

An Application of the Generalized Additive Model to Groundfish Survey Data with Atlantic Cod off the Northeast Coast of the United States as an Example

Loretta O'Brien and Paul Rago
NEFSC Woods Hole Laboratory
Woods Hole, Massachusetts USA

Introduction

Data collected on annual groundfish surveys provide information on the spatial distribution of species and the environmental conditions at the time of capture. One of the common features of survey data is a high degree of measurement error. Sampling designs attempt to reduce such error by appropriate stratification but environmental covariates introduce additional variability. Typically, analyses of these data have been done using generalized linear models (GLMs), often with the addition of covariates. Assumptions regarding linearity of the response variable (e.g. abundance), however, are often difficult to satisfy. For example, if fish distribution is governed by thermal preferences, then the expected distribution of fish along a thermal gradient will likely be unimodal, implying a nonlinear response. Difficulties can arise in the a priori specification of such nonlinear responses. One technique that can assist in the identification of appropriate nonlinear response functions is the generalized additive model (GAM). GAMs were first proposed by Hastie and Tibshirani (1986) and some of the first applications to fisheries survey data have been by Swartzman *et al.* (1992, 1994, 1995). A GAM can be utilized as a predictive model or as an exploratory method to suggest possible transformations of the data or appropriate parametric models such as a GLM. In this paper we apply the GAM to cod captured on the NEFSC trawl survey and suggest additional extensions of the model. Our orientation is tutorial with detailed guidance on application of the GAM in S-Plus.

The GAM was applied to catch, in numbers, of Atlantic cod to explore spatial trends in distribution influenced by latitude, longitude, depth and temperature. A step-wise GAM was performed to determine the best fitting model prior to applying the final GAM to the entire data set. Abundance indices of stratified mean number-per-tow and the associated variances were estimated for both the sample and fitted catch numbers.

The procedures for employing the GAM presented in this workbook are essentially 'work in progress'. The analysis can be extended and improved in several ways, for example, by transforming the predictor variables, employing bootstrapping to estimate variability, investigating the interaction of the predictor variables, or adding other predictor variables such as bottom sediment type, age class, or length groups.

Model Description and Application

A GAM is a nonparametric analog to a GLM and can be described as:

$$Y = \alpha + \sum_{j=1}^n f_j(X_j) + \epsilon$$

where the usual linear function of a covariate, $\beta_j X_j$ is replaced with f_j , an unspecified smooth function. As in the GLM it is necessary to specify the underlying error distribution of the model and the link function which relates the response variable to the predictors. The error distribution used for this application of the GAM is a Poisson, which is appropriate for describing random occurrences and count data (Zar, 1984; Sokal and Rohlf, 1981). Other distributions available in S-PLUS for the GAM include gaussian, binomial and Gamma. The link function is the log of the response variable, catch in numbers. The candidate predictor variables are the spatial variables latitude and longitude, and the environmental variables depth and bottom temperature.

In addition, the GAM requires specification of the smooth function using a scatterplot smoother such as loess (a locally weighted regression smoother), running mean, or a smooth spline. The scatterplot

smoother used in this application of the GAM is the cubic B-spline. The degree of smoothing in a scatterplot smoother, for example in a loess, is controlled by the span, which is the proportion of points contained in each neighborhood (the set of x values within a defined distance to x_j). The resulting 'smooth' characterizes the trend of the response variable as a function of the predictor variables. In S-PLUS the smooth functions of the GAM are solved by an iterative process of smoothing the partial residuals called the Gauss-Seidel iterative method or backfitting (Hastie 1992). The algorithm separates the parametric from the nonparametric part of the fit, and fits the parametric part using weighted linear least squares within the backfitting algorithm.

As an example, in a multiple predictor model with 2 variables, given an estimate $\hat{f}_1(x_1)$, $f_2(x_2)$ is estimated by smoothing the residual of $Y - \hat{f}_1(x_1)$ on x_2 . With the estimate $\hat{f}_2(x_2)$, an improved estimate of $\hat{f}_1(x_1)$ is obtained by smoothing $Y - \hat{f}_2(x_2)$ on x_1 . Smoothing is continued until $Y - \hat{f}_1(x_1)$ on x_2 is $\hat{f}_2(x_2)$ and $Y - \hat{f}_2(x_2)$ on x_1 is $\hat{f}_1(x_1)$ (Hastie and Tibshirani, 1986).

The fitting of the GAM is an iterative looping process involving the scatterplot smooth, the backfitting algorithm, and the local scoring algorithm, a generalization of the Fisher scoring procedure in a GLM. Each iteration of the local scoring algorithm produces a new working response and weights that are directed back to the backfitting algorithm which produces a new additive predictor using the scatterplot smoother (Hastie, 1992; Hastie and Tibshirani, 1986; Stat. Sci., 1993; Swartzman *et al.*, 1992).

A step-wise GAM is performed to determine the best fitting model based on the criteria of the lowest Akaike Information Criterion (AIC) test statistic. AIC is a function of both the log likelihood function and the effective number of parameters being estimated. The AIC in the step-wise GAM (Hastie, 1992) is calculated as:

$$AIC = D + 2 \text{ df } \phi,$$

where D = Deviance (residual sums of squares),
 df = effective degrees of freedom, and
 ϕ = dispersion parameter (variance).

The model with the lowest AIC is considered to have the best number of parameters to include in the final model. The deviance estimated in the model, analogous to the residual sums of squares, is a measure of the fit of the model. A pseudo coefficient of determination, R^2 , is estimated as 1.0 minus the ratio of the deviance of the model to the deviance of the null model (Swartzman *et al.*, 1992). The effect of the environmental variables alone on abundance can be determined by running a second model that does not include the spatial variables and comparing the R^2 values for the two models. The effect of the spatial trend information can also be examined by comparing the mean abundance and variance estimates (Cochran, 1977) of the sample and the fitted catch numbers.

Data description

The example data set is from the Northeast Fisheries Science Center (NEFSC) stratified random bottom trawl research surveys (Azarovitz, 1981). Catch of Atlantic cod, in numbers, was obtained from the autumn surveys from 1963–94. Stations where cod were not caught, i.e. 'zero tows', were also included. The area used in the analysis was delimited by the range of occurrence of cod within the time series. Catch numbers were adjusted for vessel and gear differences (NEFSC, 1991) and stations without a bottom temperature observation were deleted.

To run the GAM procedure on a different data set, create a file containing data for all years in the time series. The minimum fields needed are year, stratum, tow, catch in numbers, bottom temperature, depth, latitude, longitude, stratum area and counter (defined below). There can be no missing cells for any of the fields listed above. For stations where there is no catch for the species of interest, the field must be zero filled. Determine the minimum and maximum range of the species of interest in the survey. These latitude and longitude positions will be the boundaries for the GAM analysis. Exclude any records from the data file that are outside of the species' range but be sure to include the 'zero catch' tows that occur within the range. If the user has different environmental variables other than temperature and depth, those fields can be substituted or added. Latitude and longitude need to be recorded in decimal degrees, i.e. 45 degrees 35 minutes = 45.5833. Area is the size of a stratum and will be used in the estimation of the stratified mean. The 'counter' field is filled with '1' on all records, and will be also used in the estimation of the

stratified mean to count the number of tows in a stratum. Other fields, such as time, day, month, sex code, and catch weight can be on the record and these could have missing values, to be designated by a '.'.

