ICES Journal of Marine Science, 54: 741–752. 1997

Prediction of year-class strength by calibration regression analysis of multiple recruit index series

J. G. Shepherd*

Shepherd, J. G. 1997. Prediction of year-class strength by calibration regression analysis of multiple recruit index series. – ICES Journal of Marine Science, 54: 741–752.

The analysis of multiple time series of indices of recruitment to fish stocks by means of calibration regression is discussed, together with the use of the relationships so fitted for the prediction of year class strength. A simple method for the combination of the estimates derived from different index series using inverse variance weighted averages is proposed, and methods for the estimation of the overall error in the prediction are discussed. The method used has been shown to perform well in simulation tests, and is well adapted for use on real datasets with time series of variable length and missing data. It has been implemented in a computer program (RCT3, superseding RCRTINX2) which is available for operational use, and has been endorsed by the ICES Working Group on the Methods of Fish Stock Assessment as being satisfactory for operational use until more complex methods have been shown to have superior performance.

Key words: Irish Sea Plaice, weighted regression, inverse variance weighting, VPA, recruitment, prediction.

Received 9 December 1993; accepted 20 January 1997.

J. G. Shepherd: MAFF Directorate of Fisheries Research, Pakefield Road, Lowestoft, Suffolk, NR33 0HT, UK.

Introduction

The choice of a method for the analysis of recruit index data, and for the subsequent prediction of year-class strength, has been a problem for fish stock assessment (especially the preparation of catch forecasts) for many years. A variety of more or less *ad hoc* methods have been used at various times, without any clear consensus emerging, and without much discussion in the literature. The problem was addressed at some length by the ICES Working Group on the Methods of Fish Stock Assessment (hereafter referred to as the Methods Working Group) (ICES, 1984, 1987), with some further discussion in 1985 (ICES, 1986). As a result of these deliberations and the recommendations made, a simple calibration regression and combination method was implemented in the RCRTINX2 program (now superseded by RCT3) and this is now commonly used by ICES working groups. A description of the method and its rationale has until now only been available as an unpublished manuscript (Working Paper No 5 of the Methods Working Group in 1987), and this paper aims to provide a fuller and more accessible account of the method, its use, and the evaluation of the results.

The problem may be regarded as having two parts. First, the analysis of any individual recruit index series by a regression method and its use for prediction. Second, the combination of several such predictions based on different index series to obtain a best final overall estimate. This separation is not essential, as will be mentioned below in discussing alternative methods, buit it does help to clarify various aspects of the problem. Such a method is also well adapted to the analysis of data series of differing lengths, often with missing values, which is very important for practical use. A number of other facilities which are very useful in practice can also easly be provided within the framework of a weighted regression method, and these are also discussed.

Analysis and prediction using a single index series

The particular problem of analysis and prediction of concern is simply stated as ''given a time series of recruit

*Present address: Southampton Oceanography Centre, Empress Dock, Southampton, SO14 3ZH, UK.

1054–3139/97/050741+12 $25.00/0/jm970222

index values and associated VPA estimates of year-class strength, how may these data be analysed so as to yield a high quality prediction of future year-class strength?'' A regression method is an obvious candidate, but there is room for a considerable variety of choices in selecting the exact method to be used. Surprisingly, the usual predictive regression is not usually the best choice. This is because recruit index data are not usually very precise – experience suggests that standard errors in the logarithmic scale of less than 0.3 are unusual, and that values of the order of 1.0 are quite common. The problem therefore involves the analysis of imprecise data, and a fundamental assumption of predictive regression, that the explanatory variates are measured without error, is violated. In dealing with ''noisy'' data different choices about apparently small details of the method can have a substantial effect on the results obtained. The analysis of high quality data is relatively straightforward, whilst that of poor quality data requires very great care. The particular problem of the analysis of a recruit index data was addressed by the Methods Working Group (ICES, 1984), relying heavily on a working paper prepared by E. F. Harding, which has regrettably not been published in full, although a brief account is available (Harding, 1986).

The essential features of the data concerned are that:

(a) year-class strength and recruit index values are non-negative, and have a highly skewed distribution, which may often be well-approximated by the log-normal distribution (Hennemuth *et al.*, 1980; Garrod, 1983; Rothschild and Mullen, 1985; Myers *et al.*, 1990);
(b) the relationship between recruit indices and year class strength (as estimated by VPA) cannot necessarily be assumed to be linear and proportional, especially not for indices for the youngest age groups (0 and 1 group);
(c) the measurement errors in recruit indices are apparently often large, and increase with abundance (as reflected in the index value): a constant coefficient of variation is a better approximation than a constant variance, and a log-normal distribution for the errors is a more appropriate description than a normal distribution.

These features means that standard predictive linear regression is not immediately appropriate, since it assumes that the relationship is linear, that the errors in the explanatory variate (the recruit indices) are negligible, and that those in the dependent variate are normal with constant (homoscedastic) variance.

Transformations of the data are often employed to address such problems (Atkinson, 1985). These often solve one problem whilst making another worse (Gilchrist, 1984). However, in this case, the above features may all be allowed for simultaneously by carrying out a logarithmic transformation of both dependent and explanatory variables. This helps to normalise the distribution of the data, linearise the relationship between the variables, normalise the distribution of the errors, and promote homoscedasticity. Provided that a power-law relationship between the untransformed variables is an adequate model for the non-linear relationship, this simple transformation brings the problem squarely within the framework of conventional linear regression, although the dominant errors are still in the explanatory variate. To avoid the problems which occur when real data sets include zero values (even when they are otherwise approximately log-normally distributed), the common (but not very satisfactory) procedure of adding one to all values before log-transforming has been adopted throughout.

