

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI[®]

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Statistical inference about fish abundance: An approach based on research survey data

by

Noel Cadigan

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 1999

©Noel Cadigan 1999



National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-38227-3

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

Trawl research surveys for groundfish populations provide an important source of information for estimating the size of the population. The surveys usually provide an estimate of the size of a component of the population, called trawlable abundance. In this thesis we consider how to efficiently estimate trawlable abundance, and how to construct precise confidence intervals for this quantity. The mobility of the population and other aspects of the survey design motivate us to pursue a model-based approach for inferences. Another advantage of the model-based approach is the relative ease in which scientific knowledge can be incorporated into inferences. We demonstrate that this can have important consequences on the precision of inferences.

We propose a mixture-model approach for the analysis of trawl survey data. With such data it is common that extremely large catches occasionally occur. While most catches vary smoothly in terms of spatial location and depth, these large catches are quite different. In some years large catches do not occur, but it still seems likely that the dense fish aggregations that produce these catches exist. These aggregations are not sampled simply due to the small probability of sampling a very dense aggregation. Simple habitat-selection theory suggests that it is reasonable to expect different trends in abundance at favorable and less favorable locations. This is because animals tend to move to more favorable sites as space becomes available. This is the basis for our mixture-model. We use a smooth mean-function model for most trawl catches (type I's) in terms of spatial location and tow depth. For the rare large catches (type II's) we use a simple constant mean model, where the mean of the large catches is constant over time. The mixture-model combines the models for type I and type II catches. Habitat-selection is accommodated because the mean for type II catches can remain high while the mean for type I catches declines. We apply these methods to trawl research survey data for cod on the Southern Grand bank of Newfoundland.

We used a variety of graphical procedures for model building purposes. They suggest a nonparametric approach for modelling type I catches is useful. The approach we take is local likelihood estimation. A concern in model building is which covariates to include. We propose a nonparametric variable selection procedure for this purpose. We do not know exactly which catches are type II a priori, so robustness is an issue. For this we develop a robust nonparametric Negative Binomial smoother, to identify type II catches and for “robustified” variable selection. The Negative Binomial distribution is the stochastic model we use for trawl catches. We show that the mixture-model is an improvement over the model commonly used to analyze trawl survey data. We suggest approximate pivotal methods for confidence intervals for total type I and type II abundance, and use the profile likelihood to compute confidence intervals for total type I + type II abundance.

Our estimates show less interannual variation than traditional methods for estimating abundance. We feel the mixture model estimates are more consistent with the survey data. Our confidence intervals are also considerably shorter in some years than those based on $\pm 2 \times$ the design-based standard errors.

CONTENTS

1	Introduction	1
1.1	Data source: Trawl research surveys	7
1.2	Outline	9
1.3	Appendix: Figures	11
2	Review of some fisheries topics	13
2.1	Introduction	13
2.2	Brief history of the Newfoundland cod fishery	13
2.3	General biology of Atlantic cod	16
2.3.1	Stock designation of Southern Grand Bank (3NO) Atlantic cod . .	17
2.3.2	General biology of the Southern Grand Bank (3NO) Atlantic cod .	19
2.4	Some stock size assessment methods	20
2.4.1	Data sources	20
2.4.2	Total catch and effort analysis	21
2.4.3	Catch at age analyses	24
2.4.4	Catch and effort analysis	30
2.4.5	Relative indices of abundance	31
2.4.6	Catch at length analysis	34
2.4.7	Future considerations	35

3	Survey design and inference	36
3.1	Introduction	36
3.2	3NO design	37
3.2.1	Sample selection	39
3.3	Overview of design-based inferences	40
3.3.1	General design inferences	42
3.3.2	Horvitz-Thompson estimator of T	43
3.3.3	Stratified three-stage sampling	45
3.3.4	Confidence intervals	55
3.4	3NO survey results	56
3.4.1	Variations from $S3C_1S$ in the 3NO survey design	58
3.5	Model-based approach	60
3.6	Conclusions	62
3.7	Appendix A: Variance of $\hat{T}_h$, and the Horvitz-Thompson estimator of $Var(\hat{T}_h)$	63
3.7.1	Variance of $\hat{T}_h$	64
3.7.2	Horvitz-Thompson estimator of $Var(\hat{T}_h)$	66
3.8	Appendix B: Tables	68
4	Data overview and exploratory analyses	77
4.1	Introduction	77
4.2	Spatial overview	79
4.3	Spatial-temporal overview	80
4.4	Spatial-temporal dependencies	83
4.4.1	Spatial autocorrelation results	84
4.4.2	Spatial-temporal autocorrelation results	85
4.5	Covariate associations	88
4.5.1	Depth analyses	89

4.5.2	Temperature analyses	91
4.5.3	Surface light analyses	92
4.5.4	Wind strength analyses	92
4.5.5	Sea state analyses	94
4.6	Large catches	94
4.7	Summary	95
4.8	Appendix: Literature review of the effects of depth and hydrographic conditions on bottom trawl catches	98
4.9	Appendix: Figures	103
5	Model specification	119
5.1	Introduction	119
5.2	Local likelihood smoother	122
5.3	Variable selection	126
5.3.1	Brief review of parametric variable selection	127
5.3.2	Nonparametric variable selection	129
5.4	Analysis of the 3NO trawl survey data, I	135
5.4.1	Standardizing covariates	135
5.4.2	Variable smoothing	139
5.4.3	Choosing the Negative Binomial k parameter	140
5.4.4	Variable selection results	145
5.5	Robust local likelihood estimation	148
5.5.1	Basic theory of robust M-estimation	149
5.5.2	Optimal B_s -robust estimators of Poisson and Negative Binomial rate parameters	154
5.5.3	Local likelihood estimates	164
5.5.4	Robust variable selection algorithm	168

5.6	Analysis of the 3NO trawl survey data, Π	175
5.7	Identification of Type II catches	178
5.8	Smooth mean function model for type I catches	186
5.9	Comparison with the stratum-effects model	188
5.10	Summary	190
5.11	Appendix: Tables	192
5.12	Appendix: Figures	201
6	Inferences	213
6.1	Introduction	213
6.2	Mixture model estimation	214
6.2.1	Estimates of 3NO trawlable stock size, $\mu_{t_{\cdot}}$	216
6.3	Inferences	221
6.3.1	Consistency of $\hat{\mu}_t^I(\mathbf{x})$	221
6.3.2	Consistency and asymptotic normality of $\hat{\mu}_{t_{\cdot}}^I$	226
6.3.3	Confidence intervals for $\mu_{\cdot}^{II}$	233
6.3.4	Confidence intervals for $\mu_{t_{\cdot}}$	234
6.4	Simulation	237
6.5	Conclusions	241
6.6	Appendix: Figures	244
7	Summary, and future research	246
7.1	Summary	246
7.2	Future research	252
8	Bibliography	258

LIST OF TABLES

3.2a	$erb_1 = \frac{E_{\xi}[v_1(\hat{T})] - E_{\xi}[Var(\hat{T})]}{E_{\xi}[Var(\hat{T})]}$	53
3.2b	$erb^* = \frac{E_{\xi}[v^*(\hat{T})] - E_{\xi}[Var(\hat{T})]}{E_{\xi}[Var(\hat{T})]}$	54
3.1a	Strata information for NAFO Div. 3N	68
3.1b	Strata information for NAFO Div. 3O	70
3.3a	Estimated total strata abundance and standard errors for 1992	72
3.3b	Same as Table 3.3a, except for 1993	73
3.3c	Same as Table 3.3a, except for 1994	74
3.4a	Estimated totals and standard errors for the 300 series strata	75
3.4b	Estimated totals and standard errors for strata 717-728	76
4.1	Estimated autocorrelations in the 20-50 Km distance class	87
5.1a	Bandwidths that minimize the cross validation criterion for the covariate regression analyses	139
5.1b	Covariate regression correlation coefficients	140
5.2a	Simulated NB mple bias for k	143
5.2b	Estimates of the NB scale parameter k and bandwidths b that maximize a predictive likelihood function	144
5.5	Standardizing constant (A) and asymptotic relative efficiency (ARE) of the bounded-influence estimator for μ in a $N(\mu, 1)$ model	152

5.6	Standardizing constants (α and A) and asymptotic relative efficiency (ARE) of the bounded-influence estimator for σ in a $N(0, \sigma)$ model	154
5.7	Standardizing constants (α and A) and asymptotic relative efficiency (ARE) for the bounded-influence estimator of m in a $Poi(m)$ model	156
5.8	Finite-sample relative efficiency for the bounded- influence estimator of m in a $Poi(m)$ mode	156
5.9	Outlier-induced bias in the the bounded- influence influence estimator of m in a $Poi(m)$ mode	157
5.10	Asymptotic relative efficiency (ARE) of the bounded- influence estimator of m and in a $NB(m, k)$ model	159
5.11	Relative efficiency (RE) of the bounded-influence estimator of m in a $NB(m, k)$ model	160
5.12	Outlier-induced bias in the bounded- influence estimator of m in a $NB(m, k)$ model	161
5.13a	Simulated biases in the Poisson bounded- influence estimator of m applied to $NB(m, k)$ data	162
5.13b	Simulated biases in the NB bounded- influence estimator of m applied to $Poi(m)$ data	163
5.14	Wing spread data, with kernel (w_t) and robust (w_t^r) weights	173
5.16a	Estimates of the NB scale parameter k and bandwidths b that maximize a predictive likelihood function	179
5.16b	Estimates of the NB scale parameter k and bandwidths b that minimize AIC	180
5.16c	Estimates of the NB scale parameter k and bandwidths b that minimize BIC	181
5.17	$GC_{pb_{min}}$'s with $k = 1$ on the shelf, and $k = 0.5$ on the slopes	182

5.16	Type II catches	184
5.19	$GC_p b_{min}$'s for type I catches	187
5.20	Deviances and prediction error measures for Type I catches. Smoother minus stratum-effects estimates are in parentheses	189
5.3a	Cross-validation (CV) variable selection results. The Negative Binomial scale parameter is $k = 0.65$ for all subsets	192
5.3b	Same as Table 5.3a, except for generalized cross-validation (GCV)	193
5.3c	Same as Table 5.3a, except for generalized C_p (GC_p)	194
5.4a	Cross-validation (CV) variable selection results. The Negative Binomial scale parameter is $k = 5$ for all subsets	195
5.4b	Same as Table 5.4a, except for generalized cross-validation (GCV)	196
5.4c	Same as Table 5.4a, except for Generalized C_p (GC_p)	197
5.15a	Cross-validation (CV) variable selection results. The Negative Binomial scale parameter is $k = 0.65$ for all subsets	198
5.15b	Same as Table 5.15a, except for generalized cross-validation (GCV) . . .	199
5.15c	Same as Table 5.15a, except for Generalized C_p (GC_p)	200
6.1	Summary of type II catches	218
6.2	Mixture model estimates of 3NO stock size	220
6.3	Mixture model estimates of 3NO stock size, with confidence intervals . . .	235
6.4	Simulation results	239
6.5	Simulation results	240

LIST OF FIGURES

1.1	The Atlantic cod, <i>Gadus morhua</i>	2
1.2	A trawler towing an otter trawl.	8
1.3	NAFO management units	11
1.4	Depth stratification scheme	12
2.1	Cod landings in NAFO divisions 3N and 3O by country.	15
3.1	Annual estimates of trawlable abundance for the shallow water strata	58
4.1a	Spring 3NO bottom trawl research survey catches of cod	103
4.1b	Same as Figure 4.1a	104
4.2a	Time series of catches	105
4.2b	Same as Figure 4.2a	106
4.3a	Survey trajectory	107
4.3b	Same as Figure 4.3a	108
4.4a	Spatial-temporal autocorrelations in catches	109
4.4b	Spatial-temporal autocorrelations in catches	110
4.4c	Comparison of spatial and spatial-temporal autocorrelations	111
4.5a	Catches versus tow depth	112
4.5b	Time series of catches at depth intervals	113
4.6a	Catches versus bottom temperature	114
4.6b	Time series of catches at temperature intervals	115

4.7a	Distributions of catches versus wind force	116
4.7b	Depth stratified annual ratios of average catches for low versus high winds	117
4.8	Synoptic analysis of the five tows with the largest catch of cod	118
5.1a	Difference between the <i>CV</i> per observation deviance	201
5.1b	Same as Figure 5.1a	202
5.2	"Robustified" smoothing analysis of trawl wing-spread data	203
5.3a	Difference between the <i>CV</i> per observation "cleaned" deviance	204
5.3b	Same as Figure 5.3a	205
5.4a	Inverse robust weights for a smoother based on spatial location and tow depth	206
5.4b	Same as Figure 5.4a	207
5.5	Year and location of type II catches	208
5.6a	Smoother predicted type I cod catches	209
5.6b	Same as Figure 5.6a	210
5.7a	Smoother standardized type I residuals	211
5.7b	Same as Figure 5.7a	212
6.2	Annual estimates of trawlable stock abundance from the mixture model, and stratum area-weighted averages.	236
6.3	Simulation population.	238
6.1a	Comparison of model <i>f</i> -weights and design π -weights.	244
6.1b	Same as Figure 6.1a	245
7.1	Temporal persistence in stratum-average survey catches	254
7.2	Observed and predicted Canadian 3PS survey indices	255
7.3	Projection risk plot	257

Acknowledgements

I wish to thank my supervisor, Dr. Jiahua Chen, for his practical guidance and insights, and financial support, with my research. I acknowledge financial support from the Department of Fisheries and Oceans, Government of Canada. I wish to thank all my colleagues at the Northwest Atlantic Fisheries Center for their support and advice with this project. I thank my parents for keeping my “bow to the wind”. Most of all I thank my wife, Tina, for her support, encouragement, and good humor.

1. INTRODUCTION

In this dissertation we present methods to make probabilistic inferences about the size of a fish stock based on scientific survey data. The total number of fish in a stock changes with time, and it is very important to know whether the stock is increasing or decreasing, and at what rate. The total number of fish at time t , denoted as N_t , is never known exactly; however, annual research surveys are often available which provide quantitative information about components of the stock. These surveys involve random trawl samples, and they are often conducted for periods of 10 years or more. Many measurements are taken during the surveys, and the one we focus on is the number of fish caught in the trawl. Statistical methods are required to make inferences about trends in stock size from the survey data. The methods we investigate will be used to estimate the relative magnitude of N_t with high precision, and we develop confidence intervals for this quantity as well.

The fish stock we consider is the Atlantic cod (*Gadus morhua*; see Figure 1.1), located off the east coast of Newfoundland in NAFO (North Atlantic Fisheries Organization) Divisions 3N and 3O (3NO; see Figure 1.3 in the **Appendix**). This area is commonly referred to as the Grand Bank of Newfoundland. A Grand Bank cod fishery has been conducted for nearly 500 years. This stock and other Atlantic cod stocks have played a fundamental role in the development of the economy of Newfoundland and the Maritimes in Canada, as well as the northeastern United States. Despite the recent commercial collapse of this stock and several other Northwest Atlantic cod stocks, great scientific and

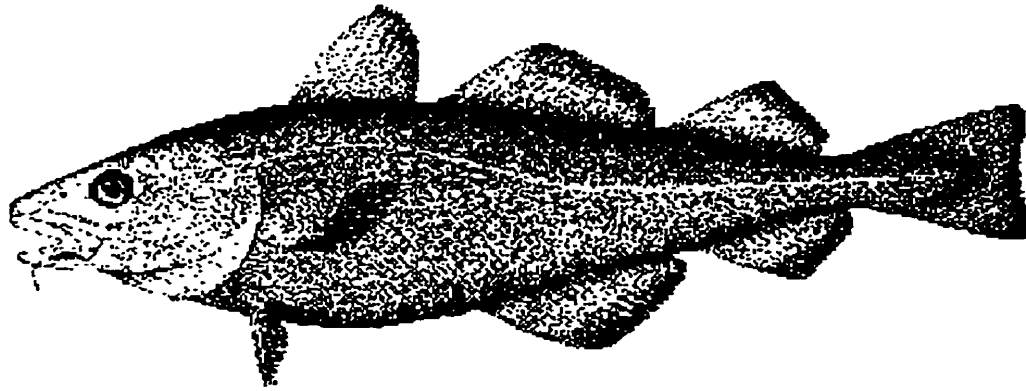


Figure 1.1: The Atlantic cod, *Gadus morhua*.

commercial interest still exists in these fish. Hence, the cod example is quite important and methodologies developed in this dissertation may well be applied in future assessments of this stock.

We first review common methods used to assess the size of fish stocks. Most of these methods require accurate estimates of the size of commercial fish catches. Next we review the survey design used to assess 3NO cod, including methods for summarizing the survey data. The basic design used is stratified random sampling. Exploratory graphical analyses are presented that suggest a mixture model is useful for analyzing the survey data. Stock size inferences are then based on the mixture model.

A common method used to analyze research survey data is based on the survey design used to obtain the data. This method is often referred to as the design-based approach. There are several problems with the application of this approach to the 3NO cod survey data. With design-based methods we must assume that fish are stationary; however, this assumption is not true for 3NO cod. Failure to meet this assumption implies that the precision of inferences may be compromised. To assess the variability of an estimator, the design-based approach requires that any two sampling units have a positive probability

of being included in the sample. This is not the case for the 3NO survey (see **Chapter 3**). Another implicit assumption in the design-based approach is that all sampling units are alike, at least for those in the same stratum. For 3NO cod extremely large catches occasionally occur that seem very different from the other catches within the stratum. Such catches should be considered as special cases because, otherwise, they have undue influence on stock size inferences. An approach for dealing with such extremes is to identify these units before conducting a survey and assign specific inclusion probabilities to the large units accordingly; however, for our case it is not possible to do this.

To address these problems we use a model-based approach to analyze the survey data. Such an approach is based on an assumed stochastic model that “generates” the data. A model-based approach can accommodate random movements of fish. This approach facilitates variance estimation in that it does not require that any two sampling units have a positive probability of being included in the sample. In fact, it does not even require that observations be collected using a sampling design with known inclusion probabilities, which is a basic assumption in the design-based approach. We can also use a separate model for large catches, according to any biologically relevant theory about their incidence. The model we use can be categorized as nonparametric regression. The fundamental assumption we make is that fish abundance varies smoothly over time, space and depth, and possibly other covariate dimensions. Many covariates are measured in the 3NO cod survey, and we use these data to construct a hopefully reliable predictive relationship between some of the covariates and fish abundance. In doing this, some interesting statistical problems arise.

It is well known that when the number of observations is limited we cannot fit a model with a large number of covariates. Otherwise the model will over-fit the data and have little predictive value, and this is important in our application. Including too few covariates may result in a poorly fitted model and possible bias. A challenging

statistical problem in our work has been to determine which covariates should be included in the model. This problem has received much attention for linear models; however, little discussion is present in the statistical literature for nonparametric models. We use a two-stage approach for this purpose. First we conduct graphical analyses that focus on potential covariate associations that seem plausible. Covariates that seem unrelated to survey abundance are removed from further analyses. The second stage involves analytic variable selection procedures performed with the covariates that appear important in the graphical analyses. These methods are used because it is difficult to graphically assess the relative strengths of covariate associations in more than two dimensions. Confounding among covariates is also a problem in the data we analyze; hence, an association between trawl catches and a covariate may be an artifact caused by the association between that covariate and some other variable that is truly related to trawl catches. In this dissertation we select covariates subsets that minimize, or nearly minimize, an estimate of mean square prediction error; that is, the mean squared difference between a random observation and its predicted value.