Bookkeeping

S-PLUS is the software needed to run this analysis. Install S-PLUS for windows on your computer before proceeding. All of the DOS files referred to in the workbook will be provided to you on a diskette to be copied into your own working directory.

1) Creating working files and directories

Once S-PLUS is loaded, working directories and files must be created. Create a subdirectory for this specific project, for example *codgam*. The subdirectory would then be *c:\spluswin\home\codgam*.

Under the *codgam* subdirectory create another subdirectory called *_data* i.e. *c:\spluswin\home\codgam_data*. The subdirectory *_data* is where all S-PLUS working files will be stored. Also, in *codgam*, create a file called *_first* and type in the following line:

```
attach ("c:\\spluswin\\home\\codgam\\_data\\", where=1)
```

The *_first* file will direct all S-PLUS work files to the subdirectory *_data*. Note the double backslash, \\, this is required to run properly in S-PLUS.

2) Creating an icon (in Windows 3.1)

Use windows utilities to create an icon that will be specific for this project. Click on FILE, then click on PROPERTIES and put in a name for the icon, like *codgam* or *nafo* or *cod*, etc. Type in the directory and executable file for S-PLUS as *c:\spluswin\cmd\splus.exe* and type in the working directory as *c:\spluswin\home\yourdir*, where *yourdir* would be the subdirectory chosen above, i.e. *codgam*, *nafo*, *cod*. If you don't like the icons that are available, click on CHANGE ICON and follow instructions.

3) Going from S-PLUS to DOS : DOS to S-PLUS

dos() brings up the DOS window in the default subdirectory.
exit closes the DOS window, brings back the S-PLUS screen.

4) Save screen output/objects to a file

An object in S-PLUS is any variable the user creates, i.e. *yvec <- c(1,2,3)* creates the object *yvec*, which is a vector of 3 numbers. Data read into S-PLUS also becomes an object:

- 1) print the object to the screen by typing the object name at the prompt : *> yvec*
- 2) highlight the output by using the left mouse button and dragging to the end of the object output
- 3) click EDIT in the menu bar, click COPY
- 4) Bring up the NOTEPAD window
- 5) click EDIT
- 6) click PASTE
- 7) edit the file as desired, then exit with SAVE AS, and save in your working subdirectory
- 8) return to S-PLUS
- 9) to print the file, go to DOS and print. (If you print the file from Notepad the first column is not printed – this is a 'bug' in Notepad.)

5) Reading in data files

Data can be read in as either an ASCII file or as an Excel spreadsheet file.

ASCII File. The catch data file, for example, *cod6394*, consists of numbers only, with blank space as a delimiter between variables.

8405	1010	1	1	1	1	84	946	40.2667	...
8405	1010	2	1	1	1	84	840	40.2167	...
8405	1010	3	1	1	1	84	247	39.8833	...
8405	1010	4	1	1	1	84	234	039.9833	...

Edit the file readdat.txt, provided to you at the beginning of the Workshop, to reflect the appropriate object/file names and column names. The columns can be in a different sequence than presented but if the same column names are used it will save time editing other files used further on in the analysis. The following is a copy of the file readdat.txt with '#' indicating comments.

```
#readdat.txt
#/c:/spluswin/home/codgam
#read in ascii file of cod data as a data frame
# data has no missing bottemp, catchnum, depth, latitude or longitude
cod6394<-read.table("c:\\spluswin\\home\\codgam\\cod6394.dat",na.strings='.',
col.names=c("cruise","stratum","tow","counter",
"haul","gearcond","year","time","lat","lon","depth","bottemp",
"svspp","sex","catchnum","catchwt","area"))
```

In S-PLUS type:

```
source("readdat.txt")
```

This command executes the S-PLUS code in the DOS file, readdat.txt, and creates the object cod6394. The object is the data set within S-PLUS. At the prompt, type the object name and the data will appear on the screen. Missing data will be designated as 'NA'. To abort scrolling to the screen, hit control C.

Reading in other files, such as a coastline, or fathom lines is done in the same way, by editing the files names and column names. The file readline.txt, provided to you, is an example of reading in a coastline or fathom line file, where longitude and latitude are simply labelled 'x' and 'y'. The creation of a coastline and fathom line file is optional. The data file would consist of two fields on each record, longitude and latitude. The data needs to be in decimal degrees. Use a line NA NA to end a polygon, for example, an island or fathom line. An example file looks like the following:

-64.242615	45.400002
-64.313263	45.396152
-64.448929	45.400002
NA	NA
-64.516998	45.000002
.	.
.	.

EXCEL File. The file will have the column names in the first row and the data will follow to the end of the file.

cruise	stratum	tow	counter	haul	gearcond	year	time	lat
8405	1010	1	1	1	1	84	946	40.26...
8405	1010	2	1	1	1	84	840	40.21.....

In S-PLUS, click FILE, then IMPORT, then OTHER FILES. Provide the object name, in this example, it is cod6394. Click OK, then provide the appropriate file name, cod6394.dat. In the spreadsheet grabber just click on DATA AREA and choose the default, which is column names in row 1 and data in the other rows. After the file has been imported, check that the object has been created by typing the S-PLUS command objects(). Contents of the object can be viewed by typing the object name, 'cod6394', without the parentheses. Missing data will be designated by a period, '.'.

Data Analysis

1) Step-wise GAM

A preliminary step-wise GAM is performed to determine the best fitting model. If the time series is relatively long, several representative years can be analyzed. The step.gam procedure needs a

GAM object initially to start the procedure, therefore the objects, `form2` and `pre63` are created prior to the `step.gam` procedure.

Edit `step3yr.txt` to reflect the appropriate objects/files. The `#` denotes a comment.

```
#step3yr.txt
#c:/spluswin/home/codgam
#stepwise gam for the beginning, middle year and end year of series
# set up formula for use in stepwise gam
form2 <- formula(catchnum ~s(depth) + s(bottemp) + lat + lon)

attach(cod6394)
#create single year objects
cod63<-cod6394[year==63,]
cod78<-cod6394[year==78,]
cod94<-cod6394[year==94,]
detach("cod6394")

#stepwise gam for 63
attach(cod63)
pre63<-gam(form2,family=poisson,na.action=na.omit)
cat("stepwise results for 1963 cod\n")
step63<-step.gam(pre63, scope=list(
  "depth" = ~ 1 + depth + lo(depth) + s(depth),
  "bottemp" = ~ 1 + bottemp + lo(bottemp) + s(bottemp),
  "lat" = ~ 1 + lat,
  "lon" = ~ 1 + lon ),
  trace=T)
detach("cod63")

#step-wise gam for 78
cat("stepwise results for 1978 cod\n")
attach(cod78)
pre78<-gam(form2,family=poisson,na.action=na.omit).....
```

