J. Northw. Atl. Fish. Sci., Vol. 27: 133–138

A Local-influence Method of Estimating Biomass from Trawl Surveys, with Monte Carlo Confidence Intervals

G. T. Evans, D. G. Parsons, P. J. Veitch and D. C. Orr
DFO Science Branch, P. O. Box 5667
St. John's, Newfoundland, Canada A1C 5X1

Abstract

The probability distribution for biomass of many marine species varies in space, partly as a function of bottom depth. A non-parametric method is described 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 sampling from the estimated distributions at the survey points enables us to compute a Monte Carlo confidence interval for the biomass. For two surveys of northern shrimp in NAFO Div. 2HJ, these methods produced confidence intervals that are narrower than those computed using methods based on stratified-random sampling and an assumed Gaussian distribution.

Key words: biomass, distribution, Monte Carlo, shrimp, trawl

Introduction

The biomass of northern shrimp off the east coast of Newfoundland and Labrador is estimated from stratified-random bottom trawl surveys carried out annually in autumn. These are multi-species surveys, and the strata have been designed to serve the needs mainly of gadoid and flatfish assessments. Standard stratified-random calculations have been used to compute unbiased estimates of the mean and variance of catches (e.g. Smith and Somerton, 1981). Confidence intervals for the integrated abundance in the stratum are computed assuming a Gaussian distribution. However, it is common to find that a few very large catches make the Gaussian assumption dubious, and lead the confidence intervals to include negative values. One might make some other assumption about the form of the probability distribution (pdf): for example that it is a delta, negative binomial or gamma distribution. However, all of these models make parametric assumptions about the probability distribution of observations larger than those observed: assumptions that are rarely well supported either by theory or by the bulk of the data. For example, in a lognormal distribution with a mode of 1, an observation of 0.01 should lead one to 'expect' a balancing observation of 100. Is this what we really think? Model-based estimates of the mean and variance can be biased if the wrong model is assumed. In another approach, Smith (1997) has produced stratified confidence intervals using bootstrapping, which replaces parametric assumptions about the pdf with the assumption that there is no possibility of values other than those observed.

An assumption common to all these estimation procedures is that the pdf is constant within each stratum. People often wish to relax this assumption as well, interpolating between the observations to obtain a smooth picture of how concentration changes with space. Such a procedure is especially attractive if the stratum boundaries are not appropriate to species being assessed. Simard *et al.* (1992) and Cadigan (1999) reported applications of continuous local weighting of observations to Atlantic Canada fisheries.

This paper extends the idea of continuous local weighting to the whole probability distribution, not just its expected value. It is driven by the view that we have no trusted statistical model either for the probability distribution for shrimp density at a point, or for how this distribution changes over space. We therefore try to be as non-parametric as we can.

Methods

A preliminary note on overloaded terms. The word 'distribution' is commonly used to refer either to a pattern in space or to a probability distribution, i.e. a pattern in the value of some arbitrary variable. In this paper, we use 'distribution' to refer only to a probability distribution; a distribution in space is called a 'spatial pattern'. The word 'density' can refer to mass of shrimp, per unit volume or area, or to a probability density. In this paper 'density' means a probability density, and mass per unit area is called 'concentration'.

The Data

Multispecies, bottom-trawl surveys have been conducted off eastern Newfoundland and Labrador annually during autumn. Since 1995, stations were sampled with a Campelen 1800 shrimp trawl. Details of the survey design and fishing protocols are given in Brodie (MS 1996). Figures 1 and 2 show the distribution of catches within Div. 2HJ, which include the shrimp fishing grounds of the Hopedale and Cartwright Channels, in 1998 and 1996, respectively, with respect to horizontal position and depth. Most good catches were within a restricted range of depths. In this analysis we assume that the distribution and spatial pattern of shrimp catches represents actual concentrations, ignoring all questions of catchability.

The Problem

Shrimp move, and the instantaneous pattern of their concentration changes with time even during a survey, and certainly well before the results of the survey can be used to guide decisions. The instantaneous pattern is one realization of a random process; and we take the view that the underlying process, the spatial pattern of probability distributions for concentration, does not change during the survey. The task is to estimate this stable pattern of probability distributions.