Even if we know exactly which covariates to include there are still difficulties due to a small number of extremely large catches. These large catches are dissimilar to other catches with similar covariate values, and this conflicts with the assumption that fish abundance varies smoothly as a function of covariates. Recall that this is also a problem in the design-based approach. These large catches should be removed when fitting the nonparametric regression, and yet we do not know which catches are clearly extreme without the regression. It is not a simple matter of deciding which catches are large through the use of a cut-off criterion because spatial clusters of large catches can occur that are consistent with nearby catches. We use robust smoothing techniques as a solution to this problem.

Large catches do provide vital information about the size of the stock however, and

they cannot be ignored. It commonly occurs that 50% or more of the total catch of cod in a survey occurs in less than 10% of the samples, and this is merely a reflection of the truly contagious nature of stock abundance. It is quite easy for a large percentage of the total stock abundance to lie between two sample sites, and yet no fish exist at these sites. The large catches are not mistakes. They are often more informative than other catches. To utilize the rich information contained in these catches we use a separate model specifically designed for this purpose.

Before analysis we do not know which large catches are inconsistent with nearby smaller catches. Visual inspection of spatial plots of the data are very useful for this purpose, but analytical methods are still required because other covariates, specifically tow depth, are also important, and we cannot visually inspect trawl catches for covariate dimensions greater than two. For this we use robust nonparametric regression to identify the large discordant catches. Robust residuals are used for this purpose. It is well known that nonparametric regression techniques are quite sensitive to outliers, and the standard non-robust fitting techniques will “mask” the discordant catches. A simple example of this is given in **Chapter 5**. Robust nonparametric regression techniques are much more useful for this purpose. Once the discordant catches have been identified the nonparametric model is then fitted to the data with the discordant catches removed.

We propose to use a very simple model for the number and size of catches discordant with a smooth mean function of covariates model. The probability of a discordant catch is assumed to be constant over time, but vary spatially over three large geographic regions within 3NO. This is consistent with life-history information about cod in this region. The mean and variance of the discordant catches are assumed to be constant over time and throughout 3NO. Part of the rationale for this assumption is that the number of discordant catches we identify is quite small, and we simply do not have enough information to fit a more complicated model than the one we have chosen. The small number of large catches

does not mean that they have no influence however.

It is easy to use the model for stock size inferences. Each year the nonparametric model is used to predict trawl catches at a large number of locations. These predictions are merely weighted averages of the observed catches for the year in question, minus those that are identified as discordant. The predictions are combined with predictions for the discordant catches. The combination is also a weighted average, where the weighting involves the probability of a discordant catch. The end result is a set of weights used to estimate total trawlable stock abundance, and in this respect our inferential methods are identical to the design-based approach. Note that the weights used in the two approaches may be very different however.

This dissertation is orientated towards the solution of an important practical problem that involves rather complex data. To reach this solution a variety of statistical methods are employed. In important analyses we have exercised great care to ensure that our estimators are as efficient as possible. For example, the nonparametric regression procedures were selected to provide a reasonable summary of trawl catches while maintaining as many “error degrees-of-freedom” as possible. Our main contributions to statistical theory involve practical issues in nonparametric variable selection, and fitting nonparametric regression models to over-dispersed count data. We have developed useful theoretical results for robust estimation with Negative Binomial data, and robust smoothing of count data. We have also developed inferential procedures for the mixture model. We have tested some of our inferential results with simulated data and found them to be good for the samples sizes in our time series. Our main contribution to fisheries theory is a method more suitable for summarizing trawl survey catches than stratum area-weighted averages, which is the current method used. We have also developed a method for dealing with large errant catches. We have assimilated a wide range of fisheries literature relevant for stock size inferences. We have also explored the importance of this literature for 3NO cod,

and where possible we have attempted to quantify the covariate relationships suggested in this literature. In doing so we were able to assess the importance of these relationships in terms of inferences about annual stock size.

The mixture model is an important contribution. It is an implementation of a simplified version of a habitat-selection model in which fish move from less favorable areas to more favorable areas as space becomes available; hence, annual stock density trends in less favorable areas are expected to be different than in more favorable areas. The model estimates, especially in 1991-1993, are more realistic than those from the design-based approach. We anticipate that the results in this dissertation will lead to improved estimators of stock size in other regions. For example, 87% of the survey biomass for 3PS (just west of 3NO; see Figure 1.3) cod in 1995 occurred in one out of more than 100 tows, and this has led to considerable uncertainty in assessing the history of this stock. The methods developed in this dissertation may improve this situation substantially.

1.1. Data source: Trawl research surveys

Trawl research surveys are conducted annually by the Canadian Department of Fisheries and Oceans (DFO). Data are collected by towing a standard research trawl (see Figure 1.2) at a constant speed for a fixed duration of time. The measure of stock size we considered is based on the number of cod caught per tow; however, a variety of other abundance and biological information for many species is also collected. Essentially the unit of measurement is a three-dimensional quadrat sample of relative stock abundance. Measurements are relative because not all fish in the path of a trawl are caught, and a portion of the local abundance will be above the trawl and also not caught. The time period chosen for analysis is 1984-1994. This period was selected because relatively few changes in the survey design occurred during these years.

The survey design used in 3NO since 1974, and in all areas that DFO surveys in the

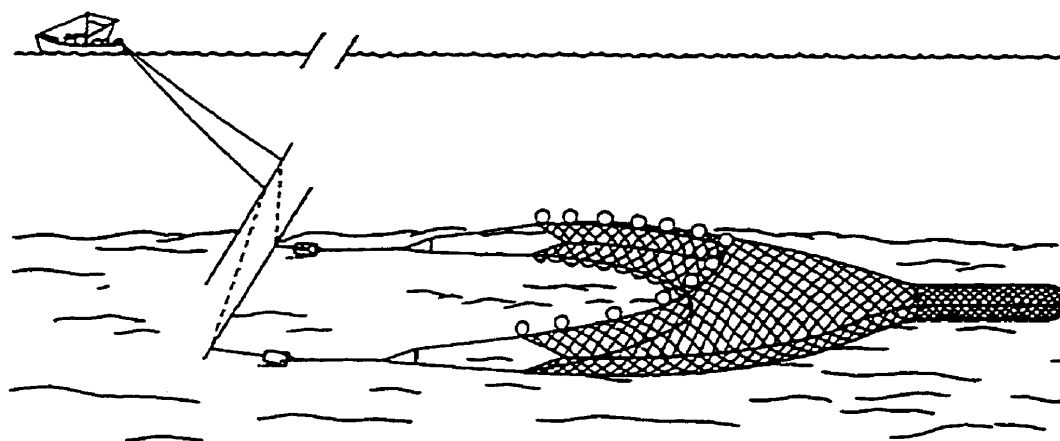


Figure 1.2: A trawler towing an otter trawl.

Northwest Atlantic Ocean, is stratified random sampling. Doubleday (1981) presents a comprehensive manual on the survey methodology. The strata boundaries for NAFO Divisions 3N and 3O are presented in Figure 1.4 in the **Appendix (Section 1.3)**. These boundaries are primarily based on depth, which is important in delineating distributional patterns of some species (Pitt *et. al.*, 1981). Each time a station is trawled many variables are measured, such as the time and location of a tow, light conditions, wind and sea states, depth of a tow, water temperature, and the number and weight of every species caught. Samples of the catch are taken to measure individual fish lengths, whole and gutted weights, liver and gonad weights, stomach contents, sexual and maturity status, and presence of parasites. Otoliths (ear bones) are also taken for age determinations in a lab. We restrict our attention only to the numbers caught per tow; however, clearly the surveys provide much more information than this.

1.2. Outline

We begin our investigations in **Chapter 2** by briefly outlining the history of the cod fishery on the Grand Bank, as well as the basic biology of cod in this region. In this chapter we also study common methods used to estimate the size of fish stocks. The advantages and disadvantages of many of these methods will be illustrated.

Design based methods for estimating stock population parameters from research survey data are often used in practice, and are considered in **Chapter 3**. The details of the 3NO survey design are first presented, followed by methods of inference for total trawlable stock abundance.

Chapter 4 involves a graphical overview of the survey data. The purpose of this chapter is to present graphic explorations of potential relationships in the data. Many of the plots focus on relationships that have been suggested in the fisheries literature on similar stocks. These plots will also be used to develop and interpret a stock dynamics model in **Chapters 5** and **6**.

Based on the previous graphical investigations a simple mixture model for trawl catches is postulated in **Chapter 5**. This involves a smooth mean function of covariates model which describes most (Type I) catches. A small number of isolated, very large, (Type II) catches are described by a different model that is a function of space only. Estimates from this model will be based on the average stock density for the two mixtures, as well as their areal extent. We compare the fits from this model with those using strata information only.

In **Chapter 6** we use the mixture model to estimate annual trawlable stock abundance. This estimation is based on predictions of trawl catches at many trawling stations. In this chapter we also develop and use methods for constructing confidence intervals for trawlable stock size. The estimates and confidence intervals are compared with those obtained in an analogous design-based approach.

A summary of the material in **Chapters 1-6** is presented in **Chapter 7**, along with future research considerations.

1.3. Appendix: Figures

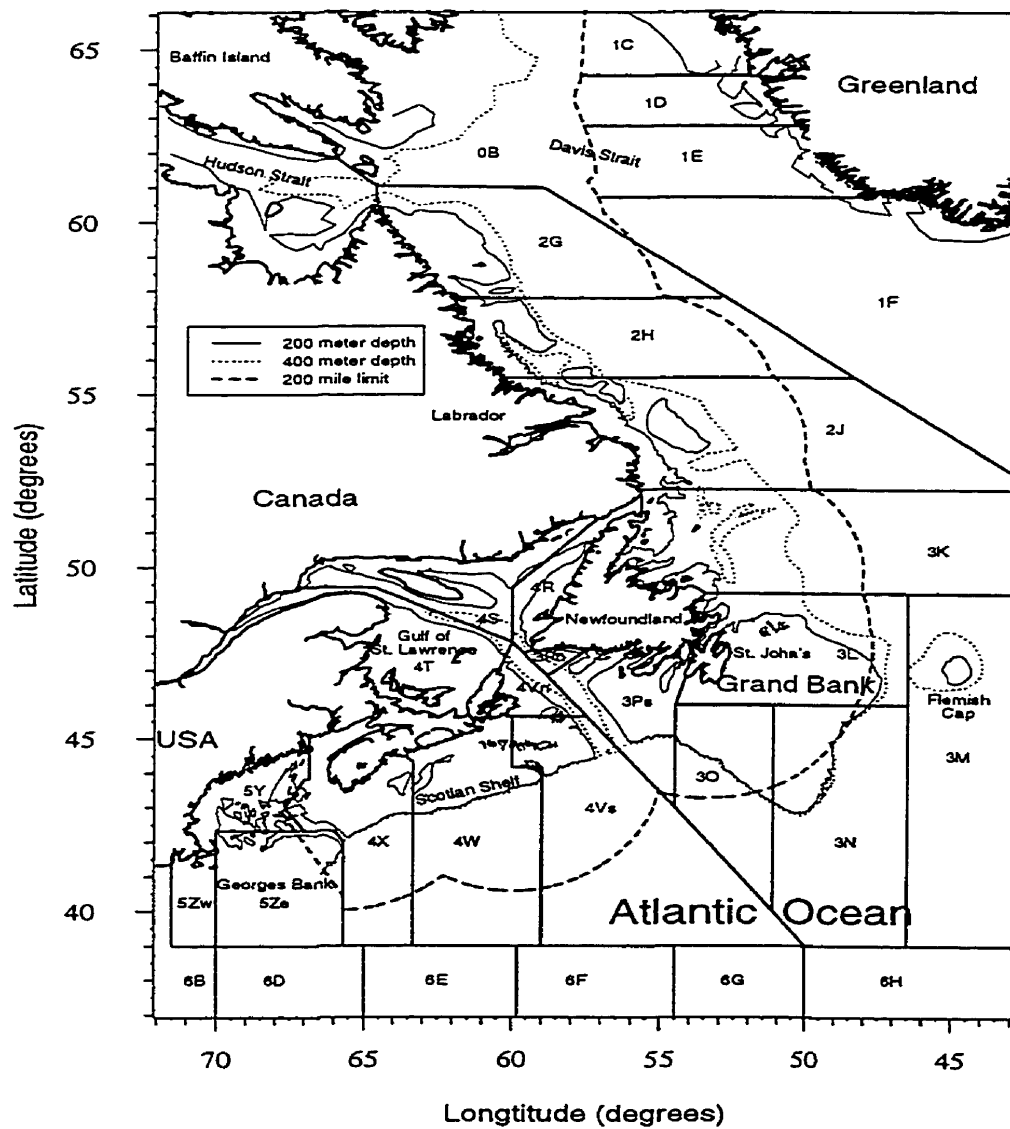


Figure 1.3: Some NAFO management units in the Northwest Atlantic ocean.

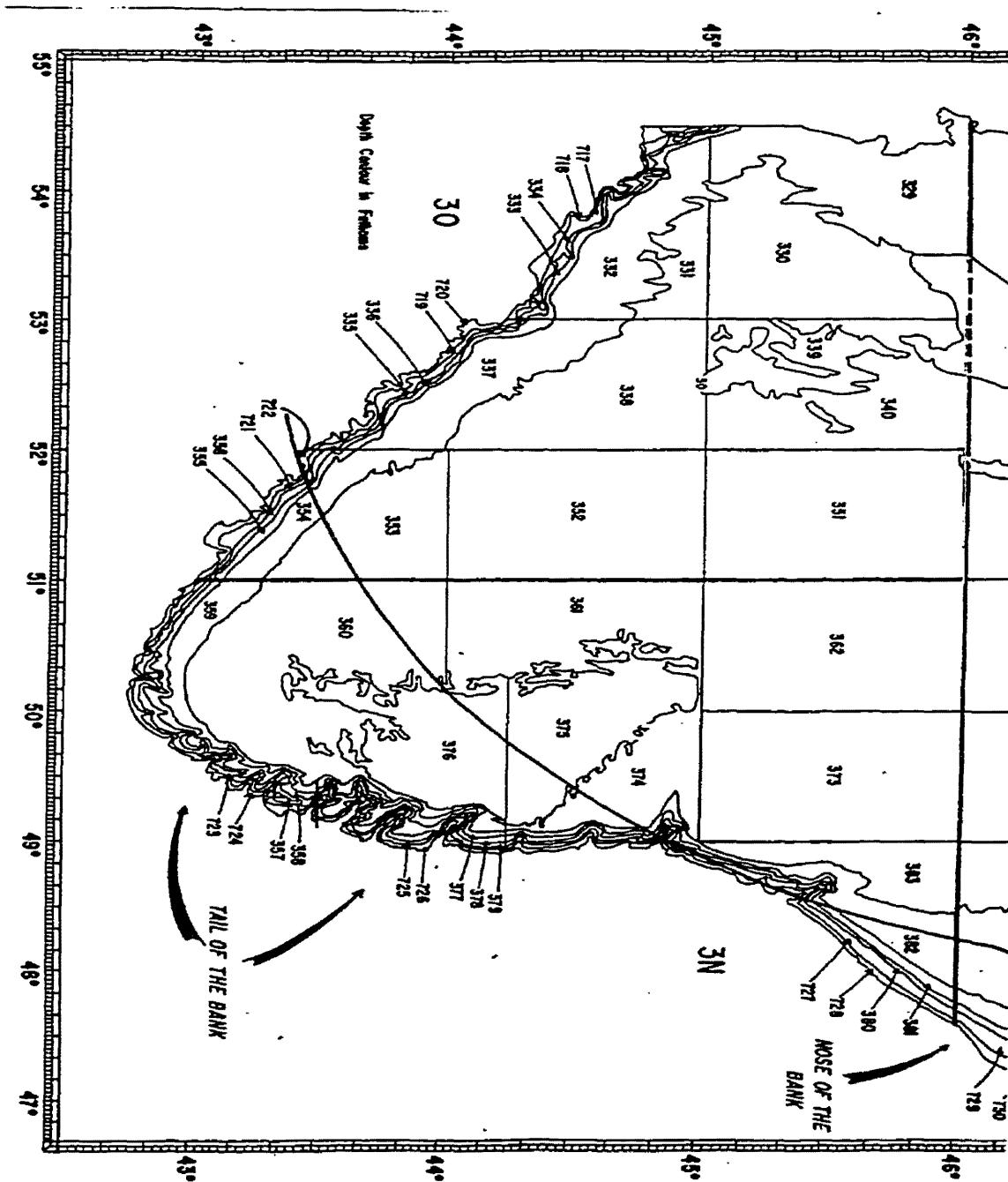


Figure 1.4: Depth stratification scheme for NAFO Divisions 3N and 30. The map orientation is north to south. The axes are in degrees of latitude and longitude. One fathom is 1.8 meters.

2. REVIEW OF SOME FISHERIES TOPICS

2.1. Introduction

In this chapter we review several fisheries topics that provide background information, and a context, to assess the methods developed in this thesis. The topics involve the history of the cod fishery on the Grand Bank (**Section 2.2**), basic biological information about cod (**Section 2.3**), and analytic methods for assessing fish stocks (**Section 2.4**). The first topic is of general interest and provides information on the historic size of the Grand Bank cod fishery. The second topic provides some life history information on cod and is useful when considering population dynamics models for this species. In data analysis it is generally useful to understand as much as possible about the subject of measurement, and this rule applies to 3NO cod. Some of the information presented in **Section 2.3** and **2.4** will be drawn upon when motivating the use of models in **Chapters 4-6**. This information is also useful in understanding the way these fish are surveyed. The third topic provides the motivation for the remainder of the dissertation, and illustrates how the methods we develop may assist when assessing the absolute size of a fish stock such as cod.

2.2. Brief history of the Newfoundland cod fishery

Records of cod fisheries in Newfoundland waters exist for the early sixteenth century (Innis, 1954) and fisheries probably existed even before then. France, Spain and, to a

limited extent, Portugal first started a Newfoundland fishery early in the 1500s. England became involved later. During the 1600s and 1700s the British and the French controlled the Newfoundland fishery. During the years 1600 to 1750 approximately 300 French vessels visited Newfoundland each year. Approximately 250 English vessels visited Newfoundland each year in the first half of the 1600s, and about 100 vessels each year in the second half (Cushing, 1989). While the Grand Bank fishery expanded in this time period the bulk of the total Newfoundland catch came from the inshore fishery; however, later in the 1700s declining catch rates prompted the English and resident fishery to move northwards, and the English fishery also moved to Labrador coastal areas. By the 19th century the English fishery in Newfoundland had diminished, and largely the residents of the Island and the French carried out the Newfoundland fishery.

In the early 20th century France and Newfoundland were the dominant nations fishing in Newfoundland waters, although France's total catch rarely exceeded half of the total catch by Newfoundland. Most of Newfoundland's catch still came from the inshore. Until the 1950s the fisheries in the Northwest Atlantic were largely based on cod; however, the advent of the factory freezer trawler resulted in a diversification of the fishery to other species (Pinhorn and Halliday, 1990). The reported landed catch of 3NO cod (Stansbury *et. al.*, 1995) 1954-1995 is presented in Figure 2.1. Catches peaked at 227 000 tonnes in 1967, but declined till the late 1970s. Following the establishment of a 200 mile resource jurisdictional limit in 1977 catches increased till the mid 1980s and have declined steadily since. In 1995 a moratorium was established on commercial fishing in 3NO. The moratorium is still in effect. The trends in cod catches in 3NO also appear in other stocks (Pinhorn and Halliday, 1990).

