



Analysis of cod catch data from Icelandic groundfish surveys using generalized linear models

Jenný Brynjarsdóttir^{a,*}, Gunnar Stefánsson^b

^a *Science Institute, University of Iceland, Dunhaga 3, 107 Reykjavík, Iceland*

^b *Marine Research Institute, P.O. Box 1390, Skúlagata 4, 121 Reykjavík, Iceland*

Abstract

Catch data from the Icelandic groundfish surveys are analyzed using generalized linear models (GLM). The main goal is to test the effects of environmental variables on the expected cod catch and to distinguish between the gamma and log-normal distributions for the error structure. Only positive catch data are included in this work, i.e. only the positive part of a delta–gamma or delta–log-normal distribution is examined. The distributions are compared via a Kolmogorov–Smirnov goodness of fit test. Polynomials are used to describe the relationship between each environmental variable and the cod catch and their effects are tested within the GLM framework (a continuous model). Finally, an attempt is made to locate temperature fronts in the ocean by estimating the temperature gradient vector at each data point. The effect of the size of this gradient vector is then tested within in the GLM framework. A stratification model with only spatial and time effects explains 80% of the variation but that comes with a high cost of degrees of freedom. Most of the tested effects are found to be significant but the continuous model only captures 45% of the total variation. The size of the temperature gradient vector is found to be statistically significant though only a small portion of the variation in the data is explained by this term.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Groundfish surveys; Generalized linear models; Log-normal distribution; Gamma distribution

1. Introduction

Groundfish trawl surveys are commonly conducted for the purpose of obtaining an average catch per tow, to be used as an indicator of stock abundance in a stock assessment process. The single most common method for estimating means or total abundances and associated variances is probably through standard formulas

for stratified random design. Myers and Pepin (1986) were the first to use the linear regression model as a method for estimating population size in groundfish surveys. The GLM has the advantage over the stratified analysis that the underlying spatial pattern of the fish density can be modeled explicitly, an aspect ignored by the stratified analysis. Also, data from all years of the survey can be analyzed at once and data from incomplete surveys can be included. Furthermore, environmental variables as explanatory covariates can be used to separate the variation in the catch rates per tow into that due to inter-annual variation in population

* Corresponding author. Tel.: +354 5255221; fax: +354 5528911.

E-mail addresses: jennyb@raunvis.hi.is (J. Brynjarsdóttir), gunnar@hafro.is (G. Stefánsson).

size and that due to differences in the environmental variables.

In general, abundance estimates from trawl surveys have large variance estimates and often the mean and variance are directly related, i.e. larger means have larger variances associated with them. A number of statistical models have been suggested for the estimation of mean catch per tow in an attempt to account for this variability. Many of these methods use skewed probability distributions, for example the delta-distribution (Pennington, 1983) where the probability of obtaining an empty tow is modeled as a Bernoulli trial but the distribution conditioned on non-zero catches is assumed to be the log-normal or the gamma distribution. In the GLM framework, the gamma distribution has frequently been used to describe the variation of skewed marine data. This includes Stefánsson (1996) for age-disaggregated haddock catch data from the Icelandic groundfish surveys, Goñi et al. (1999) for Western Mediterranean fisheries and Ye et al. (2001) for the Kuwait driftnet fishery. Another common approach is to use the log-normal distribution by log-transforming the data. Examples of this include Myers and Pepin (1986) for the American plaice, Stefánsson (1988) for cod cpue data from the Icelandic trawler reports, Lo et al. (1992) for northern anchovy data collected by aircraft and Pennington (1996) for several marine survey data.

These two error structures for use in GLMs are related by the fact that when the coefficient of variation (CV) is small it is approximately equivalent to assume gamma distributed errors with a constant CV and to assume a constant variance (σ^2) for the log-transformation. Furthermore, a GLM analysis assuming gamma distributed errors and a GLM analysis on log-transformed data assuming normally distributed errors will usually lead to the same conclusions (McCullagh and Nelder, 1989, p. 285). In the case of the log-transformation, a direct back-transformation will result in bias, but in the present case the interest is solely in testing hypotheses, where this problem does not occur. As to how small is small enough, Atkinson (1982) concludes that these two methods of analysis should provide similar results for σ^2 as large as 0.6. When the CV is large, however, these two kinds of analysis can give different results (see, for example, Wiens, 1999). Firth (1988) gives an interesting comparison of the efficiencies of parameter estimates

using the gamma model when the errors are in fact log-normally distributed versus using the log-normal model when the errors are really gamma distributed. He concludes that the gamma model performs slightly better.

An annual groundfish survey has been conducted on the Icelandic continental shelf every March since 1985 (Pálsson et al., 1989). The primary purpose of this survey is to gather data which are used in the process of stock assessment for several demersal species (Anon., 2001). The data gathered contain numerous measurements of environmental variables which allow investigation of the relationship between such variables and catch rates in order to cast light on why the fish are caught at one place rather than another. Such analyses for the Icelandic cod (*Gadus morhua*) are the subject of this analysis and the methodology selected for this work is the generalized linear model (GLM). The emphasis here is on the variability of the survey catch and not on obtaining an index of abundance, unlike the usual analysis of survey data and any previous analysis of Icelandic marine data.

It has been suggested that the food for cod, such as capelin, may aggregate in frontal regions where cold sea meets with warmer sea, i.e. where there is a sudden change in temperature (Vilhjálmsson, 1994). Therefore it could be expected that the cod would tend to be found at such temperature fronts. In light of this it would be interesting to test the effect of temperature fronts on the catch rates of cod. Here we present a method based on locally weighted regression to estimate temperature gradients and then we test the effects of these on catch rates of cod. Incorporation of temperature fronts in the analysis of groundfish catch data has not been done before, though Sakuma and Ralston (1995) found that the spatial distributions of some late larval groundfish species off central California were dependent on a temperature front.

Steinarsson and Stefánsson (1986) fitted several probability distributions to the cod catch data from the Icelandic groundfish surveys 1985–1986 and found that among tested distributions, the gamma, log-normal and negative binomial distributions gave the best fit. In light of this, and the literature mentioned above, we examine both the gamma distribution and the normal distribution for log-transformed data as possible error distributions for the linear model (the negative binomial distribution is not considered). We conduct a compar-

ison of how closely these two distributions fit the data using a Kolmogorov–Smirnov test. Only non-empty tows are analyzed here, corresponding to the positive part of a delta–log-normal or delta–gamma analysis (Stefánsson, 1996).

