Fisheries and Oceans
Science

Pêches et Océans
Sciences

C S A S

Canadian Stock Assessment Secretariat

S C É S

Secrétariat canadien pour l'évaluation des stocks

Research Document 2000/120

Document de recherche 2000/120

Not to be cited without
permission of the authors [1]

Ne pas citer sans
autorisation des auteurs [1]

Local estimation of probability distribution and how it depends on covariates

Geoffrey T. Evans

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

evans@athena.nwafc.nf.ca

[1] This series documents the scientific basis for the evaluation of fisheries resources in Canada. As such, it addresses the issues of the day in the time frames required and the documents it contains are not intended as definitive statements on the subjects addressed but rather as progress reports on ongoing investigations.

Research documents are produced in the official language in which they are provided to the Secretariat.

This document is available on the Internet at:

[1] La présente série documente les bases scientifiques des évaluations des ressources halieutiques du Canada. Elle traite des problèmes courants selon les échéanciers dictés. Les documents qu'elle contient ne doivent pas être considérés comme des énoncés définitifs sur les sujets traités, mais plutôt comme des rapports d'étape sur les études en cours.

Les documents de recherche sont publiés dans la langue officielle utilisée dans le manuscrit envoyé au Secrétariat.

Ce document est disponible sur l'Internet à:

http://www.dfo-mpo.gc.ca/csas/

ISSN 1480-4883
Ottawa, 2000
Canada

Abstract

The kernel smoothing method for obtaining a locally weighted estimate of mean value is extended to produce a locally weighted estimate of the whole probability distribution function. This has several advantages in analysing data sets where there exists no trusted theory. For example, it automatically provides the basis for performing Monte Carlo simulations. Examples of applications to several areas of fisheries science are provided.

Résumé

La méthode d'adoucissement à base de noyau central, destinée à l'estimation d'une moyenne à pesage local, est portée plus loin afin de produire un estimé à pesage local de la distribution de probabilité des observations toute entière. Ceci a plusieurs avantages pour l'analyze des données oû il n'existe aucune théorie de base. Par exemple, cette methode donne automatiquement une base pour l'exécution des simulations Monte Carlo. Plusieurs exemples de l'application de cette approche dans divers domaines de la science des pêches sont présentés.

1 Method

1.1 The question

What is the probability distribution of a random variable y, and how does it depend on another variable x? This is not simply a question about expected values: if $E(y)$ is independent of x, but the variance of y, or the probability of exceeding some important value, depends on x, then we want to know about that. It sometimes happens that x is a variable we can influence, in which case the question can be phrased as: What is the risk and what can we do about it?

The current climate of fisheries management asks us to respect uncertainty, and one thing we would be extremely uncertain about is any parametric theory either about the form of the probability distribution (such as gaussian or lognormal or delta distribution) or about how its expected value depends on x (such as straight line or Beverton-Holt hyperbola) or about how anything other than its expected value depends on x (a common assumption being that it doesn't). So we shall address the above question given a data set (collection of pairs $\{(x_i, y_i), i = 1 \ldots n\}$) and a reluctance to make additional structural or parametric assumptions.

The first thing to say about the task ahead is that it will be done badly. If we must rely on the data to tell us about the whole form and not just a few numerical parameters, then we are probably asking for more information than a small number of observations contain. Be it so: this is often the question that the world asks of us. We could ask an easier question, but if the assumptions behind it were not warranted then any gain in precision would be spurious (like looking for a wallet where the light is good instead of where it was dropped.) Instead we do the best we can with the data at hand, and try to indicate how bad this might be.

Ordinary linear regression also addresses our basic question, but it imposes the answer: "The distribution of y is gaussian and its variance is constant and its mean is a straight line function of x." This inserts a lot of information apriori, leaving only three numbers (the variance, and the intercept and slope of the line) for the data set to determine. Believing that the expected value can be described by a function of two or three parameters entails believing that observations at one extreme of the range of values of x can be useful for constraining our estimate of the distribution at the other extreme. This is a very powerful assumption, and its power ought to make us reluctant to accept it without strong reasons for believing it to be true; it should hardly be the default assumption.

