DTU Aqua: Department of Marine Fisheries
http://www.dfu.dtu.dk/

State–space Stock Assessment as simple alternative to (semi) deterministic approaches and full parametric stochastic models

Anders Nielsen
an@aqua.dtu.dk

Problems we wish to solve

- Deterministic approaches
  - Catch at age assumed known without error
  - Procedures not models
  - Convergence of a deterministic procedure
  - Ad-hoc adjustments
- Full parametric statistical models
  - Parametric F–structure (e.g. multiplicative)
  - Trade off between flexible with (too) many parameters and rigid with tractable number of parameters
  - Number of parameters increase with every new year of data added

State-space assessment models

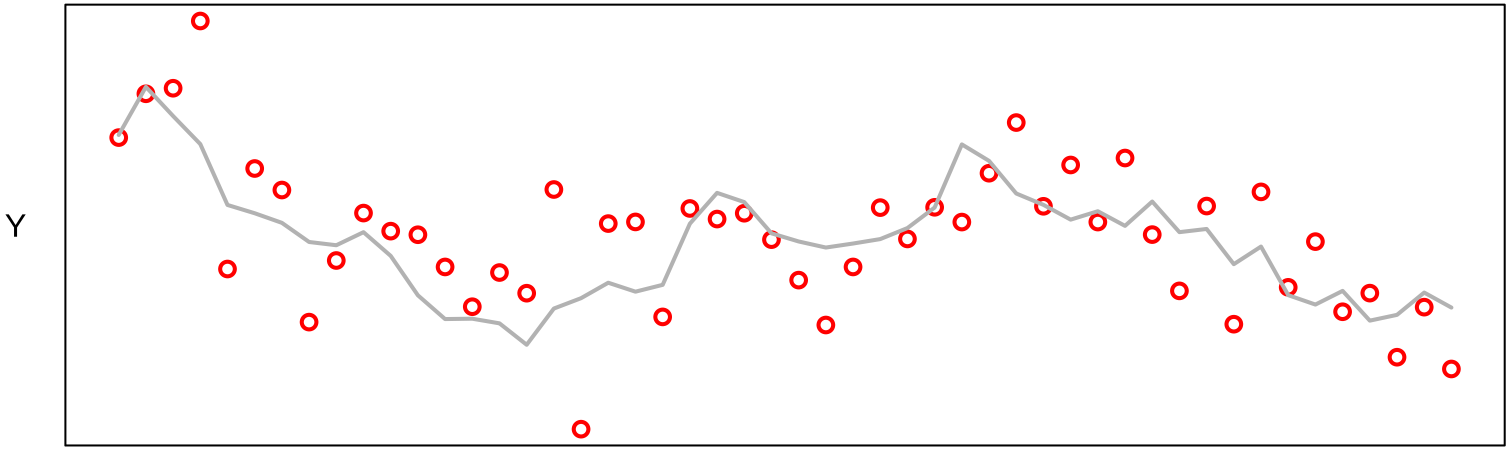
- A very useful extension to full parametric statistical models is state-space models[a]
- Introduced for stock assessment by Gudmundsson (1987,1994) and Fryer (2001)
- The reason state-space models have not been more frequently used in stock assessment is that software to handle these models has not been available
- Can give very flexible models with low number of model parameters
- For instance we can include things like:

$$F_{3,y} \ \text{ is a random walk with yearly variance } \ \sigma^2$$

- Notice:
  + Only one parameter (σ) to estimate for all $F_{3,y}$ instead of one every year
  + $F_{3,y}$ are predicted once the parameters are estimated
  + Each $F_{3,y}$ is not estimated in isolation, but as part of the F_3-series
  + This model includes the intuition behind 'F-shrinkage' and 'tapered time weights', but the amount is objectively estimated

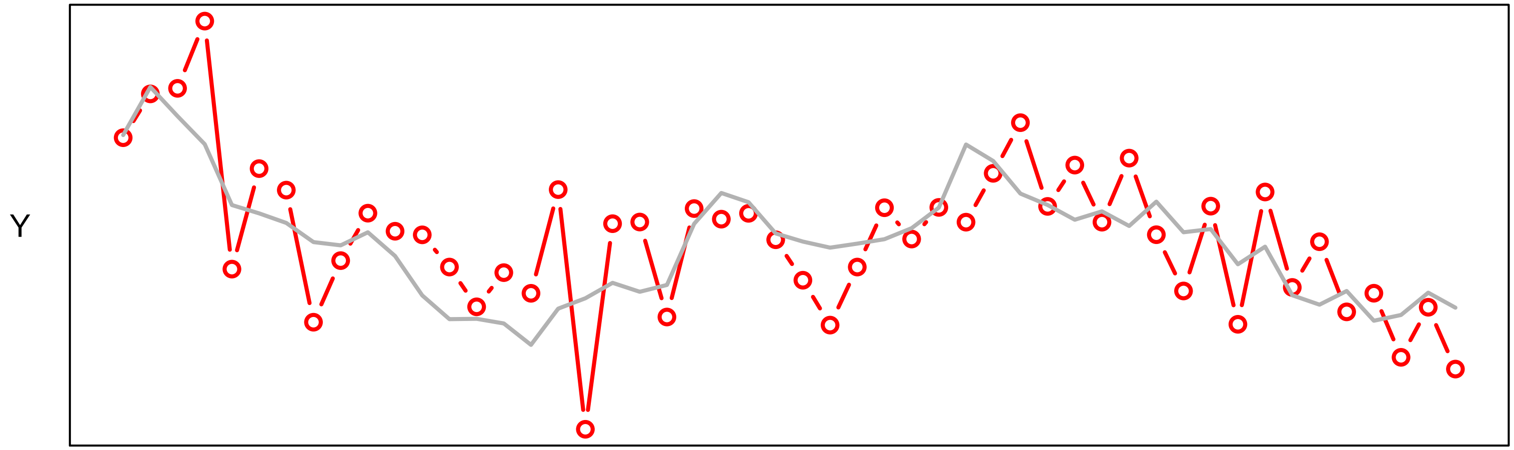
[a] a.k.a. random effects models, mixed models, latent variable models, hierarchical models, ...

Illustration of the three types of models



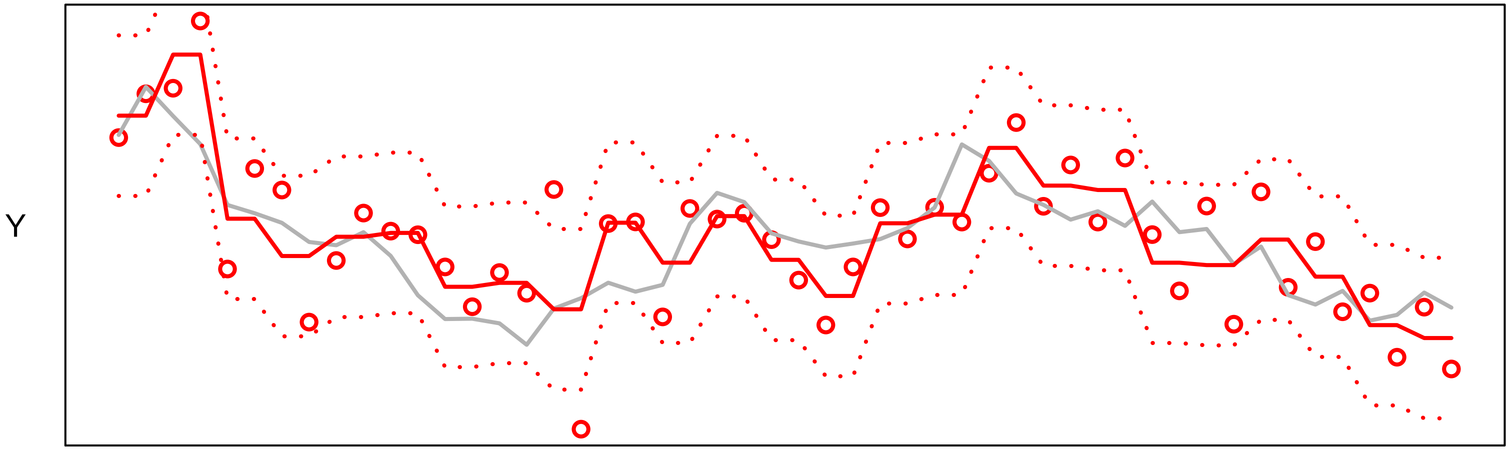
- Consider this example:
  - The true underlying F (here grey) follows a random walk with variance σ_F^2
  - But we only observe Y (here red circles) which is F+'noise' with variance σ_Y^2

Deterministic model estimates



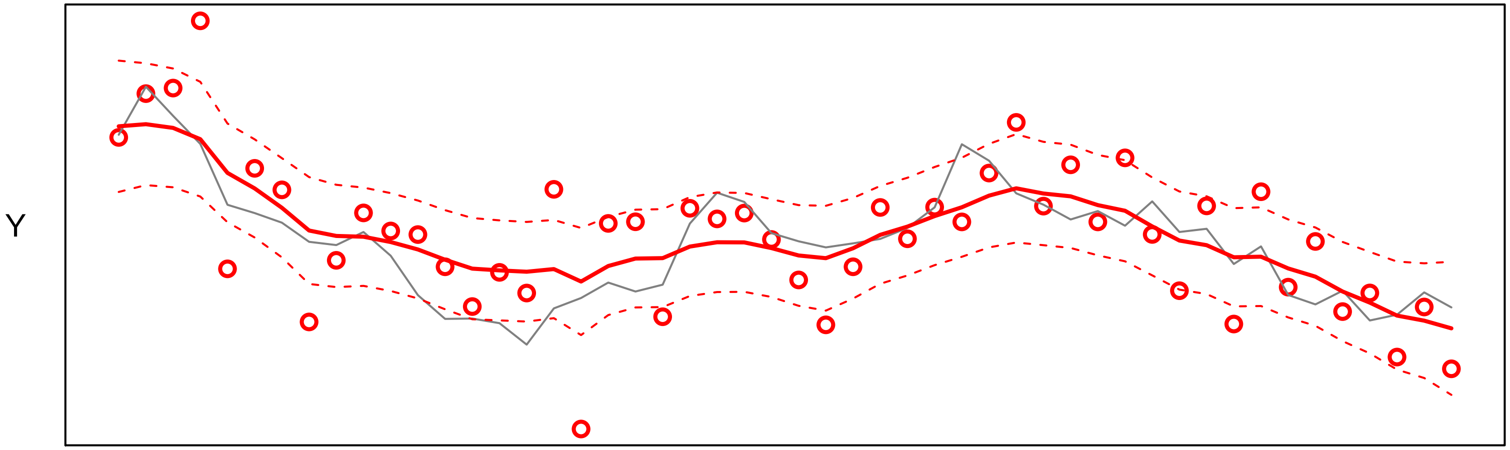
- If we assume no observation error the estimate of F is Y
- Too fluctuating
- No quantification of uncertainties

Fully parametrized statistical model estimates



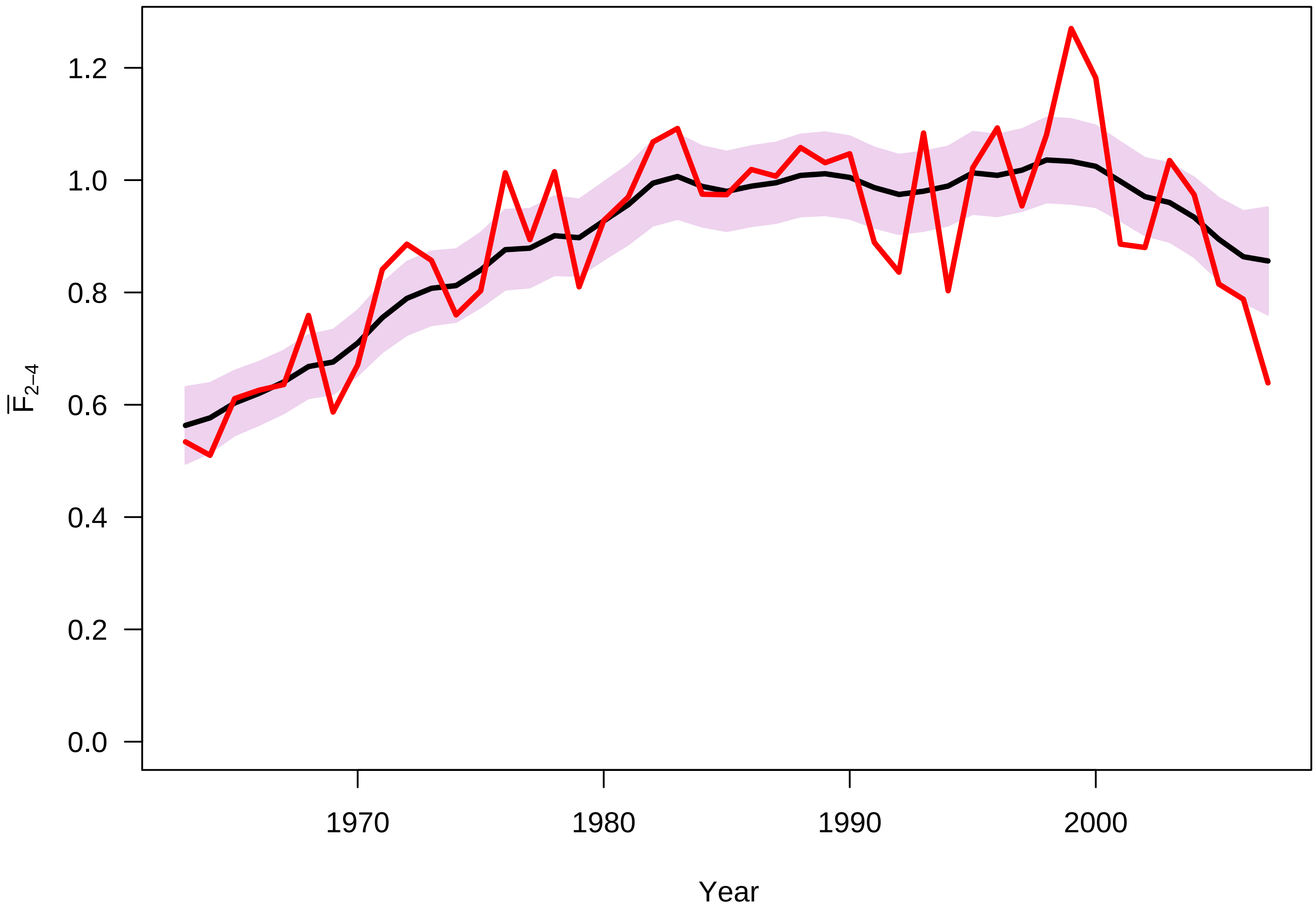
- To use a fully parametrized statistical model we first had to group the observations (here pairs, but choice is arbitrary)
- The reconstructed track appear OK
- The model contain 26 model parameters
- Uncertainties are estimated but the confidence interval seems too wide

State-space model estimates



- Consider λ as unobserved random variable
  - Estimate model parameters (σ_η and σ_ε) in marginal distribution $\int p(\lambda, Y)d\lambda$
  - Predict λ via distribution of $\lambda|Y$
- Closer reconstruction
- No artificial assumptions
- Two model parameters
- Correct coverage of the confidence interval
- Naturally this is just a simulated example, but ...