The appropriateness of a logarithmic tranformation of both variables to linearise the relationship between survey index and abundance has not, so far as I am aware, between studied in detail. However, it is well known that such a transformation often achieves approximate linearity for simple non-linear relationships which are either convex or concave, but not both (i.e. not wiggly) – see, for example, Carroll and Ruppert (1988). It has sometimes been suggested that the slopes of the relationships so obtained tend to be less than one, so that the power in the power law is also less than one. This means that the eventual year-class strength, as estimated retrospectively by VPA, tends to be less extreme than the raw indices suggest – the big year classes turn out to be not as big as expected from the relative size of the indices, and the small ones not so small. There are many possible explanations for such an effect, including density-dependent juvenile mortality rates (Myers and Cadigan, 1993), age reading errors on older fish, variations of spatial distribution with changes of abundance (Myers and Stokes, 1989; Swain and Wade, 1993) possibly coupled with inadequate survey coverage, and variations of fish behaviour with respect to survey gear. Such mechanisms warrant further study, but for practical purposes if there is significant evidence of such non-linearity, it should be allowed for (e.g. by the logarithmic transformation adopted here). It would be unwise to assume linearity (strict proportionality between the index and eventual abundance) if there is evidence of non-linearity, especially if this is of the form described above (Myers and Cadigan, 1993). However, slopes much different from one (perhaps outside the range 0.5 to 2) indicate extreme variations of apparent catchability, in conflict with the underlying ideal that an index of abundance should be proportional to the population size, and such a relationship should be viewed with scepticism, and the causes and mechanisms for it should be investigated carefully if at all possible.

The other main feature of data of this type is that both the dependent variable which one wishes to predict (year-class strength) and the explanatory variables (the recruit indices) are measured with error – the former by VPA, and the latter by research vessel surveys, etc. In fact, the conventional assumption is that VPA estimates of year class strength are precise, whereas the index measurements are subject to relatively large sampling errors. This assumption has been confirmed for some examples by factor analysis of complete datasets (Rosenberg *et al.*, 1992) though there is, of course, no guarantee that it is always valid.

The dominance of errors in the explanatory variate means that the problem approximates to one of calibration regression (Harding, 1986), about which there has been considerable controversy in the statistical literature (see, e.g. Brown, 1982). The essential point is that the best (maximum likelihood) estimate of the relationship is obtained by regressing the noisy explanatory variable on the relatively precise dependent variable. This is the reverse of the normal procedure, but is actually just a special case of functional regression, as described by Lindley (1947) and Davies and Goldsmith (1976): note that the often cited treatment of Ricker (1973) is incorrect except in rare special cases (Copas, 1972). The statistical properties of a prediction based on this relationship are however ill-defined, at least in theory, because the reciprocal of a normally distributed number is not well-behaved. The problem can be resolved by a careful treatment of the likelihood maximisation problem (Harding, 1986) but this is not readily amenable to operational use. A simpler practical solution is just to estimate the required statistical properties *post hoc*, by examining and summarising the errors arising from actually using the fitted relationship to predict the dependent variable for the historic data available. The procedure for doing this is as follows: let x denote the (error prone) explanatory variable, and y the (relatively precise) dependent variable, for which a prediction is required. This choice is usefully mnemonic, and also accords with the ''natural'' way of plotting the data. Using standard software, one needs to regress x on y, to give a relationship of the form $\hat{x} = ay + b$, in order to minimise the errors $\Sigma(x_i - \hat{x})^2$. This relationship is then used for prediction in the ''inverted'' form $\hat{y} = (x - b)/a$. The standard error of this estimate may be obtained from the residual mean square estimated from the sum of squares of the residuals of the actual predictions $\Sigma(y_i - \hat{y}_i)^2$. This is, of course, a sensible practical way of estimating the prediction errors, whatever method has been used to estimate the relationship upon which the prediction is based.

Continuing with this convention (VPA $\simeq$ y, index $\simeq$ x), the prediction formula is

$$\hat{y} = gx + h \tag{1}$$

where $g = 1/a$, and $h = -b/a$, and a and b are the slope and intercept obtained by regressing x on y as described above. It is easily shown that the residual variance is in fact

$$\sigma^2_y = g^2\sigma^2_x \tag{2}$$

where σ^2_x is the value returned by the standard procedure for the regression of x on y.

The standard error of a further individual prediction (not that of the fitted value itself) is then assumed to be given by the usual formula

$$s = \sigma_y\left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right]^{1/2} \tag{3}$$

This estimate is open to question, since the distribution of the errors is not (under the assumptions made) normal. The adequacy of this simple ''common-sense'' approximation in relation to the likelihood based confidence intervals advocated by Harding (1986) warrants further study.

An overall measure of the quality of a prediction is required if, as here, it is subsequently necessary to combine several independent predictions in a rational way. Therefore, I assume here, following a suggestion by J. Pope (pers. comm.), that the standard error of prediction is the single best indicator of the quality of a prediction. It becomes large if the relationship is a poor fit (so that the residual variance is large), or if the prediction involves a substantial extrapolation outside the range of the data (the final term in the formula above). It may still be undesirably large, even if the correlation coefficient is high, when the data have a large dynamic range. Note that the terminology in the literature for such standard errors is confused, and it is important to use the three-term expression here, which includes the residual variance associated with any observation (the first term in the bracket), as this is clearly a lower limit for the error of any prediction. Other regression diagnostics, such as the correlation coefficient and the slope of the relationship, may also be useful in evaluating the data, but the bottom line is that if s is large, the prediction is of limited value. Note that s is a logarithmic quanitty and therefore independent of the scale of measurement. As discussed below, practical experience suggests that datasets yielding values of s exceeding 0.4 or 0.5 (for log-transformed data) are troublesome, but may nevertheless have a little value if recruitment is liable to vary by more than an order of magnitude.

Apart from the possible dependence of apparent catchability on abundance, inherent in a power-law relationship, as discussed above, the simple regression model used assumes constant catchability (no change with time, or other external factors). This simple

assumption could be relaxed if there were evidence that this was necessary, e.g. by moving to a multiple regression model. Ideally, for survey-based indices, no such generalisation should be necessary. To allow for the possibility of secular changes (e.g. of stock distribution patterns, or survey methods), which could mean that old data become progressively less relevant, the practical implementation of a regression method can allow for a progressive down-weighting of old data (''tapering'') by using a weighted regression – this is discussed below. The use of a weighted regression formulation is also convenient, since it allows missing data to be handled very easily by simply assigning them zero weight in the regressions.