1.2 The local algorithm

The technique is assembled from well-known pieces. The (cumulative) probability distribution $F(y)$ is the probability that a value chosen at random will be less that y. Suppose for a moment that x is irrelevant, that all y_i are independent and identically distributed. Then one can estimate $F(y)$ from the data with the empirical distribution function: the fraction of the observed y_i less than y. The empirical distribution has equal probabilities n^{-1} at each sample value y_i, so that the cumulative distribution function (ogive) is a step function with steps of equal height at each y_i. More formally, following Davison and Hinkley (1997, eqn 2.1), the EDF is

$$\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} H(y - y_i) \tag{1}$$

where $H(z)$ is the Heaviside function: 0 for $z < 0 \quad [y < y_i]$ and 1 for $z > 0$.

Intuitively, it seems reasonable that observations made near to x should be especially influential in determining our estimate of the distribution at x. For example, think of estimating the mean value of $y(x)$ by a weighted average of the observed y_i, with the weights being greater when x_i is closer to x, so-called kernel smoothing:

$$\hat{\mu}(x) = \frac{\sum y_i w\{(x - x_i)/b\}}{\sum w\{(x - x_i)/b\}} \tag{2}$$

where w is a decreasing function of the absolute value of its argument and b, the bandwidth, describes how far the local influence extends (Davison and Hinkley 1997, eqn 7.24).

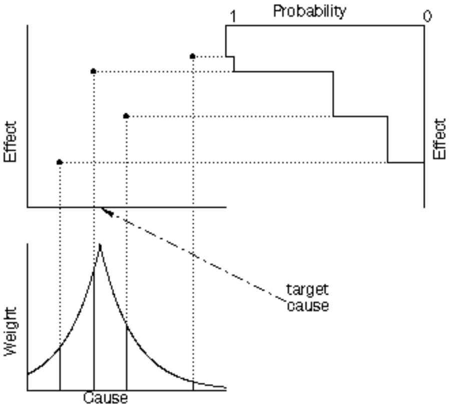

Figure 1.

Building a probability distribution. Start in the upper left with the data and a target cause for which we want to know the probability distribution of the effect. Find the weight of each datum, a decreasing function of the distance between its cause and the target cause (lower left). Use this weight as the size of the step in the cumulative distribution function at the corresponding effect (upper right, rotated 90°).

To estimate the whole probability distribution and not just its expected value, we make the natural extension in which the weights in (2) also determine the heights of the steps at each y_i in the empirical distribution (1). Thus the estimate is

$$\hat{F}_x(y) = \frac{\sum H(y - y_i) w\{(x - x_i)/b\}}{\sum w\{(x - x_i)/b\}} \tag{3}$$

so that the ogive is a step function whose steps are of different heights (Figure 1.)

This approach was first used by Evans and Rice (1988). Equation (2) is always a particular consequence of (3); (1) is a special case of (3) for $w \equiv 1$. The x and y variables are somewhat decoupled: the step sizes depend only on the distances between x and the different x_i; the step locations depend only on y_i. This provides a sort of invariance under transformation of y. There is no basic difference in the method if x is multidimensional with different bandwidths for different directions.

1.3 Local influence: shape

Two common choices for the step height function, $w(d)$, where $d = |x_i - x|/b$, are e^{-d} and $1/(1+d^2)$. If $b = \infty$ then $d = 0$ and all the step heights are the same: the identically distributed limit when x has no effect. It is often stated that the choice of w is relatively unimportant and the only thing that really matters is the bandwidth b. This may be true within the range of the data; but in practice, like it or not, we are always called on to go beyond the data, and different w's have different extrapolation properties. Consider the behaviour of $w(d_1)/w(d_2)$, where we suppose for definiteness that $d_1 < d_2$. When $x \to \infty$, the step size ratio for the exponential form is identically $e^{(x_1-x_2)/b}$. Thus when the target point moves beyond all the data and we are in the pure extrapolation realm, the probability distribution is frozen at the values it had when it passed the last data point. (True only for 1-dimesional x.) For the cauchy form, the ratio → 1 as $x \to \infty$: far from all the data they all look the same. A form that initially seemed plausible for the step height function was the gaussian e^{-d^2}; but for this form the influence of the closest point grows to dominate the distribution as $x \to \infty$. This would say that, the further y gets from the data, the smaller the variance in its estimated distribution becomes. This seems like a good reason not to use that form or, for similar reasons, forms that become zero after some finite d, and therefore make no prediction at all when x goes too far beyond the data.