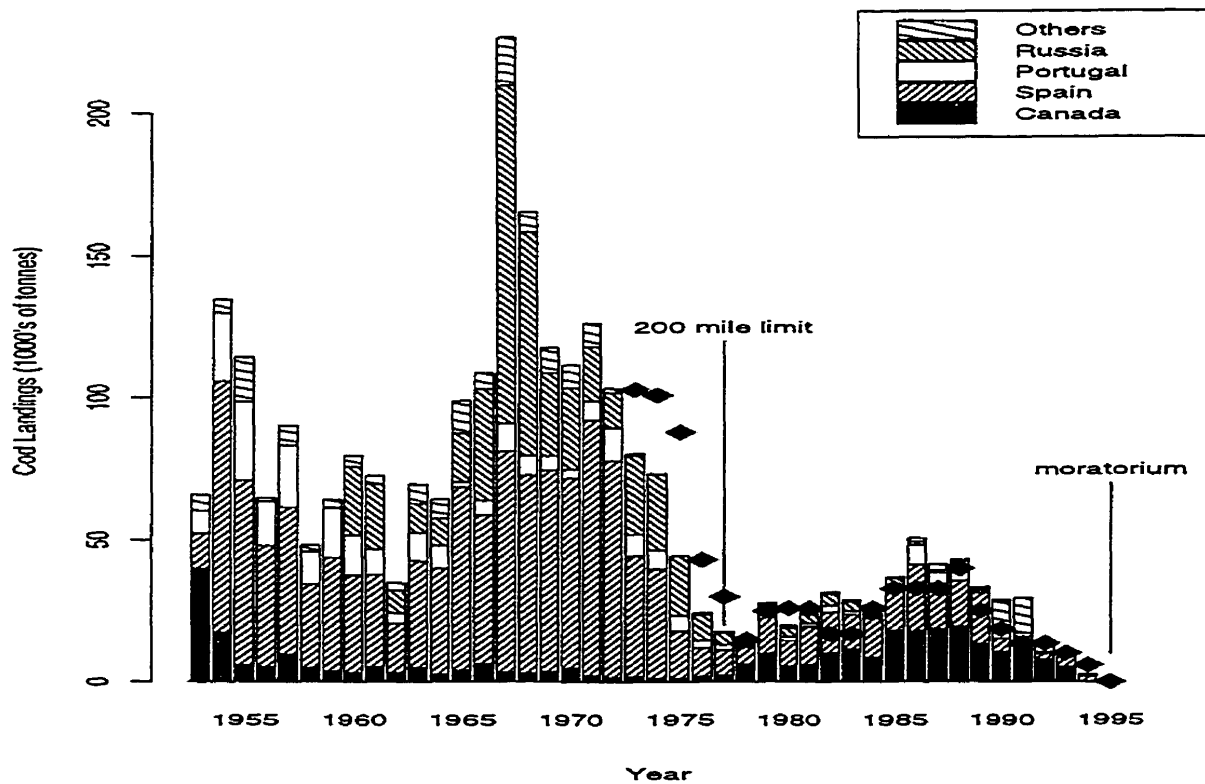


Figure 2.1: Cod landings in NAFO divisions 3N and 3O by country. The NAFO total allowable catch quota is plotted as ◆'s.

2.3. General biology of Atlantic cod

In this section we present a brief description of the biology and ecology of Atlantic cod, especially those in 3NO, which are the subject of this thesis. This information will be drawn upon occasionally throughout this dissertation, for model building purposes. An important concept is the degree of mixing between cod in 3NO and cod in other Divisions. We do not specifically use this assumption in the methods we develop; however, we refer to estimates as being of stock abundance, when they are really estimates of the abundance of fish within the 3NO region. The relationship between the abundance of cod in 3NO, i.e. those cod within the 3NO region, and 3NO cod stock abundance depends on the affinities between this stock, and other stocks, with the 3NO geographical region.

Atlantic cod is a bottom dwelling (demersal) species that ranges in distribution from Greenland down the Northwestern Atlantic continental shelf as far south as North Carolina in the United States (Scott and Scott, 1988). Cod are also found in the Northeastern Atlantic, from Iceland southward to northern Spain. Atlantic cod are found over a wide range of depths, from 0-1 kilometers. Females may lay up to 12 million eggs, but only a few survive. Females generally spawn at depths less than 400 meters and water temperatures of 3-4° Celsius. Spawning occurs as early as February in northern regions and as late as December in southern regions.

After the cod eggs are fertilized and hatch, the young (larval) cod remain in the water column for some time but eventually descend to the ocean bottom where they grow and feed on a variety of small crustaceans. As cod get older they continue to feed on small crustaceans but eventually fish, including smaller cod, become the predominant food source. Generally, cod grow faster at lower latitudes because of warmer water temperatures. The weight of cod also varies with year, age, and location. Cod can weigh up to 80 kilograms and live as long as 26 years, but commercially caught cod in the Northwest Atlantic during the 1980's were typically 5 or 6 years old.

2.3.1. Stock designation of Southern Grand Bank (3NO) Atlantic cod

The cod found on the southern portions of the Grand Bank are generally thought to comprise a stock. A fish stock has been defined by Cushing (1981, p 70) as those fish that return year after year to the same geographical region to spawn. This is essentially the same definition found in Templeman (1962, 1979). The stock concept will be important for understanding population models for 3NO cod, so we review this idea in some detail.

An early account of why cod on the southern Grand Bank comprise a stock is given by Fleming (1960). He found that:

1. fish tagging experiments “gave no evidence of inter-migration of cod between the southwestern part of the bank and the northeastern part”;
2. vertebral length averages “indicated a distinct difference between cod from the two regions”;
3. the otolith (ear bone) is somewhat different in cod from the two regions;
4. the “southwestern Grand Bank cod generally lag behind the northeastern ones in reaching sexual maturity”;
5. the “average sizes for the various age-groups of cod from the southwestern Grand Bank were generally much higher than for the corresponding age-groups of the northeastern cod”.

Fleming (1960) hypothesized that a cold layer of the Labrador current “which is in contact with the northern, eastern, and western edges of the Grand Bank at intermediate depths forms an effective barrier which prevents fish from the deeper waters of the northern area from mixing with the fish of the shallower water on the body of the bank to the south”. However, tagging experiments indicated some mixing of Southern Grand Bank cod with cod in 3Ps (see Figure 1.2).

Templeman (1962) presented evidence for the stock division of Southern Grand Bank cod:

1. the stock has a “low vertebral number; over 1 vertebral lower than the Labrador-Newfoundland stock”;

2. the stock has a low rate of infection by the parasitic copepod *Lernaeocera branchialis* compared to nearby coastal and more northerly offshore areas.

Templeman (1962) also presented evidence of other low vertebral number cod found on the northern portion of the Grand Bank which indicated the potential existence of two low vertebral number cod stocks, or that the Southern Grand Bank cod stock might range into more northern areas from time to time. Templeman (1974) presented further evidence from tagging studies where cod tagged on the southeastern part of the bank were mainly captured on “the southern half or two-thirds of the Grand Bank”, although some tagged fish were recovered on the northern portion of the bank, in 3Ps, and also in southeastern coastal areas of Newfoundland. Templeman (1981) studied cod vertebral numbers and found that the “most consistently low vertebral means, and the lowest proportions of high and highest proportions of low vertebral numbers” were found in the Southern Grand Bank cod. Templeman found that in “the northwestern part of Division 3O, the northeastern part of Division 3N and along the slope in Division 3L there was intermingling with high vertebral-count fish of the northern stock”.

The results in this section show that cod on the Southern Grand Bank usually have similar morphological characteristics, and this suggests that these fish come from the same spawning unit. The results also demonstrate that mixing of this stock with other stocks to the north and west does occur, particularly along the northern 3L boundary and the western 3Ps boundary. This information will be useful when interpreting anomalous estimates of total stock size in some years, which seem to be driven by large catches in the boundary regions of 3NO.

Delineation of 3NO boundaries

So far we have not addressed why the 3NO boundaries were constructed, and what they have to do with the Southern Grand Bank (3NO) cod stock. This is important if our

inferences are to really pertain to 3NO cod, and not merely that fraction of 3NO cod, and cod from other stocks, that were surveyed within the 3NO boundaries.

Halliday and Pinhorn (1990) describe the evolution of the NAFO Divisions from the early 1930s. The current Divisions closely resemble an earlier scheme developed in the 1950s. These Divisions were based on uniformity of size, important fishing grounds and fish stocks, simplified reporting of statistics, and conformity with existing areas. It seems that the 3Ps and 3O division boundaries were based on the distribution of important haddock stocks, while the division between 3N and 3O was based more on important commercial fisheries than any stock information.

This report does not indicate that the Division boundaries of 3NO were directly based on the distribution of Southern Grand Bank cod. This suggests that it is reasonable to expect stock mixing, especially in the boundary regions with 3Ps and 3L, and this is what the tagging studies reviewed previously have suggested. The delineation of stock boundaries is a controversial issue, especially in 3Ps, and we can do little with it here. As a matter of convenience we assume for now that our inferences are directly related to 3NO stock size. Removing this assumption will not affect our methodology, but it does affect the nature of our conclusions; that is, we may in fact be estimating the abundance of cod in 3NO, and not the abundance of the Southern Grand Bank stock. Note that this assumption is critical for most of the methods used in **Section 2.4**, because the models there assume that the reported commercial catch of cod taken in 3NO comes exclusively from a single stock.

2.3.2. General biology of the Southern Grand Bank (3NO) Atlantic cod

The Southern Grand Bank Atlantic cod inhabit the southern slope of the Grand Bank in winter and disperse northwards across the surface of the bank in the spring (Pinhorn, 1976; p 7). Spawning occurs from the latter half of April to early June (Templeman, 1979).

Cod often aggregate densely during this time. The larvae from the northern and eastern portions of the bank drift slowly southward and are retained on the southern Grand Bank, along with larvae from spawning on this bank (Templeman, 1979). Generally small cod (<40 centimeter) feed on crustaceans, medium sized cod (40–69 centimeter) feed on capelin and sand lance, and large cod (> 90 centimeter) feed more on flatfish (Gomes, 1993). The sand lance (*Ammodytes dubius*) is relatively more important in the diets of 3NO cod than of cod in more northern regions.

Templeman (1966) presented some growth data for 3NO cod. A five year-old cod averaged 50.5 centimeters in length and weighed one kilogram (gutted, head on). A twelve year-old cod during the same time period weighed on average six kilograms and measured 89 centimeters in length.

2.4. Some stock size assessment methods

Methods for assessing the size of a fish stock are reviewed in this section, with attention to the incorporation of probabilistic mechanisms. A stock assessment is concerned with the provision of advice to various management agencies about the state of a fish stock (Gulland, 1983; Hilborn and Walters, 1992).

2.4.1. Data sources

Many data sources are routinely used in stock assessments, but the most important data usually involve commercial catch and effort. Commercial catch data typically consists of the total weight of fish caught by the commercial fishery, with some sampling of the catches for lengths and ages. Effort data consists of the number of hours fished, number of vessels, etc. These data are used to compute relative indices of stock abundance and are also used in population models.

Another important source of information comes from annual research surveys. Trawl

survey data has already been discussed; however, other types of survey data are also routinely used in some stock assessments. Examples include fixed fishing gear surveys, hydroacoustic surveys, egg and larvae surveys, and, for some species, sighting and aerial surveys (see Pope, 1988; Shepherd, 1988).

A variety of experimental data sets may be used in stock assessments. Tagging data can be used to estimate stock movements and commercial fishery exploitation rates. Studies on fish condition, growth, reproductive abilities, etc. are occasionally conducted, and their results incorporated in assessments.

2.4.2. Total catch and effort analysis

The total size of commercial catches depends on the fishing effort expended by the fleet; that is, if there is no fishing then there will be no commercial catch. Effort is measured in many ways; for example, number of hours fished or number of vessels. Commercial catch and effort are basic quantities used in the assessment of many fish stocks. In the next several sections we outline some basic assessment models that use catch and effort data. The models require varying amounts of data, ranging from total catch and effort for a single year to long time series of catch-at-age and effort. These models can provide a range of information about stock size, ranging from relative estimates to absolute estimates, depending on the amount of data available and the type of fishery.

Catch per unit effort

Catch divided by effort is commonly referred to as catch per unit effort (CPUE), and is often thought to give a proportional index of stock abundance; that is, an estimate of relative stock size. It can be used as follows. Let U denote the area of a region to be surveyed, which is stratified into subregions of homogeneous fish density with known areas U_i . Let N denote the total number of fish in region U , and define N_i similarly for

region U_i . If D_i is the density of fish over U_i , then

$$N = \sum_i N_i = \sum_i U_i D_i. \quad (2.4.1)$$

Both the N_i 's and D_i 's are unknown to us, so this equation gives no information about stock size; however, it suggests a way to use CPUE to estimate relative stock size. Let C_i be the catch in U_i that resulted from known effort E_i . We assume that the fishery catch rates are proportional to local density. The proportionality is because fishing gears are usually less efficient in catching small fish, etc. This assumption may be stated as $qD_i = C_i/E_i$, where q is the unknown catchability of the fishery. Substituting this result into (2.4.1) gives

$$qN = \sum_i \frac{U_i C_i}{E_i}, \quad (2.4.2)$$

where the only unknown in (2.4.2) is qN , which is often referred to as trawlable stock size; hence, CPUE, through the use of (2.4.2), can be used to estimate relative (trawlable) stock size.

If different gear and vessel types are used then some standardization of catch and effort data is usually employed (see Gavaris, 1980; Richards and Schnute 1992). The use of CPUE is often criticized because C_i/E_i may not be proportional to D_i (see Paloheimo and Dickie, 1964; Gulland, 1983; Hilborn and Walters, 1992; Arreguín-Sánchez, 1996). This is particularly problematic for schooling species like cod, because fishing fleets can find schools and maintain high catch rates for a wide range of stock sizes. When accurate information is not available about commercial catches or effort, which is often the case, then other indices of stock size are required to estimate relative stock size. This is the case for 3NO cod, and we will use research vessel catch rates in a similar manner as above.

Biomass production models

When annual time series of commercial catch and effort (see the previous section for definitions) are available then it is possible to estimate both stock size and the catchability of the fishery. This is a potentially major improvement over the previous approach because it yields information on absolute, rather than relative, stock size. Estimates of absolute stock size are usually required for effective stock management.

In this approach stock size is expressed in terms of biomass, and biomass dynamics models (Hilborn and Walters, 1992) are used for estimation. These models are usually of the form

$$\begin{aligned}\text{next biomass} &= \text{last biomass} + \text{recruitment} + \text{growth} - \text{catch} - \text{natural mortality}, \\ &= \text{last biomass} + \text{surplus production} - \text{catch}.\end{aligned}$$

Let B_y denote the unknown biomass in year y , then

$$B_{y+1} = B_y + g(B_y) - C_y,$$

where $g(B_y)$ represents recruitment+growth−natural mortality. This can be a positive or negative amount. Hilborn and Walters (1992) considered $g(B_y) = rB_y(1 - B_y/k)$, where r and k are unknown parameters. Let $X_y = CPUE = C_y/E_y$. Hilborn and Walters (1992) rearranged this model into a linear form by assuming that $B_y = X_y/q$:

$$\frac{X_{y+1}}{X_y} - 1 = r - \frac{r}{qk}X_y - qE_y.$$

The parameters r , k , and q can be estimated using least squares.

Recently Freeman and Kirkwood (1995) and Reed and Simons (1996) considered Kalman filter approaches to estimate parameters. Their stochastic models included both

measurement error in CPUE indices and process error in the stock production model. Freeman and Kirkwood (1995) considered $X_y = qB_y + \varepsilon_y$ where ε_y is normal with mean zero and variance σ^2 . Reed and Simons (1996) considered $\log(X_y) = k + p \log(B_y) + \varepsilon_y$.

Other approaches use less empirical error assumptions. They postulate stochastic dynamics for fish capture by fishing gear, and use the dynamics to generate the distribution of commercial catches. de la Mare (1984) assumed a Poisson counting process to model commercial catches, and found “extra-Poisson” variation in 41 of 42 time series on CPUE for whales. Reed (1986) considered a simple Poisson process for commercial catches. Reed and Simons (1996) modelled the fishery as a Poisson process with a contagion effect for fish capture. Bishir and Lancia (1996) explicitly showed how probabilistic models for catches can be used for inferences; they also considered a more general catch/effort model that is useful for open populations.

Regardless of how the stochastic nature of the data is modelled, biomass production models provide only limited stock size information because they do not consider the age composition of the stock. They provide little information on stock dynamics. Realistic production models usually contain parameters which can not be estimated, or poorly so, based only on catch and effort data. In addition, effort data is often not available, and when it is available the quality is often poor. This is the case for 3NO cod. These methods do not receive much attention in the assessment of cod stocks off the East Coast of Canada. Methods based on age-structured models are generally thought to be superior to production models.

2.4.3. Catch at age analyses

The previous methods relied upon relatively basic information about stock abundance, namely total commercial catches and fleet effort. When a time series of annual age distribution information is also available then better estimators of stock size can be constructed.

Age disaggregated models are primarily used in North Atlantic fish stock assessments. The simplest method is *virtual population analysis* (VPA), in which a minimum estimate of historic stock numbers at age is the sum of subsequent catches from each cohort. An extension of this method is called *sequential population analysis* (SPA). These methods only utilize commercial catch data. Methods for the incorporation of effort data into stock size estimation methods are also available, as are related methods based on CPUE or research survey data.

Let a and y denote age and year standardized so that $a = 0, \dots, A$ and $y = 0, \dots, Y$. Let t denote time within a year, $0 \leq t \leq 1$. Let N_{ay} and C_{ay} denote stock numbers and commercial catches at age a in year y . The N_{ay} 's are usually unknown.

Virtual population analysis (VPA)

The *virtual population* is the minimum population that must have existed so that the catches $C_{0y-A}, \dots, C_{Ay}$ could occur; that is, there must have been at least $\sum_{a=0}^A C_{ay-A+a}$ fish in a cohort in year $y - A$. A *virtual population analysis* (VPA) simply involves summing cohort catches, and produces a lower bound on the size of cohorts when first exploited by a fishery. The lower bound is usually too conservative for cohorts that are still heavily fished. This approach provides no information on cohorts that are not fished. This is a concern for stocks under fishing moratoria, where the younger age classes have not been fished. This is the case for 3NO cod.

Sequential population analysis (SPA)

The absolute size of historical cohorts can be estimated by incorporating natural mortality rates of fish into a cohort model. A common exponential decay model is:

$$N_{a+1y+1} = N_{ay} e^{-\int_0^1 Z_{ay}(t) dt}, \quad (2.4.3)$$

where $Z_{ay}(t)$ is the total unknown mortality rate. This rate is assumed to be separable into $Z_{ay}(t) = F_{ay}(t) + M_{ay}(t)$, where $F_{ay}(t)$ and $M_{ay}(t)$ are fishing and natural mortality rates. Usually $M_{ay}(t) = M$, whose value is obtained from other data. If fishing mortality is constant throughout the year then it can be written in terms of annual catches as

$$\frac{C_{ay}}{N_{ay} - N_{a+1\ y+1}} = \frac{F_{ay}}{Z_{ay}}. \quad (2.4.4)$$

In (2.4.4) $N_{ay} - N_{a+1\ y+1}$ is the age-specific total decrease in stock size from year y to $y+1$. The ratio of C_{ay} to the total decrease is assumed to equal the ratio of fishing mortality to total mortality. When N_{ay} and C_{ay} are known the solution of F_{ay} in (2.4.3) and (2.4.4) can be obtained numerically. It is often approximated using

$$N_{a+1\ y+1} = (N_{ay}e^{-M_{ay}/2} - C_{ay})e^{-M_{ay}/2}. \quad (2.4.5)$$

Remark 1. In what follows let $\{X\}_{I_l:I_u\ J_l:J_u} = \{X_{ij} : i = I_l, \dots, I_u, j = J_l, \dots, J_u\}$. If only one value is given for a subscript range then it is implicitly understood that the subscript is fixed at that value. For example, $\{X\}_{0\ J_l:J_u} = \{X_{ij} : i = 0, j = J_l, \dots, J_u\}$.

