New Zealand Fisheries
Assessment Report
2006/43
November 2006
ISSN 1175-1584

Some recent problems in New Zealand orange roughy assessments

R. I. C. C. Francis

Some recent problems in New Zealand orange roughy assessments

R. I. C. C. Francis

NIWA
Private Bag 14901
Wellington

New Zealand Fisheries Assessment Report 2006/43
November 2006

Published by Ministry of Fisheries
Wellington
2006

ISSN 1175-1584

Citation:
Francis, R.I.C.C. (2006).
Some recent problems in New Zealand orange roughy assessments.
New Zealand Fisheries Assessment Report 2006/43. 65 p.

This series continues the informal
New Zealand Fisheries Assessment Research Document series
which ceased at the end of 1999.

EXECUTIVE SUMMARY

Francis, R.I.C.C. (2006). Some recent problems in New Zealand orange roughy assessments. ***New Zealand Fisheries Assessment Report 2006/43.*** **65 p.**

This report addresses four problem areas that were identified in a recent review of methods and data used in New Zealand orange roughy stock assessments. The first concerned age estimates. Analyses of all available data suggested there was substantial drift in estimates (so that ages estimated at one time may be biased relative to those estimated at a different time). This was deemed serious enough to preclude the use of most of the existing age data until the problem can be resolved. The analyses also revealed several other issues of interest or concern. There appear to be differences between age-length keys in two adjacent areas within the Chatham Rise. It is unclear why counts to the transition zone (TZ) in one sample were significantly lower than those in another sample taken from the same area (the Spawning Box) six years earlier. In some data sets, counts to the TZ were not recorded for some otoliths that must have had TZs, and the incidence of this varied greatly between data sets. This is of some concern because it appears to be connected with relative bias in age estimates.

The second problem concerned observer length frequency data, which show clear differences in mean length between trips sampled from the same area in the same year. Because only a few trips are sampled from each area in any year this causes two problems: correlation within length frequencies (LFs), and the potential for aliasing when these LFs are used in stock assessments (i.e., the assessment model may wrongly interpret a change in LFs caused by these between-trip differences as being evidence of a change in the population mean length). The extent of these problems is demonstrated and recommendations are made to avoid their adversely affecting assessments.

The third problem concerned the relationship between a_{mat}, the age at which 50% of fish are mature, and a50, that at which 50% are vulnerable to the fishery. Some recent assessments have suggested that a50 is much greater than a_{mat}. However, an examination of all observer and research data suggested that a50 is similar to, or slightly less than, a_{mat}. Other analyses showed why assessment-model estimates of a50 are sometimes unstable and also highlighted some uncertainties associated with estimates of a_{mat}. These results support the current practice of assuming a50 = a_{mat} in assessments. However, it is argued that it would be better to do this by setting a50 equal to the value of a_{mat} estimated from counts to the TZ (the so-called *sel2mat* approach), rather than the currently preferred approach of setting a_{mat} equal to the value of a50 estimated using length or age frequencies (the *mat2sel* approach).

The final problem concerned the question of whether it would be possible to continue the time series of trawl surveys in the Spawning Box. This series was discontinued in the mid 1990s when it was concluded that changes in the distribution of biomass within the survey area made it unlikely that further surveys would be sufficiently precise. An examination of data from a combined trawl-acoustic survey in 2005 found no evidence to modify this conclusion.

1. INTRODUCTION

Two workshops were held in Wellington, in October 2005 and February 2006, to review methods and data used in orange roughy stock assessments in New Zealand. These were prompted by a series of problems that had been encountered in these assessments in the preceding years. This report describes analyses that were carried out, in association with that review, to explore four quite distinct problem areas: age data, observer length data, the relationship between maturity and selectivity, and the possible continuation of a trawl survey series. Each of the following sections addresses one of those problems and is complete in itself. The work described in Section 4 was funded from MFish project ORH200504; the remainder was funded from an MFish project, SAP200504, associated with the above review.

2. AGE DATA

Until recently, only indirect use was made of age data in New Zealand orange roughy assessments. Growth and natural mortality parameters were derived from these data and used in assessment models, where they were assumed to be known without error. The first use of age frequency data in an assessment (Smith et al. 2002) involved data from only one year. It became apparent that such data could be more useful if two or more age frequencies separated by many years were available from the same area. To this end, several batches of otoliths from various areas were aged in 2003, and these data began to be used in assessments in 2004 (Dunn 2005a, McKenzie 2005). Contrary to expectations, these data did not appear to strengthen the assessments and were typically not well fitted. In 2005, the reliability of these data became doubtful when substantial differences were noticed between data sets from the same area in consecutive years (Hicks 2005b).

The aim of the analyses presented in this section was to examine all age data sets from New Zealand orange roughy for information relating to their reliability.

2.1 The data

The main data sets analysed here fall into eight sets, with each set having been read by the same institution at one time. Five sets were read by the Central Ageing Facility (CAF) (Table 1) and three by NIWA (Table 2). On two occasions some otoliths that had been read by one institution were reread by the other to allow cross-calibration (batch 163 in Table 1, and set NIWA3 in Table 2). Other data involving multiple readings of the same otoliths are discussed below in Section 2.5.

2.2 The main problem

The main problem identified in these data is that samples collected from the same area and season in three consecutive years show clear differences in their age-length keys (ALKs) (Figure 1). We can characterise these differences using a single multiplicative factor for each pairwise comparison (see Appendix 1). The differences are substantial: NECR03 = 0.82 NECR02, and NECR04 = 0.70 NECR02. This means, for example, that fish of a given length are typically 30% younger in the NECR04 sample than in NECR02.

Table 1: Description of the five sets of otoliths read by the Central Ageing Facility (CAF).

			Number of			
Set	Report date	Batch	otoliths	Label	Reader(s)	Otolith source
CAF1	Feb 03	137	413	NECR02	Corey	*San Waitaki* survey
		138	795	MEC02	Corey	7 commercial trips
CAF2	Sep 03	141	599	MEC90	Corey	market samples, 1989-91
CAF3	Nov 03	142	457	NWCR82	Corey	ktn8201
		143	551	NWCR92	Corey	tan9206
		144	572	NWCR02	Corey	tan0208
CAF4	Dec 03	154	593	NECR03.a	Corey	*Amaltal* survey
		155	800	NECR03.s	Corey	*San Waitaki* survey
CAF5	Mar 05	163	160	NECR84,90	Corey & Simon[1]	subset of NIWA2 (Table 3)
		164	302	NECR04	Corey	tan0408 (females only)

[1] Half of the otoliths were first read by each reader, and some otoliths were read a second time by the other reader. Readings were made from new CAF preparations of the sister otolith to that read by NIWA.

Table 2: Description of the three sets of otoliths read by NIWA

		Number of			
Set	Reading date	otoliths	Label	Reader(s)	Otolith source
NIWA1	Before Feb 2001	627	NECR84,90	Pete	buc8401, cor9002
NIWA2	Early 2004	762	NECR84,90	Pete & Di[1]	buc8401, cor9002
NIWA3	Early 2004	50	NECR02, MEC02	Pete	subset of CAF1 (Table 1)

[1] Pete read the cor9002 otoliths, Di read those from buc8401, and some otoliths were read a second time by the other reader. Readings were made from the original CAF preparations.

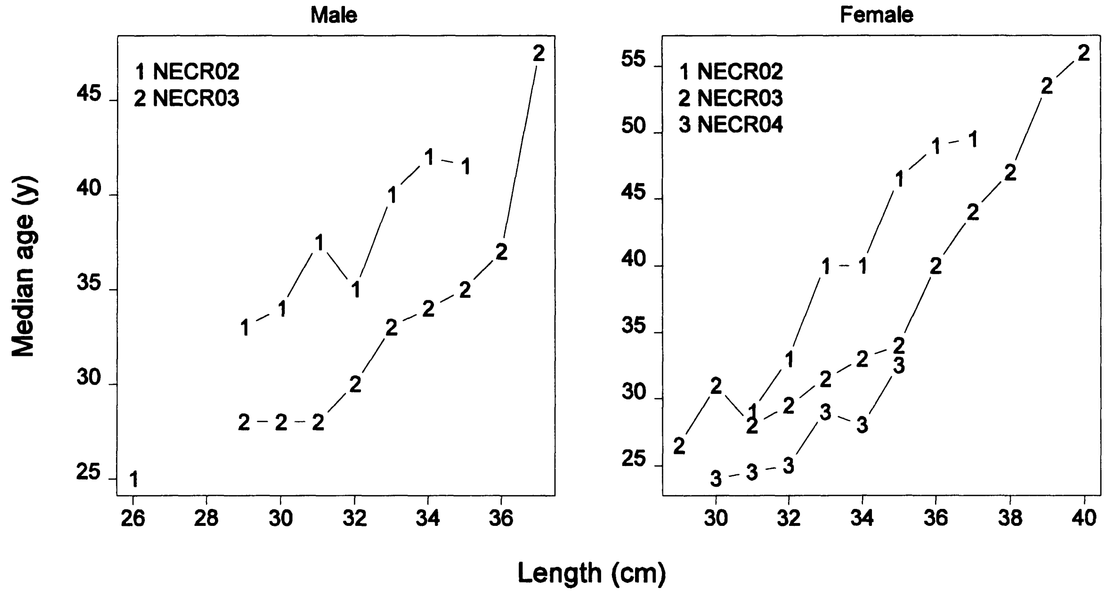

Figure 1: Differences in age-length keys (ALKs) between three samples collected from the same area (the Spawning Box – a subarea of NECR) in three consecutive years. Ages are from sets CAF1, CAF4, and CAF5 (Table 1). For each set, the plotted points are median ages for all combinations of length and sex for which there were at least 10 fish. Note that set NECR04 contained only female fish.

It seems most unlikely that these differences are real. For a slow-growing species like orange roughy it's not possible for the ALK for a population to change substantially from year to year. Apparent differences could occur if the samples came from different populations, or different parts of the same population. To avoid confounding of this sort the comparison in Figure 1 was restricted to fish caught in the Spawning Box. With this restriction all otoliths from 2002 and 2003 were from within or near the spawning plume towards the eastern end of the Box, though those from 2004 were spread right across the Box.

The simplest hypothesis that explains the patterns in Figure 1 is that there is drift in the age readings. That is, otolith readings in one set are biased relative to those in another set, and the above multiplicative factors are estimates of relative bias.

2.3 Exploring the drift hypothesis

A graphical comparison of ALKS within and between sets (Figure 2) is useful. Within-set differences are usually small. This suggests that ALKs do not usually differ much between areas or years. This, together with the fact that the CAF between-set differences are often large, adds support to the hypothesis of drift in CAF readings. The one NIWA between-set difference is small, so there is no evidence of drift in the NIWA readings (the NIWA3 set could not be included in these comparisons because of its small sample size). However, it should be added that the ability to detect drift in NIWA readings was very limited because so few sets of otoliths had been read by NIWA.

These conclusions are generally supported by the estimates of Table 3. Within-set differences are usually small (i.e., m is close to 1), as is the between-set difference for NIWA readings. However, the CAF between-set differences are often markedly larger. They suggest that readings were highest in set CAF2 and lowest in CAF5. Each between-set difference is likely to be partly real (i.e., caused by actual between-year or between-area differences in ALKs) and partly caused by relative bias in age readings. The contribution from relative bias is likely to be significant because within-set differences are usually smaller than those between sets.

There is one comparison that goes against the trend in Figure 2 and Table 3. This is the large within-set difference ($m = 0.82$) in CAF5, which is markedly greater than all the other within-set differences and similar to the larger between-set differences. It is hard to know what to make of this (between-reader differences do not seem to be responsible).

From a stock-assessment point of view there is one further comparison that is of great interest, and that is between the two otolith batches from the MEC area (Figure 3). The estimated relative difference, MEC02 = 0.86 MEC90, is not small. Whether this could be a real change in ALK is hard to say, given that the samples are 12 years apart and may be from different parts of the MEC area. It's worth noting that the difference is in the right direction (as older fish are removed in the fishing-down process we expect the median age at length to drop). However, the possibility that the difference could be at least partly due to relative bias is worrying from a stock-assessment point of view. It's of interest to note that samples covering a 20-year period from the NWCR area show no great differences (see within CAF3 comparisons in Figure 2 and Table 3).

The difficulty of interpreting these data is emphasised by the comparisons in Figure 4, which show that there appear to be geographic and/or maturity-status differences in ALKs within a single sample. The estimated differences are of similar magnitude – ECR04 = 0.91 Box04 and Imm04 = 0.88 Mat04 – and both seem to be statistically significant (the reductions in *S* were 13.9 and 15.3, respectively). It doesn't seem possible that these differences could have been caused by relative bias, so they must be real. In these comparisons, the two factors – geography and maturity – are confounded, so we cannot say whether just one or both are responsible for the differences. Two more detailed comparisons were made in which these factors were separated: ECRmat04 = 0.84 Boxmat04 (reduction in *S* = 10.0) and Boximm04 = 0.93 Boxmat04 (reduction = 2.4). These suggest that geography was the more important factor. Further comparisons were hindered by the small number of mature fish in ECR (the mature/immature sample sizes were 80/80 in Box and 98/24 in ECR).

Table 3: Pairwise estimates of the multiplicative difference, *m*, within and between sets of readings. An estimate *m* = 0.9 means that fish of a given length are typically 10% younger in group 2 than in group 1. Large reductions in *S* indicate high confidence that *m* differs from 1; where the reduction is greater than 2 the estimated difference is statistically significant (marked by *).

			Multiplicative	Reduction
Comparison	Group 1	Group 2	difference, *m*	in *S*
within CAF1	NECR02	MEC02	0.99	0.6
within CAF2	MEC89	MEC90	0.96	1.7
	MEC89	MEC91	0.92^*	6.1
	MEC90	MEC91	0.94^*	4.0
within CAF3	NWCR82	NWCR92	1.01	0.2
	NWCR82	NWCR02	0.92^*	21.5
	NWCR92	NWCR02	0.92^*	21.5
within CAF4	NECR03.a	NECR03.s	1.01	0.5
within CAF5	NECR84,90	NECR04	0.82^*	73.8
within NIWA1	NECR84	NECR90	1.06^*	9.4
within NIWA2	NECR84	NECR90	1.06^*	4.3
CAF between sets	CAF1	CAF2	1.16^*	100.3
	CAF1	CAF3	0.84^*	222.1
	CAF1	CAF4	0.83^*	403.7
	CAF1	CAF5	0.77^*	304.9
	CAF2	CAF3	0.82^*	97.1
	CAF2	CAF4	0.78^*	466.5
	CAF2	CAF5	0.75^*	88.8
	CAF3	CAF4	1.01	1.0
	CAF3	CAF5	1.00	0.0
	CAF4	CAF5	0.94^*	15.8
NIWA between sets	NIWA1	NIWA2	0.99	0.4

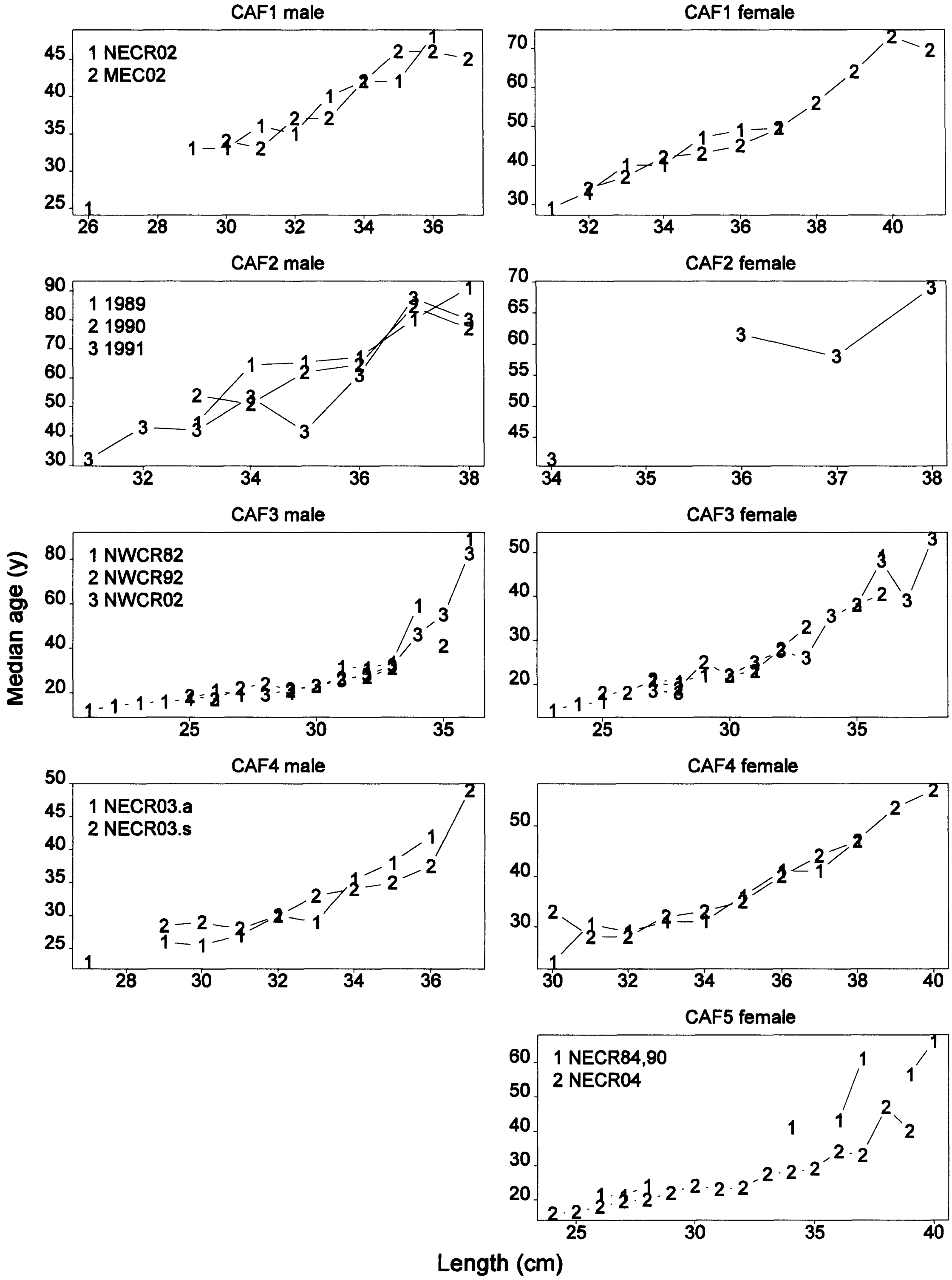

Figure 2A: Within-set comparisons of ALKs for CAF age estimates. Median ages are plotted for all length-sex combinations with at least 10 fish (or 5 fish for CAF5).

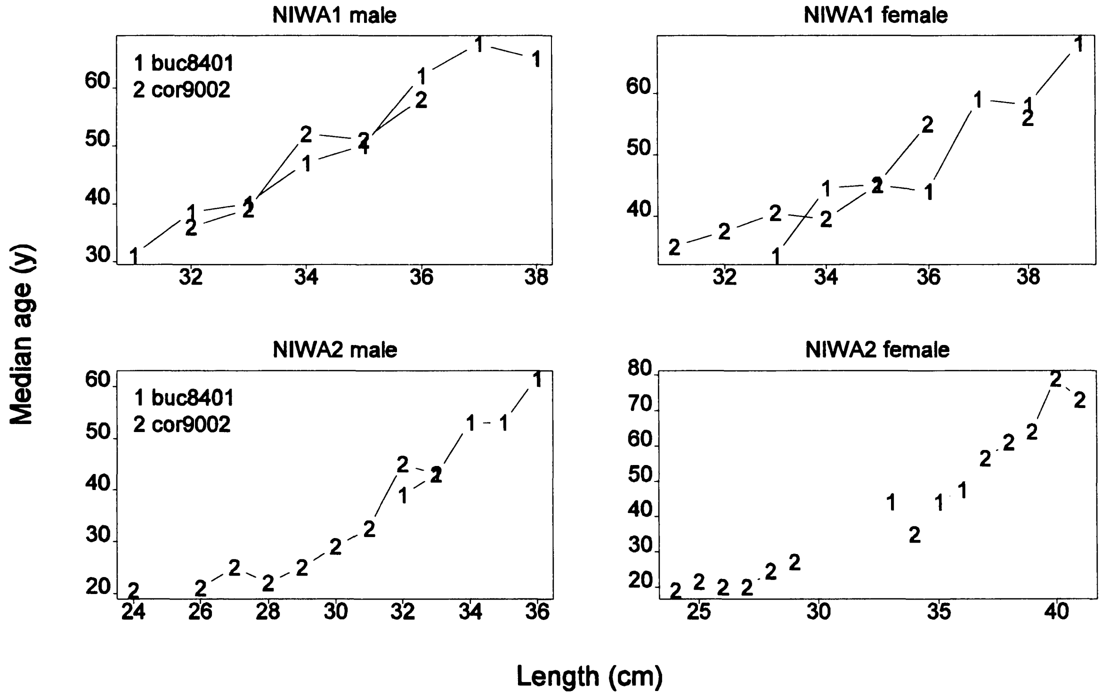

Figure 2B: Within-set comparisons of ALKs for NIWA ages. Median ages are plotted for all length-sex combinations with at least 10 fish.

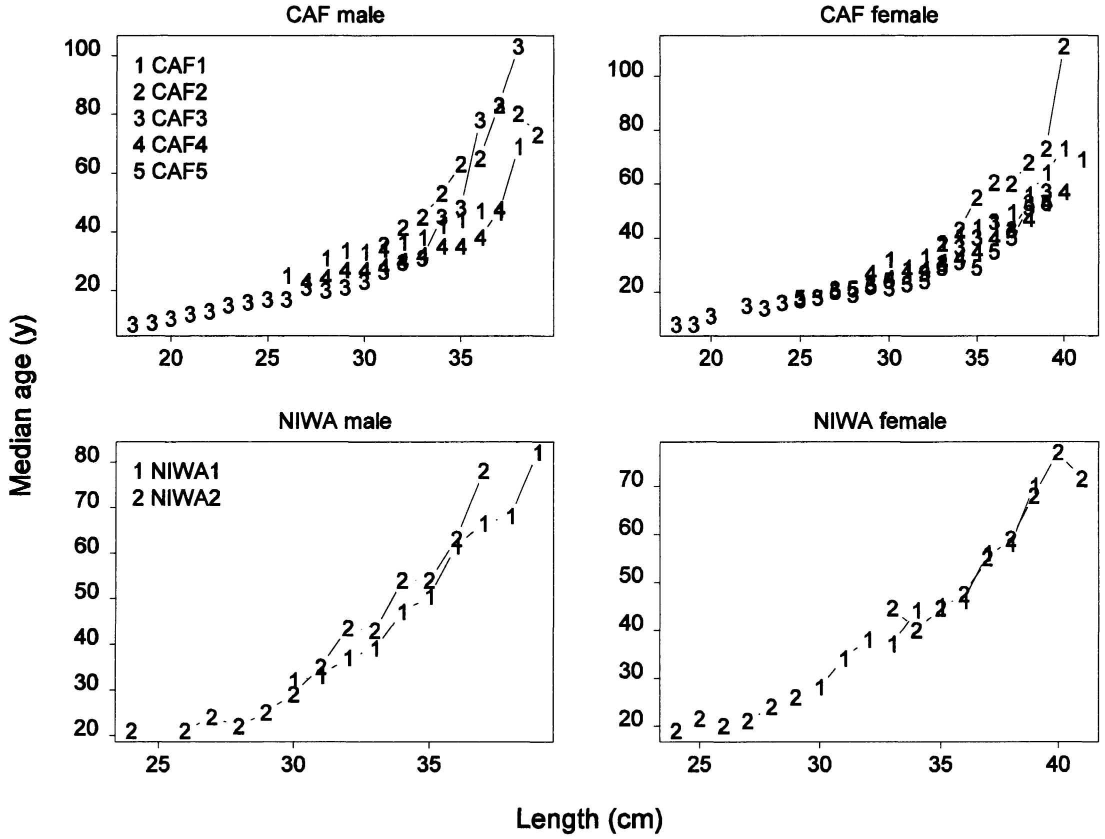

Figure 2C: Between-set comparisons of ALKs for ages from CAF (upper panels) and NIWA (lower panels). Median ages are plotted for all length-sex combinations with at least 10 fish.

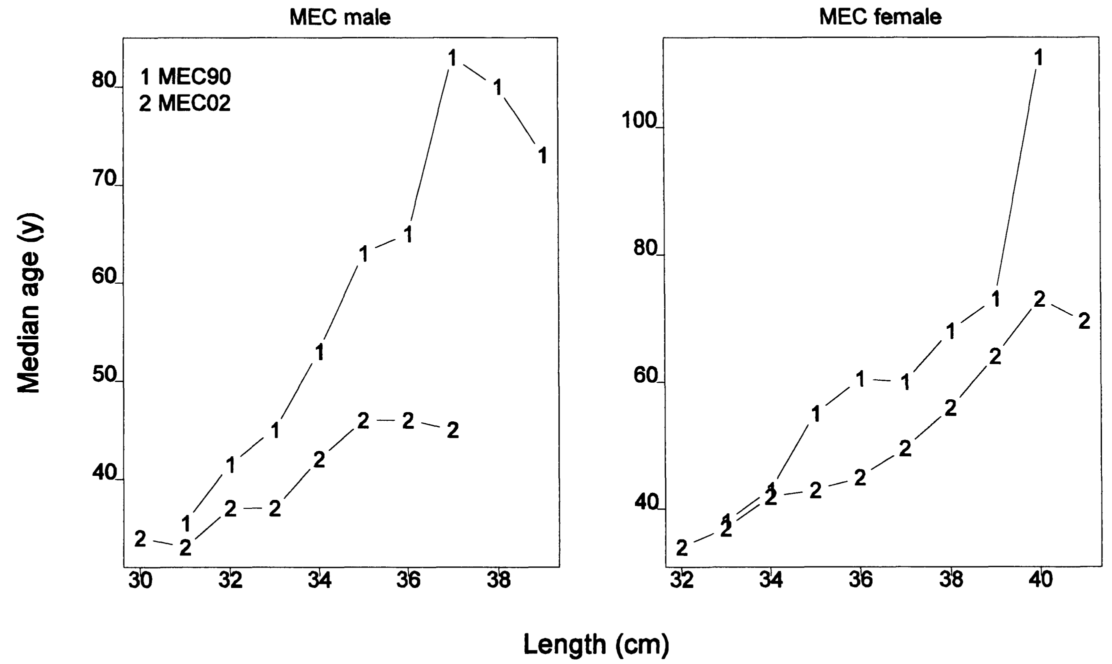

Figure 3: Differences in age-length keys (ALKs) between two samples collected from area MEC. Ages are from sets CAF1 and CAF2. For each year, median ages are plotted for all length-sex combinations with at least 10 fish.