1.4 Local influences: bandwidth

We use cross-validation to choose the bandwidth: delete each observation in turn; use the rest of the data to compute the pdf at the deleted point; observe the match between the pdf and the deleted observation; compute a measure of performance for the match, integrated over all the observations; choose the bandwidth for which the measure of performance is best. This entails a comparison between a function and a number. One way to make the comparison is to represent the function by a single number, such as its mean or median; the measure of performance could be the root mean square difference between the mean and the deleted observation, or the mean absolute difference between the median and the deleted observation. However, this measure considers only the central tendency of the distribution, and does not address the need to estimate the whole distribution. Another measure is based on the idea that each observation is a random sample from the distribution appropriate to its covariate, and the probability of the observed value in the distribution should therefore be uniformly distributed on [0,1] (Rice and Evans 1995). (If we take $b = \infty$ then the distribution would be uniform by definition, except for complications when there are repeated y values.)

1.5 Monte Carlo simulations and confidence intervals

The data set we have is only one of many that might have been generated by the true underlying random process. If we repeatedly resample from the random process, we get a collection of possible data sets. We claim to have estimated the distribution at each x, and in particular at each x_i; so, we are in a position to simulate the resampling and do Monte Carlo simulations. (This is another payoff for doing the extra work to get the whole distribution and not just its expected value, in addition to intrinsic scientific interest in the rest of the distribution.) From these simulations we can derive a whole probability distribution, including any confidence intervals we care about, for any quantity we care to compute. (This would amount to bootstrapping the residuals if they were identically distributed, and to bootstrapping the observations themselves if the bandwidth were infinite.)

This follows time-honoured statistical practice in asking the converse of the right question. Resampling estimates the distribution of data sets, and summary quantities, consistent with some truth. But we aren't given a truth; we have a data set and we want to know the distribution of truths consistent with it. Suppose, for example, that the data set contains one extreme value, which has a big effect on the computed mean. Some of the resampled data sets will not contain the extreme value, and they will make their computations on the basis that it cannot occur, which we know to be false.

Notice the importance of satisfying the test for uniformity of probabilities. It might be possible to get a good (unbiassed) measure by taking b close to zero, so that (with exponential weighting) the prediction was effectively the nearest observation. This has a variance in $y_i - y$ equal to twice the variance of a single observation. But the probability distribution asserts near certainty, the Monte Carlo resamples will be almost the same as the original sample, and the intrinsic variability would be underestimated. We detect this problem by seeing that the probabilities of the observed values are all estimated to be 1 or 0 with no intermediates.

2 Examples

I describe a number of problems in fisheries where local estimation of probability distribution has been useful. Each example also illustrates an extra point in the development of the methods, which it seemed easier to discuss in the context in which it arose rather than in the abstract description above.

2.1 Fish recruitment as influenced by spawning stock size

A precautionary approach to fish management entails addressing and embracing uncertainty – including uncertainty about the functional form of relationships – in trying to minimize the risk of poor recruitments (Rice and Evans 1988, Shelton and Morgan 1994). Figure 2a shows spawning stock biomass (SSB) and recruitment for cod off the south coast of Newfoundland (NAFO 3Ps), and the estimated median and 10th and 90th percentiles as functions of SSB. ($w(d) = e^{-d}$, b chosen to minimize cross-validated sum of squared deviations from the mean.)

