Compositional analysis of catch curve data, with an application to *Sebastes maliger*

Jon T. Schnute and Rowan Haigh

Schnute, J. T. and Haigh, R. 2007. Compositional analysis of catch curve data, with an application to *Sebastes maliger*. – ICES Journal of Marine Science, 64: 218–233.

This paper applies modern compositional analysis to catch curve data from a quillback rockfish (*Sebastes maliger*) population in British Columbia, Canada. Bubble plots and ternary diagrams portray variable age distributions and highlight distinctions between commercial and survey sample data. The models formalize important historical issues in catch curve analysis related to selectivity and recruitment variability, where a particular model corresponds to a prescribed vector of design parameters. The roles that compositional distributions (multinomial, Dirichlet, logistic-normal) can play in fishery data analysis are described, and Bayesian methods are used to examine how the distribution of a key mortality parameter depends on model choice. The framework provides a direct link between model designs and policy outcomes that depend on estimated mortalities or mortality ratios.

Keywords: age composition, catch curve, compositional analysis, Dirichlet distribution, logistic-normal distribution, mortality, quillback rockfish.

Received 9 December 2005; accepted 5 October 2006; advance access publication 8 January 2007.

J. T. Schnute and R. Haigh: Fisheries and Oceans Canada, Pacific Biological Station, 3190 Hammond Bay Road, Nanaimo, British Columbia, Canada V9T 6N7. Correspondence to J. T. Schnute: tel: +1 250 756 7146; fax: +1 250 756 7053; e-mail: schnutej@pac.dfo-mpo.gc.ca

Introduction

Estimates of mortality rates in fish populations often come from age and size composition data. Historically, techniques for obtaining these estimates played an important role in the development of quantitative fishery science. For example, Ricker (1975, pp. 29–73) devoted the second chapter of his influential handbook on the statistics of fish populations to this problem. In the simplest case, the proportion of fish declines exponentially with age, and the rate of decline gives a measure of total mortality rate $Z = F + M$, where F and M denote mortality rates attributable to fishing and natural causes. Ricker (1975, p. 33) cites Baranov (1918) for giving the name *catch curve* to the graph of log frequency against age or size. Theoretically, some right-hand portion of this curve should follow a descending straight line with slope equal to $-Z$.

Although Ricker clearly recognized the importance of catch curve analysis for stock assessment, he gave many reasons for applying the technique with caution. For example, younger fish typically experience lower selectivity in the sampling process, and variable recruitment occasionally produces strong year classes that do not fit the exponential decay predicted by theory. Time trends in recruitment and fishing mortality can introduce further complications. These problems generate a profusion of methods for selecting some portion of the age distribution where exponential decay might represent the total mortality. Quinn and Deriso (1999, Section 8.2.2) discuss these problems further and point out (p. 322) that "catch curve analysis from a single year can be a dangerous procedure that should be used only if one knows that no trend in recruitment has been present."

Changes in computing technology and statistical theory now make it possible to extend Ricker's analyses into a more comprehensive framework (Schnute, 2006). Both frequentist and Bayesian methods have evolved to take into account the many recognized sources of error simultaneously. Furthermore, new theories suggest better methods for exploring and analysing data on the composition of a population (Aitchison and Shen, 1980; Aitchison 1985, 1986). Although these techniques cannot circumvent the problems discussed above, they do make it possible to examine their consequences systematically.

The analytical framework discussed here can facilitate such investigations. It includes a family of models that admit some of the catch curve scenarios contemplated by Ricker and other authors. Statistical theory plays a major role, where the data consist of proportions y_i $(i = 1, \ldots, g)$ in group i among a total of g groups. Although the observed vector has length g, its effective length is $g - 1$, because the constraint $\Sigma_i\, y_i = 1$ determines y_g from $(y_1, \ldots, y_{g-1})$. Just as age compositions indicate properties of fish populations, mineral compositions reveal features of geological structures. Aitchison's (1986) theory of compositional data analysis, used particularly by geologists, fits well with the application here. Following his approach, we define relevant terminology, such as a *composition operator*, *subcomposition*, *amalgamation*, and *ternary diagram*. Our examples use ternary diagrams as devices for exploratory data analysis.

To assess uncertainty, we need to compare observed proportions y_i with predicted proportions p_i from a specified model. We discuss three approaches to this problem, based on the multinomial, Dirichlet, and logistic-normal distributions. We adopt an

exploratory approach by comparing results obtained from a variety of *ad hoc* trial models. If different models give similar outcomes for a fixed data set, then at least we know that the results are robust to that set of model assumptions. On the other hand, if results vary substantially with model choice, then we have discovered different biological scenarios, consistent with the available data. We use formal criteria, such as Akaike's information criterion (AIC), to guide model selection (Burnham and Anderson, 2002).

The statistical theory described here applies not only to catch curves, but more generally to age- and size-structured population models with proportion data for one or more years. Our model focuses primarily on the data-limited context associated with data from a single year. Our worked example, however, uses a richer data set on quillback rockfish (*Sebastes maliger*) from British Columbia (BC), Canada. This allows us to examine the results of single-year analyses in a broader context and to illustrate the utility of compositional techniques, such as ternary diagrams.

Table 1. Notation for the catch curve model.

Fish ages	
a	actual fish age ($k \le a \le A$)
k	youngest age of interest
A	maximum age considered (plus class)
B	maximum age used internally by the model ($B \gg A$)
Counts and proportions	
n_a	number of fish observed at age a
n_A	number of fish observed in the plus class (age $a \ge A$)
n	total sample size $n = \sum_{a=k}^{A} n_a$
y_a	observed proportion of fish at age a
p_a	predicted proportion of fish at age a
Age groups	
i	index for fish age groups
g	number of fish age groups
a_i	cut points for grouping ages; $a_0 = k - 1 < a_1 < \cdots < a_g = A$; group i includes ages a in the interval $a_{i-1} < a \le a_i$
y_i	observed proportion in group i ($i = 1, \ldots, g$)
p_i	predicted proportion in group i ($i = 1, \ldots, g$)
$\tilde{y}$	geometric mean of y_i : $(\prod_i y_i)^{1/g}$
$\tilde{p}$	geometric mean of p_i : $(\prod_i p_i)^{1/g}$
Mortality and survival	
Z	total mortality $Z = F + M$
S_a	survival from age k to age a
Selectivities	
β_a	fishery selectivity on age $a \ge k$ ($0 < \beta_a \le 1$)
β_k	selectivity on youngest age $a = k$ ($0 < \beta_k < 1$)
α	selectivity parameter ($\alpha > 0$)
b_0	age of full selectivity with $\beta_a = 1$ for $a \ge b_0$
Recruitment anomalies	
m	number of recruitment anomalies
b_h	age with anomalous recruitment ($h = 1, \ldots, m$); $k \le b_1 < \cdots < b_m < A$
ρ_h	recruitment anomaly parameter at age b_h ($h = 1, \ldots, m$)
τ	standard deviation for recruitment anomalies
R_a	relative recruitment at age a ($a = k, \ldots, A$)

Catch curve models

Our catch curve model generates theoretical proportions p_a of fish at age a, based on the combined effects of survival, selectivity, and recruitment. Table 1 summarizes the required notation, and the model itself appears in Table 2. We confine fish ages to the range $k \le a \le A$, where k is the youngest age of interest and A denotes a plus class for all ages $a \ge A$. Internally, the model computes p_a for a broad range of ages up to $B \gg A$, then uses (T2.5) to accumulate the proportion p_A in the plus class.

The total mortality parameter Z determines the exponentially decaying survival curve (T2.1), where cumulative survival S_a from age k to A has the initial value $S_k = 1$. The model assumes that fish are fully selected by the fishery or sampling programme at age b_0. For younger ages the selectivity β_a is given by the asymptotic curve (T2.2). This depends on two parameters: the selectivity β_k ($0 < \beta_k < 1$) on the youngest age k, and a positive exponent α that determines how rapidly β_a increases from β_k to 1 as a increases from k to b_0. Across this range, the selectivity curve is concave, linear, or convex when $\alpha > 1$, $\alpha = 1$, or $0 < \alpha < 1$, respectively. In our examples, we constrain α between 2 and 25 to ensure a concave selectivity curve. Schnute and Richards (1995) used a similar curve with $b_0 = A$, but we give the model here greater flexibility by allowing b_0 to be set arbitrarily.

The model assumes a constant base level of recruitment R_a to each age a, but this can be modified to include m anomalies at specified ages b_h ($h = 1, \ldots, m$). In actual samples, if a strong year class appears at age b_h, then ageing error tends to increase the proportion of fish at nearby ages a. Our model uses a single parameter τ to capture this effect, where τ increases as the influence of ageing error spreads to a broader range of nearby ages. Although we could allow a different parameter τ_h for each anomaly, a single value τ works reasonably well in our examples below. If circumstances permit, τ might be set or assigned a prior distribution from independent reader data. The model uses (T2.3) to calculate the combined effect of all recruitment anomalies, where a standard base level $R_a = 1$ is incremented by the amount

Table 2. Catch curve model with age-dependent survivals S_a, selectivities β_a, and recruitments R_a. Calculations depend on a fixed design vector $\boldsymbol{\phi}$ in (1), and the predicted proportions $p_a(\boldsymbol{\theta})$ vary with the parameter vector $\boldsymbol{\theta}$ in (2).

$$S_a = e^{-Z(a-k)}; \quad a = k, \ldots, B \tag{T2.1}$$

$$\beta_a(\beta_k, \alpha) = \begin{cases} 1 - (1 - \beta_k)\left(\dfrac{b_0 - a}{b_0 - k}\right)^{\alpha}; & a = k, \ldots, b_0 - 1 \\ 1; & a = b_0, \ldots, B \end{cases} \tag{T2.2}$$

$$R_a(\rho_1, \ldots, \rho_m, \tau) = 1 + \sum_{h=1}^{m} \rho_h \exp\left[-\frac{1}{2}\left(\frac{a - b_h}{\tau}\right)^2\right]; \quad a = k, \ldots, B \tag{T2.3}$$

$$p_a(\boldsymbol{\theta}) = \frac{S_a \beta_a R_a}{\sum_{a=k}^{B} S_a \beta_a R_a}; \quad a = k, \ldots, B \tag{T2.4}$$

$$p_A(\boldsymbol{\theta}) = \sum_{a=A}^{B} p_a(\boldsymbol{\theta}) \tag{T2.5}$$

ρ_h at age b_h and this effect is spread to nearby ages with the profile of a normal distribution.

The three model equations (T2.1–T2.3) describe the effects of survival, selectivity, and recruitment. Each is normalized to a base level by requiring that $S_k = 1$, $\beta_a = 1$ for large a, and $R_a = 1$ in the absence of recruitment anomalies. The products $S_a\beta_a R_a$ determine the profile of age proportions $p_a (a = k, \ldots, B)$. Explicitly, the transformation (T2.4) converts these positive quantities to model predictions p_a with the property $\Sigma_{a=k}^{B} p_a = 1$. Finally, the aggregation (T2.5) gives the proportion p_A in the plus class, as mentioned earlier.

In practice, an analyst applying this model would specify the design vector

$$\boldsymbol{\phi} = (k, A, B; b_0; m, b_1, \ldots, b_m) \tag{1}$$

with $m+5$ components, partitioned by semicolons into quantities associated with the age range, selectivity, and recruitment. Predictions $\{p_a\}_{a=k}^{A}$ then depend on the parameter vector

$$\boldsymbol{\theta} = (Z; \alpha, \beta_k; \tau, \rho_1, \ldots, \rho_m) \tag{2}$$

with $m+4$ components associated with mortality, selectivity, and recruitment. Because catch curve analysis focuses primarily on Z, the remaining components of $\boldsymbol{\theta}$ act essentially as nuisance parameters. As discussed in the introduction, an exploratory analysis typically involves numerous trial designs (1), such as choices b_h that correspond to prominent modes in the observed data. The model can be used to investigate how the estimates of Z vary among these potential interpretations.