Combination of estimates from multiple series

It is a common occurrence to have available more than one data series which might serve as a basis for prediction. The method of analysis described above may be applied to each in turn, to obtain a set of predictions, with estimates of their standard errors. Since one cannot know *a priori* which of these will eventually prove to be the best prediction, it seems sensible to attempt to combine all the available information in some way. This is a common statistical problem, and for certain cases optimal solutions can be derived. Among these is the technique known as the Kalman filter which may be applied to the present problem (J. G. Pope, unpublished manuscript, Working Paper No. R1 to ICES Working Group on Methods of Fish Stock Assessment 1985). This technique is, however, more general (and complicated) than is necessary for the present problem, which is really rather simple, since it is usually possible to treat the individual estimates as having independent errors. In such a case, the Kalman filter in fact reduces to something very similar to the weighted averaging procedure proposed below.

It is a standard result in statistics that under plausible assumptions the best (minimum variance, unbiased) overall estimate obtainable from a set of independent estimates of known standard errors is a weighted mean of those estimates, where the weights are taken to be the inverse of the variance of the individual estimates (see, e.g. Weatherburn, 1962). Thus, given a set of estimates $\hat{y}(j)$ with associated prediction standard errors $s(j)$, where j indexes the estimates obtained from the different index series ($j = 1,m$) the overall weighted estimate of the mean is

$$\bar{y} = \frac{\sum_j \hat{y}(j)/s^2(j)}{\sum_j 1/s^2(j)} \tag{4}$$

This result may be found in almost any practical statistics text (e.g. Davies and Goldsmith, 1976, Appendix 6A), but a particularly useful treatment is given by Topping (1962)).

It is also possible to estimate the standard error of this overall weighted mean. In fact, as pointed out by Topping (1962, pp. 91–93) one can make two independent estimates of the standard error. The first, s_{ext}, referred to as the external standard error, is obtained from the variance

$$s^2_{ext} = \frac{\sum_j \{\hat{y}(j) - \bar{y}\}^2/s^2(j)}{(m-1)\sum_j 1/s^2(j)} \tag{5}$$

This estimate is based on the actual deviations of the individual estimates from the overall mean, i.e. the observed discrepancies between the estimates, and may thus be regarded as a posterior estimate.

The second estimate, s_{int}, is obtained from the variance

$$s^2_{int} = \left[\sum_j 1/s^2(j)\right]^{-1} \tag{6}$$

This is referred to as the internal standard error, and is based only on the estimates of the individual standard errors. It is independent of the actual estimates $\hat{y}(j)$, and represents a prior estimate of what we would expect the error of the final mean to be, taking account of the known errors of the individual estimates from which it is constructed. If the $s^2(j)$ are all equal, it reduces to the usual estimate ($s/\sqrt{m}$) for the standard error of a mean.

This discussion may equally well be phrased in terms of a hierarchical one-way analysis of variance, or within sample and between sample variances, if desired (see, e.g. Davies and Goldsmith, 1976, section 6.3). In either case, one reaches the conclusion that the ratio

$$F = s^2_{ext}/s^2_{int} \tag{7}$$

may be tested as a variance ratio with m and $m-1$ degrees of freedom. If all is well, it should be close to one. If it is significantly larger than one, it indicates that there is evidence of a discrepancy between the estimates – their deviations from the mean (and each other) are larger than expected from their previously estimated errors. Conversely, if it is significantly less than one, it indicates a suspicious degree of concordance among the estimates, possibly indicating that they are not independent of each other as assumed, or that the data have been manipulated in an inappropriate way. For practical purposes, one can adopt the larger of these two estimates as the final estimate of the standard error of the overall mean prediction.

Note that in order to maintain the assumptions of normality (etc) which underlie the optimality of the inverse variance weighted mean, all these calculations should be carried out on the logarithmic estimates, without retransformation: one is therefore constructing what will (after retransformation) be a weighted geometric mean. It is arguable whether one should apply a bias correction factor ($\exp[s^2/2]$) to the final estimate. This should be done if one seeks an unbiased estimate of the ''expected'' (arithmetic mean) value, but not if one is content with an unbiased estimate of the median value for the prediction so that there is an equal chance of the prediction being too high or too low. This amounts to a choice of the loss function. The latter choice (no bias correction) is adopted here. As pointed out by Laurec and Perodou (1987) if the bias is large enough to be a problem in practice (more than 10% perhaps) then the overall prediction error will be dominated by the variance anyway, and will be so large that the prediction is of dubious value (i.e. $s \geq 0.4$). Since the geometric mean is always less than the arithmetic mean, the absence of a bias correction leads to an inherent slight tendency towards caution in the prediction, which is of practical significance only if the uncertainty is large, when this may perhaps be regarded as a useful feature.

The use of weighted means as described above is by no means the only possible method for combining the various estimates available. On the face of it, since one ends up with a linear combination of the individual estimates (Equation 4), and these are themselves linear functions of the index values, a suitable estimate ought also to be that given by standard multiple regession methods. In practice, this is not a satisfactory method, because the index series tend to be mutually highly correlated. This is especially true if they are good quality indices, in which case they are all highly correlated with VPA and each other. This leads to the well-known problem of collinearity, and near singularity (ill-conditioning) of the information matrix (Davies and Goldsmith, 1976, section 8.53). The practical effect of this is that the coefficients of the regression become very variable, and sensitive to noise in the data. There is a tendency to get a mixture of large positive and negative coefficients, so that the overall prediction becomes dependent on small differences between the individual predictions. In addition, the coefficients become very dependent on the precise dataset used, and tend to vary wildly as extra data are included, so that the weight attached to a particular index may differ greatly from one year to the next. All these are undesirable features, and make the method unsuitable for this application. The weighted average method may be regarded as being like a multiple regression in which the coefficients are constrained to be all positive and fairly consistent from one year to the next.