Example: $\overline{F}_{2-4}$ for North Sea Cod



Model

States are the random variables that we don't observe $(N_{a,y},\ F_{a,y})$

$$\begin{pmatrix} N_y \\ F_y \end{pmatrix} = T \begin{pmatrix} N_{y-1} \\ F_{y-1} \end{pmatrix} + \eta_y$$

Observations are the random variables that we do observe $(C_{a,y},\ I^{(s)}_{a,y})$

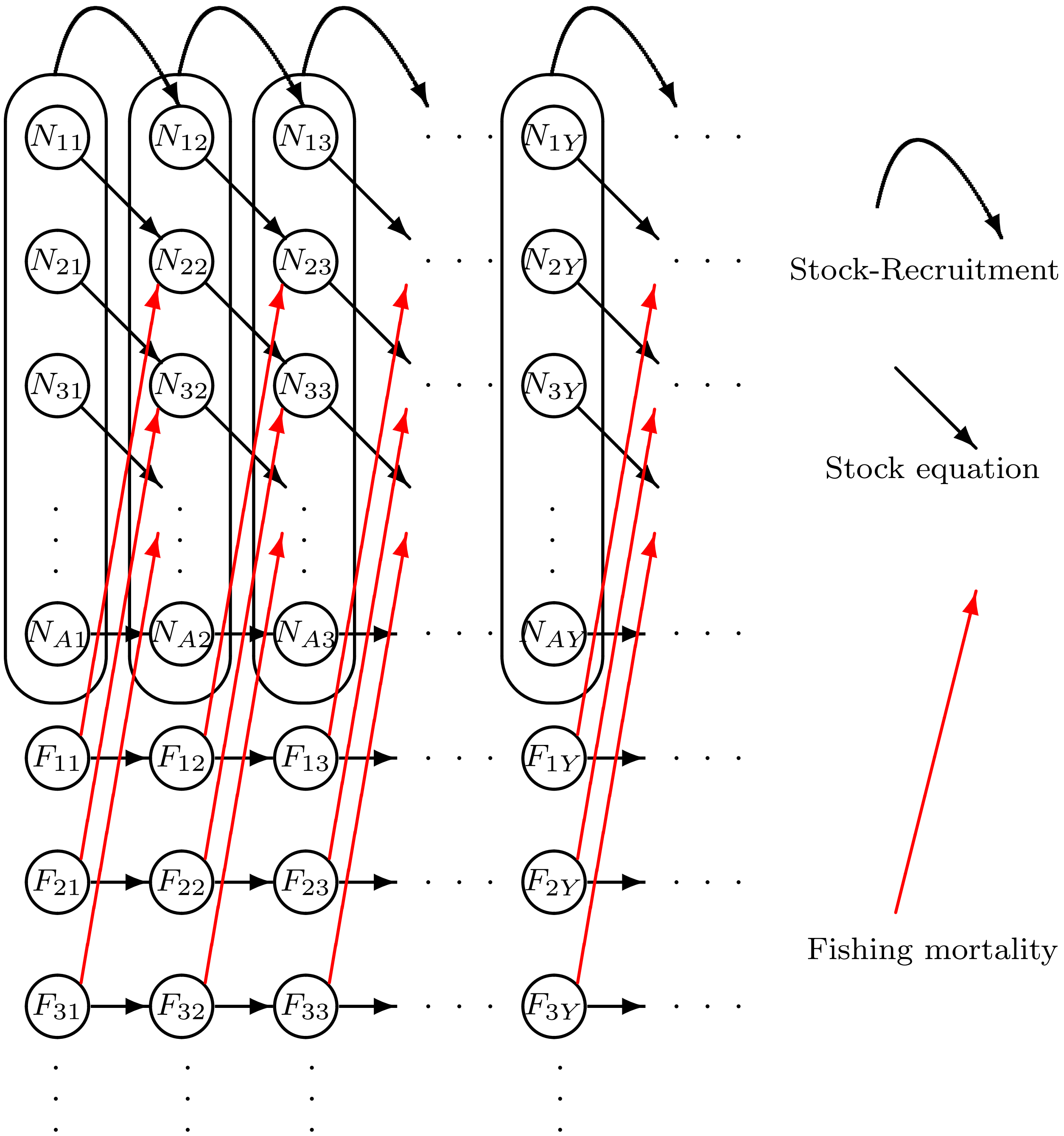
$$\begin{pmatrix} C_y \\ I^{(s)}_y \end{pmatrix} = O \begin{pmatrix} N_y \\ F_y \end{pmatrix} + \varepsilon_y$$

Model and parameters are what describes the distribution of states and observations through T, O, η_y, and ε_y.

Parameters: Survey catchabilities, S-R parameters, process and observation variances.

All model equation are as expected:

- Standard stock equation
- Standard stock recruitment (B-H, Ricker, or RW)
- Standard equations for total landings and survey indices


N_{11} N_{12} N_{13} N_{1Y}
N_{21} N_{22} N_{23} N_{2Y}
N_{31} N_{32} N_{33} N_{3Y}
N_{A1} N_{A2} N_{A3} N_{AY}
F_{11} F_{12} F_{13} F_{1Y}
F_{21} F_{22} F_{23} F_{2Y}
F_{31} F_{32} F_{33} F_{3Y}
Stock-Recruitment
Stock equation
Fishing mortality

Numerical Methods

- Kalman Filter
- Extended Kalman Filter
- Unscented Kalman Filter
- Laplace approximation
- Sampling based methods

(Numerical methods are needed to calculate the marginal distribution)

Optimization is done using AD Model Builder

Random effects in AD Model Builder

- In random effects models we have
  - Random variables we observe: x
  - Random variables we do not observe: z
  - Model parameters we want to estimate: θ
- If we had observed x and z we would have a joint likelihood $L(x, z, \theta)$
- but z is unobserved so we have to estimate θ in the marginal likelihood:

$$L(x, \theta) = \int L(x, z, \theta)dz$$

- This requires a high dimensional integral — which is difficult
- This is (part of) the reason MCMC methods are so widely used
- MCMC can be slow, difficult to judge convergence, and in tools like winBugs a prior must be assigned to everything — even when you have no prior information.
- AD Model Builder has a better solution

Laplace approximation

- Want to compute the marginal likelihood for a given θ value:

$$L(x,\theta) = \int L(x,z,\theta)dz$$

- First the joint likelihood $L(x,z,\theta)$ is optimized w.r.t. z.
- This optimization yields an estimate $\hat{z}$, and an estimated hessian $\mathcal{H}(\hat{z})$.
- Next a Gaussian approximation is assumed and the result (apart from a constant) is:

$$L(x,\theta) \approx |\det(\mathcal{H}(\hat{z}))|^{-0.5} L(x,\hat{z},\theta)$$

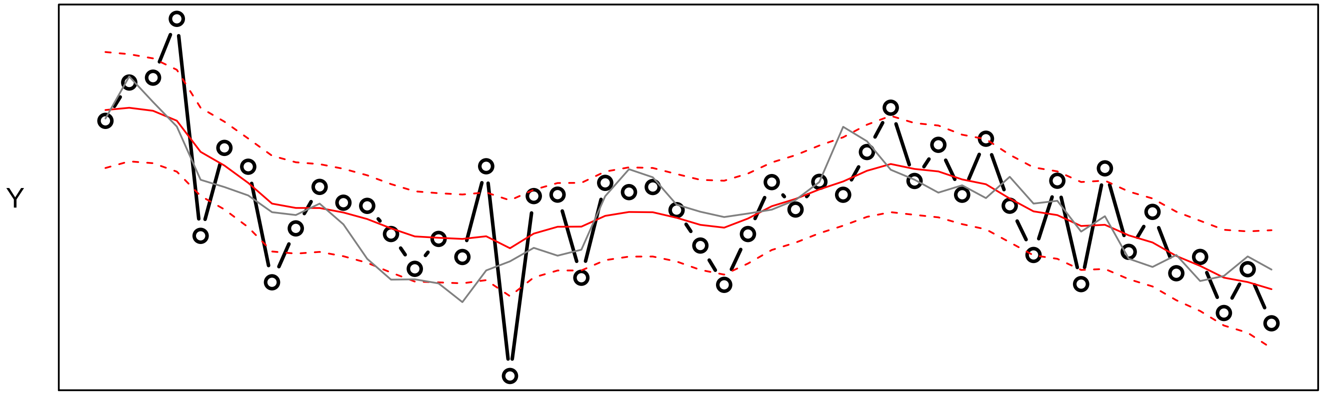
- Notice that when defined in this way $\hat{z}$ and $\mathcal{H}(\hat{z})$ and also depend on θ, which makes AD of this pretty difficult, but all solved for us in AD Model Builder.
- Actually this is all very simple to use. All we have to do is:
  - Code up the joint negative log likelihood
  - declare as `random_effects_vector z(1,n);`

```
DATA_SECTION
  init_int N
  init_vector y(1,N)

PARAMETER_SECTION
  init_number logSdLam
  init_number logSdy
  random_effects_vector lam(1,N);
  objective_function_value jnll;

PROCEDURE_SECTION
  jnll=0.0;
  dvariable var=exp(2.0*logSdLam);
  for(int i=2; i<=N; ++i){
    jnll+=0.5*(log(2.0*M_PI*var)
               +square(lam(i)-lam(i-1))/var);
  }
  var=exp(2.0*logSdy);
  for(int i=1; i<=N; ++i){
    jnll+=0.5*(log(2.0*M_PI*var)
               +square(lam(i)-y(i))/var);
  }
TOP_OF_MAIN_SECTION
  gradient_structure::set_MAX_NVAR_OFFSET(3000);
```

index	name	value	std dev
1	logSdLam	-2.3131e-01	2.8824e-01
2	logSdy	5.5298e-01	1.2554e-01
3	lam	-1.1611e+00	1.0876e+00
4	lam	-1.0768e+00	1.0052e+00
5	lam	-1.1900e+00	9.8298e-01
6	lam	-1.5622e+00	9.5522e-01
7	lam	-2.7309e+00	8.2689e-01
8	lam	-3.2436e+00	8.2301e-01
9	lam	-3.8943e+00	8.3214e-01
10	lam	-4.6716e+00	8.9801e-01
.			
.			
.			
43	lam	-5.1027e+00	8.2002e-01
44	lam	-5.5115e+00	8.2467e-01
45	lam	-5.4663e+00	8.2865e-01
46	lam	-5.8623e+00	8.2205e-01
47	lam	-6.1294e+00	8.2499e-01
48	lam	-6.6359e+00	8.2991e-01
49	lam	-7.0076e+00	8.4620e-01
50	lam	-7.4495e+00	8.9770e-01
51	lam	-7.6153e+00	9.5078e-01
52	lam	-7.8819e+00	1.1064e+00



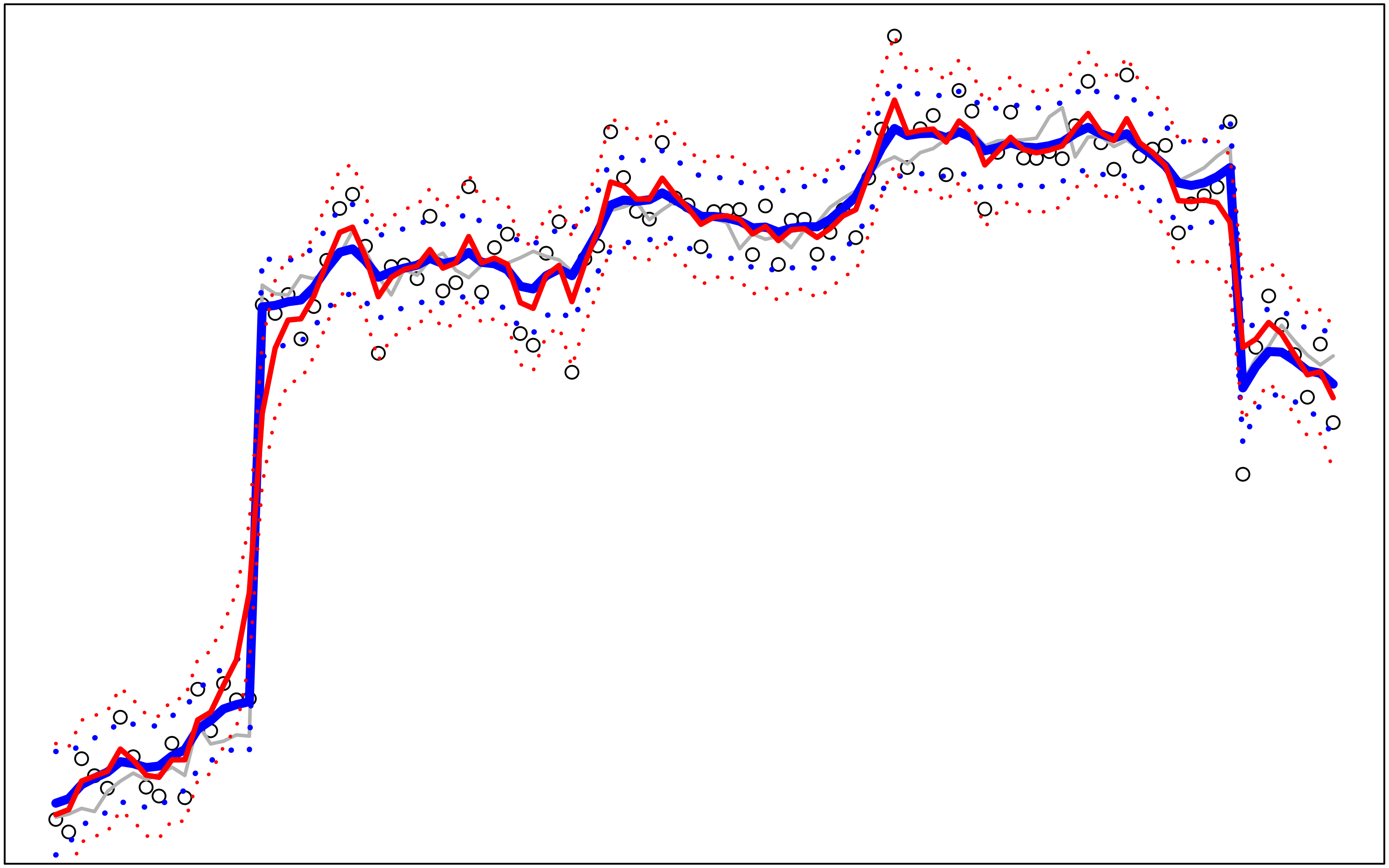
Status

- Primary model for Western Baltic Cod, Kattegat Cod, and Sole in 3A
- Exploratory model for Eastern Baltic Cod, North Sea Cod, and North Sea Sole
- Quick unsystematic tests for a few other stocks (Western Baltic spring spawning herring, North Sea Haddock, 3PS Cod, and Georges Bank Yellowtail Flounder)

From Fryer's listed disadvantages

- Requires normally distributed errors. **No, but they are still convenient.**
- Requires linear approximation of non-linear equations. **Not anymore.**
- There is some arbitrariness in the starting values. **Not anymore.**
- The likelihood can be very flat. **No change.**
- Maximum likelihood estimation can take a long time. **1-2 minutes on my laptop.**
- Initial coding is hard. **ADMB makes it easier**
- Favours status quo so struggles to pick up a collapsing stock.