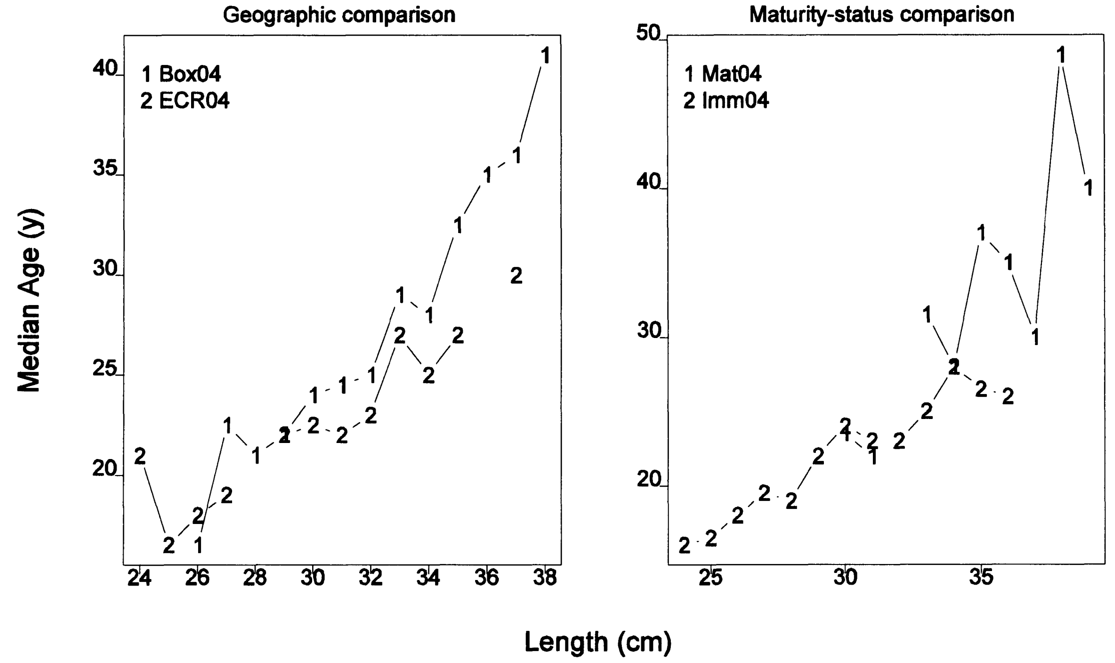

Figure 4: Differences in age-length keys (ALKs) within the NECR04 sample: a geographic difference (Box vs ECR, left panel), and a maturity-status difference (right panel). Ages are from CAF5 and all fish are females. For each year, median ages are plotted for all length-sex combinations with at least 5 fish.

2.4 Some transition zone (TZ) problems

NIWA's mean counts to the TZ vary systematically, both between trips and between sets: counts are consistently higher for trip buc8401 and for set NIWA1 (Table 4). Both between-trip-within-set and both between-set-within-trip comparisons are statistically significant (Mann-Whitney tests). The reason for these differences is unclear. The latter could be caused by drift, but not the former. Note that these counts record events (presumed maturation of fish) that occurred decades before the samples were collected so we cannot simply explain the between-trip differences by saying that fish were maturing 1 to 2 years earlier in 1990 than they were in 1984.

Table 4: NIWA's mean counts to the transition zone (TZ) by set and trip.

	buc8401	cor9002
NIWA1	30.4	29.2
NIWA2	29.3	27.3

It is not straightforward to make similar comparisons for CAF data because of CAF's different approach to recording counts to the TZ. In the NIWA data, the lack of a count to the TZ always indicated that no TZ was seen. This was not so with the CAF data. It is clear that there were many otoliths in which a CAF reader saw a TZ but did not record a count to it, presumably because its precise position was unclear. Indeed, this is sometimes (but apparently not always) recorded in a comment. Thus, for CAF readings, it is not always clear whether the reader saw a TZ. The difference between the data from the two institutions is very apparent in plots of the frequency with which counts to the TZ was recorded (Figure 5). This between-institution difference in procedure may derive, at least to some extent, from the difference in methods of preparing the otoliths. However, it is not clear why the probability that a count to the TZ would be recorded varied markedly between sets from CAF. In the remainder of this section I will use the phrase 'detect the TZ' as a shorthand meaning simply that a count to the TZ was recorded.

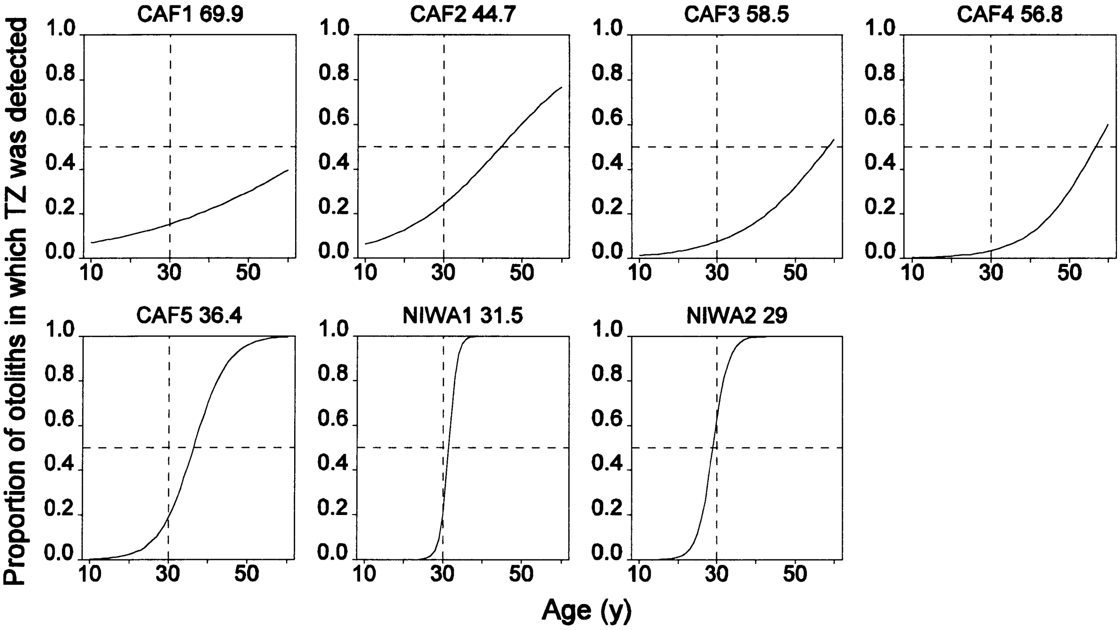

Figure 5: Proportion of otoliths in which the TZ was 'detected' (i.e., a count to the TZ was recorded) plotted against age for each set of readings. Curves are logistic functions fitted to each data set and the interpolated age at which the TZ is detected in 50% of otoliths is given above each panel. The broken lines, the same in each panel, were added to aid comparison between panels.

The second NIWA-CAF inter-calibration data set (CAF batch 163, plus the corresponding otoliths in NIWA2) shows a disturbing pattern related to the TZ. For most otoliths there was no clear bias between institutions. However, if we restrict attention to the 22 otoliths for which NIWA detected a TZ but CAF did not, we find that the CAF age was always less than the NIWA age, and sometimes substantially (the average ratio was 0.77) (Figure 6). Given that rings are more closely spaced outside the TZ than inside it, it is not surprising that a reader who does not see a TZ in an otolith will get a lower ring count than one who does. Thus it must be of concern if there is a consistent difference between how often the TZ is seen by readers from different institutions. We don't know the extent to which this happens, but whatever the case, the strong relative bias in the lower left panel of Figure 6 is of concern.

This problem did not appear in the first NIWA-CAF inter-calibration data set (set NIWA3, plus the corresponding otoliths in CAF1). NIWA detected a TZ in all these otoliths and the relative bias was not much affected by whether CAF also detected the TZ (Figure 7).

It is noteworthy that the rate of detection of TZs by CAF was much higher in CAF5 (which included the inter-calibration data) than in their earlier sets (see Figure 5), but it's not clear why. However, it *is* clear that even in this last set there were some otoliths which must have contained TZs but for which CAF recorded no count to the TZ. This must be true because the highest count to the TZ (by either institution) in the inter-calibration data set was 39 y, but using curves like those in Figure 5, applied just to the inter-calibration data, the estimated proportion of fish of age 40 y in which the TZ was detected was only 0.65 for CAF (and 0.99 for NIWA).

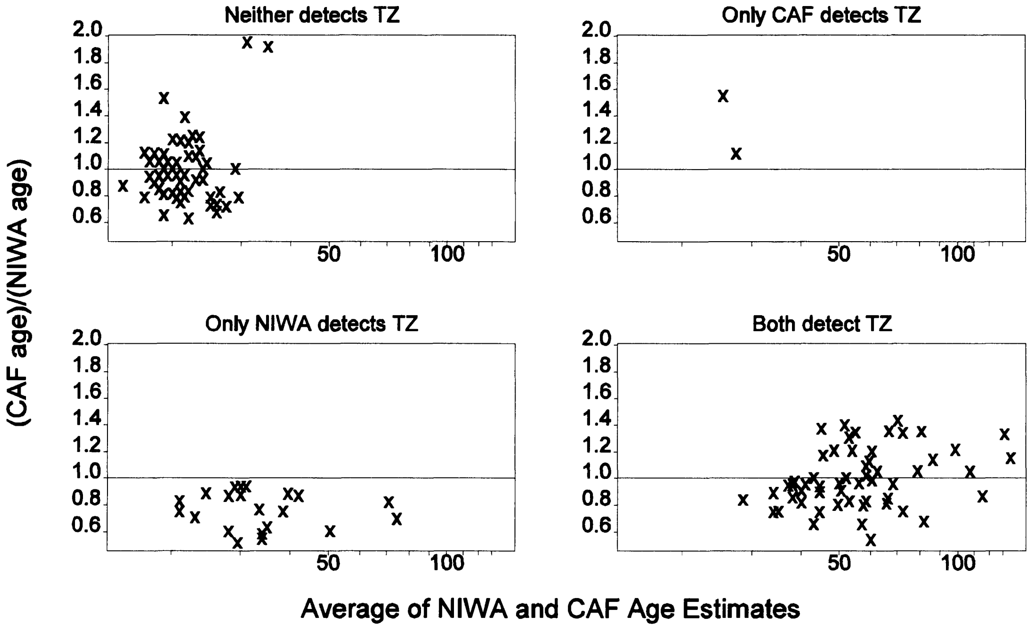

Figure 6: Effect of TZ detection on relative bias in the second NIWA-CAF inter-calibration data set. Each panel shows the ratio (CAF age)/(NIWA age), plotted against the mean age, for a different subset of the data defined by whether the TZ was detected by each reader.

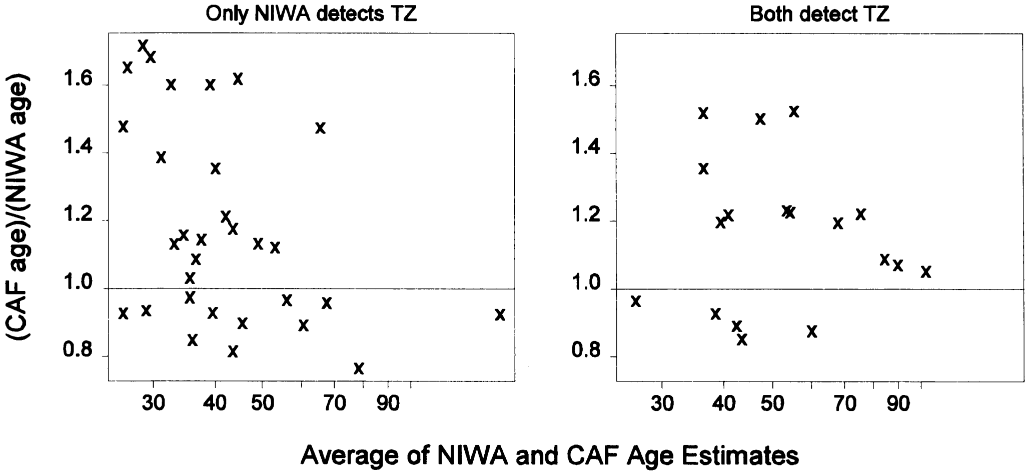

Figure 7: As for Figure 6, but using the first NIWA-CAF inter-calibration data set. Only two panels are plotted here because NIWA detected the TZ in all otoliths.

2.5 Calibration data

Calibration data sets, in which the same otoliths were read twice, are useful for estimating precision and relative bias. There are at least 10 such data sets – some involving replicate readings by the same reader, and others involving different readers from the same or different institutions (Table 5).

There should be at least one more calibration set. In the report for set CAF2 it is said that '100 otoliths were chosen from previous New Zealand collections [presumably from CAF1] and ... re-read' in order to re-familiarise the reader with orange roughy otoliths. However, these readings were not included in the data set provided by CAF. According to the associated reports, such re-familiarisation was not deemed necessary for sets CAF3 and CAF4 because the preceding sets of NZ otoliths had been read recently. Some of the data in Table 5 have already been analysed in other reports (Tracey et al. 2004, Hicks 2005a).

Table 5: Details of ten calibration data sets (in which the same otoliths were read twice).

Type of comparison	Source of otoliths	Number of otoliths	Comparison	Label
One reader, same time	CAF1	302	Corey2/Corey1	Corey.CAF1
	CAF2	125	Corey2/Corey1	Corey.CAF2
	CAF3	365	Corey2/Corey1	Corey.CAF3
	CAF4	302	Corey2/Corey1	Corey.CAF4
One reader, different time	Batch 141	60	Corey2/Corey1*	Corey2/Corey1.141
Two readers	Batch 141	61	Simon/Corey1*	Simon/Corey1.141
	Batch 163	83	Simon/Corey	Simon/Corey.163
	NIWA2	44	Pete/Di	Pete/Di
Two institutions	NIWA3	50	NIWA/CAF	NIWA/CAF.1
	Batch 163	137	NIWA/CAF	NIWA/CAF.2

*In these comparisons, the Corey1 reading was done as part of set CAF2, and the Corey2 and Simon readings were done as part of a re-familiarisation procedure for set CAF5

As might be expected, the scatter in the data is least when only one reader was involved, somewhat greater with two readers from the same institution, and even greater for readers from different institutions (Figure 8).

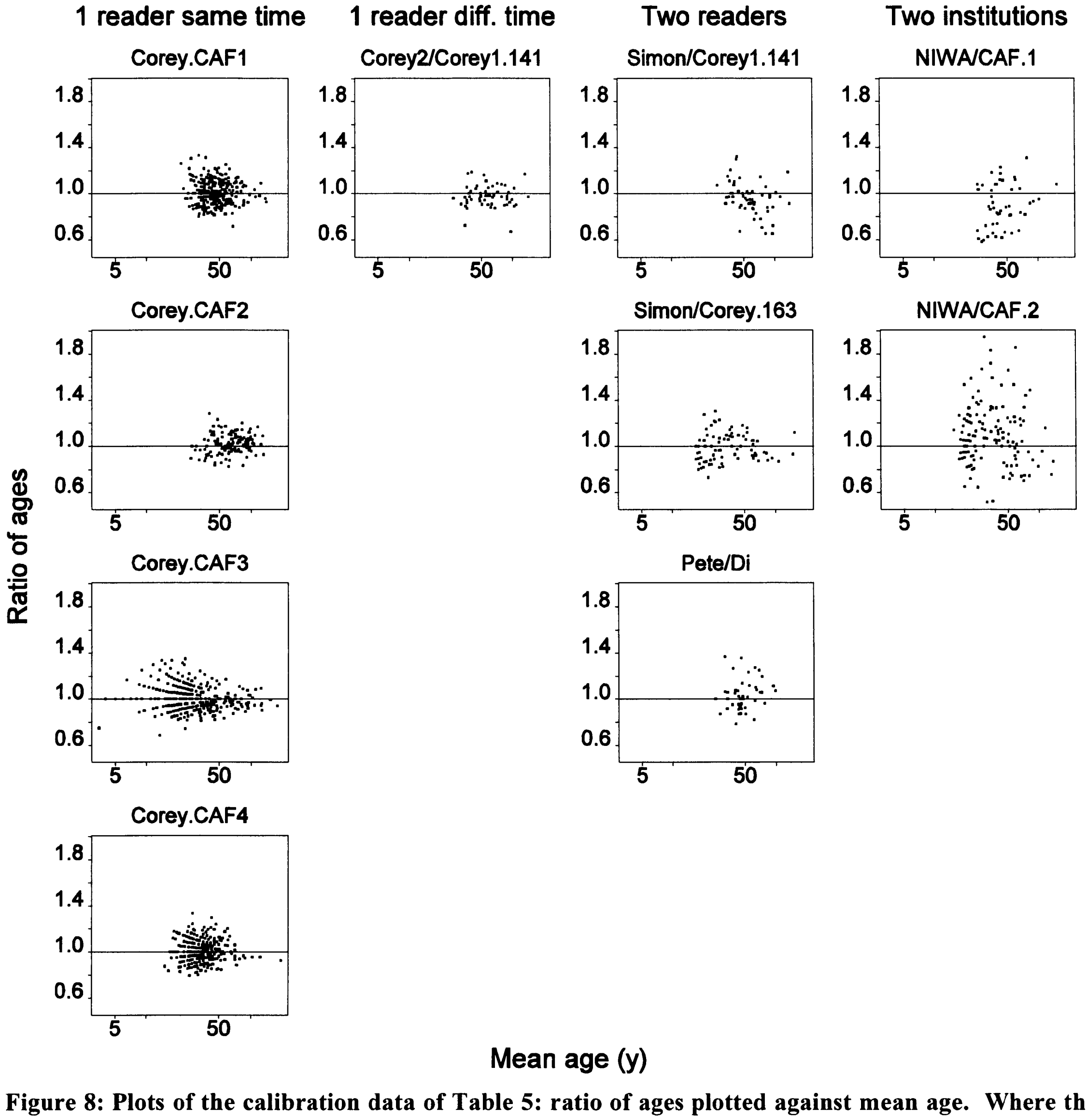

Figure 8: Plots of the calibration data of Table 5: ratio of ages plotted against mean age. Where the label above a panel does not indicate which reading was in the numerator of the ratio (i.e., in the left-most column of panels) the second reading was always in the numerator.

For each calibration data set I calculated the mean ratio of ages, and an approximate 95% confidence interval for this using a bootstrap procedure (Appendix 2). As a sensitivity analysis these calculations were repeated using a plus group at 80 y, as used in stock assessments (i.e., ages greater than 80 y were recoded as equal to 80 y and otoliths for which both readings were in the plus group were dropped).

Several patterns are apparent in these confidence intervals (shown in Figure 9), and these patterns are broadly similar whichever version of the interval is considered. First, second readings within sets CAF1-CAF4 seemed to be consistently 1% to 2% higher than first readings. Why this should have happened is unclear. Second, when otoliths from batch 141 were read again, as part of a re-familiarisation procedure for set CAF5, the later readings were 2% to 4% lower than the first readings (from set CAF2). This difference is in the same direction, but much smaller than, the 25% difference estimated between CAF2 and CAF5 using ALKs (see Table 3). Third, within both institutions, a between-reader relative bias as high as 4% was found (see comparisons Simon/Corey1.141 and Pete/Di). Finally, the two between-institution comparisons are consistent with the hypothesis of significant drift between CAF sets but not between those from NIWA. These comparisons suggest that readings in set CAF1 are about 10% higher than NIWA readings and those in CAF5 are about 10% lower than NIWA, which sums to about 20% difference between CAF1 and CAF5. This is similar to the CAF1-CAF5 difference of 23% estimated in Table 3.

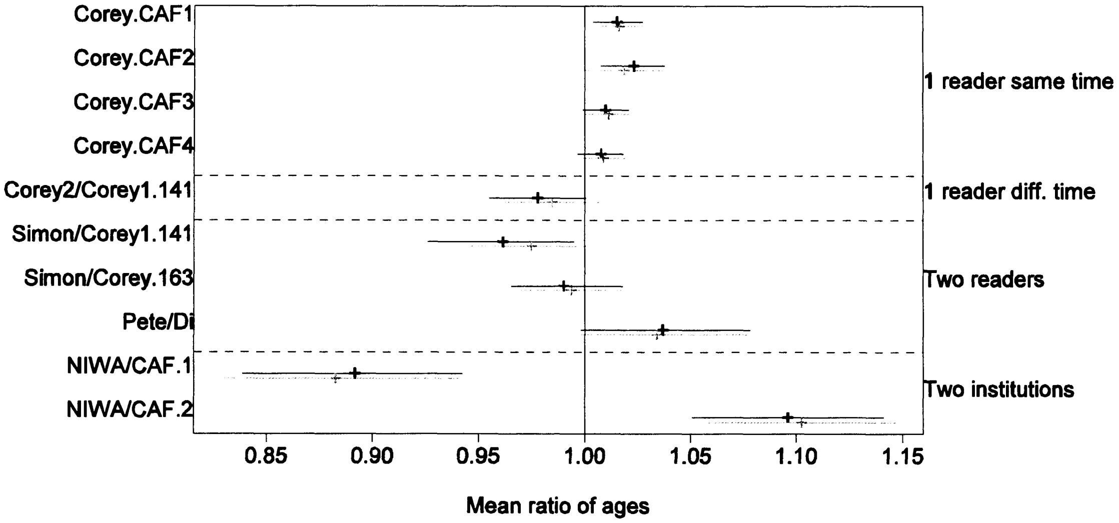

Figure 9: Mean ratio of ages ('+') with 95% confidence intervals (lines), for each of the calibration data sets in Table 5. Two versions are plotted: using all data (black), and with a plus group at 80 y (grey).

2.6 Testing the drift hypothesis

The obvious way to test the drift hypothesis would be to reread some otoliths from all 5 CAF sets. If this hypothesis is correct there will be consistent between-set differences in the comparisons between original and new readings and this would be detected, given adequate sample sizes, by a one-way analysis of variance performed on the ratio (new reading)/(original reading). If the hypothesis is false then the ANOVA would be non-significant.

If the drift hypothesis is correct then the rereading data could be used to calibrate all five CAF sets using mean ratios, like those in Figure 1. To be useful, these ratios would need to be reasonably precisely estimated. Table 6 shows the likely effect of sample size on the precision of the estimated calibration ratios.

Table 6: The effect of sample size on the precision of estimated calibration ratios. The tabulated values are the expected widths of the 95% confidence interval for a calibration ratio, calculated for different sample sizes and depending on whether the person rereading the otoliths was the same or different from the original reader (see Appendix 3 for details of the calculation of these widths).

	Sample size (number of otoliths per set)				
	50	100	150	200	250
Same reader	0.054	0.038	0.031	0.027	0.024
Different reader	0.073	0.052	0.042	0.036	0.033

For the purposes of this note I have made the simple assumption that any drift between sets is purely proportional. Any analysis of the rereading data would, of course, investigate this assumption and modify it if necessary.

2.7 Conclusions about the age data

The patterns described above are disturbing. Until at least some of them are resolved it is hard to see how any of the age data can be used in stock assessments with any confidence.

The rereading experiment described in the previous section is an obvious way forward. If it confirms the drift hypothesis and provides useful calibration data this will give us an obvious way to use the data, although this will not be without problems. Stock assessment results would be affected, to some degree, depending on which set of readings was chosen as the standard to which all other sets were calibrated. If the drift hypothesis is not confirmed, then more study on temporal and spatial heterogeneity of ALKs will be necessary.

The rereading experiment cannot solve all our problems. It does not address the huge imprecision revealed in between-institution comparisons (differences in age estimates can exceed 40% – see right-hand panels in Figure 8) or the worrying relative bias indicated by the lower left panel in Figure 6.

I have strong doubts as to whether our procedures for ageing orange roughy are yet reliable enough to allow the use of age data in stock assessments (even if the rereading experiment is carried out successfully). There are two ways to evaluate age data. One is through the sorts of analyses described in Section 2.5 above. For this approach we need to set some sort of acceptance criteria for relative bias and precision within and between readers, institutions, and sets of readings. What are reasonable criteria I don't know, but I am doubtful as to whether the large between-institution differences we have seen are acceptable. A second, and more pragmatic, approach to the evaluation of age data for use in stock assessment is to look for consistency in data collected from the same area in consecutive years. We cannot expect to use the sort of consistency check that is useful for relatively short-lived species (e.g., hoki): the ability to track strong and weak year classes over time. For longer-lived species like orange roughy we don't expect to see evidence of strong and weak year classes. However, we should expect to see only small changes from year to year in the calculated age frequencies and ALKs from the same area. Until we have age data showing this sort of year-to-year consistency we cannot be confident that the large differences we have seen between samples collected 10 or more years apart (e.g., in MEC) are real, and not simply artefacts arising from ageing problems. I think that if we are serious about using age data in orange roughy assessments we should be setting out to collect and read otolith samples from the same fishery for several consecutive years in order to see whether our ageing methodology is good enough.

3. OBSERVER LENGTH DATA

There are two important characteristics of observer-collected length frequencies (LFs) for orange roughy: they typically come from only a few trips per fishery in each year, and they show substantial between-trip variation (see below). This causes two problems when these data are used in stock assessments — correlations within the LFs, and the potential for aliasing — and these are the subject of this section.

3.1 The problem of correlation in observer LFs

3.1.1 Demonstrating the correlation problem

The roots of this problem are illustrated in Figure 10, which uses observer data from the Spawning Box.

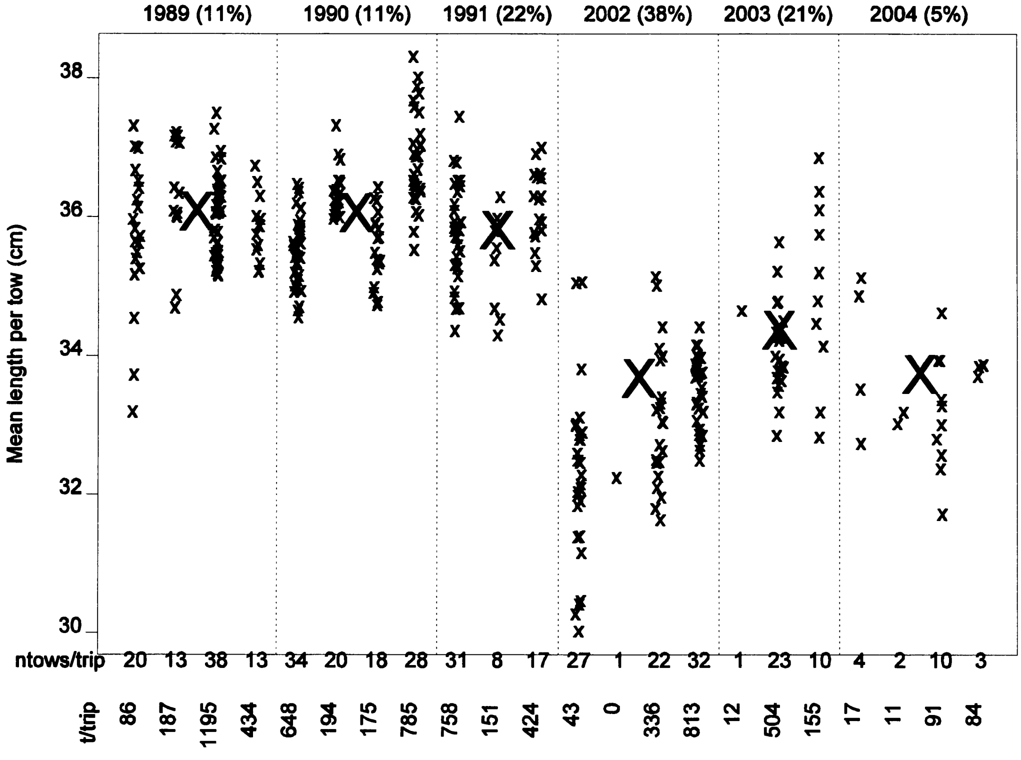

Figure 10: Evidence of between-trip differences in fish size: mean lengths for observed commercial tows in the Spawning Box for selected years (those years for which LFs were used in the 2005 assessment). Each plotted point shows the mean length of fish in one tow (adjusted to a 50:50 sex ratio) and the points in each horizontal cluster are all from the same trip (the number of tows sampled and the tonnage caught in each trip are shown below the x-axis); vertical broken lines separate trips from different fishing years, and the percentage of the total catch that was in observed tows is shown, for each year, above the plot; large Xs indicate the mean length for each year for the LFs used in the 2005 assessment. Tows without at least 20 measured fish of each sex were excluded from the plot.

The two important features to note in Figure 10 are that comparatively few trips were sampled in each year (between 2 and 4, if we ignore trips for which only one tow was sampled), and that there is evidence of considerable between-trip heterogeneity (i.e., fish caught in the same trip tend to have more similar lengths than those caught in different trips). I will show that these features cause substantial correlations in the sampling-error distributions for annual LFs generated from these data. That is to say, if p_{iy} is the estimated proportion of the catch in year y that is of length L_i, I will show that the correlation between p_{iy} and p_{jy} can be substantial. Before demonstrating this for the Spawning Box data I will consider a simple artificial example.