Another possible method, which explicitly recognises that all the estimates of abundance, including both the indices and VPA, are supposed to be related to a common underlying factor (the true abundance) is factor analysis (Rosenberg *et al.*, 1992). Their simulation tests of this method indicate that in some cases it performs slightly better than the present method, which it may eventually come to supersede, once it has been ''field-tested'' on real data. An important potential advantage of this method is that it allows for possible inaccuracies of the VPA estimates of abundance, whereas most other methods treat these as exact. This should in principle allow one to detect (and allow for) the situation where all the indices correctly estimate the size of a particular year class, but the VPA estimate is incorrect (because of an error in the catch-at-age data, or an anomalous value of natural mortality, for example). It is not yet known whether this is a significant advantage in practice.

Finally, the use of a formal maximum likelihood multiple calibration method has been explored by Laurec (Appendix E of ICES, 1987). For reasons which are not entirely clear, this method did not perform as well as the weighted average method in simulation tests (ICES, 1987), even though it is theoretically preferable.

Practical details

As with many analyses of error-prone data, the practical details of the implementation of a method may be as important as the choice of the underlying method. Several such important details are dealt with below: in each case these lead to options available to the user. Fuller guidance on the choice of options may also be found in the user's guide to the current implementation of the method (RCT3) by Darby and Shepherd (in press).

Weighting: missing data and time tapering

Virtually all real datasets of the type required for this analysis contain missing data, because the survey/index series invarably commence in different years, are occasionally not available and become available for the most recent year classes at different times of the year. This presents no difficulty for a regression method, since missing values are simply assigned zero weight, and thereby excluded from both the analysis (calibration) and the prediction. This of course requires the use of a regression routine which allows for weighting (see, e.g. Davies and Goldsmith, 1976, p. 202, for the appropriate formulae). Given this facility, it is also possible to address (to some extent) another practical difficulty, i.e., possible changes of catchability with time.

The method described above is based on a power-law relationship between recruit indices (u) and abundance (P), in the form $u = \alpha P^{\beta}$, with coefficients α and β which are constant with respect to time. This means that the catchability, $q = u/P$ is a function of abundance (unless $\beta = 1$), but is not directly dependent on time. It may vary indirectly with time, of course, if abundance (P) varies with time, as it usually does. In practice, there are always a number of reasons why catchability may also vary with time, even if abundance remains constant. For example, the spatial distribution of the stock may change with time because of climatic variations, and in the case of commercial c.p.u.e. (and possibly even research vessel surveys) there may be undetected changes in fishing gear and fishing power which are not fully allowed for. Furthermore, there may be secular changes in natural mortality between the times at which surveys are conducted, and the time to which all the estimates are referenced, because of changes in the abundance of predator stocks. All these factors could in principle be allowed for in a more complicated model. A more practicable alternative is simply to recognise that old data may no longer be fully relevant to the current situation, and down-weight, using some appropriate ''tapered'' weighting function. The tricubic weighting of Cleveland (1979) was introduced by Armstrong (1985) for this purpose in the context of VPA tuning. Such weighting functions generally have a finite range (i.e. they are strictly zero beyond some maximum range), so that the effects of outliers are localised in time. Choosing a 20 year range (for example) means that data more than 20 years old are ignored completely, and that most wight is given to those for the last 10 years or so. A suitable weighting formula is

$$w(y') = [1 - \{(y - y')/D\}^n]^n \quad (8)$$

for $(y - y') < D$, with $w(y') = 0$ otherwise, where D is the range, y is the final year, and y' the previous year for which the weight is to be determined, and n is 1 for linear, 2 for bisquare, and 3 for tricubic weighting (the latter being recommended for general use by Cleveland, 1979).

A further facility afforded by weighted regression is that one may in special circumstances apply a prior weighting to all the data for a particular survey, perhaps because this is known to be unreliable for some reason, or indeed to exclude it entirely from the analysis (by assigning zero weight) without deleting it from the data files.

When a weighted regression is adopted, a little ambiguity arises over how the standard errors of the parameters and the estimates are to be calculated (M. Nicholson, pers. comm.). There are essentially two philosophically different points of view. In one case the weights are taken to reflect prior knowledge of higher error variances of the older data. In the other they are taken to reflect a reduction in the number of (hypothetical) multiple observations at each point. Neither is necessarily correct, and regrettably the results obtained differ slightly. The latter approach is adopted here. This means that the residual variance is estimated as

$$\sigma^2_y = \Sigma[w_i(y_i - \hat{y}_i)^2]/(\Sigma w_i - 2) \quad (9)$$

and the standard error of the estimate is then given by Equation (3): the positions of n and Σw_i are reversed in the denominators in the alternative interpretation. The difference is small provided that the weights are scaled to be of the order of one.

Shrinkage toward the mean

The calibration regression employed in this analysis is a special case of a functional regression problem (Davies and Goldsmith, 1976; Chapter 7) in which it is assumed that there is a real functional relationship between the variables observed over an indefinite range thereof. No prior assumption is made about the distribution of the observations or the range over which they are distributed, which would correspond to a structural regression: (Davies and Goldsmith, loc cit, section 7.8; Snedecor and Cochrane, 1980, section 9.14).

However, for a variate such as recruitment the observations may normally be described by some reasonable probability distribution (e.g. a log-normal distribution with a logarithmic standard deviation of the order of 0.5). Thus, estimates at the extremes of, or outside, the normal range are surprising, to be regarded with some suspicion, and possibly to be discounted to some extent. It would therefore be reasonable to take account of the expected (prior) distribution of recruitment in constructing individual estimates thereof. This may be done in a number of ways, including of course a full Bayesian treatment of the problem, or treatment as a multivariate structural regression problem, or by factor analysis as mentioned above. Here, however, we adopt a simpler procedure which is in the same spirit as these (and equivalent in some circumstances). This is to include the historic mean recruitment as an additional estimate, assigning it a weight corresponding to the observed historic variance of recruitment about the mean (all after logarithmic transformation, of course). This has little effect if high quality (precise) estimates are available, since the weight attached to the mean is then small. If, however, the available estimates are imprecise (compared with the historic variability), the effect may be substantial. The final estimate is always deliberately biased towards the mean, slightly so given good quality estimates, and substantially so if they are imprecise. This bias is incurred in order to reduce the variance of the final estimate, and is worthwhile because ultimately it is

the total mean square prediction error (measured as variance plus bias squared) which best describes the quality of the estimate.