2. Data

The data analyzed here are cod catch data from the Icelandic groundfish surveys 1985–2001 conducted by the Marine Research Institute (MRI) in Iceland. A detailed description of the survey design can be found in Pálsson et al. (1989), and the survey handbook (Einarsson et al., 2002) provides detailed descriptions of the data collection methodology. The survey area comprises the Icelandic continental shelf inside the 500 m depth contour (Fig. 1), which covers the fishing grounds for the most important commercial species of demersal fish in Icelandic waters. Based on biological and hydrographic considerations, the survey area is divided into two main areas, the northern and southern areas, and ten sub areas (strata). Stations are allocated among strata in direct proportion to the area of each stratum and its estimated cod density. The whole survey area is also divided into statistical rectangles (0.5° latitude and 1° longitude) which are then subdivided into

four sub-rectangles (Fig. 1). The trawl stations within a stratum are allocated to each statistical rectangle within the stratum in direct proportion to the area of the rectangle. The station positions are selected in a semi-random way, in the sense that half of the stations in each statistical rectangle were selected randomly and half of them were located by commercial fishermen in accordance with their knowledge and experience of fishing and fishing grounds. The same stations are visited every year. Commercial fishing vessels are leased every year and the fishing gear and methods are standardized as much as possible.

The data collected can be categorized into trawl station data, trawl catch data and environmental observations. The trawl station data recorded are position, time, direction and depth of the tow, distance towed and trawling speed. Biological data include the number of fish caught, length measurements, age determination from otolith samples and sex determination. The environmental data include wind strength (Beaufort scale) and direction, air-, surface- and near-bottom temperature, weather conditions, cloud cover, wave height, ice conditions and barometric pressure. This analysis includes only stations where the cod catch is non-zero and none of the environmental measurements needed for the analysis are missing, leading to a total of 7066 observations.

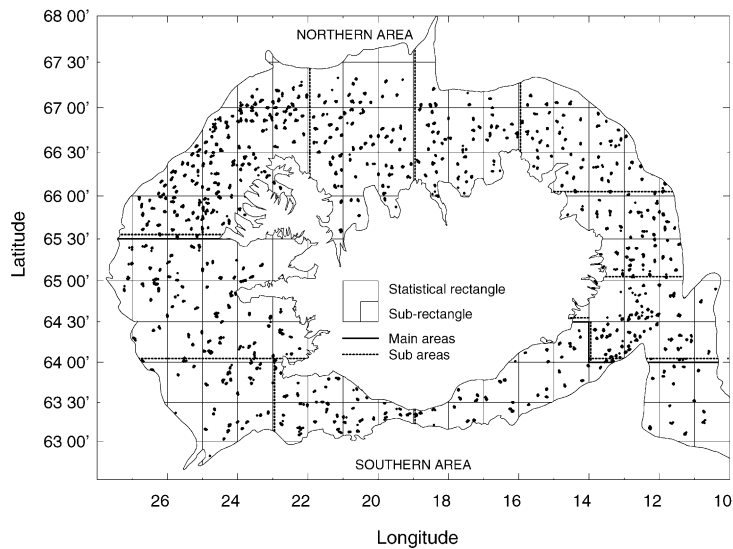


Fig. 1. Map of the survey area showing the 500 m contour line, main areas, sub areas (strata) and statistical rectangles. The points denote the (middle) locations of stations for all survey years where non-zero tows occurred.

3. Methods

The gamma and log-normal distributions share some characteristics which often make it difficult to choose between them. Both distributions have a positive probability mass only for positive values and can describe data sets for which the majority of the probability mass is at low values but there is a heavy tail to the right. They also share the same relationship between the mean and variance, i.e. the variance function:

$$\text{var}(Y) = \phi E(Y)^2 \quad (1)$$

where ϕ is a constant. This relationship differs from that for other distributions such as the normal, Poisson and the negative binomial distribution and can therefore be used to distinguish these two distributions from others. A common approach to check this relationship is to examine a plot of $\log(\text{sample variance})$ versus $\log(\text{sample mean})$ for homogeneous groups of data, see, for example, McCullagh and Nelder (1989, p. 306). If the points lie on a straight line with the slope close to 2, the gamma and log-normal distributions with fixed scale parameters can not be rejected as the true underlying distribution. Such an investigation cannot, however, distinguish between these two distributions. Data in one sub-rectangle and 1 year from the Icelandic groundfish survey can be thought to be realizations of i.i.d. variables since the environmental conditions are fairly homogeneous within a sub-rectangle. The drawback is that there are few observations for each sub-rectangle; the highest number is 7 observations, resulting in high uncertainty of the estimated means and variances for the sub-rectangles. The statistical rectangles which have up to 16 observations per rectangle are therefore also considered, but since they are four times the size of the sub-rectangles, the assumption of homogeneity is not as reliable.

A goodness-of-fit test with help of a generalized linear model is used to distinguish between the two proposed probability distributions, the gamma and log-normal distributions. Following the approach of Stefánsson and Pálsson (1997) and Stefánsson (1988), this was done by scaling the observations with the fitted values from a GLM and then performing a Kolmogorov–Smirnov test on the scaled data. Let Y_{yji} be a random variable that represents the number of cod caught in year y , sub-rectangle j and tow i . It is assumed

that either

$$Y_{yji} \sim G\left(r, \frac{\mu_{yj}}{r}\right) \quad \text{or} \\ Z_{yji} = \log(Y_{yji}) \sim N(a_{yj}, b^2) \quad (2)$$

where $N(a, b^2)$ is the normal distribution with mean a and variance b^2 and $G(r, \mu/r)$ is the gamma distribution with mean μ , variance μ^2/r and density function

$$f(y) = \frac{y^{r-1} e^{-y/\mu}}{(\mu/r)^r \Gamma(r)}, \quad y > 0 \quad (3)$$

where Γ is the Gamma function, $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$. The effects of sub-rectangles and years are assumed to be multiplicative on the original scale of number of cod and hence additive on the log scale. This leads to the log link if Y_{yji} is gamma distributed and the identity link if $\log(Y_{yji})$ is normally distributed. We fit the models:

$$\log(\mu_{yj}) = \beta_0 + \alpha_y + \beta_j + \gamma_{yj} \quad \text{and} \\ a_{yj} = \beta_0 + \alpha_y + \beta_j + \gamma_{yj} \quad (4)$$

where β_0 is the grand mean, α_y is the year effect, β_j is the spatial effect of sub-rectangles and γ_{yj} is the interaction. The error is assumed to be gamma distributed, $G(1, 1/r)$, in the first model but normally distributed, $N(0, b^2)$, in the second. For a fixed year y , the models become:

$$\log(\mu_{yj}) = \beta_0 + \beta_j \quad \text{and} \quad a_{yj} = \beta_0 + \beta_j \quad (5)$$

The goodness-of-fit test is based on the following: Firstly, a known fact is that if $X \sim G(r, \mu/r)$ then $X/\mu \sim G(r, 1/r)$. If μ_{yj} and r were known we could test whether $Y_{yji}/\mu_{yj} \sim G(r, 1/r)$ using the Kolmogorov–Smirnov test. This is done here by assuming that the fitted values $\hat{\mu}_{yj}$ and the estimated dispersion parameter $1/\hat{r}$ obtained from model (4), with gamma distributed errors, are the true parameters. Secondly, another known fact is that if $X \sim N(a, b^2)$ then $e^{X-a} \sim \text{LN}(0, b^2)$. If the true parameters were known this could be tested via the Kolmogorov–Smirnov test. This is done here by assuming that the fitted values \hat{a}_{yj} and the estimated dispersion parameter \hat{b}^2 obtained from the model (4), with normally distributed errors, are the true parameters. The D_n Kolmogorov test statistic measures the distance from the empirical and hypothesized distributions so we compare these test statistics to see which distribution better represents the data.