We use the collection of estimated probability distributions in two ways. First, integrating the expected value of the distribution over a region gives an estimate of the biomass it contains. Secondly, a Monte Carlo simulation that resamples, at every *survey* point, from the whole probability distribution estimated at that point, provides 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 here called "ogmap" (for 'ogive mapping'; Rice and Evans, MS 1995).

The Statistical Model

What is the probability distribution of a random variable *p* (shrimp concentration) and how does it depend on some covariate *q* (position)? Note in passing that this is also the problem posed by ordinary linear regression, which answers it by saying: "The distribution is Gaussian and its variance is constant and its mean is a straight-line function of *q*". In this paper we estimate an answer with much less theory about the form either of the distribution or of its spatial pattern (it is not possible to operate without any theory at all). The trawl sets are assumed to be independent random samples from the probability distributions at the locations of the sets. They are not identically distributed because of the spatial pattern of the distributions; but to make any progress at all we assume that nearby distributions are related.

We use the local, non-parametric methods introduced by Evans and Rice (1988) and given their fullest description in Evans (MS 2000). The cumulative distribution function (CDF) *F*(*p*) is the probability that a value chosen at random will be less than *p*. If the covariate *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, equation 2.1), the CDF is estimated by:

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

where *H(z)* is the Heaviside function: 0 for $z < 0$ and 1 for $z > 0$. To incorporate *q*-dependence, we replace the equal step heights 1/*n* by local weights based on kernel 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 then:

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

a step function whose steps heights, *w*, are a decreasing function of some measure *d* of the distance between q_i and *q*. (Notice that the mean of $\hat{F}_q(p)$ is obtained by replacing the function $H(p - p_i)$ by the number p_i, and that this is precisely the formula for kernel smoothing to estimate the mean (Davison and Hinkley, 1997, equation 7.24). Thus the definition of $\hat{F}_q(p)$ is an almost inevitable combination of two standard ideas. The step *sizes* depend only on the distances between *q* and the different q_i; the step *locations* depend only on p_i.