Any useful version of the model includes the survival component S_a in (T2.1), but the selectivity and recruitment anomaly components (β_a, R_a) can be turned on or off. This suggests four cases of particular interest:

Case 1:	$b_0 = k$,	$m = 0$;	(S_a)
Case 2:	$b_0 > k + 3$,	$m = 0$;	(S_a, β_a)
Case 3:	$b_0 = k$,	$m \geq 1$;	(S_a, R_a)
Case 4:	$b_0 > k + 3$,	$m \geq 1$;	(S_a, β_a, R_a)

where labels on the right indicate active components. In particular, the choice $b_0 = k$ effectively removes selectivity from the model, because then $\beta_a = 1$ for every a. Similarly, recruitment anomalies do not occur when $m = 0$, so that $R_a = 1$ for every a. If selectivity is active, the portion of the curve (T2.2) governed by the two parameters (α, β_k) covers the age range from k to $k + b_0 - 1$. The condition $b_0 > k + 3$ in Cases 2 and 4 requires that this range should include at least three ages to estimate these two parameters.

Figure 1 illustrates Case 4 for a hypothetical catch curve model when $m = 3$. Three panels show plots of the components S_a, β_a, and R_a in relation to ages in the range $k = 5$ to $A = 40$. The fourth panel shows the resulting proportions p_a for each a. Features in the final panel reflect various aspects of the underlying components, such as the three anomalous recruitments. Although the graph represents ages in the range from k to A, the model's internal calculations go up to age $B = 200$, so the graphs of S_a, β_a, and R_a are not shown for their entire range. However, the results of this extended calculation account for the plus class at age $A = 40$ in the graph of p_a. For proper calculation of p_A, age B must be chosen large enough to include fish with a very small probability of survival. For example, if ε represents a small number like 10^{-5}, then the requirement $S_B < \varepsilon$ in (T2.1) implies that

$$B > k - \frac{\log \varepsilon}{Z}. \tag{3}$$

In Figure 1, where $Z = 0.1$, this estimate suggests using $B > 120$. We have used a higher value $B = 5A = 200$, with $S_B = 3.4 \times 10^{-8}$.

Compositional data

Catch curve analysis typically begins with age measurements from a sample of n fish. Assume as before that ages a lie in the range $k \leq a \leq A$, with a plus class at age A. Then $n = \Sigma_{a=k}^{A} n_a$, where n_a denotes the number of fish with age a. The age composition data $(n_k, \ldots, n_A)$ give the observed proportions

$$y_a = \frac{n_a}{n} \tag{4}$$

of fish at each age a, where $\Sigma_{a=k}^{A} y_a = 1$.

We often need to amalgamate age proportions into g groups, based on a prescribed sequence of ages a_i $(i = 0, \ldots, g)$ with

$$k - 1 = a_0 < a_1 < \cdots < a_g = A. \tag{5}$$

By definition, group i $(i = 1, \ldots, g)$ includes all fish with age a in the range $a_{i-1} < a \leq a_i$, and

$$y_i = \sum_{a=a_{i-1}+1}^{a_i} y_a \tag{6}$$

denotes the proportion of fish in this group. We use different indices a and i to distinguish the proportion y_a at actual age a from the amalgamated proportion y_i within group i. This notation lacks mathematical elegance, because subscripts usually act as dummy variables, but it works effectively in the context here. Similarly, proportions p_a calculated from a catch curve model can be amalgamated to give the theoretical proportion p_i in group i:

$$p_i = \sum_{a=a_{i-1}+1}^{a_i} p_a. \tag{7}$$

The transition (4), from an age composition vector $(n_k, \ldots, n_A)$ to a vector of proportions $(y_k, \ldots, y_A)$, represents one of several compositional operators mentioned in the introduction. Because these transformations play an important role in the theory of compositional data analysis, we define them mathematically here. In general, let $\mathbf{u} = (u_1, \ldots, u_g)$, $\mathbf{v} = (v_1, \ldots, v_g)$, and $\mathbf{y} = (y_1, \ldots, y_g)$ denote vectors of real numbers, positive numbers, and proportions, respectively. Thus, $-\infty < u_i < +\infty$, $0 < v_i < +\infty$, $0 < y_i < 1$, and $\Sigma_{i=1}^{g} y_i = 1$. The three transformations

$$v_i = e^{u_i}, \tag{8}$$

$$y_i = \frac{v_i}{\sum_{j=1}^{g} v_j}, \tag{9}$$

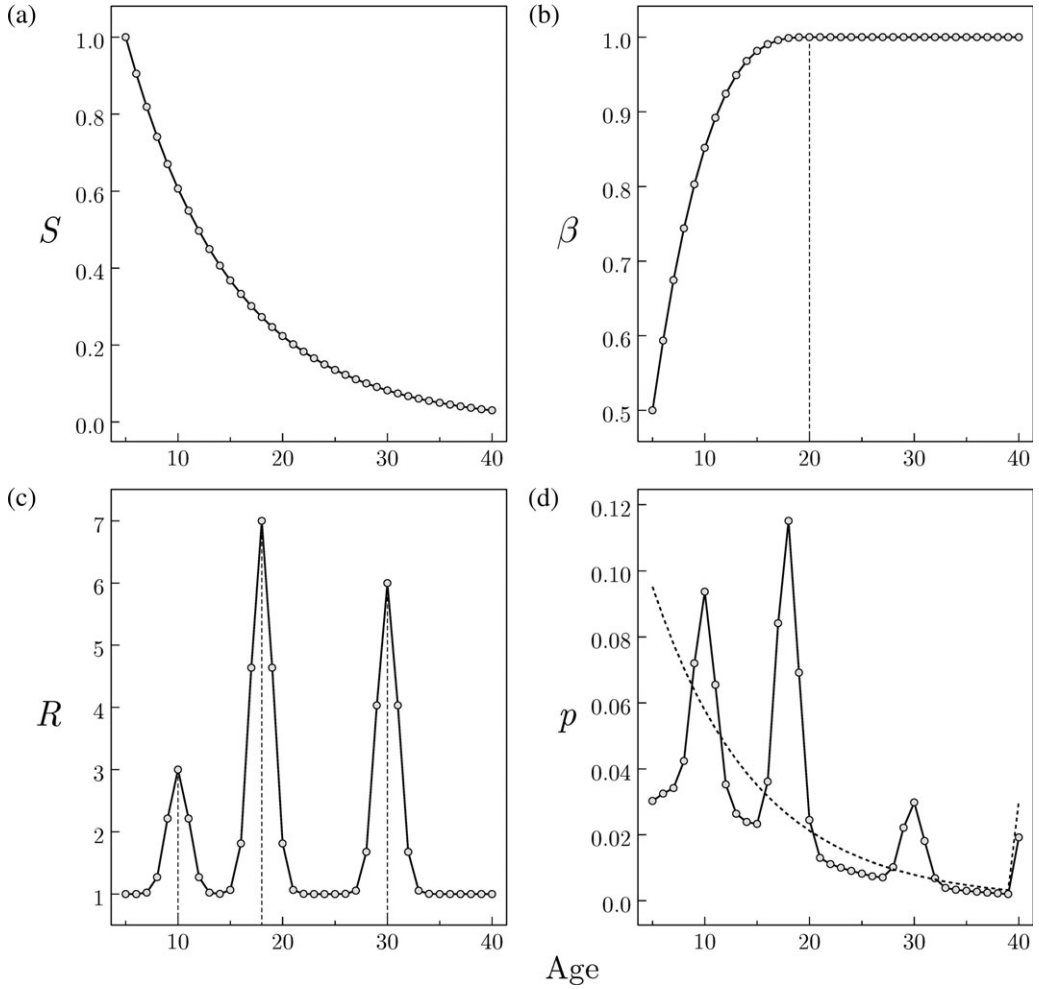

Figure 1. Plots of (a) survival S_a, (b) selectivity β_a, (c) recruitment R_a, and (d) predicted proportions p_a as functions of age a in the range $k \leq a \leq A$, based on the catch curve model (T2.1)–(T2.5). This example uses the design vector $\boldsymbol{\phi}$ in (1) specified by $k = 5$, $A = 40$, $B = 200$, $b_0 = 20$, $m = 3$, $b_1 = 10$, $b_2 = 18$, and $b_3 = 30$. The parameter vector $\boldsymbol{\theta}$ in (2) has components $Z = 0.1$, $\beta_k = 0.5$, $\alpha = 3$, $\tau = 1.5$, $\rho_1 = 2$, $\rho_2 = 6$, and $\rho_3 = 5$. Broken lines indicate (b) the age b_0 of full selectivity, (c) ages b_h with recruitment anomalies, and (d) proportions p_a based on survival only (Case 1 of the catch curve model).

$$y_i = \frac{e^{u_i}}{\sum_{j=1}^{g} e^{u_j}}, \tag{10}$$

take real numbers to positives, positives to proportions, and reals to proportions, respectively. We use both (9) and (10) to design probability models for a random proportion vector $\mathbf{y}$, such as the age proportion data in a catch curve analysis.

Aitchison and Shen (1980) used the relationship (9) to define the *composition operator* C, where $\mathbf{y} = \mathrm{C}[\mathbf{v}]$. Their theory found a natural home in geology, where C transforms a vector $\mathbf{v}$ of observed mineral components to a corresponding vector $\mathbf{y}$ of proportions. Similarly, our transformation (4) illustrates an application of C. Geologists also use *subcompositions*, defined by selecting particular components of a composition. For example, the first three components define the subcomposition $\mathbf{y}' = (y'_1, y'_2, y'_3) = \mathrm{C}[(y_1, y_2, y_3)] = (y_1, y_2, y_3)/(y_1 + y_2 + y_3)$. Our transformation (6) represents an *amalgamation*, which preserves all population components, but lumps some of them together.

When population components are amalgamated into $g = 3$ groups, vectors of proportions can be portrayed in a graph called a *ternary diagram* (Aitchison, 1986, p. 5). As illustrated in Figure 2, the diagram begins with an equilateral triangle that has vertices labelled "1", "2", and "3". A vector $\mathbf{p} = (p_1, p_2, p_3)$ of proportions can then be represented as a point within this triangle, where the perpendicular distance to the side opposite vertex "i" is proportional to p_i. The example here shows the point $\mathbf{p} = \mathrm{C}[(3, 2, 1)] = (1/2, 1/3, 1/6)$. This lies closest to vertex "1", because of the relatively large proportion p_1 with a corresponding large distance p_1 from the opposite side between "2" and "3".