In S-PLUS, execute the source file:

```
source("step3yr.txt")
```

The output that comes to the screen can be copied into a file using Notepad.

```
stepwise results for 1963 cod
Start: catchnum ~ s(depth) + s(bottemp) + lat + lon; AIC= 2053.563
Trial: catchnum ~ lo(depth) + s(bottemp) + lat + lon; AIC= 2062.681
Trial: catchnum ~ s(depth) + lo(bottemp) + lat + lon; AIC= 2071.988
Trial: catchnum ~ s(depth) + s(bottemp) + 1 + lon; AIC= 2231.88
Trial: catchnum ~ s(depth) + s(bottemp) + lat + 1; AIC= 2066.439
stepwise results for 1978 cod
Start: catchnum ~ s(depth) + s(bottemp) + lat + lon; AIC= 4162.384
Trial: catchnum ~ lo(depth) + s(bottemp) + lat + lon; AIC= 4288.749
Trial: catchnum ~ s(depth) + lo(bottemp) + lat + lon; AIC= 4205.318
Trial: catchnum ~ s(depth) + s(bottemp) + 1 + lon; AIC= 4672.877
Trial: catchnum ~ s(depth) + s(bottemp) + lat + 1; AIC= 5052.866
stepwise results for 1994 cod
Start: catchnum ~ s(depth) + s(bottemp) + lat + lon; AIC= 1196.169
Trial: catchnum ~ lo(depth) + s(bottemp) + lat + lon; AIC= 1166.502
Trial: catchnum ~ s(depth) + lo(bottemp) + lat + lon; AIC= 1238.74
Trial: catchnum ~ s(depth) + s(bottemp) + 1 + lon; AIC= 1365.951
```

```

Trial: catchnum ~ s(depth) + s(bottemp) + lat + 1; AIC= 1193.132
Step: catchnum ~ lo(depth) + s(bottemp) + lat + lon ; AIC= 1166.502
Trial: catchnum ~ depth + s(bottemp) + lat + lon; AIC= 1357.194
Trial: catchnum ~ lo(depth) + lo(bottemp) + lat + lon; AIC= 1217.213
Trial: catchnum ~ lo(depth) + s(bottemp) + 1 + lon; AIC= 1339.215
Trial: catchnum ~ lo(depth) + s(bottemp) + lat + 1; AIC= 1164.196
Step : catchnum ~ lo(depth) + s(bottemp) + lat ; AIC= 1164.196

Trial: catchnum ~ depth + s(bottemp) + lat + 1; AIC= 1389.329
Trial: catchnum ~ lo(depth) + lo(bottemp) + lat + 1; AIC= 1214.822
Trial: catchnum ~ lo(depth) + s(bottemp) + 1 + 1; AIC= 1527.697

```

2) Generalized Additive Model (GAM)

Single year

Edit [cod652.gam](#) to reflect appropriate objects/files.

```

#cod652.gam
#c:\spluswin\home\codgam

#gam for single year

# formula for final model
form2<-formula(catchnum~ s(depth) + s(bottemp) + lat + lon)

#create single year object
cod65<-cod6394[cod6394$year==65,]

#run gam analysis
cod652<-gam(form2,family=poisson,na.action=na.omit,data=cod65)

```

In S-PLUS, execute the source file:

```
source("cod652.gam").
```

Results can be viewed by typing the name of the generated object: cod652

```

> cod652
Call: gam(formula = form2, family = poisson, data = cod65, na.action = na.omit)
Degrees of Freedom: 189 total; 177.9434 Residual
Residual Deviance: 1807.371

```

Summary statistics can be viewed by typing: summary(cod652)

```

> summary(cod652)

Call: gam(formula = form2, family = poisson, data = cod65, na.action = na.omit)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.966353 -2.072729 -1.102021 -0.01000965 14.85034
(Dispersion Parameter for Poisson family taken to be 1)
Null Deviance: 3649.812 on 188 degrees of freedom
Residual Deviance: 1807.371 on 177.9434 degrees of freedom

```

Number of Local Scoring Iterations: 6				
DF for Terms and Chi-squares for Nonparametric Effects				
	Df	Npar Df	Npar Chisq	P(Chi)
(Intercept)	1			
s(depth)	1	3.1	72.29207	1.554312e-015
s(bottemp)	1	3.0	76.08867	2.220450e-016
lat	1			
lon	1			

To save output, highlight text and cut and paste into Notepad.

Description of GAM summary output

The final model chosen based on the lowest AIC in the stepwise GAM is the model in the object `form2`:
`form2<- formula(catchnum~ s(depth) + s(bottemp) + lat + lon).`

Deviance Residuals are the ordinary residuals, $y(i) - \hat{\mu}(i)$.

The dispersion parameter is equivalent to the variance, σ^2 , for the Gaussian distribution and 1 for the Poisson and binomial.

The null deviance is the deviance of the model with only the intercept term.

The residual deviance is the deviance of the full model, similar to the residual sum of squares in the linear model. A pseudo coefficient of determination R^2 can be estimated (1 minus Residual Deviance/Null Deviance) to measure the model fit (Swartzman *et al.*, 1992).

The model looped through 6 local scoring algorithms before converging.

Df are the parametric degrees of freedom from fitting the linear component for each smooth term.

Npar Df are the nonparametric degrees of freedom from fitting the smooth after the linear component is removed.

The Npar Chisq represents an approximate chi-squared test to evaluate the nonlinear contribution of the nonparametric terms, and the P(Chi) indicates the probability level.

Multiple years

Edit `codsum.gam` to reflect appropriate objects/files. The plus sign, +, denotes a continuation of the previous line in S-PLUS.

```
#codsum.gam
#c:/spluswin/home/codgam
#run gam using the 'by' statement, then
#create object with gam summaries from 63-94, for two models

#attach (cod6394)
form1 <- formula(catchnum ~ s(depth) + s(bottemp))
form2 <- formula(catchnum ~ s(depth) + s(bottemp) + lat + lon )

codsum1<-by(cod6394,cod6394$year,
+ function(cod6394)summary(gam(form1,family=poisson,na.action=na.omit, data=cod6394)))

codsum2<-by(cod6394,cod6394$year,
+ function(cod6394)summary(gam(form2,family=poisson,na.action=na.omit, data=cod6394)))
#detach("cod6394")
```

In S-PLUS, execute the source file:

```
source("codsum.gam")
```

This may take a while to execute depending on the number of years in the data file. When completed the objects 'codsum1' and 'codsum2' will have the summary output for the gam analysis for each year, from 1963–1994, for formula 1 and formula 2, respectively. Save the output to a file using Notepad.

Results can be viewed by typing the name of the generated object: codsum1

```
> codsum1
cod6394$year:63
Call: gam(formula = form1, family = poisson, data = cod6394, na.action = na.omit)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.581353  -2.933003  -2.039011  -0.2868468  16.37018
(Dispersion Parameter for Poisson family taken to be 1 )
Null Deviance: 2954.755 on 179 degrees of freedom
Residual Deviance: 2264.08 on 171.1238 degrees of freedom
Number of Local Scoring Iterations: 6
DF for Terms and Chi-squares for Nonparametric Effects
```

	Df	NparDf	NparChisq	P(Chi)
(Intercept)	1			
s(depth)	1	3.0	171.0131	0
s(bottemp)	1	2.9	124.1030	0

```
cod6394$year:64
Call: gam(formula = form1, family = poisson, data = cod6394, na.action = na.omit)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.891006  -2.18017   -1.284479   -0.4363303  12.34678
.....
```

3) Plotting GAM Results

Distribution Plots

The following file is set up to create plot files without a coastline or fathom lines. However, the commands are written in, as comments, to create a plot with a coastline and fathom lines (Fig.1). If the data files are available, modify pc65.win by removing the **##** comment symbol from the two lines:

```
lines(fathom,lty=2)
lines(coast).
```

Read in the coastline and fathom line files using instructions in the Bookkeeping section.