When $\{C\}_{0:A\ 0:Y}$ and $\{M\}_{0:A\ 0:Y}$ are known, then equations (2.4.3) and (2.4.4) can be used to derive $\{N\}_{0:A\ 0:Y}$ given $\{N\}_{0:A\ 0}$ and $\{N\}_{0\ 1:Y}$, or given $\{N\}_{A\ 0:Y-1}$, $\{N\}_{0:A\ Y}$ using the backward solution:

$$N_{a\ y} = (N_{a+1\ y+1}e^{M_{ay}/2} + C_{ay})e^{M_{ay}/2}. \quad (2.4.6)$$

The backward solution is used more often. For example, consider a population of five ages, modelled over 7 years. We represent this model as a 5×7 matrix:

$$\begin{array}{cccccc}
 \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & N_{17} \\
 \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & N_{27} \\
 \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & N_{37} \\
 \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & N_{47} \\
 N_{51} & N_{52} & N_{53} & N_{54} & N_{55} & N_{56} & N_{57}
 \end{array}$$

In this matrix the N 's are the unknowns, and the $\nwarrow$'s represent quantities that can be obtained in terms of the N 's, commercial catches, and mortalities; for example, N_{46} is obtained by plugging in N_{57} , M_{46} , and C_{46} into (2.4.6). The $\{N\}_{0:A,Y}$ are the survivors and usually must be estimated. The $\{N\}_{A,0:Y-1}$ are often modelled in terms of survivors. For example, assuming $F_{Ay} = F_{A-1,y}$ gives $N_{Ay} = N_{A-1,y}C_{Ay}/C_{A-1,y}$, and this relationship can be used recursively to express the $\{N\}_{A,0:Y}$ in terms of survivors. For the above example, the cohort matrix is filled out like

$$\begin{array}{cccccc}
 \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & N_{17} \\
 \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & N_{27} \\
 \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & N_{37} \\
 \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & N_{47} \\
 \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & N_{57}
 \end{array}$$

The $\downarrow$ arrow indicates that N_{Ay} is obtained using $\{N\}_{0:A-1,y}$. The survivors cannot be estimated from catch data alone because the method of solving for F_{ay} 's uses all available "degrees of freedom" in the data. It is the survivors, or the current population age distribution, that is of most interest, and required for management purposes; hence, an SPA on its own is not that useful. Several solutions have been proposed that involve the

use of statistical models for estimation purposes.

Separable sequential population analysis

One of the simplest methods used to provide relative information about survivors involves, in addition to (2.4.3), an assumption about the separability of fishing mortality; that is,

$$F_{ay} = s_a f_y, \quad (2.4.7)$$

where s_a is a parameter accounting for the differential exploitation of age classes in the commercial fishery and f_y is a parameter that accounts for yearly differences in the overall stock exploitation by the fishery. Here we assume again that fishing mortality is constant throughout the year. The separability assumption reduces the number of unknown fishing mortalities in (2.4.4) from AY to $A+Y$. This allows the catches to also be used to estimate stock numbers at age.

The separability assumption (2.4.7) can be used with (2.4.3) and (2.4.4) to express catches in terms of stock numbers,

$$C_{ay} = N_{ay} s_a f_y (1 - e^{-Z_{ay}}) / Z_{ay},$$

where $Z_{ay} = M_{ay} + s_a f_y$. Recursive substitutions of $N_{ay} = N_{a-1, y-1} e^{-Z_{ay}}$ can be used to show that

$$C_{ay} = s_a f_y N_{0, y-a} e^{-\sum_{i=0}^{a-1} Z_{i, y-a+i}} (1 - e^{-Z_{ay}}) / Z_{ay}. \quad (2.4.8)$$

When the catches and mortalities are known then $\{s\}_{0:A}$, $\{f\}_{0:Y}$, $\{N\}_{0:A,0}$, and $\{N\}_{0,1:Y}$ can be estimated using (2.4.8). Estimation is required because (2.4.8) will not hold exactly. There are a total of $2(Y+A)+3$ unknown parameters to estimate. Using these estimates

we can compute all the $\{N\}_{0:A,0:Y}$. Many estimation procedures have been proposed. For example, Doubleday (1976) used least squares based on log catches because he felt the log transformation would stabilize error variance, which he assumed to be proportional to the mean. Pope and Shepherd (1982) considered a similar estimation procedure. Shepherd and Nicholson (1986) examined the estimation of (2.4.8) within the GLIM setting, where the $E(C_{ay})$ is modelled as the right hand side in (2.4.8).

Shepherd and Nicholson (1986) show that two of the parameters in a *separable SPA* are not identified without additional information. Many kinds of parameter constraints have been proposed. For example, Shepherd and Nicholson (1986) suggested constraining $s_0 = f_0 = 1$. Deriso *et. al.* (1985) constrained the s_a 's by modelling them in terms of a and two unknown parameters using a gamma function, where the gamma curve was scaled so that s_a was one for older ages. Gudmundsson (1994) constrained s_a 's so that $\sum_{a=0}^{a_m < A} \log(s_a) = 0$, and $s_{a \geq a_m} = s$. These procedures are arbitrary and produce only relative estimates of stock size. For example, if we fixed $s_o = f_o = 0.1$ then estimates of stock size will be 10 $\times$ that obtained if we fix $s_o = f_o = 1$. The *separable SPA* is useful for trend information only.

In summary, the assumption about the separability of fishing mortality produces relative estimates of current stock status, whereas an *SPA* without this assumption produces no such information - not even reliable trend information for cohorts still heavily fished. As mentioned in **Section 2.4.2** what we really want are absolute estimates of current stock size. The solution to this problem is often obtained by incorporating additional information into a model to produce reliable and robust absolute estimates of current stock size.

2.4.4. Catch and effort analysis

In this section we review methods involving an age-disaggregated *separable SPA* in which effort is used to reduce the number of unknown commercial exploitation parameters, f_y 's.

The annual effort, E_y , of the commercial fishing fleet in obtaining catches can be used in a *separable SPA* to eliminate the f_y unknown parameters by assuming that $f_y \propto E_y$; hence, $F_{ay} = s_a E_y$, the s_a is different than in (2.4.8) because it absorbs the proportionality constant. If Z_{ay} in (2.4.8) is small so that $\log \{(1 - e^{-Z_{ay}})/Z_{ay}\} \doteq -Z_{ay}/2$, and also $s_a = s$ and $M_{ay} = M$ then (2.4.8) can be written as a linear model (Paloheimo, 1980) by substituting sE_y for F_{ay} :

$$\log \left(\frac{C_{ay}}{E_y} \right) = \gamma_{y-a} + sX_1(a, y) + MX_2(a),$$

where $\gamma_{y-a} = \log(sN_{0y-a})$, $X_1(a, y) = -\sum_{i=0}^{a-1} E_{y-a+i} - E_y/2$, and $X_2(a) = -(a + 1/2)$. Recruitment estimates to the fishery are obtained using $N_{0k} = e^{\gamma_k}/s$. Paloheimo (1980) advocated estimation based on least squares using $\log$ CPUE.

Several papers have been published where the stochastic nature of the catch/effort data has been incorporated in various ways. Fournier and Archibald (1982) assumed that the annual total catch was lognormally distributed, and that the catch at age conditional on total catch was multinomially distributed. They used a log-linear stochastic model for fishing mortalities in terms of effort and a parametric function for age effects, but also included year effects; that is,

$$\log(F_{ay}) = \beta_1 g(a) + \beta_2 g^2(a) + \log(f_y E_y) + \delta_y,$$

where $\{\delta\}_{0:Y}$ are iid normal random variables, and $g(a)$ is a known function of a . This is also basically the same approach taken in Deriso *et. al.* (1985). The stock synthesis approach (see Methot, 1989) incorporates additional sources of data beyond effort. For

example, Method's (1989) Northern Anchovy example incorporated auxiliary data related to M_{ay} 's and s_a 's.

While effort may be used to reduce the number of unknown parameters in a *separable SPA*, it is still the case that such estimates produce only **relative** estimates of stock size because one parameter is still not identified based on catch data alone. For reasons already discussed we require methods to produce absolute estimates of stock size. Methods will be discussed in the following section in which the unknown *SPA* parameters are estimated using an index of stock abundance, i.e. an index proportional to stock size, such as standardized CPUE or survey estimates of trawlable stock abundance. This results in absolute estimates of stock size.

2.4.5. Relative indices of abundance

A wide variety of stock size indices are used to assist in the estimation of absolute current stock size using *SPA*'s. An index R_{ay} is some measure that is assumed to be related to stock size in some quantifiable way. Many indices are used, and their relationships with stock size are diverse. One example is the average catch from a research survey. The basic relationship usually involves an assumption about proportionality.

The analysis of relative abundance indices usually proceeds with the stochastic model

$$g(R_{ay}) = g(q_{ay}N_{ay}) + \epsilon_{ay}, \quad (2.4.9)$$

where g is a known link function, q_{ay} is the unknown constant of proportionality between R_{ay} and N_{ay} , and ϵ_{ay} is a stochastic error term. The q_{ay} 's and N_{ay} 's are functions of a smaller number of unknown parameters; for example, $q_{ay} = q_a$ for all y . If N_{ay} is defined by (2.4.3) and (2.4.4) and the F_{Ay} 's are somehow constrained in terms of survivors, as is usually the case, then $N_{ay} = N_{ay}(\{N\}_{0:A \setminus Y})$; that is, all the N_{ay} 's can be computed using just the numbers in the final year (Y). Note that many other types of functions for the

q_{ay} 's and N_{ay} 's are routinely used. The unknown parameters are usually estimated by least squares or maximum likelihood.

Many approaches that fall within the framework of (2.4.9) have been published. Collie and Sissenwine (1983) considered the identity function for the g 's and estimators that minimized $\sum_y \sum_a \epsilon_{ay}^2$ plus a process error component in their age-structured population dynamics model. Mendelssohn (1988) also considered the identity function for g and assumed additive errors, but he specified an objective function based on the notion that the underlying population process (the N_{ay} 's) is stochastic and should be treated as missing data. An adaptive framework encompassed by (2.4.9) is used within DFO (see Gavaris, 1988), and is based on the unseparated *SPA*. In this framework g is usually either the identity or log function, and errors are additive. Several abundance indices are routinely used in the objective function. The error terms are assumed to be independent and identically distributed, and estimates are produced that minimize a weighted error sum of squares. Myers and Cadigan (1995a) have generalized this procedure somewhat by allowing the ϵ_{ay} 's to be correlated within-years.

CPUE indices

Commercial catches standardized by fleet effort, or catch per unit effort (CPUE), is often used as a relative index of stock abundance, and we have mentioned some problems associated with this use. Gavaris (1988) described two difficulties:

1. "potential bias arising from the non-random distribution of fishing locations", and
2. "the amalgamation of information from all cohorts necessitating the use of indirect methods to obtain abundance at age".

The first problem has proven to be quite insidious for schooling species because catch rates may not reflect abundance if changes in abundance correspond to changes in the number and size of schools but not the density of fish within a school. In this case efficient fish finding devices may enable catch rates to remain high even as fish densities decline.

The important errors in commercial catches are often more systematic than stochastic; for example, unreported discarding of fish at sea. These types of errors are difficult to accommodate statistically. Although this problem will affect an *SPA* no matter how parameters are estimated because an *SPA* is based on commercial catches, the effect of the systematic catch errors will be greater if estimates are based on CPUE abundance indices. These problems are hopefully not present when using research survey relative abundance indices if the survey follows a reasonable protocol and has adequate coverage.

Research survey indices

The use of research survey data to provide relative indices of stock abundance has become standard within DFO. This is the motivation for studying research survey data in this dissertation. The relative abundance indices used are simply estimates of population mean trawlable stock size, and it is this quantity that we focus on estimating in the remainder of the dissertation. We have shown how estimates of trawlable stock size can be used to produce estimates of absolute stock size; however, precise stochastic descriptions of trawlable stock size are currently not available, and this has naturally lead to difficulties in estimating absolute stock size.

Many of the problems with using CPUE data are either absent or reduced when using survey data; for example, misreporting of catches can be eliminated. Also, changes in catchability due to technology can be controlled to yield more reliable relative abundance indices; however, changes in catchability due to biological and hydrographic reasons may still exist.

In recent years some difficulties have arisen in *SPA* analyses using survey data. One difficulty is year effects in the indices that are unrelated to stock size. These can lead to misleading inferences (Myers and Cadigan, 1995a,b). Another problem involves a systematic change over time, or a retrospective pattern, in population estimates from

SPA regressions with research survey data. That is, quite often population estimates for a particular year decline as additional years of survey data and catches are used to produce the estimates (see Figure 1 in Sinclair *et. al.*, 1991; Figure 6 in Myers *et. al.*, 1997). Note that this problem cannot solely be attributed to the survey indices; it could also be related to deficiencies in the commercial catch data or other assumptions.

As with all methods relying on commercial catches, difficulties arise for stocks under commercial fishing moratoria, such as 3NO cod. However, absolute estimates for such stocks can still be produced using the relationship between known historic stock size from *SPA* and historic survey indices, and applying this relationship to current survey indices (or whatever indices are used). With the *separable SPA* this was not possible.

Despite these problems the method of estimating *SPA* parameters using research survey indices is one of the most common methods used to estimate stock sizes in the North Atlantic ocean, and certainly off the east coast of Canada. It is hoped that the methods we develop will improve *SPA* estimation by providing a precise stochastic description of the research survey indices that stock size estimation is based on.

2.4.6. Catch at length analysis

We would be remiss not to mention that length-structured methods are also used in fish stock assessments, especially in tropical fisheries where age-structured data may be difficult to obtain. Lengths are more easily measured for fish than are ages; in fact, age compositions are often estimated from known length compositions of catches and estimated length to age conversion formulae constructed from samples of the catch. The fundamental methods used in length-structured assessments are reviewed by Rosenberg and Beddington (1988). Examples of length-structured stock assessments methods are given in Fournier and Doonan (1987) and Sullivan (1992). A basic difference in length-structured models compared to age-structured models is that from year to year fish lengths

increment differently for different individuals, whereas ages increment by one for all individuals. Annual length-structured models require length transition matrices.

2.4.7. Future considerations

A diverse array of methods for estimating stock size have been presented. These methods range from a simple *VPA* to complicated models involving multiple data sources and associated stochastic relationships with underlying (latent) stock dynamics. The major improvements in stock assessment models historically have involved the addition of new data that reflect some aspect of the fish stock, such as the incorporation of research survey relative abundance indices. This trend will persist in the future. The demand for risk-based stock management will require even better understanding of the stochastic relationships between fish stocks and the data available for their assessment. This is another motivation for studying research survey data in this dissertation. New sources of data will likely involve improved research surveys that more completely cover the range of a stock, as well as more detailed commercial catch statistics. In addition, multi-species approaches to stock assessments will require joint modelling of fish stocks, their prey, and their predators.

3. SURVEY DESIGN AND INFERENCE

3.1. Introduction

A simple and often effective statistical method for analyzing survey data is the survey sampling approach. This is the method currently used by DFO to produce summary statistics from research survey data. In the sampling approach we assume that the population consists of stationary sampling units. When a unit is selected for sampling we measure characteristics about the unit. Statistical inference about population characteristics is then based on how the samples are collected, and the observed values in the sample. To save costs, improve precision, and sometimes for practical considerations, various survey designs are often used in different applications. In **Section 3.2** we give a detailed description of the design used in the 3NO research surveys, and also give the rationale for this design.

Once the sample has been collected we need to choose a standardized statistical analysis method. In the sampling literature there are two broad classes of analysis methods: design- and model-based. The assumptions under the design-based approach are that the population consists of fixed units. We obtain random outcomes because the sample units are selected randomly. The model-based approach assumes that the responses from the sampling units are themselves random. The additional randomness from the design is less relevant in this approach. Recently, discussions have appeared in the sampling literature about model-assisted approaches, which are model-based and yet also have some good

design-based properties. It is hoped that such an approach will be good when the model is correct, and not fail badly when the model is not valid. In this chapter we focus on how to construct a design-based estimate of a population total, and confidence intervals for this quantity, using the 3NO survey design.

In **Section 3.3** a brief account of the design-based approach is given. This section describes the Horvitz-Thompson (HT) estimator, and related quantities. The HT estimator, or expansion estimator, gives an algorithm for constructing a design-unbiased estimator based on a known probability sampling design (see **Section 3**). In **Section 3.4** we apply this approach to the 3NO survey data. Unfortunately unbiased estimators of the variance of HT estimators do not exist for the 3NO survey design. This is because some pairs of units in the population have zero probability of appearing in the sample. We discuss the utility of the design-based approach for the 3NO survey design in **Section 3.5**. Although we do not recommend the design-based approach for the analysis of the 3NO survey data, we do feel that such an approach provides useful benchmarks to assess model-based results developed in later chapters. The conclusions from this chapter are presented in **Section 3.6**.

3.2. 3NO design

In this section we present the basic rationale behind the survey design used in 3NO, along with the relevant design details required to construct inferences. Common variations in the design are also discussed.

Stratified random surveys began on the Grand Bank in 1971, and good coverage throughout strata has generally been achieved since 1980. A stratified design was adopted to improve the precision of estimators by exploiting trends in groundfish abundance with hydrographic and/or bathymetric conditions (Doubleday, 1981). Strata are based primarily on depth because depth zonation is important in delineating distributional patterns

for many species, although cod have a relatively wide depth range (Pitt *et. al.*, 1981). The areal divisions of strata were made to correspond to NAFO division boundaries. The strata boundaries along the top of the bank were generally drawn along lines of latitude and longitude. The Southeast and Southwest slopes of the bank are strongly influenced by northward flowing Gulf Stream water; hence, strata in these regions were made as small as possible to facilitate detailed analyses of trawl catches. Boundaries for many of the strata are presented in Figure 1.4; however, additional strata (covering depths to 1.4 kilometers) have been added to the survey since this map was drawn. Relevant strata information is presented in Tables 3.1a,b in the **Appendix B (Section 3.8)**; these tables are based on Doubleday (1981) and Bishop (1994).

Timing is also important in the survey design. The survey should take place at the same time of the year, or even better at the same migration stage of important stocks. This is to ensure that the same portion of the stock is surveyed from year to year so that the annual survey results are comparable. The seasonal timing of the survey in 3NO throughout 1984-1994 is presented in Figures 4.2a,b in **Chapter 4**. The timing has primarily been determined by the survey vessel availability, and this has varied somewhat throughout the years. The annual survey duration is also important. The surveys should be conducted as quickly as possible to minimize the effects of fish movements. Such movements can result in counting schools of fish more than once, and missing schools that should have been detected. The survey design should also be balanced throughout all hours of a day because of potential day/night vertical stock migrations. In 3NO this has approximately occurred because the survey vessel has operated on a 24-hour basis throughout 1984-1994; hence, approximately the same number of day and night tows occur from year to year.

Bottom type is an important determinant in the distribution and abundance of a stock (Doubleday, 1981). Unfortunately, the survey gear cannot be towed on rough bottoms and

this introduces a potential source of bias if stock density is not the same in trawlable and untrawlable regions. Ideally the best way to deal with this is to delineate the rough bottom areas in 3NO, and then restrict sampling and inferences to trawlable areas; however, maps of rough bottom areas do not exist. In practice, if an untrawlable site (station) is selected in a sample then a nearby trawlable station is substituted; hence, the presence of untrawlable areas also affects the probability that a trawlable site is sampled. The sample selection procedure is discussed in the next section.

