COMBINATION OF RECRUIT INDICES BY WEIGHTED AVERAGES USING RCT3: A USER'S GUIDE

J G Shepherd and C D Darby
Ministry of Agriculture, Fisheries & Food
Lowestoft
Suffolk NR33 OHT

A) Method and Data file

The method is explained by Shepherd (1991). Data must be supplied on a file in one of the formats illustrated in Table 1, the 'old' format or Table 2, the 'new' format. The files may be created in a text editor or exported as comma separated files from EXCEL, LOTUS123 or Super Calc 5\4\3. Titles and data are separated by spaces or commas.

Line 1. The title may be up to 80 characters long.

Line 2. The number of surveys and number of yearclasses must be specified, together with the number of the column in which the VPA estimates for each yearclass are located (it is conventional to use column 2).

If the file is in the 'new' format

Line 3. The column name 'YEAR' followed by 'VPA', if the vpa results are in column 2, and the survey identifying codes, each enclosed in 'single quotes'. The program will use only the first 6 characters of each code. If the VPA data are not in column 2 the 'VPA' column heading should be placed in the appropriate position in the survey sequence.

Line 4. (New format. Line 3. Old format.)

The survey data. These are entered in free format with spaces or commas as delimiters. Column 1 contains the survey date, column 2 is usually the VPA results. These are followed by the survey data following the sequence of the identifying codes. Data are normally entered in their usual

form (the program can accept logged rather than arithmetic quantities, but this is not recommended). Missing values are identified by -11.

If using the 'old' format the lines following the data contain the identifying codes for the surveys. These must be appended, one per line and left-justified delimiting quotes are not required.

B) Questions and answers

The program is executed by typing RCT3, and answering the questions. Upper or lower case characters may be used, and only one character is necessary for yes/no responses (more will be ignored). All responses must be terminated by a carriage return (or Enter, or whatever). The responses to the questions are mostly obvious, but the following additional notes may be useful.

1. Output medium : Output can be directed to either a disk file (D) or directly to the printer (P). The output file name can be any combination of 40 characters including drive:\path\filename. The program will check for the existence of an output file with the given name and require confirmation if it is to overwrite the old file.

2. Data file name : 40 characters with no special extension type assumed. The full filename is required (including drive:\path\filename if the file is not in default path).

3. Input Format : The program will print the title of the data file and ask for the input format. The default option is the 'old' format.

4. Analysis parameters : The program offers the user the choice of two methods for entering the answers to questions 5 - 10.

a) A configuration file which contains a title followed by the answers to questions 5 - 10 on consecutive lines. These consist of either a numeric value or 'Y'/'N'. (Note: The answer to question 8. must be provided regardless of the answer to question 7.) The choice allows customised configuration files to be created for individual stocks or species. An example is provided in Table 3.

b) A series of default answers for questions 5 - 10, selected automatically if there is no configuration file or if there is an error in the configuration file format.

The users configuration file entries or the default selections are printed to the screen. If the list conforms to the required answers, typing 'Y' will skip to question 11. Typing 'N' will allow the selections to be altered. Each question has a default option associated with it which is indicated in the text and is selected by typing Return/Enter as a response. ***Questions which require a numeric value will default to zero if Return/Enter is typed.***

5. Regression type : C (for calibration) is the default (following the recommendations of the Methods WG 198). P (for predictive) and F (for functional) are also available but not recommended.

6(a) Tapered time weighting : this permits the progressive down-weighting of old data, which is usually a good idea. The shape of the taper may be specified. Typing 0 will give uniform weight up to a sharp (knife-edge) cut-off, and may be used to exclude old data but weight all recent data equally (there is thus no need to delete data from the file to achieve this). 1 gives a linear taper, which seems

sensible. 3 gives the Cleveland tri-cubic taper which is popular in some circles. 2 is included for completeness (these tapers are all members of the same family) and gives a bisquare weighting which will make users of robust regression feel at home. In general the choice of taper among options 1, 2 or 3 doesn't seem to make much difference.