A temperature surface over the survey area is fitted using locally weighted regression (loess) on bottom temperature (which is measured in each station in the survey) to obtain estimates of temperature gradients. The temperature estimates of a fine grid are then used to obtain an estimate of the magnitude of the temperature gradient vector at each station, which is then tested for a significant relation to cod catch in a GLM. A grid containing the entire survey area was constructed by distributing 101 points equally over the latitude range and 101 points over the longitude range (a total of 10 201 points). A second-order loess smoother in latitude and longitude was used to obtain a bottom temperature surface over the survey area, a different one for each year. The model for a given year is an additive model:

$$T_i = g(\text{lat}_i, \text{lon}_i) + \epsilon_i \quad (6)$$

where T_i represents the bottom temperature, ϵ_i is assumed to be normally distributed with zero mean and constant variance and g is the loess smoother. The scope parameter f was set to 0.2. For each year of the survey, model (6) was fitted to a dataset, containing the recorded bottom temperatures, latitudes and longitudes for that year along with the grid points, which were given the temperature value of 0° C. A point in these datasets ($t_i, \text{lat}_i, \text{lon}_i$) was given a weight of 1 if it was a record from the survey data but a weight of 10^{-10} if it was a grid point. The effect of these weights is that they are simply multiplied by the built-in weights of the loess smoother. The grid points therefore have negligible effect on the fitted values, except at points that are far from the survey data, where the fitted temperatures are not needed anyway. On the other hand, the survey data are dominant for the fitted values at the grid points and hence the fitted values at the grid points provide a smooth surface of the temperature over the survey area for every year. Once the estimated temperature $\hat{t}_i = \hat{g}(\text{lat}_i, \text{lon}_i)$ has been obtained for the grid points for each year of the survey, the squared length of the gradient vector is estimated using:

$$\|\widehat{\nabla}g(\text{lat}_i, \text{lon}_i)\| = \sqrt{(\hat{g}(\text{lat}_{i+1}, \text{lon}_i) - \hat{g}(\text{lat}_i, \text{lon}_i))^2 + (\hat{g}(\text{lat}_i, \text{lon}_{i+1}) - \hat{g}(\text{lat}_i, \text{lon}_i))^2} \quad (7)$$

where i and $i + 1$ are adjacent points on the grid. Finally, a tow in the survey data is assigned the gradient length value of the grid point that is nearest to the position of the tow. This gradient value is then used as a covariate in a continuous GLM model.

The analyses presented here are partly performed in R, a free statistical software package (<http://www.r-project.org>) and partly in S-PLUS (Venables and Ripely, 2002). R is available from the Comprehensive R Archive Network, <http://cran.r-project.org>.

4. Results

4.1. The variance function

In order to check assumption (1), the log sample variances and log sample means are calculated for homogeneous groups of data (Fig. 2). As noted above, each sub-rectangle/year combination contains at most seven observations and many of these contain one or two observations. Only those sub-rectangle/year combinations that contain five or more ($n_j \geq 5$) observations are included in this part of the analysis to reduce the variances of the estimates of the means and variances. This reduced data set contains 1289 observations and 241 different sub-rectangle/year combinations.

A weighted linear regression, with weights $1/n_j$, of the log sample variance on the log sample mean gives a slope of $\beta = 2.23 \pm 0.05$ (mean \pm standard error) and an intercept of $\alpha = -1.6 \pm 0.2$. The points are close to a straight line and the regression has an R^2 value of 0.88, but the two-sided t -test for the hypothesis $H_0 : \beta = 2$ is rejected at the 5% level of significance since $(\hat{\beta} - 2)/\hat{\sigma}_{\beta_2} = (2.23 - 2.00)/0.05 = 4.60$ (P -value is less than 10^{-5}). The results for statistical rectangles were similar, the regression gives a slope of $\beta = 2.23 \pm 0.05$ and an intercept of $\alpha = -1.1 \pm 0.3$ and the hypothesis $H_0 : \beta = 2$ is rejected at the 5% level of significance.

In the following analysis it will nevertheless be assumed that the mean-variance relationship (1) is valid for the data at hand and the gamma and log-normal

distributions are proposed for describing the data. The sample mean and variance are estimated from a small group of data and have associated uncertainty which are not accounted for in the regression above.

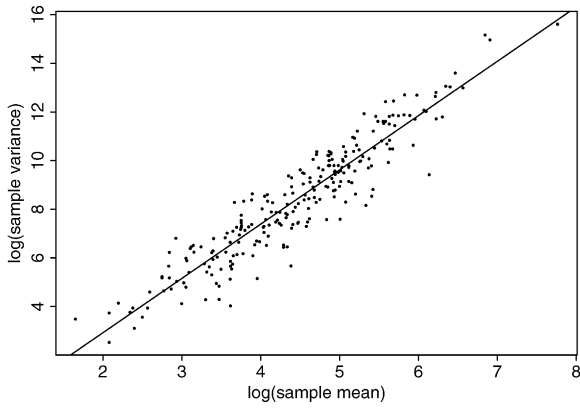


Fig. 2. Scatter plot of $\log(s_j^2)$ vs. $\log(\bar{y}_j)$ for every sub-rectangle j that has five or more observations (a total of 241 points). The regression line $\log(s^2) = 2.23 \log(\bar{y}) - 1.6$ is also included.

4.2. Goodness of fit

Models (4) and (5) were fitted separately for each of the proposed distributions. Only data from sub-rectangles with five or more observations were used (a total of 1289 observations) to obtain more reliable estimates of the parameters.

Both hypothesis $H_0 : W_{yji} = Y_{yji} / \hat{\mu}_{yj} \sim G(\hat{r}, 1/\hat{r})$ and $H_0 : W_{yji} = e^{\log(Y_{yji}) - \hat{a}_{yj}} \sim \text{LN}(0, \hat{b}^2)$ are rejected when all data are used and parameters estimated from model (4). The Kolmogorov statistic for the gamma distribution is $D_n = 0.075$ and $D_n = 0.056$ for the log-normal distribution but the 95% quantile of the distribution of D_n is $1.36/\sqrt{n} = 1.36/\sqrt{1289} = 0.038$. Fig. 3

Table 1
The D_n test statistics of the Kolmogorov test per year

Year	Gamma	Log-normal	$d(n)_{0.95}$
1985	0.106	0.110	0.152
1986	0.121	0.079	0.151
1987	0.103	0.119	0.151
1988	0.093	0.084	0.157
1989	0.104	0.096	0.157
1990	0.095	0.074	0.152
1991	0.105	0.072	0.151
1992	0.103	0.071	0.154
1993	0.104	0.100	0.152
1994	0.109	0.104	0.156
1995	0.100	0.080	0.156
1996	0.142	0.113	0.156
1997	0.129	0.097	0.158
1998	0.105	0.078	0.177
1999	0.087	0.058	0.164
2000	0.113	0.083	0.158
2001	0.138	0.098	0.158

shows the cumulative distribution functions (CDFs) for the hypothesized distributions $G(1.07, 0.93)$ and $\text{LN}(0, 1.05)$ along with the empirical CDFs of corresponding the W_{yji} .