Consider an artificial fishery in which there are two types of tow: type A, which catches larger fish, and type B, which catches smaller fish (Figure 11A). Now suppose we have an estimated LF from this fishery that was formed by sampling 100 fish from each of two tows selected at random (the two tows could both be of type A, or both of type B, or one of each type). To study the sampling uncertainty I simulated 300 such LFs. From these replicate LFs it was straightforward to calculate c.v.s and show that the effective sample size, N_{eff}, was 105 (Figure 11B). Our concern here is with correlations between LF proportions. In stock-assessment models these are usually assumed to be either zero (when lognormal or Coleraine errors are assumed) or small and negative (with multinomial errors – Figure 11C). From the simulated LFs I estimated the actual correlations and found that these were often large and could be either positive or negative (Figure 11D). As an example of how such large correlations are generated, consider the proportions at lengths 32 cm and 40 cm in Figure 11A. If both the tows we sampled were of type B then p_{32} would be small and p_{40} would be large, whereas if both tows were of type A they would be the other way round. Thus we would expect a negative correlation between these proportions (its actual value is –0.73).

An important effect of these correlations is that uncertainty in mean length is much larger than is implied by the effective sample size. From the effective sample size of 105 the assessment model would infer a standard error of 0.27 cm for the mean length (this is the s.d. of the mean of the simulated LFs divided by $105^{0.5}$); however, our bootstrap samples show that the true s.e. is more than three times larger than this: 0.89 cm. Thus, our effective sample size of 105 correctly represents the uncertainty in individual LF proportions, but understates the uncertainty in the location (mean) of the LF. It might be thought that we could reduce the size of the LF correlations by increasing the bin size for the LFs, but this is not so (Figure 11E).

Now, returning to the data in Figure 10. I bootstrapped these data to generate 300 LFs for each year. The bootstrapping involved resampling at the levels of trips, tows, and fish. Thus, for example, in generating an LF for 1990 the first step was to select four trips at random, with replacement, from the four trips that were sampled in that year (a full description of the bootstrapping procedure is given in Appendix 4). The correlations between LF proportions in each year were often large, and broadly similar in structure to those from the artificial example, though there were substantial differences in estimated structure between years (Figure 12). As with the artificial example, the actual s.e.s for mean length are much greater (by a factor of between 1.7 and 4.1) than those inferred from the multinomial distribution (Table 7).

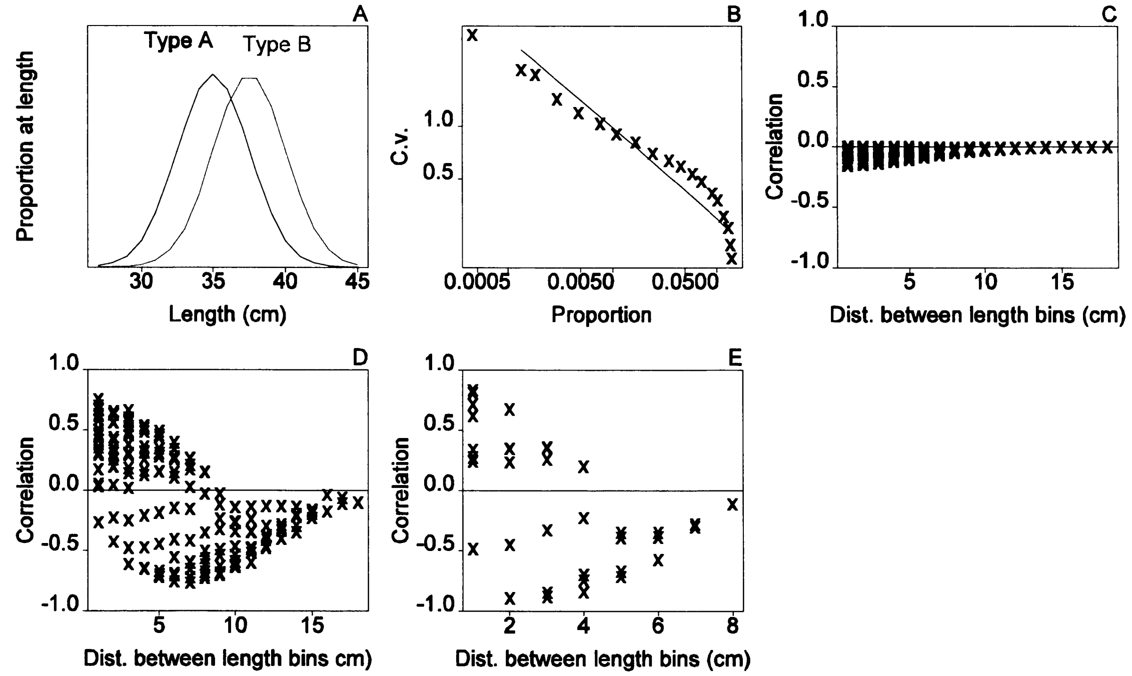

Figure 11: Artificial example illustrating the problem of between-trip heterogeneity in fish size: A, assumed length distributions in the two types of tow; B, estimated c.v.s for simulated LFs (the fitted line is that for a multinomial distribution with sample size N = 105); C, correlation structure for the LF, assuming a multinomial distribution (the correlation between two multinomial proportions, p_i and p_j, is $-(p_ip_j/[(1-p_i)(1-p_j)])^{0.5}$) ; D, actual correlation structure for the LF; E, correlation structure for the LF when bin sizes were increase by a factor of about 2 (the number of length bins was reduced from 19 to 9).

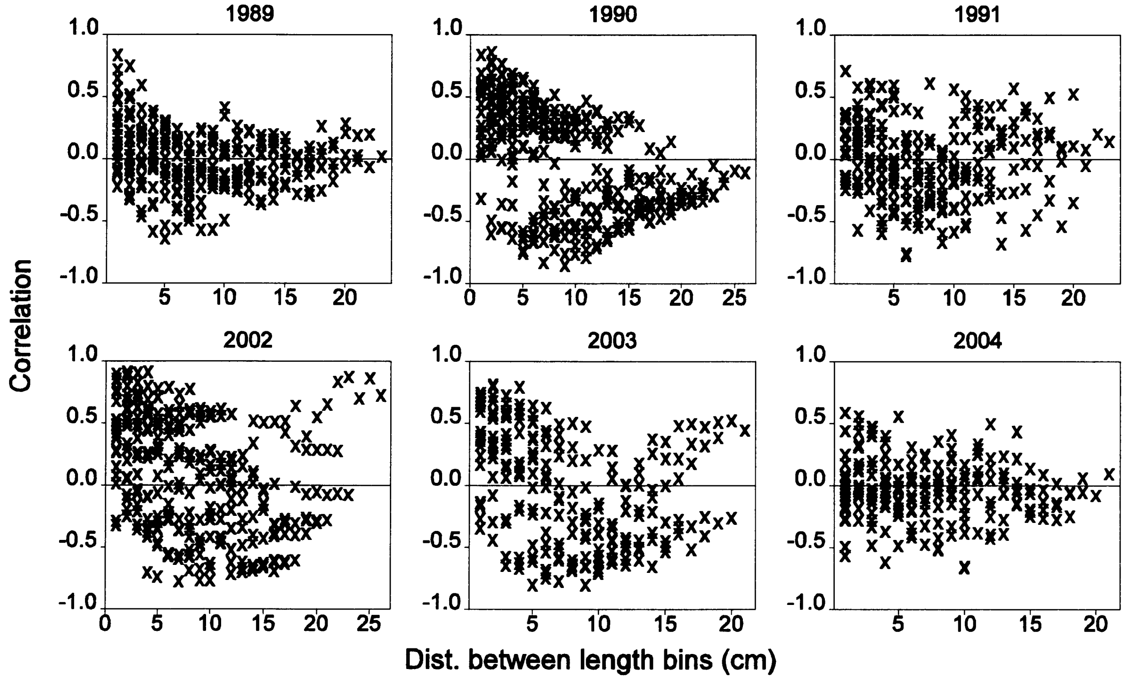

Figure 12: Correlation structures for Spawning Box LFs, estimated for each year from bootstrap replicate LFs.

Table 7: Estimates of the standard error of mean length for annual LFs based on the Spawning Box data in Figure 10. Two estimates are given for each year: one is based on the multinomial distribution and an effective sample size (estimated as in Figure 11B), and the other is the actual value calculated directly from the bootstrapped LFs.

	1989	1990	1991	2002	2003	2004
Based on multinomial	0.064	0.079	0.085	0.107	0.116	0.148
Actual	0.147	0.325	0.172	0.337	0.384	0.257

3.1.2 Three alternative responses to the correlation problem

I consider three possible responses to the correlation problem. The first is the simplest: *check error size*. This requires no change to the assumed error distribution for these LFs, which is currently what is called in CASAL the Coleraine structure (a robustified multivariate normal with zero correlations). What might change are the effective sample sizes, which determine the assumed sizes of errors. We want to be sure that these sample sizes are not so large as to mislead the stock-assessment model by implying that the mean lengths were more precisely known than is indicated by the actual s.e.s in Table 7. This appears not to have been a big problem in the 2005 Spawning-Box assessment because row D of Table 8 is much less than row C in all years but one. It should be remembered that we want row D to be less than row C because the latter includes observation error only, whereas the former includes both observation and process error. An obvious difficulty is that we don't have any idea how large the process error might be.

The other problem with the *check error size* response is that the assumed error structure is grossly wrong because it ignores substantial correlations. It's hard to say what effect this might have on the assessment.

Table 8: Four different versions of sample sizes for LFs for the spawning Box: A, actual (number of fish measured by observers); B, effective sample size for proportions (estimated as in Figure 11B); C, effective sample size to correctly represent uncertainty in mean length (calculated by multiplying the sample sizes at B by the square of the ratio of the two s.e.s in Table 7); and D, as used in the 2005 assessments (these were set at 1% of the actual sample sizes following Smith et al. 2002).

Version	1989	1990	1991	2002	2003	2004
A, Actual[1]	8815	9099	5075	7500	3263	1947
B, Effective size for proportions	1633	1171	1111	728	697	392
C, Effective size for mean lengths	282	69	255	74	60	127
D, Used in 2005 assessment	88	92	52	47	33	21

[1] These actual sample sizes differ, mostly slightly, from those used in setting effective sample sizes in the assessment; the reason for the big difference in 2002 is that the values used in the assessment were, inadvertently, based on calendar years, rather than fishing years.

A second possible response to the correlation problem is what I call *full error*. This involves modifying the error structure for the LFs by including the correlations. This response has the disadvantage of being very cumbersome because it would require much larger data files. We would have to provide, for each year, a full covariance matrix ($0.5(m^2 + m)$ independent numbers, where m = the number of bins in the LF) instead of a single sample size. A better reason for rejecting the full-error response is the fact that the correlations are very poorly estimated, which I will show after discussing the third response, which I call *change data*.

Each LF for the Spawning Box is represented by 27 proportions (one for each length bin from 19 cm to 45 cm, inclusive). When we look at the size of the correlations between these numbers (as shown in Figure 12) we might ask whether each LF contains enough information to require 27 numbers in order to describe it. I will show that most of the information is captured in just three descriptive statistics: the mean, s.d., and skewness (using the conventional definition of skewness based on the third moment – see p. 42 in Johnson et al. (1992)). Thus it might be sensible to change the LF data that we present to the stock assessment model from 27 proportions to these 3 statistics.

To measure the information content of these three statistics I will borrow the conventional information measure used in linear regression: percent variance explained. Suppose we are using some measure of temperature to predict year-class strengths (YCSs) in a fish stock. We say that temperature explains 60% of the variance if $100(V_1 - V_2)/V_1 = 60$, where V_1 is the variance of the YCSs and V_2 is the variance of the residuals when the YCSs are regressed against temperature. We could also say that this measure of temperature contains 60% of the information in the YCSs. In the context of the LFs, the set of bootstrap replicate LFs are analogous to the YCSs, so the total variance in this set is V_1. We can measure this total variance as the sum of the variances of the principal components of the set (we need to use principal components so that the variances we are adding are of quantities that are independent). If we are trying to measure the information content of some LF statistic (e.g., the mean length) we first adjust each of the replicate LFs so that they all have the same value of this statistic. This removes all the information carried by this statistic, so the adjusted set of LFs is analogous to the set of regression residuals, and its total variance is V_2. Having calculated these two variances we can now measure the information contained in our LF statistic as a percent variance explained: $100(V_1 - V_2)/V_1$.

Applying this approach I successively adjusted the bootstrap LFs from the Spawning Box data for the mean, s.d., and skewness (Figure 13 shows the results for one year) and calculated the percent variance explained at each stage (see Appendix 5 for a description of exactly how the adjustment was made for each statistic). I found that these three statistics typically explained about 80% of the variance in the Spawning Box data (Table 9).

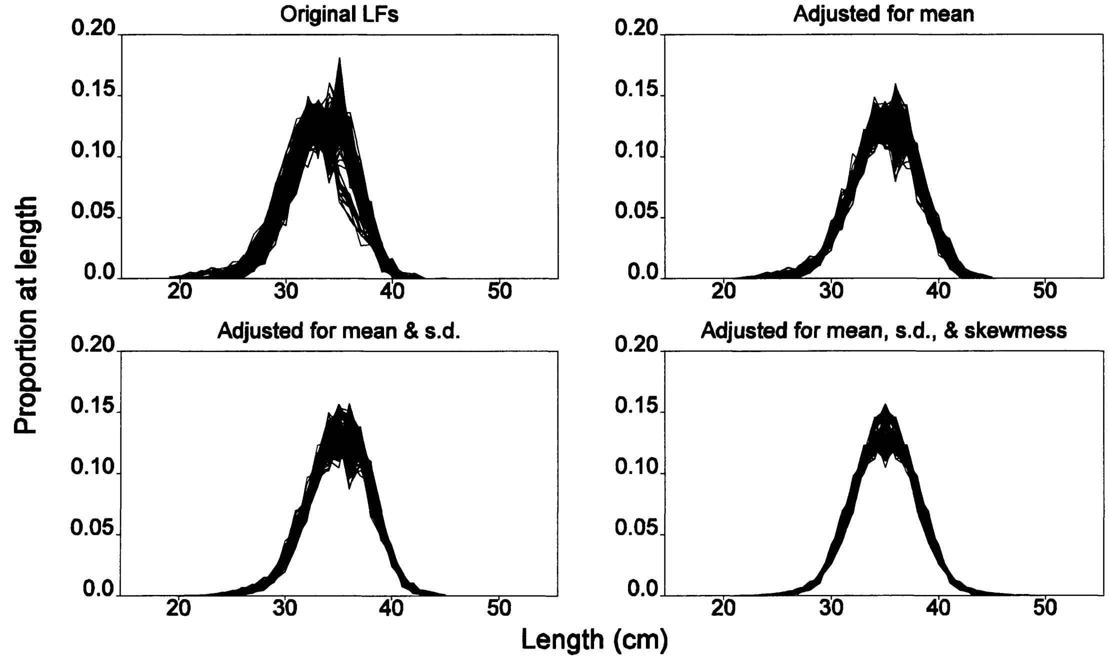

Figure 13: Illustration of the reduction in variance as a set of bootstrap replicate LFs for year 2002 in the Spawning Box is progressively adjusted to remove variation in the mean, s.d., and skewness. Each panel shows a set of 300 LFs; the original bootstrap replicate LFs are shown in the top left panel; in each succeeding panel the LFs are adjusted for one more statistic, as described above the panel.

Table 9: Cumulative percent variance explained by three LF statistics (mean, s.d., and skewness) for the Spawning Box LF data of Figure 10.

Year	Mean	Mean & s.d.	Mean, s.d., & skewness
1989	34.1	50.0	78.3
1990	67.3	70.7	85.5
1991	46.8	61.3	79.0
2002	57.7	66.0	83.1
2003	61.0	70.5	85.8
2004	48.2	62.1	81.8

Changing the data we input to the stock assessment model from LF proportions to the three statistics does not mean that we can ignore correlations. The correlations estimated for the Spawning Box data are sometimes quite large (Table 10). However, we might expect these correlations to be better estimated than those for the proportions because they use all the data, and not just that from single length classes. This expectation appears to be correct. The correlations between LF proportions sometimes changed substantially when data from one trip were omitted, whereas the correlations between our three statistics changed much less (Figure 14). As it is a matter of chance which trips are observed it does not make sense to use an error structure for our LF data that is strongly dependent on which trips were observed.

Table 10: Estimated pairwise correlations amongst mean, s.d., and skewness for the Spawning Box LF data of Figure 10.

	Mean & s.d.	Mean & skewness	S.d. & skewness
1989	-0.50	-0.05	-0.43
1990	-0.31	0.75	-0.54
1991	0.12	-0.50	-0.31
2002	-0.83	0.35	-0.46
2003	0.39	-0.31	-0.32
2004	-0.38	-0.02	0.41

For the Spawning Box, the decreasing trend in mean length is accompanied by an increase in s.d., but no clear trend for skewness (Figure 15).

It is worth asking whether we might do better to use some other statistics in place of mean, s.d., and skewness. I explored only two types of alternative for the Spawning Box data – quantiles and principal components – and did not succeed in finding anything better. I first tried quantiles at equally spaced probabilities (i.e., for *n* quantiles I considered those at probabilities 1/(*n*+1) , 2/(*n*+1), ... , *n*/(*n*+1), so for *n* = 3 these are quartiles). This was no improvement: it took at least 7 quantiles to explain as much variance as our three statistics (Table 11A). I then fixed *n* at 3 and experimented with changing the spacing of the quantiles. This did not provide much improvement (Table 11B).

Principal components are just linear combinations of LF proportions. Depending on the year, it required between 4 and 6 components to explain at least 80% of the variance in the sets of replicate LFs. Thus, these components are not as useful as our three statistics.

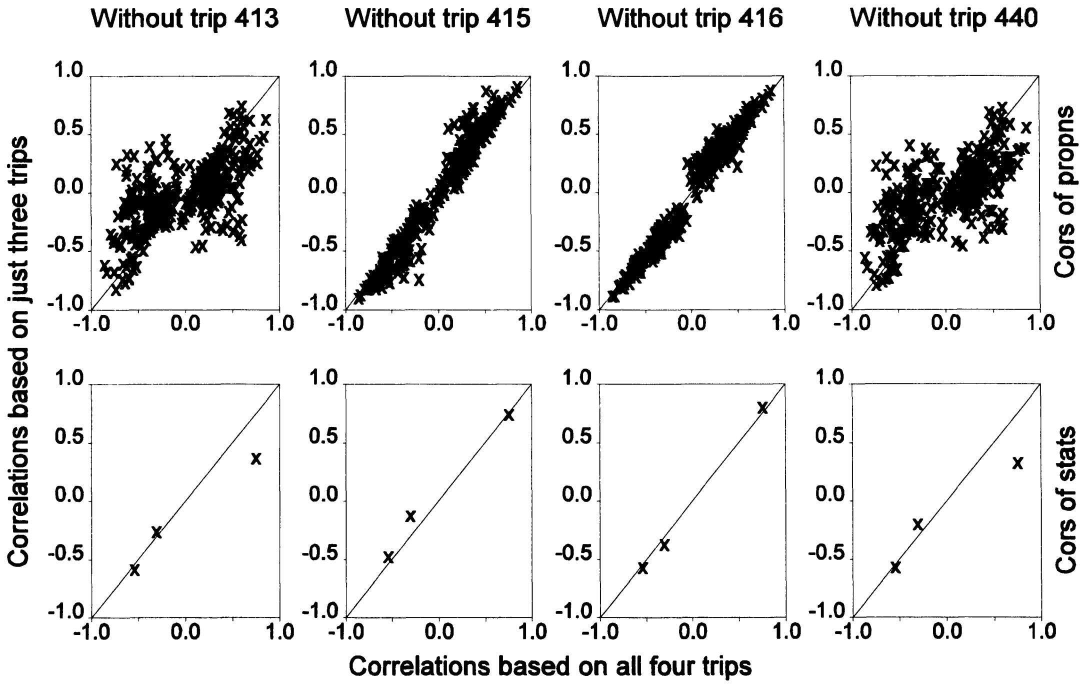

Figure 14: Illustration of the relative robustness of two types of LF correlations: those between LF proportions (upper panels) and those between three LF statistics – mean, s.d., and skewness (lower panels). The data used are those for the four trips observed in the Spawning Box in 1990 (as plotted in Figure 10) and each panel compares correlations calculated using data from all four trips (horizontal axis) against those based on just three trips (vertical axis). Note that the changes in correlations were greater when the trip that was omitted was one with a large total catch (the catches for each trip are shown in Figure 10, where the trips are ordered by trip number).

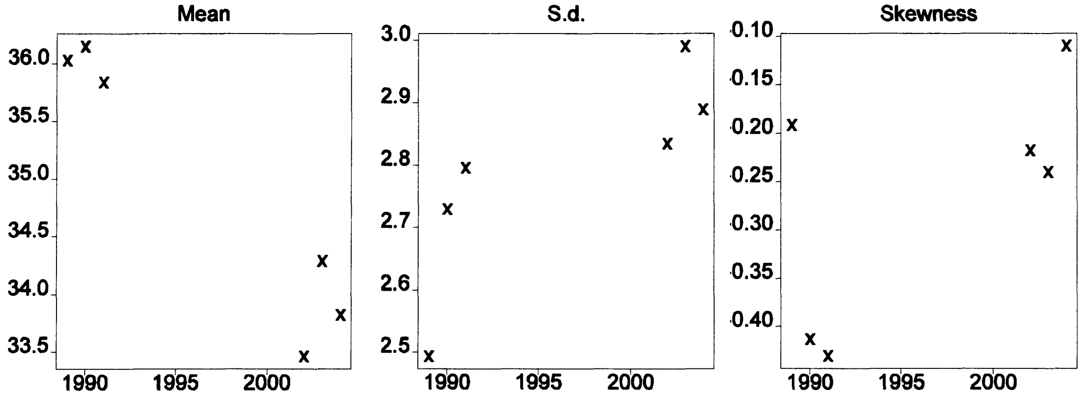

Figure 15: Estimates of mean, s.d., and skewness for observer LF data from the Spawning Box.

Table 11: Percent variance explained by quantiles for the Spawning Box LF data of Figure 10: A, 3, 5, or 7 quantiles with equally-spaced probabilities; B, 3 symmetric quantiles with probabilities at *a*, 0.5, and 1–*a*, for various values of *a*.

A		Number of quantiles		
		3	5	7
	1989	53.4	72.1	79.5
	1990	57.6	73.8	77.6
	1991	45.9	63.8	69.5
	2002	60.3	74.0	87.5
	2003	61.5	75.2	85.7
	2004	26.0	41.3	53.9

B		*a*=0.35	*a*=0.30	*a*=0.25	*a*=0.20	*a*=0.15	*a*=0.10	*a*=0.05
	1989	49.2	53.6	53.4	54.9	44.9	39.0	27.5
	1990	53.6	54.2	57.6	59.7	59.5	54.0	51.7
	1991	45.8	47.1	45.9	42.5	39.0	37.8	34.2
	2002	61.8	62.8	60.3	57.5	50.1	45.4	42.5
	2003	63.6	63.4	61.5	62.7	62.2	60.7	58.6
	2004	26.2	27.2	26.0	24.1	21.6	20.8	21.1

3.1.3 What should we do about age frequencies?

The correlation problem will obviously affect any age frequencies (AFs) which, like all the AFs used in the 2005 assessment, are calculated using an age-length key. The correlations in the LFs will induce correlations in AFs (Figure 16). (The correlations in Figure 16 are meant to be indicative only. They were estimated using the bootstrap replicate LFs for 2004 and a single age-length key based on all the age data from the Spawning Box. Strictly speaking, I should have restricted the age data to 2004 and bootstrapped them. This would certainly have changed the estimated correlations but there is no reason to believe that it would have made them any smaller.)

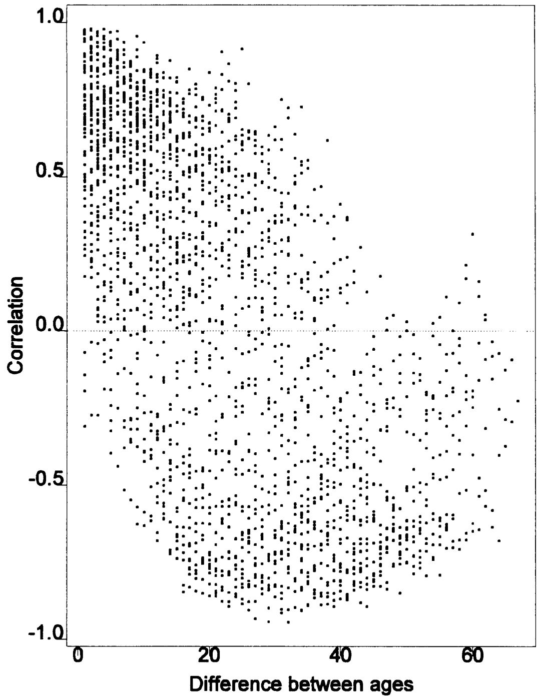

Figure 16: Estimate correlation structure for an AF for the Spawning Box in 2004.

Rather than trying to deal with these correlations I propose that the simplest solution is to avoid AFs in the stock assessment. This can be done by presenting the assessment model with two independent data sets: the LFs (either proportions or summary statistics) and the age-length data from which the age-length key is constructed. It is important that the likelihood for the age-length data should recognise that the otoliths were not sampled at random. In CASAL, this can be done using by labelling the observations as random_at_size.

3.2 The problem of aliasing

Aliasing occurs when we attribute some pattern in our data to the wrong cause. The pattern of interest in the current context is between-year changes in LFs, and there are three main causes of this pattern: sampling error (difference between our estimated LF for the catch and the true LF); a change in the overall selectivity of the fleet; and a change in the fish population. In reality, we can expect that all three causes will be active every year. However, in practice, we usually ignore changes in selectivity, so our assessment model will attribute all LF changes that are larger than expected from the LF error distributions to changes in the age structure of the population (since we don't have enough data to allow the model to estimate changes in growth rate over time). The primary danger for us is that our assessment model will wrongly infer a change of population age structure from the LF data.

To avoid aliasing of this sort we need to understand what is driving the observed variability in observer length data. Figure 17 illustrates an investigation of this for East Hills tows. The upper two panels show a clear seasonal effect (with mean lengths more than 1 cm greater during the June/July spawning season) and a suggestion of smaller fish on the hill Erebus. A closer look at the data suggested that the spawning season would better be defined as running from 28 May to 27 July (julian days 239 to 299) (panel C). Also, the suggestion that fish on Erebus are smaller seems to be an artefact caused by the fact that all but one tow on that hill occurred outside the spawning season (i.e., hill location was aliasing for season!). When restricted to the non-spawning season, the data show only a small (and statistically non-significant) difference between Erebus and the other hills (panel D). Other plots (not shown) showed no obvious relationship between mean length and time of day, starting depth, or target species.

What are the consequences of this relationship between fish size and season? First, the selectivity of this fishery will change from year to year if the percentage of the catch that is taken in the spawning season fluctuates. The data show that, after the first three years, this fluctuation was relatively minor, with the percentage varying between 0 and 20 (Figure 18A). We can get a rough idea of how much this variation is likely to affect the mean length of the catch if we assume that all fish caught during the spawning season have length 35.3 cm and those outside this season have length 34.0 cm (these values are taken from the horizontal broken lines in Figure 17C). The annual mean lengths inferred from this assumption can be read from the right axis of Figure 18A, which shows a relatively small fluctuation of about 0.3 cm after the first three years. How we should deal with this variation in selectivity depends on the circumstances. If none of the LF data we are using come from the first three years we can probably safely ignore the variation. If, however, we want to use data from both the first year and some later years then we could standardise our LFs to remove the variation. To do this we would calculate separate spawning and non-spawning LFs for each year and then take a weighted mean of these with, say, a 15:85 weighting (this represents what we think the LFs would have looked like if 15% of the catch had been caught during the spawning season in every year). If we don't do this standardisation, aliasing will occur: the model will interpret the changes in LFs that are caused by selectivity changes as being due to changes in population age structure.

