Northwest Atlantic Fisheries Organization

Serial No. N4143 **NAFO SCR Doc. 99/72**

SCIENTIFIC COUNCIL MEETING – SEPTEMBER 1999
(Joint NAFO/ICES/PICES Symposium on Pandalid Shrimp Fisheries)

A Non-parametric Method of Estimating Biomass from Trawl Surveys with Monte Carlo Confidence Intervals

by

G. T. Evans, D. C. Orr, D. G. Parsons and P. J. Veitch

DFO Science Branch, P. O. Box 5667
St. John's, Newfoundland, Canada A1C 5X1
Evans@athena.nwafc.nf.ca

Abstract

The probability distribution for biomass of many marine species varies in space, partly as a function of bottom depth. We describe a non-parametric method for using trawl survey data to estimate the probability distribution at any point in the survey region whose bottom depth is known. Integrating the expected value of the distribution over the region provides an estimate of the biomass in the region. Repeated resampling from the estimated distributions at the survey points enables us to compute a Monte Carlo confidence interval for the biomass. When we apply these methods to northern shrimp in NAFO Divisions 2HJ, we obtain confidence intervals that are narrower than those computed using methods based on random-stratified sampling and an assumed Gaussian distribution.

Introduction

How should we use bottom trawl surveys to infer a possible range of values for shrimp biomass? Figures 1 and 2 show the distribution of catches during autumn surveys of NAFO Divisions 2HJ in 1998 and 1996 respectively, (a) horizontally and (b) by depth. Most good catches are within a restricted range of bottom depths. In interpolating between the survey sets, it makes sense to take account of information about bottom depth. There is, however, some reluctance to compute on the basis of the fixed strata, depth-based though they be, that are used in random-stratified estimates of groundfish abundance.

We shall assume that survey catch rates are a reliable measure of local instantaneous shrimp concentration. Nevertheless, shrimp move, so that by the time a survey is complete we would no longer expect the concentrations at the trawl set locations to be exactly what we measured there. So for practical purposes we should think not about the concentration at a point but about the probability distribution for concentration. This is the object that nature presents us with, and our data set consists of one random sample from the distribution at each set of a set of survey points. The distributions at different points are not identical; but we shall assume that distributions at nearby points are related. Using this assumption, we compute, from the data of a particular survey, an estimate of the probability distribution for shrimp abundance (mass/km^2) at any point in the region at which bottom depth is known. We then use this collection of estimated distributions in two ways. First, integrating the expected value of the distribution over a region gives an estimate of the biomass it contains. Secondly, we perform Monte Carlo simulation by resampling, at every *survey* point, from the whole probability distribution estimated at that point, to obtain a new simulated survey and thence a new abundance estimate; an ensemble of many such estimates provides a probability distribution for the estimated abundance. The whole method is called ogmap (for 'ogive mapping').

2. **Statistical Methods**

To estimate the probability distributions from a collection of relatedly but not identically distributed observations, we use the local, non-parametric methods introduced by Evans and Rice (1988). We face the following statistical question: what is the probability distribution of a random variable p (shrimp biomass density) and how does it depend on some other variable q (position)? How do we estimate it without any theory about the form either of the distribution or of the (spatial) dependence?

The cumulative distribution function (CDF) $F(p)$ is the probability that a value chosen at random will be less than p. If q were irrelevant and all p_i were independent and identically distributed, then $F(p)$ could be estimated from the data with the empirical distribution function: the fraction of the observed p_i less than p. The CDF is a step function with steps of equal height $1/n$ at each p_i, where n is the number of samples. More formally, following Davison and Hinkley (1997, eq. 2.1), the CDF is:

$$\hat{F}(p) = \frac{1}{n}\sum_{i=1}^{n} H(p - p_i)$$

where $H(z)$ is the Heaviside function: 0 for $z < 0$ and 1 for $z > 0$. When we turn to the possibility of q-dependence, we have to estimate not a single distribution but instead a separate one for each possible value of q. To compute the local influence of q on p, we generalize the idea of locally weighted estimates of expected value by kernal smoothing, assuming that the nearer an observation is to the target q, the more relevant it is for estimating the distribution at q. The estimate of the CDF is:

$$\hat{F}_q(p) = \frac{\sum H(p - p_i) w(d(q, q_i))}{\sum w(d(q, q_i))}$$

a step function whose steps heights, w, are a decreasing function of some measure d of the distance between q_i and q. (If we replace the function $H(p - p_i)$ by the number p_i, we have a local weighted estimate of the mean of the distribution at q (Davison and Hinkley, 1997, eq. 7.24). The step sizes depend only on the distances between q and the different q_i ; the step locations depend only on p. We use the weighting function $w(d) = e^{-d}$ and the distance function:

$$d^2 = \frac{(x - x_i)^2 + (y - y_i)^2}{S_h^2} + \frac{(z - z_i)^2}{S_v^2},$$

where (x, y, z) are the longitude, latitude and depth of the target point and (x_i, y_i, z_i) of the I^{th} survey point. S_h and S_v are horizontal and vertical distance scales (Davison and Hinkley (1997) refer to them as bandwidths) that describe how far local influence extends: for an increase in horizontal distance of S_h , or of vertical distance S_v, the step height decreases by a factor of $1/e$.

2.1 Choice of distance scales

It remains to choose the S_h and S_v that give as accurate a representation as possible of the probability of distributions. As is common, we use leave-one-out cross validation – delete each observation in turn, make a prediction from the rest of the data and compare it with the deleted observation. There are (at least) two things we can compute for comparison: (1) the difference between the observation and some point prediction like the mean or median of the computed distribution; (2) the (cumulative) probability of a value no larger than the observation. The observation should be a random sample from the distribution. We can't test a single number for randomness; but, if the observation is random from the distribution, its cumulative probability is uniformly distributed on [0,1]. Thus we ask if the set of all the cross-validated probabilities is U[0,1] (Rice and Evans, 1995).

This uniformity requirement is in fact more important than obtaining a small point prediction error. We are going to use Monte Carlo simulation from the computed distributions; so, if they are estimated too narrow, the simulation will overestimate the accuracy. Another way to look at it: the desire for a small squared prediction error is a matter of convenience – we hope that the estimated pdf turns out to be usefully narrow; the desire for an acceptable C^2 is a matter of correctness – we need to estimate the correct pdf, however inconvenient it turns out to be. The symptom of a distribution estimated too narrow is too many probabilities close to either 0 or 1. We therefore use a C^2 test designed specifically to detect such a pattern, based on a grouping of probabilities into 3 groups: 0-0.2, 0.2-0.8, 0.8-1.

Large scale factors lead to uniform distributions and, typically, to larger prediction errors – although the prediction error is much less sensitive than the distribution error. So we choose the narrowest bandwidths that produce an acceptable C^2. By 'acceptable' we do not mean simply that it is impossible to reject it at some stringent level like 0.95. We wish to produce our best estimate of the width of the distribution, not the narrowest one we think we can get away with. So an acceptable C^2 for this grouping with 2 degrees of freedom is not much greater than 2 (the expected value). We found that S_h = 30 km and S_v = 25 m was acceptable for each of the 1996, 1997 and 1998 bottom trawl surveys of Hopedale-Cartwright.

2.2 **Integration**

We form a network of triangles covering the region, with vertices whose bottom depths are known, derived from bottom trawl surveys combined over the last 3 years that have built up a good collection of position and depth. There is no need for the network to be regular. For each triangle, we compute the expected value of the distribution at every vertex, and then integrate the expected value for shrimp mass within a triangle using bilinear interpolation. The expected value of the biomass in the whole region is then the sum over all triangles.

3 **Application to shrimp in 2HJ** (Hopedale-Cartwright)

We applied the method to autumn bottom trawl surveys in NAFO Divisions 2HJ, which includes the shrimp fishing grounds to the Hopedale and Cartwright Channels. Trawls were made with a Campelen 1800 shrimp trawl with a lined codend.

We present results for 1998, when there are no obvious outliers among the catches (the largest catch was less than twice the third largest), and for 1996 when the two largest catches were 6.9 and 2.8 times the third largest. Maps of estimated biomass density are presented in Fig. C, and distributions of the resampled Monte Carlo biomass estimates in Fig. D. Table 1 shows the point estimates, medians, and upper and lower limits of the 95% confidence interval, both for ogmap and for the stratified-random Gaussian inferences that have been used until now for estimating these stocks. (The Monte Carlo estimates differ from those reported in Parsons *et al.* (1999), which used a preliminary guess at vertical bandwidth that was subsequently determined to be too wide.) Even for 1998, when stratified random methods seem to work well (Parsons *et al.*, 1999), the ogmap confidence intervals are smaller. This is not implausible: ogmap is not committed to gaussian distributions, and we can in principle take account of finer spatial detail than the fixed stratification can.