The general procedure (known as ''shrinkage'' for obvious reasons) is discussed in a slightly different context by Copas (1983). In fact it is easily demonstrated (A. Laurec, pers. comm.) that when one has only a single recruitment estimate the procedure of calibration regression and shrinkage to the mean is precisely equivalent to doing a predictive regression in the first place. However, this is not true for multiple estimates. The difference is easily seen by considering the result of using many sets of bad data. For predictive regression, one obtains a set of regression lines of near zero slope, and therefore of predictions near the mean, with prediction standard errors similar to the historic standard deviation of the data. Combining these would lead to a final estimate very close to the mean, with a moderately small standard error ($1/\sqrt{n}$ of the individual ones). This standard error could be made arbitrarily small by including lots more bad data sets, which is unreasonable. By contrast, calibration regression on bad data leads to extreme predictions with enormous standard errors. Combining these would give a random number with a very large standard error. Applying the shrinkage by including the mean leads to a final estimate close to the mean, with a prediction standard error equal to the historic standard deviation, which is perfectly reasonable.

The efficacy of calibration with shrinkage has been confirmed by simulation testing (Rosenberg *et al.*, 1992; ICES, 1993a), and extensive practical experience since the method was introduced in 1987. All methods work well on high quality data, and multiple predictive regression, and calibration without shrinkage, give poor results on poor quality data.

Dealing with short time series

It is a matter of common observation that predictions from short time series are often wrong. For this reason, it is common practice to disallow predictions from regressions based on very few points (for example, fewer than five). This procedure is, however, not entirely satisfactory. Firstly, the choice of the minimum number is essentially arbitrary, but may be quite influential if it causes the results of a particular short time series to be either included or excluded, especially if it is apparently well-correlated. Secondly, it is in practice difficult to justify ignoring data, possibly obtained at great expense, for five years or more, before suddenly accepting them as valid and useful. A more progressive and statistically satisfactory procedure is required.

The reason for the difficulty is well-known: the fitted model is very sensitive to chance errors in any observation when few are available. The model fits some of the noise as well as the signal, and the goodness-of-fit for the hindcast (the fit to the construction dataset) is therefore always unrealistically good, compared to that obtained for the forecast (the fit to the validation dataset), and the model obtained (being partly fitted to the noise) is always less appropriate for the future than one thinks. The estimated residual variance can be corrected for this. The standard result for least squares (Seber, 1977) is that the variance should be inflated by a factor $(n+2)/n$, but Copas (1983) suggests that (in the context of multiple regression) this should be modified to $(n-1)/(n-3)$. This implies that the prediction variance from a regression based on three points is infinite, which is surprising, and would imply that such a prediction be discounted completely. An intermediate inflation factor of $n/(n-2)$ seems more plausible, and may (optionally) be applied in calculating the prediction standard errors. This has the effect of downweighting the results based on short time series compared with those for longer ones, so that no arbitrary restriction needs to be placed on the minimum number of points to be used (other than the fundamental minimum of three required to estimate a regression and its residual variance).

Setting a minimum standard error

Even with the adjustment described above, it is still possible on occasion for the prediction to be dominated by one data series, because by chance this just happens to have a very high correlation with VPA for the data available. This is of course most likely to occur with a short data series, but even for series of moderate length (up to 10 years), this can cause considerable variation in the weights applied to the predictions from different series from one year to the next, as one or the other just happens to have the best correlation.

The residual standard errors obtained can also be unrealistically small. Although the VPA estimates are treated as exact in this analysis, they are of course subject to errors because of sampling variability in the catch data. Various analyses, including separable VPA (Pope and Shepherd, 1982) and multiplicative models (Shepherd and Nicholson, 1991) suggest that catch-at-age data, and therefore VPA estimates of recruitment, rarely have coefficients of variation of less than 20%, and often more. It is unreasonable to believe an estimated residual standard error of a prediction which is less than the estimated standard error of the VPA recruitment estimates themselves, as the whole calibration is based on these. To allow for this, it is recommended that any such small error estimate be replaced by a *de minimis* figure, chosen to reflect the probable errors of the VPA estimates: in the absence of any other information the choice of 0.2 (in logarithmic units) reflecting a 20% coefficient of variation may be a suitable lower limit. Where the catch-at-age data are of poor quality

Table 1. Recruitment indices from Irish Sea plaice stock used to illustrate RCT3 analysis. NWGFS and EWGFS refer to North Wales Ground Fish Survey and England & Wales Ground Fish Survey, respectively. The age group, month of survey and survey series name in ICES (1993b) are given for each recruitment series.

		NWGFS					Irish	EWGFS	
Year class	VPA age 1	age 0 Oct (ssoct 0)	age 1 Jun (ssjun 1)	age 1 Oct (ssoct 1)	age 2 Jun (ssjun 2)	age 2 Oct (ssoct 2)	age 1 May (irmay 1)	age 1 Sep (ewsep 1)	age 2 Sep (ewsep 2)
1974	11 180				352	473			
1975	17 254		308	726	1775	1711	8.18		
1976	19 167	78	877	190	1648	650	14.56		
1977	23 226	32	641	1110	1744	3018	6.06		
1978	20 768	237	348	4046	5588	1161	19.09		
1979	15 585	757	3003	2330	1925	1897	3.37		
1980	8497	17	98	323	940	844	3.4		
1981	21 525	18	585	3125	1371	1538	12.9		
1982	21 330	1250	1195	4061	1796	2358	22.18		
1983	22 422	262	1983	2995	2208	1683			
1984	16 235	508	2635	2649	2281	970	17.9		
1985	18 995	430	2520	2246	1959	2145	19.71		
1986	20 025	1033	2074	4886	4264	2945	29.71		297
1987	10 945	173	2624	4053	2961	914	38.78	12 727	111
1988	5797	397	506	553	610	134	14.01	5998	69
1989		31	438	271	480		9.65	24 855	140
1990		216	873				8.31	11 052	
1991							40.37		

(e.g. because of known low levels of sampling) a higher choice in the range 0.3 to 0.5 would be more appropriate.