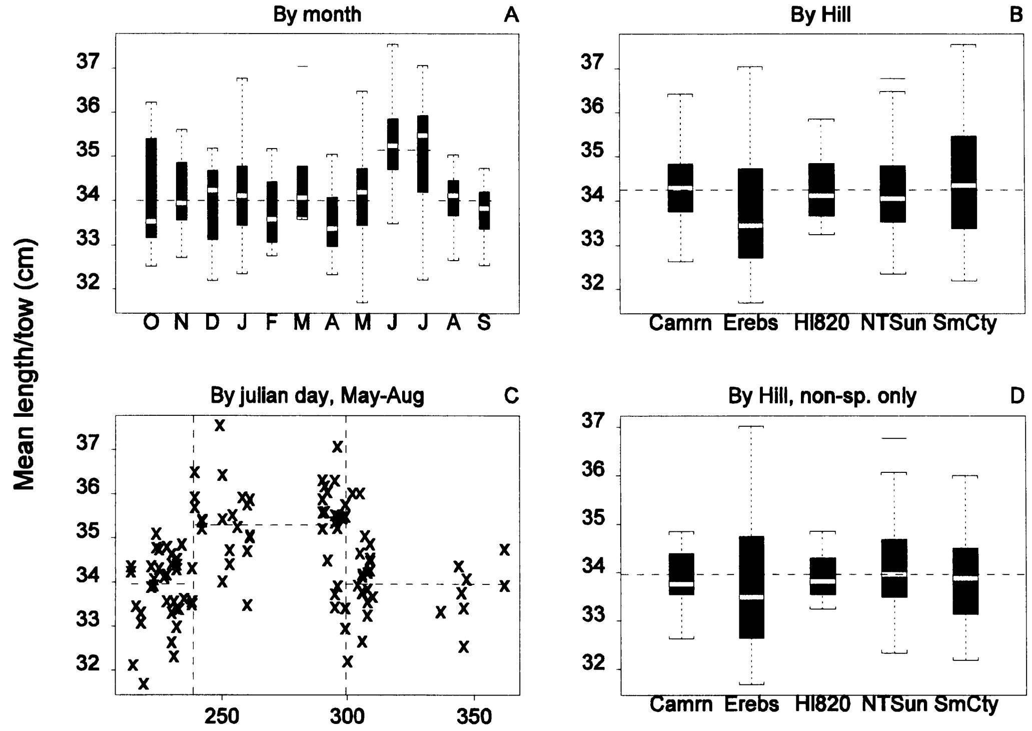

Figure 17: Mean length per tow (adjusted to a 50:50 sex ratio) for tows in the East Hills plotted against A, month; B, hill; C julian day (May-Aug only) with vertical lines indicating the presumed extent of the spawning season; and D, hill (restricted to non-spawning period, as defined in panel C). Horizontal broken lines show the average mean length per tow, either for all plotted data, or for the portion indicated. Tows plotted are restricted to those with duration not exceeding 30 minutes, starting positions within 3 n. mile. of one of the four named hills, and at least 20 measured fish of each sex.

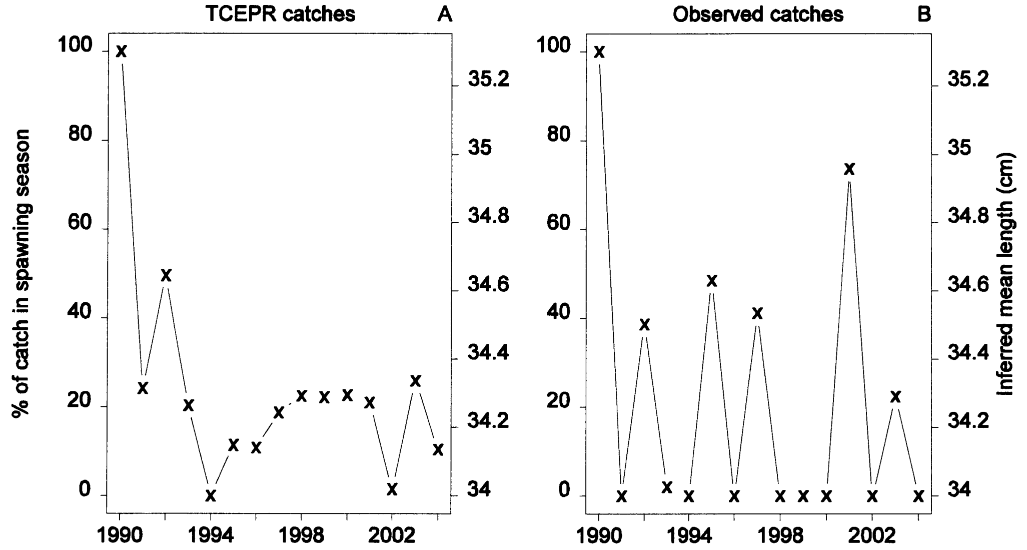

Figure 18: Percentage of East Hills catch in spawning season (left axis), and the associated inferred mean length (right axis), calculated from A, TCEPR data; and B, observed catches.

The second consequence of the fish size-season relationship for East Hills relates to the fact that observer sampling is patchy. Variation in the percentage taken during the spawning season was much greater for the observed catch, and showed little correlation with that for the total catch (Figure 18B). If we don't adjust for this sampling variation, it will be interpreted by the model as being caused by changes in population age structure. The obvious way to adjust is by stratifying by season. Sample size becomes a problem here. It is usual to set a sample size threshold for each year (e.g., at least 20 tows from two or more vessels and three or more trips was required in the 2005 assessment). This threshold should now apply to each season in each year, which means that more years will be rejected.

It is of interest to note that for the Andes complex there is no indication of a change of mean size during the spawning season (Figure 19), although there have not been many observed tows in June and July in this area (3 in June and 18 in July). This adds support to the hypothesis that this is not a major spawning area. In fact, if anything the seasonal pattern in this plot suggests that larger fish leave this area at the beginning of the calendar year, returning in October (though this pattern could be aliasing for fleet movement within the Andes area – this needs more study).

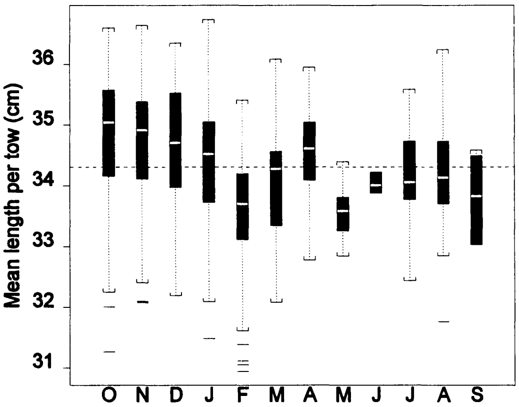

Figure 19: Mean length per tow (adjusted to a 50:50 sex ratio), for tows in the Andes complex, plotted against month. Tows plotted are restricted to those with duration not exceeding 30 minutes and at with least 20 measured fish of each sex. The broken line shows the average of all mean lengths.

We were lucky that we found only one cause of variation in the East Hills length data. It happens that there is another hill nearby, Dreamtime (about 40 n. mile to the northwest), on which fish are substantially larger (Figure 20A). Fishing in this area is too limited to indicate whether mean size varies seasonally (Figure 20B). By good fortune, Dreamtime is in a separate management area (Arrow Plateau). Had it been in the same management area as the East Hills we would have had to consider this size difference in our analysis. The effect on selectivity was not great, because this hill has never contributed a significant percentage of the catch (Figure 21A). However, it did contribute more than 50% of the observed catch in two years (Figure 21B), which would have a very strong effect on our calculated LFs if this fact was ignored.

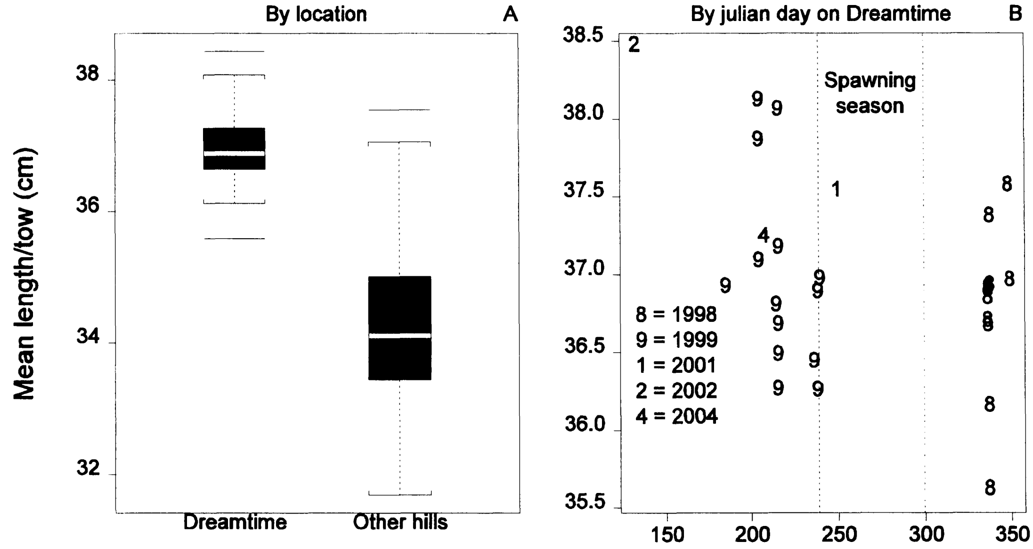

Figure 20: Mean length per tow (adjusted to a 50:50 sex ratio), for tows in the area including East Hills and Dreamtime, plotted against A, location; and B, julian day (dreamtime only, with vertical lines indicating the spawning season inferred in Figure 17C). Restrictions on tows plotted as for Figure 17.

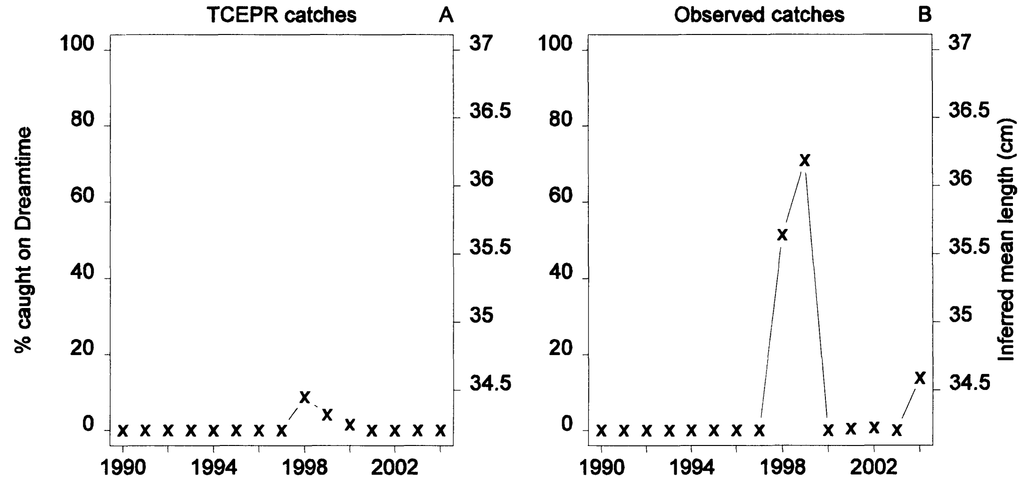

Figure 21: Percentage of E. Hills + Dreamtime catch caught on Dreamtime (left axis), and the associated inferred mean length (right axis), calculated from A, TCEPR data; and B, observed catches.

Considering the problems identified in Section 3.1, it is of interest to ask how much between-trip variation in fish size in the East Hills area is explained by season. Figure 22 shows that quite a lot is explained, particularly in 1992, 1995, and 2003, but not all of it (see 1993 and 1994). This plot also shows the paucity of observer data for this area. If we insist on at least 20 observed tows per year then we are left with only two years (1992 and 1995), and no year satisfies this criterion for both seasons. Note that almost half the observed catch in 1995 came from one spawning-season trip with four tows.

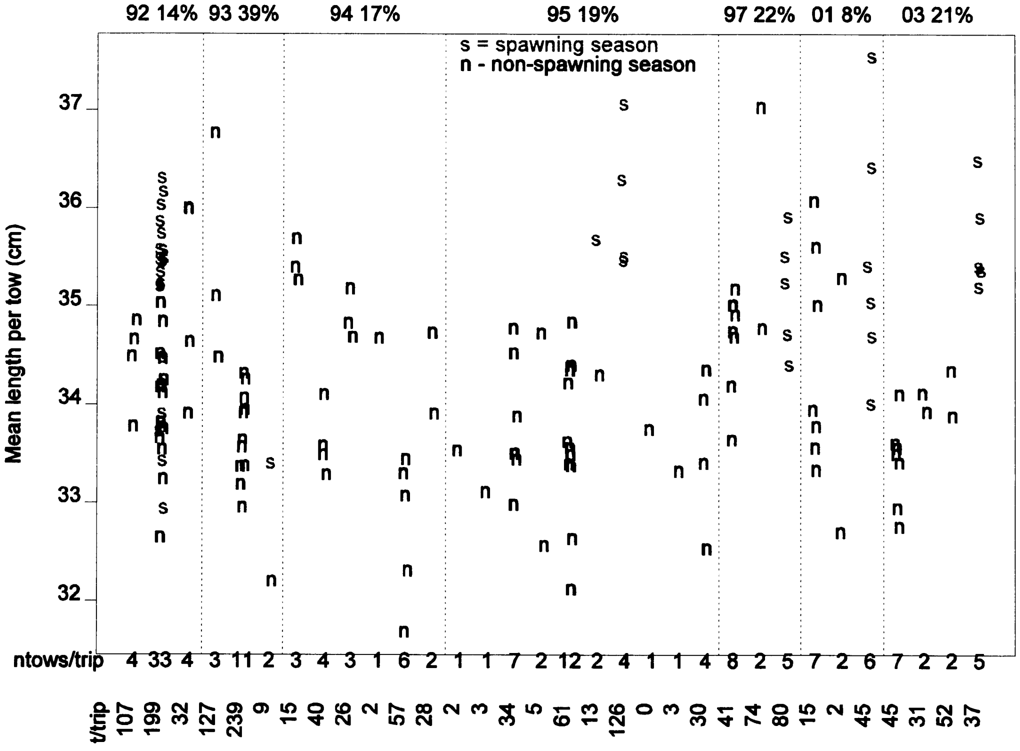

Figure 22: Mean lengths for observed commercial tows in the East Hills for years in which at least 15 tows were observed. Plotting conventions as for Figure 10, except that the season for each tow (spawning or non-spawning) is identified by the plotting symbol.

3.1 The potential for aliasing in the Spawning Box

The problem of potential aliasing is much more complicated in the Spawning Box because mean length varies with latitude, longitude, depth, and julian day, and has changed substantially between the pre- and post-closure periods (Figure 23). It is of interest to note that the geographical variation in mean length is broadly similar in the two periods, with mean length being greatest for latitude near 42.83 S, longitudes near 183.2 E, and depths less than 950 m. Of course, it is likely to be a combination of factors that is driving variability in mean length. For example, in the pre-closure period the increase in mean length towards the end of the season was associated with a shift into shallower water towards the south (Figure 24). This pattern was not evident in the post-closure data. On the basis of Figure 23 it appears likely that there really has been a substantial drop in the mean length of the population between the two periods. This drop is an important signal to give to the stock-assessment model.

There is clearly some scope for aliasing to bias our estimate of the magnitude of that drop. This could occur if the spatio-temporal pattern of fishing changed between the two periods in such a way as to change the overall selectivity of the fleet. Fishing in the latter period appears to have been somewhat more to the west and earlier in the season than it was pre-closure (left panels, Figure 24), and both of these changes seem likely to reduce the mean lengths of fish caught. Thus, the drop in mean length in the population may be exaggerated if this change in selectivity is not allowed for. However, the shift to earlier in the season was not as great as is suggested by the observer data (compare bottom two panels, Figure 25). So sampling error may make the change in selectivity appear greater than it actually was.

Our ability to correct for these biases is limited by the complexity of the factors causing within-year variability in mean length (as shown in Figure 23), the relatively sparse observer coverage, and the spatio-temporal changes. We should probably drop some tows (particularly those before day 220) to reduce the effect of temporal changes. Then we could stratify the data into a few (probably no more than two or three) strata, chosen (using tree-based models?) to maximise between-stratum variability in mean length. This stratification is not easy, because of the need to have a reasonable number of observed tows in each stratum in each year. If sensible strata can be found, these will help reduce bias due to sampling error and changes in selectivity. There is some consolation in the observation that the magnitude of any bias seems likely to be small compared to the change in mean length between the pre- and post-closure periods.

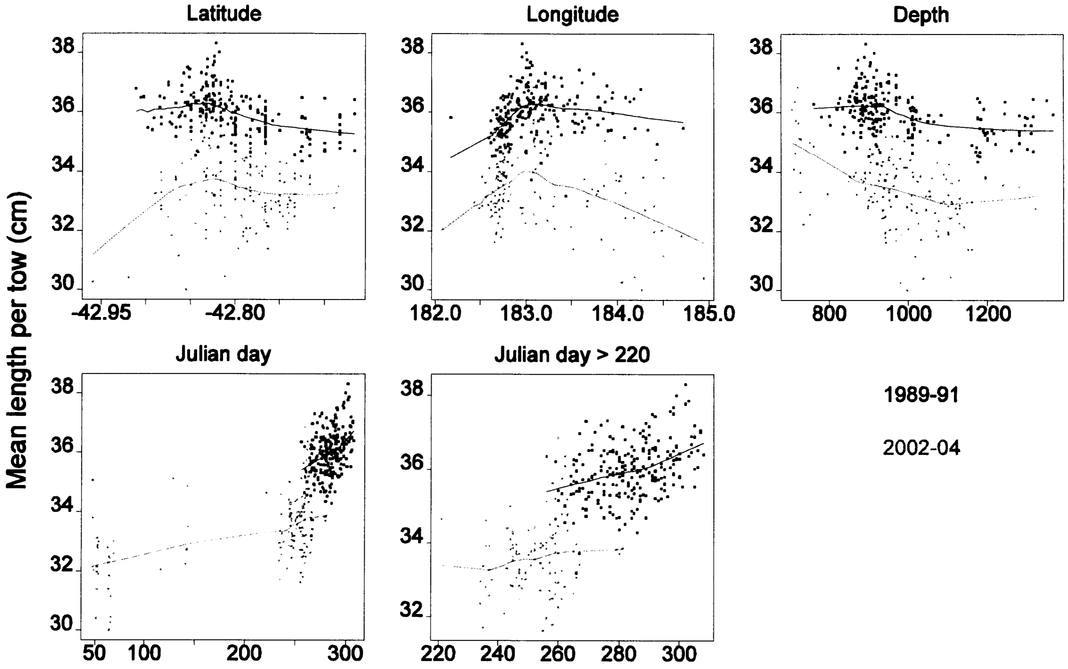

Figure 23: Mean length per tow plotted against latitude, longitude, depth, and julian day for the Spawning Box tows shown in Figure 10, with the two distinct periods of the data identified by colour: black for the early, pre-closure, period (1989–91), grey for the later, post-closure, period (2002–04). The lines are lowess curves fitted to the data for each period. To show more detail, the last panel is restricted to tows after julian day 220 in each year.

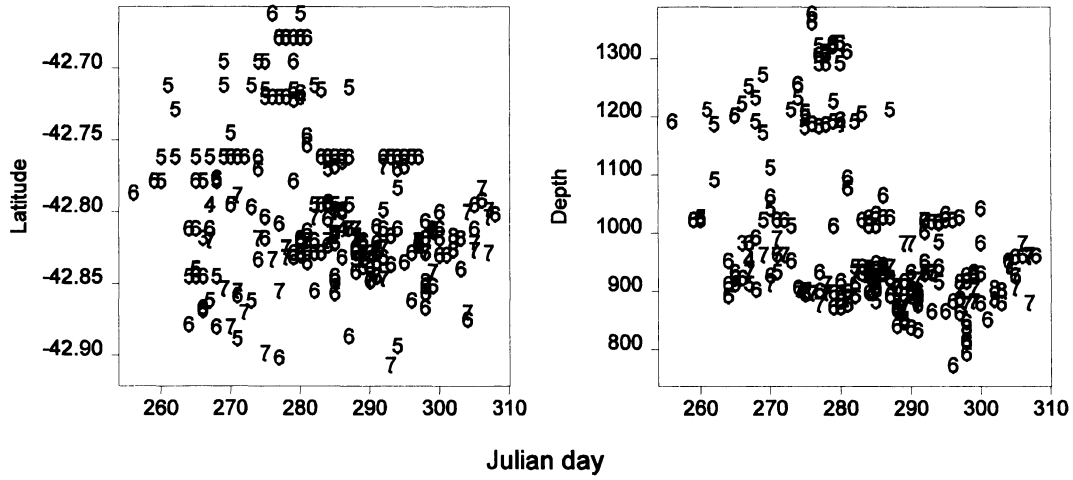

Figure 24: Illustration, for the pre-closure period, of the shift to shallower water to the south towards the end of the fishing season. Each plotted point represents one observed tow from the period 1989–91 and the plotting symbol indicates the estimated mean length for that tow, rounded to the nearest cm ('5' = 35 cm, '6' = 36 cm, etc).

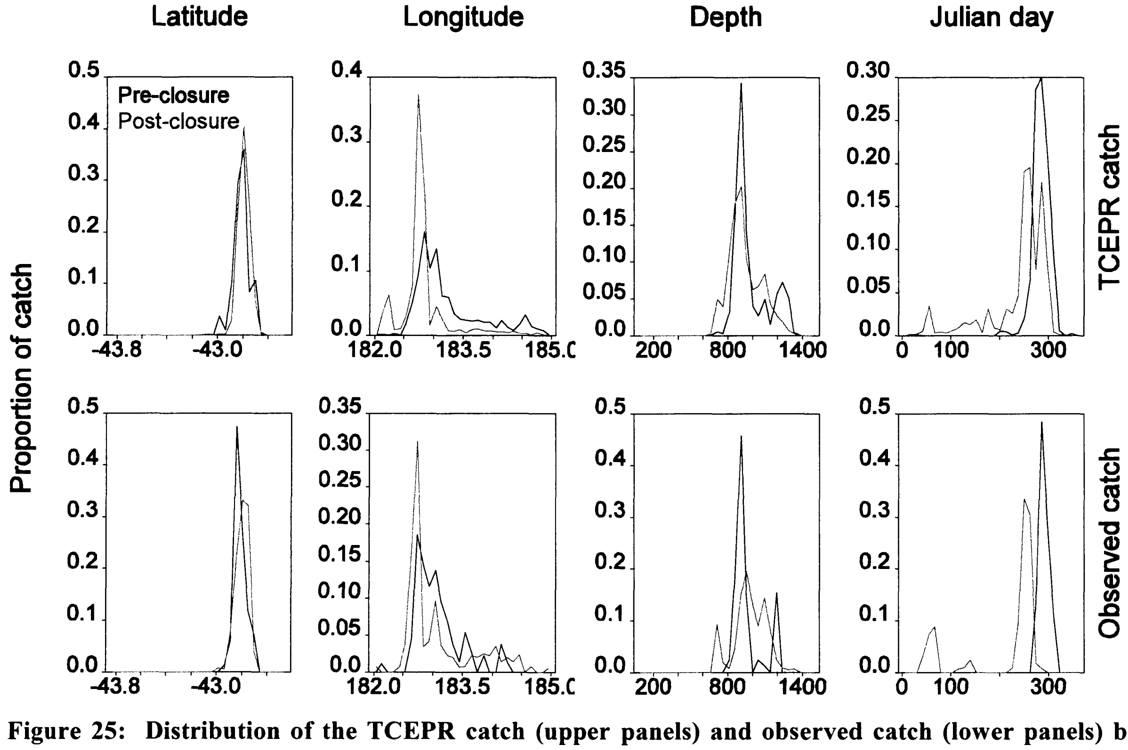

Figure 25: Distribution of the TCEPR catch (upper panels) and observed catch (lower panels) by latitude, longitude, depth, and julian day for three pre-closure years combined (1989–91, black lines) and three post-closure years combined (2002–04, grey lines).

3.3 Recommendations on the use of observer LF data

1. *Where observer LF data are used in the assessments they should be entered as proportions at age, and effective sample sizes should be consistent with bootstrap-derived estimates of uncertainty in mean length.*

Comments: This is what I have called, in Section 3.1.2, the *check error size* response to the problem of LF correlations. I suspect that it may be inferior to *change data*, but the latter response would require non-trivial modifications to the stock-assessment program(s), which would be difficult to achieve in time for the 2006 assessments. Also, before adopting the *change data* approach it might be worth carrying out analyses like those for Table 9 for other orange roughy stocks.

It is important that the bootstrap LFs are constructed in the same way as the actual LFs (i.e., they should include any stratification or standardisation that is deemed necessary to avoid aliasing – see Recommendation 3). Thus, bootstrapping should be done after an investigation of aliasing. From the bootstrap LFs we can easily calculate, for each year, an effective sample size, $N_{\mathrm{eff,\,obs}}$, which takes account of the effect of observation error on uncertainty in mean length (see Appendix 6). The effective sample size used in the assessment should be smaller than $N_{\mathrm{eff,\,obs}}$ to allow for some process error, but it's not clear how much smaller. I suggest arbitrarily setting $N_{\mathrm{eff}} = 0.5\ N_{\mathrm{eff,\,obs}}$.

2. *AF data based on age-length keys and observer LFs should not be used in the assessment. Instead, the LF data and age-length data should be entered separately, using an appropriate likelihood for the latter (i.e., the random_at_size likelihood).*

Comment: Of course, the age-length data should not be used at all until the problems detected in the age data have been resolved.

3. *Before observer LF data are used in the assessments they should be analysed to determine whether aliasing could be a significant problem. Where this problem is deemed significant, the data should be adjusted, if possible, to minimise aliasing. If such adjustment is not possible (because of small sample sizes and/or patchy sampling) the data should not be used. The adjustment could involve any or all of three measures: stratification (by factors like season or area that are associated with variation in mean size); standardisation (e.g., the 15:85 seasonal weighting suggested above for the East Hills data); and dropping of problematic years.*

Comments: It is difficult to be prescriptive about how these analyses should be done. The detection of potential aliasing is essentially an exploratory analysis, and how this should be done will depend on the data and what patterns exist in them. Some examples of possible analyses are given above. These analyses have focussed on the most salient feature of the LFs – the mean. In principle, we should look for changes in other features (e.g., the s.d.) However, I suspect that most changes of interest will involve a change in mean length. Note that stratification to avoid aliasing may require the dropping of years in which not all strata were adequately covered by observers.

There will be occasions when it is possible to determine what is causing a change in LFs. As an extreme example, if we have data from just two years and the area of fishing is completely different in each year, we cannot say whether any observed change in the LFs is caused by spatial heterogeneity in size distributions or a temporal change in the overall population. It will often be prudent to omit data from an area or season which is fished only in some years (e.g., the pre-spawning fishing in the Spawning Box in the post-closure years – see Figure 22).