The hypothesis can not be rejected when model (5) is fitted to the data for each year separately since the D_n test statistics are all lower than the critical 95% quantile (Table 1). The log-normal distribution leads to a lower values for the test statistic for all years except 1985 and 1987, indicating that the log-normal distribution is closer to the distribution of the data. Therefore, the log-normal distribution is used in the following analysis.

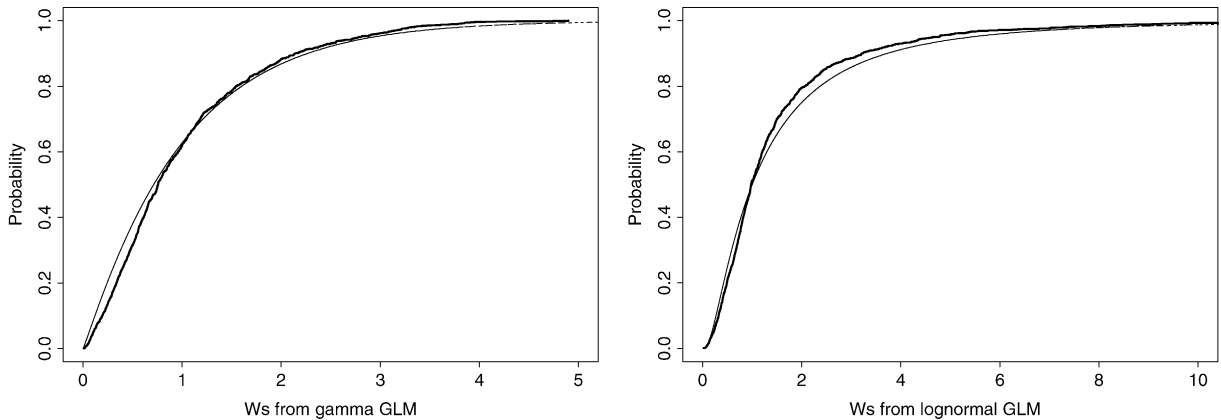


Fig. 3. CDFs for the hypothesized distributions $G(1.07, 0.93)$ and $\text{LN}(0, 1.05)$ (the thinner lines) along with the empirical CDFs of corresponding W_{yji} (the thicker lines).

4.3. A stratification model

The following model was fitted to log-transformed data:

$$\log(Y_{yji}) = \beta_0 + \alpha_y + \beta_j + \gamma_{yj} + \epsilon_{yji} \quad (8)$$

The parameter β_0 is the intercept (the grand mean), α_y is the year effect, β_j is the spatial effect of sub-rectangles and γ_{yj} is the interaction. The errors ϵ_{yji} are assumed to be normally distributed with zero mean and constant variance. This is the same model as (4) but in this case all data are used to fit the model (7066 observations). The sub-rectangle factor represents a spatial effect, the effect of the habitation conditions of that area for the cod. These conditions are likely to be controlled by environmental effects such as depth, temperature, amount of food available, etc. The year factor represents the inter-annual fluctuations in the cod stock size and in catchability. This basic model (without interactions) is a fairly standard method of analysis for the purpose of stock assessment. An interaction between sub-rectangles and years represents the difference in variation between sub-rectangles for different years. In other words, the habitation conditions do not change in the same way between years in two different sub-rectangles.

The analysis of variance is shown in Table 2. The terms are added sequentially (first to last). This model accounts for 80.4% of the total variation using $3653/7065 = 51.7\%$ of the total degrees of freedom. Both the main effects of sub-rectangles and years are significant and also the interaction between them. Most of the explained variation comes from the sub-rectangle effect, 49.4%, while the year effect explains only about 3%. The interaction between sub-rectangles and years is a considerable part of the explained variation, 28.2%,

but also uses the majority of the total degrees of freedom of the model.

The residuals are plotted to identify model inadequacy (Fig. 4). A scatter plot of standardized residuals versus the fitted values shows no obvious structure (Fig. 4a). Residuals plotted against the fitted values indicate that the model does not capture the smallest and largest observations (Fig. 4b). A normal probability plot of the residuals reveals that the residuals are not normally distributed (Fig. 4c). This is mostly due to the fact that many sub-rectangle/year combinations contain only one observation, resulting in a residual of zero. The standardized residuals plotted against year indicate heterogeneous variability (Fig. 4d). The standard deviations for each year (connected squares in Fig. 4d) are less than unity. The standard deviations for 1999–2001 are half those for the other years and the range of residuals for these years is smaller than for the other years.

4.4. A continuous model

Using sub-rectangles and years as factors does not provide any information or explanation of the data other than the fact that the expected cod catch depends on location and time. A more informative model is one that relates the expected catch to environmental variables that can be expected to effect the behavior of cod on biological grounds. These environmental covariates can be thought of as substitutes for the sub-rectangles effect as they explain why the fish is more likely to be at one place rather than another.

The covariates used in the continuous model were selected by looking at box plots of the response versus each of the environmental variables and some of the trawl station data (e.g. Fig. 5). These plots can, of

Table 2
Analysis of variance table for the stratification GLM of log-transformed cod catch data

Source of variation	d.f.	SS	% expl.	SS/d.f.	F-test	P-value
Sub-rectangles	305	8719.4	49.4	28.59	28.14	<2.2E-16
+ Years	16	481.3	2.7	30.08	29.61	<2.2E-16
+ Interaction	3332	4972.8	28.2	1.49	1.47	<2.2E-16
Total model	3653	14173.4	80.4			
Residuals	3412	3465.8		1.016		
Total	7065	17639.2				

The terms are added sequentially (first to last). 80.4% of the total variation is explained by this model. Both the main effects of sub-rectangles and years are significant and so is the interaction between them.