The key question is, of course, *do* 95% of the confidence intervals computed in this manner in fact contain the true value of total biomass? This has not yet been investigated. As a larger and longer term project, it would be rewarding to design a large class of ways the ocean might behave, and of designs for sampling it and interpreting the observations, and observing the performance.

References

DAVISON, A. C. and D. V. HINKLEY, 1997. Bootstrap Methods and Their Application. Cambridge University Press. 582 p.

EVANS, G. T. and J. C. RICE. 1998. Predicting Recruitment from Stock Size without the Mediation of a Functional Relation. *ICES J. Cons.*, **44**:111-122.

RICE, J. C. and G. T. EVANS. 1995. Ogive Mapping: a non-parametric use of spatial data. Working Paper for the ICES Cod and Climate Change Database Workshop, Nov. 1995, Woods Hole.

PARSONS, D. G., P. J. VEITCH, and G. T. EVANS, MS 1999. Resource status of northern shrimp (*Pandalus borealis*) off Baffin Island, Labrador and northeastern Newfoundland – second interim review. *Can. Stock. Ass. Sec.*, Res. Doc. No. 112, 53 p.

Table 1. The single best (point) estimate of biomass, and the median and confidence limits, for biomass of shrimp (thousands of tons) in Hopedale-Cartwright in 1998 and 1996, as described by ogmap with Monte Carlo resampling and by the traditional random-stratified calculations (strap).

	1998		1996	
	ogmap-mc	strap	ogmap-mc	strap
0.25	61	50	66	-66
0.5	84	86	153	192
point	86	86	155	192
0.975	110	121	335	451