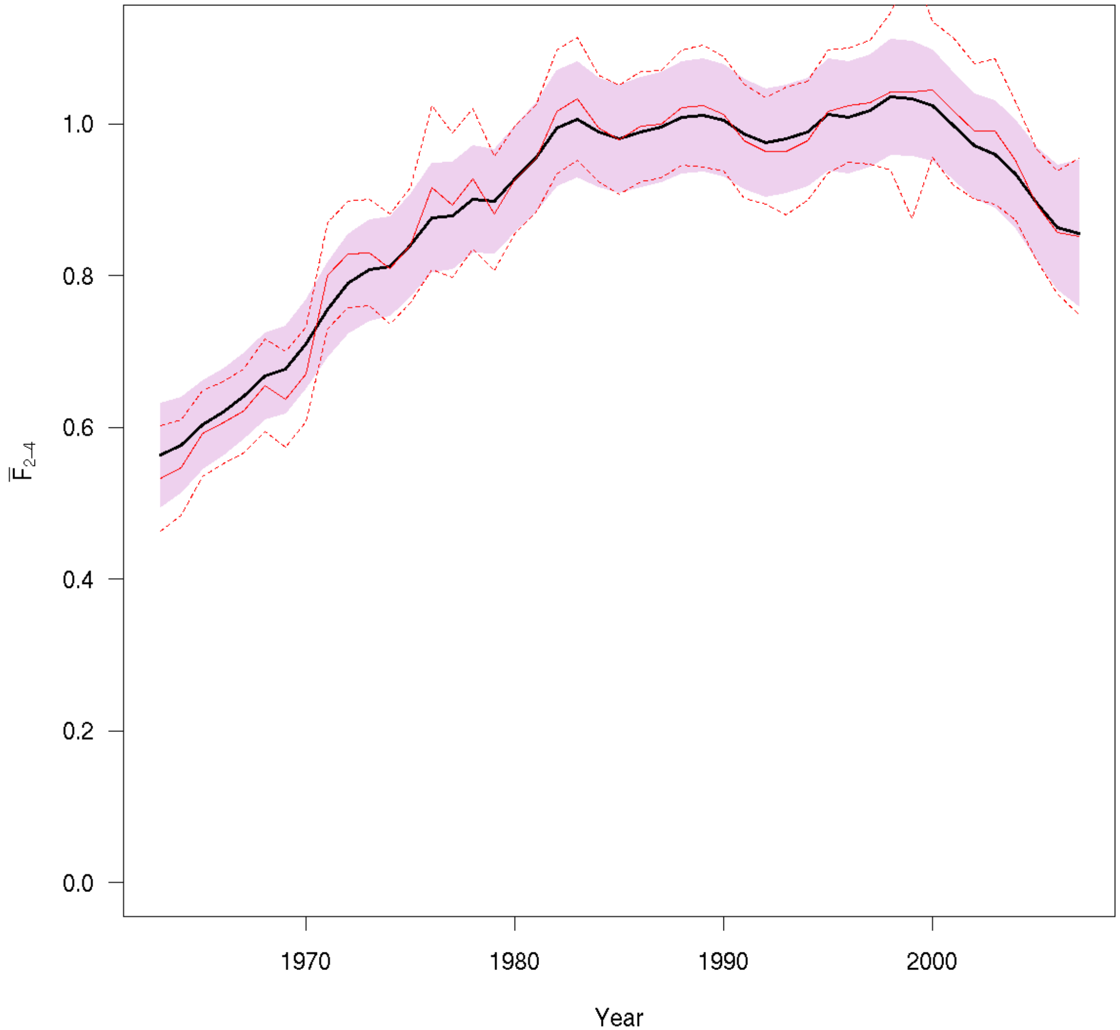
Allow sharp jumps



- In the standard model $\Delta \log F_y = \log F_y - \log F_{y-1}$ is assumed Gaussian
- Instead use a mixture, such as: $\Delta \log F_y \sim (1-p)\mathbf{N}(.,.) + p\mathbf{t}_1(.,.)$

Allow sharp jumps - results

- Allowing the **t**-jump-fraction p to be estimated.
  No change.
- Forcing $p = 10\%$.
  No visible change.
- Forcing $p = 30\%$.
  Visible change,
  but nothing dramatic



Correlated Random Walks

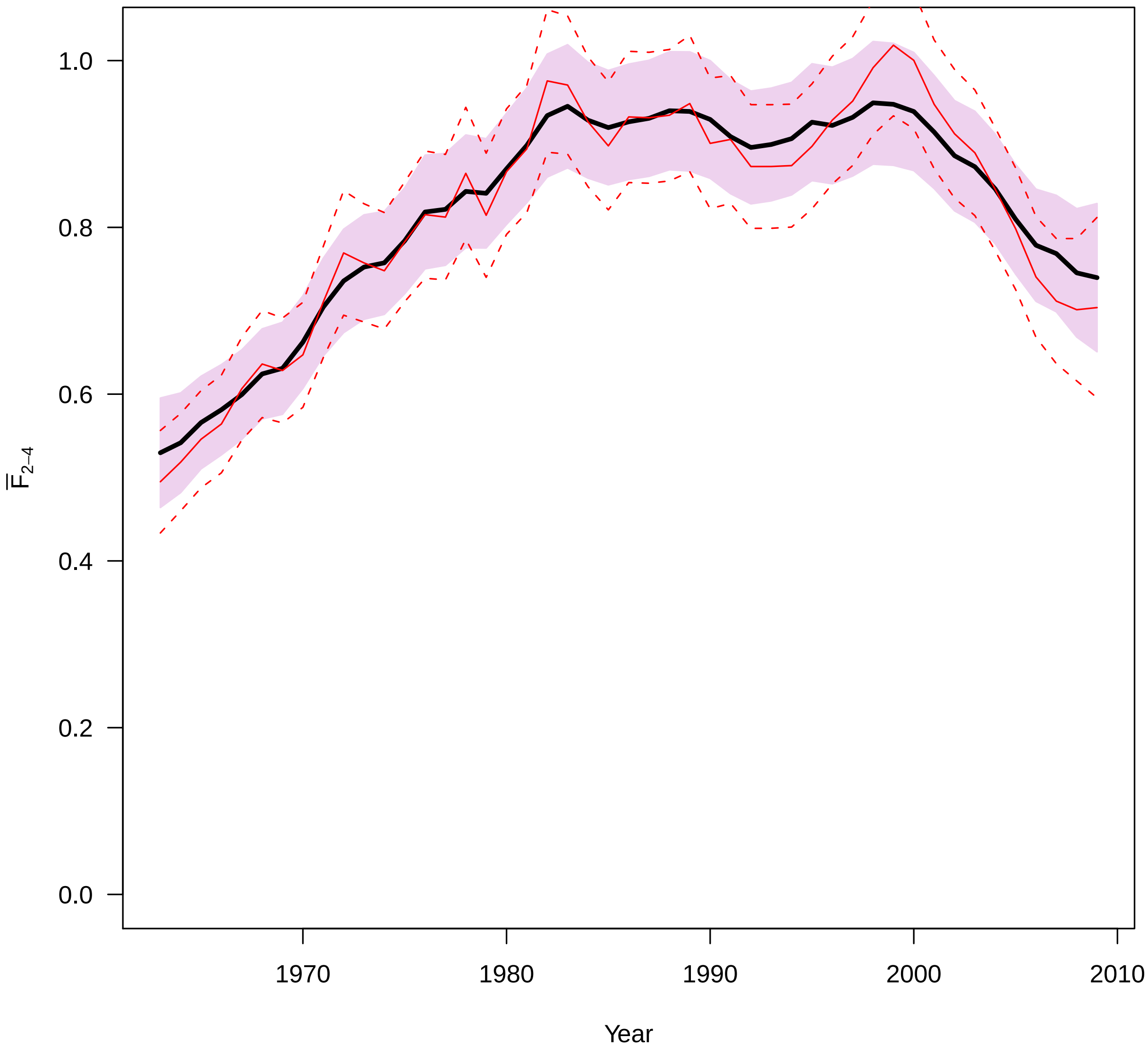
- Instead of independent random walks for F at different ages, we can allow those random walks to be correlated

$$\Delta \log(F) \sim \mathcal{N}(0, \Sigma)$$

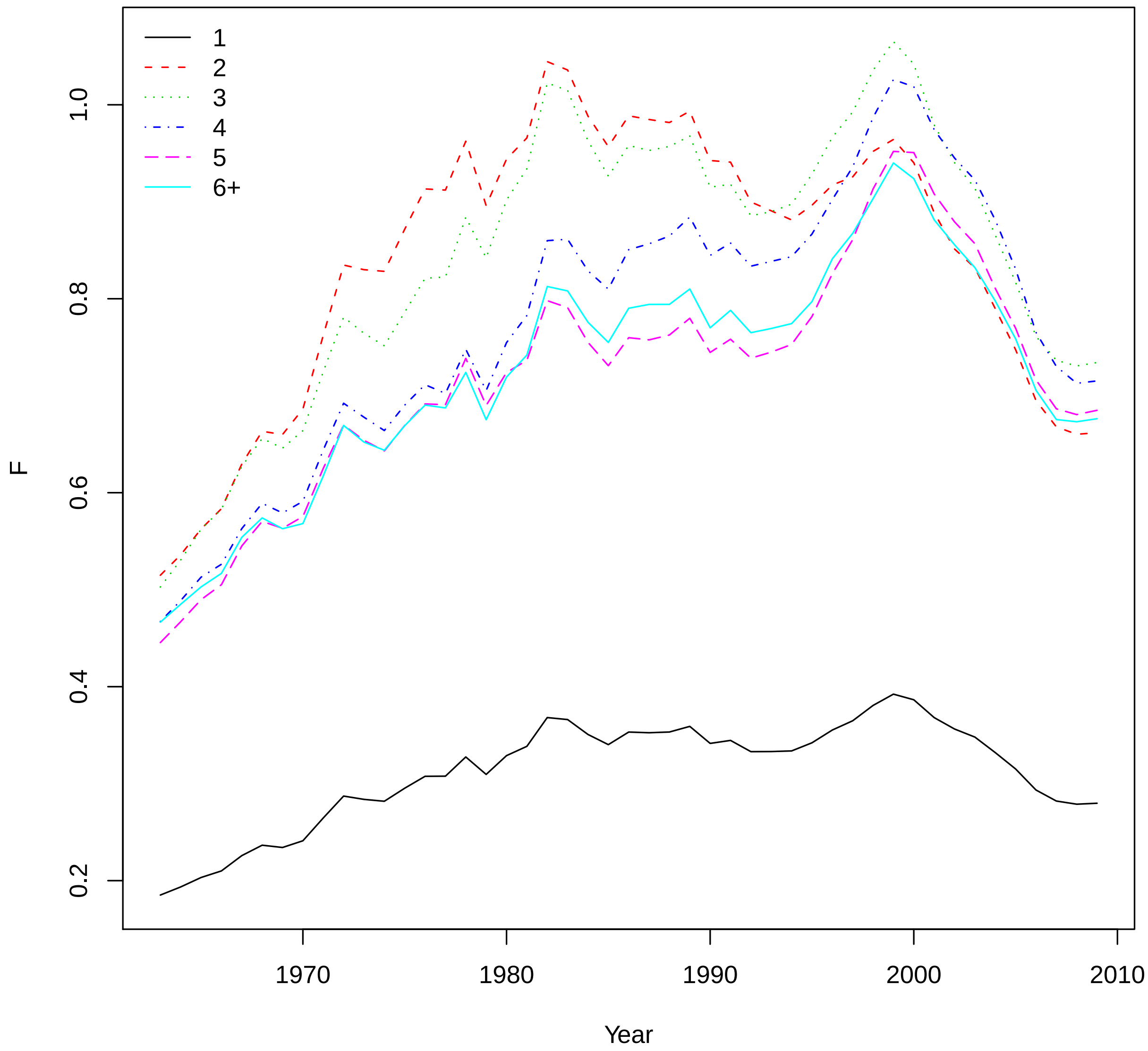
- The covariance matrix Σ is defined via the random walk variances, and the correlation coefficients $\rho_{i,j} = \Sigma_{i,j} / \sqrt{\Sigma_{i,i}\Sigma_{j,j}}$
- We assume the very simple structure

$$\rho_{i,j} = \begin{cases} 1, & \text{for } i = j \\ \rho, & \text{otherwise} \end{cases}$$

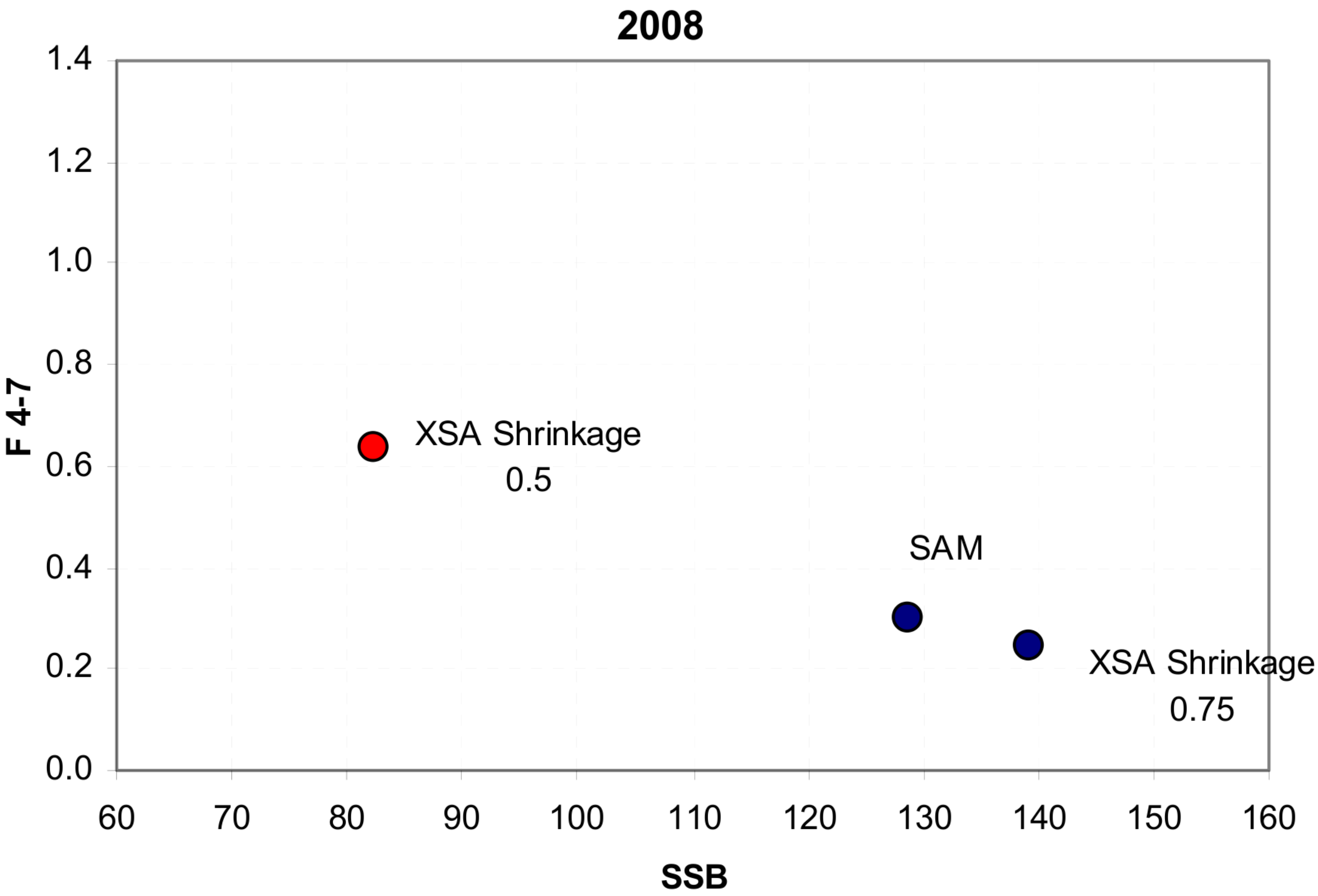
Correlated Random Walks — results



Correlated Random Walks — results



Avoiding ad-hoc choices — Eastern Baltic Cod



- Using the State-space Assessment Model (SAM) gives us an objective criteria

Features of the State-space assessment model

- Statistical model
- Consistent treatment of all $N_{a,y}$
- Random walk fishing mortality ($\log F_{a,y} = \log F_{a,y-1} + e_{a,y}$)
- Allows selectivity to evolve
- Maximum likelihood estimation of model parameters
- Estimation of uncertainties are an integrated part of the model
- Prediction is straight-forward
- Built-in 'F-shrinkage' and 'tapered time weights'
- Nicely handles missing observations
- Room for additional features

Web interface - Why?

- Scientific software is a way communicate ideas
- Peer review process is important
- Should be possible for all involved to:
  - see all details of the implementation[a]
  - run it themselves
  - experiment with data
  - experiment with model assumptions
  - run the same version
- The interface makes it all one step easier
- Will make update assessment very easy

[a] including all the "invisible" fixes they have had to include to get their models to work!

Web interface - How?