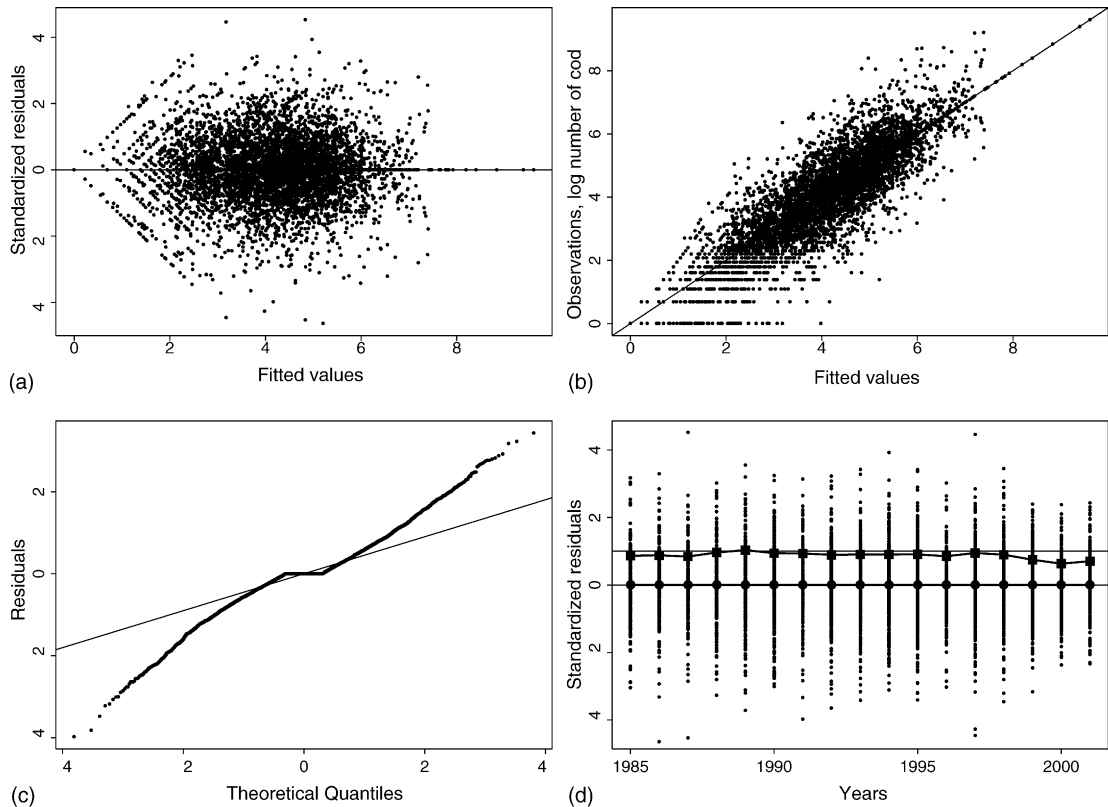


Fig. 4. Analysis of residuals of the stratification model. (a) Scatter plot of standardized residuals vs. the fitted values. (b) Scatter plot of observations vs. fitted values along with the $y = x$ line. (c) Normal probability plot of residuals. (d) Standardized residuals per year. The sample mean and standard deviation for each level are indicated with connected circles and squares.

course, only show one covariate at a time and can therefore only give hints as to appropriate covariates and the functional forms for the relationships between the covariates and the response. Plots such as Fig. 5 lead to consideration of the following covariates:

- poly(Depth, 2).
- poly(Bottom temperature, 2).
- Surface temperature.
- Air temperature.
- poly(Wave height, 2).
- poly(Latitude, Longitude, Year, 4).
- poly(Towing time, 2).
- poly(Towing length, 2).
- factor(Vessel).

The notation ‘poly(x_1, \dots, x_p, n)’ denotes an orthonormal polynomial of degree n in the variables

x_1, \dots, x_p . The notation ‘factor(Vessel)’ denotes that the vessel effect is qualitative. As before, the response is the log-transformed number of cod and the errors are assumed to be normally distributed. All of the data (7066 observations) is used in the analysis. Covariates are added to the model using a stepwise procedure based on the maximum decrease in AIC, with the constraint that each additional covariate explains at least 0.5% of the total variation. Only the environmental variables are considered initially because they are the variables of primary interest. Trawl station data are examined only after all possible environmental variables are considered for possible inclusion in the model. The analysis of variance (ANOVA) is shown in Table 3. The terms are added sequentially (first to last). Table 3 also shows the percentage change in the residual sum of squares (RSS) if one term is removed from the model (columns 10 and 11) and the percentage of total vari-

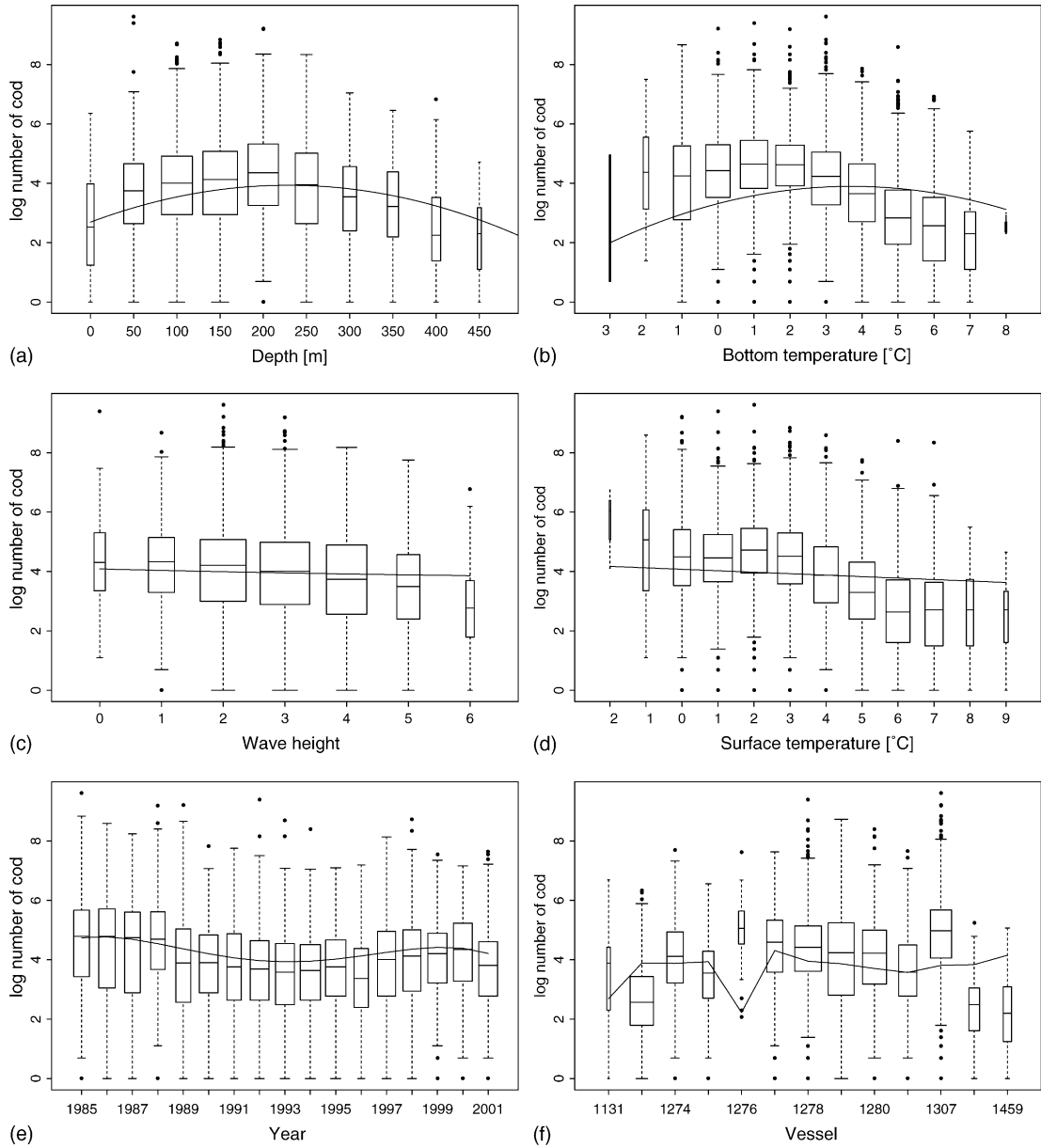


Fig. 5. Box plot of log number of cod in each tow vs. (a) depth, (b) bottom temperature, (c) wave height, (d) surface temperature, (e) year, and (f) vessel identification number. The boxes show the middle 50% of the data and the middle line shows the median. Dotted lines are drawn to the extreme points but are not made longer than 1.5 times the height of the box and data outside that range are shown with points. The width of the boxes is made proportional to the square root of the number of observations for the box. The curves show the estimated effects on log number of cod for all other variables hold fixed.