This modification does introduce an element of subjectivity into the analysis, but the effect is in fact to tend to equalise the weights attached to the available good quality datasets, without affecting the downweighting of the poor quality ones, and is therefore relatively benign. Experience suggests that this is an important safeguard against occasional extreme predictions due to chance configurations of noise.

Exclusion of poor quality datasets

Where a dataset is found to have poor predictive utility, it is arguable whether or not it should be excluded from further analysis entirely. To do so however would require that one set some sort of quality threshold, and most such thresholds are arbitrary to some extent. An appropriate selection is not therefore straightforward (and may be controversial). In addition, if a sharp cut-off is used (so that a dataset is either excluded or included with full weight), difficulties may arise over marginal cases which may be included one year and excluded the next (or vice-versa). Some form of robust estimation procedure based on a progressive down-weighting (see Mosteller and Tukey, 1977) might be appropriate, and would warrant further investigation. At present the option to exclude data is avialable, but should be used cautiously, unless other datasets of substantially higher utility are available too. The example discussed below includes several datasets of dubious utility, which have been retained for illustrative purposes, but in practice should probably be discarded.

Use of method and interpretation of results

The method as described is implemented by a computer program named RCT3: copies of this (for IBM compatible PC) are available from the author on request. This is functionally almost identical to the earlier version RCRTINX2, widely used by ICES working groups, except for the inclusion of the forecast/hindcast variance correction factor (see ''Dealing with short time series''), some cosmetic changes to input and output formats, and some improvements to the user interface. A User's Guide for the program is available (Darby and Shepherd, in press): this gives details of file formats, and guidance on the selection of user options.

An example dataset is given in Table 1, for the Irish Sea plaice stock, from ICES (1993b). This dataset includes a fairly large number of indices, some of which are poor predictors of recruitment. It has been chosen because it illustrates the performance of the method in down-weighting poor data, and employing the shrinkage toward the mean, and because it was also analysed by the ICES Working Group (ICES, 1993a).

Table 2. Results of RCT3 analysis for a single year class for the Irish Sea plaice data in Table 1.

	Regression					Prediction			
Survey series	Slope	Intercept	Standard error	R^2	No. of points	Index value	Predicted value	Standard error	Prediction weights
ssoct 0	0.73	5.85	1.12	0.080	12	5.99	10.24	1.318	0.037
ssjun 1	1.05	2.44	1.12	0.076	13	6.23	8.95	1.325	0.037
ssoct 1	0.85	3.25	0.81	0.136	13	6.32	8.64	1.011	0.063
ssjun 2	1.41	−0.99	0.78	0.150	14	6.42	8.07	1.055	0.058
ssoct 2	0.92	3.01	0.39	0.415	14	4.91	7.54	0.753	0.113
irmay 1	1.70	5.13	1.29	0.061	12	2.71	9.74	1.514	0.028
				VPA Mean:			9.75	0.311	0.664

The result of the analysis using the standard default options is given in Table 2. For each year class for which analysis is requested, the results of the calibration regression and resulting prediction are given for each index series: these are identified by the mnemonic code at the left-hand side. The slope of the log-log regression (VPA/index), its intercept, residual standard error, and the correlation r-squared, are given first, together with the number of points used. The (log-transformed) value of each index for the year class in question, the predicted value, its prediction standard error, and associated weight in the final overall mean are given next. Finally, the historic weighted mean over the VPA series, and its associated prediction standard error and weighting are given (NB: the prediction standard error includes the residual error of an individual observation). Note that all relevant quantities are given in logarithmic units where appropriate, and that the prediction weights have been normalised so that they sum to 1.0.

In evaluating these results, it should be noted that for a high quality prediction one is ideally looking for a slope near unity, a small residual standard error, and a value of r-squared near unity. These conditions are related but distinct, and may be met in any combination (or not at all). The acid test is, however, the size of the prediction standard error, which controls the weighting process.

In this example, it is clear that none of the indices performs very well. The slopes for all except the ''irmay'' series are in the range 0.7 to 1.4, which is acceptable, but all the residual standard errors except that for ''ssoct'' exceed 0.5, which may be taken as a rough boundary between the useful and the dubious (a standard error greater than 1.0 may be taken as indicating that the data are virtually useless). Similarly, all the r^2 values are small (less than 0.5), with even the best (that for ''ssoct'') only reaching a little more than 0.4. Not surprisingly, the prediction standard errors are large, with even the best being about 0.75, corresponding to a prediction of very dubious utility. This is in fact several times the historic standard error of (log) recruitment about its mean (0.311), so the mean is given more weight (0.664) than anything else in the overall prediction.

The final overall weighted means and their associated internal and external standard errors (and their variance ratio) are given in a summary table (Table 3), together with the historic VPA values for comparison, where available. The variance ratio may be used as the basis of an F-test, to determine whether or not the various indices are consistent or discrepant.

In this case none of the variance ratios exceeds two, so there is little evidence of discrepancy. Most in fact are quite small (less than 0.5), indicating a surprising degree of concordance. This is probably because the indices include several for different ages from the same survey cruises (three from October, and two from June). This may be due to variations between surveys which affect different age groups similarly (survey effects), which are quite likely to occur.

Up to 1986 (with the exception of 1980, which was poorly predicted), recruitment was close to the mean, and the heavily shrunk predictions (in which the mean is given a high weight) were reasonably satisfactory. Year classes since 1987 seem to be below average. The predictions reflect this (except for 1987) but the recent VPA estimates are of course themselves uncertain.