4. MATURITY AND SELECTIVITY

In early orange roughy assessments in New Zealand it was always assumed, as an approximation, that the maturity and (commercial) selectivity ogives were the same (Francis & Robertson 1990). That is, only mature fish are caught by the commercial fishery, and all mature fish are equally selected. This was assumed because the early fisheries were primarily on spawning concentrations. Initially, no estimates of the maturity ogive were available, so this was set equal to the selectivity ogive, which was inferred from the ascending limb of commercial length frequencies (see section 3.5.1 of Clark & Francis 1990). Later, the maturity ogive was inferred from counts to the transition zone in otoliths (Doonan 1994, Francis & Horn 1997) and the selectivity ogive was set equal to this.

The first time the selectivity and maturity ogives were allowed to differ in an assessment was in 1999, where the maturity ogive was estimated (outside the assessment model) from transition zone data, and the selectivity was estimated using observer LFs (Hilborn et al. 1999). In this assessment, a problem arose: the age at which 50% of fish were commercially selected, a50, was estimated to be about 20 y greater than the age at 50% maturity, a_{mat}. The suggestion that a50 >> a_{mat} seemed (and still seems) highly implausible. In spawning fisheries it seems reasonable to assume a50 = a_{mat} and, in non-spawning fisheries, where smaller fish are sometimes caught, we would expect a50 < a_{mat}.

This problem has occurred in most subsequent orange roughy assessments when selectivity and maturity ogives were allowed to differ. It occurred whether the selectivity was estimated from LFs or AFs. For example, the 2004 Plenary Report lists examples where the estimated a50 exceeded a_{mat} by between 6.5 y and 19.2 y (see table 4, p. 325, in Annala et al. 2004). A common consequence of this was that the current vulnerable biomass (that selected by the selectivity ogive) was often estimated to be much smaller than the current mature biomass (in the examples just cited the vulnerable biomass ranged from 15% to 56% of the mature biomass). The implausibility of these results made the Deepwater Working Group decide that, as an interim measure, the maturity and selectivity ogive should be forced to be equal in orange roughy assessments. The maturity data were considered to be 'indirect', in that they were based on an assumption about the significance of the transition zone. In contrast, the selectivity data (commercial LFs or AFs) seemed more direct. Thus, when both maturity and selectivity data were available, it was decided that the maturity ogive should be set equal to the estimated selectivity (this has been designated by the label *mat2sel*), but when only maturity data were available the opposite should be done (*sel2mat*) (Annala et al. 2004).

A second problem arose. Estimates of a50 were sometimes not robust, in that they could change substantially when a seemingly unrelated model assumption was changed (see examples below). Why this happened was unclear.

In this section I present several analyses that relate to these two problems. This work was funded under MFish project ORH2005/04, for which the only objective was

> *To examine gonad and otolith samples and/or to conduct new analyses to investigate the discrepancy between maturity and vulnerability ogives observed in 2004 stock assessments for orange roughy.*

4.1 Observer and research length and maturity data

Female length and gonad-stage data were extracted from the observer (*obs_lfs*) and research (*trawl*) databases. Data on males were ignored because the focus of this study was on commercial selectivity, and the observers do not collect gonad-stage data for males. Stations at which fewer than 20 females were measured and staged were also ignored. The following ancillary data were extracted for each tow: trip code, station number and position, date, and catch weight of orange roughy. For some analyses, the data were grouped geographically into 17 areas which were defined partly on the basis of management areas and partly by the way the tow positions clustered (Figure 26). More than 200 000 fish from 4512 tows in 393 trips were selected from the observer data, and there were more than 150 000 fish from 1862 tows in 68 research trips (see Table 12 for a breakdown by area). Most of the observer data were from the Scientific Observer Programme (SOP), but about 10% of the fish from the period 1998–2002 (and 3% overall) were from Orange Roughy Management Company (ORMC) observers.

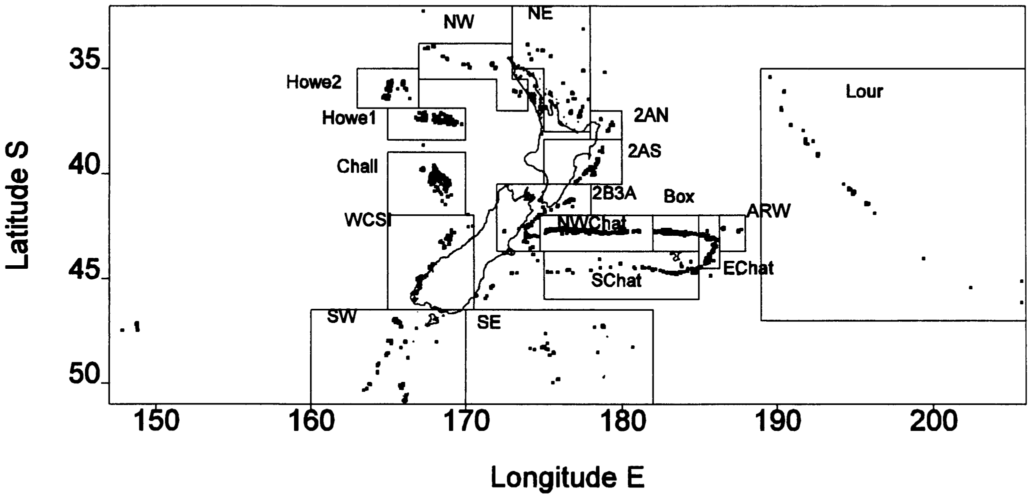

Figure 26: Positions of tows in which least 20 female orange roughy were measured and staged by observers, and the areas into which they were grouped for some analyses.

Different gonad-staging schemes have been used in the two data sets (Table 13). Most of the research data used the 8-stage deepwater scheme (labelled DW in *trawl*), but all six trips since the beginning of 2004 have used the new orange roughy (OR) scheme, which is the same as DW, except for the addition of stage 9 (all tows from these trips were in June or July). Six tows, all from the series of research trips by F.V. *Wanaka* in 1985 and 1986, were dropped from the research database because they used a different staging scheme (WK).

Table 12: Numbers of tows and fish, by area, that were selected from the observer (*obs_lfs*) and research (*trawl*) databases. Area definitions are shown in Figure 26.

	Observer data		Research data	
Area	Tows	Fish	Tows	Fish
NWChat	348	15 529	443	34 422
Box	586	29 054	605	61 091
EChat	807	38 512	190	15 502
SChat	363	18 387	68	4 840
Chall	401	18 644	75	5 756
Howe1	362	15 527	0	0
Howe2	192	8 522	0	0
SE	83	3 440	1	27
SW	236	10 925	29	1 530
WCSI	83	3 259	1	22
NW	152	5 581	7	502
NE	203	8 796	63	4 698
2AN	48	2 127	70	5 901
2AS	169	8 029	140	11 259
2B3A	126	5 553	168	12 858
ARW	125	6 004	0	0
Lour	148	5 707	0	0
Other	80	2 823	2	1 263
All areas	4512	206 419	1862	159 671

Table 13: Female gonad-staging schemes used by observers and for research tows, and the definitions of the categories 'spawning' and 'adult/juvenile' used in this study.

	Observer			Research		
Stage	Description	Spawning	Adult/juvenile	Description	Spawning	Adult/juvenile
1	Immature/Resting		J	Immature/resting		J
2	Maturing		A	Early maturation		J
3	Mature/Ripe	+	A	Mature		A
4	Running Ripe	+	A	Ripe	+	A
5	Spent	+	A	Running ripe	+	A
6				Spent	+	A
7				Atretic		A
8				Partially spent	+	A
9				Mature and resting[1]		A

[1] implies no reproductive activity this season (this stage has been used only since 2004)

Our interest in this study is to distinguish between 'adult' and 'juvenile' fish (I use these terms rather than the more usual 'mature' and 'immature' to avoid any confusion with the particular use of the latter pair of terms in the two gonad-staging schemes). In principle, we should be able to designate each of the gonad stages in Table 13 as juvenile or adult simply on the basis of the definitions of these stages. However, the stages in our databases were assigned by a wide range of people, most of whom are not specialists in this field. It is very likely that different stages would sometimes have been assigned had all staging been done by specialists (and there may not even be unanimity amongst specialists). For this reason I used a combination of the stage definitions and patterns in the data to decide which gonad stages should be designated as adult. The most clear-cut stages are those which show definite evidence that a fish will spawn soon, or has spawned recently (these are labelled 'spawning' in Table 13). These are obviously adult stages. Fish of observer stage 2 and research stage 3 were also classed as adults because their length frequencies were so similar to those of spawning fish (Figure 27). Fish of research stages

7 (atretic) and 9 (mature and resting) lay within, but towards the lower end, of the spawning length distribution. Their definitions suggest that they should be included as adults, and so they were. However, it's worth noting that there were so few fish of these stages that they made no visible difference to the length frequency of adult fish. The similarity between the length frequencies from SOP and ORMC observers (compare panels A and B of Figure 27) indicate that two groups of observers used similar interpretations of the staging scheme.

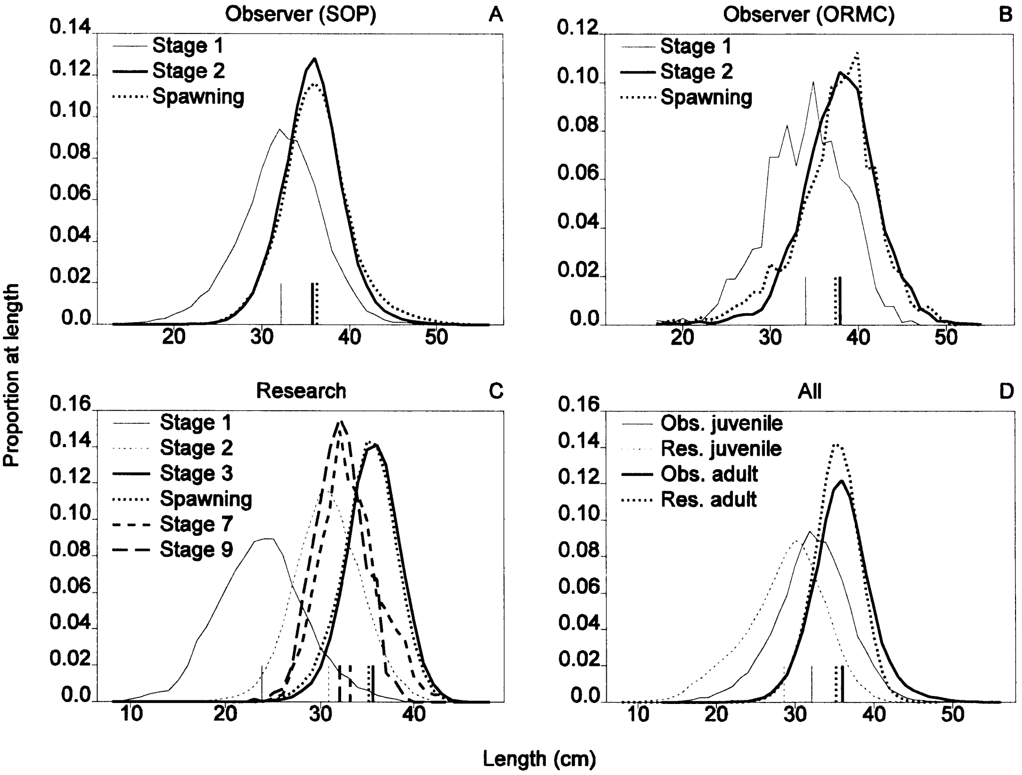

Figure 27: Length frequencies by selected gonad stages, or groups of stages, for data from: A, SOP observers; B, ORMC observers; C, research trips; and D, all data. The vertical line segments at the bottom of each panel show the means of the length frequencies. For reasons of stability, data for each length frequency were combined without scaling by catch weight.

Plots of seasonal changes in the proportions of fish by stage (Figure 28) are also useful. The substantial drop in the proportion of observer stage 2 fish during the spawning season supports the inclusion of these fish as adults. Seasonal changes in the proportion of research stage 3 fish are consistent with the usual interpretation that this is a stage that adult females enter early in the calendar year when they are preparing to spawn. The high proportion of observer stage 1 in September suggests that, in this month (and possibly the following few months), this stage may include some fish that spawned in the previous winter. Thus, this stage may include some adult fish. The same may be true of research stage 2, which also rises sharply after spawning. However, it's unlikely that adult fish dominate either of these stages, because of the difference between their length frequencies and those of spawning fish. Therefore I have classified observer stage 1 and research stage 2 as juvenile, despite some evidence of contamination with adults. In Figure 27D, the fact that the right-hand limb of the juvenile length frequency from observer data is to the right of that from research data suggests that contamination is worse for the observer data. The similarity between the two adult length frequencies in the same plot is reassuring.

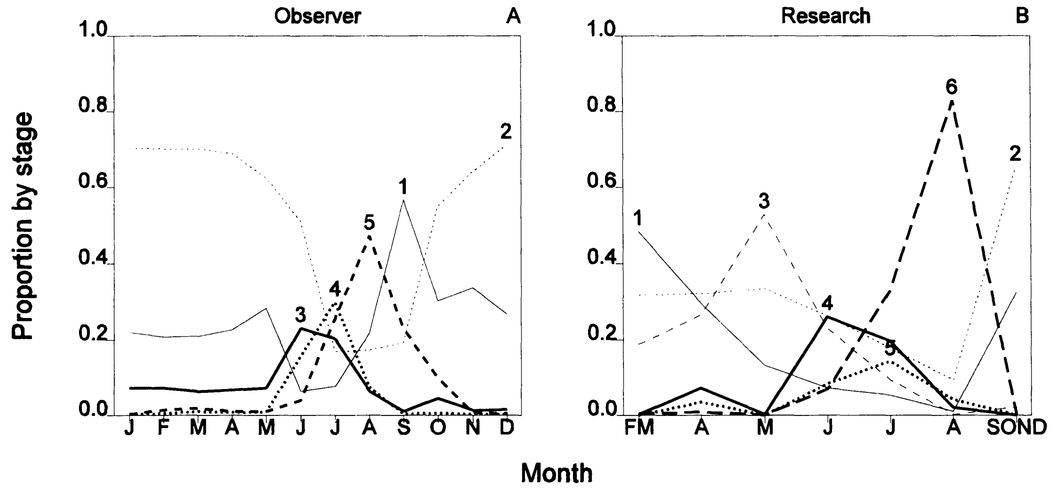

Figure 28: Monthly variation in the proportions of fish in each gonad stage for data from A, observers; and B, research trips. In panel B, stage 8 fish (partially spent) were included with stage 5 (spent); months with few data were lumped; and stages 7 and 9 (whose proportions rarely exceeded 0.01) were not plotted. For reasons of stability, data were combined in each month without scaling by catch weight.

The proportion, by weight, of adults in each observed catch was estimated as $p = \sum_i \left(L_i^b n_{ai}\right) / \sum_i \left(L_i^b n_i\right)$, where n_{ai} is the number of adults amongst the n_i staged fish of length L_i and b = 2.71 is the standard exponent in the length-weight relationship for orange roughy (Sullivan et al. 2005). Note that I am assuming, in the lack of evidence to the contrary, that this proportion is the same for males and females. This proportion varied widely, but was typically greater than 0.8, and was higher in the spawning months of June and July (Figure 29A). Monthly average proportions were calculated as $\sum_t p_t C_t / \sum_t C_t$, where p_t and C_t are the proportion adult and (orange roughy) catch weight for the *t*th tow in the month. The proportion was about 0.97 in the spawning months and, ignoring the three months following spawning (presumably affected by adult contamination), was quite stable in the non-spawning months at about 0.85. Averaged across whole years the proportion adult has fluctuated around 0.87 over the last 10 years (Figure 29B).

The picture becomes more complicated when we look at the observer data in more detail. The seasonal pattern, when broken down by area (Figure 30), becomes much more variable, in part because of smaller, and highly variable, sample sizes (Table 14). When we focus on the anomalous month of September (where the proportion adult was so low in Figure 29A) we see that the total sample for the month was dominated by a minor area (ARW, which contributed 40% of the observed catch), and that the proportion adult was high in one important area (NWChat). It is clearly difficult to know what the proportion adult would have been had the entire September catch been observed. If, instead of looking at just a single month, we compare September to November (where proportions adult are lower in Figure 29A) with December to May (where they are higher and more stable) we see that it is in only 4 of our 14 areas that the proportions are much lower in the former period (Figure 31). Why these areas stand out from the others is unclear.

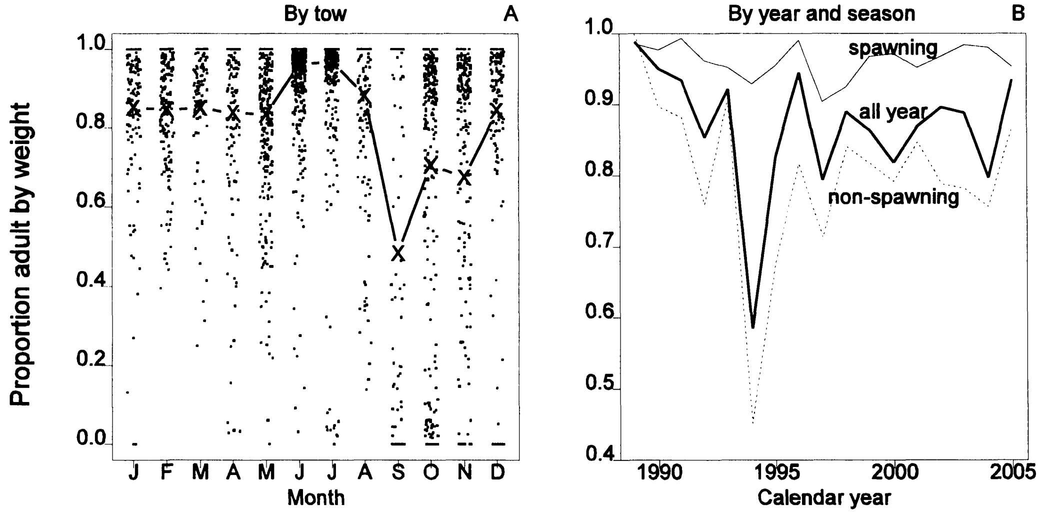

Figure 29: Proportion (by weight) of adult fish in the catch: A, by tow (each point represents one tow, but tows catching less than 2 t are omitted, for clarity, and 'X' represents the mean proportion, weighted by catch, for each month); and B, mean proportions (weighted by catch) by year and season (where the spawning season is defined as June and July).

Table 14: Total weight (t) of orange roughy in observed tows by area and month.

Area	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
NWChat	101	64	50	30	213	690	417	11	159	578	186	183
Box	6	25	0	12	301	4 571	4 767	290	0	2	113	18
EChat	875	753	218	375	623	216	424	924	252	1 029	879	1 588
SChat	237	258	202	156	252	32	13	108	80	509	514	249
Chall	0	0	0	1	6	810	1 078	40	6	0	0	0
Howe1	1	26	57	15	342	965	366	0	30	5	0	0
Howe2	12	49	28	10	21	351	636	3	0	1	21	0
SE	0	1	3	129	6	14	3	2	52	414	2	0
SW	40	213	132	63	87	172	289	30	69	54	248	211
WCSI	12	4	35	0	82	60	4	22	12	0	0	0
NW	71	56	273	0	105	59	88	5	0	199	83	1
NE	8	4	13	14	56	723	179	1	0	28	11	0
2AN	0	7	2	0	132	479	54	0	0	5	2	0
2AS	0	6	36	63	32	268	34	0	0	49	5	42
2B3A	49	0	86	84	42	2	0	25	0	62	9	22
ARW	75	41	42	82	146	137	33	285	466	132	0	0
Lour	31	35	32	5	49	335	221	114	31	0	0	0
Other	0	13	122	32	191	0	6	177	0	82	30	0

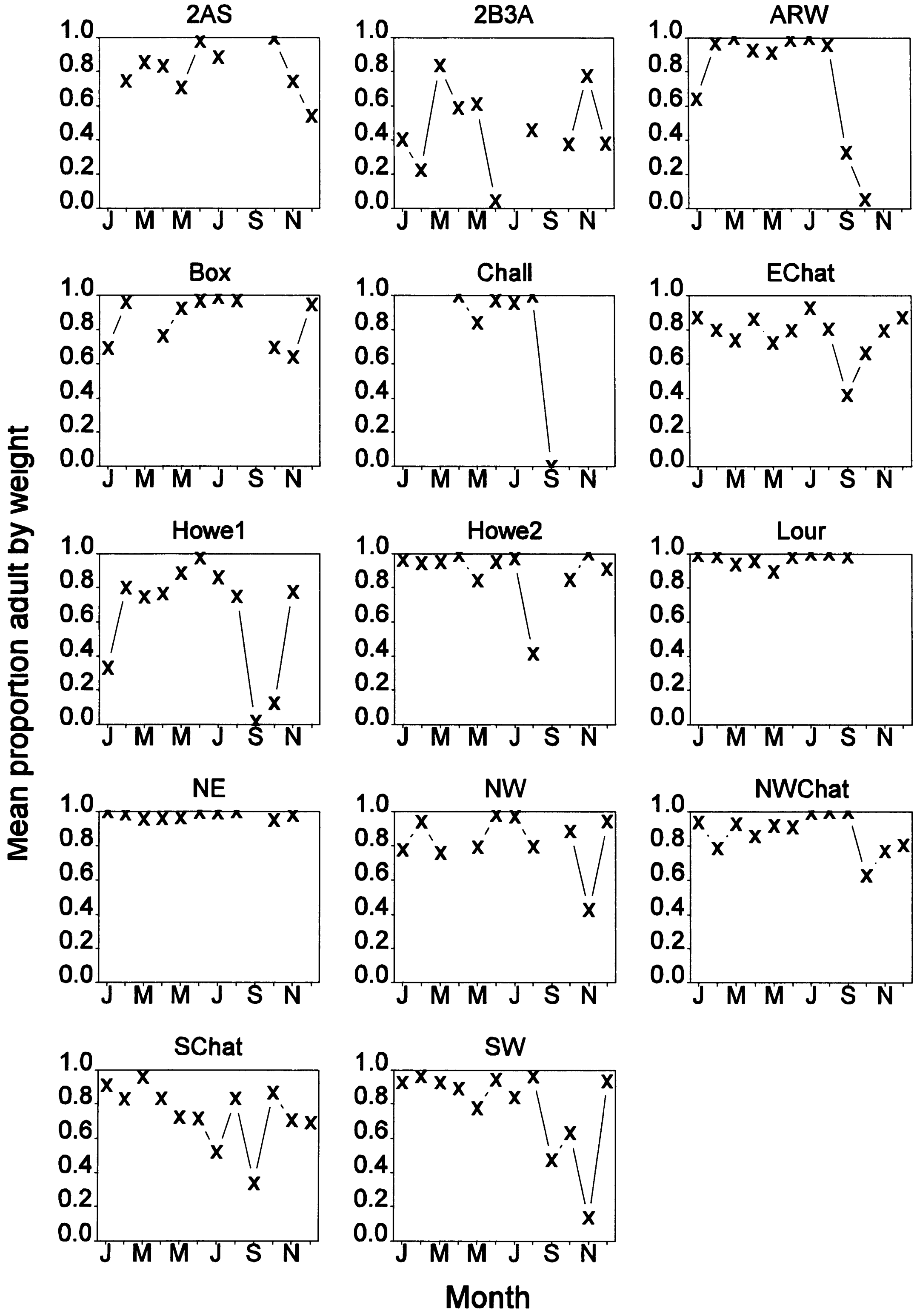

Figure 30: Proportion (by weight) of adult fish in the catch, by month (all years combined) and area (the areas are shown in Figure 26; those with fewer than 100 observed tows are omitted).

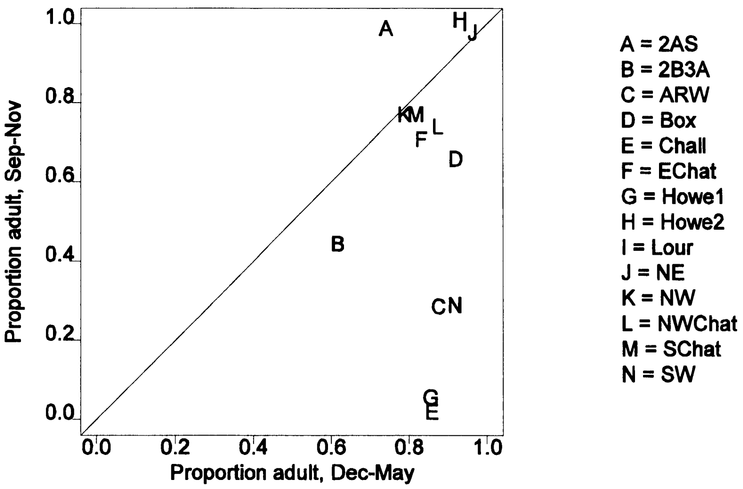

Figure 31: Comparison of the proportion adult in the catch, by area, in two parts of the non-spawning period: December to May (x-axis) and September to November (y-axis).

4.1.1 Errors in gonad staging

There are certainly errors in the above gonad stage data. Two recent studies have shown, using microscopic examination of ovaries, that macroscopic staging is sometimes wrong. Our interest here is to see what effect such errors are likely to have on the above estimates of proportion of the commercial catch that is adult.

The first study (Doonan et al. 2004) is not much use to us. It focussed on how well macroscopic staging can distinguish between the different phases of active spawning. The second study, by Grimes et al. (2005), considered all stages and thus is of some use. This study did not explicitly compare macroscopic and histological stages but, after discussion with the authors, I concluded that their sample was about 22% (= 44/203) adult histologically, and 34% (= 68/203), macroscopically (Paul Grimes, Matt Dunn, NIWA, pers. comm.). Of particular note was the fact that macroscopic stage 6 was particularly bad, with only 14 of 42 such fish being classed as adult histologically. However, we should not try to generalise from these results because the sample used in this study was not randomly selected, and mis-classification rates are certain to vary strongly with season and gonad stage. The most we can say is that there is some evidence that macroscopic staging over-estimates the proportion adult.

Thus we have two conflicting pieces of information about adult/juvenile classification errors. Seasonal patterns in the incidence of juveniles suggest we may under-estimate the proportion adult, but work by Grimes et al. (2005) suggests the opposite. I did find one recent paper, for a different species (Kattegat cod, *Gadus morhua*), that the proportion adult was strongly over-estimated by macroscopic staging. This made the estimated female spawning biomass too high by up to 35% (Vitale et al. 2006).

4.1.2 Wide-area surveys

Amongst the many random trawl surveys for orange roughy, there are four which are of particular interest for this study (Table 15). These all covered wide areas, which approximate the total area for the stock, and took place during those months in which it is possible to identify which females are going to spawn in the coming winter (i.e, those with gonad stage > 2). Thus, for each of these surveys we can estimate an LF for the spawning females in that stock for that year. These LFs are much more statistically respectable than those presented above because we are able to correctly weight contributions from individual tows (using catch rates and stratum areas) and each LF is thus representative of a whole stock (although it may be affected by the vessel selectivity). Thus it is reassuring that these LFs are quite similar to those calculated for adult females from (unweighted) observer data in the corresponding areas (Figure 32). In the comparisons in Figure 32 I restricted the observer data to the corresponding survey areas because I found that there was substantial between-area variation in the median size of adult fish (Figure 33).

Table 15: Details of four wide-area random-trawl surveys that occurred before the spawning season.

Year	Trip code	Stock	Dates	Reference
1992	tan9203	Mid-East Coast (MEC)	5 Mar – 2 Apr	Grimes (1994)
1993	tan9303	MEC	16 Mar – 10 Apr	Grimes (1996a)
1994	tan9404	MEC	16 Mar – 10 Apr	Grimes (1996b)
1994	tan9406	NW Chatham Rise	21 May – 6 Jun[1]	Tracey & Fenaughty (1997)

[1] These dates are just for the NW Chatham Rise area; this survey also covered other areas.