Why does this work? Figure 2 illustrates the proof. We need to show that a single constant c can be used to scale any vector $\mathbf{p}$ of three proportions and represent it in this way. Suppose that the equilateral triangle has side length a, altitude $\sqrt{3}a/2$, and area $\sqrt{3}a^2/4$. Three dotted lines from the vertices to the point representing $\mathbf{p}$ separate the entire triangle into three component triangles, with base a, altitude cp_i, and area $acp_i/2$ ($i = 1, 2, 3$). Equating the area of the triangle to the sum of these three

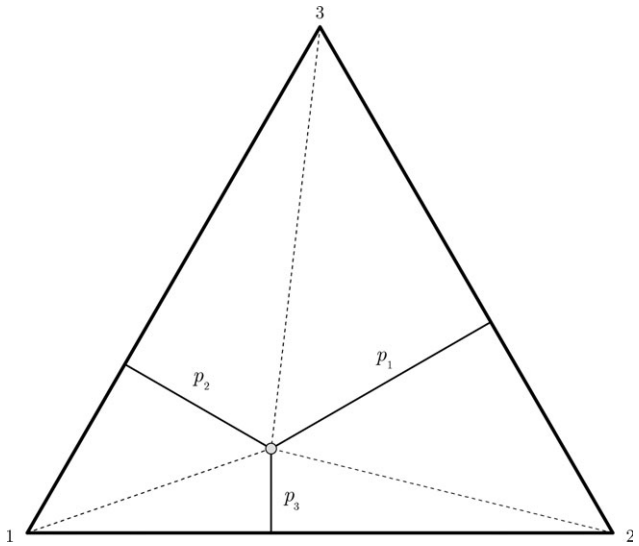

Figure 2. Ternary diagram for compositions amalgamated into $g = 3$ groups. The indicated point represents a vector of proportions. Solid lines perpendicular to the sides of an equilateral triangle have lengths proportional to p_i ($i = 1, 2, 3$). Dotted lines facilitate a proof that this method actually works, given the constraint $\Sigma_{i=1}^{3} p_i = 1$.

component areas gives

$$\frac{\sqrt{3}a^2}{4} = \frac{ac}{2}(p_1 + p_2 + p_3) = \frac{ac}{2}, \tag{11}$$

independent of the vector $\mathbf{p}$ of proportions. Consequently, the constant $c = \sqrt{3}a/2$ scales the proportions p_i to the perpendicular distances shown in Figure 2.

Ternary diagrams have the additional property that a straight line from one vertex to the opposite side corresponds to proportions $\mathbf{p}$ with a constant odds ratio between the other two components. For example, a line through vertex "3" to the side joining "1" and "2" represents proportions with a constant odds ratio p_1/p_2. Equivalently, the subcomposition $\mathrm{C}[(p_1, p_2)]$ remains the same along this line. Readers inclined towards geometry can prove this by drawing the figure and observing that any two points on the line define similar triangles with a common ratio p_1/p_2. Appendix A gives the formulae needed to construct ternary diagrams in a graphical computer language like R or S–PLUS (R Development Core Team 2005; Venables and Ripley, 2000).

Compositional statistics

The catch curve model in Table 2 predicts the proportion p_a of fish at each age a, and samples from a survey or commercial fishery give observed proportions y_a. Although the observations might be 0 for some ages a, a suitable amalgamation (6) guarantees that each age group i has a positive observed proportion y_i. The corresponding amalgamation (7) gives comparable predictions p_i. Statistical analysis requires a probability distribution $P(\mathbf{y} \mid \mathbf{p})$ that generates observation vectors $\mathbf{y} = (y_1, \ldots, y_g)$ from the predicted vector $\mathbf{p} = (p_1, \ldots, p_g)$. The distribution should have properties that link random observations to the predictions, such as the expected value

$$E[\mathbf{y}] = \mathbf{p}. \tag{12}$$

Table 3. The multinomial distribution as a statistical model for the vector of observed proportions $\mathbf{y}$, given the predicted proportions $\mathbf{p}$ and the sample size n.

$\mathbf{v} \sim \mathcal{M}(\mathbf{p}, n)$	(T3.1)
$\mathbf{y} = \dfrac{\mathbf{v}}{\sum_{j=1}^{g} v_j} = \dfrac{\mathbf{v}}{n}$	(T3.2)
$P(\mathbf{y} \mid \mathbf{p}, n) = \dfrac{n!}{(ny_1)!(ny_2)! \ldots (ny_g)!} \prod_{i=1}^{g} p_i^{ny_i}$	(T3.3)
$\ell(\mathbf{p} \mid \mathbf{y}, n) = -n \sum_{i=1}^{g} y_i \log p_i$	(T3.4)
$\mathrm{E}[y_i] = p_i$	(T3.5)
$\mathrm{Var}[y_i] = \dfrac{p_i(1 - p_i)}{n}$	(T3.6)
$\mathrm{Cov}[y_i, y_j] = -\dfrac{p_i p_j}{n}$ for $i \neq j$	(T3.7)

Perhaps the simplest probability model for catch curve analysis is the multinomial distribution (Table 3). Given a sample size n and probability vector $\mathbf{p}$, the multinomial model (T3.1) generates a vector $\mathbf{v} = (n_1, \ldots, n_g)$ representing the number n_i of fish observed in each age group i, where $n = \Sigma_i n_i$. The composition (T3.2) transforms these observations to proportions. The function $P(\mathbf{y} \mid \mathbf{p}, n)$ in (T3.3) gives the explicit probability of observing $\mathbf{y}$, given $\mathbf{p}$ and n. Statistical inference depends on the negative log-likelihood function

$$\ell(\mathbf{p}) = -\log P + K, \tag{13}$$

where K represents any convenient constant that does not depend on the parameters $\mathbf{p}$. In particular, the distribution (T3.3) and constant $K = \log n! - \Sigma_i \log[(ny_i)!]$ give the function $\ell(\mathbf{p} \mid \mathbf{y}, n)$ in (T3.4). The observations y_i ($i = 1, \ldots, g$) have means (T3.5), variances (T3.6), and covariances (T3.7) determined by the predictions p_i and sample size n, where (T3.5) corresponds to the relationship (12).

In spite of its attractive simplicity, the multinomial distribution has at least two limitations that keep it from dealing adequately with catch curve analysis. First, it assumes a definite sample size n and generates proportions y_i that necessarily take discrete fractional values in the set

$$\left\{\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}\right\}. \tag{14}$$

If the data y_i come from more complex sampling schemes, such as a stratified random sample, n may not be properly defined, and the observations y_a can take arbitrary values in the interval [0, 1]. Second, according to (T3.5)–(T3.6), the variances $\mathrm{Var}[y_i]$ should become small as the sample size n becomes large and the observations y_i should become close to their expected values $E[y_i] = p_i$.

Table 4. The Dirichlet distribution as a statistical model for the vector of observed proportions **y**, given the predicted proportions **p** and an effective sample size n. The model applies when **y** comes from the composition (T4.2) of independent variates (T4.1) drawn from gamma distributions with a common scale parameter $n > 0$. The approximate modal estimate $\hat{n}$ in (T4.5) depends on Stirling's approximation (17) to the gamma function.

$$v_i \sim \mathcal{G}(\text{shape} = np_i, \text{scale} = n) \tag{T4.1}$$

$$y_i = \frac{v_i}{\sum_{j=1}^{g} v_j} \tag{T4.2}$$

$$P(\mathbf{y} \mid \mathbf{p}, n) = \frac{\Gamma(n)}{\Gamma(np_1)\Gamma(np_2)\ldots\Gamma(np_g)} \prod_{i=1}^{g} y_i^{np_i - 1} \tag{T4.3}$$

$$\ell(\mathbf{p}, n \mid \mathbf{y}) = \sum_{i=1}^{g} [\log\Gamma(np_i) - np_i \log y_i] - \log\Gamma(n) \tag{T4.4}$$

$$\hat{n} \approx \frac{g-1}{2} \left(\sum_{i=1}^{g} p_i \log \frac{p_i}{y_i} \right)^{-1} \tag{T4.5}$$

$$E[y_i] = p_i \tag{T4.6}$$

$$\text{Var}[y_i] = \frac{p_i(1-p_i)}{n+1} \tag{T4.7}$$

$$\text{Cov}[y_i, y_j] = -\frac{p_i p_j}{n+1} \quad \text{for } i \neq j \tag{T4.8}$$

In other words, the data should fit the catch curve model almost perfectly if the sample size is large enough. In reality, the model itself is not entirely appropriate, and a realistic analysis requires a distribution that allows potentially large errors, regardless of sample size.

The composition transformation (9) suggests a general technique for modelling stochastic proportions **y**. Start by generating a random vector **v** with positive components; then set $\mathbf{y} = \text{C}[\mathbf{v}]$. The Dirichlet distribution (Table 4) results from this approach, where the vector **v** has components v_i drawn independently from the gamma distributions

$$G(v_i \mid p_i, n) = \frac{1}{n^{np_i}\Gamma(np_i)} v_i^{np_i - 1} e^{-p_i} \tag{15}$$

with shape parameters np_i and a single scale parameter n, as indicated in (T4.1). In our formulation, n plays the role of an effective sample size, where **y** becomes less variable as n increases. The model allows any value of $n > 0$, and uncertainty increases dramatically as $n \to 0$. The Dirichlet probability distribution (T4.3) formally resembles the multinomial distribution (T3.3), given the linkage

$$n! = \Gamma(n+1) \tag{16}$$

between the factorial and gamma functions. Essentially, the roles of **p** and **y** are reversed between the two distributions. For this reason, in a Bayesian framework, the Dirichlet is the conjugate prior distribution for the parameters of the multinomial (Gelman *et al.*, 1995, p. 482). The moment formulae (T4.6)–(T4.8) and (T3.5)–(T3.7) also bear a striking resemblance to each other.

The Dirichlet distribution improves the multinomial for catch curve analysis in two key ways. First, the proportions y_i are now continuous variables with $0 < y_i < 1$, rather than discrete fractions in the set (14). Second, the Dirichlet parameter n can actually be estimated from the observed data **y**, unlike the prescribed value n for the multinomial. Estimation depends on the negative log-likelihood (T4.4), where (13) defines ℓ from P with $K = \Sigma_i \log y_i$. Stirling's approximation

$$\log\Gamma(z) \approx \left(z - \frac{1}{2}\right) \log z - z + \frac{\log(2\pi)}{2} \tag{17}$$

implies the approximate modal estimate (T4.5) for n (Appendix B), where $\hat{n} \to \infty$ as the data conform more closely to the model ($y_i \to p_i$ for each i).

A third model for compositional data starts with a random vector **u** of real numbers, which are then transformed by (10) into a vector **y** of proportions. In particular, a multivariate normal vector **u** generates a logistic-normal distribution (Table 5) for **y**. Our formulation uses the simplifying assumption (T5.1) that the components u_i are drawn independently with mean $\log p_i$ and common variance σ^2. A more general approach would allow **u** to have an arbitrary covariance matrix, but the application here does not require this added level of complexity. In Appendix C, we explain precisely how the model here fits into the broader context of compositional distributions. We also prove that assumptions (T5.1) and (T5.2) give the relatively simple probability distribution (T5.3). This, combined with definition (13), gives $\ell(\mathbf{p}, \sigma \mid \mathbf{y})$ in (T5.4), where $K = -[(g-1)/2]\log(2\pi) - \sqrt{g}(\Sigma_i \log y_i)$.

Exact values (T5.6)–(T5.8) for the means and covariances of log odds ratios $\log(p_i/p_j)$ can be computed from the parameters **p** and σ, as can the modal estimate (T5.5) for σ. Notice that $\hat{\sigma} \to 0$ as $y_i \to p_i$ for each i, so that a good model fit corresponds to a small value of σ. Analytically, the moments (T5.9)–(T5.11) have only approximate values when σ is small (Appendix C), in contrast with the exact moments (T3.5)–(T3.7) for the multinomial and (T4.6)–(T4.8) for the Dirichlet. Aitchison (1986, p. 64) argues that log odds ratios have greater statistical relevance than the proportions themselves, as we discuss further in Appendix C.