Attempts are made to standardize the survey measurement procedures. The vessel speed is kept at 3.5 knots (nautical miles per hour), or 6.5 kilometers per hour. The duration of a tow is almost always 30 minutes; hence, a tow covers a distance of 1.75 nautical miles, or 3.24 kilometers. The net width is generally determined by the net wing spread, which is usually taken at the nominal value of 45 feet, or 13.7 meters; hence, the nominal swept area of a tow is 44388 meter². Later in this chapter the swept area is used to approximate the design weights. In reality the wing spread varies from tow to tow, and has been shown to generally increase with depth (see Koeller, 1991); hence, swept area varies, and probably increases with depth. This implies that the design weights also vary, and the variation may be systematic with depth.

The direction of a tow is determined so that the tow depth is maintained at a constant level, and otherwise the tow direction is along the course leading to the next survey station.

3.2.1. Sample selection

The procedures for determining where to sample in 3NO are presented in this section. This essentially defines the survey design. Let H be the number of strata in 3NO, and let n be the number of tows in a year. Define n_h similarly for stratum h , and let A_h be the surface area of this stratum. The amount of vessel time that is available determines n . At least two tows are allocated to each stratum, after which the remaining samples

are allocated to strata in proportion to A_h . The station location is determined randomly within a stratum using a three-stage procedure.

Each stratum is subdivided into primary units that are approximately 120 Km². Each primary unit is further broken down into 10 equal area secondary units, which are called fishing units. These units are used to determine the location of a tow. The number of fishing units in each stratum is presented in Tables 3.1a,b. The three-stage selection procedure is as follows:

Stage 1. Randomly select without replacement n_h of the primary units.

Stage 2. Randomly select one fishing unit from each of the selected primary units.

Stage 3. Select a tow station within the fishing unit.

Stage 3 is determined at sea. Typically the location and direction of a tow within a fishing unit is chosen to maintain a constant tow depth and speed, and to be as close as possible to the midpoint of the fishing unit. The rationale for the three-stage selection procedure is that nearby fishing units will have similar fish densities and, consequently, more information can be gained by spreading out the sampled tow stations.

3.3. Overview of design-based inferences

The framework for design-based inferences is presented in this section. We begin by considering a simple design, then show how the Horvitz-Thompson (HT) estimator can be used for unbiased estimation with an arbitrary probability sampling design. We then apply the general procedures to the 3NO sampling design. In addition, we consider a useful variance decomposition for the HT estimator of a population total. This decomposition illustrates a problem with the 3NO sampling design. Finally, we illustrate problems with variance estimation for the 3NO design using a simple population model.

Consider a finite population of N elements labelled $1, 2, \dots, N$. Let $\mathcal{U} = \{1, \dots, N\}$ denote the elements in the population. Associated with each element in the population is a measurement y_i , for which the population total is $T = \sum_{i=1}^N y_i$. All of the elements in the population cannot be measured, so T is estimated using a simple random sample without replacement (SRSWOR) of n elements. An intuitively reasonable estimator of T is $\hat{T} = N\bar{y}$, where $\bar{y}$ is the sample mean.

A probability sampling scheme is essential for reliable design-based inferences. Särndal *et. al.* (1992; pg. 8) define a probability sample as

1. a set of samples, $\mathcal{S} = \{s : s \subset \mathcal{U}\}$, that are possible to obtain with the sampling procedure,
2. a known probability of selection, $p(s)$, associated with each sample s ,
3. a nonzero probability of selection for every element in the population, and
4. one sample is selected using a random sampling mechanism through which each s receives exactly the probability $p(s)$.

Design-based inferences are made with $p(s)$. For example, with SRS there are $\binom{N}{n}$ possible samples to take, each with equal probability. It is easy to show in this case that $E(\hat{T}) = T$, and $Var(\hat{T}) = N(N - n)S^2/n$, where $S^2 = \sum_{i=1}^N (y_i - \bar{Y})^2 / (N - 1)$ is the population variability in the y 's. Confidence intervals are quite often used to summarize the uncertainty in $\hat{T}$.

A confidence interval $CI(s) = [T_L(s), T_U(s)]$ is a random interval such that $T_L(s) \leq T_U(s)$ for every s (Särndal *et. al.*, 1992). The coverage probability, or the confidence level, of the interval is

$$\Pr[CI(s) \ni T] = \sum_{s \in \mathcal{S}_o} p(s),$$

where $\mathcal{S}_o = \{s : s \in \mathcal{S}, p(s) > 0, \text{ and } T \in [T_L(s), T_U(s)]\}$. The required coverage

probability is pre-specified, and is usually close to one. A pivotal statistic for T is usually needed to construct an exact confidence interval, but is often not available for most parameters, and approximate methods are required. Asymptotic-normal and bootstrap methods are two common procedures used to produce approximate confidence intervals.

If the sampling distribution of $\hat{T}$ is asymptotically normal with mean T and variance $Var(\hat{T})$, and $v(\hat{T})$ is a consistent estimator for $Var(\hat{T})$, then $CI(s) = \hat{T} \pm Z_{1-\alpha/2} \sqrt{v(\hat{T})}$ is a confidence interval with approximate coverage probability $1 - \alpha$, where $Z_{1-\alpha}$ is the upper α th quantile from a standard normal distribution. Hence, variance estimation is important in setting asymptotic confidence intervals. The accuracy of the approximation will be greatly influenced by the skewness of the distribution of $y_1, \dots, y_N$, among other things.

Bootstrap methods rely on approximating the usually unknown distribution of $\hat{T}$ with a distribution generated from the observed sample, and inferences are based on this approximate distribution. Efron and Tibshirani (1993) give a general introduction to the bootstrap, and Booth *et. al.* (1994) consider the application of the bootstrap to finite population sampling problems.

3.3.1. General design inferences

In this section inferential procedures for an arbitrary probability sampling design are presented, followed by specific procedures for stratified simple random sampling (SSRS) and stratified three-stage cluster sampling (S3CS) designs. The latter design is most similar to that currently used by DFO, while the summary statistics currently produced by DFO are based on the first design.

The relevant design information is contained in $p(s)$; however, many calculations can be performed more easily using the sample inclusion probabilities for the population elements. The inclusion probability for the i th element in $\mathcal{U}$ is the probability that this element

appears in the sample. That is, for all $s \in \mathcal{S}$

$$\pi_i = \Pr(s \ni i) = \sum_{s: i \in s} p(s).$$

Another way of defining π_i is through the sample indicator variable

$$I_i = \begin{cases} 1, & i \in s, \\ 0, & \text{otherwise.} \end{cases}$$

π_i is merely the $\Pr(I_i = 1)$, or equivalently $E(I_i)$. In a similar manner, define the joint inclusion probability for the i th and j th elements in $\mathcal{U}$ as

$$\pi_{ij} = \Pr(i \& j \in s) = \sum_{s: i, j \in s} p(s).$$

Note that $\text{Var}(I_i) = \pi_i(1 - \pi_i)$, and $\text{Cov}(I_i, I_j) = \pi_{ij} - \pi_i\pi_j$.

3.3.2. Horvitz-Thompson estimator of T

The Horvitz-Thompson (HT) estimator of T is given by

$$\hat{T} = \sum_{i \in s} y_i / \pi_i.$$

The HT estimator is sometimes referred to as the π expansion estimator (for more discussion see Särndal *et. al.*, 1992). This estimator is widely used. By construction $\hat{T}$ is

design unbiased; that is, $E(\hat{T}) = T$. The variance of $\hat{T} = \sum_{i=1}^N I_i y_i / \pi_i$ is

$$\begin{aligned} Var(\hat{T}) &= \sum_{i=1}^N \left(\frac{y_i}{\pi_i} \right)^2 Var(I_i) + \sum_{i=1}^N \sum_{j \neq i}^N \left(\frac{y_i y_j}{\pi_i \pi_j} \right) Cov(I_i, I_j), \\ &= \sum_{i=1}^N \left(\frac{y_i}{\pi_i} \right)^2 \pi_i (1 - \pi_i) + \sum_{i=1}^N \sum_{j \neq i}^N \left(\frac{y_i y_j}{\pi_i \pi_j} \right) (\pi_{ij} - \pi_i \pi_j). \end{aligned}$$

For designs with fixed sample sizes $\pi_i(1 - \pi_i) = -\sum_{j \neq i}^N (\pi_{ij} - \pi_i \pi_j)$, and this can be used to show that

$$Var(\hat{T}) = \frac{1}{2} \sum_{i=1}^N \sum_{j=i}^N \Omega_{ij} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = - \sum_{i=1}^N \sum_{j=1}^N \Omega_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}, \quad (3.3.1)$$

where $\Omega_{ij} = \pi_i \pi_j - \pi_{ij}$.

Provided all $\pi_{ij} > 0$ an unbiased estimator of $Var(\hat{T})$ is

$$v(\hat{T}) = \frac{1}{2} \sum_{i=1}^n \sum_{j=i}^n \omega_{ij} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2,$$

where $\omega_{ij} = \Omega_{ij} / \pi_{ij}$. The notation v is used to denote an estimator of Var . No unbiased estimator of $Var(\hat{T})$ exists if any of the $\pi_{ij} = 0$. This is the case for the 3NO research survey design, as will be shown later.

As an example, consider HT estimators for a SRSS design. Let $N_{h..}$ denote the size of stratum h , and let $n_{h..}$ denote the corresponding sample size. This notation is used to be consistent with the next section. In this case $\pi_i = n_{h..} / N_{h..}$ when observation i is in stratum h . Also, $\pi_{ij} = n_{h..}(n_{h..} - 1) / N_{h..}(N_{h..} - 1)$ when $i \neq j$ are in the same stratum. Otherwise, $\pi_{ij} = \pi_i \pi_j$. The HT estimator is

$$\hat{T}_s = \sum_{h=1}^H N_{h..} \bar{y}_h,$$

and

$$Var(\hat{T}_s) = \sum_{h=1}^H (N_{h..} - n_{h..}) N_{h..} S_h^2 / n_{h..},$$

where $S_h^2 = \sum_{i=1}^{N_{h..}} (y_i - \bar{Y})^2 / (N_{h..} - 1)$. The same formula is given in Cochran (1977).

3.3.3. Stratified three-stage sampling

Recall that the survey design used in 3NO is basically stratified three-stage cluster sampling (S3CS); that is, a trawl station is (i) selected from a fishing unit, which is (ii) randomly selected from a primary unit, which is also (iii) randomly selected in a stratum. Each primary unit has ten contiguous fishing units; hence, a primary unit is a geographic cluster of fishing units. Fishing units are not subdivided into disjoint trawl stations, as was discussed in Section 3.2.1; however, the traditional approach used by DFO to analyze trawl survey data is based on assuming disjoint trawl stations within fishing units are sampled. In this chapter we also assume disjoint trawl stations within fishing units; hence, we consider a fishing unit to be a geographic cluster of trawl stations. This is the finite-population sampling design that most resembles the 3NO survey design. Design-based inferences should take into account the sampling restrictions imposed by the primary and fishing units, which is why we consider the S3CS design. In this section HT estimators of T , denoted as $\hat{T}_c$, are derived for a S3CS design. The subscript c is used to indicate that this estimator is based on a cluster sampling design.

Some notation for a multi-stage design is introduced next. Let N_h denote the number of primary sampling units in stratum h , and let $\mathcal{U}_h = \{1, \dots, N_h\}$. Define N_{hi} and $\mathcal{U}_{hi}$ similarly for secondary sampling units within the i th primary unit, and let $\mathcal{U}_{hij}$ be the set of population elements in the j th secondary unit within the i th primary unit. In terms of the 3NO survey design, $\mathcal{U}_{hij}$ is assumed to be a finite population of disjoint trawl

stations within the j th fishing unit in the i th primary unit in stratum h . Let n_h , n_{hi} , and n_{hij} be the size of the samples taken from $\mathcal{U}_h$, $\mathcal{U}_{hi}$, and $\mathcal{U}_{hij}$. For the 3NO survey design $n_{hi} = n_{hij} = 1$ for all i, j . We develop results for more general designs than stratified three stage one-per-cluster sampling (S3C₁S) because we wish to investigate the performance of two variance estimators for S3C₁S, and compare these results with other S3CS designs.

The sampling theory to be developed is based on the assumption that strata and clusters are sampled independently, and with the same design regardless of which other units are sampled. The primary unit inclusion probabilities are $\pi_{hi} = n_h/N_h$, $i \in \mathcal{U}_h$. The secondary unit inclusion probabilities conditional on the i th primary unit being selected are $\pi_{j|hi} = n_{hi}/N_{hi}$, $j \in \mathcal{U}_{hi}$. The population element inclusion probabilities conditional on the j th secondary unit being selected within the i th primary unit are $\pi_{k|hij} = n_{hij}/N_{hij}$, $k \in \mathcal{U}_{hij}$. The unconditional inclusion probabilities for the population elements can be computed based on the independence assumptions, and are:

$$\pi_k = \pi_{hi}\pi_{k|i} = \pi_{hi}\pi_{j|hi}\pi_{k|hij}, \quad k \in \mathcal{U}_{hij}.$$

Second order inclusion probabilities can be found in a similar manner. All we require to compute HT-estimates of T , and associated variances, are the first- and second-order inclusion probabilities; however, it is useful to further simplify these estimators for computational and theoretical purposes.

The HT estimator of T is

$$\hat{T}_c = \sum_{k \in s} y_k / \pi_k = \sum_{h=1}^H \hat{T}_h.$$

That is, $\hat{T}_c$ is merely the sum of estimated stratum totals. Note that $\hat{T}_h = \sum_{i \in s_h} \hat{T}_i / \pi_i$ is also a weighted sum of estimated primary unit totals, for those units selected for sampling (s_h). Because strata are sampled independently from each other $Var(\hat{T}_c) =$

$\sum_{h=1}^H Var(\hat{T}_h)$. We show in **Appendix 3 (Section 3.7)** that $Var(\hat{T}_h)$ can be written as

$$Var(\hat{T}_h) = V_{h1} + V_{h2} + V_{h3},$$

where

$$\begin{aligned} V_{h1} &= N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{T_h}^2}{n_h}, \\ V_{h2} &= \frac{N_h}{n_h} \sum_{i=1}^{N_h} N_{hi}^2 \left(1 - \frac{n_{hi}}{N_{hi}}\right) \frac{S_{T_{hi}}^2}{n_{hi}}, \text{ and} \\ V_{h3} &= \frac{N_h}{n_h} \sum_{i=1}^{N_h} \frac{N_{hi}}{n_{hi}} \sum_{j=1}^{N_{hi}} N_{hij}^2 \left(1 - \frac{n_{hij}}{N_{hij}}\right) \frac{S_{hij}^2}{n_{hij}}. \end{aligned}$$

$S_{T_h}^2$ and $S_{T_{hi}}^2$ are the population variances of the primary and secondary unit totals; that is,

$$\begin{aligned} S_{T_h}^2 &= \sum_{i=1}^{N_h} (T_{hi} - T_h/N_h)^2 / (N_h - 1), \text{ and} \\ S_{T_{hi}}^2 &= \sum_{j=1}^{N_{hi}} (T_{hij} - T_{hi}/N_{hi})^2 / (N_{hi} - 1). \end{aligned}$$

S_{hij}^2 is the population variance within secondary sampling units,

$$S_{hij}^2 = \sum_{k=1}^{N_{hij}} (y_k - \bar{Y}_{hij})^2 / (N_{hij} - 1).$$

This expression for $Var(\hat{T}_c)$ is partitioned into variability related to primary and secondary sampling unit totals, and population variability within secondary units. We show in the **Appendix A in Section 3.7** that an unbiased estimator of $Var(\hat{T}_c)$ is $v(\hat{T}_c) = v_1 + v_2 + v_3$,

where $S_{T_h}^2$, $S_{T_{hi}}^2$, and $S_{hi_j}^2$ are replaced by their sample equivalents,

$$\begin{aligned} s_{T_h}^2 &= \sum_{i=1}^{n_h} (\hat{T}_{hi} - \hat{T}_h/N_h)^2 / (n_h - 1), \\ s_{T_{hi}}^2 &= \sum_{j=1}^{n_{hi}} (\hat{T}_{hij} - \hat{T}_{hi}/N_{hi})^2 / (n_{hi} - 1), \text{ and} \\ s_{hi_j}^2 &= \sum_{k=1}^{n_{hi_j}} (y_{hjk} - \bar{y}_{hi_j})^2 / (n_{hi_j} - 1). \end{aligned}$$

Variance estimation with one-per-cluster sampling

The problem with sampling primary units only once is that no information is available about $S_{T_{hi}}^2$ or $S_{hi_j}^2$. In this section the bias for two estimators of $Var(\hat{T}_c)$ is considered. One estimator has been proposed by Särndal *et. al.* (1992) when within-cluster variability is unknown. The other estimator is based on assuming simple random sampling within strata. This is the estimator currently used by DFO.

The first estimator we consider is

$$v_1 = \sum_{h=1}^H \sum_{i \in s_h} \sum_{i' \in s_h} \left(\frac{\pi_{hii'} - \pi_{hi}\pi_{hi'}}{\pi_{hii'}} \right) \frac{\hat{T}_{hi} \hat{T}_{hi'}}{\pi_{hi} \pi_{hi'}},$$

and was suggested by Särndal *et. al.* (1992), especially when $n_h \ll N_h$. Under S3SC sampling it can be shown that the $bias(v_1) = -\sum_{h=1}^H n_h(V_{h2} + V_{h3})/N_h$. If $n_h \ll N_h$ then the bias relative to $Var(\hat{T}_c)$ will usually be small. For a single stratum design the relative bias is

$$\begin{aligned} \frac{bias(v_1)}{Var(\hat{T}_c)} &= \frac{n}{N} \left\{ 1 + \frac{(N-n)S_T^2}{\sum_{i=1}^N N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{T_i}^2}{n_i} + \sum_{i=1}^N \frac{N_i}{n_i} \sum_{j=1}^{N_i} N_{ij}^2 \left(1 - \frac{n_{ij}}{N_{ij}}\right) \frac{S_{ij}^2}{n_{ij}}} \right\}^{-1} \\ &= O\left(\frac{n}{N}\right). \end{aligned}$$

This variance estimator has the undesirable property that it can be negative for some samples (see Särndal *et. al.*, 1992, p 47). In addition, for the 3NO surveys it often occurs that $n_h = N_h$, and this approximation is not that useful. This is because many strata have only 2-5 primary units ($\#$ fishing units/10; see Table 3.1a,b), and the sample size allocated to these strata is the same as the total number of primary units; that is, $n_h = N_h$.

If we assume that within-strata samples are collected at random without replacement, then the HT-estimator of T is $\hat{T}_s$, which was presented in the last paragraph in **Section 3.2**. An unbiased estimator of $Var(\hat{T}_s)$ for this design is

$$v^* = \sum_{h=1}^H (N_{h..} - n_{h..}) N_{h..} s_h^2 / n_{h..},$$

where $s_h^2 = \sum_{i \in n_h} (y_i - \bar{y}_h)^2 / (n_h - 1)$. This is the estimator currently used by DFO for estimating the $Var(\hat{T}_c)$. The bias in $v^*(\hat{T}_c)$ is not particularly revealing. In the next section we use a simple random-effects population model to assess the biases in v_1 and v^* . The random-effects model is constructed to have features broadly similar to the 3NO cod stock.

Bias in v_1 and v^* for some simple population models The bias in v_1 and v^* depend on the y values in the population, particularly the variability of responses within primary and secondary sampling units. In this section we assess the bias in v_1 and v^* under a population model that has some similar characteristics to the 3NO cod population. That is, in our model population, observations within secondary units are more similar than those between secondary units, and observations within primary units are more similar than those between primary units. We will show in **Chapters 4 and 5** that 3NO survey catches are correlated spatially. The correlations we impose in our population model have this structure, because primary and secondary units in the 3NO design are geographic clusters of tow stations. For a fixed sampling design and sample size, we

compute model expectations of $bias(v_1)$ and $bias(v^*)$ for specific parameterizations of our model. This gives us some understanding about potential problems with the variance estimator currently used by DFO (v^*), which, in part, motivate us to explore new model-based procedures for inferences about stock size.