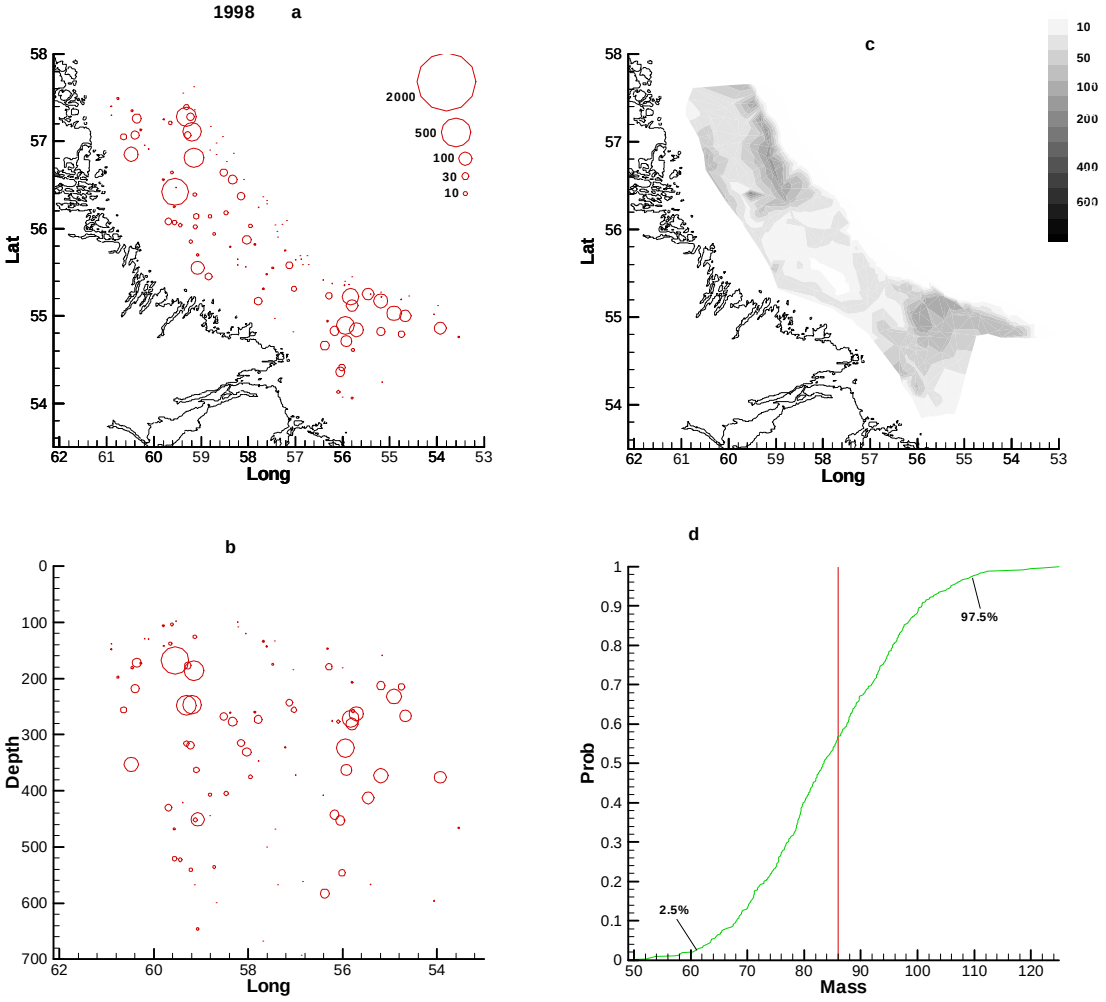

Fig. 1. 1998 survey, Div. 2HJ – Shrimp catches (kg/hr) plotted (a) with latitude and longitude; (b) with depth (m) and longitude. The symbol *area* is proportional to the catch. (c) Map of interpolated expected biomass values. (d) Cumulative distribution of Monte Carlo biomass estimates in the region. The vertical line is the single best estimate.