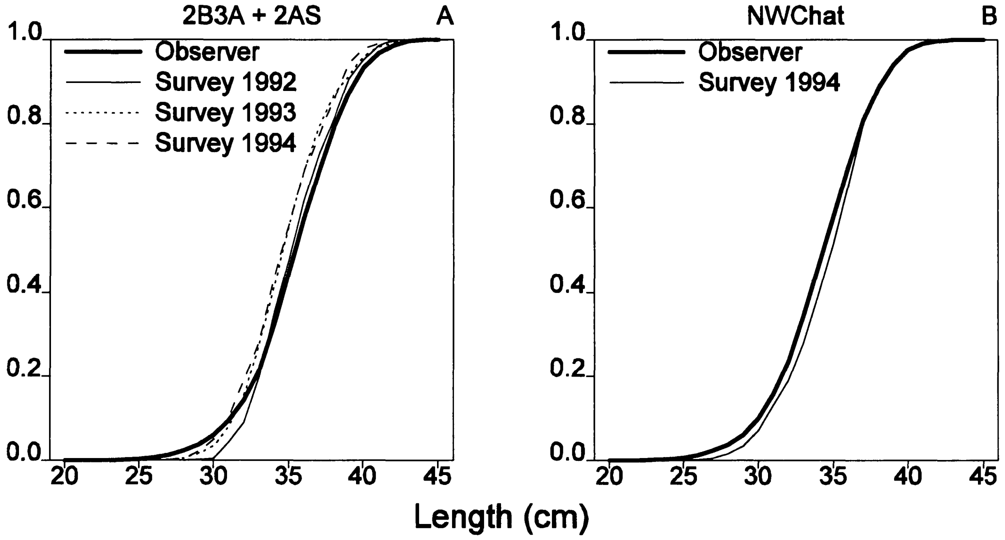

Figure 32: Comparison of cumulative length distributions for spawning females estimated from wide-area surveys (thin lines) with those for adult females from the observer data in corresponding areas (heavy lines) (observer areas used for the Mid-East Coast stock were 2AS and 2B3A).

From these surveys we can also calculate the proportion adult for each length, and thus L50, the length at which 50% of females are adult (Figure 34). These estimates are 2 to 3 cm larger than the estimates of mean fish length at onset of maturity made by Horn et al. (1998) (Table 16).

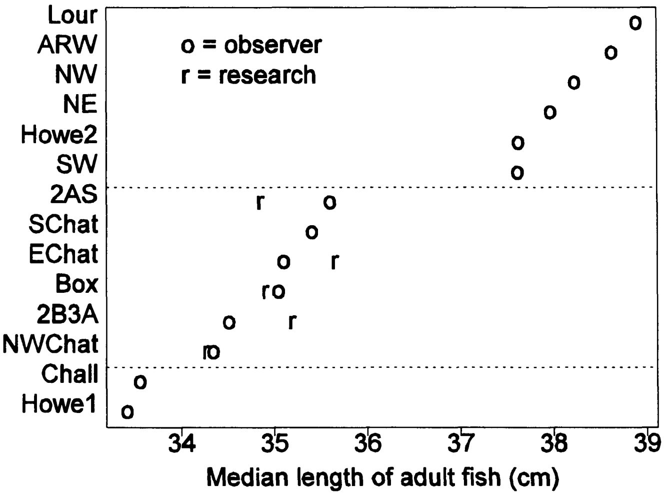

Figure 33: Median length of adult fish by area from observer and research data. The data are unweighted, and medians based on fewer than 10 tows are not plotted. Areas are as shown in Figure 26. The horizontal broken lines separate the areas into three apparent groups on the basis of similarity of median length.

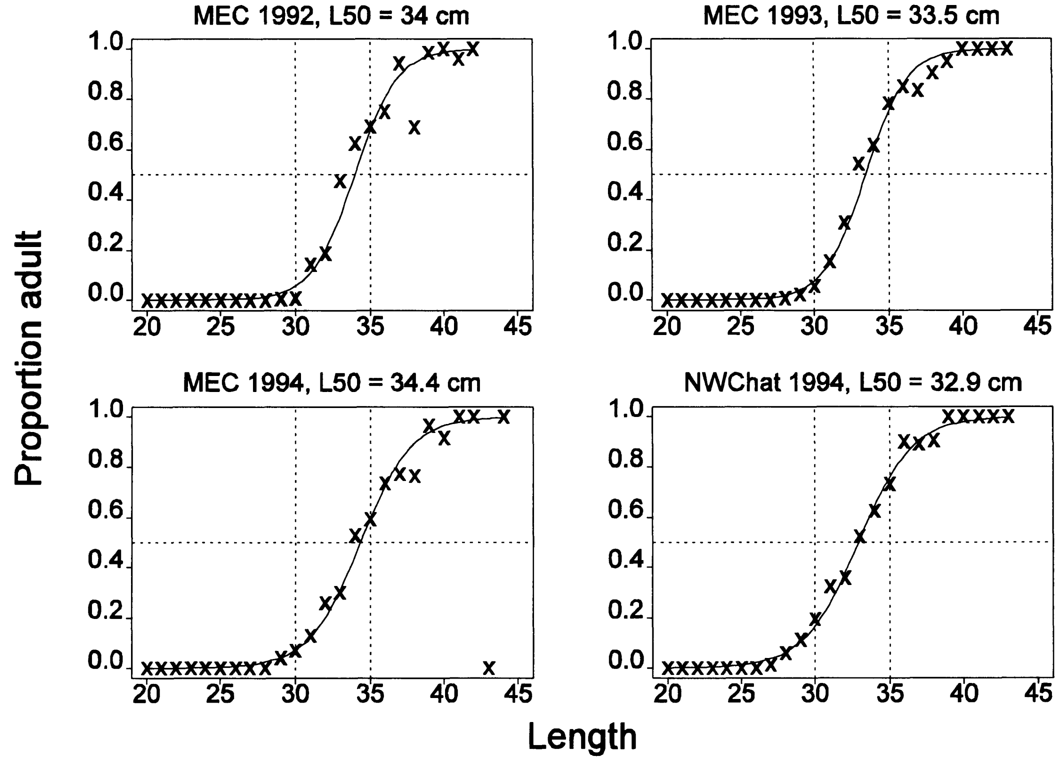

Figure 34: Estimated proportion adult by length for the four wide-area surveys of Table 15. The plotted points were estimated from the wide-area surveys (weighting observations appropriately by catch rates and stratum areas) and the solid line is a logistic curve fitted to these points. The estimated length at which 50% are adult (L50) is written above each panel. The dotted lines (same in each panel) are added to aid comparison between panels.

Table 16: Comparison between estimates of two variants of the length at maturity for females: L50 (length at which 50% of females are adult) and L_{mat} = mean length at onset of maturity.

Area	Quantity	Estimate(s) (cm)	Source
MEC	L50	33.5–34.4	Figure 34
Ritchie	L_{mat}	30.75	table 4 of Horn et al. (1998)
NWChat	L50	32.9	Figure 34
Chatham Rise	L_{mat}	30.75	table 4 of Horn et al. (1998)

The difference between the areas covered by the bottom two samples in Table 16 does not seem to be problematic. The sample labelled 'Chatham Rise' contained 29 otoliths from NWChat and 180 from other parts of the Chatham Rise. However, there was no significant different difference between the counts to the transition zone in these two areas (Mann-Whitney test, P = 0.16). Neither was there a significant different difference between all the counts in this sample and those in CAF batches 142–144, which were all from NWChat (Mann-Whitney test, P = 0.40).

4.2 The effect of model assumptions on the estimated selectivity curve

A puzzling aspect of several recent orange roughy assessments is the way the estimated commercial selectivity curve sometimes changed substantially when a seemingly unrelated model assumption was changed. In this section I examine two examples of this behaviour to shed some light on this phenomenon.

The examples are from the 2005 assessments of two stocks: Andes and Spawning Box. In both assessments the only data that obviously contained information about the commercial selectivity were observer LF samples from the commercial catch (there were no age data from the commercial catch). The range of estimates of a50 in the examples was 5.2 y amongst three Andes runs, and 7 y between two Box runs (Figure 35). One consequence of this wide range of estimates is that the amount of cryptic biomass (biomass that is mature but not vulnerable) varied from a lot to nothing: the vulnerable biomass in 2005 lay between 27% and 102% of the mature biomass (Table 17).

Table 17: Some details of the five model runs shown in Figure 35. B_0 = virgin biomass, a50 = age at 50% vulnerability for the commercial fishery, a50.surv = age at 50% vulnerability for the trawl survey, cv1 and cv2 = c.v.s for length at age, M = natural mortality, beta = curvature parameter for the relationship between CPUE and biomass, YCS = year-class strengths (i.e., stochastic recruitment).

Stock	Run	Parameters estimated	a50 (y)	$B_{\text{vulnerable, 2005}}$ (% $B_{\text{mature, 2005}}$)
Andes	Beta1	B_0, a50, cv1, cv2	32.8	50
	EstM	B_0, a50, cv1, cv2, M	35.2	27
	EstBeta	B_0, a50, cv1, cv2, beta	30.0	97
Box	acous.hiabs	B_0, a50, a50.surv, cv1, cv2	36.5	66
	acous.hiabsR	B_0, a50, a50.surv, cv1, cv2, YCS	29.5	102

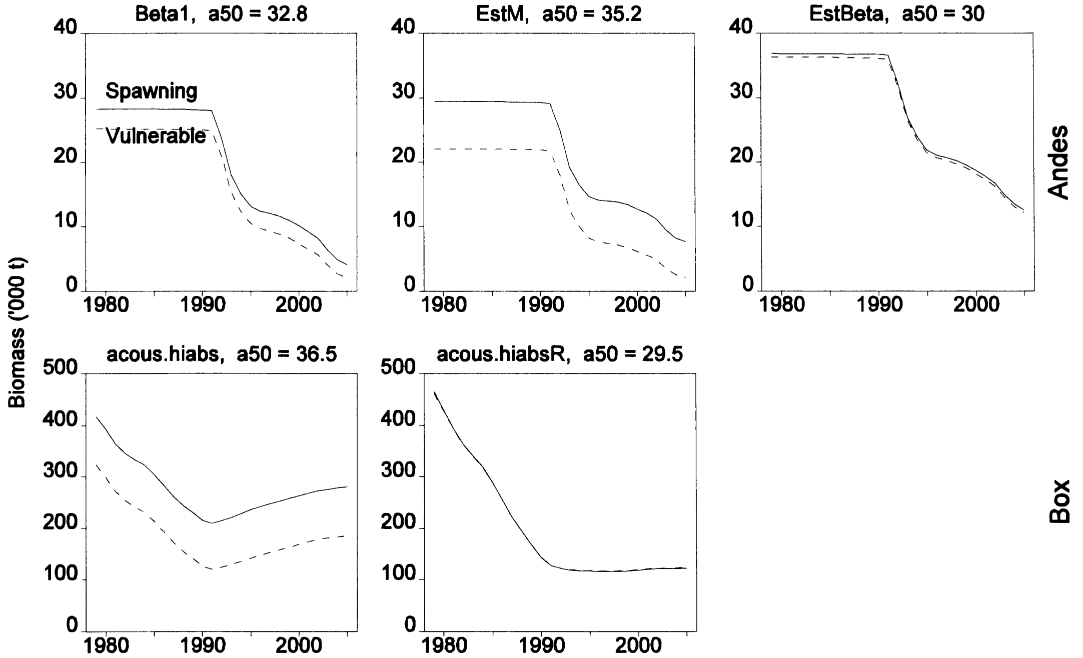

Figure 35: Five model runs from the 2005 northeast Chatham Rise assessment illustrating how the model estimate of a50 (the age at 50% selectivity for the commercial fishery) changed with model assumptions. Each panel shows estimated trajectories for spawning and vulnerable biomass for one run: top row, three runs from the Andes assessment; bottom row, two runs from the assessment of the Spawning Box plus eastern flats. The name of the run and the estimate of a50 are shown above each panel.

The first question to ask is which data set is determining the estimate of a50 in each run. We presume that it is the observer LFs, but this needs to be confirmed. This I did by constructing posterior profiles for a50. That is, for each run I estimated the best fit for each of a series of values of a50. I then plotted curves that show how the contribution to the objective function from these LFs changes with a50. The minimum of each curve is at the value of a50 that best fits the observer LFs. The main thing to notice about these minima is that they are always at or near the overall best value for a50 (plotted as 'o' in Figure 36). This means that, as we presumed, the observer LFs are the main determinants of our estimate of a50.

As an aside, we can look at the other components of the objective function to see which were most influential in the two cases where the overall best value of a50 was not the same as that which gave the best fit to the observer LFs. There are three components of interest: one for Andes (CPUE) and two for Box (acous and AF.surv) (Figure 37). All caused the estimate of a50 to be slightly less than that which gave the best fit to the observer LFs.

Having confirmed that it is primarily the observer LFs that determine a50 in these runs I next asked why these data produced such different estimates of a50 in different runs. For most of the runs, the answer is clear when we look at the expected mean lengths from the profiles (Figures 38, 39). In three of our five runs (EstBeta in Andes and both Box runs) it is not possible to find a value of a50 for which the estimated mean lengths match the trend in the observed mean lengths. This means that a value of a50 which fits the LF in the first year will not fit well in the last year. Thus, for these runs, the best value of a50 is a compromise. Because the nature of the compromise is different in each run, we get different estimate of a50. It is perhaps useful to think of these estimates of a50 as being the 'least bad', rather than the 'best'. For the Andes run Beta1, the slopes of the observed and expected declines in mean length match fairly well and the model is able to select a value of a50 which best goes through these data. For run EstM, all values of a50 produce about the same trend in expected mean length (bottom panel, Figure 38) which is why the curve for this run in Figure 36 is so much flatter than those for the other runs. Presumably, the choice of best a50 here is determined by differences in the shapes of the expected LFs, rather than differences in their means.

For the record, I include plots of the way in which other model parameters varied with a50 in the profiles (Figure 40).

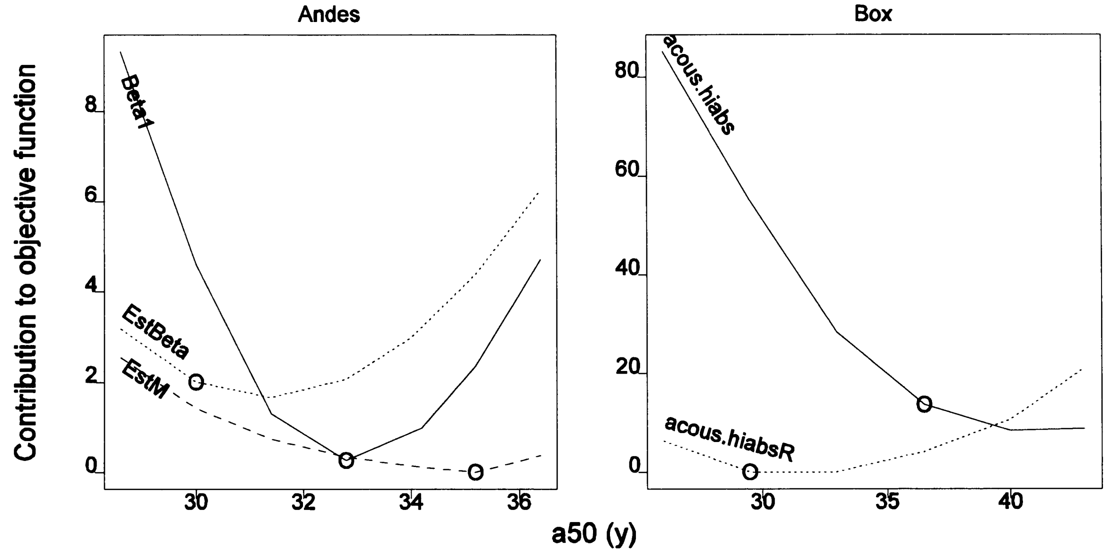

Figure 36: Contribution of observer LFs to the objective function in posterior profiles on the parameter a50 based on each of the five model runs of Figure 35. In each panel there is one line for each profile; all lines have been shifted vertically by the same amount so that the lowest value amongst all of them is zero; and 'o' indicates the best estimate of a50 in that profile.

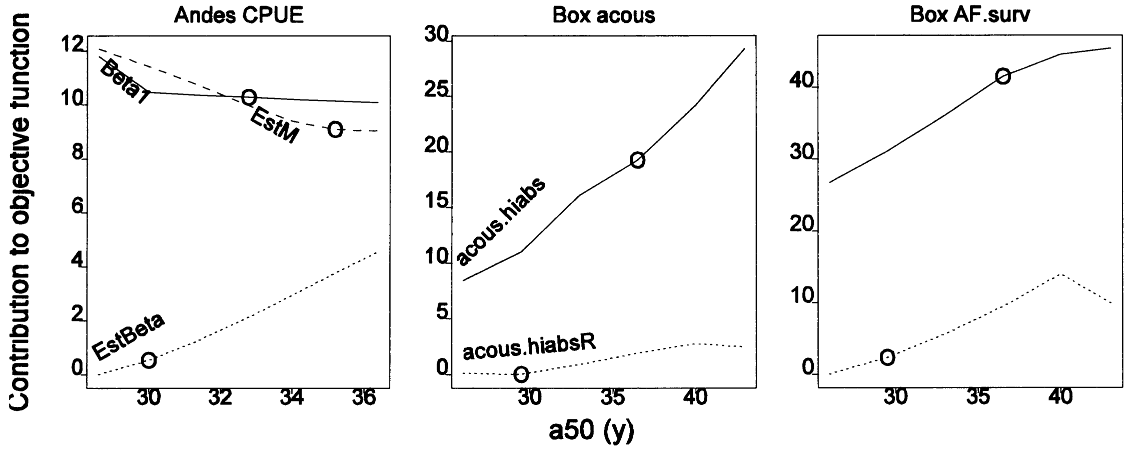

Figure 37: Other significant components of the objective function (besides those from the observer LFs) in posterior profiles on the parameter a50 based on each of the five model runs of Figure 35. In each panel there is one line for each profile; all lines have been shifted vertically by the same amount so that the lowest value amongst all of them is zero; and 'o' indicates the best estimate of a50 in that profile. The left panel shows the contribution of CPUE in the three Andes profiles (it made a significant difference only for the run EstBeta); the other panels show two components (acous and AF.surv) that were significant in the acous.hiabs run in the Box.

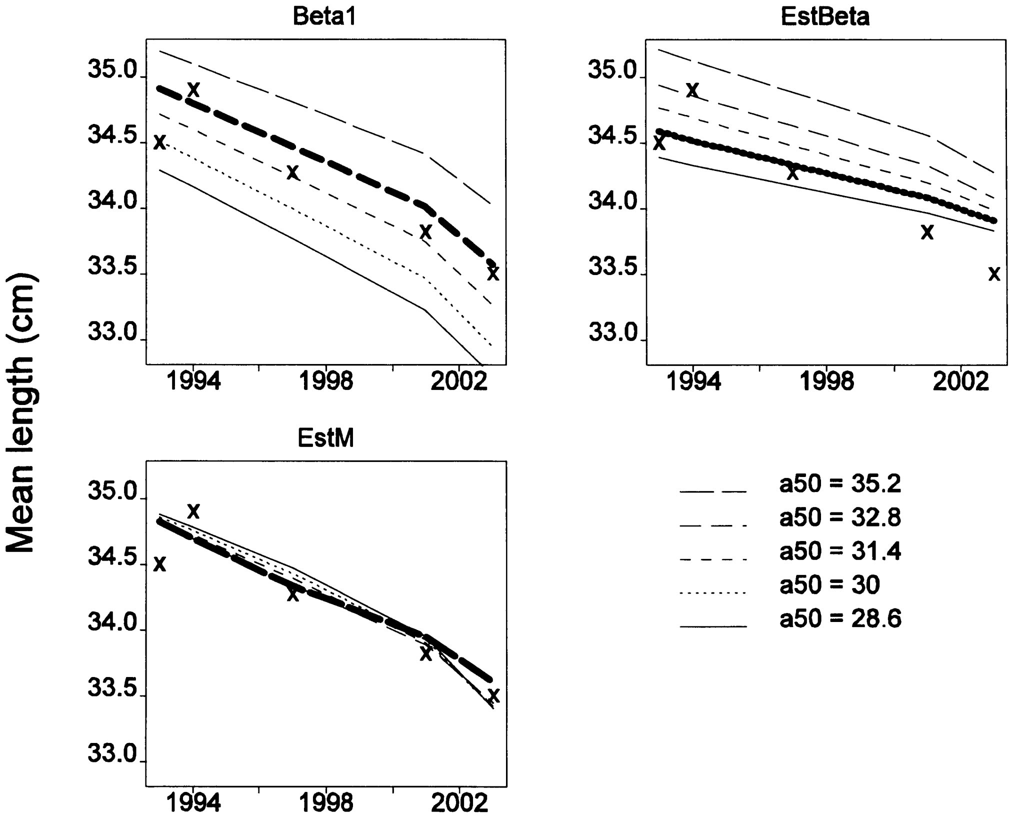

Figure 38: Observed ('x') and estimated (lines) mean lengths of the observer LFs for selected parts of the profiles for the three Andes runs. Each line in a panel shows the estimated mean lengths for a specific value of a50 and the thick line in each panel is the best fit.

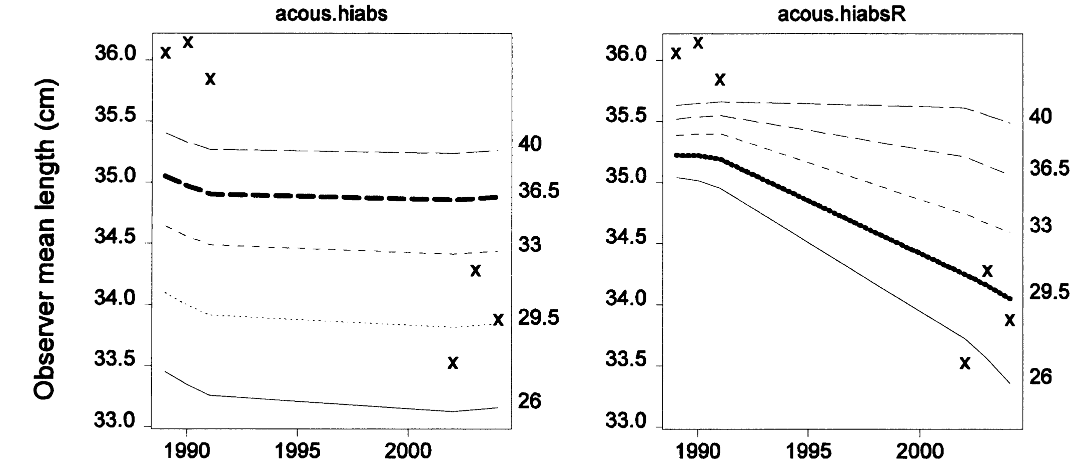

Figure 39: Observed ('x') and estimated (lines) mean lengths of the observer LFs for selected parts of the profiles for the two Box runs. Each line in a panel shows the estimated mean lengths for a specific value of a50; the value of a50 for each line is shown to the right of the plots; and the thick line is the best fit.

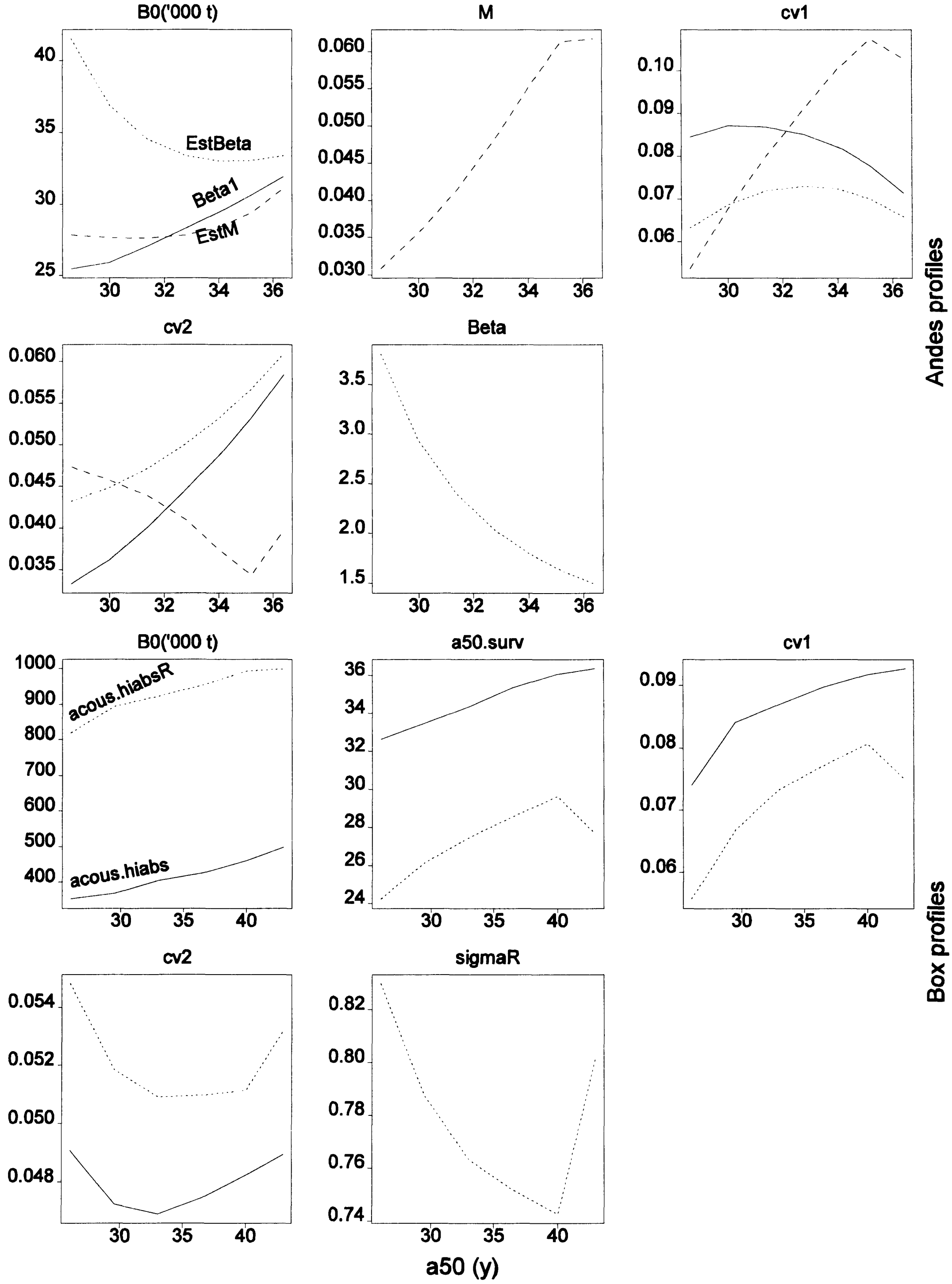

Figure 40: Variation of other model parameters (those besides a50) in the profiles. The upper two rows are for the Andes profiles and the lower two rows are for the Box profiles. There were too many YCS parameters to plot for Box run acous.hiabsR; instead, the standard deviation of the log(YCS) estimates, here labelled sigmaR, was plotted.

4.3 Other sources of uncertainty

4.3.1 Conflict between age-length data sets

One interesting aspect of the 2005 model runs for the Box was the poor fit to the survey LFs. The model estimated a trend in mean length which was similar in slope to that shown in these data but about 0.7 cm higher (Figure 41A). One reason for this is that the observed LFs typically had a long tail to the left, whereas those predicted by the model did not (Figure 41B).