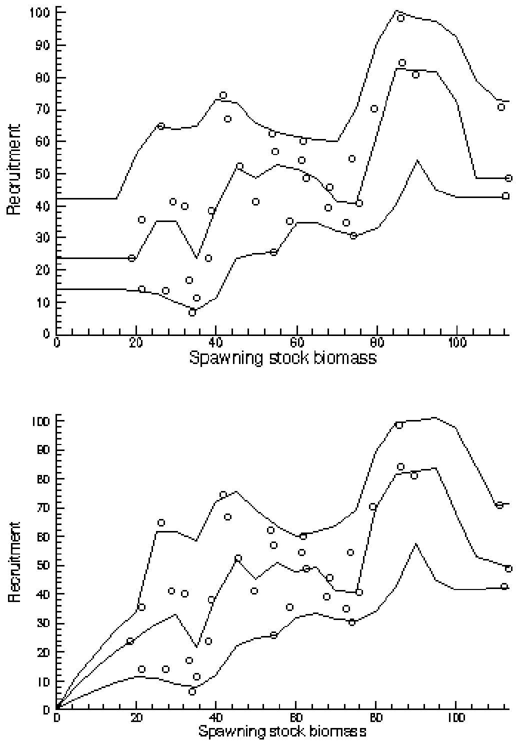

Figure 2.

Recruitment probability distributions. The squares denote observed spawning stock biomass and recruitment for the 3Ps cod stock. Lines denote, from top to bottom, the recruitment that is exceeded with probability 0.1, 0.5 and 0.9. (a) The data are untransformed. (b) The data are transformed to be multiples of the best Beverton-Holt fit and then analysed according to the methods described here; the lines are transformed back to the original units.

A weakness of this analysis is that the probability of a given recruitment remains constant between the smallest observed SSB and 0, whereas we believe that recruitment goes to 0 as SSB goes to 0. One way around this problem is to say that there are certain things we know about recruitment: it goes to zero when stock goes to zero and remains bounded as stock goes to ∞; so, let us build a way to be parametric about what we know while remaining non-parametric about our ignorance: the details of what happens in between (where there are data). We can fit a parametric curve – say a Beverton-Holt hyperbola – to the data, create a new data set consisting of the ratios of observed recruitments to the recruitments predicted for the corresponding stock sizes, and compute the probability distribution locally for this transformed data set, and back-transform it to the original recruitment variable (Fig 2b). The curves in the middle of the data can still deviate from any parametric form as the data dictate; while beyond the data, where only our prejudices can dictate, the curves follow the dictates of our prejudice.

If it happened that the distribution around the residuals was constant, then the best estimated b for the residual data would probably be large — perhaps greater than the range in the data. Conversely, if b were still estimated to be small compared to the range in the data, then this would be evidence that the pattern in the distribution was not simply a pattern in its expected value.

2.2 Nucleic acid ratios in larval fish

The ratio of RNA to DNA in cells is an indication of cell activity, and might indicate overall condition and survival probability. However, condition might have little to do with survival if predators take good- and poor-condition larvae indiscriminately. If condition is important for survival, then we might expect to see a removal of the lowest tail of the probability distribution for condition as larval age increases. The ecological effect of interest concerns the possible change in scatter (say the difference between the 10^{th} and 90^{th} percentile) with age, and not the mean (the average larva is dead).

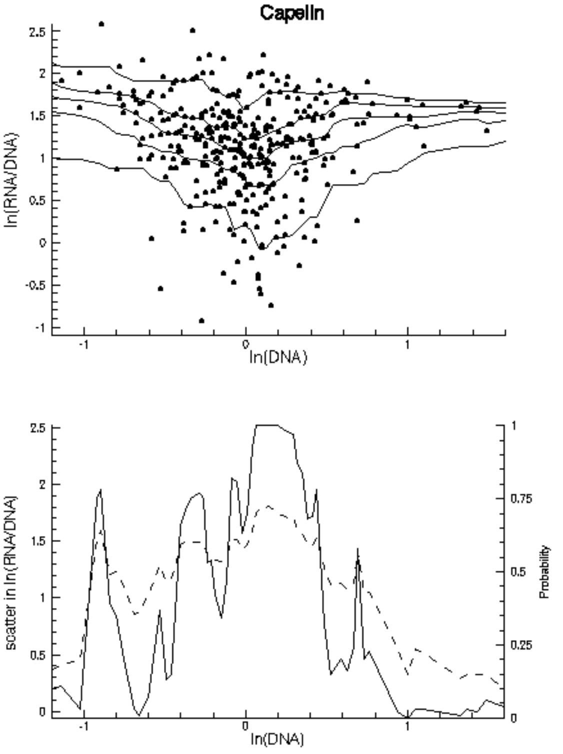

Figure 3.