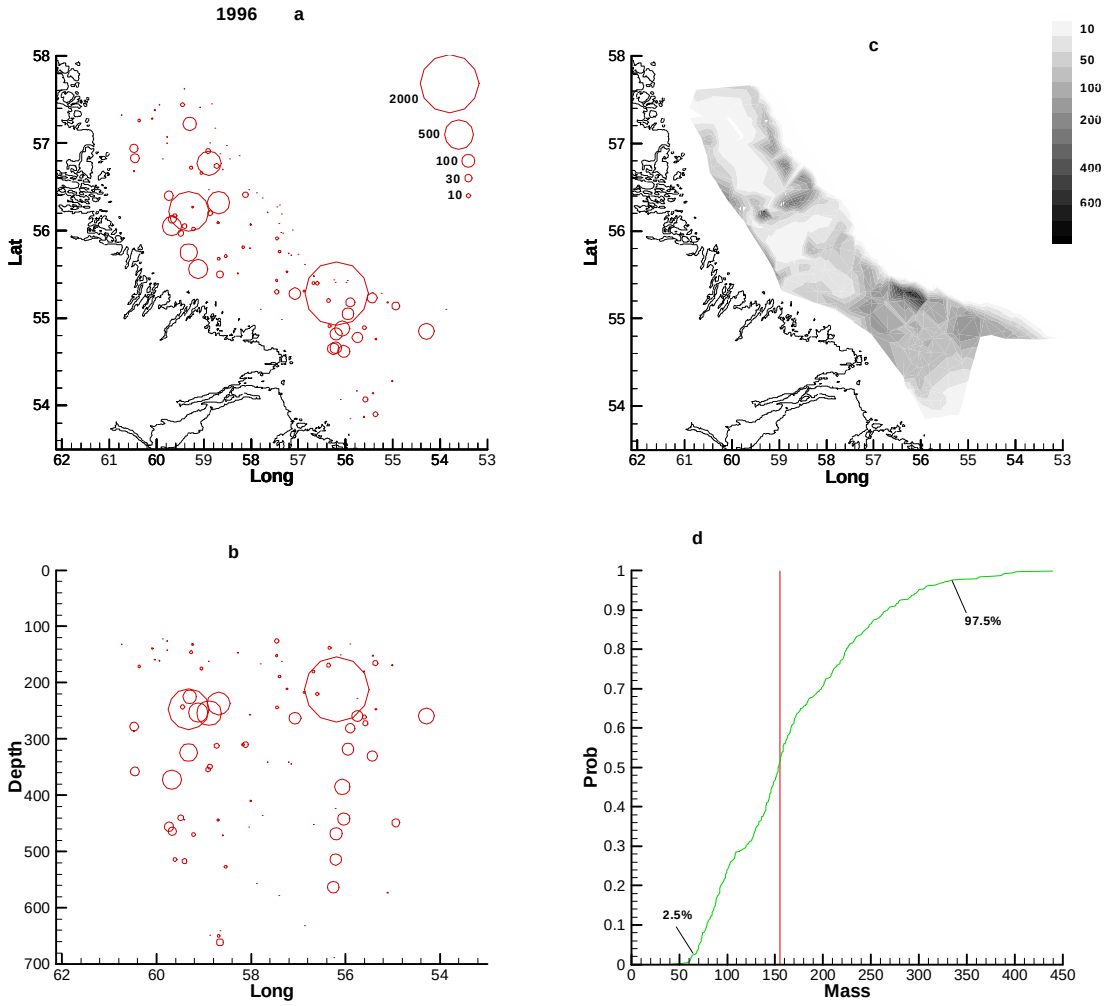

Fig. 2. 1996 survey, Div. 2HJ. Shrimp catches (kg/hr) plotted (a) with latitude and longitude; (b) with depth (m) and longitude. The symbol *area* is proportional to the catch. (c) Map of interpolated expected biomass values. (d) Cumulative distribution of Monte Carlo biomass estimates in the region. The vertical line is the single best estimate.

Afterword

A comparison of two methods for making estimates and assigning confidence intervals

G. T. Evans

Cadigan (1999) has addressed much the same question in a different way. In this afterword I present the similarities and differences between Cadigan's work and mine. The intent is not to judge either one, but to show different assumptions and procedures that may have been contemplated for much the same problem – in essence to get a better view of the problem by seeing it in two lights.

In addition to differences in assumptions, there is a huge difference in thoroughness. The thesis of Cadigan contains much work on precise justifications of certain moves, and tests by simulation, wheras I simply state thinhgs I feel are plausible. Moreover, this note is not about the whole of Cadigan (1999) but only about the points of contact with mine.

The view of the world

Both: An assumed stochastic model "generates" the data at any point in the region [3]. (Numbers in [ ] are references to pages in Cadigan, 1999.) The parameters that describe the model vary smoothly in space, but the catches produced by the model may be discontinuous in space.

The view of the task

Both: The task is to make efficient probabilistic inferences about the stock integrated over the region.

The form of the pdf for catch at a single place

Evans: no prior assumption

Cadigan: a 5-parameter distribution, the mixture of two Negative Binomials. The mean of one of the distributions is much larger than the other, and that component of the mixture is by far the scarcer.

The form of how the pdf changes with position

Evans: no prior assumption beyond smoothness.

Cadigan: The survey region is coarsely stratified into shelf and two slope regions. Within each stratum 4 of the parameters are constant; some parameters are constant across strata. No prior assumption about the fifth, the mean of the small-mean, type I distribution.

How the pdf changes with year

Evans: no relation between different years.

Cadigan: The parameters of the large-mean, type II distribution, and the mixing proportion, are constant over a 10-year period. For some applications, the Negative Binomial overdispersion parameter for the type I distribution is constant over the 10-year period.

Covariates

Evans: horizontal position and depth, by fiat.

Cadigan: Other covariates are evaluated; only horizontal position and depth are found to be useful.

How neighbouring distributions are related

Evans: kernel smoothing for the whole pdf. Two constant bandwidths, for (isotropic) horizontal position and for depth.

Cadigan: Kernel smoothing for the type I mean. Within strata, 2 constant bandwidths, for (isotropic) horizontal position and for depth. There is no influence between strata.

Kernel weighting function

Evans: double exponential; Cadigan: gaussian. The two would have different effects only when extrapolating far beyond the range of the covariates, which we both avoid.

Both: the contours of equal weight are ellipsoids.

Bandwidth selection

Cadigan: Cross-validation based on mean square prediction error or some similar concept. Bandwidths chosen on the basis of simple cross-validation tend to be too narrow, based on analysis of goodness-of-fit and residual plots; there is a decision to favour criteria that lead to wider ones. Bandwidths vary from year to year. [145]

Evans: Cross-validation based additionally on the requirement that the probabilities of the deleted observations in the predicted distributions be uniformly distributed on [0,1] – an objective way to require large bandwidths. Bandwidths are chosen to be acceptable for three years of surveys.

Both: The choice of the bandwidth has only a small effect on the estimated mass in the region, but it affects the width of the confidence intervals.

Confidence intervals

Cadigan: Approximate confidence intervals computed based on properties of the Negative Binomial distribution [221-237]. Statistical efficiency (narrow confidence intervals) is an explicit consideration.

Evans: Based on Monte Carlo resampling from the estimated pdfs at the survey points. Thus it computes the range of data sets consistent with a particular assumed truth. The real question is of course the converse: what is the range of truths consistent with a particular data set? Statistical efficiency is not explicity considered.

CADIGAN, N. G. 1999 Statistical inference about fish abundance: An approach based on research survey data. Ph. D. thesis in Statistics, University of Waterloo, Waterloo, Ontario, Canada. 268 p.