We deal with a simple population, and also a simplified *S3CS* design. Recall that the population of trawl stations in the 3NO design are grouped in spatial clusters. Stock abundance at tow stations in the same cluster will generally be more similar than abundance at tow stations in different clusters; that is, the within-cluster correlation in trawl abundance should be positive. We assume that the amount of correlation depends on whether tow stations are in the same fishing unit or primary sampling unit. A random-effects model is used to generate a population with this correlation structure. We feel the correlation structure should be constant across strata; therefore, expected biases will be additive across strata, at least in sign, if the same sampling design is used in every strata. This is because biases across strata are weighted averages of biases within strata, where all the weights are positive. For this reason we only consider a single stratum population.

Consider a random population of y 's, where

$$y_k = \xi_i + \xi_{j|i} + \xi_{k|ij},$$

The ξ 's are independent random variables, where ξ_i is an effect common to all members in primary unit i , $\xi_{j|i}$ is an effect common to all members in the j th secondary unit within the i th primary unit, and $\xi_{k|ij}$ represents the randomness of individuals in primary unit i and secondary unit j . Let E_ξ and Var_ξ denote expectations and variances with respect to the ξ 's. Without loss of generality assume that $E_\xi(\xi_i) = E_\xi(\xi_{j|i}) = E_\xi(\xi_{k|ij}) = 0$ for all i, j, k . Also, $Var_\xi(\xi_i) = \alpha\sigma^2$, $Var_\xi(\xi_{j|i}) = \beta\sigma^2$, and $Var_\xi(\xi_{k|ij}) = \sigma^2$, where $\alpha, \beta \geq 0$.

This model allows for positive correlations in the y 's within clusters; that is,

$$\begin{aligned} \text{Within secondary units:} \quad & \text{Corr}_\xi(y_k, y_{k'}; k, k' \in \mathcal{U}_{ij}, k \neq k') = \rho''' = \frac{\alpha + \beta}{\alpha + \beta + 1}, \\ \text{Between secondary units:} \quad & \text{Corr}_\xi(y_k, y_{k'}; k \in \mathcal{U}_{ij}, k' \in \mathcal{U}_{ij'}, j \neq j') = \rho'' = \frac{\alpha}{\alpha + \beta + 1}, \\ \text{Between primary units:} \quad & \text{Corr}_\xi(y_k, y_{k'}; k \in \mathcal{U}_{ij}, k' \in \mathcal{U}_{i'j}, i \neq i') = \rho' = 0. \end{aligned}$$

This population model is used to help quantify the biases in v_1 and v^* . To simplify formulae we consider a simplified version of *S3CS* in which primary and secondary sampling units are all of the same size; that is, K primary units each contain L secondary units that consist of M sampling units. A simple random sample without replacement (SRSWOR) of k primary units is selected. Within each primary unit a SRSWOR of l secondary units is selected, and within each secondary unit a SRSWOR of m sampling units is selected for measurement. The total number of sampling units in the population is denoted as $N = KLM$, and the total sample size as $n = klm$. Let $f_n = n/N$, $f_k = k/K$, $f_l = l/L$, and $f_m = m/M$. The following results are required to compute model-expected biases:

- $E_\xi(S_{ij}^2) = \sigma^2$,
- $E_\xi[S_{Ti}^2] = \sigma^2 M(M\beta + 1)$,
- $E_\xi[S_T^2] = \sigma^2 LM(LM\alpha + M\beta/L + 1)$, and
- $E_\xi[S^2] = \sigma^2 [1 + \alpha(N - ML)/(N - 1) + \beta(N - M)/(N - 1)]$.

We use $M = 270$ and $L = 10$ in our comparisons. The value for L is always used in the 3NO surveys, and the value of M is typical, being the nominal fishing unit area of 12 Km² divided by the nominal trawl swept area of 0.0444 Km². We consider designs with $K = 3$ and $K = 70$; these values are approximately lower and upper bounds on the

number of primary clusters found in 3NO strata. The sampling designs we explore are

$$\begin{array}{ll}
 K = 70, & n = 8, \quad \text{(i) } k = 8, l = 1, m = 1, \\
 . & . \quad \quad \quad \text{(ii) } k = 2, l = 2, m = 2, \\
 . & . \quad \quad \quad \text{(iii) } k = 1, l = 1, m = 8, \\
 K = 3, & n = 2 \quad \text{(iv) } k = 2, l = 1, m = 1, \\
 . & . \quad \quad \quad \text{(v) } k = 1, l = 1, m = 2.
 \end{array}$$

The values $n = 2$ and 8 are lower and upper bounds on the within-strata number of stations towed in 3NO surveys. Four populations are considered:

$$\begin{array}{ll}
 \text{A: } & \rho''' = 0.1, \quad \rho'' = 0.05, \\
 \text{B: } & \rho''' = 0.5, \quad \rho'' = 0.1, \\
 \text{C: } & \rho''' = 0.9, \quad \rho'' = 0.5, \\
 \text{D: } & \rho''' = 0.99, \quad \rho'' = 0.99.
 \end{array}$$

Note that sampling schemes (i) and (iv) are typical in 3NO. Analysis of spatial autocorrelation in catches presented in **Section 4.1 of Chapter 5** suggest that the population of y values in 3NO may resemble A or B the most.

Recall that $\text{bias}(v_1) = E[v_1 - \text{Var}(\hat{T}_c)]$. Define the model-expected relative bias in v_1 as $\text{erb}_1 = E_\xi[\text{bias}(v_1)]/E_\xi[\text{Var}(\hat{T}_c)]$. This expression can be written as:

$$\text{erb}_1 = -f_k \frac{\beta(1 - f_l)m + 1 - f_l f_m}{1 - f_n + \lambda(\alpha, \beta)},$$

where

$$\lambda(\alpha, \beta) = \alpha(1 - f_k)lm + \beta(1 - f_k f_l)m.$$

If $\alpha = \beta = 0$ (i.e. $\rho'' = \rho''' = 0$) then $erb_1 = -f_k(1 - f_l f_m)/(1 - f_n)$ which is $O(f_k)$. The expected relative biases for the five sampling designs (i-v) and four populations (A-D) described above are presented in Table 3.2a. As expected erb_1 is relatively small for sampling schemes i, ii, and iii in which f_k is small, and this is more so when the within-

Table 3.2a. $erb_1 = \frac{E_{\xi}[v_1(\tilde{T})] - E_{\xi}[Var(\tilde{T})]}{E_{\xi}[Var(\tilde{T})]}$.

Sampling Scheme	Populations			
	A	B	C	D
(i)	-0.11	-0.10	-0.06	-0.00
(ii)	-0.03	-0.02	-0.01	-0.00
(iii)	-0.01	-0.01	-0.01	-0.00
(iv)	-0.62	-0.61	-0.47	-0.02
(v)	-0.32	-0.29	-0.18	-0.00

cluster correlation is large. The results in this table suggest that $v_1(\hat{T}_c)$ can be very poor when n is close to K . In fact, if $n = K$ then $v_1(\hat{T}_c) = 0$ always! Hence, v_1 is a poor variance estimator when the number of primary units sampled is relatively large; that is, when f_k is large.

For our simplified design v^* can be written as

$$v^* = (N - n)Ns^2/n.$$

Under SRSWOR v^* is an unbiased estimator of $Var(\hat{T}_s)$, the SRSWOR variance of $\hat{T}_s = N\bar{y}$; however, v^* is biased for $Var(\hat{T}_c)$, the cluster sampling variance of $\hat{T}_c$, even though $\hat{T}_s = \hat{T}_c$ in our simplified $S3CS$ design. If we express s^2 as

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{Y})^2 - \frac{n}{n-1} \frac{1}{N^2} (\hat{T} - T)^2, \end{aligned}$$

then under the simplified $S3CS$ design it is straight-forward to show that

$$E(s^2) = \frac{n}{n-1} \left\{ \frac{N-1}{N} S^2 - \frac{1}{N^2} Var(\hat{T}) \right\}.$$

Using this result we can show that

$$\begin{aligned} Bias(v^*) &= (N-n)NE(s^2)/n - Var(\hat{T}) \\ &= \frac{(N-n)(N-1)}{n-1} S^2 - \left(\frac{n}{n-1} \right) \left(\frac{N-1}{N} \right) Var(\hat{T}_c). \end{aligned}$$

Using the expressions for $E_\xi(S_{ij}^2)$, etc., that we developed earlier, the model-expectation of $bias(v^*)$ is

$$erb^* = \frac{E_\xi[bias(v^*)]}{E_\xi[Var(\hat{T}_c)]} = \left(\frac{n}{n-1} \right) \left(\frac{N-1}{N} \right) \left[\frac{(1-f_n)\kappa(\alpha, \beta) - \lambda(\alpha, \beta)}{1-f_n + \lambda(\alpha, \beta)} \right],$$

where

$$\kappa(\alpha, \beta) = \alpha(N - ML)/(N - 1) + \beta(N - M)/(N - 1).$$

When $\alpha = \beta = 0$ then $erb^* = 0$. Values of erb^* are presented in Table 3.2b for the

Table 3.2b. $erb^* = \frac{E_\xi[v^*(\hat{T})] - E_\xi[Var(\hat{T})]}{E_\xi[Var(\hat{T})]}$.

Sampling Scheme	Populations			
	A	B	C	D
(i)	0.00	0.02	0.07	0.13
(ii)	-0.09	-0.47	-0.74	-0.85
(iii)	-0.35	-0.89	-0.99	-1.00
(iv)	0.01	0.10	0.56	1.94
(v)	-0.11	-0.64	-0.94	-0.99

same populations and sampling designs as in Table 3.2a. As expected erb^* is small for

population A, in which the y values have relatively little within-cluster correlation. It is also generally small for one-per-cluster sampling schemes (i and iv) when correlations are not large. However, the results illustrate that $v^*(\hat{T}_c)$ can be a very poor estimator if the within-cluster correlation is large (Populations C and D).

The model-based comparisons of v_1 and v^* suggest that for the 3NO survey $v^*(\hat{T}_c)$ provides more reliable estimation of $Var(\hat{T}_c)$ than does $v_1(\hat{T}_c)$ because of the poor performance of the latter estimator when f_k is large. If the within-cluster correlation is not large, i.e. < 0.5 for secondary units, then our results suggest that $v^*(\hat{T}_c)$ may provide reasonable variance estimation; however, further investigation is required to better understand the bias in v^* for 3NO cod.

If auxiliary information about α and β is available then several procedures can be employed to improve variance estimation. For example, if the variance components model is reasonable for the population then σ^2 could be estimated from primary sampling units and then $Var(\hat{T}_c)$ could be estimated using a “plug-in” estimator. Procedures for constructing optimal variance estimators could also be employed (see Wolter, 1985; p 253-255). The spatial nature of the clusters in the 3NO survey suggest that a spatial model could provide information about within-cluster variability. This is because tow stations that are spatially close together do occur; for example, it is possible for two fishing units to be adjacent if they are from different primary units. The spatial structure of the survey data is investigated in **Chapter 4**; however, the spatial approach to variance estimation would require more detailed information about the spatial locations of primary and secondary sampling units, which is not readily available.

3.3.4. Confidence intervals

Some asymptotic results for stratified cluster sampling designs are presented in Krewski and Rao (1981). Bootstrap methods for constructing confidence intervals based on these

designs are considered by Rao and Wu (1988), Sitter (1992), and Rao *et. al.* (1992). Unfortunately all of the published methods are for designs where each cluster has more than one element sampled, and are therefore not directly applicable to the 3NO survey. It may also be possible to construct confidence bounds for one-per-cluster sampling designs using v^* if it generally holds that $v^*(\hat{T}_c) \geq \text{Var}(\hat{T}_c)$; however, we do not investigate this further.

A more detailed stochastic model for 3NO survey trawl catches is required so that confidence intervals for estimated population totals can be constructed. Such a model is developed in **Chapters 4 and 5**.

3.4. 3NO survey results

In this section we compute estimates of the annual total trawlable stock abundance, $\hat{T}$, and v^* for the 3NO survey data. The standard application of design-based inference requires that the stock is temporally stationary within the time frame of the survey. A member of the population is a trawl quadrat, or the area swept by the trawl. We consider the population to be a finite set of disjoint quadrats, and the y 's are the trawlable number of fish in the water column above a quadrat. The number of quadrats in a stratum is computed by dividing the total area of the stratum (A_h) by the nominal swept area of the trawl. The swept area used in standard DFO software is 0.04569 Km². This would seem to be based on a trawl width of 13.7 meters and a tow distance of 3.33 kilometers. The validity of these assumptions will be discussed later in this section.

The first order inclusion probabilities are based on the following quantities.

$$\begin{aligned} N_h &= \frac{A_h}{10 \times \text{Area per fishing unit}}, \\ N_{hi} &= 10 \text{ for all } h \text{ and } i, \\ \text{and } N_{hij} &= \frac{\text{Area per fishing unit}}{\text{Swept area of the trawl}}. \end{aligned}$$

Hence, $\pi_k = n_h 0.04569 / A_h$, $k \in \mathcal{U}_h$. The inclusion probabilities are identical within strata, but different between strata. The HT estimator of the population total is

$$\hat{T} = \sum_{h=1}^H N_h \bar{y}_h = \sum_{h=1}^H \hat{T}_h,$$

where $N_h = A_h / 0.04569$. We approximate $\text{Var}(\hat{T})$ using v^* .

For illustrative purposes we present estimated strata totals and standard errors in Tables 3.3a-c for 1992-94 in **Appendix B (Section 3.8)**. The notation $s(\hat{T}_h)$ denotes $v^{*1/2}$. Prior to 1991 the 700 series strata, or deep water strata, were not sampled. Note the large estimated abundances in stratum 332 and 337 in 1993. The combined totals for shallow and deep water strata are presented in Tables 3.4a,b. The estimated totals for the shallow water strata are plotted in Figure 3.1. Overall it is apparent that trawl catches have declined between 1984-86 and the 1990's. Estimated abundances are anomalous in 1987 and 1993. These high estimates are produced by large $\hat{T}_h$'s in a few strata. In 1993, 88% of the estimated abundance for the shallow water strata occurred in two strata (see Table 3.3b) which cover 6% of the 3NO region, while in 1987, 57% of the abundance occurred in three strata which cover 20% of the region. Trawl catches in these strata were also quite variable, as reflected in the large standard errors for $\hat{T}$ in for these years. These results will be discussed further in following chapters.

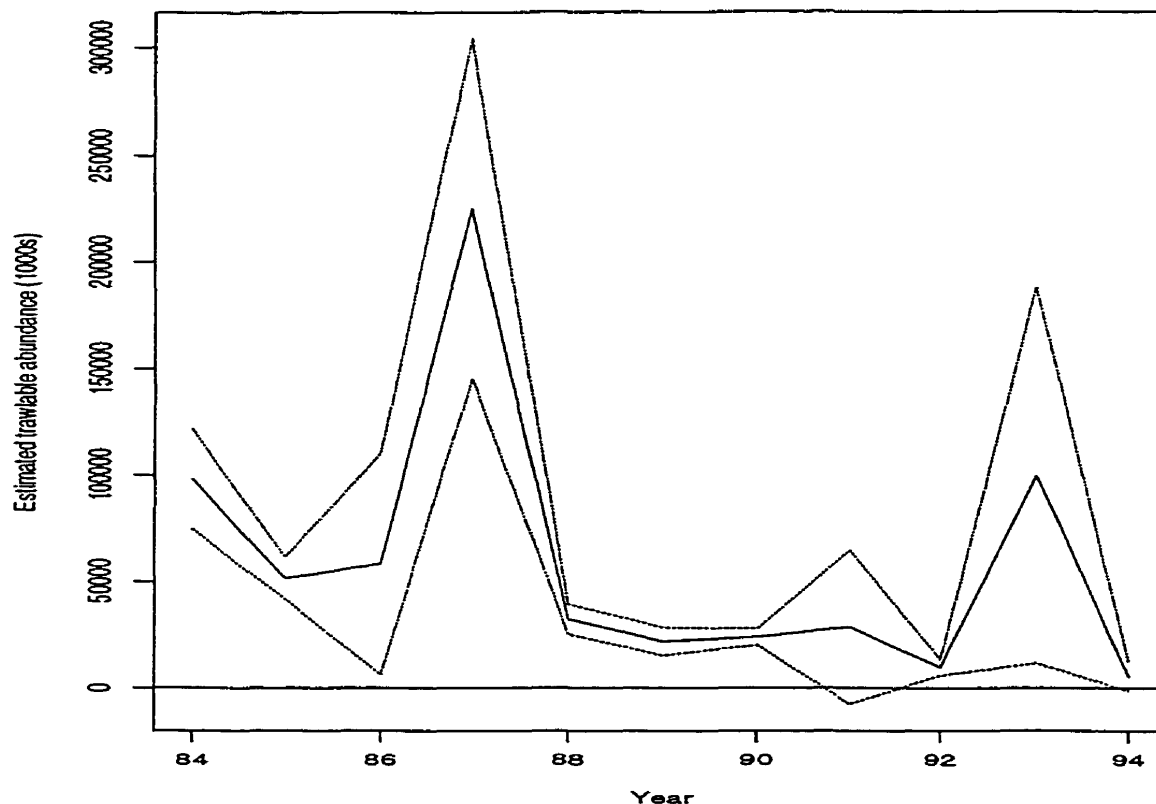


Figure 3.1: Annual estimates of trawlable abundance for the shallow water strata in NAFO Divisions 3N and 3O. Dotted lines connect ± 2 standard errors.

3.4.1. Variations from $S3C_1S$ in the 3NO survey design

There are problems in using design-based inferences with the 3NO survey, and these problems are in addition to the inferential difficulties with the $S3C_1S$ design. The problems essentially involve variations in the proportion of the 3NO division that is surveyed from year to year. In this section we discuss the major problems, including how they can introduce bias in $\hat{T}_c$.

Early in this chapter we mentioned a type of non-response problem that occurs when

the bottom type in a tow station is not conducive to sampling. The location and size of the untrawlable areas is not known in enough detail to be used in the survey design. Generally all that can be done is to substitute a nearby trawlable station for an untrawlable station. The decision to substitute has to be made at sea because that is when it is determined, using echo sounders, if a station can be towed. The presence of untrawlable stations increases the selection probabilities of stations near untrawlable areas, and can cause bias if the stock composition varies with bottom type.

A related problem is caused by the non-random selection of a tow station within fishing units. Trawling for fish is much more successful if the tow depth does not vary much; hence, tow stations are selected so that a constant depth is towed. Tow stations are also selected so that the tow direction is towards the next fishing unit to be sampled. This saves vessel time, and allows for more sampling. Nonetheless this does not constitute probability sampling from a finite population because a disjoint population of trawl stations is not used as a basis for sampling. In fact, samples are selected from an infinite population of partially overlapping trawl quadrats, where the probability that a point within a fishing unit is sampled by the trawl is not constant. This probability will be relatively low towards the periphery of the fishing unit, and is also related to bottom depth. Points that are close to the contour of constant depth passing through the center of the fishing unit will be sampled more often by the fishing gear than points that are far away from this contour. We do not explore this further.

Another assumption necessary for the direct application of design-based sampling theory is that the trawl catches all trawlable fish without error. This is tantamount to assuming that the availability of fish to the trawl and the catchability of the trawl are constant and equal everywhere. These assumptions are discussed in more detail in **Chapter 4**, but they do not seem valid.