Estimation and model selection

Our catch curve models generate predicted proportions $p_i(\boldsymbol{\theta})$ for a specified parameter vector $\boldsymbol{\theta}$. We can compare these predictions with observations y_i, using the multinomial, Dirichlet, or logistic-normal distributions. At best, our models represent only approximations to reality with

$$y_i \approx p_i(\boldsymbol{\theta}) \quad \text{for } i = 1, \ldots, g. \tag{18}$$

As we have discussed, the multinomial model implies that the approximation (18) approaches equality as the sample size n increases. (The variances (T3.6)–(T3.7) approach 0 as $n \to \infty$.) In effect, the multinomial allows only for measurement error, which should become negligibly small with large sample sizes.

To address this limitation, various authors have suggested the alternatives discussed here: logistic-normal (Schnute and Richards, 1995) or Dirichlet (Williams and Quinn, 1998). Each of

Table 5. The logistic-normal distribution as a statistical model for the vector of observed proportions **y**, given the predicted proportions **p** and a standard deviation σ. The model applies when **y** comes from the logistic transformation (T5.2) of independent variates (T5.1) drawn from normal distributions with a common standard deviation σ. Calculations involve the geometric means $\tilde{y}$ and $\tilde{p}$ of **y** and **p**, respectively. The modal estimate $\hat{\sigma}$ in (T5.5) is exact. The approximations (T5.9)–(T5.11) apply when σ is small.

$$u_i = \log p_i + \sigma\epsilon_i \quad \text{where } \epsilon_1 \sim \mathcal{N}(0,1) \tag{T5.1}$$

$$y_i = \frac{e^{u_i}}{\sum_{j=1}^{g} e^{u_j}} \tag{T5.2}$$

$$P(\mathbf{y}\,|\,\mathbf{p},\sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{g-1}\left(\sqrt{g}\prod_{i=1}^{g} y_i\right)^{-1} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{g}\left(\log\frac{y_i}{\tilde{y}} - \log\frac{p_i}{\tilde{p}}\right)^2\right] \tag{T5.3}$$

$$\ell(\mathbf{p},\sigma\,|\,\mathbf{y}) = (g-1)\log\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{g}\left(\log\frac{y_i}{\tilde{y}} - \log\frac{p_i}{\tilde{p}}\right)^2 \tag{T5.4}$$

$$\hat{\sigma}^2 = \frac{1}{g}\sum_{i=1}^{g}\left(\log\frac{y_i}{\tilde{y}} - \log\frac{p_i}{\tilde{p}}\right)^2 \tag{T5.5}$$

$$\mathrm{E}[\log(y_i/y_j)] = \log(p_i/p_j) \tag{T5.6}$$

$$\mathrm{Var}[\log(y_i/y_j)] = 2\sigma^2 \quad \text{for } i \neq j \tag{T5.7}$$

$$\mathrm{Cov}[\log(y_i/y_j), \log(y_j/y_k)] = \sigma^2 \quad \text{for } i \neq j \neq k \neq i \tag{T5.8}$$

$$\mathrm{E}[y_i] \approx p_i \tag{T5.9}$$

$$\mathrm{Var}[y_i] \approx \sigma^2 p_i^2\left(1 - 2p_i + \sum_{i=1}^{g} p_i^2\right) \tag{T5.10}$$

$$\mathrm{Cov}[y_i, y_j] \approx -\sigma^2 p_i p_j\left(p_i + p_j - \sum_{i=1}^{g} p_i^2\right) \quad \text{for } i \neq j \tag{T5.11}$$

these distributions involves an extra parameter related to the approximation (18), which becomes more exact as the logistic-normal σ decreases or the Dirichlet n increases. Effectively, σ and n leave room for the approximation (18) to include both measurement and process error. Even very precise observations y_i need not closely match the predictions $p_i(\boldsymbol{\theta})$. Nature's "true" model might not lie in the family defined by Table 2.

Let Θ denote the complete parameter vector: ($\boldsymbol{\theta}$, σ) for the logistic-normal model or ($\boldsymbol{\theta}$, n) for the Dirichlet model. Equations (T4.4) and (T5.4) define the negative log-likelihood function $\ell(\Theta\,|\,\mathbf{y})$ for a given data set **y**, as illustrated for the Dirichlet distribution by the calculation

$$\ell(\Theta\,|\,\mathbf{y}) = \ell(\boldsymbol{\theta}, n\,|\,\mathbf{y}) = \ell(\,p(\boldsymbol{\theta}), n\,|\,\mathbf{y}).$$

From (13), the function ℓ defines the posterior distribution

$$P(\Theta\,|\,\mathbf{y}) \propto P(\mathbf{y}\,|\,\Theta)P_0(\Theta) \propto e^{-\ell(\Theta|\mathbf{y})}P_0(\Theta), \tag{19}$$

where $P_0(\Theta)$ denotes the prior distribution for Θ.

Given a data set **y**, we want to explore models that correspond to various choices of design vector $\boldsymbol{\phi}$. From a Bayesian perspective, we could generate a posterior sample (Appendix D) of Θ from (19) for each model in question. As suggested in the discussion of catch curve models, it might suffice to look at the resulting sample distribution of Z, where all other parameters are considered nuisance parameters. Analysis would be guided by investigating how model choice influences the Z distribution.

A frequentist approach typically begins by finding the maximum likelihood estimate $\hat{\Theta}$ that minimizes the negative log-likelihood (Appendix D), where

$$\ell(\hat{\Theta}\,|\,\mathbf{y}) = \min_{\Theta}\ell(\Theta\,|\,\mathbf{y}). \tag{20}$$

As discussed by Burnham and Anderson (2002, pp. 61–66), this can be used to calculate Akaike's information criterion

$$\mathrm{AIC} = 2\ell(\hat{\Theta}\,|\,\mathbf{y}) + 2N \tag{21}$$

and a *corrected* criterion

$$\mathrm{AIC_c} = 2\ell(\hat{\Theta}\,|\,\mathbf{y}) + 2N\left(\frac{g-1}{g-N-2}\right), \tag{22}$$

where N and g denote lengths of the vectors Θ and **y**, respectively. The derivation of (21) comes from a theory in which the "true" model of nature lies outside the parametric family that defines $\ell(\Theta\,|\,y)$, a theory appropriate to the application here.

Based on information theory, the AIC measures relative distances of various proposed models to the (unknown) true model. Intuitively, the calculation (21) balances model fit (a low minimum in (20)) with model complexity (a large number N of parameters Θ). The minimum decreases as N increases, and the final term in (21) acts like a penalty associated with the number of parameters. Given a set of proposed models, the one with lowest AIC seems most credible.

The corrected $\mathrm{AIC_c}$ in (22) also involves the effective number $g-1$ of observations **y**. For large values g, the AIC and $\mathrm{AIC_c}$ are essentially the same, but the penalty in (22) increases as g decreases. Intuitively, when the number of observations is small, it takes a greater improvement of fit to justify adding a new parameter to the model. Our examples illustrate situations in which the model selected by AIC differs from that selected by $\mathrm{AIC_c}$.

Application to quillback rockfish

The genus *Sebastes* comprises a diverse group of marine fish generally known as rockfish. More than 60 rockfish species live in the northeast Pacific Ocean (Love *et al.*, 2002). According to Hart (1973), 35 of these inhabit waters along Canada's west coast in BC. The BC commercial fishery captures at least 28 rockfish species (Yamanaka *et al.*, 2004). In particular, quillback rockfish (*S. maliger*) belong to a group loosely termed "inshore rockfish" owing to their prevalence in shallow waters (0–200 m). They occupy rocky reefs, where they are caught primarily by hook and

line. Like most rockfish, quillback can live a long time, with ages recorded up to 90 years (Munk, 2001).

Historically, large populations of quillback rockfish occurred in BC coastal waters between Vancouver Island and the mainland, including the Strait of Georgia. In the late 1970s, fishers began targeting rockfish in response to developing local markets for live fish. Easily accessible populations in the Strait of Georgia became depleted, and fishing pressure shifted north to more remote areas. The fishery reached Johnstone Strait by the mid-1980s, and this prompted the introduction of a survey to target that population while it was still relatively unfished (Richards and Cass, 1987).

Handline surveys have been conducted in Johnstone Strait and adjacent waterways (126°37′W to 126°53′W, 50°32′N to 50°39′N) since 1986. Yamanaka and Richards (1993) describe surveys conducted in 1986, 1987, 1988, and 1992. In 2001, the Rockfish Selective Fishery Study (Berry, 2001) targeted quillback rockfish for experiments on improving survival after capture by hook and line. The resulting data subsequently have been incorporated into the survey data series. The most recent survey in 2004 essentially repeated the 1992 survey design.

Commercial handline fishery samples have been collected from a larger region (126°35′W to 127°39′W, 50°32′N to 50°59′N) in the years 1984–1993, 1996, and 2000–2001. Bubble plots of age proportions (Figure 3) portray all available survey data, plus commercial data (if available) from years when no surveys were conducted (1984–1985, 1989–1991, 1993, 1996, and 2000). We include the commercial data partly to complete the time-series and partly to compare data from the two sources. Commercial samples sometimes reinforce major recruitment anomalies identified by survey samples (e.g. the 1985 year class). In some years, however, a low number of commercial specimens (e.g. 1991 with $n = 50$) shows at best highly smeared age structure patterns.

Overall, Figure 3 suggests the decline of older year classes from 1984 to 2004, with a tendency for surveys to capture younger fish than the commercial fishery. Data early in the time-series indicate the consistent presence of fish with ages up to 50 and occasionally beyond. By 2004, the age distribution lies almost entirely below age 30. This apparent decline is consistent with the history of the fishery, in which 1986 was the first year of a new license category that applied direct quotas to quillback rockfish (Yamanaka *et al.*, 2004). Previously, the species was caught only as bycatch in other fisheries. The 1986 survey gives baseline data from a relatively unfished population that can be compared with data from the fished population sampled by the 2004 survey.

Ternary diagrams (Figure 4) show age proportions amalgamated into three groups (ages 0–12, 13–20, 21–69)

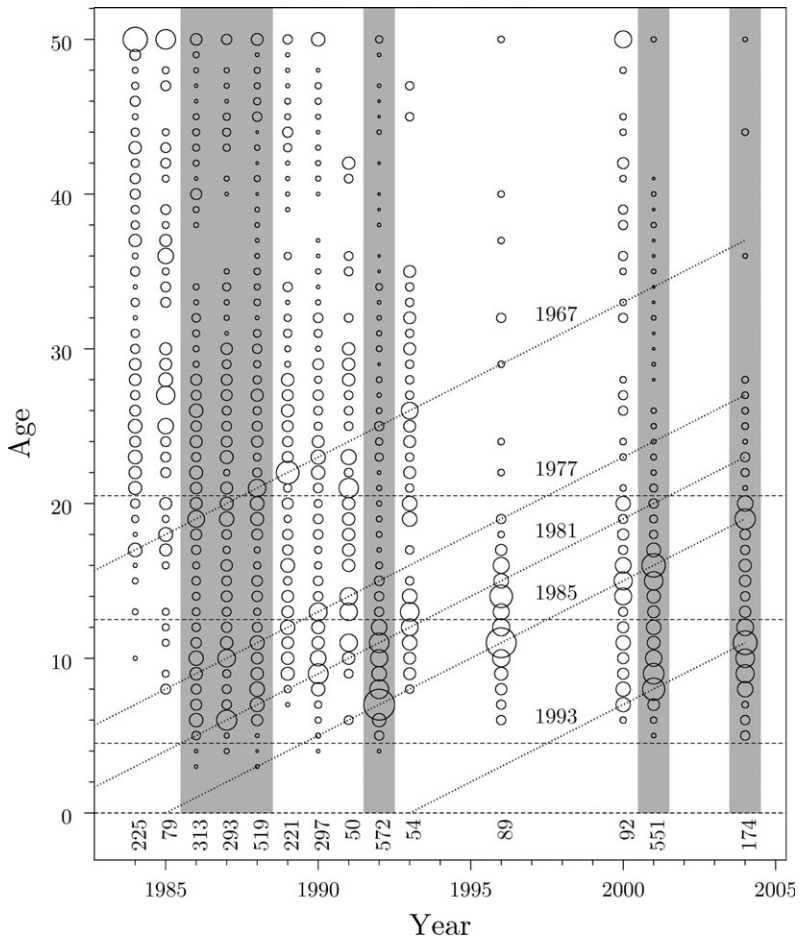

Figure 3. Bubble plot representing observed age proportion y_a in various years for quillback rockfish in Johnstone Strait, BC, Canada. Background shading indicates columns with data from surveys, and the remaining data come from the commercial fishery. Diagonal lines highlight possible strong cohorts born in the years indicated. Numbers below the horizontal line at age 0 show the number n of fish aged each year.