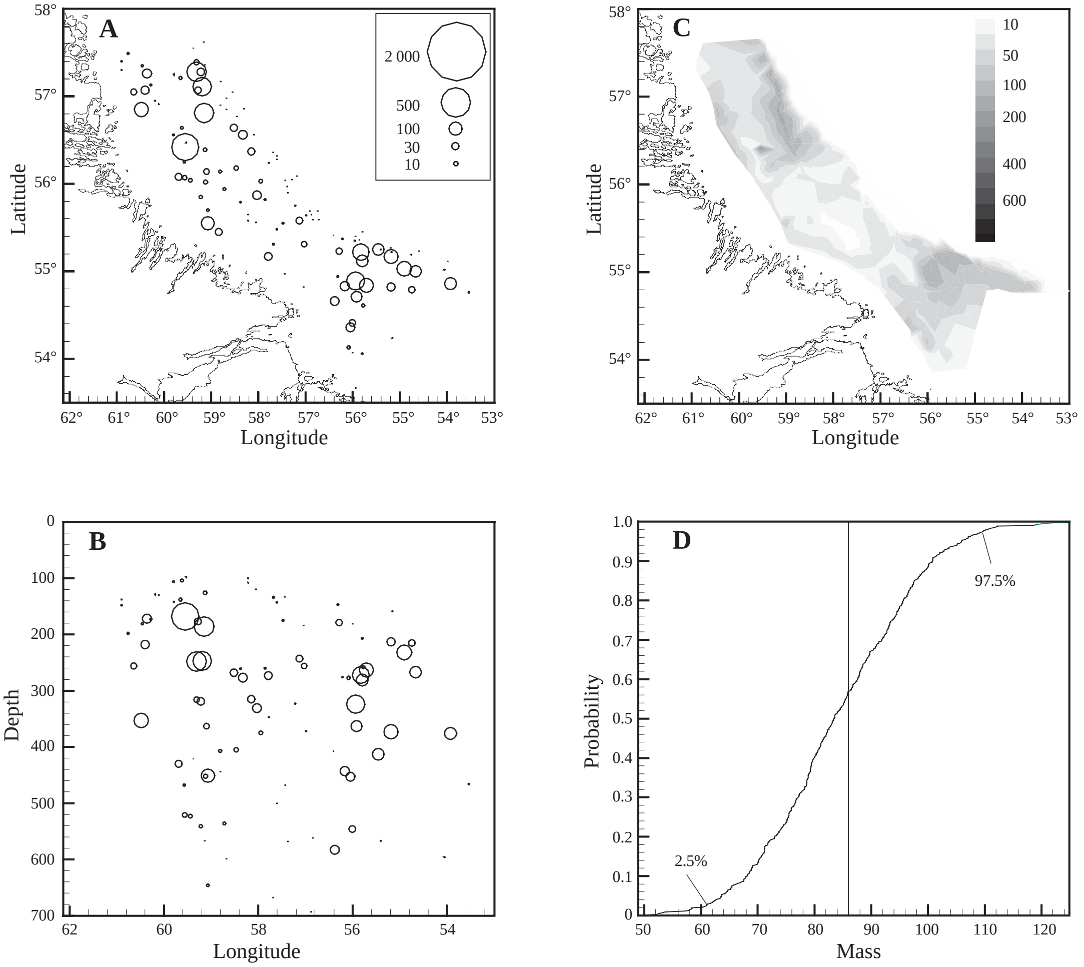

Fig. 1. (**A**) Catches of shrimp in Div. 2JH in 1998. The area of the symbol is proportional to the catch in the set located at the center of the symbol. (**B**) The catches of panel (A) plotted as a function of depth and longitude. (**C**) Map of the estimated mean value of the probability distribution for catch as a function of space. (**D**) The Monte Carlo distribution for expected mass integrated over the region.

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, or 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 $1/e$.

Choice of bandwidths. It remains to choose the S_h and S_v that give as accurate a representation as possible of the probability distributions. As is common, we use jackknife cross validation, which deletes each observation in turn, predicts it from the rest of the data, and compares the prediction with the deleted observation. There are (at least) two measures of performance: (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. A single number cannot be tested for randomness; but, if the observation is