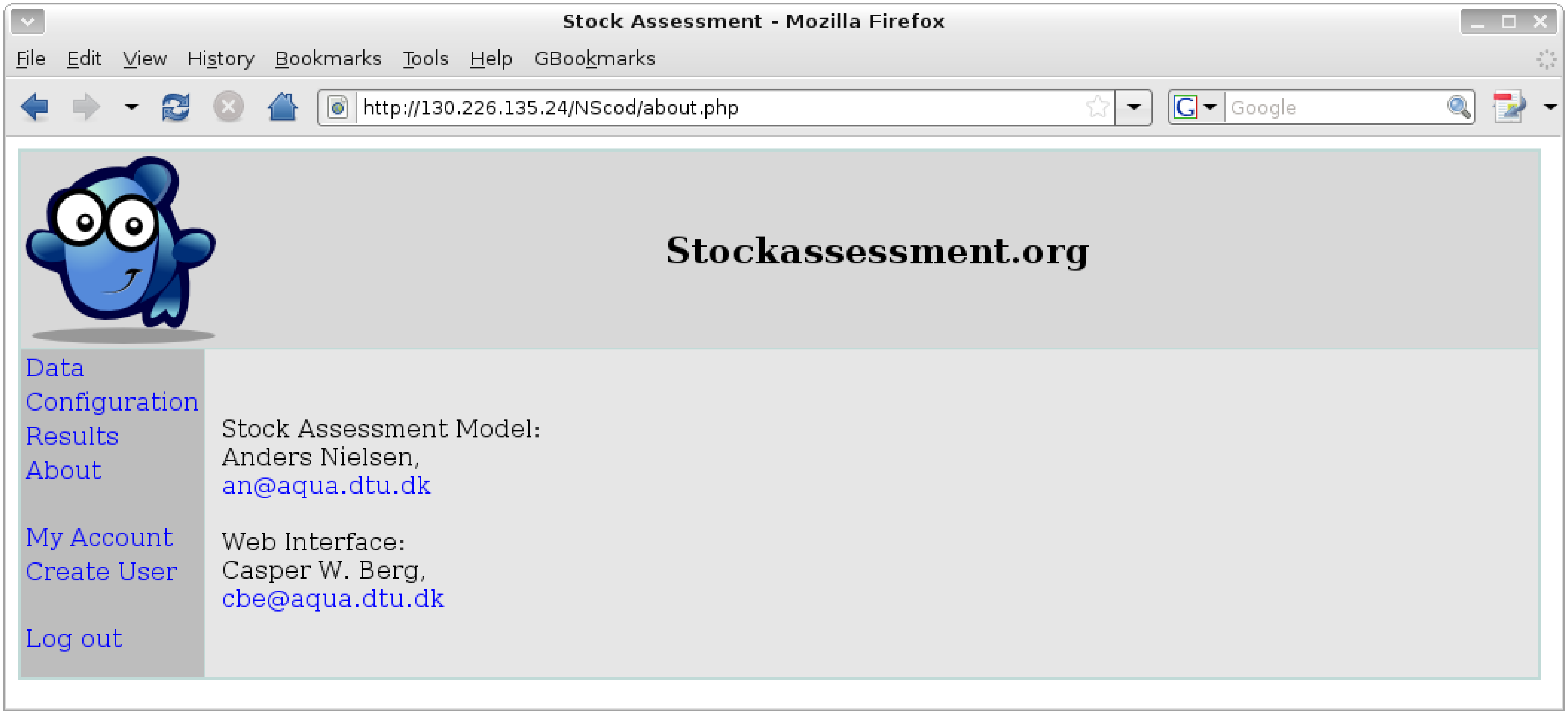
- Send an email[a] to request an account
- We will send a password back
- Go to the stock page:

http://www.kcod.stockassessment.org	http://www.wbcod.stockassessment.org
http://www.sole3a.stockassessment.org	http://www.ebcod.stockassessment.org
http://www.nscod.stockassessment.org	http://www.wbssher.stockassessment.org
http://www.sole4.stockassessment.org	http://www.plaice3a.stockassessment.org

- Log in to your new account

[a] an@aqua.dtu.dk or cbe@aqua.dtu.dk

Web interface - How?



(small demo?)

Three scenarios for North Sea Cod

$$\log(C_{a,y}S_{a,y}) = \log\left(\frac{F_{a,y}}{Z_{a,y}}(1-e^{-Z_{a,y}})N_{a,y}\right) + \varepsilon_{a,y}$$

- Use total landings as reported

$$S_{a,y} = 1$$

- Separate scaling each year

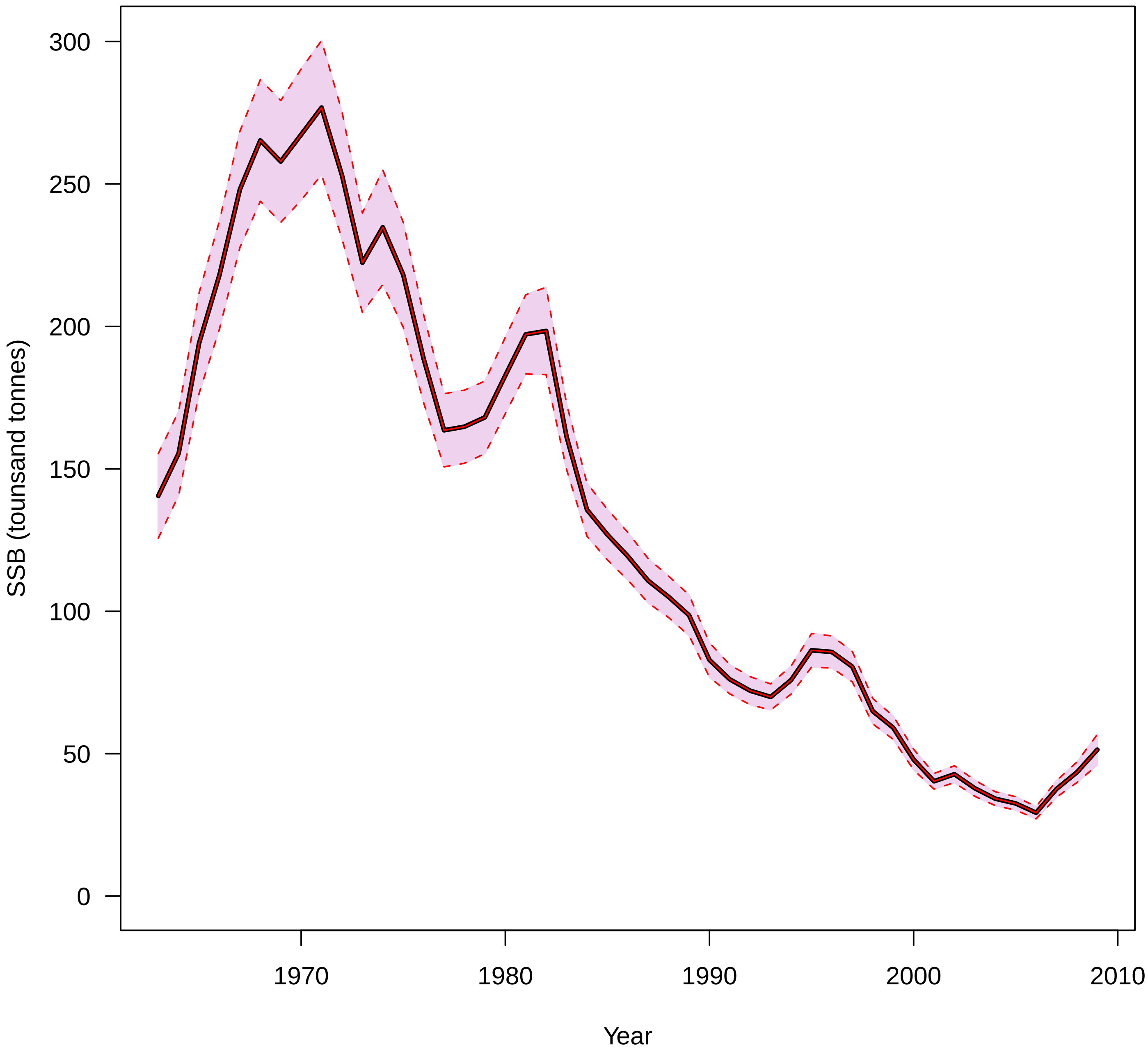
$$S_{a,y} = \begin{cases} 1, & y < 1993 \\ \tau_y, & y \geq 1993 \end{cases}$$

- Separate scaling each year, and in three age classes

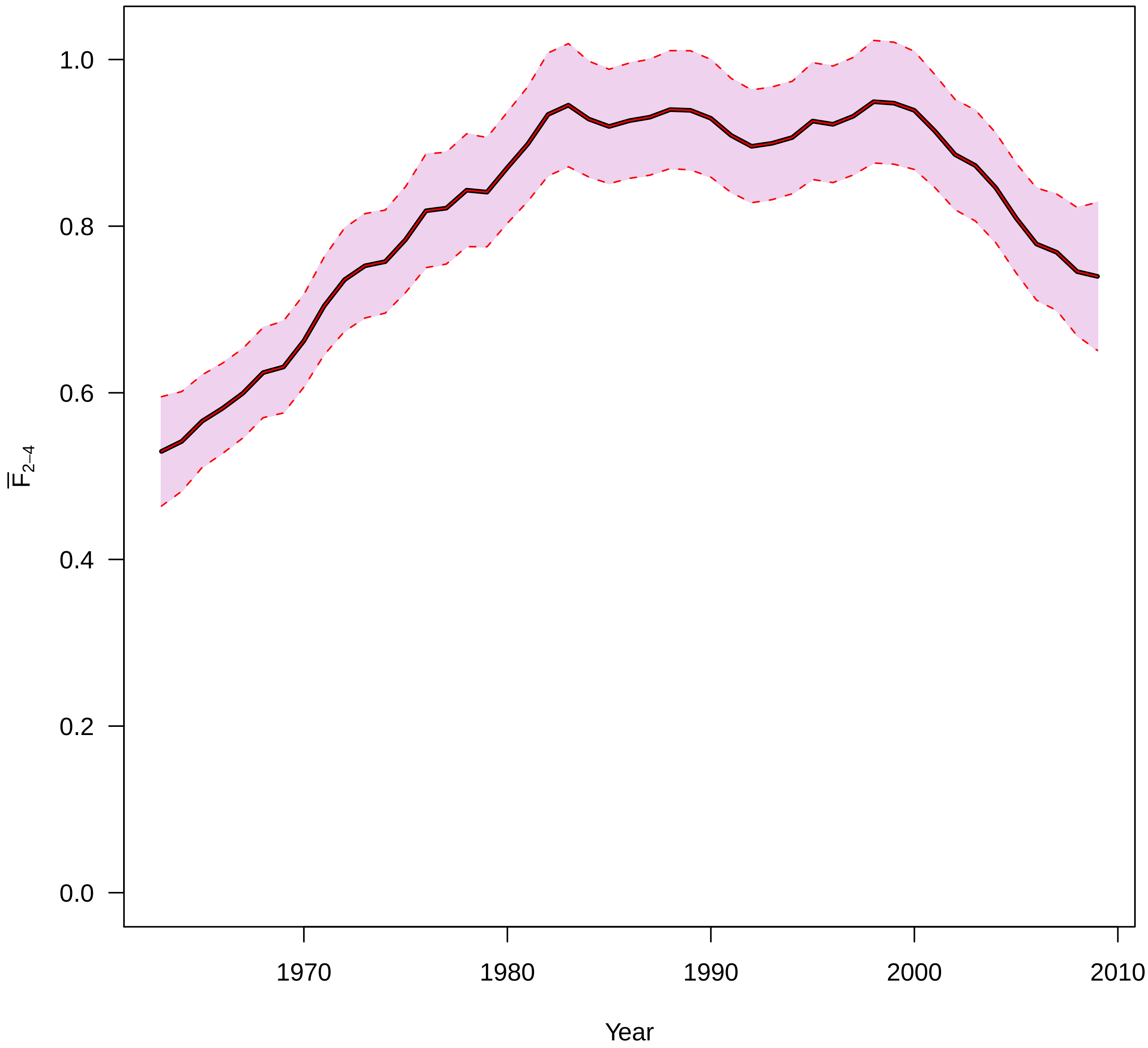
$$S_{a,y} = \begin{cases} 1, & y < 1993 \\ \tau_y^{(1)}, & y \geq 1993 \;\&\; a = 1 \\ \tau_y^{(2)}, & y \geq 1993 \;\&\; a = 2 \\ \tau_y^{(3+)}, & y \geq 1993 \;\&\; a \geq 3 \end{cases}$$