Table 3
Analysis of variance table for the continuous GLM of log-transformed cod catch data

Source of variation	d.f.	SS (sequence)	% explained	AIC	F-test	P-value	SS (add1)	% decrease of RSS	SS (drop1)	% increase of RSS
poly(Bottom temperature, 2)	2	3640.7	20.6	4837	1311.4	< 2.2E-16	3640.7	20.6	178.2	1.8
+ poly(Depth, 2)	2	724.4	4.1	4465	260.9	< 2.2E-16	893.9	5.1	268.1	2.8
+ Surface temperature	1	290.9	1.6	4311	209.6	< 2.2E-16	3178.1	18.0	14.5	0.1
+ poly(Wave height, 2)	2	130.2	0.7	4243	46.9	< 2.2E-16	327.2	1.9	11.7	0.1
+ poly(Latitude, Longitude, Year, 4)	34	2984.6	16.9	2444	63.2	< 2.2E-16	7252.4	41.1	2292.0	23.5
+ factor(Vessel)	12	135.4	0.8	2371	8.1	2.205E-15	3698.9	21.0	135.4	1.4
Total model	53	7906.2	44.8							
Residuals	7012	9733.0								
Total	7065	17639.2		6466						

The terms are added sequentially (first to last) and they are all significant.

ation that each term explains when it is the only (or first) term in the model (columns 8 and 9). All of the covariates considered lead to reductions in AIC. However, three of them do not reduce the sum of squares by 0.5% or more (Table 4).

The environmental variables explain 27% of the variation, and when the polynomial in latitude, longitude and year is added, these terms together explain 43.9% of the variation, slightly more than the latitude, longitude and year polynomial alone (41.1%). A model containing only the 4th degree polynomial in latitude, longitude and year is comparable to the stratification model (8) in the sense that both models include only spatial and time effects and the interaction between space and time. The polynomial model does not reduce the variation as much as the stratification model but considering the number of degrees of freedom (34 versus 3653) it performs very well. The bottom and surface temperatures explain 20.6% and 18.0% of the total variation when fitted separately. When both bottom temperature and depth are included in the model, the additional variance explained by surface temperature

is only 1.6% indicating that the effect of surface temperature on the location of cod is captured by bottom temperature and depth. On the other hand, this 1.6% is highly significant and may well be an indication of the behavior of pelagic prey such as capelin.

There is considerable confounding between the vessel effect and the location effect since each vessel tends to be sent to a similar area. The confounding is not complete, however, since the boundaries between areas covered by the vessels are not fixed and, in some cases, vessels get moved to new areas, or temporarily replaced by other vessels. This is seen in the two ANOVA tables. The effect of location as a single factor, explains 49% of the variation (Table 2) whereas the effect of vessel alone explains only 21% (Table 3). The year effect only contributes in a minor way to explaining variability, but the interaction between year and location appears quite important. When looking at the marginal effects in the final model (last column in Table 3), it is seen that the vessel effect appears much less important than the location–year effect combined.

Table 4
Analysis of variance for those terms not included in the model

Term	d.f.	SS	% explained	AIC	F-test	P-value	SS (add1)	% decrease
None				2371				
Air temperature	1	3.2	0.0	2370	2.30	0.1295	862.3	4.9
Poly(Towing time, 2)	2	51.2	0.3	2337	18.54	9.313E-09	218.4	1.2
poly(Towing length, 2)	2	50.9	0.3	2338	18.41	1.061E-08	218.3	1.2

For each term the additional variance explained is shown along with the AIC. Columns 8 and 9 show the SS and the percentage decrease of RSS when each of the terms are fitted separately.

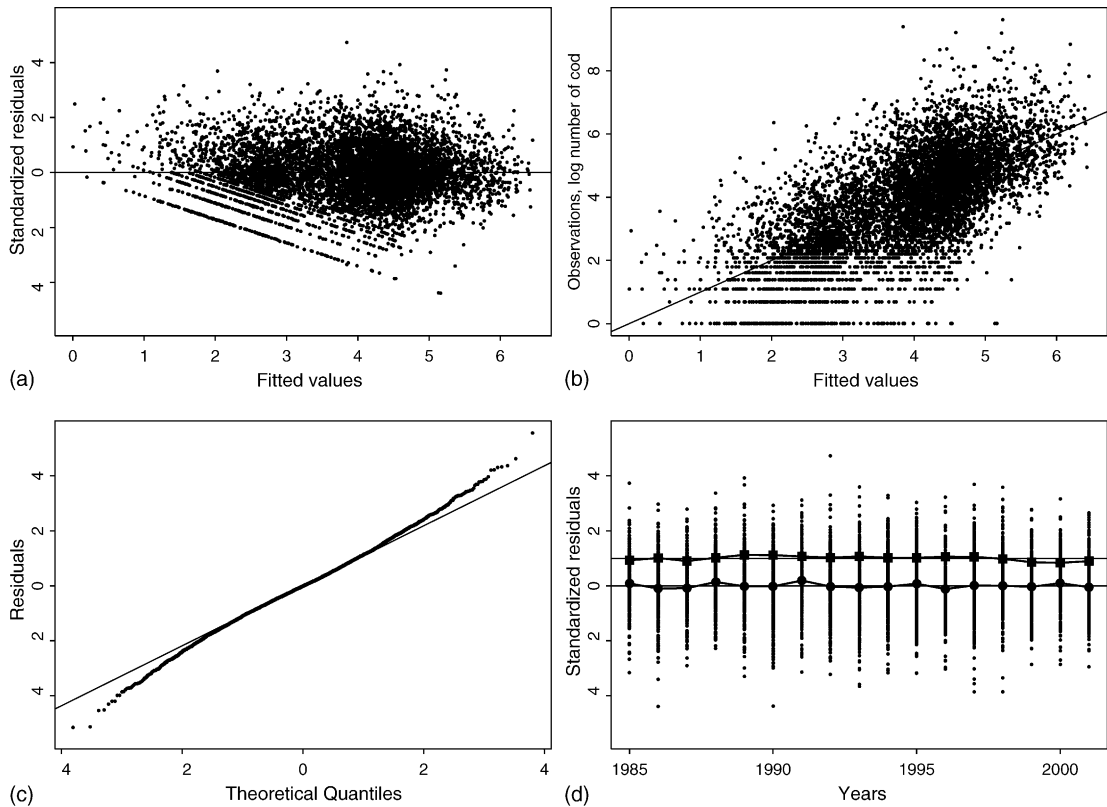


Fig. 6. Analysis of residuals for the continuous model. (a) Scatter plots of standardized residuals vs. the fitted values. (b) Scatter plot of observations vs. fitted values along with the $y = x$ line. (c) Normal probability plot of residuals. (d) Standardized residuals per year. The sample mean and standard deviation for each level are indicated with connected circles and squares.