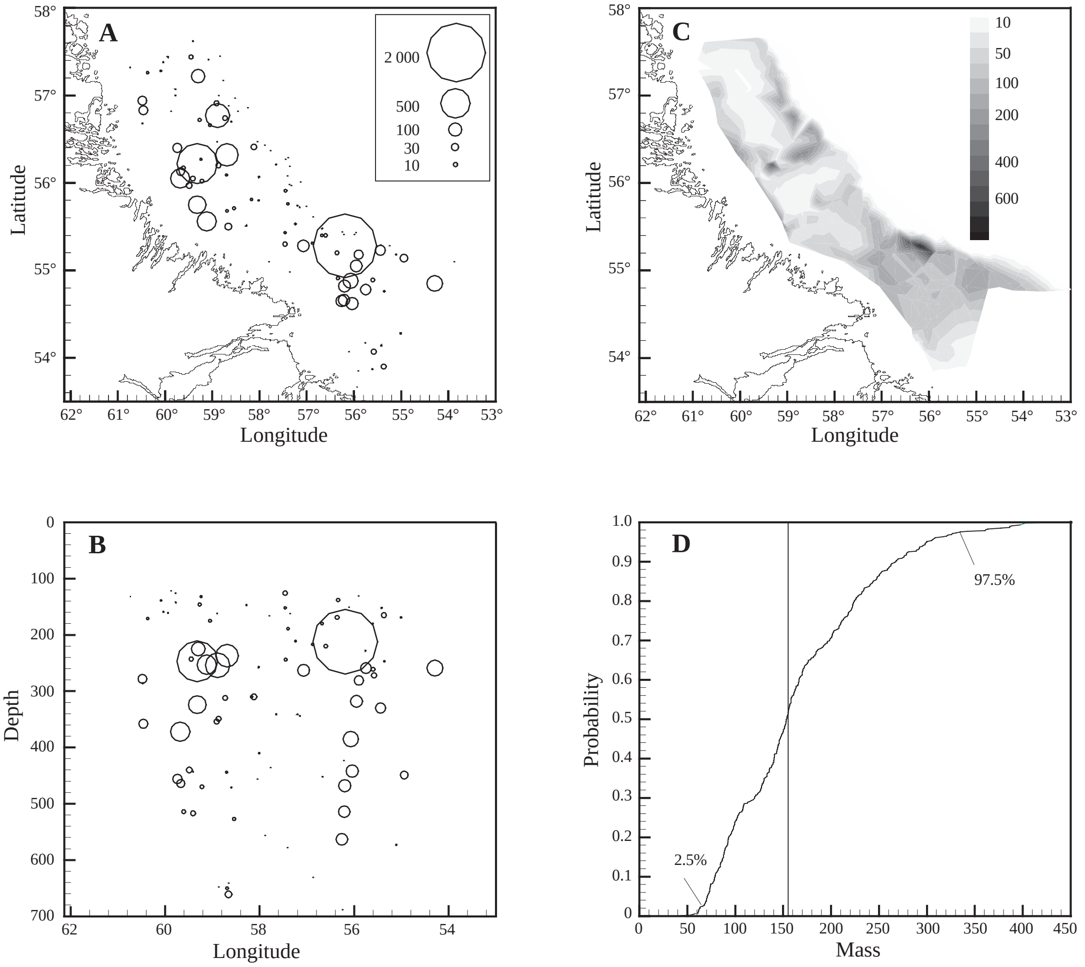

Fig. 2. (**A**) Catches of shrimp in Div. 2JH in 1996. The area of the symbol is proportional to the catch in the set located at the center of the symbol. (**B**) The catches of panel (A) plotted as a function of depth and longitude. (**C**) Map of the estimated mean value of the probability distribution for catch as a function of space. (**D**) The Monte Carlo distribution for expected mass integrated over the region. The abscissa in Fig. 2D extends to higher values, reflecting the uncertainty due to the two very large catches.

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, MS 1995).

This uniformity requirement is in fact more important than obtaining a small point prediction error. If the variances of the distributions are underestimated, then Monte Carlo simulation using them will overestimate the accuracy of estimates. 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 χ^2 is a matter of correctness – we need to estimate the correct pdf, however inconvenient it turns out to be. The uniformity requirement provides a guard against the risk of over-fitting if we tried to match the predicted central value alone.

The symptom of underestimating the width of the distribution is that too many probabilities are close to either 0 or 1. We therefore use a χ^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 bandwidths 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 χ^2. By 'acceptable' we do not mean simply that it is impossible to reject it at some stringent probability level like 0.95. We need to produce our best estimate of the width of the distribution, not the narrowest one we think we can get away with. So instead of accepting any χ^2 less than 6 (the 0.05 value with 2 degrees of freedom), we accept only those values not much greater than 2 (the expected value). We found that S_h = 30 km and S_v = 25 m was acceptable for the 1996 and 1998 bottom trawl surveys of Hopedale-Cartwright.

Integration. We cover the region with a Delaunay triangulation from vertices whose positions and depths were measured during bottom trawl surveys in 1996–98. 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.

Results

Results for shrimp in Div. 2HJ (Hopedale-Cartwright) are presented for 1998 (Fig. 1), when there were no unusually large values among the catches (the largest catch was less than twice the third largest), and for 1996 (Fig. 2), when the two largest catches were 6.9 and 2.8 times the third largest. Maps of estimated biomass density are presented in Fig. 1c and 2c, and distributions of the re-sampled Monte Carlo biomass estimates in Fig. 1d and 2d. 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. The Monte Carlo estimates differ slightly from those reported in Parsons *et al.* (MS 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.,* MS 1999), the ogmap confidence intervals are smaller. This is not implausible: ogmap is not committed to Gaussian distributions, and it can in principle take account of finer spatial details (more information on covariates) than the fixed stratification.