P = 0.000%

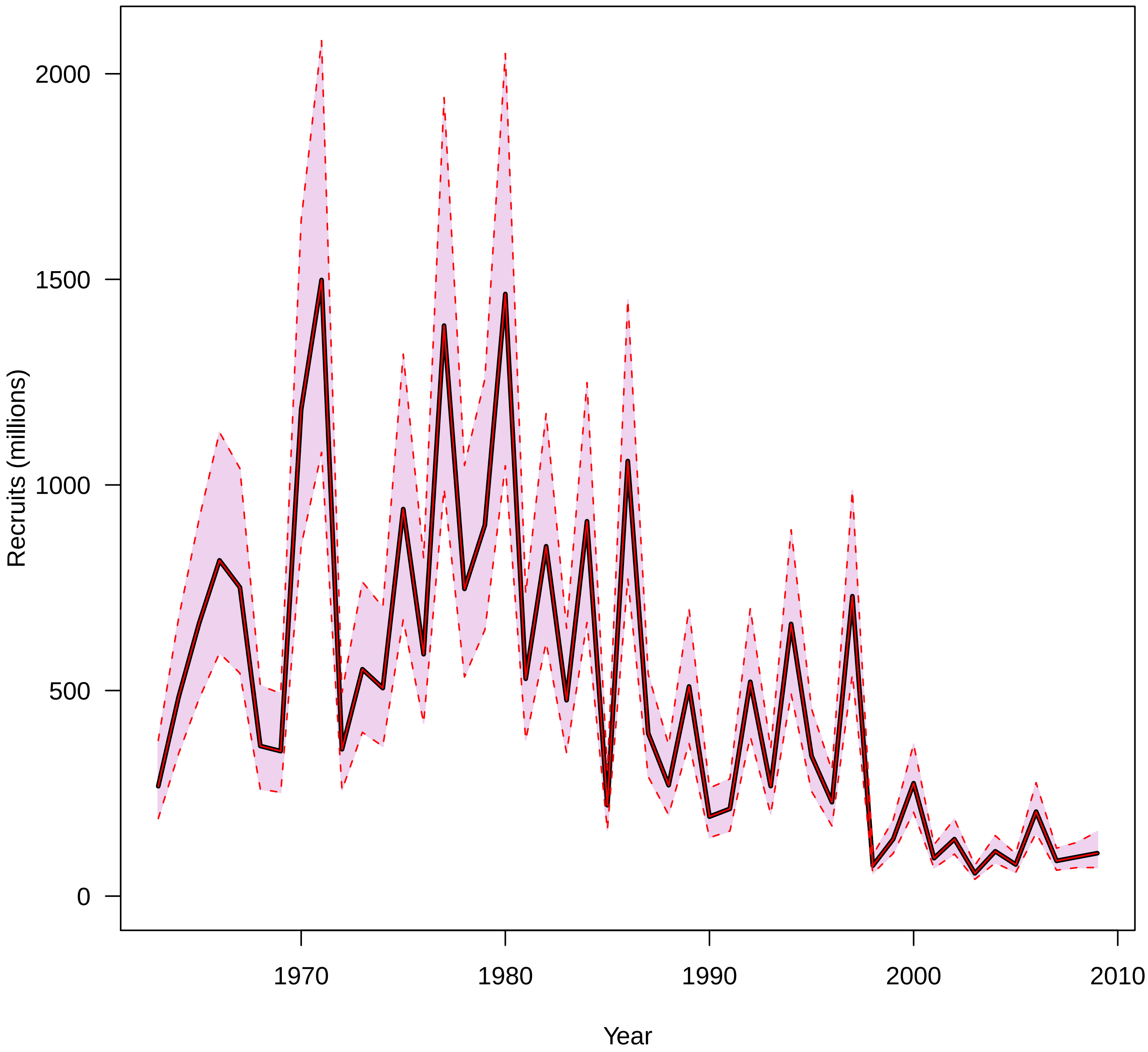
P = 17.4%

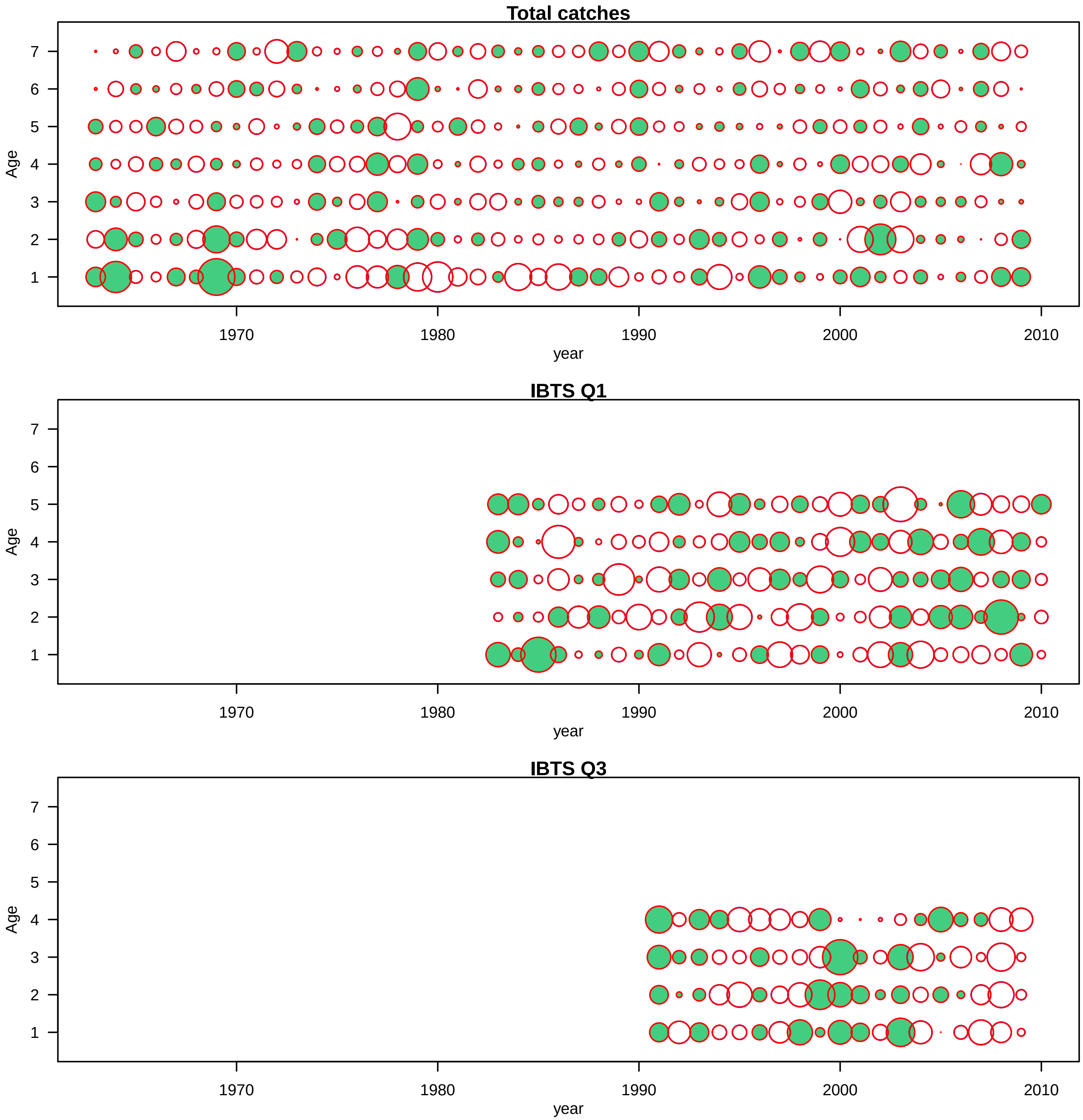


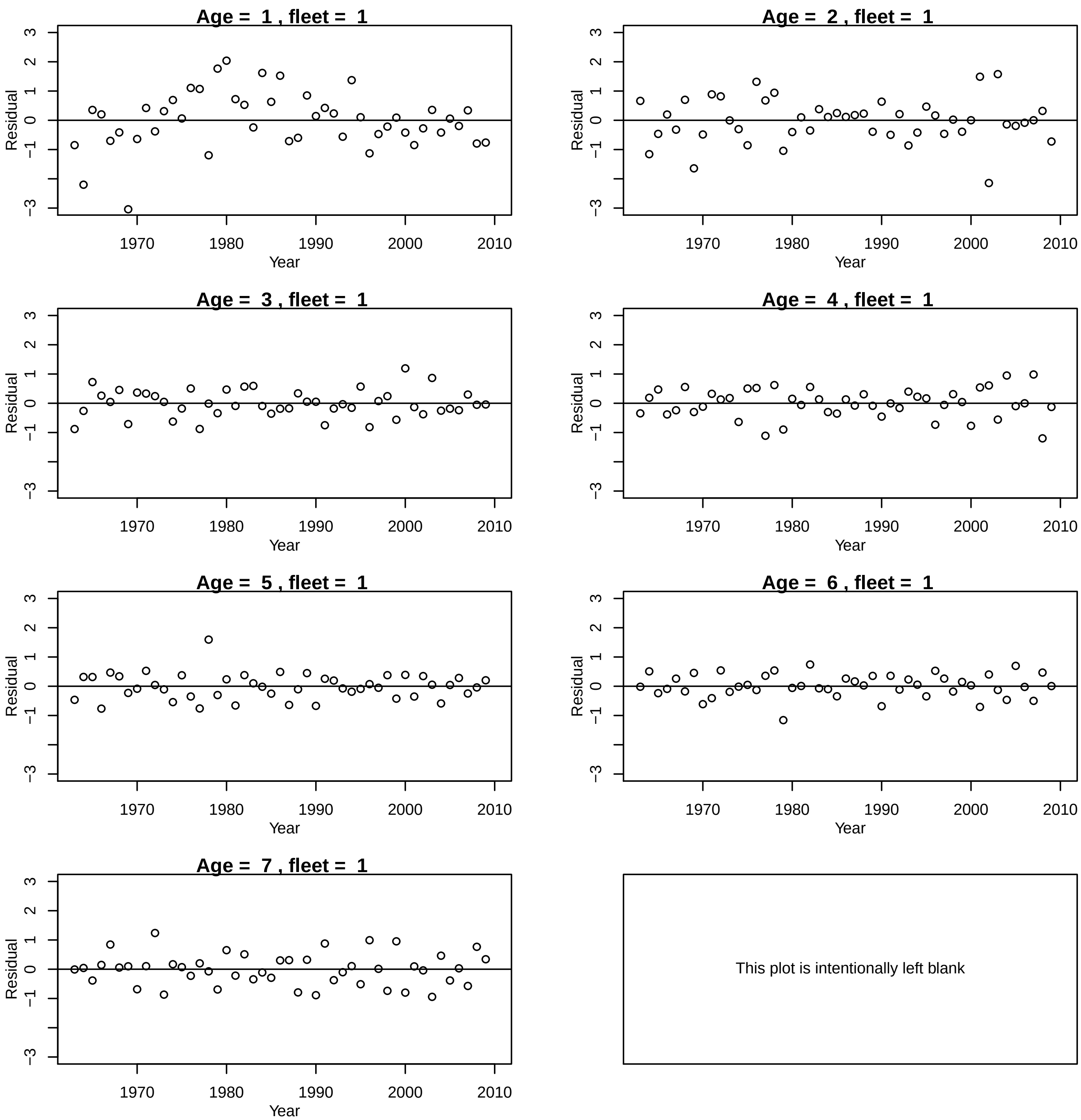
300
250
200
150
100
50
0
SSB (tounsand tonnes)
1970
1980
1990
2000
2010
Year

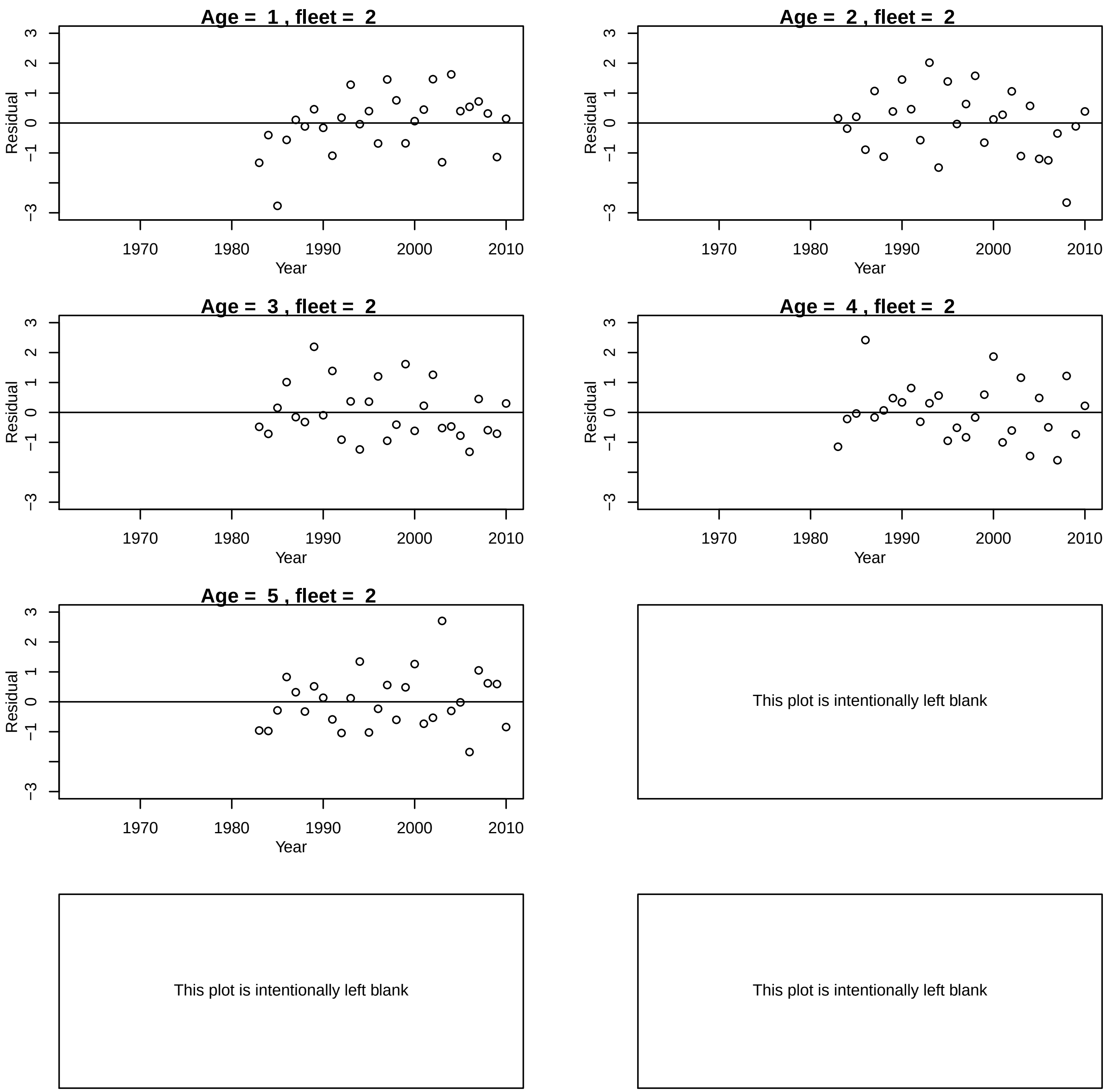


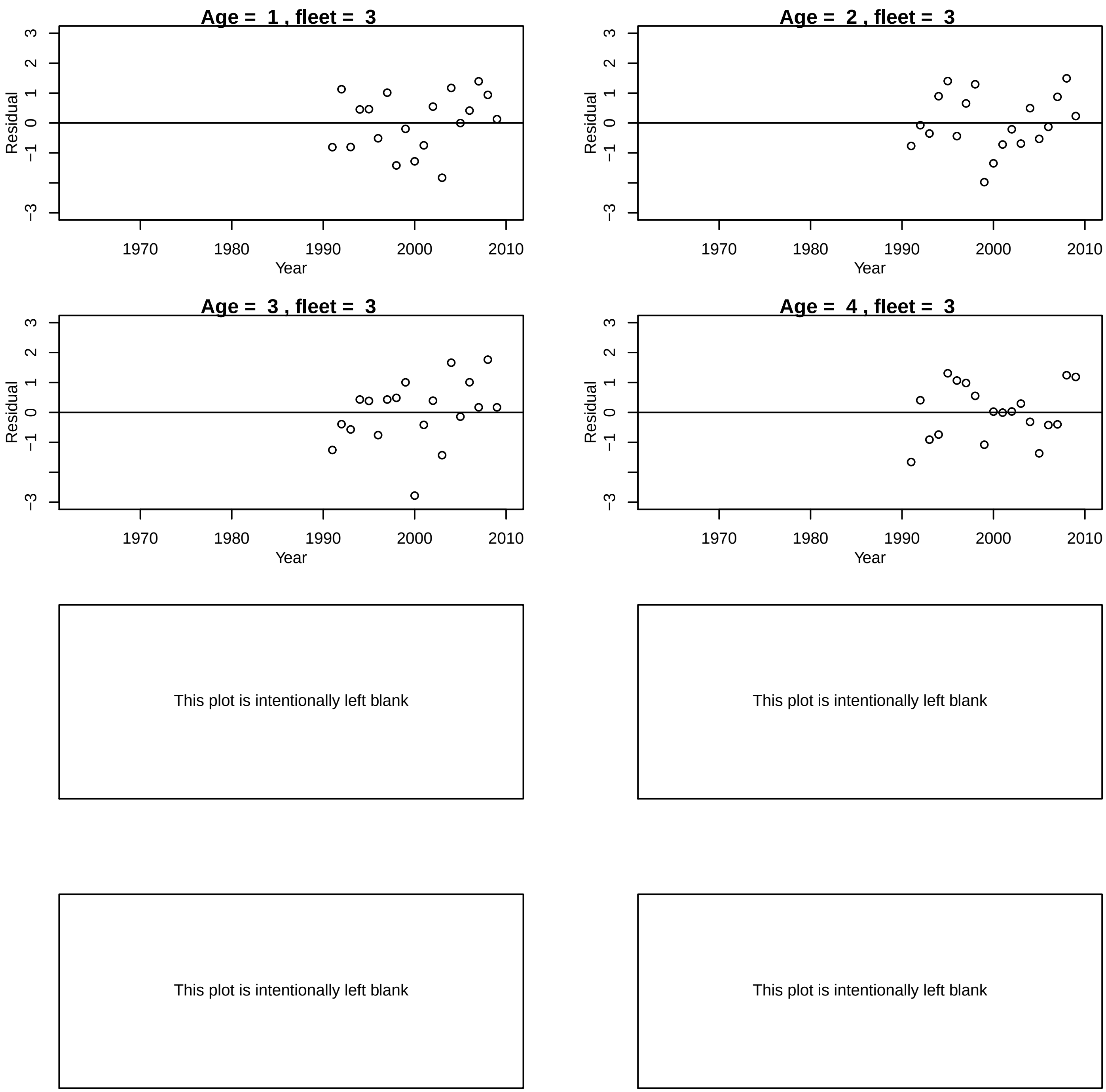
1.0
0.8
0.6
0.4
0.2
0.0
$\overline{F}_{2-4}$
1970
1980
1990
2000
2010
Year

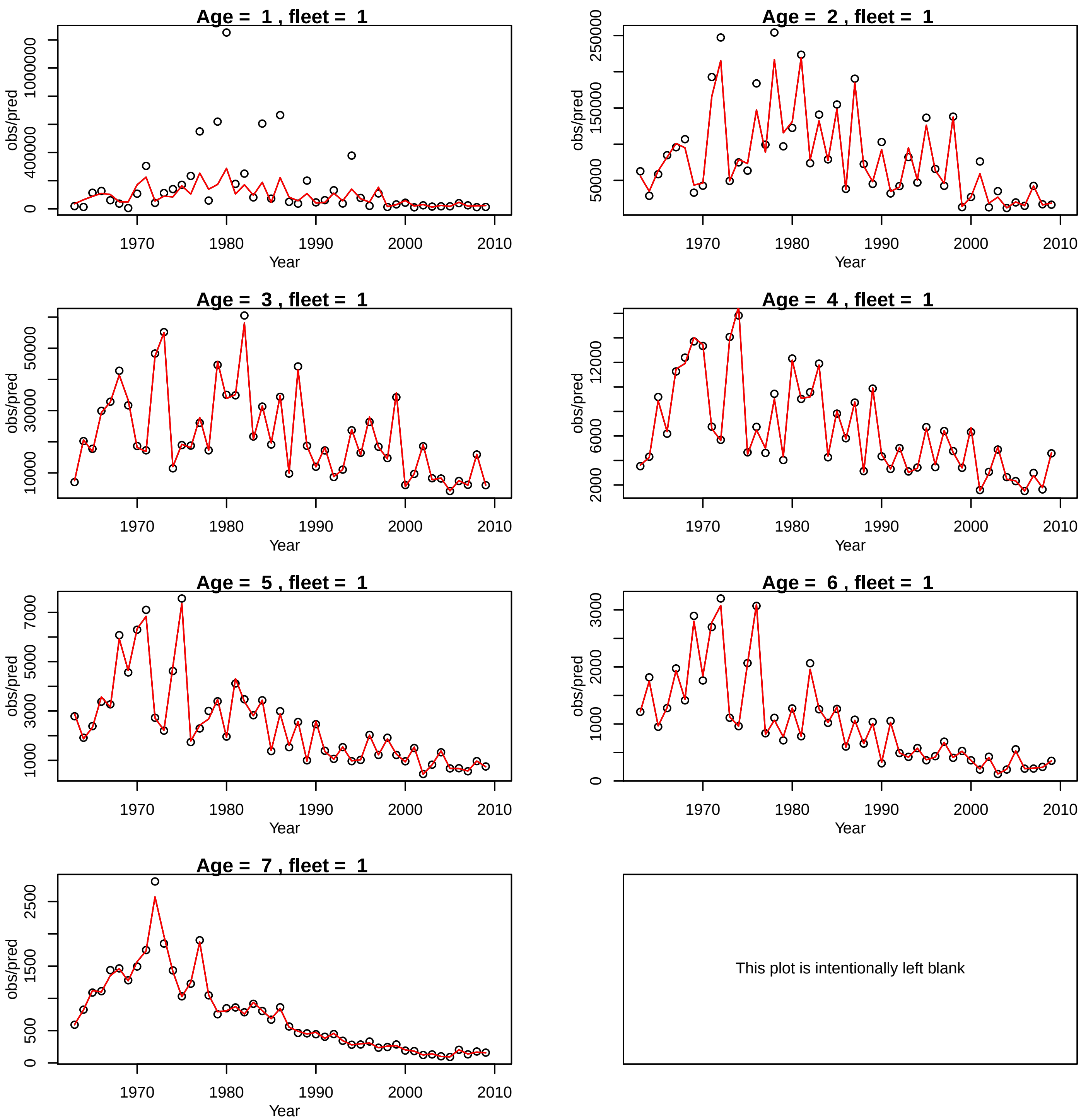

Recruits (millions)
2000
1500
1000
500
0
1970
1980
1990
2000
2010
Year


Total catches
Age
year
IBTS Q1
Age
year
IBTS Q3
Age
year
1970
1980
1990
2000
2010

