288	0.097	0.124	0.124
Std. Errors	0.064	0.120	0.132	0.106	0.165	0.117
Proportions	0.112	0.810	0.078			
Std. Errors	0.066	0.189	0.145			
Proportions	0.131	0.869				
Std. Errors	0.048	0.048				
Proportions	1.000					
Means	13.567	15.246	16.983	18.259	19.384	21.006
Std. Errors	0.241	0.247	0.164	0.376	0.413	1.427
Means	13.545	17.017	21.090			
Std. Errors	0.140	0.462	1.799			
Means	13.533	17.468				
Std. Errors	0.145	0.248				
Means	16.943					
Std. Errors	0.175					
Start. Values	0.500	0.500	0.500	0.500	0.500	0.500
Sigmas	0.506	0.552	0.493	0.219	0.490	1.174
Std. Errors	0.165	0.447	0.295	0.238	0.551	0.698
Start. Values	1.000	1.000	1.000			
Sigmas	0.369	2.009	1.260			
Std. Errors	0.252	0.536	0.891			
Start. Values	0.500	2.500				
Sigmas	0.443	2.218				
Std. Errors	0.156	0.187				
Start. Values	2.500					
Sigmas	2.481					
Std. Errors	0.131					
Chi-square	25.564	11.767	10.618			0.575
P Value	0.143	0.760	0.643			0.966

[1] Starting values for proportions = 1/k, for means = modal lengths.

Table 2. Estimated parameters[1] of normal components from the Macdonald and Pitcher analysis[2] (A - starting sigmas = 0.7, 0.7 and 0.8. B - starting sigmas = 0.6, 1.3 and 0.6.)

Component	Proportion		Mean		Sigma	
	A	B	A	B	A	B
1	0.251	0.192	18.804	18.635	0.663	0.570
Std. Errors	0.068	0.054	0.254	0.182	0.163	0.136
2	0.336	0.757	21.020	21.697	0.692	1.320
Std. Errors	0.293	0.072	0.456	0.226	0.421	0.153
3	0.413	0.052	22.805	23.521	0.870	0.258
Std. Errors	0.256	0.053	0.654	0.215	0.299	0.458

[1] Starting values for both runs: proportions = 1/k, means = modes.
[2] A:Chi-square = 5.044, P = 0.655. B:Chi-square = 5.063, P = 0.652.

Table 3. Estimated parameters[1] of normal components from the Macdonald and Pitcher analysis[2] (A = Leading 0, B = Trailing 0, C = A + B).

Component	Proportion			Mean			Sigma		
	A	B	C	A	B	C	A	B	C
1	0.237	0.231	0.220	18.75	18.73	18.70	0.628	0.625	0.598
Std. Err.	0.071	0.079	0.082	0.25	0.27	0.26	0.155	0.163	0.153
2	0.380	0.453	0.497	21.05	21.18	21.22	0.769	0.862	0.941
Std. Err.	0.364	0.384	0.425	0.60	0.65	0.73	0.527	0.616	0.698
3	0.383	0.316	0.283	22.87	23.04	23.10	0.847	0.765	0.745
Std. Err.	0.321	0.329	0.365	0.79	0.77	0.80	0.334	0.296	0.299

[1] Starting values for all runs same as example A, Table 2.

[2]

	Chi-square	P
A:	5.306	0.724
B:	5.963	0.651
C:	6.189	0.721

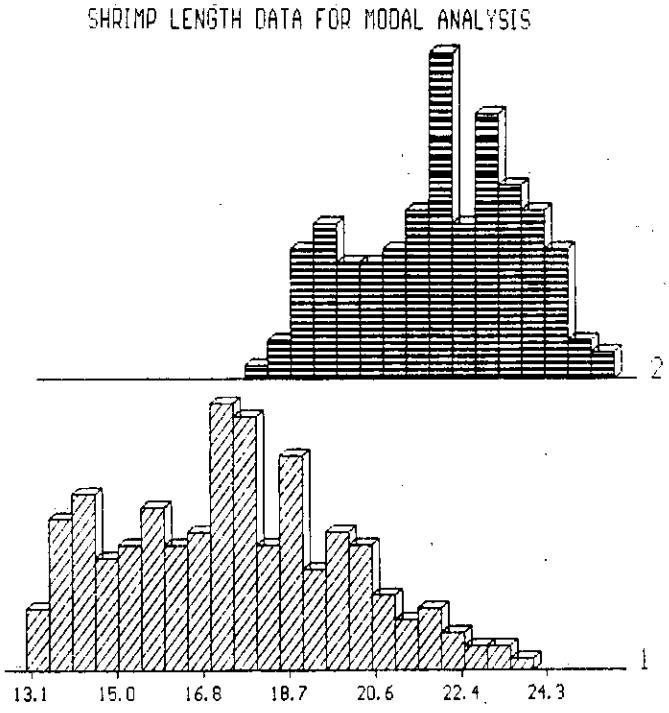

Figure 1. Shrimp length frequency data for number of components.

Figure 2. Length frequency data for different standard deviations.