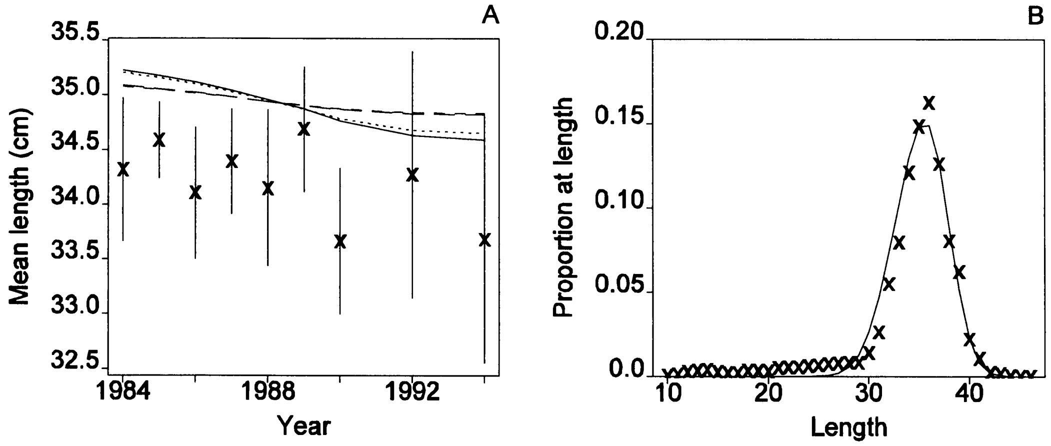

Figure 41: Two aspects of the fit to survey LFs in some runs from the 2005 assessment: A, comparison between observed mean lengths (plotted as 'x', with 95% confidence intervals indicated by the vertical lines) and those estimated by the model for four alternative runs (curved lines); comparison between the observed LF for 1987 ('x') and that estimated by the model in one run (line) (the estimated LFs from the other runs were very similar).

I suspected that another reason was that there was a conflict between the von Bertalanffy curve used in the assessment and the age and length data used in constructing the survey AFs (age frequencies) for 1984 and 1990. These AFs were constructed using age-length keys (ALKs) (Hicks 2005b). I recreated these ALKs, with rows indexed by length bin and columns by age, and a plus group at age 80. By definition, each row of an ALK must sum to 1. For each year, I then multiplied each column of the ALK by the estimated survey LF for that year, creating a matrix in which each column holds the estimated numbers at length for an age class. From this, I calculated a mean length for each age in each year. I then compared these mean lengths at age with those from von Bertalanffy curve and found that they lay below the curve for much of the age range (Figure 42).

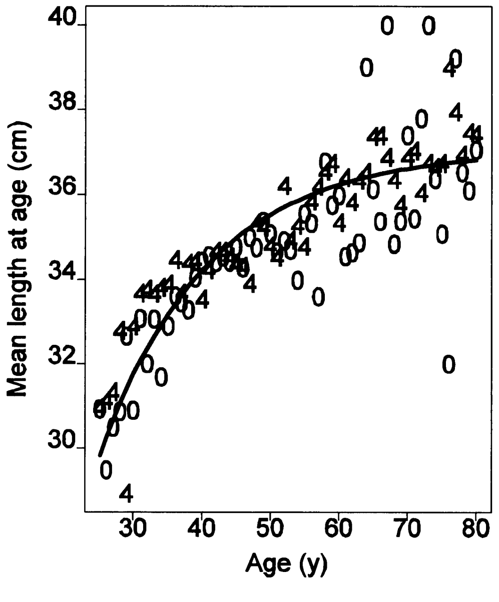

Figure 42: Comparison of mean lengths at age derived from survey AFs ('4' = 1984, '0' = 1990) with those from the von Bertalanffy curve (line). The von Bertalanffy curve has been shifted down by 0.5 cm to allow for the fact that lengths were floored in constructing ALKs.

4.3.2 Conflict between estimates of age at maturity for Ritchie Banks

There is a striking difference between two data sets estimating age at maturity for Ritchie Banks (Figure 43). The difference is highly significant (P < 0.001, Mann-Whitney test) and emphasises the uncertainty surrounding maturity ogives used in orange roughy stock assessments.

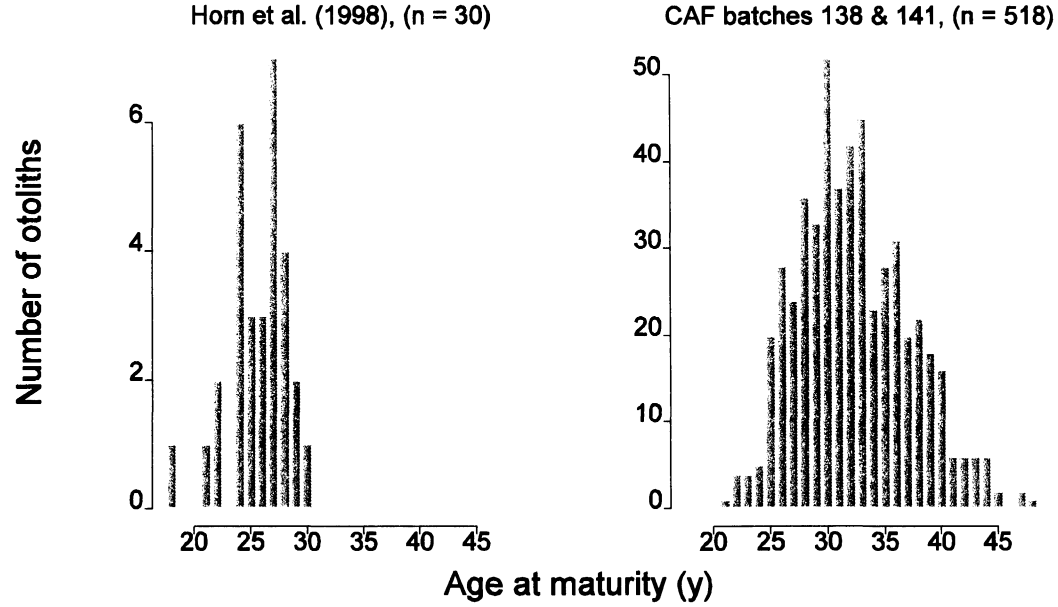

Figure 43: Estimates of age at maturity from counts to the transition zone in two samples of otoliths from the Ritchie Banks region.

4.3.3 Bias in estimates of age at maturity

The usual method of estimating the mean age at maturity from a sample of otoliths is simply to calculate the mean *agetoTZ* in the sample (where *agetoTZ* is the number of rings inside the TZ, if there is one).

An overlooked source of negative bias in this method is illustrated by the curved line in Figure 44A. In this sample, *agetoTZ* ranged from 15 to 43. To see how this bias arises, consider all fish if age 30 y, say, in the sample. These fish fall into two groups: those with a TZ, and those without. The former group will have age at maturity less than or equal to 30, and will contribute to the overall mean; the latter group will not contribute to the mean, but their age at maturity must exceed 30. The omission of the latter group thus produces a negative bias. One simple way of avoiding this bias would be to omit all otoliths whose age at capture was less than the maximum *agetoTZ*. For the sample shown, the cutoff is at 43 y, and omitting fish younger than this increased the estimated mean age at maturity from 29.3 y to 30.0 y.

Another possible source of bias is the fact that the TZ is not easy to recognise until there are several rings outside it. This is evident from the fact that there tend to be few otoliths in which there are only one or two rings outside the TZ (Figure 44B). Thus, in our present example we might, to be on the safe side, increase the cutoff age by say 5 y (this increased the estimated mean age at maturity by only 0.1 y).

This method of avoiding bias is simple, but it seems a pity to throw away data. An alternative, though much more complicated, approach which might be worth pursuing would be to use a model-based estimator which does not ignore those otoliths without a TZ, and takes into account the difficulty of detecting a TZ with few rings outside it.

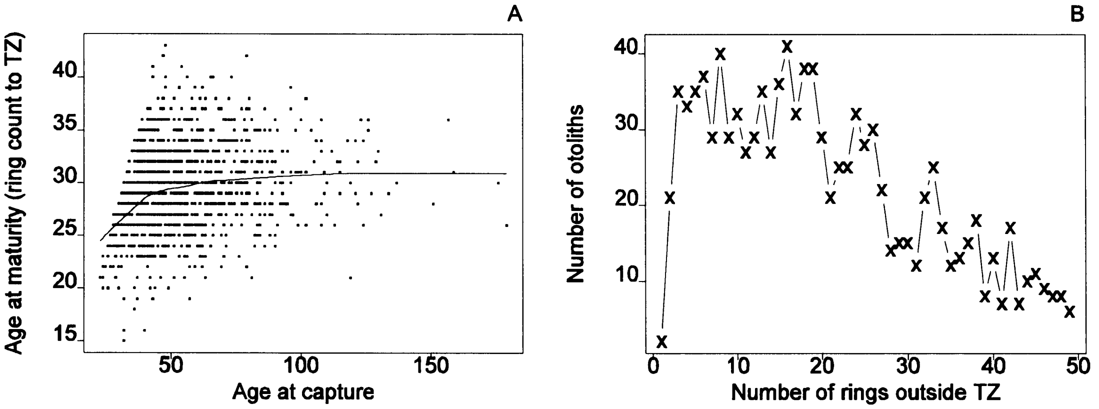

Figure 44: Illustration, using otoliths from the Spawning Box, of two potential sources of bias in estimates of age at maturity: A, age at maturity plotted against age at capture (each plotted point is one otolith, and a lowess line is plotted to show the trend which can cause bias); and B, demonstration that the TZ is often not detected when there are few rings outside it (for clarity, data for otoliths with more than 50 rings outside the TZ were omitted from this panel).

4.4 Conclusions and recommendation on selectivity and maturity

The above analyses do not support the idea that the commercial selectivity ogive is well to the right of the maturity ogive. If anything it appears to be slightly to the left because, on average, about 87% of the commercial catch is adult. Thus, it would seem sensible to continue the recent practice of forcing the two ogives to be equal.

In deciding whether this should be achieved by setting maturity equal to selectivity (*sel2mat*) or vice versa (*mat2sel*), there are three points to consider. First, neither can be estimated with confidence. Estimates of maturity outside the assessment model are still very uncertain (see the differences in Table 16 and Figure 43, and the problems discussed in Section 4.3.3) and estimates of selectivity inside the model are unstable. The analyses of Section 4.2 give us an understanding of the source of this instability, but they do not provide us with a way of avoiding it. The second point is that it doesn't seem to matter much which approach we choose because estimates of the main stock-assessment output – current biomass as $\%B_0$ – are not very sensitive to this choice (Table 18). The final point is that *sel2mat* is much simpler to implement because it allows us to avoid the problems associated with observer length data (see Section 3), which are used either directly (as LFs) or indirectly (in constructing AFs via and age-length keys) to estimate the commercial selectivity ogive.

My recommendation, based on these points, is that the *sel2mat* approach be used in orange roughy assessments for the meanwhile. This means avoiding the use of observer length data in these assessments, which will save the time currently spent first in preparing such data, and then in trying to avoid or explain the poor fits that are so often associated with them. I believe this is unlikely to reduce the quality of the assessments as I am unaware of any orange roughy assessment which has been improved by the use of these data.

Table 18: Pairs of estimates of stock status (B_{current} as $\%B_0$) from recent assessments. For each assessment, the two estimates given are from runs which differ only in whether they assume *sel2mat* or *mat2sel*.

	Year of	B_{current} $(\%B_0)$		
Stock	assessment	*sel2mat*	*mat2sel*	Source
Mid-East Coast	2004	22	18	Dunn (2005a), table 10[1]
Northwest Chatham Rise	2006	13	11	McKenzie (2006), tables 2 & 3
Andes	2006	23	23	Dunn (2006), slide 7[2]
Northeast hills	2006	13	13	Dunn (2006), slide 14[2]
Spawning Box & NE flats	2006	53	53	M. Dunn, NIWA, pers. comm.

[1] Runs *Seltomat* and *Beta1*, [2] Runs *sel2mat.short*, *mat2sel.short*

5. REVIEW OF SPAWNING-BOX TRAWL DATA FROM THE 2004 SURVEY

The series of trawl surveys in the Spawning Box (1984–94) was discontinued because it appeared to be no longer possible to achieve acceptable coefficients of variation (c.v.s) (Francis 1996). During the first workshop of the *Review of Methods and Data Used in Orange Roughy Stock Assessments* (held in Wellington 10–12 October 2005) the reviewers asked about the possibility of carrying out another Spawning Box trawl survey to continue this series. In this section, I review the Spawning-Box trawl data from the 2004 combined acoustic-trawl survey (Dunn 2005b) to see what light they shed on this question.

5.1 Problems with the original series

The main problem with the original series of trawl surveys was that the biomass c.v.s became unacceptably high (Figure 45A). Another feature of some (but lesser) concern was that the overall sex ratio in these surveys started to diverge substantially from its earlier values near 50:50 (Figure 45B).

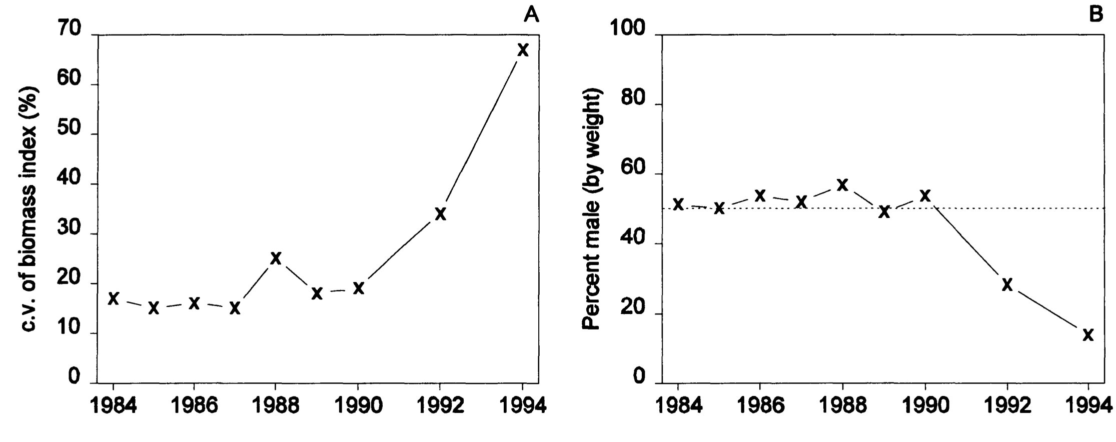

Figure 45: Two problems experienced in the original series of Spawning Box trawl surveys: A, the substantial increase in the c.v. of the biomass index (this is the primary problem), and B, the deviation of the sex ratio from its earlier value near 50:50.

Another survey was carried out in 1995, and this was designed to gain some understanding of the reason for these two problems. The main activity was four repeated surveys of the key strata (those contributing most to the biomass estimates in the preceding surveys), which allowed a characterisation of how catch rates and sex ratios varied over time. Data from all surveys in this area, together with sex-ratio data from the commercial fishery, were analysed with a focus on the two problems (Francis 1996).

With regard to the first problem, it was concluded that it would not be possible to design future random trawl surveys to avoid high c.v.s. The main reason for this was that, although the incidence of high catch rates had not changed systematically over time (usually, catch rates exceeded 10 t/n.mile in 2% to 4% of tows) there had been a dramatic reduction in the area in which moderate to high catches could be obtained (Figure 46). In other words, the distribution of fish had gradually become more and more clumped. Further, Francis (1996) said

> *It is not possible to solve this problem by intensifying the survey coverage in the high density area because this would inevitably result in more large catches, which are already a problem because of limited processing capacity. Nor is it a good idea to reduce the incidence of large catches by shortening the duration of tows when the netsonde shows that a large tonnage has been caught. There is good reason to believe that this practice, although carried out for very sensible reasons, will bias the biomass estimate upwards.*

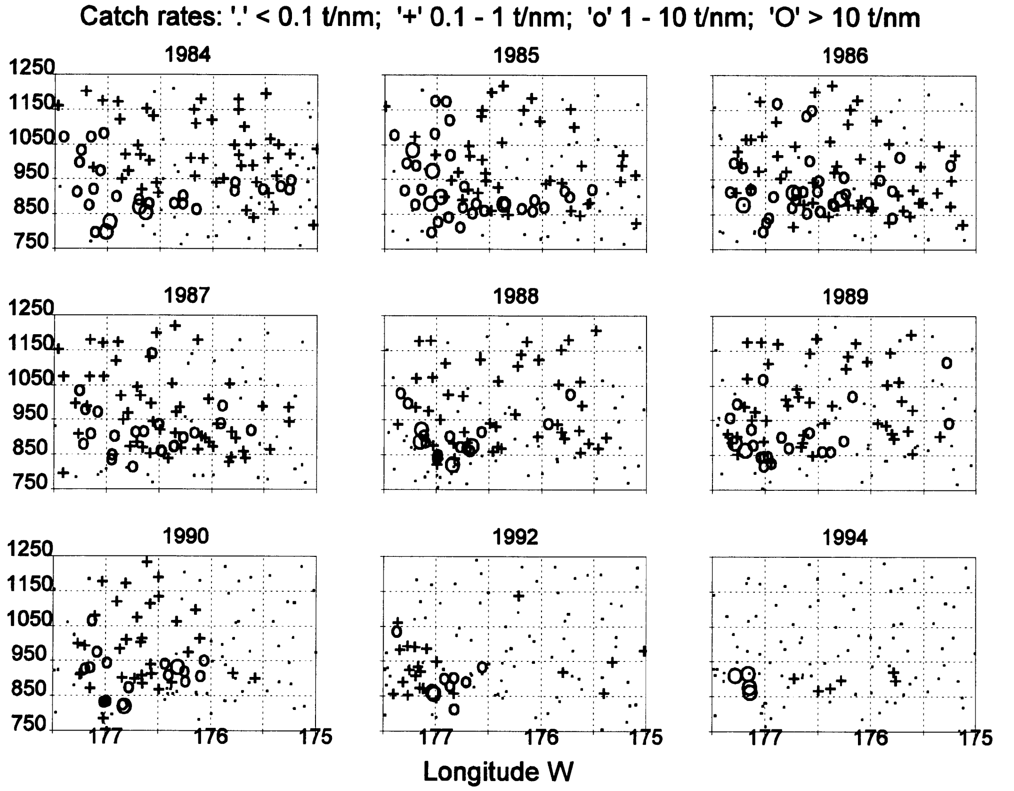

Figure 46: Geographical distribution of high and low catch rates for each survey in the original series. Broken lines are stratum boundaries. (This is figure 12 in Francis 1996)

With regard to the sex-ratio problem, Francis (1996) reached the following conclusions.

> *There is strong evidence that the pattern, within the spawning season, of when and where extreme sex ratios are likely to be caught is now [in the mid 1990s] different from what it was in the 1980s. ... The intensive 1995 survey found that females predominated over more than 90% of the survey area, and that this predominance increased during the survey period (10 July to 4 August). There was a small area (about 110 km^2) [called the 'mixed area'], which appeared to contain a large proportion of the biomass, in which sex ratios were highly variable and did not strongly favour either sex.*
>
> *It seems more likely than not that the population that spawns in the Spawning Box is now [in the mid 1990s] dominated by females. ... However, the changes described above make it difficult to estimate [the current] sex ratio precisely and thus be confident that it has changed.*

5.2 Comparisons with the 2004 data

A comparison of the 2004 data with that from the original series is hindered because of several differences in the design of these surveys (Table 19). Note that these differences were deliberate. When the 2004 survey was designed it was decided, after some discussion, that there would be no attempt to make the trawl portion of this survey a continuation of the original series. Changes were made from the earlier design when it was felt that these would improve results from the new combined acoustic-trawl survey.

Table 19: Comparison of some aspects of the 2004 survey with those from the original series (1984–94).

	Original series (1984–94)	2004 survey
Dates	9–29 July (typical)	9–14 July
Number of tows	105–122	26
High-density area include?	Yes	No
Net	ORH net	Ratcatcher
Distance towed (target)	3 n. mile	1.5 n. mile

There are two important differences. First, the change of nets is significant because there is reason to believe that the catchability of orange roughy is much higher with the ratcatcher. During the 2005 acoustic survey of the northwest Chatham Rise (TAN0509), a series of 14 ratcatcher tows was repeated (between 1 and 9 days later, with a median time difference of 3.5 days) using the ORH net. Catch rates were consistently higher with the ratcatcher, by a factor of about 4.3 on average (Figure 47). The second important difference is that the 2004 survey avoided the high-density area (which was covered by acoustics).

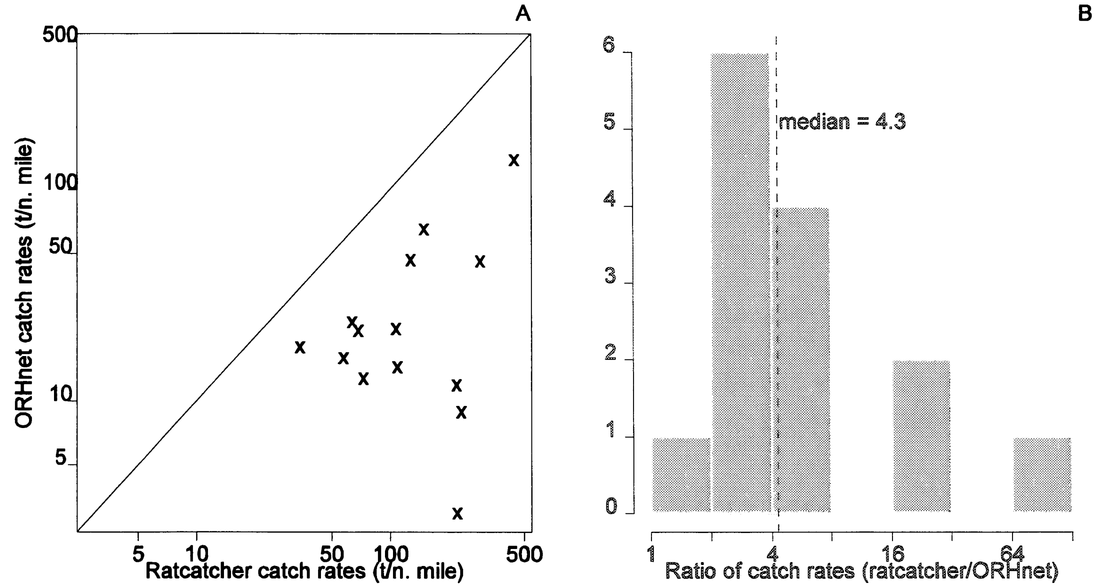

Figure 47: Comparison, from 14 paired tows in the 2005 northwest Chatham Rise survey (TAN0509), of orange roughy catch rates from two nets, ratcatcher and ORH net.

5.2.1 Comparison of catch rates

The distribution of catch rates in 2004 (scaled down by a factor of 4.3 to adjust for the difference in nets) was similar to that in 1994 except that the latter included a few large catch rates from the high-density area not covered by the 2004 survey (Figure 48).

What is not known is what catch rates would have been obtained in the high-density area if the survey had covered this. If they would have been similar to those obtained in earlier years there is every reason to believe that the biomass estimate from a trawl survey covering the whole Spawning Box in 2004 would have had a high c.v., like those in 1992 and 1994.

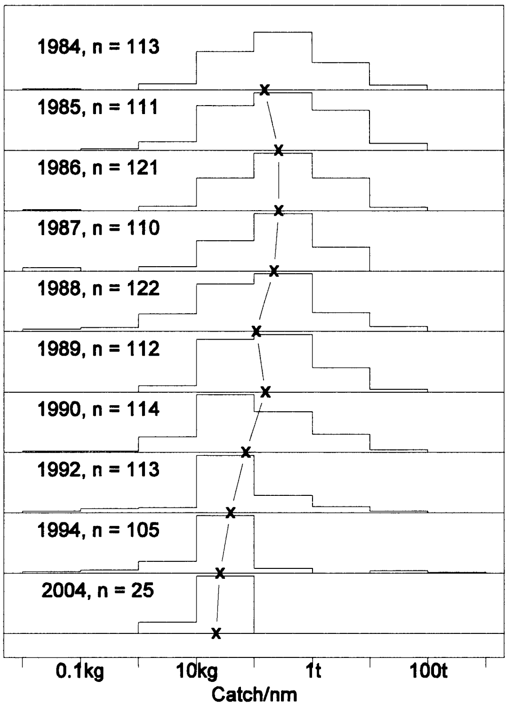

Figure 48: Catch-rate histograms (on a logarithmic scale) for the original trawl-survey series (1984–94) and the 2004 survey (Spawning Box tows only). 'x' marks the median catch rate in each survey. Catch rates for the 2004 survey have been divided by 4.3 to correct for net differences and one zero catch has been omitted.

5.2.2 Comparison of sex ratios

In the intensive 1995 survey, sex ratios showed a complex pattern, varying with location, date, and (to a lesser extent) catch rate. The very limited data from the 2004 survey show no clear change from this pattern (Figure 49).

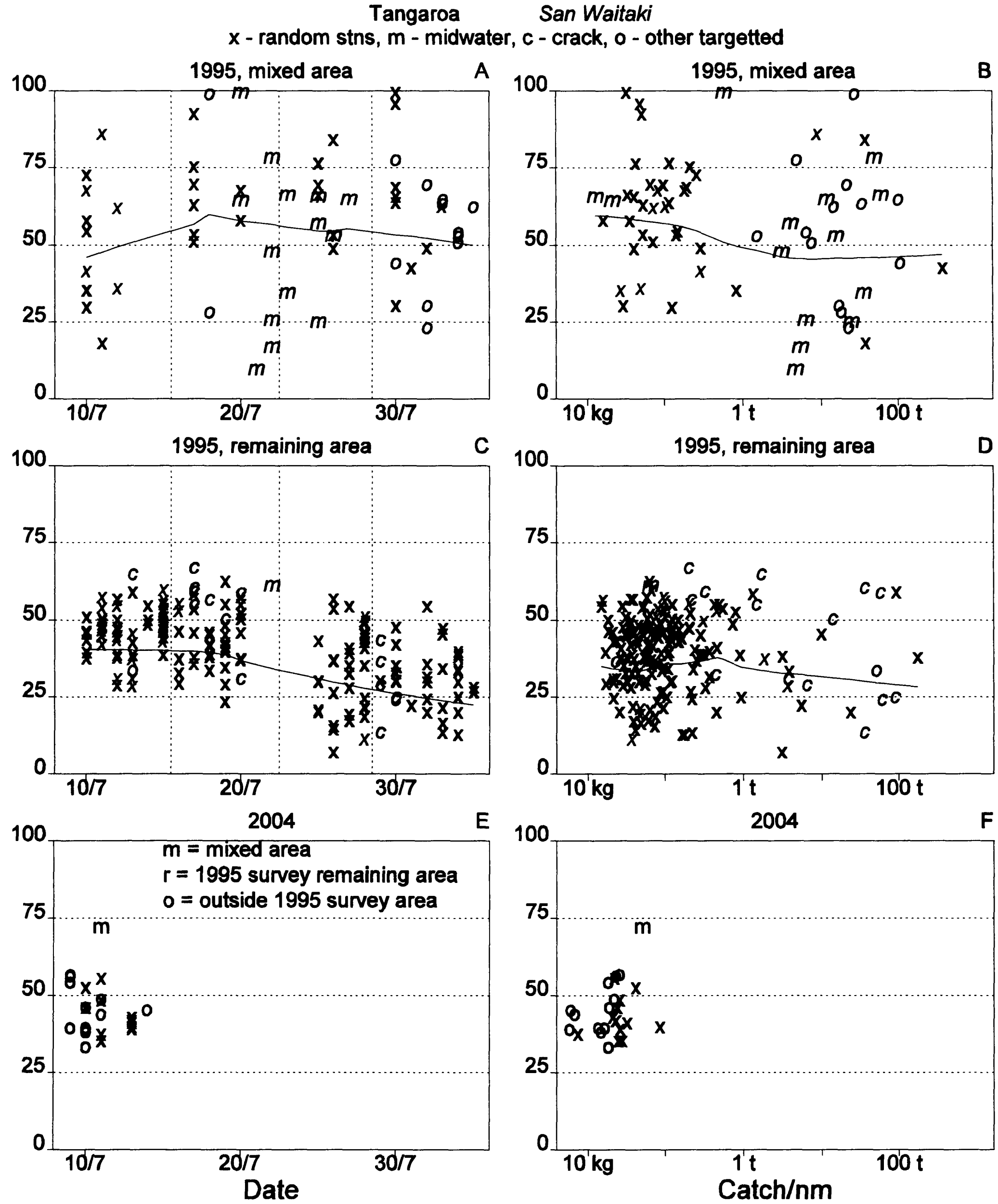

Figure 49: Comparison of Spawning Box sex-ratio data from 1995 and 2004 surveys: plots of percent male against date (left panels) and catch/n. mile (right panels; 2004 catch rates divided by 4.3 to correct for net differences). Plotting symbols are coded by vessel and type; tows with ≤ 20 fish are excluded; a lowess line has been fitted to the data in some panels to guide the eye. Vertical lines in panels A and C separate the four snapshots carried out in 1995. Panels A–D are as in figure 23 of Francis (1996), except that percent male is calculated by number, rather than by weight.