If you want black and white prints of the plots, you can set up the amount of shading using the following instructions. In the S-Plus tool bar, click on 'Options' – 'Colors'. Then click 'Copy Scheme', and give a new scheme name like 'gam'. Click on 'Modify colors', then 'Images'. Click on Box 1 under 'Feature Colors'. Make box 1 white by having Red, Green, Blue (RGB) all set at 255. Click on Box 2, click 'Insert Shades' (type in number of shades you want, e.g. 15. Click on Box 3, make it black by having RGB set at 0. Delete any boxes above Box 3. Click ok, and then click save. If you want this as the default scheme click that box.

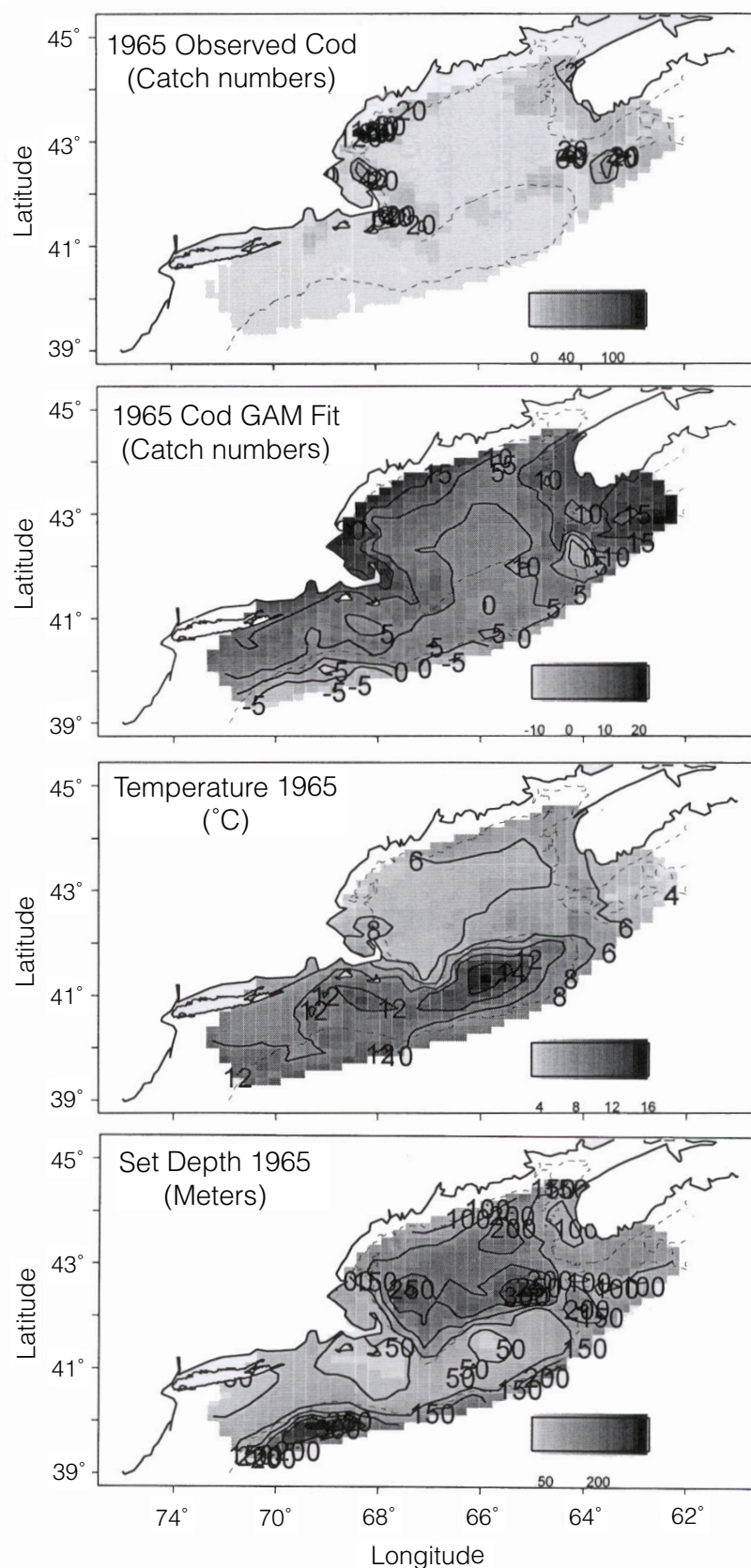


Fig. 1. Interpolated observed and GAM-fitted catch number of Atlantic cod and corresponding bottom temperature and average depth for 1965.

Edit pc65.win to reflect appropriate objects/files.

```
#pc65.win
#c:\spluswin\home\codgam
#plot file for cod observed,fitted,temperature,depth data
#without coastline or ##with coastline
#use following command to print hard copy
#win.printer()
attach(cod6394)
cod65<-cod6394[year==65,]
detach("cod6394")
attach(cod65)
#use following command for screen output
win.graph()
#sets up 4 plots per page
par(mfrow=c(2,2))

#OBSERVED PLOT
l<-interp(-lon,lat,catchnum)
image(i,xlab="longitude",ylab="latitude",xlim=c(-74.9999,-63.0000))
contour(i,add=T,nlevels=7)
title(main="1965 Observed Cod", sub="Catch Numbers")
zcat<-(catchnum)
image.legend(zcat,x=-67,y=40.1,siz=c(.7,.2),cex=.4)
##lines(fathom,lty=2)
##lines(coast)

#FITTED PLOT
#run gam before plot
cod652<-gam(form2,famliy=poisson,na.action=na.omit,data=cod65)
l<-interp(-lon,lat,cod652$fitted)
image(i,xlab="longitude",ylab="latitude",xlim=c(-74.9999,-63.0000))
contour(i,add=T,nlevels=7)
title(main="1965 Cod GAM Fit",sub="Catch Numbers")
zcat<-(cod652$fitted)
image.legend(zcat,x=-67,y=40.1,siz=c(.7,.2),cex=.4)
##lines(fathom,lty=2)
##lines(coast)

#PLOT OF TEMP
it<-interp(-lon,lat,bottemp)
image(it,xlab="longitude",ylab="latitude",xlim=c(-74.9999,-63.0000))
contour(it,add=T,nlevels=5)
title(main="Temperature 1965",sub="Degrees C")
zcat<-(bottemp)
image.legend(zcat,x=-67, y=40.1,siz=c(.7,.2),cex=.4)
##lines(fathom,lty=2)
##lines(coast)

#PLOT OF DEPTH
id<-interp(-lon,lat,depth)
image(id,xlab="longitude",ylab="latitude",xlim=c(-74.9999,-63.0000))
contour(id,add=T,nlevels=5)
title(main="Depth 1965",sub="Meters")

zcat<-(depth)
image.legend(zcat,x=-67,y=40.1,siz=c(.7,.2),cex=.4)
##lines(fathom,lty=2)
##lines(coast)
```

```
detach("cod65")
#use following line if creating hard copy
#dev.off()
```