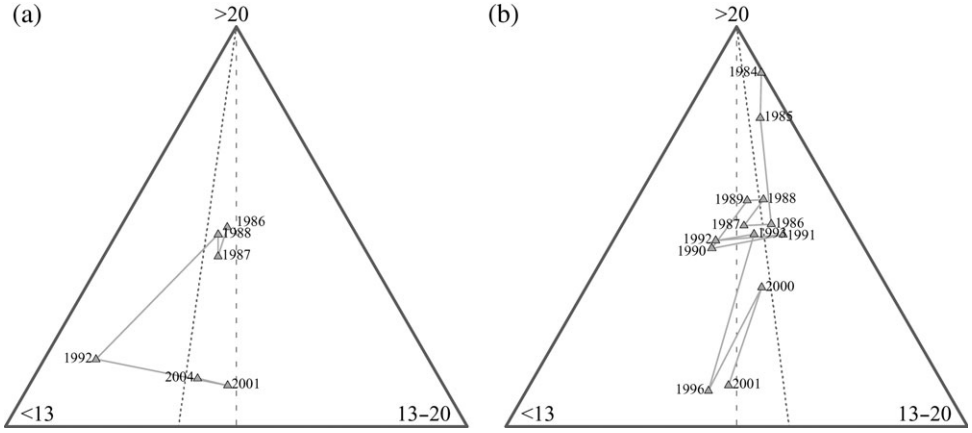

Figure 4. Ternary diagrams for data from (a) surveys and (b) commercial fisheries, including some commercial data not represented in Figure 3. Ages a are amalgamated into three groups defined by $0 < a \leq 12$, $12 < a \leq 20$, $20 < a \leq 69$. Within each panel, a dashed line indicates points with equal proportions in the first two age groups. A dotted line represents points where the ratio between these two age groups equals the geometric mean of ratios in the observed data.

corresponding to fish that could be considered young, mid-aged, and old. We distinguish survey data (Figure 4a) from commercial data (Figure 4b), where Figure 4b portrays commercial data from all available years. The time trend in both panels shows a pronounced movement away from vertex 3 (old fish). A large influx of new recruits from the 1985 year class appears as a high proportion p_1 near vertex 1 for the survey year 1992 (Figure 4a). This shift to younger fish does not appear in the commercial data (Figure 4b), probably because the fishery is less selective for young fish.

Each panel of Figure 4 shows a vertical dashed line from vertex "3" to the base of the triangle. Along this line, the proportions p_1 and p_2 of young and mid-aged fish are equal (i.e. $p_1/p_2 = 1$). Dotted lines are also drawn from vertex 3 to indicate a constant ratio $p_1/p_2 = \tilde{p}_1/\tilde{p}_2$, where $\tilde{p}_i$ denotes the geometric mean of observed proportions p_i in group i. All points on this line have a constant ratio between young and mid-aged fish. The survey shows a larger proportion of young fish, so that the dotted line lies left of the dashed line in Figure 4a. In contrast, the dotted line lies right of the dashed line in Figure 4b, indicating a larger proportion of mid-aged fish in the fishery.

We use the start and end years, 1986 and 2004, of the survey series to illustrate the models described here. Table 6 lists design parameters $\boldsymbol{\phi}$ for three examples. Two of these use the 1986 data, with either $m = 2$ or $m = 4$ recruitment anomalies, and the third uses the 2004 data with $m = 2$ anomalies. Our analyses involve the model parameters $\boldsymbol{\theta}$, plus an additional parameter n for the Dirichlet distribution or σ for the logistic-normal. We constrain each parameter to lie in a specified interval (Table 7). From a frequentist perspective, these constraints bound the region of interest in parameter space. In a Bayesian context, they define uniform prior distributions.

For two reasons, we like to start our analyses using uniform priors. First, in this context, the maximum likelihood estimate corresponds exactly to the posterior mode. Second, a uniform prior makes it fairly easy to see the impact of the prior on the posterior. The data provide at least some information about a parameter if its posterior distribution lies well within the prior interval. By contrast, the data appear uninformative if the posterior lies rather evenly scattered across the interval.

As described earlier, the model in Table 2 has four cases associated with possible combinations of survival, selectivity, and recruitment anomalies. Combining these with the three examples in Table 6 gives a total of ten distinct analyses, six for the 1986 data and four for the 2004 data. Tables 8 and 9 list maximum likelihood parameter estimates obtained from the Dirichlet and logistic-normal distributions, respectively. Figure 5 provides some insight into the biological interpretation of the estimates in Table 8. Vertical bars in Figures 5a, 5c, and 5e represent the observed age proportions y_a for the three examples in Table 6, where both Figures 5a and 5c portray the 1986 data. Curves in each panel show estimated age proportions $\hat{p}_a$ associated with the four cases of the model. Corresponding curves in Figures 5b, 5d, and 5f portray the cumulative probability distributions for the data and the proportions estimated from each case of the model.

Rockfish age structure data typically suggest multiple recruitment anomalies, as in the bubble plot of Figure 3. The 1986 survey data shown in Figures 5a and 5c suggest more anomalies than the 2004 survey data in Figure 5e, although the latter spans a smaller group of ages because of the reduced presence of old fish. Our three examples (Table 6) are motivated by patterns in the data. Cumulative model estimates $\hat{p}_a$ match the cumulative observations y_a in 1986 with $m = 4$ anomalies (Figure 5d) and in 2004

Table 6. Design parameters $\boldsymbol{\phi}$ in (1) for three examples based on quillback rockfish data from surveys in 1986 and 2004.

Example	Year	k	A	B	b_0	m	b_1	b_2	b_3	b_4	g
1	1986	5	60	200	20	2	10	19	–	–	37
2	1986	5	60	200	20	4	10	19	26	40	37
3	2004	5	28	200	20	2	11	19	–	–	15

Table 7. Constraints on the parameters $\boldsymbol{\theta}$, n, and σ used in all analyses in this paper.

Parameter	Z	β_k	α	ρ_i	τ	n	σ
Minimum	0	0	2	0	0	10	0
Maximum	1	1	25	20	5	1000	3

Table 8. Maximum likelihood estimates of parameters $\boldsymbol{\theta}$ in (2) and n from the Dirichlet distribution for the three examples in Table 6 and Cases 1–4 of the catch curve model. Analyses apply to survey data from 1986 (Examples 1 and 2) and 2004 (Example 3). AIC and AIC_c values are computed from (21)–(22). For comparison with $\hat{n}$, examples 1 and 2 have sample size $n = 313$; example 3 has $n = 174$.

Example	Case	$\hat{Z}$	$\hat{\beta}_k$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	$\hat{\rho}_4$	$\hat{\tau}$	$\hat{n}$	ℓ_{min}	AIC	AIC_c
1	1	0.056	–	–	–	–	–	–	–	97	41.6	87.2	87.6
	2	0.066	0.34	2.00	–	–	–	–	–	114	37.5	83.0	84.3
	3	0.055	–	–	0.55	2.59	–	–	0.64	127	37.0	84.0	86.0
	4	0.063	0.41	2.00	0.61	1.87	–	–	0.57	143	33.9	81.7	88.7
2	3	0.055	–	–	0.70	2.36	3.47	1.39	1.80	200	27.5	68.9[a]	72.9[b]
	4	0.058	0.62	25.00	0.43	2.03	3.08	1.27	1.64	204	27.0	72.0	78.9
3	1	0.095	–	–	–	–	–	–	–	26	20.6	45.3	46.4
	2	0.157	0.03	2.00	–	–	–	–	–	54	13.7	35.4	39.9
	3	0.084	–	–	5.62	6.18	–	–	1.59	135	7.03	24.1	31.6[b]
	4	0.132	0.10	3.93	2.23	4.64	–	–	0.67	274	1.94	17.9[a]	36.6

[a]Selected by AIC.
[b]Selected by AIC_c.

Table 9. Maximum likelihood estimates of parameters $\boldsymbol{\theta}$ in (2) and σ from the logistic-normal distribution for the three examples in Table 6 and Cases 1–4 of the catch curve model. Analyses apply to survey data from 1986 (Examples 1 and 2) and 2004 (Example 3). AIC and AIC_c values are computed from (21)–(22).

Example	Case	$\hat{Z}$	$\hat{\beta}_k$	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\rho}_2$	$\hat{\rho}_3$	$\hat{\rho}_4$	$\hat{\tau}$	$\hat{\sigma}$	ℓ_{min}	AIC	AIC_c
1	1	0.050	–	–	–	–	–	–	–	0.64	1.83	7.67	8.03
	2	0.060	0.32	2.00	–	–	–	–	–	0.59	−0.86	6.28	7.57
	3	0.047	–	–	0.47	2.33	–	–	0.58	0.60	−0.39	9.21	11.2
	4	0.056	0.38	2.00	0.61	1.71	–	–	0.51	0.56	−2.56	8.88	12.9
2	3	0.048	–	–	0.46	1.74	2.58	1.68	1.87	0.49	−7.68	−1.35[a]	2.65[b]
	4	0.050	0.57	25.00	0.27	1.53	2.34	1.57	1.75	0.48	−8.14	1.72	8.65
3	1	0.103	–	–	–	–	–	–	–	0.78	3.48	11.0	12.0
	2	0.162	0.02	2.00	–	–	–	–	–	0.56	−1.17	5.66	10.1
	3	0.083	–	–	4.85	6.43	–	–	1.66	0.35	−7.65	−5.30	2.20[b]
	4	0.135	0.09	3.86	1.98	4.62	–	–	0.69	0.25	−12.3	−10.7[a]	7.99

[a]Selected by AIC.
[b]Selected by AIC_c.

with $m = 2$ anomalies (Figure 5f). Estimates of Z appear fairly robust to model choices, with $\hat{Z}$ varying from 0.047 to 0.060 in 1986 (Table 8, Examples 1 and 2) and from 0.083 to 0.162 in 2004 (Table 8, Example 3). In particular, these estimates suggest a higher total mortality Z in 2004.

For the 1986 survey data, the AIC and AIC_c values in Tables 8 and 9 suggest choosing Case 3 with $m = 4$ recruitment anomalies. Both the Dirichlet and logistic-normal models lead to the same conclusion, although with somewhat different parameter estimates. The 1986 data set has a relatively large number of observations from $g = 37$ age groups (Table 6). By contrast, the survey data set for 2004 has only $g = 15$ groups, because of the reduced number of fish ages present in the sample. The AIC indicates Case 4 with $m = 2$ anomalies, but the AIC_c suggests Case 3. Intuitively, the number of observations is too small to justify the extra selectivity parameters (α, β_k). Again, the Dirichlet and logistic-normal models produce the same conclusions.