Fig. 6 examines the residuals for the continuous model. The patterns are generally the same as for the stratification model. The tails in Fig. 6c, particular for low residual values, seem to be thicker than could be expected for a normally distributed variable. Unlike the stratification model, there is no evidence for reduced residual standard deviations after 1999 in Fig. 6d and, in fact, the residual standard deviations are generally larger than unity.

4.5. Estimated temperature gradient

The data used to estimate a temperature surface for each year are all the groundfish survey data from 1985 to 2001 where the bottom temperature was measured. Empty tows are included when fitting the temperature surface to increase the number of temperature measurements. The total number of observations for this anal-

ysis is 8705 or 512 per year on average. However, the temperature data provided by the survey is very sparse and model (6) only explains (on average) 16% of the total variation and uses (on average) 24.7 approximate degrees of freedom. In other words, the surface is too smooth to capture local changes in temperature. Fig. 7 shows contour plots of the smoothed temperature surface obtained by model (6) for 2 years of the survey. The temperature surface exhibits the same pattern each year, the bottom temperature is higher (mostly $> 3^{\circ}\text{C}$) south and east of Iceland than north and west of the country (mostly $< 3^{\circ}\text{C}$).

A third degree polynomial in gradient length (see box plot in Fig. 8) is included along with the previous environmental covariates and trawl station data to explain the variability of cod catch rates. This new covariate is significant although its contribution to variance reduction is small (Table 5). The gradient term explains

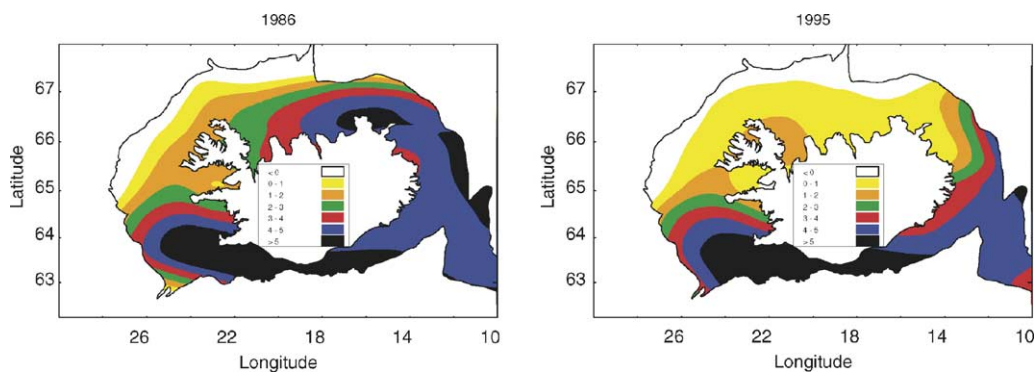


Fig. 7. Contour plots for 1986 and 1995 of the smoothed temperature surface inside the 500 m depth contour. The temperature is higher in darker areas than in brighter areas.

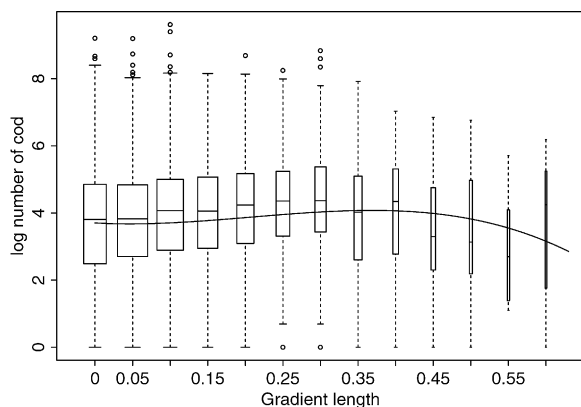


Fig. 8. Box plot of log number of cod in each tow vs. the length of the estimated gradient vector at each location. The smooth curve shows the estimated effects on log number of cod for all other variables held fixed.

1.4% of the total variance as the first term in a model. It is included as the fourth term in the model, adding an additional 1.1% to the explained variance. Removing the gradient term from the final model increases the residual sum of squares by 0.6%. The residuals for this model exhibit the same patterns as those for the continuous model (Fig. 6).

5. Discussion

The test of probability distributions revealed that the log-normal distribution mimicked the variance structure in the data slightly better than the gamma distribution. However, there seems to be room for improvement here since the regression of log sample variance

Table 5

Analysis of variance table for the continuous GLM of log-transformed cod catch data where a polynomial in gradient length has been included

Source of variation	d.f.	SS (sequence)	% explained	AIC	F-test	P-value	SS (add1)	% decrease of RSS	SS (drop1)	% increase of RSS
poly(Bottom temperature, 2)	2	3640.7	20.6	4837	1318.9	< 2.2E-16	3640.7	20.6	174.5	1.8
+ poly(Depth, 2)	2	724.4	4.1	4465	262.4	< 2.2E-16	893.9	5.1	271.4	2.8
+ Surface temperature	1	290.9	1.6	4311	210.8	< 2.2E-16	3178.1	18.0	19.0	0.2
+ poly(Gradient, 3)	3	189.7	1.1	4213	45.8	< 2.2E-16	250.7	1.4	59.5	0.6
+ poly(Wave height, 2)	2	123.5	0.7	4148	44.7	< 2.2E-16	327.2	1.9	11.7	0.1
+ poly(Latitude, Longitude, year, 4)	34	2861.5	16.2	2407	61.0	< 2.2E-16	7252.4	41.1	2223.0	23.0
+ factor(Vessel)	12	135.1	0.8	2333	8.2	1.894E-15	3698.9	21.0	135.1	1.4
Total model	56	7965.8	45.2							
Residuals	7009	9673.4								
Total	7065	17639.2		6466						

on the log sample mean lead to a slope higher than 2. An avenue of future research could be to assume the variance function $V(\mu) = \mu^{2.23}$ for the data and estimate the model parameters by constructing a quasi-likelihood function. Other variance functions are of course also possible. For example, $V(\mu) = \mu + \mu^2/k$, which is the variance function for the negative binomial distribution. How these different error structures would affect the results of this study is unclear. In the case of gamma versus log-normal, however, it turns out that the main effects are robust to the choice of the underlying distribution, all the same variables are significant but the gamma model only explains 38.7% of the total deviance compared to 44.8% for the log-normal model.

The models in this paper have been designed to explain the “log-normal” part of the “delta models”, leaving aside the various Bernoulli or binomial models for the “delta” part. The models with the greatest number of parameters are able to explain some 80% of the variation in the (logged) data, which is considerable given the nature of catch data. These models include location (sub-rectangle) and year as factors along with their interaction (the “stratification model”).