In S-Plus, execute the source file:

```
source("pc65.win")
```

Smooth plots

To examine plots of the fitted smooth functions (Fig. 2) type in the following **bolded** commands.

```
> win.graph()           [opens a graphics window. To close type: dev.off() ]
> par(mfrow=c(2,2))      [sets up a 4 panel plot; c(1,2) would set up a 2 panel plot]
> plot.gam(cod652,ask=T) [ask=T enables interactive plotting]
Make a plot selection (or 0 to exit):
  1: plot: s(depth)
  2: plot: s(bottemp)
  3: plot: lat
  4: plot: lon
  5: plot all terms
  6: residuals on
  7: rug off
  8: se on
  9: scale (0)
 10: browser()
Selection: 5           [ A four panel plot comes up in the graphics window]
      s(depth)      s(bottemp)      lat      lon
      14.689      22.9188      7.227613      2.114185
Make a plot selection (or 0 to exit):
  1: plot: s(depth)
  2: plot: s(bottemp)
  3: plot: lat
  4: plot: lon5: plot all terms
  6: residuals on
  7: rug off
  8: se on
  9: scale (0)
 10: browser() [enables executing of other commands to change title, number of panels,etc.]
Selection: 10
browser: plot.gam(cod652, as = T)
b> title("Atlantic cod 1965")
```

```
b> 0
Make a plot selection (or 0 to exit).....
```

A two panel graph (Fig. 3) can be created by using the command **par(mfrow=c(1,2))**, then running plot.gam as described above.

4) Stratified Mean

The GAM reduces the variability in the observed indices by incorporating the influence of environmental variables. The following section describes the S-Plus code necessary to compute means and standard errors for the survey based on the observed and fitted GAM catch numbers.

Single year

Edit codmn65.txt to reflect appropriate objects/files.

```
#codmn65.txt
# calculate stratified mean for COD strata 13–25, observed and fitted
cod65<-cod6394[cod6394$year==65,]
attach(cod65)

#OBSERVED
####
#CALCULATE ARITHMETIC MEAN
codmn65<-sapply(split(catchnum,stratum),mean)
codvr65<-sapply(split(catchnum,stratum),var)
nk<-sapply(split(counter,stratum),sum)
#SET UP STRATA SET
indy<-as.numeric(names(codmn65))
indy1<-indy >=1130 & indy <=1250
strset<-as.logical(indy1)

#indy2<-indy >=1290 & indy <=1300
#strset<-as.logical(indy1+indy2)

#CREATE AREA VARIABLE, NOT REALLY A MEAN
varea<-as.vector(area)
meanarea<-sapply(split(varea,stratum),mean)
#STRATIFIED MEAN AND VARIANCE CALCULATION
cod65mo<-sum( (codmn65[strset]*meanarea[strset]))/sum(meanarea[strset])
areasum<-sum(meanarea[strset])
sqarea<-areasum*areasum

codvr65st<- (1/nk[strset])*codvr65[strset]
codvr65mo<-(1/sqarea) * sum( ((meanarea[strset]^2) * codvr65st),na.rm=T)
####
#FITTED MEAN AND VARIANCE CALCULATIONS
codmn65f<-sapply(split(cod652$fitted,stratum),mean)
codvr65f<-sapply(split(cod652$fitted,stratum),var)

cod65mf<-sum((codmn65f[strset]*meanarea[strset]))/sum(meanarea[strset])
codvr65sf<- (1/nk[strset])*codvr65f[strset]
codvr65mf<-(1/sqarea) * sum( ((meanarea[strset]^2) * codvr65sf), na.rm=T)
```

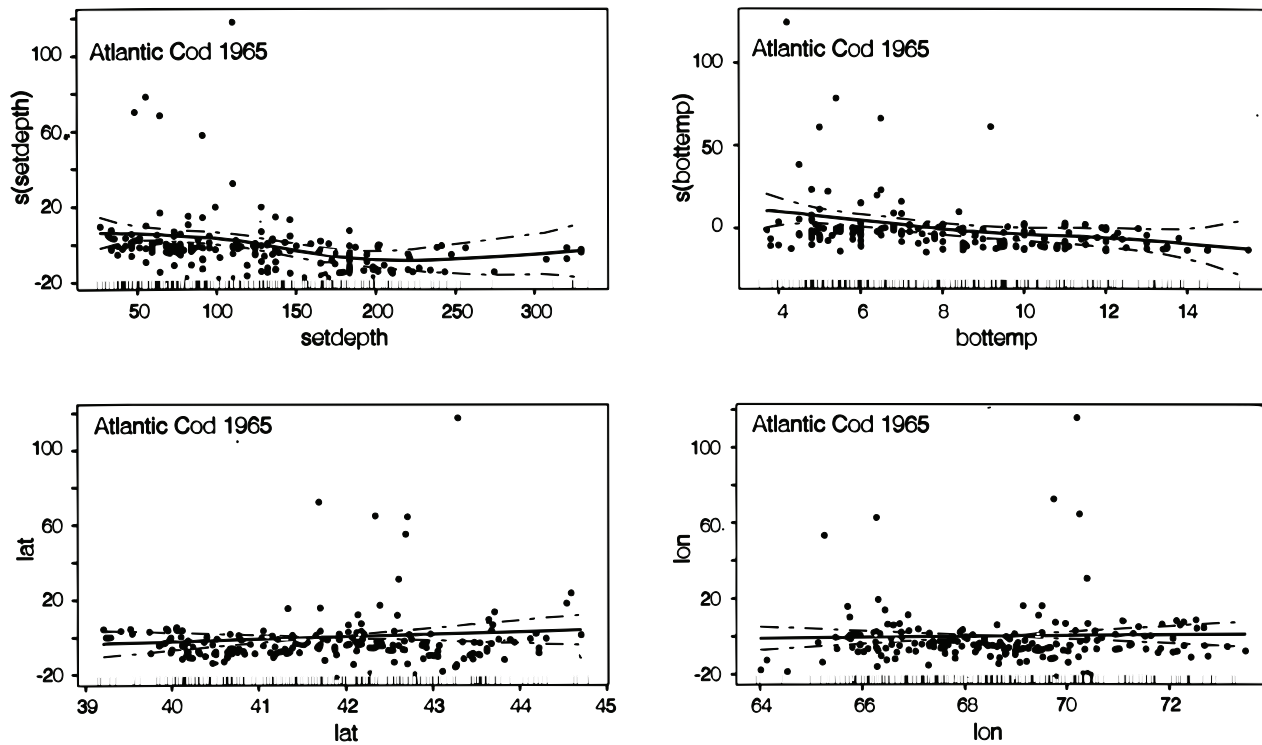


Fig 2. Scatterplot smooths of set depth, bottom temperature, latitude, and longitude for model 2 with 95% confidence intervals (dashed lines) and residuals (dots). The y-axis is scaled to zero and the rugplot on the x-axis indicates number of observations.

X/Y Plots

The following commands will create an x-y plot (Fig. 4) of interim files in the calculation of the stratified mean. The generated plot can be printed by clicking FILE, then PRINT.

```
> win.graph()
plot(codmn65f,codmn65f,type="l",xlab="",ylab="", xlim=c(-10,45),ylim=c(-10,45))
par(new=T)
plot(codmn65, codmn65, type="l",xlab=' ',ylab=' ',xlim=c(-10,45),ylim=c(-10,45))
par(new=T)
plot(codmn65,codmn65f,main=" Strata Mean – 1965 Cod",xlab="observed mean", ylab="fitted
mean",xlim=c(-10,45), ylim=c(-10,45))
> dev.off()
```

The observed mean and the fitted mean for individual strata for 1965 are plotted in Figure 4. The fitted mean is similar to the observed mean when abundance is less than 10 fish per tow. When the observed mean is greater than 20 fish per tow, however, the fitted mean is always less than the observed. This may be an indication that other variables besides temperature and depth are influencing cod distribution at higher abundance.