Figure 6 represents a Bayes sample drawn from the Dirichlet posterior distribution for Example 3 (1986 data), Case 4. Unlike the point estimates listed in Table 8 and portrayed in Figure 5, this analysis examines the uncertainty in all seven parameters (Z, β_k, α, ρ_1, ρ_2, τ, n). Modal values from Table 8 appear as triangles in Figure 6, and these often lie near the edge of the scatterplot. Mean parameter values, shown as squares, are generally in a more central position. Scatterplots for the selectivity parameters (α, β_k) indicate very little structure, with each parameter varying almost like the uniform prior distribution across the range specified in Table 7. This result reinforces the preference for Case 3, based on the AIC and AIC_c in Table 8.

Histograms in the diagonal panels of Figure 6 show varying degrees of skewness in the model parameters, where the distribution of Z appears almost normal. Loess lines suggest that Z is inversely related to all other model parameters except n. In particular, the perceived value of Z will decrease as the parameters (ρ_1, ρ_2, τ) become larger, corresponding to higher and broader recruitment anomalies.

In the framework here, a choice of design parameters $\boldsymbol{\phi}$ determines a particular model structure. That, combined with a data set, a Dirichlet or logistic-normal error distribution, and a choice of prior distributions implies a posterior distribution for Z. How much does the choice of model influence this distribution?

Figures 7a and 7b answer this question for Examples 2 and 3 (the 1986 and 2004 survey data), respectively. Four cases of the model, coupled with a choice of Dirichlet or logistic-normal error, give eight outcomes for each example. We portray the

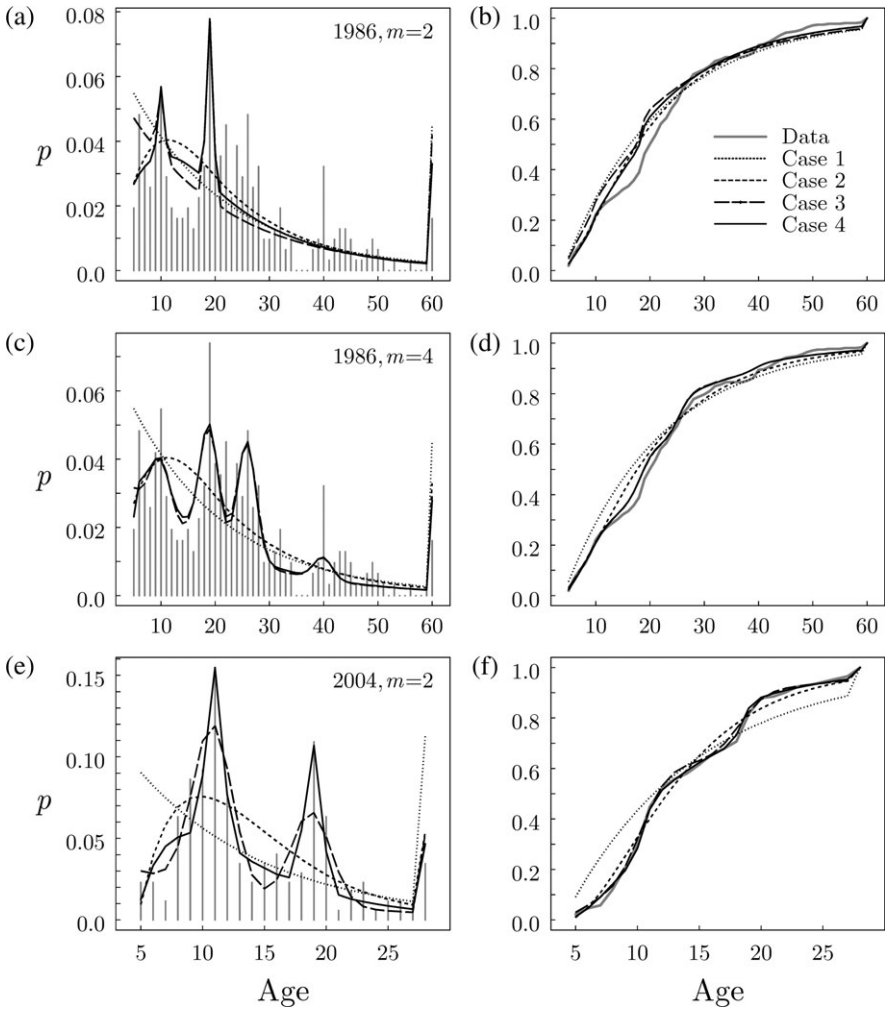

Figure 5. Observed and estimated proportions y_a and $\hat{p}_a$ from the posterior mode of the Dirichlet distribution for Examples 1–3 in Table 6, and Cases 1–4 of the catch curve model. (a), (c), (e) Results for Examples 1–3, respectively, where vertical lines represent y_a, and curves show $\hat{p}_a$ for Cases 1–4 (dotted, short-dashed, long-dashed, solid). (b), (d), (f) Cumulative curves corresponding to the proportions in (a), (c), (e).

resulting Z distributions as boxplots, which often overlap considerably and therefore suggest similar conclusions about Z. Logistic-normal estimates usually have a wider distribution than those obtained from the Dirichlet. Consistent with Figure 6, the mode $\hat{Z}$ sometimes lies in the periphery of the Z distribution, typically in the upper range for the examples here. Case 4 of the model, with all components present, tends to give the most consistent posteriors between the two error assumptions.

Figure 7 also allows us to address an essential question for the biology of this stock. Has Z changed between 1986 and 2004? Dotted lines in Figures 7a and 7b represent the mean of all Z values in both panels. Boxplots from 1986 lie almost entirely below this line, whereas those from 2004 lie predominantly above. This suggests a higher Z in 2004 than in 1986, and we carry the analysis one step further. For each year, we have equal sized, independent samples from the two distributions. Thus, we can obtain a sample distribution for the ratios

$$r = \frac{Z_{2004}}{Z_{1986}}. \tag{23}$$

These lie almost entirely above 1 for all models and choices of error distribution (Figure 7c). Given the assumptions of our catch curve model, Z has apparently increased by a factor near 2 for *S. maliger* in Johnstone Strait over the two decades since a quota fishery was initiated.

Discussion

Our model formalizes issues with catch curve analysis raised by Ricker (1975, Chapter 2). Different choices of the design vector $\boldsymbol{\phi}$ produce different posterior distributions for Z. If these all appear similar, then the information on Z is robust to the choice of model. If not, then the analyst can examine reasons why different models lead to different interpretations. Graphs such as those in Figure 5 act as diagnostic tools for assessing model fit at the maximum likelihood estimate or posterior mode. Age distribution plots (e.g. Figure 5a) and cumulative curves (e.g. Figure 5b) show how well the model fits the observed data.

Both our theory and the worked examples show the relevance of techniques from compositional data analysis. Visual methods, such as bubble plots and ternary diagrams, help reveal patterns in

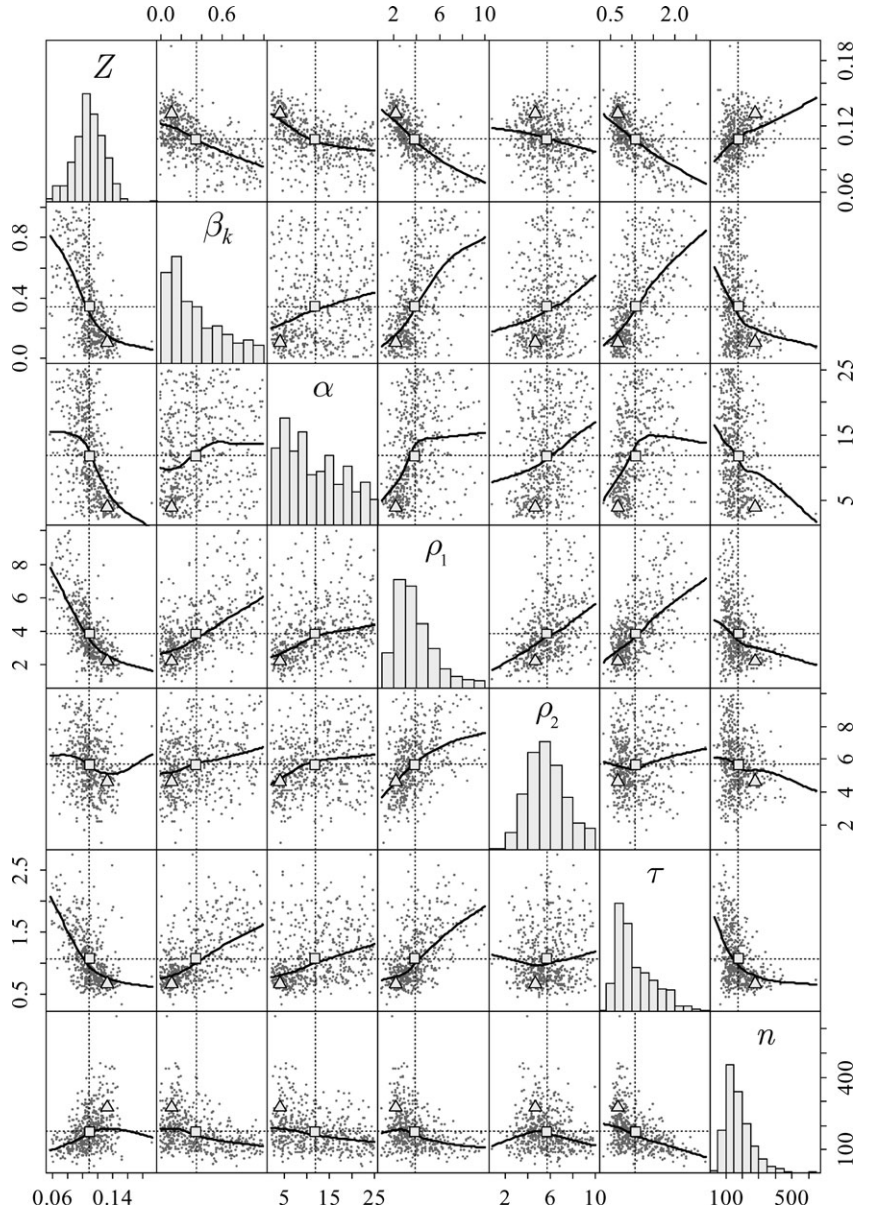

Figure 6. Pairs plot showing 500 sample points from the posterior Dirichlet distribution for Example 3, Case 4. These come from a systematic subsample of every 200th point in a convergent Markov chain with length 100 000. Panels on the diagonal contain histograms representing the distributions of individual parameters. Other panels show a loess line through the scatterplots for parameter pairs. Triangles and squares designate points corresponding to the posterior mode and mean, respectively.

the data. The Dirichlet and logistic-normal distributions give rigour and biological meaning to the results, where estimates n and σ provide measures of model error. A good model fit corresponds to a high value of n or a low value of σ. Appendix D describes software to implement our analyses.

The model in Table 2 deals only with scenarios limited to constant effects of survival, selectivity, and recruitment. It could be extended in various ways, such as allowing a distinct standard deviation τ_h for each recruitment anomaly at age b_h. The ages b_h could also be treated as free parameters, similar to parameters that determine modes of a mixture distribution (Schnute and Fournier, 1980). Historical patterns could be represented as systematic trends in natural or fishing mortality. Such models might be used to examine the problems with catch curve analysis discussed by Quinn and Deriso (1999), cited in the introduction.