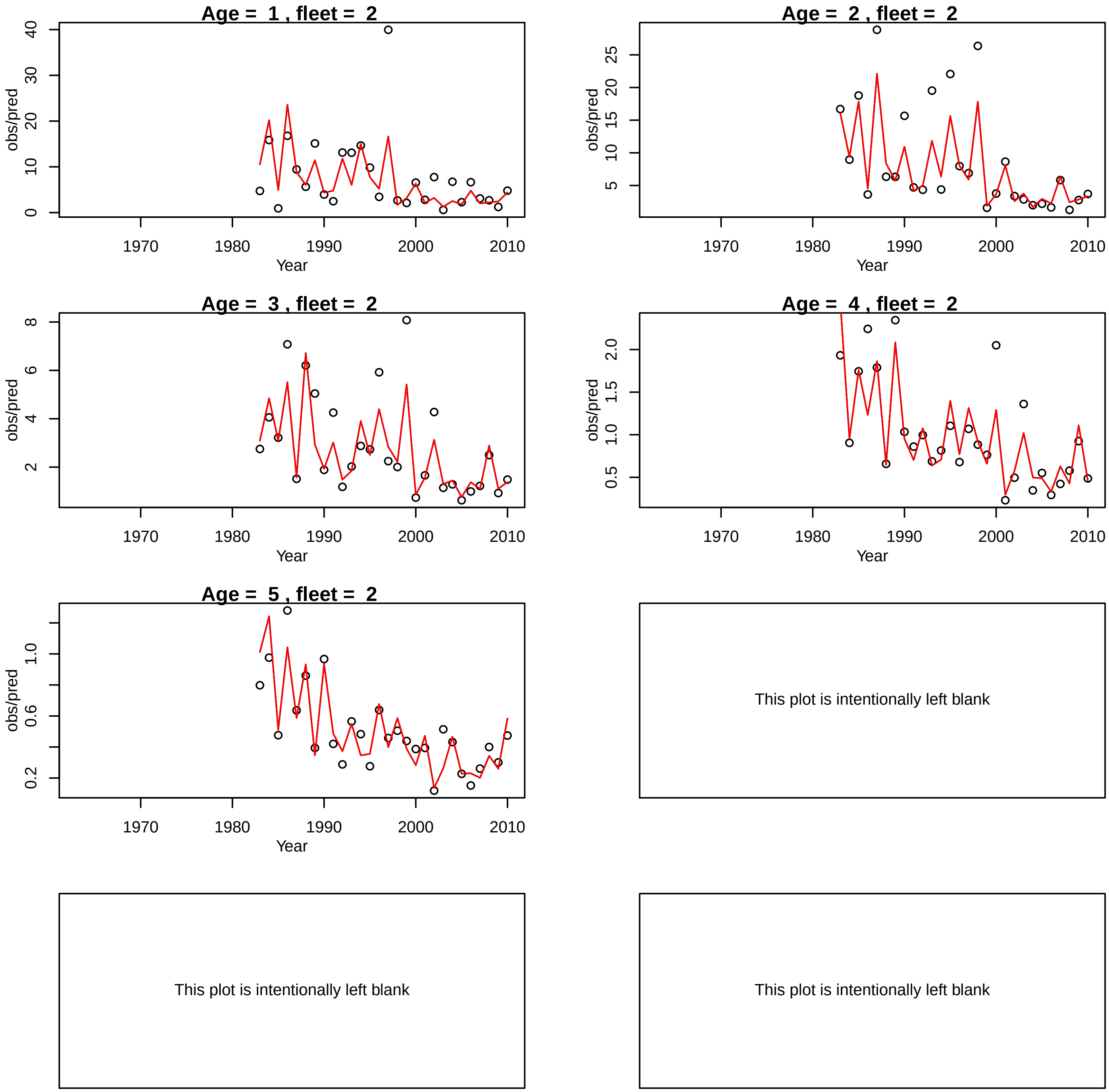

Age = 1 , fleet = 1
Age = 2 , fleet = 1
Age = 3 , fleet = 1
Age = 4 , fleet = 1
Age = 5 , fleet = 1
Age = 6 , fleet = 1
Age = 7 , fleet = 1
Residual
Year
−3
−1
0
1
2
3
1970
1980
1990
2000
2010
This plot is intentionally left blank

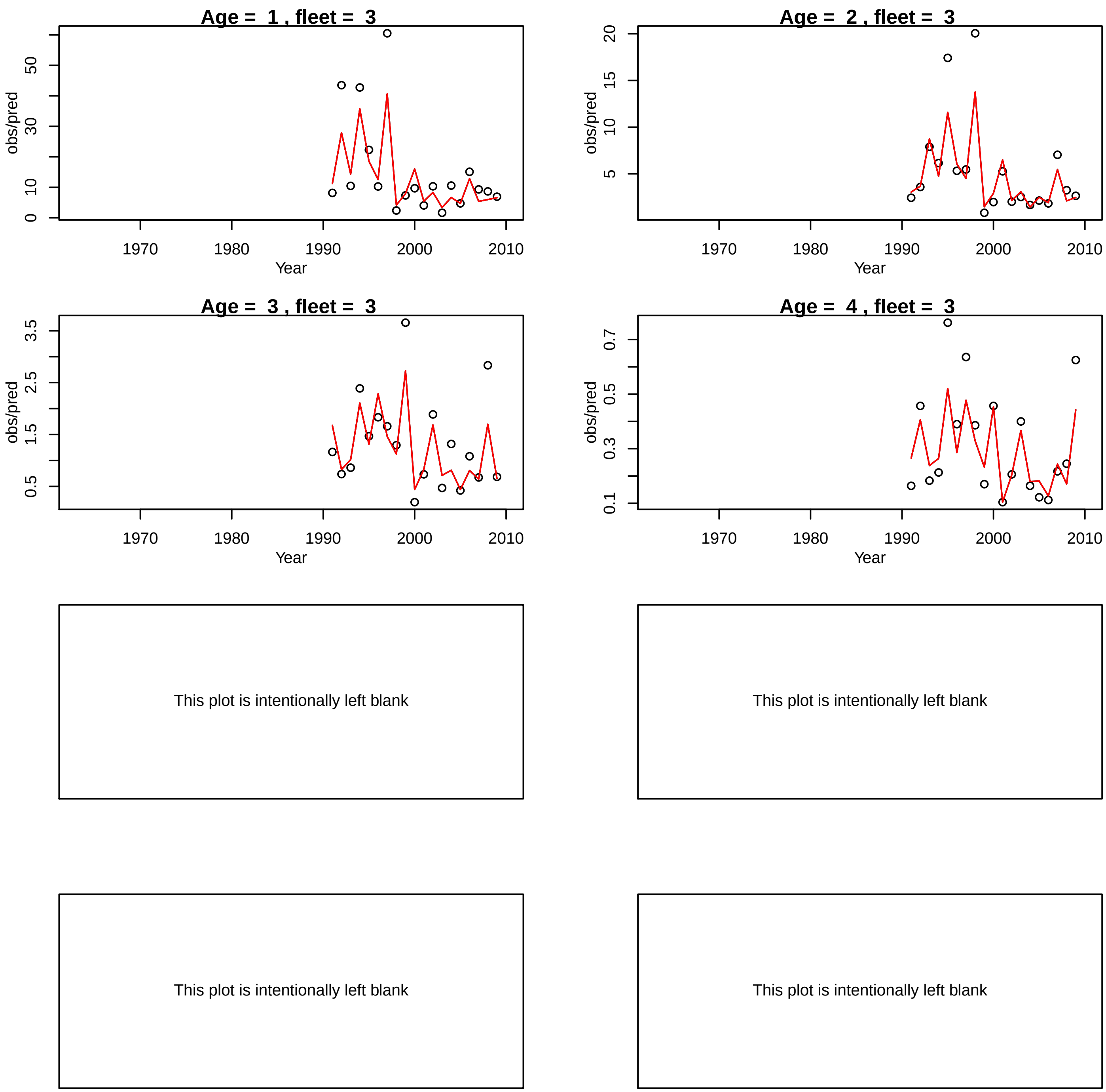

Age = 1 , fleet = 2
Age = 2 , fleet = 2
Age = 3 , fleet = 2
Age = 4 , fleet = 2
Age = 5 , fleet = 2
Residual
Year
1970
1980
1990
2000
2010
−3
−1
0
1
2
3
This plot is intentionally left blank
This plot is intentionally left blank
This plot is intentionally left blank

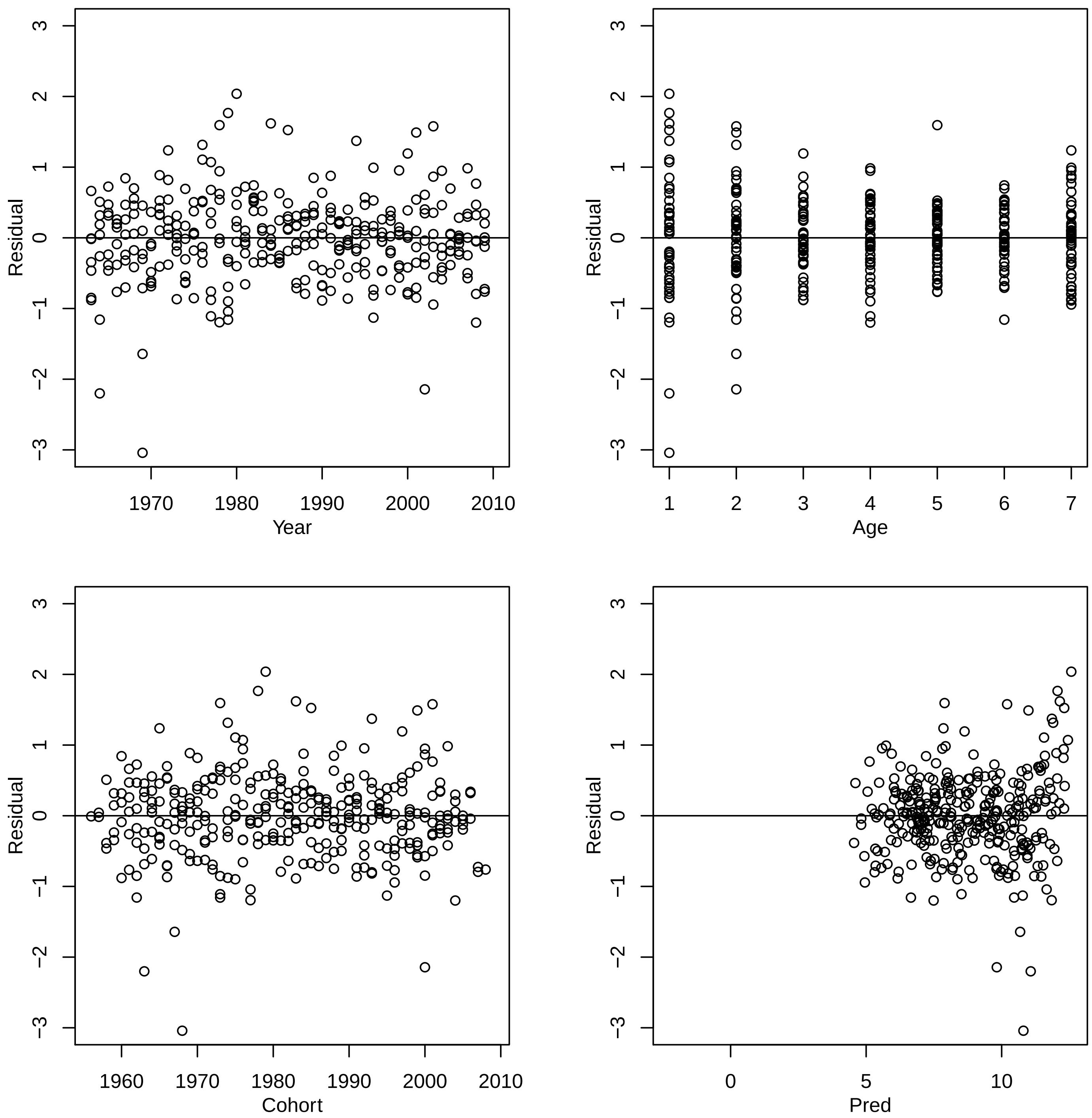

Age = 1 , fleet = 3
Age = 2 , fleet = 3
Age = 3 , fleet = 3
Age = 4 , fleet = 3
Residual
3
2
1
0
−1
−3
1970
1980
1990
2000
2010
Year
This plot is intentionally left blank
This plot is intentionally left blank
This plot is intentionally left blank
This plot is intentionally left blank