These results, and those for normal predictive regression, and unshrunk calibration, are illustrated in Figure 1. The tendency of unshrunk calibration to over-predict changes is clear, as is that of conventional predictive regression to yield results excessively close to the mean, when presented (as here) with multiple datasets of low predictive utility. The shrunk calibration predictions are the most satisfactory: Given the low correlations of these indices with VPA it is not surprising that the results are not very impressive. For an analysis of more extensive comparisons with simulated data see Rosenberg *et al.* (1992), and with real data, see ICES (1993a).

Table 3. Retrospective analysis of results from RCT3 analysis of Irish Sea plaice data for several year classes.

Year class	Weighted average prediction	Log wtd. aver. prediction	Internal standard error	External standard error	Variance ratio	VPA	Log VPA
1980	19 064	9.86	0.14	0.13	0.86	8498	9.05
1981	17 318	9.76	0.24	0.12	0.24	21 525	9.98
1982	22 476	10.02	0.24	0.17	0.52	21 330	9.97
1983	20 960	9.95	0.23	0.11	0.23	22 422	10.02
1984	21 177	9.96	0.20	0.13	0.44	16 236	9.69
1985	21 935	10.00	0.20	0.10	0.26	18 996	9.85
1986	26 388	10.18	0.19	0.15	0.56	20 026	9.90
1987	2 1004	9.95	0.18	0.15	0.65	10 946	9.30
1988	11 166	9.32	0.25	0.33	1.66	5797	8.67
1989	8798	9.08	0.34	0.35	1.06		
1990	14 683	9.59	0.42	0.15	0.13		
1991	16 234	9.69	0.45	0.43	0.91		

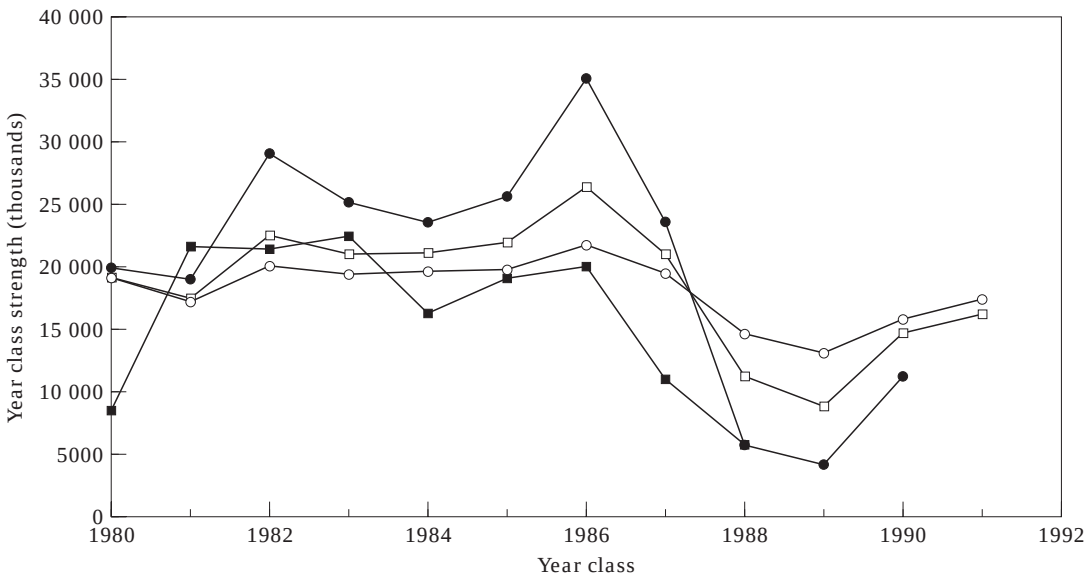

Figure 1. Comparison of year-class strengths for Irish Sea plaice from RCT3 analysis. (■) VPA, (□) shrunk calibration, (●) unshrunk calibration, (○) unshrunk prediction.

Discussion

The method described here is conceptually and computationally simple, and sufficiently flexible that a number of important practical details can be incorporated without difficulty. It may, if necessary, be implemented using any regression program or statistics package, but the use of a specific computer program is more convenient. It has been tested by simulation methods (ICES, 1987; Rosenberg *et al.*, 1992), and also by practical application over more than five years. No serious defects have as yet been identified, and most of the problems which have been encountered have been due to inadequate or misleading data. It is not suggested that this method is optimal, but it does seem to be adequate for catch forecasting purposes when the strength of recruiting year classes is an important factor. It also has the merit of providing a standard and objective framework for an analytical procedure in which there are otherwise many pitfalls for the unwary.

It should be noted that the method assumes that the errors in the various index series are mutually uncorrelated (the indices themselves are of course ideally highly correlated), and that the retrospective VPA estimates of year-class strength may be regarded as exact for practical purposes. In addition, it is assumed that the

catchability (as expressed by the coefficients of the calibration regression) is constant with respect to time. None of these assumptions is likely to be precisely correct, and safeguards against modest violations of the last two are incorporated, through the down-weighting of old data, and the imposition of a minimum prediction standard error.

As with any computational process that has been deliberately designed for convenient operational use, there is an inherent danger that it may be treated by the user as a ''sausage machine'', paying insufficient attention to the quality of either the raw materials or the product. It is therefore most important that users of this method pay careful attention to the regression diagnostics provided, and seek reasons for anomalous behaviour. They should also exercise considerable caution in accepting predictions from datasets which do not conform to the assumptions made in designing the method (particularly any yielding regressions slopes far from one, especially if they are negative) and routinely examine the retrospective analyses of past performance, which are supplied as a matter of course.

Finally, it should be noted that the procedure used is closely related to that employed in the calibration (''tuning'') of virtual population analysis itself. The special features are the need to focus on all available data for particular year classes, especially very recent data which becomes available shortly before the analysis is carried out, and to allow for a possible non-linear relationship (variation of catchability with abundance), as commonly observed for very young fish. It is arguable that the same methods would indeed be directly applicable to the VPA tuning problem, and developments along these lines would probably be worth pursuing.