As models become more complex, the number N of parameters increases. The AIC provides a guide for evaluating model fit, given the value of N. Moreover, the AIC_c takes account of the effective number $g-1$ of observations. Statistical analysis requires more observations than parameters; the denominator on the right side of (22) is positive only if $g-1 > N+1$. This constraint highlights an essential issue for catch curve analysis. The available data may not be adequate to estimate all parameters of interest. We propose the framework here as an exploratory tool for investigating potential biological interpretations of limited data sets. The examples in

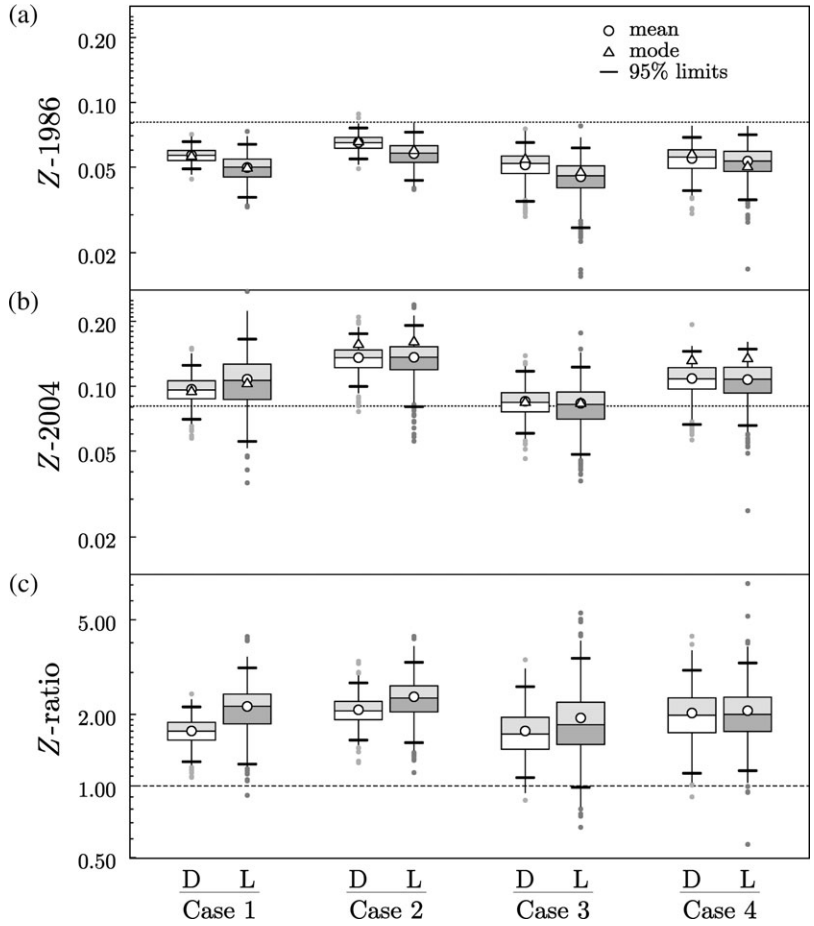

Figure 7. Boxplots of 500 sample values Z from the posterior distribution for (a) Example 2 (1986) and (b) Example 3 (2004). As in Figure 6, these come from a systematic subsample of a convergent Markov chain with length 100 000. Each panel shows boxplots for Cases 1–4, based on the Dirichlet (D) and logistic-normal (L) error distributions. A dotted line represents the mean $\bar{Z}$ of all samples in (a) and (b). (c) Boxplots of 500 ratios Z_{2004}/Z_{1986} from values Z in (a) and (b). These lie almost entirely above the value 1, highlighted by a dashed line. Bounds marked as "95% limits" correspond to the 2.5 and 97.5% quantiles of the posterior distribution.

Tables 8 and 9 and in Figure 7 by no means exhaust the possibilities for the 1986 and 2004 survey data sets.

The likelihood (T3.3) for the multinomial distribution can be calculated even if $y_i = 0$ for some observations i ($0! = 1$ and $p^0 = 1$ for $p > 0$). However, for reasons discussed earlier, the multinomial fails to give an appropriate statistical description of the data vector $\mathbf{y}$. The more realistic Dirichlet and logistic-normal distributions both require $y_i > 0$ for every observation i. We use an amalgamation (one of several common compositional transformations) to obtain data that meet this requirement. A good general strategy gives as many groups as possible, subject to a specified minimum value for the amalgamated proportions y_i. To some extent, an amalgamation smooths the observed data by combining small values of the observed proportions y_a. Excessive smoothing corresponds to using a small value of g, which weakens a test based on the AIC_c.

More complete data, such as the age structures for two decades in Figure 3, suggest using a full catch-at-age analysis that tracks the population through time. For example, Schnute and Richards (1995) describe such analyses with a logistic-normal model for the observed age proportions. Here we have focused on separate analyses for each year, with a comparison between years early and late in the time-series. If we assume that the 1986 stock had experienced very little fishing mortality ($F = 0$), then the mortality ratio

$$r = \frac{F + M}{M} = 1 + \frac{F}{M} \tag{24}$$

in (23) relates simply to the ratio between fishing and natural mortality. In particular, the value $r = 2$ corresponds to a policy of setting fishing mortality F at the same level as natural mortality M.

The Bayesian approach in Figures 6 and 7 illustrates a focus on a single key parameter in a model with many additional nuisance parameters. In this case, the total mortality Z or the ratio (24) act like reference parameters that could be linked closely with policy decisions. Schnute and Haigh (2006) describe a similar theory of management strategies, where the allowable catch quota Q acts as the key parameter. Catch curve analyses might play a similar strategic role in helping define policies associated with a target mortality ratio F/M. The framework here provides a direct link between model designs, encapsulated by design vectors $\boldsymbol{\phi}$, and policy outcomes that depend on estimated mortalities or mortality ratios.

Acknowledgements

We thank the teams of field workers who, under the leadership of Laura Richards and Lynne Yamanaka, collected data on inshore rockfish along the coast of BC, Canada. Carl Schwarz provided useful background on the history and application of compositional data analysis. Terry Quinn, Franz Mueter, and a third reviewer (anonymous) provided helpful comments leading to an improved final draft.

References

Aitchison, J. 1985. A general class of distributions on the simplex. Journal of the Royal Statistical Society, Series B, 47: 136–146.

Aitchison, J. 1986 (reprinted in 2003). The Statistical Analysis of Compositional Data. The Blackburn Press, Caldwell, NJ. 416 pp.

Aitchison, J., and Shen, S. M. 1980. Logistic-normal distributions: some properties and uses. Biometrika, 67: 261–272.

Baranov, F. I. 1918. [On the question of the biological basis of fisheries.] Izvestâ Otdela rybovodstva i naučno-promyslovyh issledovanii, 1(1): 81–128 (in Russian; cited after Ricker, 1975).

Berry, M. D. 2001. Area 12 (Inside) Rockfish Selective Fishery Study. Science Council of British Columbia, Project Number FS00–05.

Burnham, K. P., and Anderson, D. R. 2002. Model Selection and Multivariate Inference: a Practical Information—Theoretic Approach, 2nd edn. Springer Science and Business Media, Inc., New York, NY. 488 pp.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. 1995 (reprinted in 2000). Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton, FL. 526 pp.

Hart, J. L. 1973 (reprinted in 1988). Pacific fishes of Canada. Bulletin of the Fisheries Research Board of Canada, 180. 740 pp.

Love, M. S., Yoklavich, M., and Thorsteinson, L. 2002. The Rockfishes of the Northeast Pacific. University of California Press. 405 pp.

Munk, K. M. 2001. Maximum ages of groundfishes in waters off Alaska and British Columbia and considerations of age determination. Alaska Fisheries Research Bulletin, 8(1): 12–21.

Quinn II, T. J., and Deriso, R. B. 1999. Quantitative Fish Dynamics. Oxford University Press, New York, NY. 542 pp.

R Development Core Team. 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Richards, L. J., and Cass, A. J. 1987. 1986 Research catch and effort data on nearshore reef-fishes in British Columbia statistical areas 12, 13, and 16. Canadian Manuscript Report of Fisheries and Aquatic Sciences, 1903. 119 pp.

Ricker, W. E. 1975. Computation and interpretation of biological statistics of fish populations. Bulletin of the Fisheries Research Board of Canada, 191. 382 pp.

Schnute, J. T. 2006. Curiosity, recruitment, and chaos: a tribute to Bill Ricker's inquiring mind. Environmental Biology of Fishes, 75: 95–110.

Schnute, J. T., Boers, N., and Haigh, R. 2004. PBS Mapping 2: User's Guide. Canadian Technical Report of Fisheries and Aquatic Sciences, 2549. 126 pp.

Schnute, J. T., Couture-Beil, A., and Haigh, R. 2006. PBS Modelling 1: User's Guide. Canadian Technical Report of Fisheries and Aquatic Sciences, 2674. 111 pp.

Schnute, J., and Fournier, D. 1980. A new approach to length-frequency analysis: growth structure. Canadian Journal of Fisheries and Aquatic Sciences, 37: 1337–1351.

Schnute, J. T., and Haigh, R. 2006. Reference points and management strategies: lessons from quantum mechanics. ICES Journal of Marine Science, 63: 4–11.

Schnute, J. T., and Richards, L. J. 1995. The influence of error on population estimates from catch-age models. Canadian Journal of Fisheries and Aquatic Sciences, 52: 2063–2077.

Venables, W. N., and Ripley, B. D. 2000. S Programming. Springer, New York, NY. 264 pp.

Williams, E. H., and Quinn, T. J. 1998. A parametric bootstrap of catch-age compositions using the Dirichlet distribution. Fishery stock assessment models for the 21st century. Alaska Sea Grant College Program, AK-SG-98-01: 371–384.

Yamanaka, K. L., Lacko, L. C., Lochead, J. K., Marin, J., Haigh, R., Grandin, C., and West, K. 2004. Stock assessment framework for inshore rockfish. Canadian Science Advisory Secretariat, Research Document, 2004/068. 63 pp.

Yamanaka, K. L., and Richards, L. J. 1993. 1992 Research catch and effort data on nearshore reef-fishes in British Columbia Statistical Area 12. Canadian Manuscript Report of Fisheries and Aquatic Sciences, 2184. 77 pp.

Appendix A

Ternary diagrams

This appendix gives the formulae required to draw ternary diagrams, such as those in Figures 2 and 4. The proofs, not given here, depend on standard arguments from analytical geometry. In *xy*-coordinates with a 1:1 aspect ratio, start by drawing the equilateral triangle with vertices (0, 0), (1, 0), and $(1/2, \sqrt{3}/2)$. This corresponds to the choices a=1 and $c = \sqrt{3}/2$ in (11). The point **p** has coordinates

$$(x, y) = \frac{1}{2}\left(2p_2 + p_3, \sqrt{3}p_3\right). \tag{A.1}$$

For each vertex "i", let (x_i, y_i) denote the point on the opposite side of the triangle, where a perpendicular line connects that side to $\mathbf{p} = (x, y)$. Then

$$(x_1, y_1) = \frac{1}{4}\left(3 + x - \sqrt{3}y, \sqrt{3}(1 - x) + 3y\right), \tag{A.2}$$

$$(x_2, y_2) = \frac{1}{4}\left(x + \sqrt{3}y, \sqrt{3}x + 3y\right), \tag{A.3}$$

$$(x_3, y_3) = (x, 0). \tag{A.4}$$

For distinct indices (i, j, k) and a given odds ratio $r_{jk} = p_j/p_k$, define **p** by the coordinates

$$p_i = 0, \quad p_j = \frac{r_{jk}}{1 + r_{jk}}, \quad p_k = \frac{1}{1 + r_{jk}}. \tag{A.5}$$

Then a line through vertex "i" with constant odds ratio p_j/p_k connects "i" to the point (x, y) in (A.1) determined by **p** in (A.5).