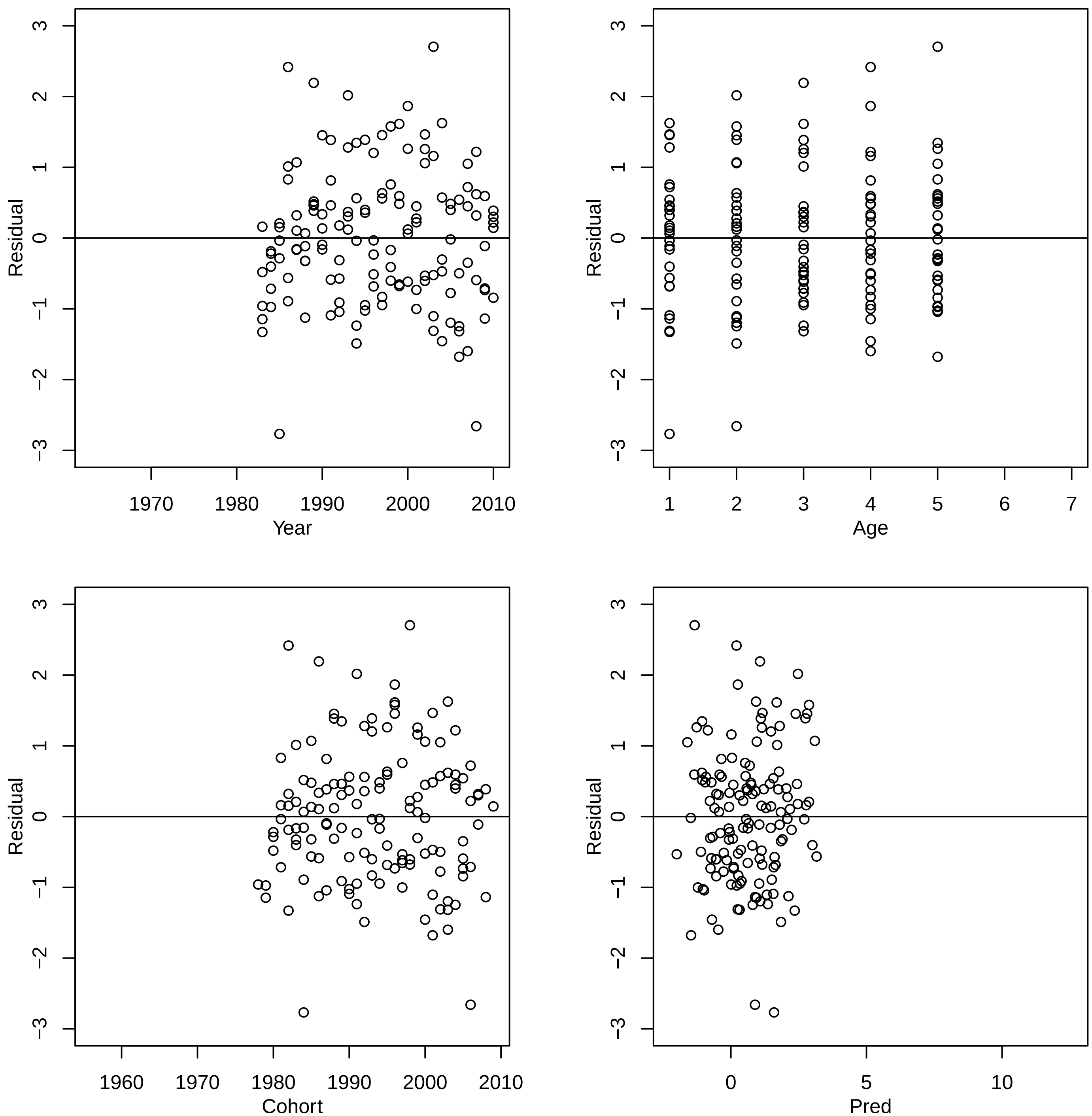


Age = 1 , fleet = 1
obs/pred
0
400000
1000000
1970
1980
1990
2000
2010
Year
Age = 2 , fleet = 1
obs/pred
50000
150000
250000
Year
Age = 3 , fleet = 1
obs/pred
10000
30000
50000
Year
Age = 4 , fleet = 1
obs/pred
2000
6000
12000
Year
Age = 5 , fleet = 1
obs/pred
1000
3000
5000
7000
Year
Age = 6 , fleet = 1
obs/pred
0
1000
2000
3000
Year
Age = 7 , fleet = 1
obs/pred
0
500
1500
2500
Year
This plot is intentionally left blank

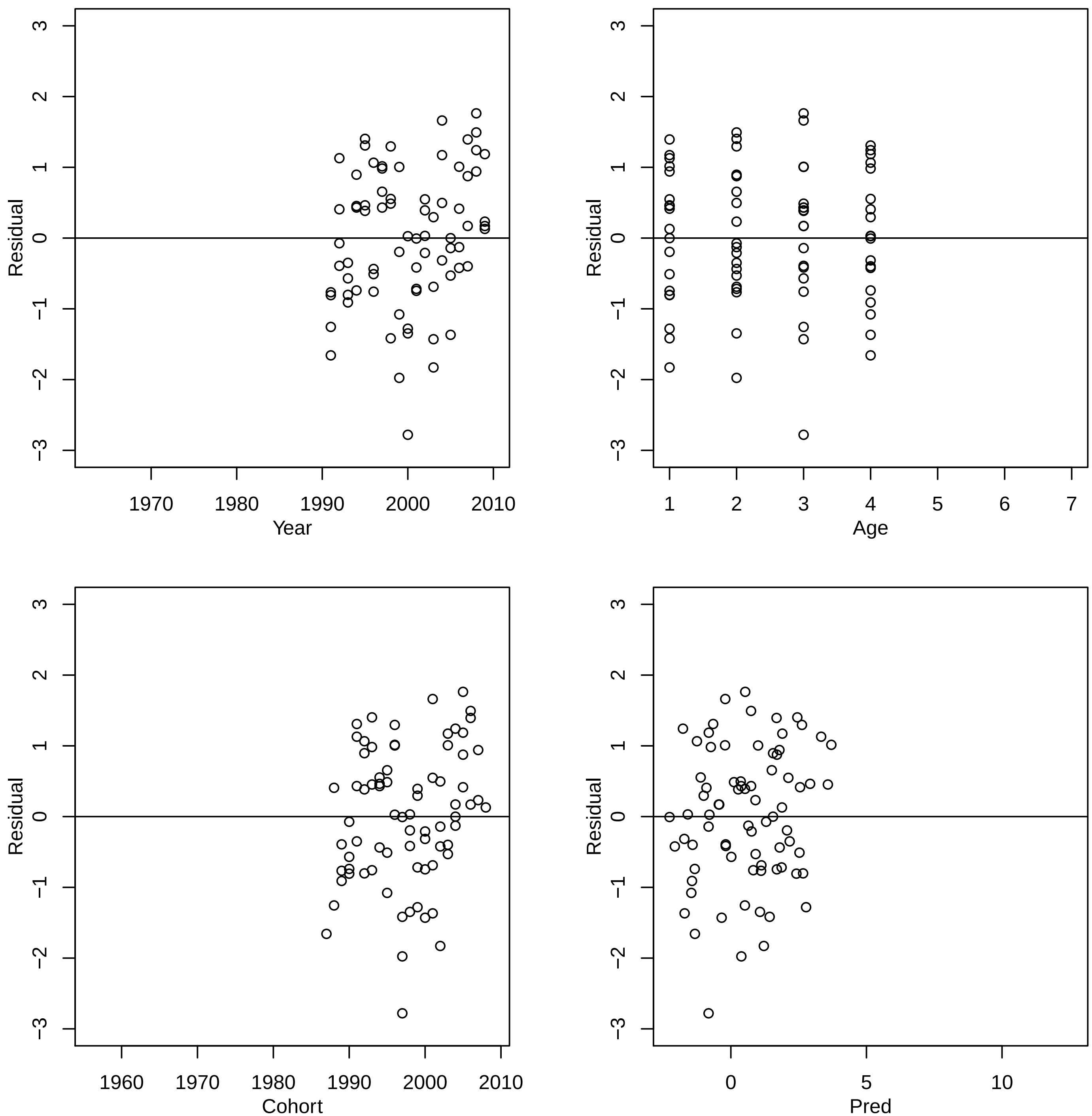

Age = 1 , fleet = 2
Age = 2 , fleet = 2
Age = 3 , fleet = 2
Age = 4 , fleet = 2
Age = 5 , fleet = 2
obs/pred
Year
1970
1980
1990
2000
2010
This plot is intentionally left blank
This plot is intentionally left blank
This plot is intentionally left blank

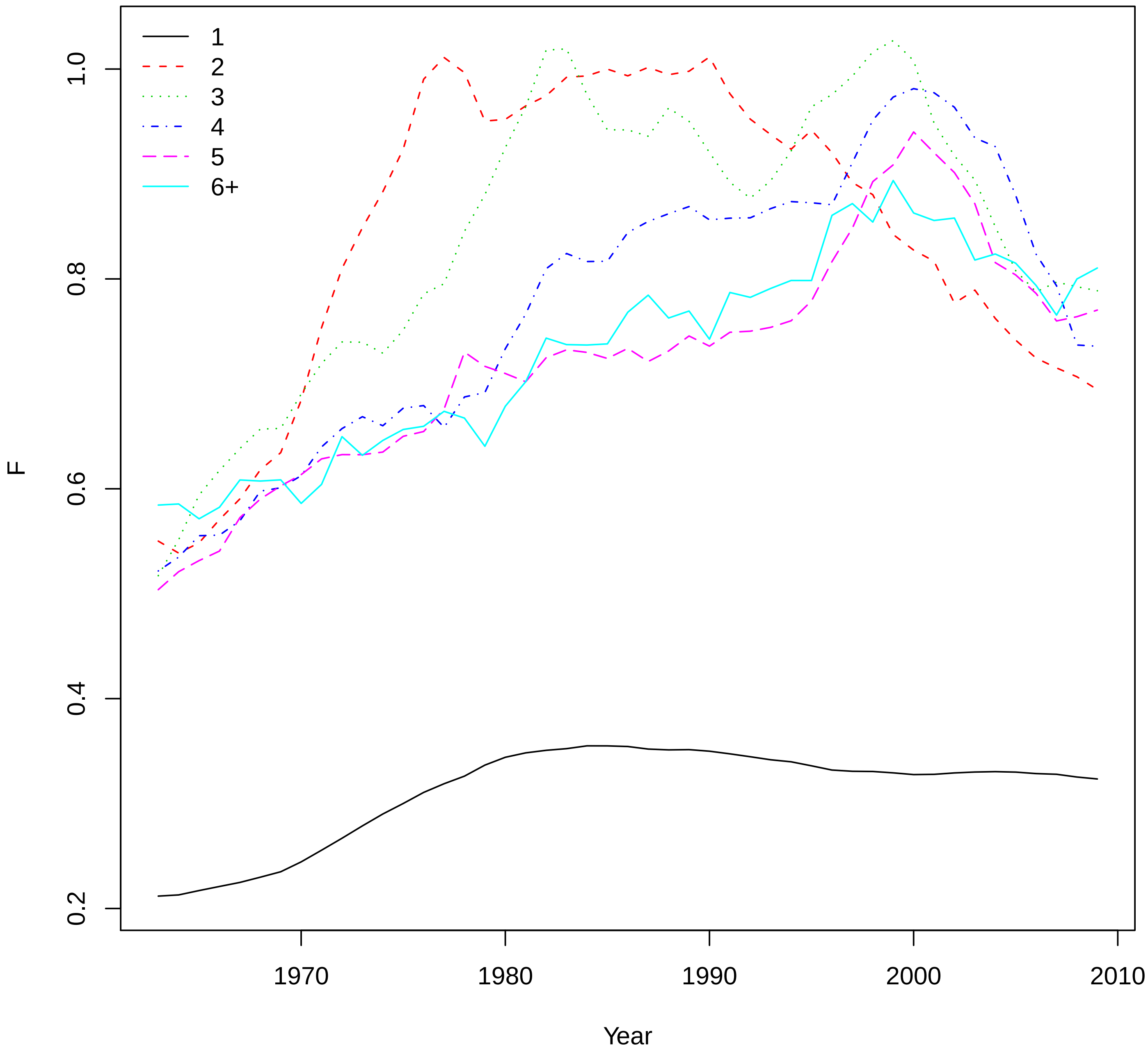

Age = 1 , fleet = 3
Age = 2 , fleet = 3
Age = 3 , fleet = 3
Age = 4 , fleet = 3
obs/pred
Year
1970
1980
1990
2000
2010
0
10
30
50
5
10
15
20
0.5
1.5
2.5
3.5
0.1
0.3
0.5
0.7
This plot is intentionally left blank
This plot is intentionally left blank
This plot is intentionally left blank
This plot is intentionally left blank


Residual
Year
Age
Cohort
Pred
1970
1980
1990
2000
2010
1
2
3
4
5
6
7
1960
0
5
10
3
2
1
0
−1
−2
−3

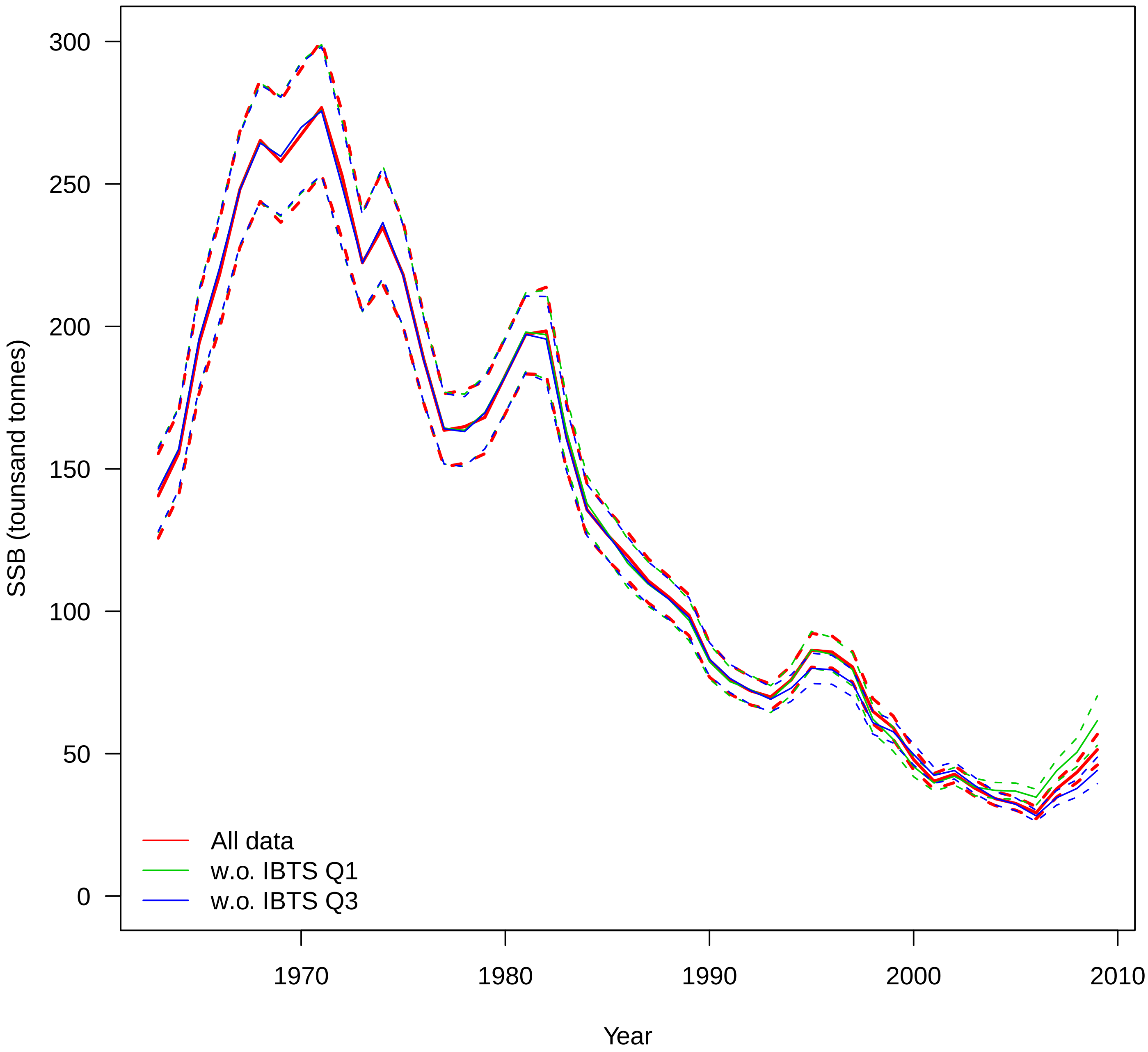


Residual
Year
1970
1980
1990
2000
2010
Age
1
2
3
4
5
6
7
Cohort
1960
Pred
0
5
10
3
2
1
0
−1
−2
−3