Several issues arise from this conceptually simple model, but since these kinds of models are commonly used to extract an annual index of abundance, it must be noted how important the interaction term between year and location is, at least for this species in these waters. In this case, it is quite possible that the change in distribution over time implied by this interaction might create havoc when attempts are made to use simple multiplicative models to obtain an abundance index, without an interaction term.

Another point to note is the heterogeneity of variances on log scale. Although the standard deviation on log scale appears to be close to unity for most years, it appears to decline considerably in the last 2 years (Fig. 4). This can only happen when the variance within several sub-rectangles drops considerably, and this indicates the absence of large catches in the last years. This interesting phenomenon may indicate a change in aggregation mechanisms either at low population abundance or due to environmental effects.

On the other hand, when attempting to use extensive information on environmental variables (the “continuous model”), it is seen that these models are unable to explain the same amount of variation as the factor-based models, and even when all available environmen-

tal variables are included, there is still a need to include location. Since the fish are expected to aggregate in different locations due to changes in the environment rather than simply geographic location, it would appear that there are important environmental variables which are not collected on most surveys. One unexplored variable is stomach content data, which could be an indicator of food supply. Such data have been gathered routinely during Icelandic groundfish surveys, albeit at varying levels of intensity.

The drop in residual standard deviation is not seen in the continuous model. This implies that there is estimated high variability at least in some sub-rectangles where the observed variability around the mean is low. The danger is that this may overestimate the mean in several locations in the last years, so if this type of model is used to obtain an abundance index, it may fail to show the actual decline of the population.

The temperature gradient estimated from the smoothed temperature field is related significantly to abundance. However, this effect is of minor importance in terms of the amount of explained variation. It would seem to be of considerable importance to improve on this methodology since it appears to be a well-established fact among fishermen that changes in temperature tend to be indicators of fish schools.

The models considered here can be extended considerably in different ways. For example, the degrees of polynomials can be increased and more interaction terms can be included. The method used here has been to decide on the degree through visual inspection and only obvious interaction terms have been included. When considering the final continuous model, the interaction between bottom temperature and depth is an obvious candidate to test. This term is indeed a significant addition to the final model, giving a further reduction of 1.2% in the sum of squared errors. It is not at all clear, however, what meaning can be attached to the product of temperature and depth, but it is clear that there is room for further investigation of these data.

In this analysis, possible covariates were examined by means of box plots. Polynomials of different degrees were suggested for the relationship with the response based on these plots. The disadvantage of this method is, however, that it is difficult to guess the most appropriate degree of such polynomials. This is especially the case for combined effects, i.e. when the poly-

mial contains more than one variable. Another disadvantage is that the estimated surface of such a polynomial regression becomes very limited since the fit of one data point depends on all the data. Generalized additive models (GAMs) provide a method where less restrictive functions of covariates can be used when modeling data because no rigid parametric assumptions are made about the dependence of the response on the covariates. The loess smoother used in this study is an example of these types of models. GAMs are an important area of future research for the analysis of groundfish survey data, but were not used in this study partly due to the lack of distribution theory for formal model tests (Hastie and Tibshirani, 1990, p. 155).

Acknowledgments

The analyses presented in this paper were a part of a M.Sc. Thesis. (Brynjarsdóttir, J., 2002. Statistical analysis of cod catch data from Icelandic groundfish surveys. Department of Engineering, University of Iceland, Reykjavík.) This work was supported in part by EU grant QLK5-CT199-01609.

References

- Anon., 2001. State of marine stocks in Icelandic waters 2000/2001, prospects for the quota year 2001/2002. MRI Technical Report 87. Marine Research Institute, Reykjavík.
- Atkinson, A.C., 1982. Regression diagnostics, transformations and constructed variables. *J. R. Statist. Soc. B Met.* 44, 1–36.
- Einarsson, S.T., Jónsson, E., Björnsson, H., Pálsson, J., Schopka, S.A., Bogason, V., 2002. Handbók um stofnmælingu botnfiska á Íslandsmiðum. Hafrannsóknastofnunin (The Marine Research Institute), Iceland.
- Firth, D., 1988. Multiplicative errors: log-normal or gamma? *J. R. Statist. Soc. B Met.* 50, 266–268.
- Goñi, R., Alvarez, F., Adlerstein, S., 1999. Application of generalized linear modelling to catch rate analysis of Western Mediterranean fisheries: the Castellón trawl fleet as a case study. *Fish. Res.* 42, 291–302.
- Hastie, T.J., Tibshirani, R.J., 1990. *Generalized Additive Models*. Chapman & Hall/CRC, London.
- Lo, N.C.H., Jacobson, L.D., Squire, J.L., 1992. Indexes of relative abundance from fish spotter data based on delta-lognormal models. *Can. J. Fish. Aquat. Sci.* 49, 2515–2526.
- McCullagh, P., Nelder, J.A., 1989. *Generalized Linear Models*, 2nd ed. Chapman & Hall/CRC, London.
- Myers, R.A., Pepin, P., 1986. The estimation of population size from research surveys using regression models. ICES, C.M. 1986/D:9.
- Pálsson, O.K., Jónsson, E., Schopka, S.A., Stefánsson, G., Steinarsson, B.Æ., 1989. Icelandic groundfish survey data used to improve precision in stock assessment. *J. Northwest Atl. Fish. Sci.* 9, 53–72.
- Pennington, M., 1983. Efficient estimators of abundance, for fish and plankton surveys. *Biometrics* 39, 281–286.
- Pennington, M., 1996. Estimating the mean and the variance from highly skewed marine data. *Fish. Bull. US* 94, 498–505.
- Sakuma, K.M., Ralston, S., 1995. Distributional patterns of late larval ground-fish off central California in relation to hydrographic features during 1992 and 1993. *Cal. Coop. Ocean. Fish.* 36, 179–192.
- Stefánsson, G., 1988. A statistical analysis of Icelandic trawler reports, 1973–1987. ICES, C.M. 1988/D:13.
- Stefánsson, G., 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. *ICES J. Mar. Sci.* 53, 577–588.
- Stefánsson, G., Pálsson, O.K., 1997. Statistical evaluation and modelling of the stomach content of Icelandic cod (*Gadus morhua*). *Can. J. Fish. Aquat. Sci.* 53, 89–93.
- Steinarsson, B., Stefánsson, G., 1986. Comparison of random and fixed trawl stations in Icelandic groundfish surveys and some computational considerations. ICES, C.M. 1986/D:13.
- Venables, W.N., Ripely, B.D., 2002. *Modern Applied Statistics with S*, 4th ed. Springer, New York.
- Vilhjálmsón, H., 1994. The Icelandic capelin stock. *J. Mar. Res. Inst.* XIII, 1–281.
- Wiens, B.L., 1999. When log-normal and gamma models give different results: a case study. *Am. Statist.* 53, 89–93.
- Ye, Y., Al-Husaini, M., Al-Baz, A., 2001. Use of generalized linear models to analyze catch rates having zero values: the Kuwait driftnet fishery. *Fish. Res.* 53, 151–168.