Appendix B

Proofs: Dirichlet distribution

Gelman *et al*. (1995, pp. 476–477) give the expressions (T3.3) and (T4.3) for the multinomial and Dirichlet distributions. Their version of the Dirichlet uses parameters α_i equivalent to np_i here. They also give moment formulae equivalent to (T4.6)–(T4.8). Applying Stirling's approximation (17) to the expression (T4.4)

for ℓ gives (after algebraic simplification)

$$\ell \approx \ell' = n \sum_{i=1}^{g} p_i \log \frac{p_i}{y_i} - \frac{g-1}{2} \log n + K,$$

where the constant $K = (1/2)[(g-1) \log 2\pi - \Sigma_i \log p_i]$ does not depend on n. The estimate $\hat{n}$ that minimizes this approximation satisfies the derivative condition $d\ell'/dn = 0$, that is

$$\sum_{i=1}^{g} p_i \log \frac{p_i}{y_i} - \frac{g-1}{2\hat{n}} = 0. \tag{B.1}$$

Solving (B.1) for $\hat{n}$ gives (T4.5).

Appendix C

Proofs: logistic-normal distribution

The logistic-normal distribution (T5.3) appears less commonly in the literature than the Dirichlet, particularly in the simplified form given here. At the end of this appendix, we sketch the key steps in its derivation. The underlying model (T5.1) and (T5.2) implies that

$$\log\left(\frac{y_i}{y_j}\right) = \log\left(\frac{p_i}{p_j}\right) + \sigma(\varepsilon_i - \varepsilon_j). \tag{C.1}$$

Taking expected values of both sides of (C.1) gives the result (T5.6). Similarly, for $i \neq j$

$$\mathrm{Var}\left[\log\left(\frac{y_i}{y_j}\right)\right] = \sigma^2 \,\mathrm{Var}[\varepsilon_i - \varepsilon_j] = 2\sigma^2,$$

because ε_i and ε_j are independent. This proves (T5.7). For three distinct indices (i, j, k), another moment formula comes from the calculation

$$\begin{aligned}\mathrm{Cov}\left[\log\left(\frac{y_i}{y_k}\right), \log\left(\frac{y_j}{y_k}\right)\right] &= \sigma^2 \mathrm{E}[(\varepsilon_i - \varepsilon_k)(\varepsilon_j - \varepsilon_k)] \\ &= \sigma^2 \mathrm{E}[\varepsilon_i\varepsilon_j - \varepsilon_i\varepsilon_k - \varepsilon_j\varepsilon_k + \varepsilon_k^2] = \sigma^2,\end{aligned}$$

again because ε has independent components. This proves (T5.8).

Aitchison (1986, p. 64) argues forcefully that log odds ratios, like those in (T5.6)–(T5.8), have more relevance to statistical analysis than the original proportions. Although the moment formulae (T4.6)–(T4.8) for the Dirichlet might look attractive, the notion of covariance seems a bit inappropriate for variates such as y_i confined to the interval (0, 1). In an amusing analogy, he points out that a barbecue, although an excellent tool in the wide open spaces of North America (comparable to real space in g dimensions), might not be a suitable concept for cooking in a cramped Hong Kong apartment (comparable to the simplex associated with g proportions). For small values σ in the model here, a power series expansion in the nonlinear transformation (T5.1) and (T5.2) gives

$$y_i \approx p_i \left[1 + \sigma\left(\varepsilon_i - \sum_{j=1}^{g} p_j \varepsilon_j\right)\right]. \tag{C.2}$$

The approximate moments (T5.9)–(T5.11) follow from (C.2), after some algebra not shown here. Readers can test these results using a small value σ in the simulation (T5.1) and (T5.2).

We come finally to the derivation of the likelihood (T5.3). Because the observed vector $\mathbf{y}$ sums to 1, the probability distribution properly belongs to a region in $g-1$ dimensions. The subvector $\mathbf{y}_{-g} = (y_1, \ldots, y_{g-1})$ defines a suitable region, namely the simplex where $y_i > 0$ and $\Sigma_{i=1}^{g-1} y_i < 1$. (By convention, a negative subscript indicates removal of a particular component from a vector.) We transform $\mathbf{y}$ to an equivalent vector $\mathbf{z}$ of real numbers with $z_g = 0$, where the transformation and its inverse are

$$z_i = \log\left(\frac{y_i}{y_g}\right), \tag{C.3}$$

$$y_i = \frac{e^{z_i}}{\sum_{j=1}^{g} e^{z_j}}. \tag{C.4}$$

These also define a map between $\mathbf{y}_{-g}$ and $\mathbf{z}_{-g}$, where it is understood that $y_g = 1 - \Sigma_{i=1}^{g-1} y_i$ in (C.3) and $z_g = 0$ in (C.4).

It follows from (C.1) that

$$z_i = \log\left(\frac{p_i}{p_g}\right) + \sigma(\varepsilon_i - \varepsilon_g). \tag{C.5}$$

The assumed normality of ε implies that $\mathbf{z}_{-g}$ is multivariate normal. Aitchison (1986, p. 115, Formula 6.3) shows that

$$\begin{aligned}P(\mathbf{y}_{-g}) = (2\pi)^{-(g-1)/2} |\mathbf{V}|^{-1/2} \tilde{y}^{-g} \\ \exp\left[-\frac{1}{2}(\mathbf{z}_{-g} - \boldsymbol{\mu})^T \, \mathbf{V}^{-1} (\mathbf{z}_{-g} - \boldsymbol{\mu})\right],\end{aligned} \tag{C.6}$$

where $\boldsymbol{\mu}$ and $\mathbf{V}$ denote the mean vector and covariance matrix of $\mathbf{z}_{-g}$. These have dimensions $(g-1)\times 1$ and $(g-1) \times (g-1)$, respectively. The result (C.6) takes account of the Jacobian

$$\frac{\partial \mathbf{z}_{-g}}{\partial \mathbf{y}_{-g}} = \left(\prod_{i=1}^{g} y_i\right)^{-1} = \frac{1}{\tilde{y}^g}. \tag{C.7}$$

Aitchison (1986, p. 77) calls $\mathbf{V}$ the *logratio covariance matrix*, although his notation differs somewhat from ours.

The moment formulae (T5.6)–(T5.8) and the definition (C.3) allow us to evaluate $\boldsymbol{\mu}$ and $\mathbf{V}$ explicitly. In particular, (T5.6) gives

$$\mu_i = \log\left(\frac{p_i}{p_g}\right). \tag{C.8}$$

Similarly, (T5.7) and (T5.8) imply that $V_{ii} = 2\sigma^2$ and $V_{ij} = \sigma^2$ when $i \neq j$. In matrix notation,

$$\mathbf{V} = \sigma^2(\mathbf{I}_{g-1} + \mathbf{J}_{g-1}), \tag{C.9}$$

where $\mathbf{I}_{g-1}$ is the identity matrix, $\mathbf{J}_{g-1}$ is a matrix with all entries 1, and subscripts emphasize that these matrices have dimension $(g-1) \times (g-1)$. Moreover, the determinant and inverse of $\mathbf{V}$ can be calculated explicitly to give

$$|\mathbf{V}| = g\sigma^{2(g-1)}, \tag{C.10}$$

$$\mathbf{V}^{-1} = \sigma^{-2}\left(\mathbf{I}_{g-1} - \frac{1}{g}\mathbf{J}_{g-1}\right). \tag{C.11}$$

The result (C.11) allows the quadratic form in the exponential factor of (C.6) to be written explicitly in terms of the residuals $\mathbf{x} = \mathbf{z}_{-g} - \boldsymbol{\mu}$:

$$\mathbf{x}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{x} = \sum_{i=1}^{g-1}\sum_{j=1}^{g-1} x_i x_j\left(\delta_{ij} - \frac{1}{g}\right) \tag{C.12}$$

$$= \sum_{i=1}^{g-1} x_i^2 - \frac{1}{g}\sum_{i=1}^{g-1} x_i \sum_{j=1}^{g-1} x_j$$

$$= \sum_{i=1}^{g} x_i^2 - \frac{1}{g}\left(\sum_{i=1}^{g} x_i\right)^2 \tag{C.13}$$

$$= \sum_{i=1}^{g} x_i^2 - \frac{g^2\bar{x}^2}{g} \tag{C.14}$$

$$= \sum_{i=1}^{g} x_i^2 - g\bar{x}^2 = \sum_{i=1}^{g}(x_i - \bar{x})^2. \tag{C.15}$$

The symbol δ_{ij} in (C.12) refers to the usual indicator variable, with $\delta_{ii} = 1$ and $\delta_{ij} = 0$ when $i \neq j$. By definition, $z_g = 0$ in (C.3), and similarly (C.8) implies $\mu_g = 0$. Thus $x_g = z_g - \mu_g = 0$, and this extension of $\mathbf{x}$ allows us to extend the range of summation from $g-1$ to g in (C.13). The mean $\bar{x}$ in (C.14) refers to the full vector $(x_1, \ldots, x_g)$, where

$$\bar{x} = \bar{z} - \bar{\mu} = \log\left(\frac{\tilde{y}}{y_g}\right) - \log\left(\frac{\tilde{p}}{p_g}\right). \tag{C.16}$$

Because $x_i = z_i - \mu_i = \log(y_i/y_g) - \log(p_i/p_g)$, it follows from (C.16) that

$$x_i - \bar{x} = \log\left(\frac{y_i}{\tilde{y}}\right) - \log\left(\frac{p_i}{\tilde{p}}\right). \tag{C.17}$$

Combining (C.6), (C.10), (C.12)–(C.15), and (C.17) gives the logistic-normal distribution in (T5.3).

Appendix D

Implementation methods

Analyses such as those described here require appropriate software. We used "AD Model Builder" (http://otter-rsch.com/admodel.htm) to obtain modal parameter estimates and to generate Bayes posterior distributions. This software package primarily requires code to calculate the negative log-likelihood ℓ, as defined here by (T4.4) and (T5.4). Constraints such as those in Table 7 can readily be incorporated. The package automates calculation of the posterior mode, and it generates posterior Markov chain samples that start at the mode. For the examples presented here, formal tests (not shown) indicated rapid convergence. We believe that samples of size 500 drawn systematically from a chain of length 100 000 give reasonable representations of the posterior, although other examples might require longer chains. In part, we restricted posterior sample sizes to produce legible graphics in Figure 6. We used R and S–PLUS (Venables and Ripley, 2000) to generate all Figures.

Users interested in software available without charge can investigate our library "PBS Modelling" (Schnute *et al.*, 2006) available on the Comprehensive R Archive Network (CRAN, http://cran.r-project.org/). We include an example that uses native R code to obtain maximum likelihood estimates. We also generate posterior samples using the R "BRugs" library and the program "OpenBUGS" (Bayesian inference Using Gibbs Sampling, http://mathstat.helsinki.fi/openbugs/). This includes native support for the Dirichlet distribution. Furthermore, as described in Appendix C, suitable transformations convert the logistic-normal distribution into a multivariate normal distribution, which is also supported by "OpenBUGS".

Another CRAN package "PBS Mapping" (Schnute *et al.*, 2004) supports the fixed aspect ratios needed to produce ternary diagrams from the formulae in Appendix A.

doi:10.1093/icesjms/fsl024