Finally, our assumption about the temporal stationarity of fish within the time frame

of the survey is not valid. Cod are mobile, and migrate great distances. Little is known about the rates of daily 3NO cod movements, or the directions at the time of the survey, although the rates and directions of movement will vary from year to year, which further complicates the issue. If the majority of cod are in a directed migration at the time of the survey then it is possible that the survey will either miss or double count large amounts of fish, depending on the survey tract. We conduct analyses in **Chapters 4 and 5** that attempt to detect these temporal effects. If this is the case then the utility of the surveys for inferences about 3NO cod stock size will be greatly diminished, because large scale movements can introduce considerable bias in any year, and lead to a very noisy annual time series of trawlable stock abundance. At the least we can expect random local movements that will increase the variability of stock size estimators, but hopefully not introduce much bias in any year because local undercounting of fish is counteracted by local overcounting.

The importance of these deviations from the S3C₁S design will undoubtedly depend on the distribution of y 's in the population. It is quite likely that populations exist in which these deviations could have drastic effects. Populations also exist for which these deviations will be inconsequential. It seems likely that deviations in the sampling design will have more of an effect on inferences in populations with long-tailed distributions such as 3NO cod; however, these effects cannot be specified more precisely with the information that is available.

3.5. Model-based approach

In this section we discuss more details about model-assisted and model-based approaches to the analysis of the 3NO survey data. We have used a simple population model in **Section 3.3.3** for illustrative purposes; however, this was a measurement error model and may not be appropriate for trawl catches, which are counts. More analyses are necessary

before we propose a model, because of the complex stochastic nature of the survey data (heavy-tailed, spatially-temporally correlated, covariates associations). We will present these analyses in the next two chapters. In this section we merely wish to specify what role the sampling design will play in model-based inferential procedures developed in the remainder of this dissertation.

There are shortcomings in standard design-based inference methods for 3NO survey trawl catches, and the resolution of these shortcomings requires additional population assumptions. The assumptions essentially involve the spatial and temporal variability of the stock; that is, further knowledge is required about (i) the within-cluster spatial variability of survey catches, (ii) the population distribution in untrawlable areas, and (iii) the temporal stationarity of the stock within the survey period. Furthermore, to make inferences about total trawlable stock size from survey data, a relationship has to be assumed between trawl catches and local stock size. Realistic assumptions involve a stochastic relationship between trawl catches and fish abundance. Hence, we need to specify a stochastic model that relates trawl catches to local abundance, and which also relates local abundance throughout 3NO.

Statistical models are used in a variety of ways in the analysis of complex survey data, ranging from using models to assist with design-based inferences to using models solely for inferences. In the predictive inference approach a model is used to predict the y values for the $N - n$ elements in a population that are not observed. Smith (1990) has used this approach with trawl research survey data. However, the problem of estimating trawlable stock abundance is not one of predictive inference. Our problem is better described as follows.

Assume that the trawl catch at tow station i in year t can be adequately described by a random variable Y_{ti} . Let $\mathcal{N}_{ti}$ be a random variable for the number of fish in the water column above station i . The distribution of Y_{ti} involves $\mathcal{N}_{ti}$ and other parameters. Our

aim is to construct inferences about $\sum_i \mathcal{N}_{ti}$ based on a complex sample from $Y_{t1}, \dots, Y_{tn}$. If fish movements are important during the survey then for modelling we have to use a time scale finer than years. The model for the Y 's developed in **Chapters 5 and 6** is assumed for now to be sufficiently detailed (contain strata effects, temporal effects, ...) that the sampling design is non-informative (see Pfeffermann, 1993; Breckling *et. al.*, 1994; Godambe, 1996). Essentially this means that the design does not depend on the y values, nor does it contain important information about parameters in the distribution of the Y_{ti} 's.

Even with a stochastic model for the Y 's and a non-informative sampling design the role of the design in inferences is still controversial. Various requirements for design-unbiasedness or consistency have led to a variety of procedures for incorporating design-weights into parameter estimation and inferences (see Pfeffermann, 1993). The notion behind requiring estimators to be design-unbiased or consistent is that these requirements are model-independent; consequently, such estimators have model-robust inferential characteristics. This has been advocated for fish stock abundance estimators (Smith, 1990). However, it is generally recognized that design-unbiased estimators can have severe conditional biases, as Royall and Cumberland (1981) have shown for ratio estimators. Coupled with the problems of applying design-based methods to the 3NO survey, the requirement of design-unbiasedness does not seem useful in model-based analyses of the 3NO survey data.

3.6. Conclusions

In this chapter we considered a design-based analysis of the 3NO survey data. This is the standard approach used by DFO to analyze survey results, and we therefore choose this as our starting point in constructing inferences about trawlable stock size. We produced HT-estimates of trawlable stock size, and these showed that cod in Divisions 3NO have

generally declined in abundance since 1984, although anomalies in this trend occurred in 1987 and 1993. Estimated standard errors are quite high in 1993 and 1994, with coefficients of variation (CV's) around 50%, although in other years the CV's are smaller.

We identified several problems with the application of design-based methods to the 3NO survey. These involved the mobility of the population, subjectiveness in the sample design, and the inability to unbiasedly estimate the variance of HT-estimators of important population quantities, like trawlable stock abundance. We propose to use a stochastic superpopulation model for local stock abundance, as well as a model for the fish capture process, to construct inferences about 3NO trawlable stock abundance based on samples from the 3NO research survey. This model will be developed in **Chapters 4 and 5**, and inferences presented in **Chapter 6**.

3.7. Appendix A: Variance of $\hat{T}_h$, and the Horvitz-Thompson estimator of $Var(\hat{T}_h)$

In this section we develop the variance of $\hat{T}_h$ in terms of cluster variance components. We also develop an unbiased estimator of $Var(\hat{T}_h)$. We drop the h subscript for simplicity. We use conditional expectations to develop the variance of $\hat{T}$, and the estimator of $Var(\hat{T})$. Let E_I denote expectation with respect to primary unit samples, E_{II} denote expectation with respect to secondary unit samples given the primary units, and E_{III} denote expectation within secondary units. Conditional variances are defined similarly. First we find $Var(\hat{T}_h)$.

3.7.1. Variance of $\hat{T}_h$

To partition $Var(\hat{T})$ we use

$$Var(\hat{T}) = Var_I \left[E_{II} \left\{ E_{III}(\hat{T}) \right\} \right] + E_I \left[Var_{II} \left\{ E_{III}(\hat{T}) \right\} \right] + E_I \left[E_{II} \left\{ Var_{III}(\hat{T}) \right\} \right].$$

Recall that $\hat{T}$ can be written as

$$\hat{T} = \sum_{i=1}^n \sum_{j=1}^{n_j} \sum_{k=1}^{n_{ij}} \frac{y_k}{\pi_i \pi_{j|i} \pi_{k|ij}},$$

where $\pi_i = n/N$, $\pi_{j|i} = n_i/N_i$, and $\pi_{k|ij} = n_{ij}/N_{ij}$ are the primary, secondary, and within-cluster inclusion probabilities. It is straightforward to show that $E_{III}(\hat{T}_h) = \sum_{i=1}^n \pi_i^{-1} \sum_{j=1}^{n_j} \pi_{j|i}^{-1} T_{ij}$, where $T_{ij} = \sum_{k=1}^{N_{ij}} y_k$. Also, $E_{II} \left\{ E_{III}(\hat{T}_h) \right\} = \sum_{i=1}^n \pi_i^{-1} T_i$, where $T_i = \sum_{j=1}^{N_i} T_{ij}$. $Var_{III}(\hat{T}_h)$ requires π_{kl} , $k, l \in \mathcal{U}_{ij}$, which is $\pi_{kl} = \pi_i \pi_{j|i} \pi_{kl|ij}$. Let

$$\Delta_{kk'|ij} = \pi_{k|ij} \pi_{k'|ij} - \pi_{kk'|ij}.$$

Using (3.3.1) we can show that

$$Var_{III}(\hat{T}_h) = \frac{1}{2} \sum_{i=1}^n \pi_i^{-2} \sum_{j=1}^{n_j} \pi_{j|i}^{-2} \sum_{k=1}^{N_{ij}} \sum_{k'=1}^{N_{ij}} \Delta_{kk'|ij} \left(\frac{y_k}{\pi_{k|ij}} - \frac{y_{k'}}{\pi_{k'|ij}} \right)^2,$$

and

$$E_I \left[E_{II} \left\{ Var_{III}(\hat{T}_h) \right\} \right] = \frac{1}{2} \sum_{i=1}^N \pi_i^{-1} \sum_{j=1}^{N_i} \pi_{j|i}^{-1} \sum_{k=1}^{N_{ij}} \sum_{k'=1}^{N_{ij}} \Delta_{kk'|ij} \left(\frac{y_k}{\pi_{k|ij}} - \frac{y_{k'}}{\pi_{k'|ij}} \right)^2. \quad (3.7.1)$$

This is the within-cluster variance component. Let $\Delta_{jj'|\mathbf{i}} = \pi_{j|\mathbf{i}}\pi_{j'|\mathbf{i}} - \pi_{jj'|\mathbf{i}}$. In a similar manner we can show that

$$Var_{II} \left\{ E_{III}(\hat{T}_h) \right\} = \frac{1}{2} \sum_{\mathbf{i}=1}^n \pi_{\mathbf{i}}^{-2} \sum_{j=1}^{N_{\mathbf{i}}} \sum_{j'=1}^{N_{\mathbf{i}}} \Delta_{jj'|\mathbf{i}} \left(\frac{T_{ij}}{\pi_{j|\mathbf{i}}} - \frac{T_{ij'}}{\pi_{j'|\mathbf{i}}} \right)^2,$$

$$E_I \left[Var_{II} \left\{ E_{III}(\hat{T}_h) \right\} \right] = \frac{1}{2} \sum_{\mathbf{i}=1}^N \pi_{\mathbf{i}}^{-1} \sum_{j=1}^{N_{\mathbf{i}}} \sum_{j'=1}^{N_{\mathbf{i}}} \Delta_{jj'|\mathbf{i}} \left(\frac{T_{ij}}{\pi_{j|\mathbf{i}}} - \frac{T_{ij'}}{\pi_{j'|\mathbf{i}}} \right)^2, \text{ and} \quad (3.7.2)$$

$$Var_I \left[E_{II} \left\{ E_{III}(\hat{T}_h) \right\} \right] = \frac{1}{2} \sum_{\mathbf{i}=1}^N \sum_{\mathbf{i}'=1}^N \Delta_{\mathbf{ii}'} \left(\frac{T_{\mathbf{i}}}{\pi_{\mathbf{i}}} - \frac{T_{\mathbf{i}'}}{\pi_{\mathbf{i}'}} \right)^2. \quad (3.7.3)$$

The second-order inclusion probabilities are $\pi_{\mathbf{ii}'} = n(n-1)/N(N-1)$, $\pi_{jj'|\mathbf{i}} = n_{\mathbf{i}}(n_{\mathbf{i}}-1)/N_{\mathbf{i}}(N_{\mathbf{i}}-1)$, and $\pi_{kk'|\mathbf{ij}} = n_{\mathbf{ij}}(n_{\mathbf{ij}}-1)/N_{\mathbf{ij}}(N_{\mathbf{ij}}-1)$. These results can be used to simplify (3.7.3), (3.7.2), and (3.7.1) to give

$$\begin{aligned} Var_I \left[E_{II} \left\{ E_{III}(\hat{T}_h) \right\} \right] &= \binom{N}{n} \binom{N-n}{N-1} \sum_{\mathbf{i}=1}^N (T_{\mathbf{i}} - T/N)^2, \\ E_I \left[Var_{II} \left\{ E_{III}(\hat{T}_h) \right\} \right] &= \frac{N}{n} \sum_{\mathbf{i}=1}^N \binom{N_{\mathbf{i}}}{n_{\mathbf{i}}} \binom{N_{\mathbf{i}}-n_{\mathbf{i}}}{N_{\mathbf{i}}-1} \sum_{j=1}^{N_{\mathbf{i}}} (T_{ij} - T_{\mathbf{i}}/N_{\mathbf{i}})^2, \text{ and} \\ E_I \left[E_{II} \left\{ Var_{III}(\hat{T}_h) \right\} \right] &= \frac{N}{n} \sum_{\mathbf{i}=1}^N \frac{N_{\mathbf{i}}}{n_{\mathbf{i}}} \sum_{j=1}^{N_{\mathbf{i}}} \binom{N_{\mathbf{ij}}}{n_{\mathbf{ij}}} \binom{N_{\mathbf{ij}}-n_{\mathbf{ij}}}{N_{\mathbf{ij}}-1} \sum_{k=1}^{N_{\mathbf{ij}}} (y_k - \bar{Y}_{\mathbf{ij}})^2. \end{aligned}$$

If we define

$$\begin{aligned} S_T^2 &= \sum_{i=1}^{N_h} (T_i - T/N)^2 / (N - 1), \\ S_{T_i}^2 &= \sum_{j=1}^{N_{hi}} (T_{ij} - T_i/N_i)^2 / (N_i - 1), \text{ and} \\ S_{ij}^2 &= \sum_{k=1}^{N_{hij}} (y_k - \bar{Y}_{ij})^2 / (N_{ij} - 1), \end{aligned}$$

then we can express $Var(\hat{T})$ as

$$Var(\hat{T}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_T^2}{n} + \frac{N}{n} \sum_{i=1}^N N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_{T_i}^2}{n_i} + \frac{N}{n} \sum_{i=1}^N \frac{N_i}{n_i} \sum_{j=1}^{N_i} N_{ij}^2 \left(1 - \frac{n_{ij}}{N_{ij}}\right) \frac{S_{ij}^2}{n_{ij}}.$$

This is the result used in **Section 3.3.3**.

3.7.2. Horvitz-Thompson estimator of $Var(\hat{T}_h)$

We can also use conditional expectations to partition $Var(\hat{T})$ as

$$Var(\hat{T}) = Var_{III} \left\{ E_{II} \left[E_I(\hat{T}) \right] \right\} + E_{III} \left\{ Var_{II} \left[E_I(\hat{T}) \right] \right\} + E_{III} \left\{ E_{II} \left[Var_I(\hat{T}) \right] \right\}.$$

The following results are used in deriving $v(\hat{T}) = \widehat{Var}(\hat{T})$: $E_I(\hat{T}) = \sum_{i=1}^N \hat{T}_i$, $E_{II} \left[E_I(\hat{T}) \right] = \sum_{i=1}^N \sum_{j=1}^{N_i} \hat{T}_{ij}$,

$$Var_{III} \left\{ E_{II} \left[E_I(\hat{T}) \right] \right\} = \sum_{i=1}^N \sum_{j=1}^{N_i} (N_{ij} - n_{ij}) N_{ij} S_{ij}^2 / n_{ij}, \quad (3.7.4)$$

$$Var_{II} \left[E_I(\hat{T}) \right] = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^{N_i} \sum_{j'=1}^{N_i} \Delta_{jj'|i} \left[(\hat{T}_{ij}/\pi_{j|i}) - (\hat{T}_{ij'}/\pi_{j'|i}) \right]^2, \text{ and} \quad (3.7.5)$$

$$Var_I(\hat{T}) = \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N \Delta_{ii'} \left[(\hat{T}_i/\pi_i) - (\hat{T}_{i'}/\pi_{i'}) \right]^2. \quad (3.7.6)$$

We can write $Var(\hat{T})$ as (3.7.4) + $E_{III}(3.7.5)$ + $E_{III}\{E_{II}(3.7.6)\}$. The HT expanded estimator of this quantity is

$$\begin{aligned} v(\hat{T}) = & \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{(N_{ij} - n_{ij})N_{ij}}{\pi_i \pi_{j|i}} \frac{s_{ij}^2}{n_{ij}} + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} \left(\frac{\Delta_{jj'|i}}{\pi_i \pi_{jj'|i}} \right) \left(\frac{\hat{T}_{ij}}{\pi_{j|i}} - \frac{\hat{T}_{ij'}}{\pi_{j'|i}} \right)^2 \\ & + \frac{1}{2} \sum_{i=1}^n \sum_{i'=1}^n \left(\frac{\Delta_{ii'}}{\pi_{ii'}} \right) \left(\frac{\hat{T}_i}{\pi_i} - \frac{\hat{T}_{i'}}{\pi_{i'}} \right)^2. \end{aligned}$$

Note that

$$\begin{aligned} \sum_{i=1}^n \sum_{i'=1}^n \left(\frac{\Delta_{ii'}}{\pi_{ii'}} \right) \left(\frac{\hat{T}_i}{\pi_i} - \frac{\hat{T}_{i'}}{\pi_{i'}} \right)^2 &= 2(N - n) \frac{N}{n} \sum_{i=1}^n (\hat{T}_i - \hat{T}/N)^2 / (n - 1), \text{ and} \\ \sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} \left(\frac{\Delta_{jj'|i}}{\pi_i \pi_{jj'|i}} \right) \left(\frac{\hat{T}_{ij}}{\pi_{j|i}} - \frac{\hat{T}_{ij'}}{\pi_{j'|i}} \right)^2 &= 2(N_i - n_i) \frac{N_i}{n_i} \sum_{i=1}^{n_i} (\hat{T}_{ij} - \hat{T}_i/N_i)^2 / (n_i - 1). \end{aligned}$$

Using these results we can rewrite $v(\hat{T})$ as

$$v(\hat{T}) = N^2 \left(1 - \frac{n}{N} \right) \frac{s^2(\hat{T}_{hi})}{n} + \frac{N}{n} \sum_{i=1}^n N_i^2 \left(1 - \frac{n_i}{N_i} \right) \frac{s^2(\hat{T}_{ij})}{n_i} + \frac{N}{n} \sum_{i=1}^n \frac{N_i}{n_i} \sum_{j=1}^{n_i} N_{ij}^2 \left(1 - \frac{n_{ij}}{N_{ij}} \right) \frac{s_{ij}^2}{n_{ij}},$$

where $s_{ij}^2 = \sum_{k=1}^{n_{ij}} (y_k - \bar{y}_{ij})^2 / (n_{ij} - 1)$, $s_{T_i}^2 = \sum_{j=1}^{n_i} (\hat{T}_{ij} - \hat{T}_i/N_i)^2 / (n_i - 1)$, and $s_T^2 = \sum_{i=1}^n (\hat{T}_i - \hat{T}/N)^2 / (n - 1)$. This last expression is the result we used in **Section 3.3.3** for each stratum.

3.8. Appendix B: Tables

Table 3.1a. Strata information for NAFO Div. 3N. Numbers
in () are for 1994. Number of strata is 35.

Stratum	Depth Range (Meters)	Area (Km^2)	Number Fishing Units	Number Trawl Stations	Km^2 per Fishing Unit
357	275-366	561	40	12285	14.0
358	185-274	770	50	16854	15.4
359	93-183	1441	110	31536	13.1
360	57-91	10240	840 (810)	224122	12.2 (12.6)
361	57-91	6342	480	138803	13.2
362	57-91	8625	720	188766	12.0
373	57-91	8625	720	188766	12.0
374	57-91	3186	240	69738	13.3
375	<56	5452	420	119327	13.0
376	<56	5130	400	112286	12.8
377	93-183	342	30	7491	11.4
378	185-274	476	40	10412	11.9
379	275-366	363	30	7940	12.1
380	275-366	397	30	8689	13.2
381	185-274	623	50	13633	12.5
382	93-183	2214	180	48465	12.3
383	57-91	2307	190	50487	12.1

Table 3.1a. Continued.