Acknowledgements

This work was carried out at the Fisheries Laboratory, Lowestoft, and it is a pleasure to acknowledge the contribution made by many helpful discussions with colleagues there, and with members of the ICES Working Group on the Methods of Fish Stock Assessment, especially A. Laurec and A. Rosenberg, and helpful comments on the manuscript by S. J. Smith and an anonymous referee. The implementation of the method in the program RCT3 was carried out by Chris Darby.

References

Armstrong, D. W. 1985. Catchability Analysis. Appendix 2, pp. 151–159, *In* Report of the ICES North Sea Roundfish Working Group, ICES CM 1985/Assess 9.

Atkinson, A. C. 1985. Plots, Transformations and Regressions. Clarendon Press, Oxford. 282 pp.

Brown, P. J. 1982. Multivariate calibration. Journal of the Royal Statistical Society B, 44: 287–321.

Carroll, R. J. and Ruppert, D. 1988. Transformation and Weighting in Regression. Chapman & Hall, London. 249 pp.

Cleveland, W. S. 1979. Robust locally weighted regressions and smoothing scatter plots. Journal of the American Statistical Society, 74: 829–836.

Copas, J. B. 1972. The likelihood surface in the linear functional relationship problem. Journal of the Royal Statistical Society B, 34: 274–278.

Copas, J. B. 1983. Regression, prediction and shrinkage. Journal of the Royal Statistical Society B, 45: 311–354.

Darby, C. and Shepherd, J. G. in press. Combination of recruit indices by weighted averages using RCT3: a user's guide.

Davies, O. L. and Goldsmith, P. L. 1976. Statistical Methods in Research and Production (4th ed.). Longman, London. 478 pp.

Garrod, D. J. 1983. On the variability of year class strength. Journal du Conseil International pour l'Exploration de la Mer, 41: 63–66.

Gilchrist, W. 1984. Statistical Modelling. J. Wiley & Sons, Chichester and New York. 339 pp.

Harding, E. F. 1986. Modelling: the classical approach. The Statistician, 35: 115–134.

Hennemuth, R. C., Palmer, J. E., and Brown, B. E. 1980. A statistical description of recruitment in eighteen selected fish stocks. Journal of Northwest Atlantic Fisheries Science, 1: 101–111.

ICES 1984. Report of ICES Working Group on the Methods of Fish Stock Assessment (1984). ICES CM 1984/Assess 19, 56 pp; also reprinted in Cooperative Research Report for Conseil International pour l'Exploration de la Mer, 133, 56 pp (1985).

ICES 1986. Report of the ICES Working Group on the Methods of Fish Stock Assessment (1985). ICES CM 1986/Assess 10. 92 pp.

ICES 1987. Report of the ICES Working Group on the Methods of Fish Stock Assessment (1987). ICES CM 1987/Assess 24. 107 pp.

ICES 1993a. Report of the ICES Working Group on Methods of Fish Stock Assessment, Copenhagen (1993). ICES CM 1993/Assess 12.

ICES 1993b. Report of the Assessment of Northern Shelf Demersal Stocks. ICES CM 1993/Assess 2.

Laurec, A. and Perodou, J. B. 1987. Regard statistiques et informatiques sur l'analyse des puissances de pêche et des abondances apparentes. ICES CM 1987/D: 9.

Lindley, D. V. 1947. Regression lines and the linear functional relationship. Proceedings of the Royal Statistical Society, B9: 218–244.

Mosteller, F. and Tukey, J. W. 1977. Data Analysis and Regression. Addison-Wesley, New York. 588 pp.

Myers, R. A., Blanchard, W., and Thompson, K. R. 1990. Summary of North Atlantic Fish Recruitment 1942–1987. Canadian Technical Report of Fisheries and Aquatic Sciences, No. 1743. 108 pp.

Myers, R. A. and Cadigan, N. G. 1993. Density-dependent juvenile mortality in marine demersal fish. Canadian Journal of Fisheries and Aquatic Sciences, 50: 1576–1590.

Myers, R. A. and Stokes, T. K. 1989. Density-dependent habitat utilization of groundfish and the improvement of research surveys. ICES CM 1989/D: 15.

Pope, J. G. and Shepherd, J. G. 1982. A simple method for the consistent interpretation of catch-at-age data. Journal du Conseil International pour l'Exploration de la Mer, 40: 176–184.

Ricker, W. E. 1973. Linear regressions in fishery research. Journal of the Fisheries Research Board of Canada, 30: 409–434.

Rosenberg, A. A., Kirkwood, G. P., Cook, R. M., and Myers, R. A. 1992. Combining information from commercial catches and research surveys to estimate recruitment: a comparison of methods. ICES Journal of Marine Science, 49: 379–387.

Rothschild, B. J. and Mullen, A. J. 1985. The information content of stock-and-recruitment data and its non-parametric classification. Journal du Conseil International pour l'Exploration de la Mer, 42: 116–124.

Seber, G. A. F. 1977. Linear Regression Analysis. J. Wiley & Sons, New York. 486 pp.

Shepherd, J. G. and Nicholson, M. D. 1991. Multiplicative modelling of catch-at-age data and its application to catch forecasts. Journal du Conseil International pour l'Exploration de la Mer, 47: 284–294.

Snedecor, G. W. and Cochran, W. G. 1980. Statistical Methods. Seventh Edition. Iowa State University Press. 507 pp.

Swain, D. P. and Wade, E. J. 1993. Density-dependent geographic distributions of Atlantic cod (*Gadus morhua*) in the Southern Gulf of St. Lawrence. Canadian Journal of Fisheries and Aquatic Science, 50: 725–733.

Topping, J. 1962. Errors of observation and their treatment (3rd edition), Institute of Physics. Chapman & Hall, London. 119 pp.

Weatherburn, C. E. 1962. A First Course in Mathematical Statistics. Cambridge University Press. 277 pp.