Discussion

The intuition behind using a continuous approach rather than rigid stratum boundaries is that a location near a stratum boundary ought to be more like a nearby (taking depth into account) observation in an adjacent stratum than like an observation at the other end of its own stratum. Particularly for shrimp surveyed under a groundfish stratification, this intuition seems appropriate. However, if observations are assigned to some strata out of proportion of their area (because of a wish to have at least two observations per stratum for variance calculations, or deliberately oversampling strata known to have high variance), then an analysis that does not confine the influence of observations within stratum boundaries can be biased.

Our approach differs from kriging, which sees its task as estimating the particular realization of the random process. Kriging makes different assumptions and asks a different question. The variance of the difference between two observations is assumed to be a function only of their (possibly vector) separation. When there are large regions where the shrimp catch, and therefore the variance between nearby catches, is predictably zero, and other regions of moderate and occasionally high catch with high variance between nearby stations, we would not wish to make this assumption of intrinsic stationarity (Bailey and

TABLE 1. The single best (point) estimate of biomass, and the median and confidence limits, for biomass of shrimp ('000 tons) in Hopedale-Cartwright in 1998 and 1996, as determined by ogmap with Monte Carlo resampling and by the traditional random-stratified calculations (strap).

	1998		1996	
	ogmap-mc	strap	ogmap-mc	strap
0.025	61	50	66	-66
0.5	84	86	153	192
point	86	86	155	192
0.975	110	121	335	451

Gatrell, 1995, p. 162). The objective in kriging is to estimate a particular realization of a stochastic spatial process: "to add a local (error) component to our prediction … in addition to the mean" (Bailey and Gatrell, 1995, p. 183). We believe that the particular realization is not stable enough to be worth estimating; only the spatial pattern of the probability distribution is of interest.

The approach of Cadigan (1999) is in some ways similar to ours. He also assumes that there is an underlying spatial pattern in pdf to be estimated, and that the trawl sets are independent random samples from related distributions. The main difference is that he takes a parametric form for the pdf and assumes that all but one of the parameters are constant over large areas, whereas we allow everything about the pdf to vary spatially. Cadigan (1999) also assumes that there is a small probability of large catches that is constant, not only within a survey but also between surveys in different years. Under this assumption he gets a more reliable handle on the occasional large catches – worthwhile if true. (As with all parametric assumptions, more information put in, if it is true, leads to more coming out.)

There is *no* right way to treat very few very large catches. If large catches are influential for estimates of total abundance (and in many fisheries studies they are), then an accurate estimate of abundance depends on an accurate knowledge of the pdf of such catches. It is not possible to get such an accurate knowledge from very few observations, no matter how clever we are.

It would be possible to include covariates other than position and depth, if they were known at enough points to make the triangulation and integration possible. Bottom type is a good candidate (and is being investigated now for sessile species). Bottom temperature is not, because it is not known at points that were not surveyed.

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.

References

BAILEY, T. C., and A. C. GATRELL. 1995. Interactive Spatial Data Analysis. Longman, 413 p.

BRODIE, W. MS 1996. A description of the 1995 fall groundfish survey in Divisions 2J3KLNO. *NAFO pCR Doc.*, No. 27, Serial No. N2700, 7 p.

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.

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

EVANS, G. T. MS 2000. Local estimation of probability distribution and how it depends on covariates. *Can. ptock. Ass. Res. Doc.*, No. 120, 11 p.

EVANS, G. T., and J. C. RICE. 1988. Predicting recruitment from stock size without the mediation of a functional relation. *ICEp J. Cons.*, **44**: 111–122.

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. ptock. Ass. Res. Doc.*, No. 112, 53 p.

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

SIMARD, Y., P. LEGENDRE, G. LAVOIE, and D. MARCOTTE. 1992. Mapping, estimating biomass, and optimizing sampling programs for spatially autocorrelated data: case study of the northern shrimp (*Pandalus borealis*). *Can. J. Fish. Aquat. pci.*, **49**: 32–45.

SMITH, S. J. 1997. Bootstrap confidence limits for groundfish trawl survey estimates of mean abundance. *Can. J. Fish. Aquat. pci.*, **54**: 616–630.

SMITH, S. J., and G. D. SOMERTON. 1981. STRAP: a user-oriented computer analysis system for groundfish research trawl survey data. *Can. Tech. Rep. Fish. Aquat. pci.*, **1030**, 66 p.