Answering 'N' to this question will cause all data on the file to be used and weighted equally: this, and use of uniform (option 0) weighting, can lead to substantially different results if the taper range does not equal the number of years in the data set.

6(b) Taper range. : If one chooses to use a taper, the range (in years) must be specified. This need not correspond to the actual length of the data series, but should correspond to the time over which data is thought to become useless. Data up to half the taper length old gets substantially more weight than that in the older half - so choosing 20 years as suggested means that one is concentrating on the last 10 years data, which seems reasonable.

7. Shrinkage towards the Mean : This enables one to take account of the historic distribution of yearclass strengths, and inject an element of caution about extreme predictions (see report of Methods WG 1987). If one selects this option (as recommended) the mean recruitment is included as an extra estimate, weighted in accordance with the standard deviation about that mean. If there are any good surveys this will have very little effect. If all the indications are imprecise, however, it will lead to more conservative estimates. In particular, any surveys giving estimates whose standard error is larger than s.d. about the mean will get very little weight, which seems sensible!

If this option is rejected, a pure calibration estimate is used (which is perfectly OK given any decent data).

8. Exclusion of surveys with SEs greater than that of mean : This is not an option if one has selected shrinkage towards the mean. Otherwise it permits the exclusion of datasets which appear to be useless. In the long run it would be sensible to delete such datasets from the file, although this should not be done casually, because some indication of high or low yearclasses may be useful and usable, even if imprecise. Shrinkage to the mean is a better choice, excluding only really bad datasets.

9. Minimum log S.E. : This permits one to insert an estimate of the minimum log S.E. (CV as a fraction) one would be prepared to believe of any survey (because of known sampling variability for example). This "under-pinning" of the estimated S.E.s (since it is used to replace any lower estimates when weightings are calculated) prevents (for example) a very high weight being given to a "survey" which just happens to have an excellent correlation with only a few points. The suggested value of 0.2 seems to be a reasonable choice in practice (and better estimates are hard to come by!) The same value is applied to all surveys.

10. Minimum Number of Years for Regression : The absolute minimum is 3 (because the method needs to estimate standard errors about the regressions), but a larger number (4 or 5, say) may be chosen if one wishes to be cautious about new datasets entering the analysis. The high standard error of predictions from regressions based on few points is already taken into account, so it is not essential to use this option (see also comments relating to question 9).

If the default answers were selected at the beginning of the question and answer section the program skips to this point.

11. Prior Weights. : This permits the results of a particular survey to be downweighted when the predictions derived from the individual regressions are combined in a weighted average. Note that this does not affect the weight the survey data are given in the regressions, nor (therefore) the actual estimate from any individual survey. This is not usually necessary or desirable, but may be used to exclude the results of a particular survey completely (by choosing zero as the weight), or to allow for indices which are suspect or very provisional.

Do not use this facility unless you must, because it provides considerable scope for confusion and subjective meddling. If you do select it, you are requested to supply the prior weights for each survey (in the order they occur in the data). These weights should normally lie in the range 0 to 1.

12. First Yearclass for Prediction : The program will carry out a retrospective analysis of all years from the first yearclass in the data set + the minimum number of points required for a regression. This is very instructive for the first few times you examine a particular set of data. Once you know where it's at, you will probably wish to examine the predictions only for the most recent yearclasses.

13. Surveys already logged : This permits you to inform the program if you have included data of which logarithms have already been taken by mistake (or even deliberately). If this is the case the program requests the number of logged surveys and their position in the survey sequence.

TABLE 1. The 'old' data file format.

NORTH SEA COD AS 1-GROUP : POST 1976, 0 & 1 GROUP DATA

5 11 2 (No. of surveys, No. of yearclasses, VPA column no.)