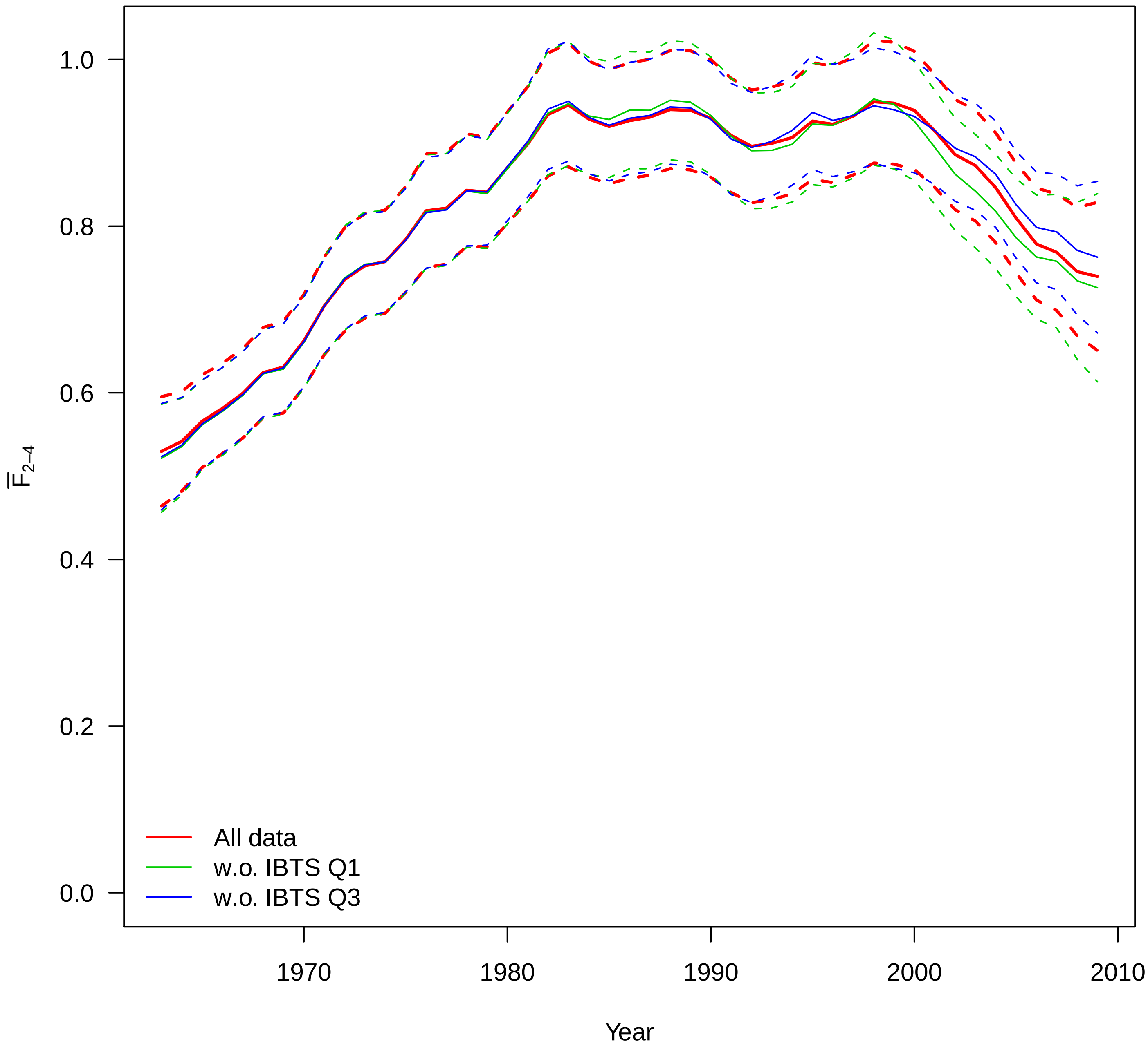

Residual
Year
1970
1980
1990
2000
2010
Age
1
2
3
4
5
6
7
Cohort
1960
Pred
0
5
10



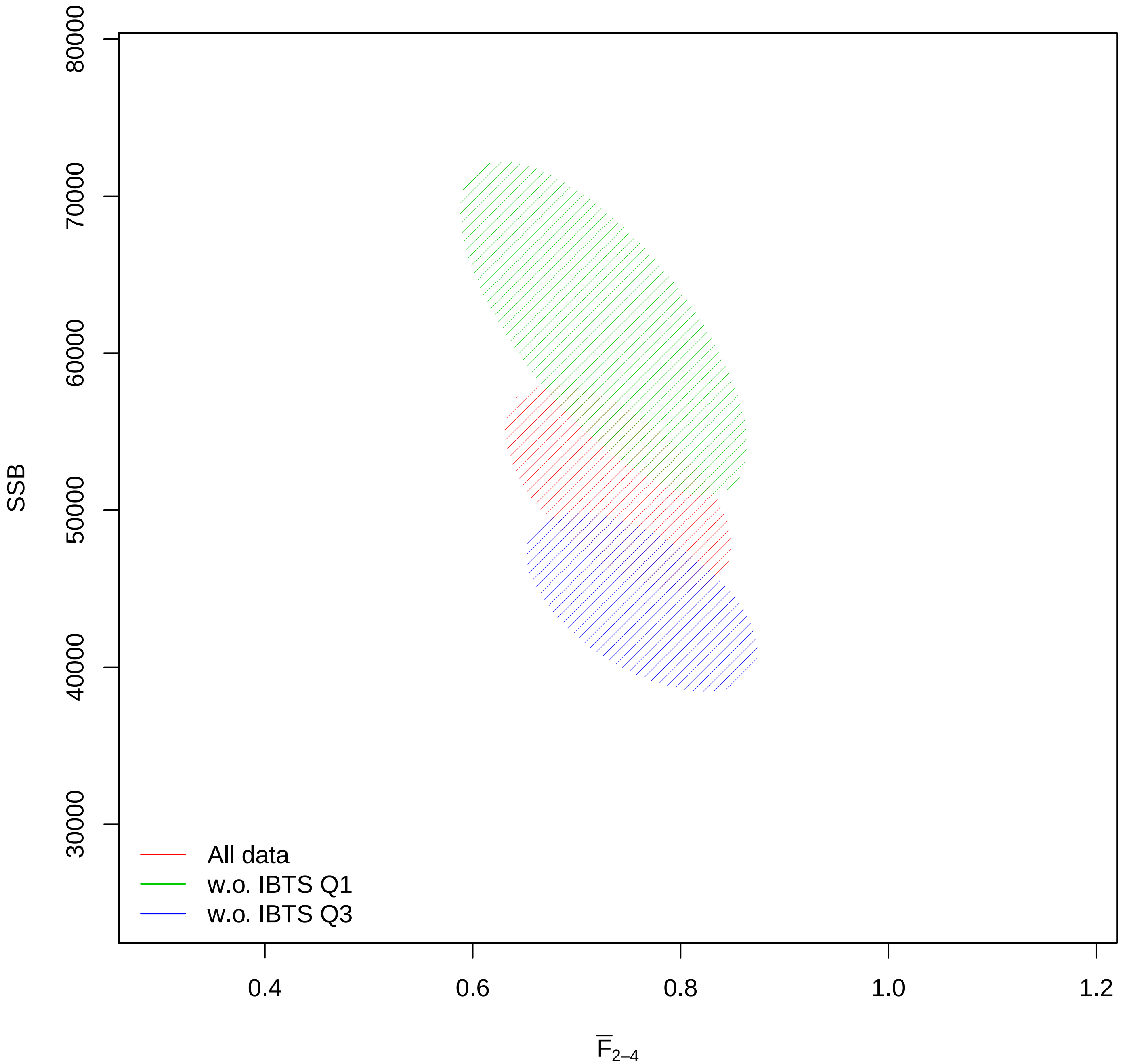
1
2
3
4
5
6+
F
0.2
0.4
0.6
0.8
1.0
1970
1980
1990
2000
2010
Year



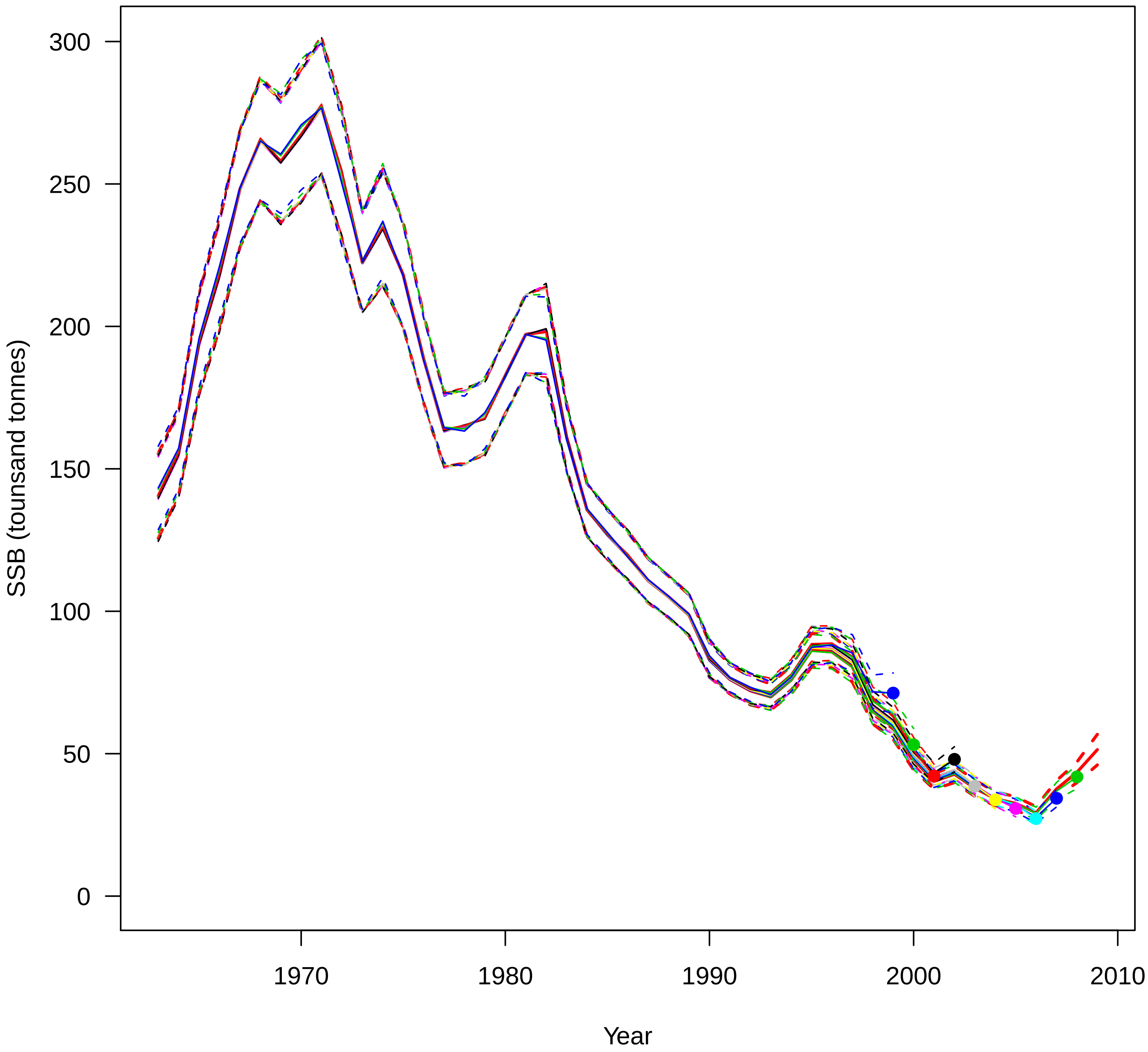
SSB (tounsand tonnes)
300
250
200
150
100
50
0
1970
1980
1990
2000
2010
Year
All data
w.o. IBTS Q1
w.o. IBTS Q3

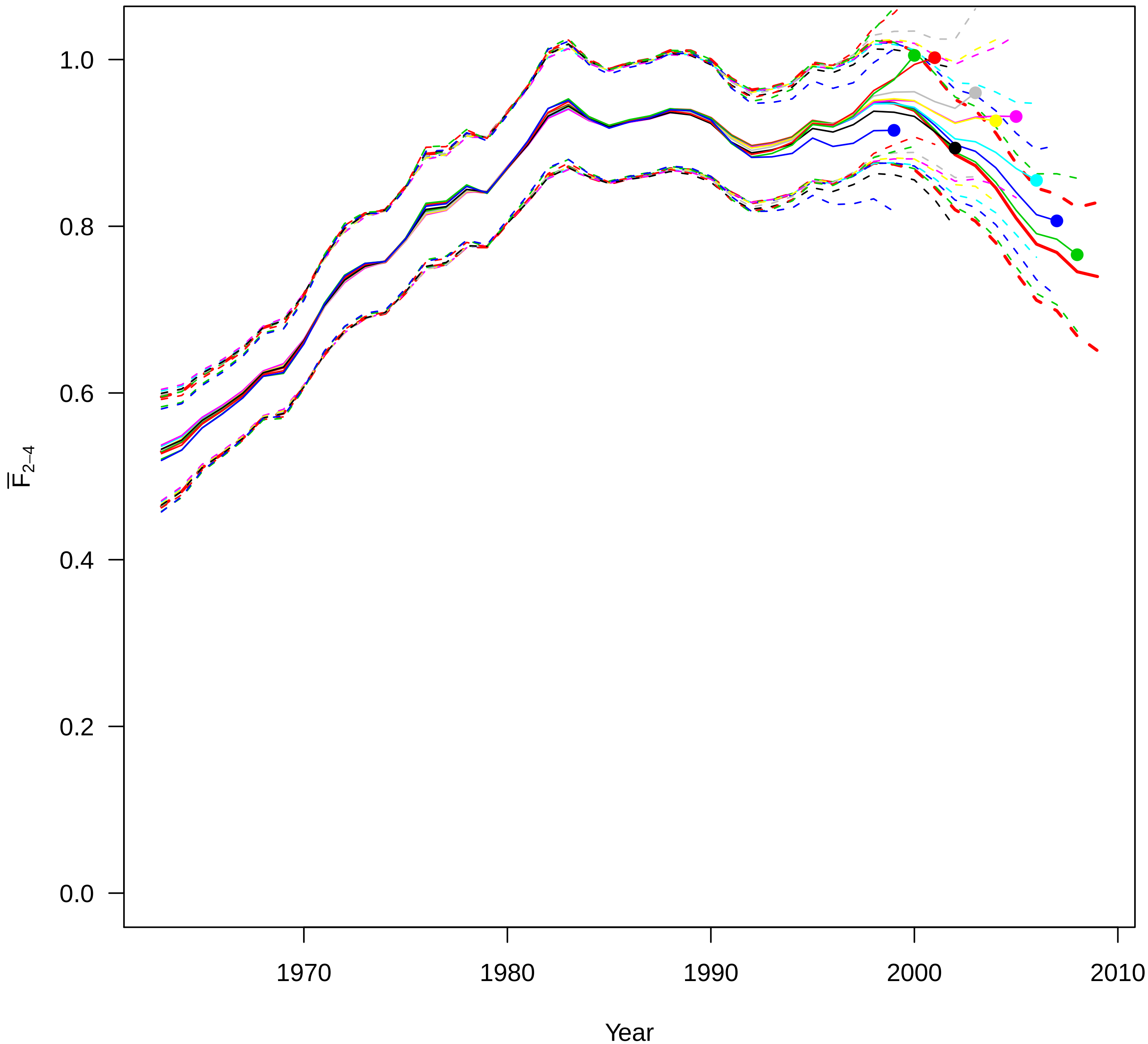


1.0
0.8
0.6
0.4
0.2
0.0
$\overline{F}_{2-4}$
1970
1980
1990
2000
2010
Year
All data
w.o. IBTS Q1
w.o. IBTS Q3



SSB
80000
70000
60000
50000
40000
30000
0.4
0.6
0.8
1.0
1.2
$\bar{F}_{2-4}$
All data
w.o. IBTS Q1
w.o. IBTS Q3


300
250
200
150
100
50
0
SSB (tounsand tonnes)
1970
1980
1990
2000
2010
Year


1.0
0.8
0.6
0.4
0.2
0.0
$\overline{F}_{2-4}$
1970
1980
1990
2000
2010
Year

Exercises

Exercise 1: Unallocated mortality for North Sea Cod

- One of the more controversial issues for North Sea Cod is the estimation of the so-called unallocated mortality (could be black landings, wrong natural mortality, or something else)
- It enters the catch equation as:

$$\log(C_{a,y}S_{a,y}) = \log\left(\frac{F_{a,y}}{Z_{a,y}}(1-e^{-Z_{a,y}})N_{a,y}\right) + \varepsilon_{a,y}$$

- A few different options can be imagined
  - Use total landings as reported

$$S_{a,y} = 1$$

  - Separate scaling each year

$$S_{a,y} = \begin{cases} 1, & y < 2003 \\ \tau_y, & y \geq 2003 \end{cases}$$

– Separate scaling each year, and in three age classes

$$S_{a,y} = \begin{cases} 1, & y < 2003 \\ \tau_y^{(1)}, & y \geq 2003 \ \& \ a = 1 \\ \tau_y^{(2)}, & y \geq 2003 \ \& \ a = 2 \\ \tau_y^{(3+)}, & y \geq 2003 \ \& \ a \geq 3 \end{cases}$$

- Discuss what these different options means in terms of the fishery
- Browse the on-line interface. Try to change something (some catches, some weights, or some options) and see what happens when you run.
- Try to run these different configurations via the on-line interface and carry out the statistical significant tests
- What option is the most appropriate?