Condition and survival of larval capelin. The abscissa is a rough measure of age; the ordinate indicates condition. Lines denote the condition that is exceeded with probability 0.1, 0.3, 0.5, 0.7 and 0.9. In the bottom panel, the dashed line is the difference between the 0.1 and 0.9 lines of the top panel; the solid line (right-hand scale) is the probability of getting a difference smaller than that if deviations from the modelled median are assigned at random. Thus the scatter around 0.2 is significantly larger than average, that above 1 is significantly less than average.

We seek evidence for a statistically significant change in scatter, and use a randomization test of significance. (Randomizing is technically easier than bootstrapping here; and Romano (1989) suggests that the answers are probably at least as good.) As in the previous example, we have reason for analysing a data set of residuals. If there is a trend, then randomizing the raw data will produce a spuriously large scatter. We therefore first compute the locally estimated median and randomize the residuals from that. We find (Fig 3) that the scatter for larval capelin is significantly higher than average for young larvae and significantly lower than average for old larvae, suggesting that poor-condition larvae do indeed have a lower probability of survival (Pepin et al. 1999).

2.3 Shrimp biomass estimates from trawl surveys

The data come from stratified random surveys whose strata were designed for other species. It is desired to use these data to map shrimp distributions. The x variable is multidimensional: latitude, longitude, bottom depth; we choose a partially isotropic b with one bandwidth for horizontal distance and another for bottom depth. The presence of occasional very large catches makes the gaussian assumption, and confidence intervals derived from it, untenable. The large proportion of zeros and other fixed small values make the test statistic for uniformity more delicate to compute.

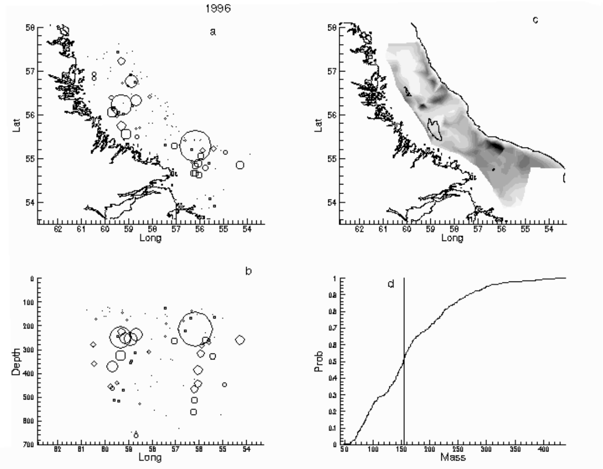

Figure 4.

Raw data from shrimp surveys off labrador, plotted as a function of (a) horizontal position and (b) depth. (c) Map of expected value of mass density. (d) Monte Carlo probability distribution for total mass in the region.

Figures 4a and 4b show the raw data from 2 different perspectives. The data are dominated by two large catches; this makes it hard to do good science, which depends on good regularities. Having said that, we still try to do the least silly things we can. It is, for example, silly to try to predict the biggest sets by cross-validation, and a sum of squares criterion might lead us to distort what we can fit in a vain search for what we cannot (compare non-robust and robust regression). One approach is to pay attention to probabilities instead of values. Another approach is to modify the data set *for cross-validation, bandwidth-choosing purposes only.* Make a new, censored data set in which the outliers are replaced by values which are still the largest in the data set but no longer outliers; use this data set to estimate b based on cross-validated sum of squares; then analyse the real data set with this b.

Figure 4c shows the map of estimated mean value of the probability density; it also indicates the 600 m depth contour. Figure 4d shows the Monte Carlo distribution of integrated abundance estimates (Evans et al. 1999).