1976	709	-11	-11	36.7	6818	-11
1977	426	1559	-11	12.9	2372	-11
1978	454	1679	-11	9.9	2265	-11
1979	800	1856	-11	16.8	5150	163.8
1980	271	1006	43.2	2.9	1232	46.9
1981	566	7963	176.8	9.2	3234	83
1982	265	254	26.9	3.9	1541	21.8
1983	552	9595	121.5	15.2	6122	121.3
1984	93	45	1.3	.9	419	3.6
1985	-11	798	143.6	17.2	3464	111.2
1986	-11	96	37	9.6	-11	-11

EGFSO

DGFSO

IYFS1

EGFS1

DGFS1

TABLE 2. The 'new' data file format.

7
5. NORTH SEA COD AS 1-GROUP : POST 1976, 0 & 1 GROUP DATA

(7) 11 2 (No.Surveys, No.Yearclasses, VPA column)

'YEAR'	'VPA'	'EGFSO'	'DGFSO'	'IYFS1'	'EGFS1'	'DGFS1'
1976	709	-11	-11	36.7	6818	-11
1977	426	1559	-11	12.9	2372	-11
1978	454	1679	-11	9.9	2265	-11
1979	800	1856	-11	16.8	5150	163.8
1980	271	1006	43.2	2.9	1232	46.9
1981	566	7963	176.8	9.2	3234	83
1982	265	254	26.9	3.9	1541	21.8
1983	552	9595	121.5	15.2	6122	121.3
1984	93	45	1.3	0.9	419	3.6
1985	-11	798	143.6	17.2	3464	111.2
1986	-11	96	37	9.6	-11	-11

Table 3. An example of an RCT3 configuration file which can be customised for individual stocks or species. The notes enlosed in [ ] are ignored by RCT3 and can be used to comment the answers.

```
"North Sea Cod as 1-Group Configuration File"
'C'   [5. Regression type 'C','P','F']
'Y'   [6. Tapered weighting Y/N]
3     [6(a). Taper power INTEGER (usually <3)]
20    [6(b). Taper range INTEGER]
'N'   [7. Shrink to mean Y/N]
'Y'   [8. Exclude surveys with SE greater than mean Y/N]
3     [9. Minimum number of regression points INTEGER]
0.2   [10. Minimum S.E. REAL]
```

Table 4. The results file generated by RCT3 using the data contained in Tables 1 or 2.

Analysis by RCT3 ver3.1 of data from file :

A:TABLE2.CSV

"NORTH SEA COD AS 1-GROUP : POST 1976, 0 & 1 GROUP DATA"

Data for 5 surveys over 11 years : 1976 - 1986

Regression type = C
Tapered time weighting applied
power = 3 over 20 years
Survey weighting not applied
Final estimates shrunk towards mean
Minimum S.E. for any survey taken as .20
Minimum of 3 points used for regression

Forecast/Hindcast variance correction used.

Yearclass = 1979

	I----------Regression----------I				I----------Prediction---------I			
Survey/ Series	Slope	Inter- cept	Std Error	Rsquare	No. Pts	Index Value	Predicted Value	Std Error	WAP Weights
EGFS0									
DGFS0									
IYFS1	.44	4.98	.12	.912	3	2.88	6.25	.315	.210
EGFS1	.45	2.60	.06	.977	3	8.55	6.44	.169	.520
DGFS1									
					VPA Mean =		6.25	.277	.271

Yearclass = 1980

	I----------Regression----------I				I----------Prediction---------I			
Survey/ Series	Slope	Inter- cept	Std Error	Rsquare	No. Pts	Index Value	Predicted Value	Std Error	WAP Weights
EGFS0	4.21	-25.01	.18	.885	3	6.91	4.10	1.727	.021
DGFS0									
IYFS1	.86	3.88	.41	.466	4	1.36	5.05	1.426	.030
EGFS1	.61	1.31	.16	.860	4	7.12	5.68	.434	.328
DGFS1									
					VPA Mean =		6.36	.315	.621

Yearclass = 1981

	I----------Regression----------I				I----------Prediction---------I			
Survey/ Series	Slope	Inter- cept	Std Error	Rsquare	No. Pts	Index Value	Predicted Value	Std Error	WAP Weights
EGFS0	1.82	-7.18	.26	.813	4	8.98	9.17	1.706	.012
DGFS0									
IYFS1	.60	4.67	.27	.781	5	2.32	6.06	.450	.169
EGFS1	.65	.97	.14	.930	5	8.08	6.26	.232	.636
DGFS1									
					VPA Mean =		6.21	.434	.183