Discussion

The influence of the environment alone on stock distribution can be evaluated by comparing the pseudo- R^2 for models with and without latitude and longitude variables. The small differences between R^2 in the two models for cod indicate that specific location generally has little effect on distribution. The R^2 for both models increased over time (Fig. 5) which corresponds to a decrease in stock size over the same time

period (Fig. 6). The increase in pseudo- R^2 to values greater than 50% in the latter half of the time series indicates that temperature and depth exerted a greater influence on cod distribution as the stock declined. The environment becomes more significant as the stock becomes aggregated and occupies preferred temperatures and depths. This contrasts with the earlier part of the time series when there was higher abundance and the stock was distributed over a wide range of temperatures and depths. These results are consistent with the predictions of the "basin model" of MacCall (1990).

The results of the GAM, the fitted catch numbers that incorporate the effect of environment on trends in abundance, can be used to derive stratified mean indices of abundance and associated variances. These enhanced estimates are important as they are employed in the calibration of sequential population or virtual population analyses used to estimate stock biomass. The GAM may also provide a means of objectively dealing with the problem of outliers in trawl surveys. Comparison of the stratified mean abundance between the observed and the fitted GAM cod catch shows the same trend (Fig. 6) with similar mean values. The fitted mean, however, is slightly less than the observed for the majority of the time series. The estimated variance is more precise for the GAM fitted numbers (Fig. 7) than the observed numbers. Slightly lower means from the fitted GAM estimates can be attributed to the downweighting of large catches. The degree to which very large catches are downweighted is controlled by the smoothing parameter in the spline and/or loess functions. An optimal choice for the smoothing parameter is unknown but simulation modelling may provide some insights. Future simulation work should also consider the ability of the GAM to "recover" the true underlying relationship between abundance and a covariate and the consequences of interactions between covariates (e.g. temperature and depth).

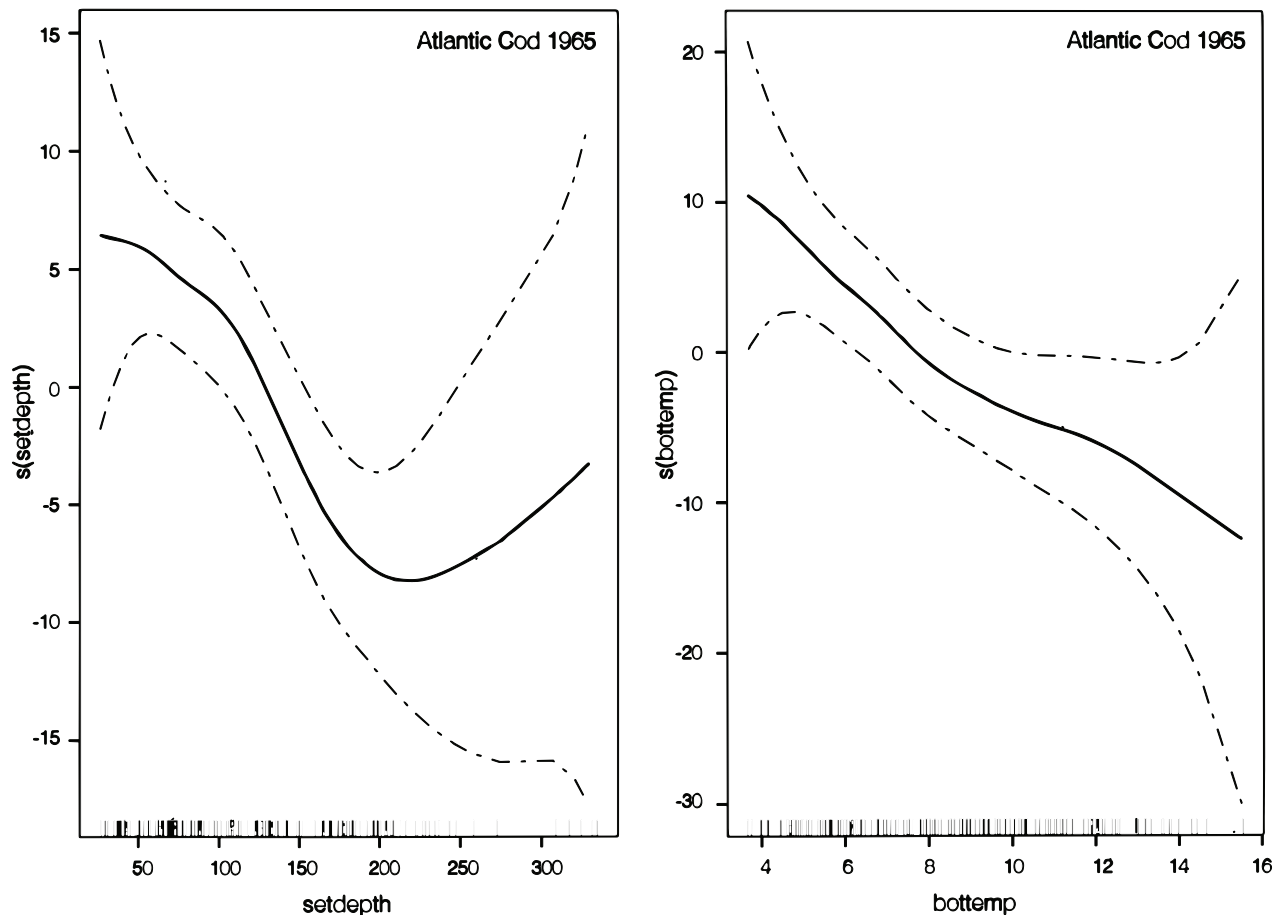


Fig. 3. Scatterplot smooths of set depth, bottom temperature, for model 2 with 95% confidence intervals (dashed lines). The y-axis is scaled to zero and the rugplot on the x-axis indicates number of observations.