Kriging is another non-parametric technique that is sometimes used to make maps and infer abundance from data like these. It 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 is predictably zero and so is the variance between nearby catches, and other regions of moderate and occasionally high catch with high variance between nearby stations, one would not wish to make this assumption of intrinsic stationarity (Bailey and 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). For a species like shrimp that will move a lot between the time of the survey and the time when management decisions are made, I suggest that the particular realization is not stable enough to be worth estimating; what interests us is the spatial pattern of the probability distribution.

2.4 First indications of a year-class

When a year-class is first seen, it is often with a relatively noisy index (and at the very least it is with only one index rather than several successive years'). If it comes out especially high, was this because the year-class was especially large, or especially catchable (available to that year's survey)? From past data we can form a probability distribution both for the year-class strength (possibly making use of the methods of §2.1) and for the catchability and then, if we assume catchability is independent of stock size, form a Bayes update of the prior distribution for recruitment.

In more detail: let r be a recruitment random variable, i a population index random variable, and q a catchability random variable. Then Bayes rule states that

$$\mathrm{P}(r = R \mid i = I) = \frac{P(i = I \mid r = R)\,\mathrm{P}(r = R)}{\mathrm{P}(i = I)}.$$

Now, if $r = R$ then the statements $i = I$ and $q = I/R$ are the same. Thus

$$\mathrm{P}(i = I \mid r = R) = \mathrm{P}(q = I/R \mid r = R).$$

If we further assume that catchability is independent of recruitment, then

$$\mathrm{P}(q = I/R \mid r = R) = \mathrm{P}(q = I/R)$$

and

$$\mathrm{P}(r = R \mid i = I) = \frac{\mathrm{P}(q = I/R)\,\mathrm{P}(r = R)}{\mathrm{P}(i = I)}.$$

These calculations apply just the same whether the P's are probabilities or probability densities. But the computations will not work with the finite probabilities (step function ogives) described in equation (3), because almost certainly there will never be a new index for which values of $q = I/R$ and $r = R$ both appeared in the past data set. Fortunately, we don't have to work with the raw empirical distribution; there are non-parametric ways of smoothing it. For example, we can build a piecewise linear ogive by joining the midpoints of each step by straight line segments, and extending the first (last) segment to 0 (1).

In Monte Carlo simulation with infinite b, a piecewise linear ogive leads to a generalization of bootstrapping that may or may not already have been invented.

References

Bailey, T.C. and A.C. Gatrell. 1995. Interactive Spatial Data Analysis. Longman. 413pp.

Davison, A.C. and D.V. Hinkley. 1997. Bootstrap Methods and their Application. Cambridge U. Press. 582pp.

Evans, G.T., D.C. Orr, D.G. Parsons and P.J. Veitch. 1999. A non-parametric method of estimating biomss from trawl surveys with Monte Carlo confidence intervals. NAFO SCR Doc. 99/72, Serial N4143. 8pp.

Evans, G.T. and J.C. Rice. 1988. Predicting recruitment from stock size without the mediation of a functional relationship. J. Cons. int. Explor. Mer. 44:111-122.

Pepin, P., G.T. Evans and T.H. Shears. 1999. Patterns of RNA/DNA ratios in larval fish and their relation to survival in the field. ICES J. mar. Sci. 56:697-706.

Rice, J.C. and G.T. Evans. 1988. Tools for embracing uncertainty in the management of the cod fishery of NAFO divisions 2J+3KL. J. Cons. int. Explor. Mer. 45:73-81.

Rice, J.C. and G.T. Evans. 1995. Ogive mapping: a nonparametric use of spatial data. Working paper for the ICES Cod and Climate Change Database Workshop. Nov 1995, Woods Hole.

Romano, J.P. 1989. Bootstrap and randomization tests of some nonparametric hypotheses. Ann. Stat. 17:141-159.

Shelton, P.A. and M.J. Morgan. 1994. An analysis of spawner biomass and recruitment of cod (*Gadus morhua*) in Divisions 2J and 2KL. NAFO Sci. Counc. Stud. 21:67-82.