Yearclass = 1982

Survey/ Series	Slope	Inter-cept	Std Error	Rsquare	No. Pts	Index Value	Predicted Value	Std Error	WAP Weights
EGFSO	1.01	-1.55	.80	.250	5	5.54	4.04	2.113	.006
DGFSO									
IYFS1	.62	4.65	.28	.718	6	1.59	5.64	.478	.127
EGFS1	.66	.92	.13	.924	6	7.34	5.79	.212	.642
DGFS1	.92	2.09	.21	.935	3	3.13	4.97	.884	.037
						VPA Mean =	6.23	.392	.188

Yearclass = 1983

Survey/ Series	Slope	Inter-cept	Std Error	Rsquare	No. Pts	Index Value	Predicted Value	Std Error	WAP Weights
EGFSO	.55	2.03	.50	.481	6	9.17	7.10	.937	.027
DGFSO	.46	3.97	.13	.955	3	4.81	6.17	.377	.169
IYFS1	.64	4.60	.25	.778	7	2.79	6.39	.376	.171
EGFS1	.74	.29	.15	.914	7	8.72	6.73	.239	.423
DGFS1	.70	3.16	.26	.872	4	4.81	6.53	.543	.082
						VPA Mean =	6.13	.435	.128

Yearclass = 1984

Survey/ Series	Slope	Inter-cept	Std Error	Rsquare	No. Pts	Index Value	Predicted Value	Std Error	WAP Weights
EGFSO	.47	2.57	.47	.475	7	3.83	4.36	1.028	.041
DGFSO	.51	3.80	.13	.938	4	.83	4.22	.594	.123
IYFS1	.64	4.60	.23	.785	8	.64	5.01	.432	.233
EGFS1	.70	.55	.20	.833	8	6.04	4.78	.411	.257
DGFS1	.67	3.26	.23	.863	5	1.53	4.27	.712	.086
						VPA Mean =	6.15	.408	.261

Yearclass = 1985

Survey/ Series	Slope	Inter-cept	Std Error	Rsquare	No. Pts	Index Value	Predicted Value	Std Error	WAP Weights
EGFSO	.43	2.85	.39	.775	8	6.68	5.73	.547	.058
DGFSO	.43	4.13	.12	.981	5	4.97	6.28	.210	.390
IYFS1	.80	4.20	.29	.859	9	2.90	6.51	.404	.106
EGFS1	.78	-.07	.21	.921	9	8.15	6.25	.286	.211
DGFS1	.60	3.57	.19	.957	6	4.72	6.39	.296	.197
						VPA Mean =	5.96	.669	.039

Yearclass = 1986

Survey/ Series	Slope	Inter-cept	Std Error	Rsquare	No. Pts	Index Value	Predicted Value	Std Error	WAP Weights
	I----------Regression----------I					I----------Prediction---------I			
EGFSO	.43	2.87	.39	.778	8	4.57	4.82	.618	.072
DGFSO	.43	4.13	.12	.982	5	3.64	5.70	.198	.690
IYFS1	.80	4.19	.29	.862	9	2.36	6.09	.395	.177
EGFS1									
DGFS1									
					VPA Mean =		5.95	.675	.061

Year Class	Weighted Average Prediction	Log WAP	Int Std Error	Ext Std Error	Var Ratio	VPA	Log VPA
1979	572	6.35	.14	.07	.24	801	6.69
1980	423	6.05	.25	.27	1.15	271	5.61
1981	518	6.25	.19	.19	1.04	566	6.34
1982	333	5.81	.17	.14	.71	265	5.58
1983	661	6.50	.16	.12	.59	553	6.32
1984	157	5.06	.21	.31	2.27	94	4.54
1985	531	6.28	.13	.08	.35		
1986	305	5.72	.17	.17	1.03		