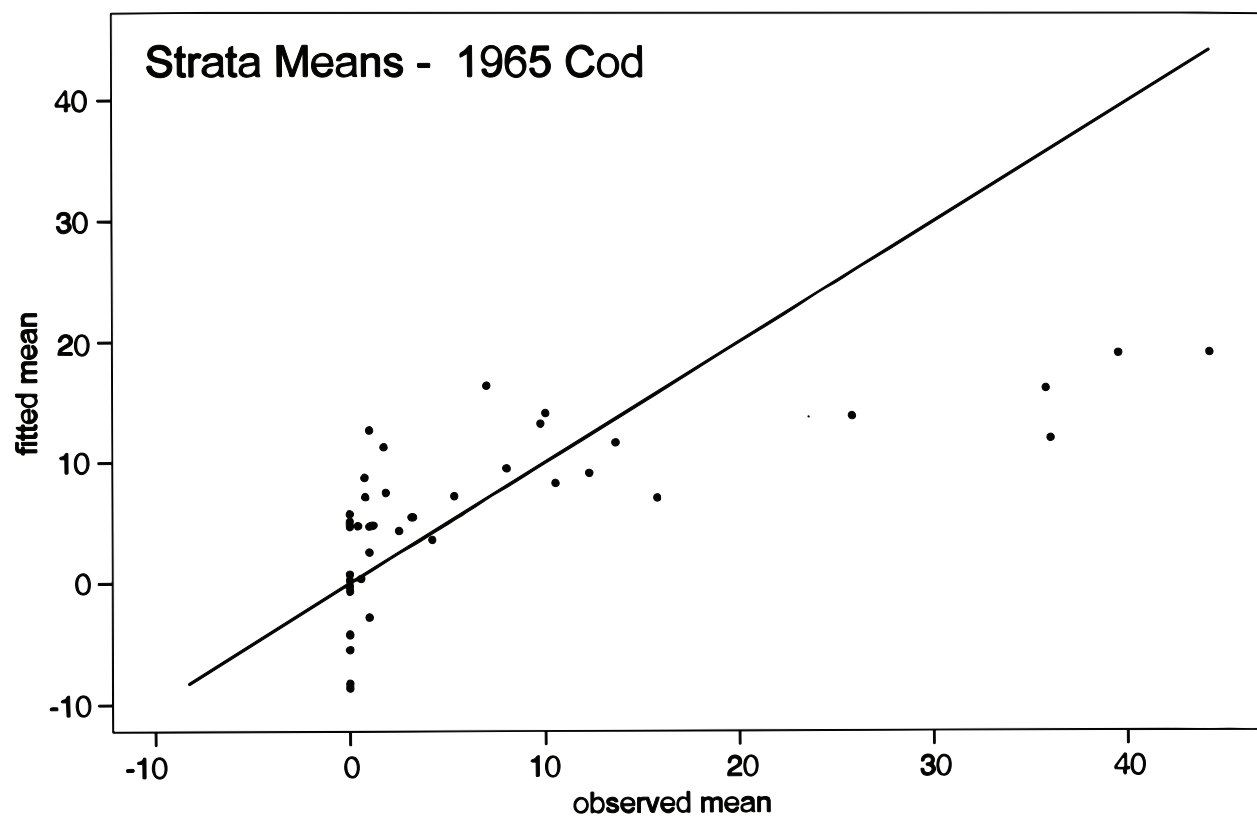


Fig. 4. The observed and fitted means for offshore strata 1–40 for Atlantic cod, 1965.

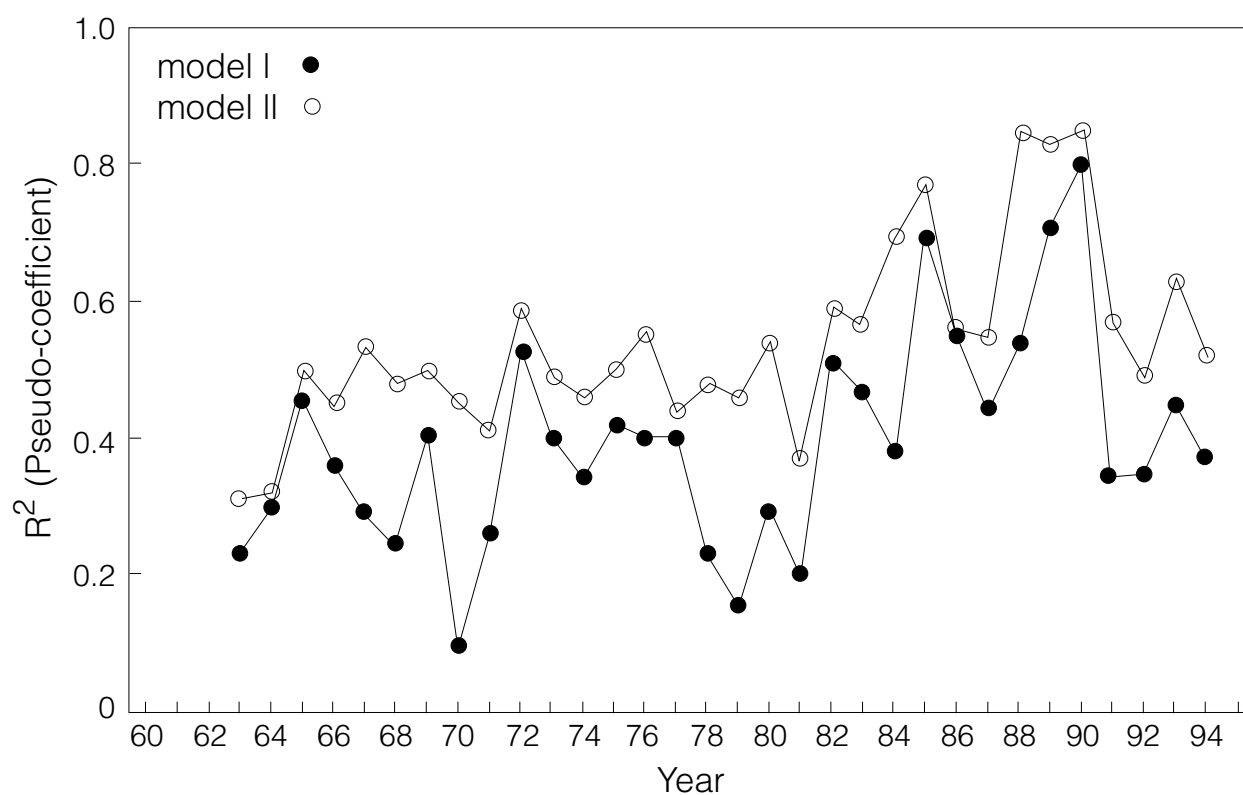


Fig. 5. The pseudo-coefficient values, R^2 , for model I (depth and temperature) and model II (depth, temperature, latitude, and longitude), 1963–94.