5.3 Conclusion and recommendation on surveys

The data from the 2004 survey are very limited, but provide no evidence to modify the conclusions of Francis (1996). Thus I recommend that we do not carry out another Spawning Box trawl survey to continue the series which ended in 1994.

6. ACKNOWLEDGMENTS

I am grateful to Allan Hicks for his help in sorting out the age data, and to Matt Dunn and Andy McKenzie for providing various stock assessment-related material for Section 4.

7. REFERENCES

Annala, J.H.; Sullivan, K.J.; Smith, N.W.M.; Griffiths, M.H.; Todd, P.R.; Mace, P.M.; Connell, A.M. (2004). Report from the Fishery Assessment Plenary, April 2004: stock assessments and yield estimates. 690 p. (Unpublished report held in NIWA library, Wellington.)

Clark, M.R.; Francis, R.I.C.C. (1990). Revised assessment of the Challenger Plateau (QMA 7A) orange roughy fishery for the 1989-90 fishing year. New Zealand Fisheries Assessment Research Document 90/1. 17 p. (Unpublished report held in NIWA library, Wellington.)

Doonan, I.J. (1994). Life history parameters of orange roughy: estimates for 1994. New Zealand Fisheries Assessment Research Document 94/19. 13 p. (Unpublished report held in NIWA library, Wellington.)

Doonan, I.J.; Tracey, D.M.; Grimes, P.J. (2004). Relationship between macroscopic staging and microscopic observations of oocyte progression in orange roughy during and after the mid-winter spawning period, Northwest Hills, Chatham Rise, July 2002. *New Zealand Fisheries Assessment Report 2004/6.* 28 p.

Dunn, M.R. (2005a). CPUE analysis and assessment of the Mid-East Coast orange roughy stock (ORH 2A South, 2B, 3A) to the end of the 2002-03 fishing year. *New Zealand Fisheries Assessment Report 2005/18.* 35 p.

Dunn, M.R. (2005b). Results from the NIWA wide area (background) trawl survey of the North_East Chatham Rise 2004. WG-DW-04/86. 10 p. (Unpublished report held by Ministry of Fisheries, Wellington.)

Dunn, M.R. (2006). Assessment runs for the Andes 2/06 and Eastern hills 2/06. WG-DW-06/38. 5 p. (Unpublished report held by Ministry of Fisheries, Wellington.)

Francis, R.I.C.C. (1996). Orange roughy sex ratios and catchrate distributions in the Chatham Rise Spawning Box. New Zealand Fisheries Assessment Research Document 96/13. 28 p. (Unpublished report held in NIWA library, Wellington.)

Francis, R.I.C.C.; Horn, P.L. (1997). Transition zone in otoliths of orange roughy (*Hoplostethus atlanticus*) and its relationship to the onset of maturity. *Marine Biology 129*: 681–687.

Francis, R.I.C.C.; Robertson, D.A. (1990). Assessment of the Chatham Rise (QMA 3B) orange roughy fishery for the 1989/90 and 1990/91 fishing years. New Zealand Fisheries Assessment Research Document 90/3. 27 p. (Unpublished report held in NIWA library, Wellington.)

Grimes, P.J. (1994). Trawl survey of orange roughy between Cape Runaway and Banks Peninsula, March-April 1992 (TAN9203). New Zealand Fisheries Data Report 42. 13 p.

Grimes, P.J. (1996a). Trawl survey of orange roughy between Cape Runaway and Banks Peninsula, March-April 1993 (TAN9303). New Zealand Fisheries Data Report 76. 18 p.

Grimes, P.J. (1996b). Trawl survey of orange roughy between Cape Runaway and Banks Peninsula, March-April 1994 (TAN9403). New Zealand Fisheries Data Report 82. 12 p.

Grimes, P.J.; Dunn, M.R.; Tracey, D.M. (2005). Histological preparations of female orange roughy ovaries. Final Research Report for Ministry of Fisheries Research Project ORH2004-02 Objective 6. 20 p. (Unpublished report held by Ministry of Fisheries, Wellington.)

Hicks, A. (2005a). Between-reader and between-lab ageing errors for orange rough otoliths aged at NIWA and CAF. WG-DW-05/23(revised). 18 p. (Unpublished report held by Ministry of Fisheries, Wellington.)

Hicks, A. (2005b). Revised proportions at age estimates for NECR orange roughy. WG-DW-05/74. 9 p. (Unpublished report held by Ministry of Fisheries, Wellington.)

Hilborn, R.; Starr, P.J.; Ernst, B. (1999). Stock assessment of the northeast Chatham Rise orange roughy. WG-DW-99/57. 45 p. (Unpublished report held by Ministry of Fisheries, Wellington.)

Horn, P.L.; Tracey, D.M.; Clark, M.R. (1998). Between-area differences in age and length at first maturity of the orange roughy *Hoplostethus atlanticus. Marine Biology 132*: 187–194.

Johnson, N.L.; Kotz, S.; Kemp, A.W. (1992). Univariate discrete distributions. John Wiley & Sons, New York. 565 p.

Manly, B.F.J. (1997). Randomization, bootstrap and Monte Carlo methods in biology. 2nd edition. Chapman & Hall, London.

McKenzie, A. (2005). Standardised CPUE analysis and stock assessment of the northwest Chatham Rise orange roughy stock (part of ORH 3B). Final Research Report for Ministry of Fisheries Research Project ORH200302 Objectives 1,2,4. 42 p. (Unpublished report held by Ministry of Fisheries, Wellington.)

McKenzie, A. (2006). Additional model runs for the northwest Chatham Rise orange roughy stock assessment. WG-DW-06/41. 20 p. (Unpublished report held by Ministry of Fisheries, Wellington.)

Smith, A.D.M.; Punt, A.E.; Wayte, S.E.; Starr, P.J.; Francis, R.I.C.C.; Stokes, T.K.; Hilborn, R.; Langley, A. (2002). Stock assessment of the northeast Chatham Rise orange roughy for 2001. *New Zealand Fisheries Assessment Report 2002/25*. 30 p.

Sullivan, K.J.; Mace, P.M.; Smith, N.W.M.; Griffiths, M.H.; Todd, P.R.; Livingston, M.E.; Harley, S.J.; Key, J.M.; Connell, A.M. (2005). Report from the Fishery Assessment Plenary, May 2005: stock assessments and yield estimates. 792 p. (Unpublished report held in NIWA library, Wellington.)

Tracey, D.M.; Fenaughty, J.M. (1997). Distribution and relative abundance of orange roughy on the Chatham Rise, May-July 1994. *New Zealand Fisheries Technical Report 44*. 43 p.

Tracey, D.M.; Francis, R.I.C.C.; George, K.; Horn, P.L.; Hart, A.; Marriot, P. (2004). Age composition of orange roughy in the Northeast Chatham Rise spawning box from Otago Buccaneer 1984 and Cordella 1990 samples. Final Research Report for Ministry of Fisheries Research Project MOF2003/03J (Unpublished report held by Ministry of Fisheries, Wellington.)

Vitale, F.; Svedäng, H.; Cardinale, M. (2006). Histological analysis invalidates macroscopically determined maturity ogives of Kattegat cod (*Gadus morhua*) and suggests new proxies for estimating maturity status of individual fish. *ICES Journal of Marine Science 63*: 485–492.

Appendix 1: Estimating the relative difference between two age-length keys

This appendix describes a simple method for quantifying the difference between two ALKs (age-length keys) as a single multiplicative factor. An ALK describes the distribution of ages within each of a number of cells, with each cell being defined by a pair of subscripts: l indexing length classes, and s indexing sex. For the ls cell in the ALK for the jth data set, suppose we have ages $\{A_{ijls}$: $i = 1,\ldots,n_{jls}\}$. The distribution of ages within a cell tends to be skewed, so we will work in log space, defining $a_{ijls} = \log(A_{ijls})$ and letting $a_{\bullet jls} = \text{mean}_i\left(a_{ijls}\right)$. If the variance of log age in the ls cell is the same in the two data sets we can estimate it by

$$\hat{\sigma}_{ls}^2 = \left[\frac{\sum_{i,j}\left(a_{ijls} - a_{\bullet jls}\right)^2}{n_{1ls} + n_{2ls} - 2}\right]$$

and the difference between the two means, $a_{\bullet 2ls} - a_{\bullet 1ls}$, will have variance $\hat{\sigma}_{ls}^2\left[\left(1/n_{1ls}\right)+\left(1/n_{2ls}\right)\right]$. Thus, a natural statistic for measuring the difference between the mean log ages in the ls cell is given by

$$S_{ls} = \frac{a_{\bullet 2ls} - a_{\bullet 1ls}}{\left[\left(1/n_{1ls}\right)+\left(1/n_{2ls}\right)\right]^{0.5}\hat{\sigma}_{ls}}$$

and $S = \sum_{ls} S_{ls}^2$ measures the overall difference between the two ALKs. (Note that if the age distribution within each cell is lognormal, S_{ls} is approximately distributed as a standard normal, and S has a χ^2 distribution).

We assume that the difference between the two ALKs is simply proportional. Specifically, if the typical age of a fish of a given length in data set 1 is A then we assume that the typical age for fish of the same length in data set 2 is mA. Our aim is to estimate m and we can do this using a simple search procedure. For any trial value of m, we adjust all age estimates in data set 2 by multiplying by m, and then calculate the statistic S measuring the difference between the two ALKs after this adjustment. The best estimate of m is that which minimises S.

A measure of how confident we are that there really is a difference between the ALKs (i.e., that m differs from 1) is given by the decrease in S as m changes from 1 to the estimated value. Roughly speaking, the difference is significant if the change is more than 2.

Appendix 2: Bootstrap confidence intervals

There are many different ways of calculating bootstrap confidence intervals for a statistic. The method used for the intervals plotted in Figure 9 is one of the simplest and is due to Efron (see p. 41 of Manly 1997). Three alternative methods (labelled Hall, bias-corrected, and accelerated bias-corrected by Manly (1997)) were tried and found to produce very similar results.

Appendix 3: Calculating Table 6

In calculating the values in Table 6 it was assumed that the width of a 95% confidence interval for the calibration ratio would be inversely proportional to the square root of the sample size (as standard errors are). The confidence intervals in Figure 9 were used to scale this relationship. Thus, the expected confidence-interval width for a sample size of *m* was calculated as $\text{mean}_i(w_i n_i^{0.5}) m^{-0.5}$, where, for the *i*th comparison in Figure 9, w_i is the width of the confidence interval and n_i is the sample size. For the first row in Table 6 the mean was calculated using just the first (upper) five comparisons in Figure 9 (all those based on just one reader); the second row of Table 6 was based on the next three comparisons (all those using two readers from the same institution).

Appendix 4: Bootstrap simulation of observer LFs

This appendix describes the bootstrapping procedure that was used to generate annual LFs from observer data. The data for each year were restricted as follows:

– an observed tow was acceptable only if at least 20 fish from each sex were measured
– a trip was acceptable only if at least 2 acceptable tows were sampled from it.

The following procedure was used to generate a set of 300 replicate LFs for one year.

1. Select one trip at random from that year.
2. Select one tow at random from the selected trip
3. Select a sample of fish from the selected tow, where the sample size is the same as in the original sample, and sampling is random with replacement.
4. Construct an LF (proportions at length by sex), with a 50:50 sex ratio (so the sum of proportions for each sex is equal to 0.5)
5. Use the length-weight relationship (W(kg) = $9.21 \times 10^{-05} \times L(\text{cm})^{2.71}$) to calculate the mean weight of fish in this sample.
6. Scale the LF up to the catch by multiplying it by the ratio of the weight of the total catch from the sampled tow to this mean weight.
7. Repeat steps 2 to 6 as many times as there were tows sampled originally from the selected trip.
8. Repeat steps 1 to 7 as many times as there were trips sampled originally in that year.
9. Sum all LFs from step 6, and then convert the LF values from numbers to proportions. The result represents one estimate of the LF from the total catch in the Box in that year.
10. Repeat steps 1 to 9 to generate 300 LFs.

Appendix 5: Adjusting LFs for various statistics

This appendix describes the procedures that were used in Section 3.1.2 to adjust bootstrap LFs for various LF statistics (mean, s.d., skewness, and quantiles) so as to determine how much information was contained in those statistics. For each statistic I give a brief description of the adjustment procedure and then the Splus function used for this adjustment. In some cases the adjustment required additional length bins at one or both ends of the LF.

Denote the proportion at length L_i in the original and adjusted LFs to be p_i and p_i', where, for simplicity, the L_i are assumed to be consecutive integers.

A5.1 Adjusting for the mean

This adjustment changes the mean of the LF from μ_{old} to μ_{new}, while maintaining the shape of the LF. This is straightforward if $t = \mu_{\text{new}} - \mu_{\text{old}}$ is an integer: we simply set $p_i' = p_{i-t}$. When t is not an integer, we make the adjustment in two steps. First we construct an LF which has the same proportions but adds t to the length for each bin. This LF has exactly the right mean (μ_{new}) but the lengths are now non-integers. We then rebin this LF into the original length bins. To see how this works, suppose t = 1.7 cm and that the proportion of fish in the original 30 cm bin (which includes fish with lengths $30 \leq L < 31$) was 0.1. This proportion of 0.1 is first shifted to apply to a new bin covering lengths $31.7 \leq L < 32.7$, and then, when we rebin it, is broken into 0.03 for the 31 cm bin and 0.07 for the 32 cm bin.

For the purposes of measuring information content it doesn't actually matter what value we use for μ_{new} as long as it is the same for all bootstrap replicates. That is, the total variance of the set of adjusted set of LFs is independent of μ_{new}. I chose to use an integer value (35 cm) throughout because this made the subsequent adjustment for skewness simpler.

The Splus function to adjust the means follows.

```
"Std.mn"<-
function(mat, new.mu)
{
# Transforms all LFs in a matrix to have the same specified mean,
# adding new bins to the LFs if necessary
#
# mat - matrix in which the columns are LFs (and so sum to 1) and the
#       row names define the centres of the length bins
# new.mu - desired mean length of new LFs
#
        nLF <- ncol(mat)
        Ls <- as.numeric(dimnames(mat)[[1]])
        old.mus <- apply(mat * Ls, 2, sum)
        shifts <- new.mu - old.mus
        shifts.whole <- floor(shifts)
        shifts.frac <- shifts - shifts.whole
        if(max(shifts.whole) > 0) {
                nn <- max(shifts.whole)
                new.Ls <- (Ls[1] - nn):(Ls[1] - 1)
                mat <- rbind(matrix(0, nn, nLF, dimnames = list(paste(new.Ls),
                        NULL)), mat)
                Ls <- c(new.Ls, Ls)
        }
        if(min(shifts.whole) < 0) {
                nn <- abs(min(shifts.whole))
                new.Ls <- (Ls[len(Ls)] + 1):(Ls[len(Ls)] + nn)
                mat <- rbind(mat, matrix(0, nn, nLF, dimnames = list(paste(
                        new.Ls), NULL)))
                Ls <- c(Ls, new.Ls)
        }
        temp <- function(pp, wh)
        {
                n <- len(pp)
                if(wh == 0)
                        pp
                else if(wh > 0)
                        c(rep(0, wh), pp[1:(n - wh)])
                else c(pp[(abs(wh) + 1):n], rep(0, abs(wh)))
        }
        for(i in 1:nLF) {
                new.ps <- (1 - shifts.frac[i]) * temp(mat[, i], shifts.whole[i]
                        ) + shifts.frac[i] * temp(mat[, i], shifts.whole[i] + 1
                        )
                mat[, i] <- new.ps/sum(new.ps)
        }
        mat
}
```

A5.2 Adjusting for the s.d.

This adjustment changes the s.d. of the LF from σ_{old} to σ_{new}, while maintaining the mean of the LF and its shape. As with the adjustment for the mean, this is done in two steps. First we construct an LF which has the same proportions as the original LF but length bins with non-integer widths, $\sigma_{\text{new}}/\sigma_{\text{old}}$. This LF has the same mean as before and exactly the right s.d. (σ_{new}). As before, we rebin this LF into the original length bins. The value of σ_{new} used was that for the LF obtained by averaging over all the bootstrap LFs.

This algorithm is not perfect because the rebinning slightly changes the s.d. of the LF. However, when the algorithm is applied several times the resultant s.d.s become closer and closer to σ_{new} (I found that just two applications were adequate). The fact that the resulting s.d.s weren't exactly equal to σ_{new} means only that our estimates of percent variance explained are slightly too small.

The Splus function to adjust the s.d.s follows.

```
"Std.sd"<-
function(mat, new.sd)
{
# Transforms all LFs in a matrix to have a specified s.d. without changing
# their means.
#
# Calculation is in two steps:
# 1. Create histogram with existing probabilities and new length bins
#    (whose bounds are newLslo & newLshi) which has required mean and s.d.
# 2. Rebin this histogram into old bins
#
# mat - matrix in which the columns are LFs (and so sum to 1) and the
#       row names define the centres of the length bins
# new.sd - desired s.d. of transformed LF
#
        nLF <- ncol(mat)
        Ls <- as.numeric(dimnames(mat)[[1]])
        nbin <- len(Ls)
        old.bwidth <- Ls[2] - Ls[1]
        oldLslo <- Ls - 0.5 * old.bwidth
        oldLshi <- Ls + 0.5 * old.bwidth
        for(k in 1:nLF) {
                mu <- sum(mat[, k] * Ls)
                old.sd <- sqrt(sum(mat[, k] * (Ls - mu)^2))
                new.bwidth <- (new.sd/old.sd)
                newLslo <- mu + new.bwidth * (Ls - mu - 0.5 * old.bwidth)
                newLshi <- mu + new.bwidth * (Ls - mu + 0.5 * old.bwidth)
                loindx <- round(((oldLslo - mu)/new.bwidth) + ((mu - Ls[1])/
                        old.bwidth) + 1)        # oldLslo[i] is in loindx[i]th new bin
                hiindx <- round(((oldLshi - mu)/new.bwidth) + ((mu - Ls[1])/
                        old.bwidth) + 1)        # oldLshi[i] is in hiindx[i]th new bin
                ps <- mat[, k]
                new.ps <- rep(0, nbin)
                for(i in 1:nbin) {
                        if(loindx[i] == hiindx[i]) {
                                if(loindx[i] > 0 & loindx[i] <= nbin)
                                  new.ps[i] <- (ps[loindx[i]] * old.bwidth)/
                                    new.bwidth
                                else new.ps[i] <- 0
                        }
                        else {
                                if(loindx[i] > 0 & loindx[i] <= nbin)
                                  new.ps[i] <- (ps[loindx[i]] * (newLshi[loindx[
                                    i]] - oldLslo[i]))/new.bwidth
                                if(hiindx[i] > 0 & hiindx[i] <= nbin)
                                  new.ps[i] <- new.ps[i] + (ps[hiindx[i]] * (
                                    oldLshi[i] - newLslo[hiindx[i]]))/
                                    new.bwidth
                                no.complete.bins <- hiindx[i] - loindx[i] - 1
                                if(no.complete.bins > 0) {
                                  for(j in 1:no.complete.bins) {
                                    indx <- loindx[i] + j
                                    if(indx > 0 & indx <= nbin)
                                      new.ps[i] <- new.ps[i] + ps[indx]
                                  }
                                }
                        }
                }
                mat[, k] <- new.ps/sum(new.ps)
        }
        mat
}
```

A5.3 Adjusting for skewness

This adjustment changes the skewness of an LF without altering its mean or s.d. Initially, I intended to make the adjustment so that the skewness became the same as that for the LF obtained by averaging over all the bootstrap LFs. However, I couldn't find an algorithm which achieved this. What was very simple though was to remove the skew (without changing the mean or s.d.). This algorithm required that the mean of the LF was an integer, and so was only ever applied after the adjustment for the mean. If the mean of the LF was L_m then the adjustment simply averages pairs of proportions that are equi-distant from this mean, so $p'_i = p'_{2m-i} = 0.5\left(p_i + p_{2m-i}\right)$.

The Splus function to remove skew follows.

```
"Rm.skew"<-
function(mat, mu)
{
# Removes any skew from all LFs in a matrix without changing their means
# or sds, adding new bins to the LFs if necessary
#
# All LFs are assumed to have the same mean, mu, which should be the
# mid-pt of one of the length bins.  In the output LFs this will be the
# central bin.
#
# mat - matrix in which the columns are LFs (and so sum to 1) and the
#       row names define the centres of the length bins
#
        nLF <- ncol(mat)
        Ls <- as.numeric(dimnames(mat)[[1]])
        indx <- match(mu, Ls)
        nbin <- len(Ls)
        nabove <- nbin - indx
        nbelow <- indx - 1
        if(nbelow < nabove) {
                nn <- nabove - nbelow
                new.Ls <- (Ls[1] - nn):(Ls[1] - 1)
                mat <- rbind(matrix(0, nn, nLF, dimnames = list(paste(new.Ls),
                     NULL)), mat)
                Ls <- c(new.Ls, Ls)
        }
        else if(nabove < nbelow) {
                nn <- nbelow - nabove
                new.Ls <- (Ls[len(Ls)] + 1):(Ls[len(Ls)] + nn)
                mat <- rbind(mat, matrix(0, nn, nLF, dimnames = list(paste(
                     new.Ls), NULL)))
                Ls <- c(Ls, new.Ls)
        }
        nbin <- len(Ls)
        revmat <- matrix(mat[rev(1:nbin),  ], ncol = nLF, dimnames = dimnames(
                mat))
        mat <- 0.5 * (mat + revmat)
        mat
}
```

A5.4 Adjusting for quantiles

This adjustment modifies an LF to make its quantiles for specified probabilities, π_1, π_2, ..., π_m, take specified values, l_1, l_2, ..., l_m. To do this, the LF is first broken into m+1 pieces: the piece to the left of l_1, the piece between l_1 and l_2, ..., and the piece to the right of l_m. Typically, each piece will consist of a number of complete length bins and then a partial bin at one or both ends. For example, if l_{j-1} = 30.3 and l_j = 32.4, then the jth piece will consist of 0.7 of the 30 cm bin, plus the whole 31 cm bin, plus 0.4 of the 32 cm bin, and the proportion of the LF in this piece will be $q_j = 0.7p_i + p_{i+1} + 0.4p_{i+2}$, where L_i = 30 cm. Since we want the proportion in this piece to be π_j, it makes sense to multiply the proportions for each bin and partial bin in this piece by π_j/q_j. This almost works. The quantiles of the resulting LF aren't quite right because the two parts of each partial bin are multiplied by different amounts. However, repeated application of the algorithm gradually moves the quantiles towards their desired values.

The Splus function to adjust for quantiles follows. In this, the algorithm is repeated as many times as it takes to ensure that the actual quantiles are within a specified tolerance of the desired value. The tolerance I used in making Table 22 was 0.1 cm.

```
"adj.quantile"<-
function(ps, Ls, probs, quants, tol, max.iter = 20)
{
# Adjusts an LF so that its quantiles are approx. equal to specified values
#
# The adjustment is done by proportionately scaling up or down the proportions
# for bins that lie between the quantiles. Because quantiles occur part-way
# through a bin the adjustment algorithm doesn't produce exactly the right
# quantiles.  However, it seems that iterating the algorithm gradually moves
# the quantiles towards the right values.
#
# ps - proportions for LF (1 per bin)
# Ls - lengths for LF (1 per bin)
# probs - probabilities for which quantiles are set
# quants - the quantiles to be adjusted to
# tol, max.iter - the algorithm is repeated until either the maximum absolute
#       difference between the actual and specified quantiles is less than
#       tol or the number of iterations exceeds max.iter
#
        nquant <- len(quants)
        dprobs <- diff(c(0, probs, 1))
        whol <- floor(quants)
        fract <- quants - whol
        pieces <- list()
        niter <- 1
        maxdiff <- 2 * tol
        while(niter <= max.iter & maxdiff > tol) {
# Split LF into nquant+1 pieces, with overlap at the joins
                if(niter == 1) {
                        for(i in 1:(nquant + 1)) {
                                sel <- if(i == 1) Ls <= whol[i] else if(i == (
                                  nquant + 1))
                                  Ls >= whol[i - 1]
                                else Ls >= whol[i - 1] & Ls <= whol[i]
                                selLs <- Ls[sel]
                                selps <- ps[sel]
                                nsel <- len(selps)
                                pieces[[i]] <- list(ps = selps, Ls = selLs)
                        }
                }
                else {
                        for(i in 1:(nquant + 1))
                                pieces[[i]]$ps <- adjps[is.in(Ls, pieces[[i]]$
                                  Ls)]
                }
                for(ll in sort(unique(whol))) {
                        indx <- (1:(nquant + 1))[unlapply(pieces, function(x, y
                              )
                        is.in(y, x$Ls), ll)]
                        fact <- diff(c(0, fract[whol == ll], 1))
                        for(j in indx) {
                                sel <- match(ll, pieces[[j]]$Ls)
                                pieces[[j]]$ps[sel] <- pieces[[j]]$ps[sel] *
                                  fact[match(j, indx)]
                        }
                }
# Rescale each piece to match the specified values in probs
                for(i in 1:(nquant + 1))
                        pieces[[i]]$ps <- (pieces[[i]]$ps * dprobs[i])/sum(
                                pieces[[i]]$ps)
        # Recombine nquant + 1 pieces to make adjusted LF
                adjps <- rep(0, len(ps))
                for(i in 1:(nquant + 1)) {
                        sel <- is.in(Ls, pieces[[i]]$Ls)
                        adjps[sel] <- adjps[sel] + pieces[[i]]$ps
                }
# Calculate maximum absolute diff between the actual and specified quantiles
                maxdiff <- max(abs(quants - quantile.cts(adjps, Ls, probs)))
                niter <- niter + 1
        }
        if(niter > max.iter)
                cat("Max. iterations reached, maxdiff = ", signif(maxdiff, 3),
                        "\n")
        adjps
}
```

Appendix 6: Calculation of Neff, obs

This appendix describes the calculation (referred to in Recommendation 1 of Section 3.3) of the observation-error effective sample size, $N_{\text{eff, obs}}$, from a set of bootstrap LFs.

Let p_{ik} be the proportion of fish of length L_i in the kth bootstrap LF, for k = 1,..., N. Then the mean of the kth LF is $\bar{L}_k = \sum_i p_{ik} L_i$, and the s.d. of these means is $s_{\bar{L}} = \left[\sum_k (\bar{L}_k - \bar{L})^2 / (N-1)\right]^{0.5}$, where $\bar{L} = \left(\sum_k \bar{L}_k\right)/N$ is the overall mean.

The average of all the bootstrap LFs is defined by $p_i = \left(\sum_k p_{ik}\right)/N$. This LF has mean $\mu = \sum_i p_i L_i$ and s.d. $\sigma = \left[\sum_i p_i (L_i - \mu)^2\right]^{0.5}$. The effective sample size must satisfy $s_{\bar{L}} = \sigma / N^{0.5}_{\text{eff, obs}}$, therefore $N_{\text{eff, obs}} = (\sigma / s_{\bar{L}})^2$.