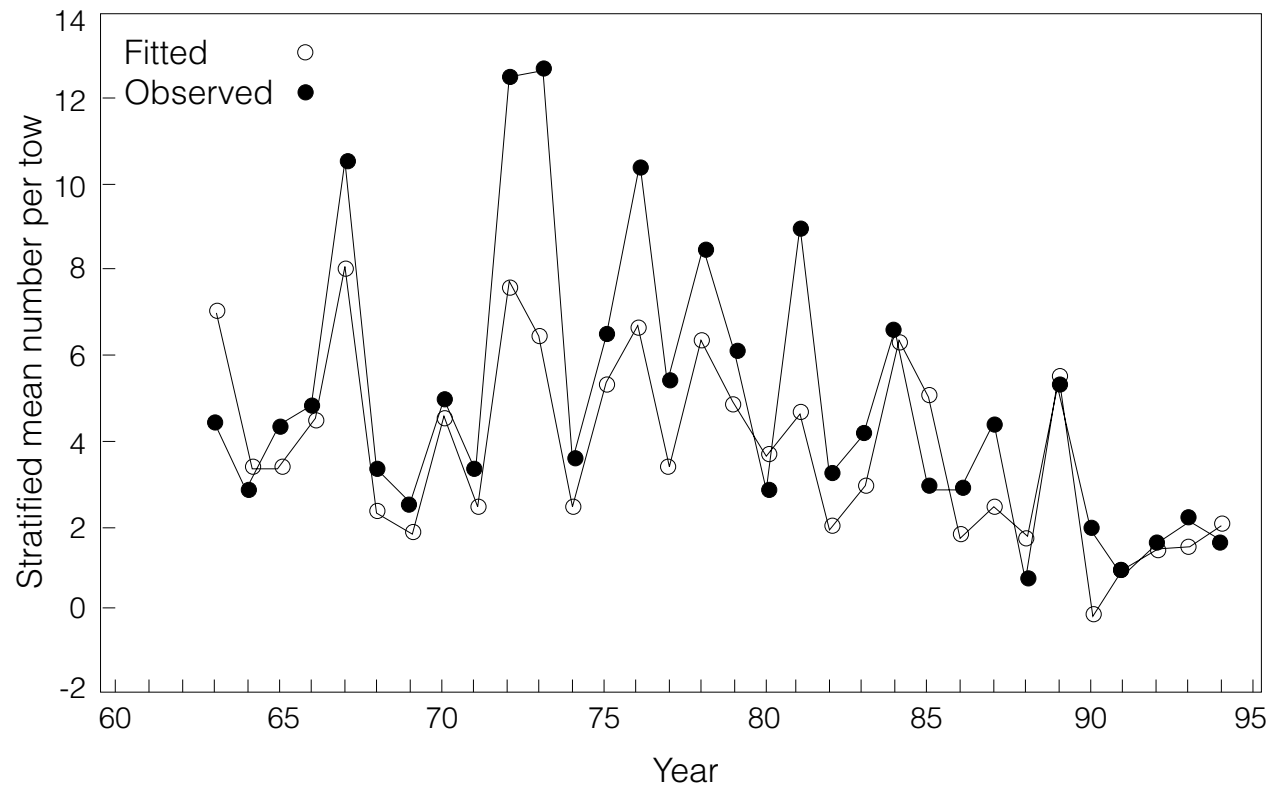


Fig. 6. Stratified mean number per tow for observed and GAM-fitted catch numbers for Atlantic cod, 1963–94.

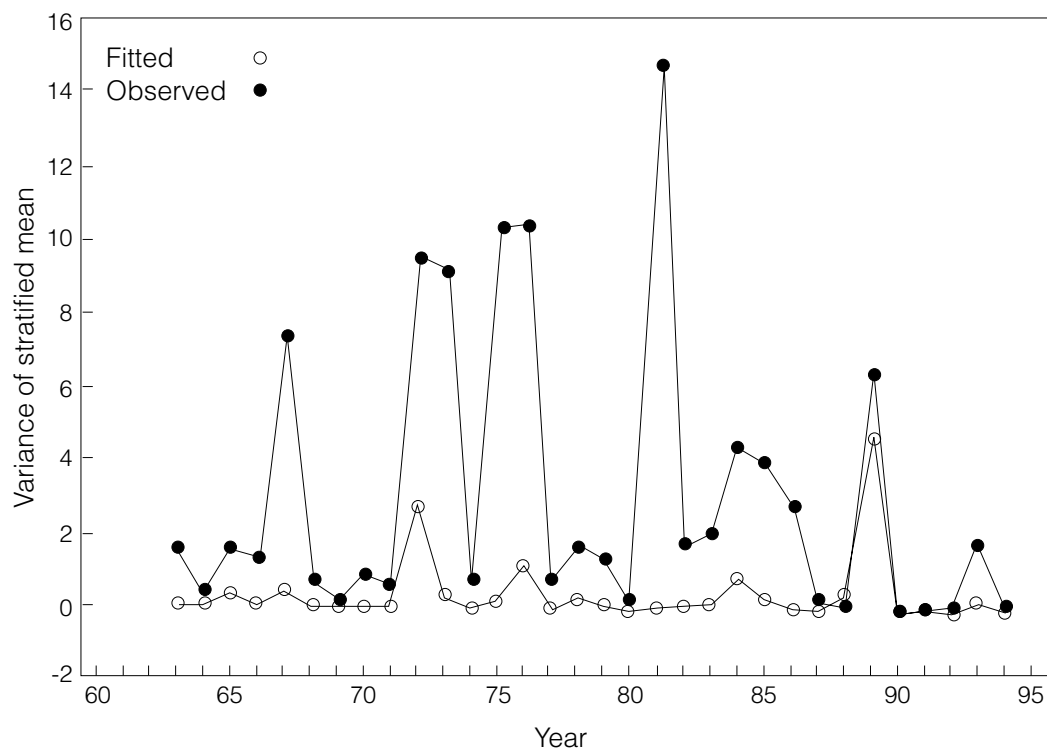


Fig. 7. Variance of the stratified mean number per tow for observed and GAM-fitted catch numbers for Atlantic cod, 1963–94.

The GAM reduces some of the problems of model mis-specification that are inherent in a linear model. The unexplained variation in the model may be due to un-modeled variables that can easily be added to the GAM. Additional oceanographic variables (e.g. salinity, tidal fronts, thermocline depth) and factors affecting foraging and aggregation, such as bottom type, spawning season, and co-occurring species and any interactions between these variables may explain more of the variance in distribution. Cross-validation of the model within years and across years could also be used to build a better model. The utility of the GAM can be expanded beyond just providing insight into distributions affected by environment. The GAM results could be used for selection and management of closed areas as stocks expand or contract in their distribution. Sensitivity analyses could be performed using the prediction capability of the GAM to forecast distributions for different scenarios of environment or recruitment.

References

- AZAROVITZ, T. R. 1981. A brief historical review of the Woods Hole Laboratory trawl survey time series. *In: Bottom trawl surveys*. Doubleday, W. G. and D. Rivard (eds). *Can. Spec. Tech. Publ. Fish. Aquat. Sci.*, **58**: 62–67.
- COCHRAN, W. J. 1977. Sampling techniques. John Wiley and Sons. New York. 428 p.
- HASTIE, T. J. 1992. Generalized additive models *In: Statistical models in* S. Chambers, J. M. and T. J. Hastie (eds). Wadsworth and Brooks, Pacific Grove.
- HASTIE, T., and R. TIBSHIRANI. 1986. Generalized additive models. Chapman and Hall, London. 335 p.
- MacCALL, A. D. 1990. Dynamic geography of marine fish populations. Univ. of Washington Press. 153 p.
- NORTHEAST FISHERIES SCIENCE CENTER. 1991. Report of the 12th NE Regional Stock Assessment Workshop (12th SAW) Spring 1991.
- SOKAL, R. R., and F. J. ROHLF. 1981. Biometry. W. H. Freeman and Company, San Francisco, 859 p.
- STATISTICAL SCIENCES, Inc. 1993. S-PLUS for Windows Reference Manual, Version 3.1 Seattle:Statistical Sciences, Inc.
- SWARTZMAN, G., C. HUANG, and S. KALUZNY. 1992. Spatial analysis of Bering Sea groundfish survey data using generalized additive models. *Can. J. Fish. Aquat. Sci.*, **49**: 1366–1378.
- SWARTZMAN, G., E. SILVERMAN, N. WILLIAMSON. 1995. Relating trends in walleye pollock (*Theragra chalcogramma*) abundance in the Bering Sea to environmental factors. *Can. J. Fish. Aquat. Sci.*, **52**: 369–380.
- SWARTZMAN, G., W. STUETZLE, K. KULMAN, and M. POWOJOWSKI. 1994. Relating the distribution of pollock schools in the Bering Sea to environmental factors. *ICES J. Mar. Sci.*, **51**: 481–492.
- ZAR, J. H. 1984. Biostatistical Analysis. Prentice-Hall, Inc. Englewood Cliffs. 718 p.

Blank