Stratum	Depth Range (Meters)	Area (Km^2)	Number Fishing Units	Number Trawl Stations	Km^2 per Fishing Unit
723	367-549	530	50	11611	10.6
724	550-731	424	40	9288	10.6
725	367-549	359	30	7865	12.0
726	550-731	246	20	5393	12.3
727	367-549	548	50	11985	11.0
728	550-731	534	40	11685	13.3
752	732-914	459	40	10038	11.5
753	915-1097	472	40	10337	11.8
754	1098-1280	616	50	13483	12.3
755	1281-1463	1318	110	28839	12.0
756	732-914	363	30	7940	12.1
757	915-1097	349	30	7641	11.6
758	1098-1280	339	30	7416	11.3
759	1281-1463	435	40	9513	10.9
760	732-914	527	40	11536	13.2
761	915-1097	585	50	12809	11.7
762	1098-1280	726	60	15880	12.1
763	1281-1463	893	70	19551	12.8

Table 3.1b. Strata information for NAFO Div. 3O. Numbers in () are for 1994. Number of strata is 36.

Stratum	Depth Range (Meters)	Area (Km^2)	Number Fishing Units	Number Trawl Stations	Km^2 per Fishing Unit
329	93-183	5890	450	128915	13.1
330	57-91	7150	540	156481	13.2
331	57-91	1561	120	34158	13.0
332	93-183	3583	280	78428	12.8
333	185-274	517	40	11311	12.9
		(503)		(11011)	(12.6)
334	275-366	315	20	6891	15.7
		(329)	(30)	(7191)	(11.0)
335	275-366	199	20	4345	9.9
336	185-274	414	30	9064	13.8
337	93-183	3245	250	71012	13.0
338	57-91	6496	500	142173	13.0
339	93-183	2002	170	43821	11.8
340	57-91	5873	490	128540	12.0
			(500)		(11.7)
351	57-91	8625	720	188766	12.0
352	57-91	8830	720	193260	12.3
353	57-91	4388	340	96031	12.9
354	93-183	1622	130	35506	12.5
355	185-274	353	30	7715	11.8
356	275-366	209	20	4569	10.4

Table 3.1b. Continued.

Stratum	Depth Range (Meters)	Area (Km^2)	Number Fishing Units	Number Trawl Stations	Km^2 per Fishing Unit
717	367-549	318	30	6966	10.6
		(568)	(50)	(12435)	(11.4)
718	550-731	380	30	8315	12.7
		(459)	(40)	(10038)	(11.5)
719	367-549	260	20	5693	13.0
720	550-731	359	30	7865	12.0
721	367-549	260	20	5693	13.0
722	550-731	318	30	6966	10.6
764	732-914	359	30	7865	12.0
765	915-1097	424	30	9288	14.1
766	1098-1280	493	40	10787	12.3
767	1281-1463	541	40	11835	13.5
768	732-914	339	30	7416	11.3
769	915-1097	472	40	10337	11.8
770	1098-1280	438	30	9588	14.6
771	1281-1463	599	50	13109	12.0
772	732-914	462	40	10112	11.6
773	915-1097	438	30	9588	14.6
774	1098-1280	462	40	10112	11.6
775	1281-1463	530	40	11611	13.3

Table 3.3a. Estimated total strata abundance and standard errors for 1992. Values are in 1000s.

3N				3O			
h	n_h	$\hat{T}_h$	$s(\hat{T}_h)$	h	n_h	$\hat{T}_h$	$s(\hat{T}_h)$
357	2	301.0	92.1	329	8	48.3	23.6
358	2	1221.9	446.6	330	10	15.6	15.6
359	2	47.3	47.3	331	2	17.1	17.1
360	14	112.1	65.4	332	5	1301.9	828.0
361	8	34.7	22.7	333	2	130.1	5.7
362	11	34.3	23.0	334	2	172.3	117.1
373	10	0.0	0.0	335	2	178.1	143.4
374	5	13.9	13.9	336	2	761.4	163.1
375	6	79.6	39.8	337	4	1580.0	1229.0
376	7	0.0	0.0	338	6	853.0	543.2
377	2	0.0	0.0	339	2	43.8	0.0
378	2	171.8	67.7	340	5	25.7	25.7
379	2	1123.5	805.8	351	10	151.0	78.6
380	2	56.5	30.4	352	8	773.0	247.7
381	2	6.8	6.8	353	4	192.1	39.2
382	3	0.0	0.0	354	2	71.0	71.0
383	2	0.0	0.0	355	2	30.9	0.0
723	2	1625.5	1230.7	356	2	134.8	70.8
724	2	13.9	4.6	717	2	108.0	59.2
725	1	70.8	.	718	2	0.0	0.0
726	2	0.0	0.0	719	2	85.4	39.8
727	2	281.6	233.7	720	2	70.8	70.8
728	2	0.0	0.0	721	2	6159.8	6050.6
				722	2	31.3	31.3

Table 3.3b. Same as Table 3.3a, except for 1993.

3N				3O			
h	n_h	$\hat{T}_h$	$s(\hat{T}_h)$	h	n_h	$\hat{T}_h$	$s(\hat{T}_h)$
357	2	12.3	12.3	329	6	107.4	51.7
358	2	707.9	67.4	330	7	44.7	28.9
359	2	189.2	126.1	331	2	0.0	0.0
360	11	20.4	20.4	332	4	49801.8	22967.4
361	8	225.6	82.7	333	2	175.3	62.2
362	8	47.2	47.2	334	2	413.5	358.3
373	8	0.0	0.0	335	2	232.5	219.4
374	3	0.0	0.0	336	2	131.4	95.2
375	6	39.8	25.2	337	2	37494.3	37493.8
376	6	0.0	0.0	338	6	355.4	159.0
377	2	3.7	3.7	339	2	21.9	21.9
378	2	88.5	57.3	340	6	64.3	28.7
379	2	55.6	47.6	351	9	62.9	31.5
380	1	8150.3	.	352	7	441.7	205.4
381	1	0.0	.	353	4	144.0	114.3
383	1	0.0	.	354	2	266.3	159.8
723	2	11.6	0.0	355	2	65.6	57.9
724	1	9.3	.	356	2	130.2	52.5
725	1	0.0	.	717	2	539.9	205.5
726	2	0.0	0.0	718	2	41.6	16.6
727	2	131.8	71.9	719	2	102.5	96.8
728	2	0.0	0.0	720	2	3.9	3.9
				721	2	108.2	91.1
				722	2	0.0	0.0

Table 3.3c. Same as Table 3.3a, except for 1994.

3N				3O			
h	n_h	$\hat{T}_h$	$s(\hat{T}_h)$	h	n_h	$\hat{T}_h$	$s(\hat{T}_h)$
357	2	104.4	104.4	329	5	3377.6	3313.4
358	2	33.7	16.9	330	5	0.0	0.0
359	2	0.0	0.0	331	2	0.0	0.0
360	8	0.0	0.0	332	4	0.0	0.0
361	4	34.7	34.7	333	2	958.0	473.4
362	6	0.0	0.0	334	2	201.3	143.8
373	7	0.0	0.0	335	2	71.7	2.2
374	3	0.0	0.0	336	2	194.9	31.7
375	4	0.0	0.0	337	3	47.3	23.7
376	4	0.0	0.0	338	6	71.1	48.6
377	2	3.7	3.7	339	2	0.0	0.0
378	2	31.2	20.8	340	4	0.0	0.0
379	2	43.7	43.7	351	1	0.0	.
380	2	0.0	0.0	352	7	82.8	57.5
381	2	6.8	6.8	353	4	0.0	0.0
382	2	0.0	0.0	354	2	0.0	0.0
383	2	0.0	0.0	355	2	11.6	11.6
723	2	11.6	11.6	356	2	123.4	109.6
724	2	9.3	9.3	717	2	628.0	441.4
725	2	15.7	7.9	718	2	210.8	210.8
726	2	0.0	0.0	719	2	179.3	42.7
727	2	41.9	30.0	720	1	188.8	.
728	3	0.0	0.0	721	2	5.7	5.7
752	2	0.0	0.0	722	2	0.0	0.0
756	2	0.0	0.0	764	2	0.0	0.0
760	2	0.0	0.0	772	2	0.0	0.0

Table 3.4a. Estimated totals and
standard errors for
the 300 series strata
Numbers are in 1000s.

Year	n	$\hat{T}_h$	$s(\hat{T}_h)$
84	116	98012.1	11794.6
85	178	51384.7	4892.7
86	203	58082.9	25994.3
87	191	224518.3	39621.8
88	161	32093.8	3526.4
89	195	21807.2	3281.3
90	175	23842.4	1922.1
91	183	28535.4	18060.4
92	160	9683.6	1872.2
93	134	99493.7	43973.1
94	113	5397.9	3355.3

Note: Strata 335 not sampled
in 1990.

Strata 382 not sampled
in 1993.

Table 3.4b. Estimated totals
and standard errors
for strata 717-728.
Numbers are in
1000s.

Year	n	$\hat{T}_h$	$s(\hat{T}_h)$
91	24	2809.0	1531.6
92	23	8447.2	6179.8
93	22	948.7	255.6
94	24	1291.1	492.2

4. DATA OVERVIEW AND EXPLORATORY ANALYSES

4.1. Introduction

As discussed in **Chapter 1**, we are interested in using a mixture model to analyze 3NO research survey trawl catches. We hope to improve upon the design-based approach explored in **Chapter 3** by using a model-based approach. We will assume that trawl catches vary smoothly, as a function of some covariates. We hope that the model-based approach will utilize more information contained in the survey data, and therefore give more efficient inferences about trawlable stock size than the design-based method. Many aspects of the model have to be determined by the data. One important aspect is which covariates to include in the model. A model with too few covariates may result in inflated error variances and biased inferences. A model with too many covariates may result in inefficient inferences. For our model we consider seven covariates routinely measured in the 3NO research surveys that could be correlated with trawl catches (see below).

Variable selection is an important topic in many data analyses, and there has been a great deal of research published about this topic. Automatic methods are commonly used, but they can be confusing, and easily misused. These methods are often not that effective on their own when used with large complicated data sets. Before using such methods it is always a good idea to plot the data and do exploratory analyses. This is the purpose of this chapter. Using primarily graphical analyses, we will reveal important

features in the 3NO data set that help us build a more justifiable model. We also use these analyses to suggest a tentative set of covariates to include in our model. Some of these covariates are correlated with each other; hence, their relationship with catches may be an artifact of their relationship with other variables that are truly related to trawl catches. Including such covariates in a model can result in less reliable inferences, so it is still important to assess the tentative set of covariates using analytic variable selection procedures that adjust for between-covariate dependencies. This is the subject of Chapter 5.

We have already suggested that catches for tows that are geographically close together will tend to be more similar than for tows that are far apart; hence, the latitude and longitude of tows are two important covariates to consider. In Section 4.2 we present spatial plots of trawl catches of cod throughout 1984-94. These plots are useful for determining the importance of the spatial location of a trawl in predicting catches. We have also mentioned problems in estimating stock abundance because of the mobility of cod; that is, there is a potential for under counting, and double counting. In Section 4.3 we explore the within-year temporal aspect of the surveys, and in Section 4.5 we further explore spatial-temporal dependencies in catches using empirical autocorrelations. We also present a literature review in the Appendix (Section 4.9) that focuses on covariate relationships that have been published for similar stocks. Some of these are explored in Section 4.5, including whether trawl catches are influenced by (i) depth of tows, (ii) water temperature during tows, (iii) surface light conditions at the start of tows, and (iv) wind and (v) sea states during tows. Graphics are presented that focus on covariate relationships with the within-year distribution of catches, and also the between-year distribution of catches stratified by covariates levels. Finally, we present detailed graphical analyses of the ten tows that precede and follow a tow with a large catch, and we do this for the five largest catches throughout 1984-1994. The mixture model we propose

in Chapter 5 is primarily designed to accommodate very large catches, and detailed graphics are useful when assessing the adequacy of this model.

The main conclusions from our graphical analyses are that important spatial and depth dependencies in trawl catches exist. Such dependencies are common in groundfish stocks. These covariates are also useful for predicting trawlable stock abundance because we know their population values; hence, we will base our model for trawl catches on spatial location and tow depth. Associations between trawl catches and water temperature, time between tows, and wind force also appear to exist; however, these covariates are correlated with spatial location and tow depth, and further analyses are required to assess if they are important beyond their association with location and depth. Covariate associations do not have a simple form, suggesting that a nonparametric approach to functional modelling is required.

Note that all figures in this chapter are presented in the **Appendix (Section 4.8)**, so we will not refer to their placement further.

4.2. Spatial overview

Spatial plots of cod catches are presented in Figures 4.1a,b. NAFO division boundaries are indicated by heavy solid lines in these figures. These plots describe an important aspect of the survey data, which is the magnitude and geographic location of catches. We will use this type of graphic extensively in this thesis for a variety of purposes, e.g. residual plots, etc. Some important features in Figures 4.1a,b are:

1. The location of tows is somewhat spread out, which is caused by the three stage sampling design. There is also a trend in the 1990s towards surveying in the vicinity of the 300 meter depth contour, which is the result of adding deep-water strata to the survey in 1991.

2. Catches have declined in the 1990s compared to the 1980s. We have already seen evidence of this in **Chapter 3**. The proportion of trawls with a cod catch has also changed. For 1984-94 the percentages for the shallow water strata are

73, 75, 79, 88, 82, 79, 82, 63, 46, 46, and 25.

3. There is spatial persistence in catches. This can be utilized for predicting catches outside the sampled sites; that is, spatial location seems like a good predictive covariate.
4. The distribution of catches is often very long-tailed. We will treat very large catches separately from other catches using a mixture model.
5. Catches can abruptly change over short spatial distances around the 300 meter depth contour. This is because depth also changes rapidly in this region, and cod are often distributed along restricted depth gradients; hence, if a nearby tow is at a very different depth then it is likely that the catch will be quite different as well. This suggests that our model for trawl catches should include spatial location and tow depth as predictive covariates.

4.3. Spatial-temporal overview

Cod are a highly migratory species. They can migrate hundreds of kilometers from offshore spawning regions to inshore feeding regions. This has long been recognized to cause problems in surveying cod, and many other stocks. Annual variations in the timing of surveys or the timing of migrations can result in changes in the proportion of the stock contained within the survey area, i.e. changes in the proportion of the stock surveyed. At best the survey can only provide information about fish abundance within the survey

region at the time of the survey. Surveys cannot provide abundance information about stock components outside the survey region. Even for the simpler problem of counting fish within a region, the mobility of fish can cause problems. It takes several weeks to survey the 3NO region, and during this time schools of fish move. This movement creates the possibility of counting schools more than once, or missing them. This is a problem when surveying mobile populations. We shall explore the potential for this problem using several types of analyses of the within-year timing of the surveys.

Time series plots of catches are presented in Figures 4.2a,b. These plots indicate the start time and end time of the annual surveys, as well as the catch value for each tow. The square root scale is chosen so that variability in the smaller catches is discernible. Temporal autocorrelation is apparent, although this is confounded with spatial autocorrelation, as discussed below. The survey duration ranged from 10 days in 1984 to 39 days in 1990, although in 1990 most of the survey stations were sampled within 23 days. In the 1990s there has been a trend to survey later in the year. Annual variations in survey timing may not have as much impact on 3NO survey catches as in other regions because tagging information suggests that fish in this region do not frequently move into other regions. Tagging data do indicate some seasonal movements of fish from the west in 3PS and north from 3L into 3NO; hence, some catches at the boundaries of 3NO may be from other stocks, and this will be influenced by the seasonal timing of the surveys. Examining the length distribution of such catches may assist in determining the stock affinity of the fish, but this is not pursued in this dissertation.

Space-time trajectories of the 3NO spring surveys are depicted in Figures 4.3a,b. The usual trajectory of a survey begins in 3O and ends in 3N. Tows located close together in space can either be close together in time (generally) or not (rarely), but tows far apart in space are never close together in time because of the relatively slow speed of the survey vessel. This confounds spatial and temporal dependencies in trawl catches, and

makes it difficult to determine if the distance between tows is a better predictor of trawl catches than the time between tows. The distinction is quite important however, because temporal dependencies raise the possibility of double counting or under counting, and lead to great uncertainty in inferences.

Figure 4.3a and Figure 4.1a provide an illustration of the importance of within-survey timing. In 1986 an anomalously large catch of cod occurred in one tow in 3O. Another tow close by did not catch much cod. This could indicate large spatial variability in abundance; however, examining Figure 4.3a suggests another explanation. The tow with a large catch of cod was taken on the 3rd of May, while the nearby tow with little cod catch was taken on the 23rd of April. These tows were taken 10 days apart, and it is easy to imagine that the large concentration of cod was not as near the much smaller concentration on the 23rd of April. The next year provides another illustration. The large catches of cod in the Northeastern part of 3O are similar to large catches in the Northwestern part of 3N, but the tows in these Divisions were taken 13 days apart. This might indicate a large stationary concentration of cod that covered both corners of 3N and 3O. It might also indicate a smaller concentration of cod that was in the Northeastern corner of 3O around April 24, but moved to the Northwestern corner of 3N by May 7, and was counted again in the survey. This would constitute double counting. The problem here is that in examining such instances usually two or more interpretations exist which suggest different courses of action, ranging from doing nothing to deleting some tows. Information on the average rates and directions of fish movements would greatly assist in interpreting these data; however, this type of scientific information is lacking for cod in 3NO. The best we can do is to quantify the spatial and temporal dependencies in trawl catches, which is done in the next section. This will merely give an indication of the potential of the problem; however, we will not be able to address the problem of survey timing in this dissertation. Nonetheless, we should be aware of it.

4.4. Spatial-temporal dependencies

In the previous section we alluded to confounding between the distance and time between tows; hence, there is also potential confounding between spatial and temporal dependencies in survey catches. In this section we attempt to quantify whether dependencies in trawl catches are more related to distance or time between tows. Temporal dependencies are of particular concern in trend estimation for the reasons discussed in the previous section.

We use empirical autocorrelations to assess the spatial-temporal dependencies in the data. Spatial-temporal autocorrelations are compared to purely spatial autocorrelations to assess if the time between tows is important beyond the spatial distance in determining differences in catches. These autocorrelations are merely sample correlation coefficients for pairs of observations that are separated by fixed spatial and temporal distances. If significant positive temporal autocorrelation exists then we would expect the correlation between catches for all tows separated by spatial distance g to be less than the correlation for all tows separated by spatial distance g and a short interval of time. We have seen evidence of this already in 1986 (see 3O catches in Figure 4.1a), where large variations in catches in a spatially small region could, in part, be explained by time between tows.

We use the formulae in Cliff and Ord (1981) to compute autocorrelations. Let y_{it} denote the number of cod caught in a tow in year t at spatial location $\mathbf{x}_{it}$ (latitude and longitude) and temporal location z_{it} (hours or minutes from start of year), $i = 1, \dots, n_t$. An estimate of spatial autocorrelation at distance g is

$$\rho_t(g) = \frac{\sum_{i,j \in A_t(g)} (y_{it} - \bar{y}_t)(y_{jt} - \bar{y}_t) / |A_t(g)|}{\sum_{i=1}^{n_t} (y_{it} - \bar{y}_t)^2 / n_t}, \quad (4.4.1)$$

where $A_t(g) = \{(i, j) : \|\mathbf{x}_{it} - \mathbf{x}_{jt}\| = g, i, j = 1, \dots, n_t\}$, and $\|\mathbf{u}_{p \times 1}\| = \sqrt{\sum_{i=1}^p u_i^2}$. The number of distinct pairs in $A_t(g)$ is denoted by $|A_t(g)|$. Typically there are few

pairs of observations exactly at distance g , so spatial autocorrelations are defined by pairs of observations that are separated by a distance that falls within an interval centered about g . Equation (4.4.1) is based on the assumption that spatial dependencies are of an isotropic nature; that is, the autocorrelation depends only on the distance between tows and not direction.

We estimate the spatial-temporal autocorrelation in catches separated by distance interval g and temporal interval h by

$$\rho_t(g, h) = \frac{\sum_{i,j \in B_t(g,h)} (y_{it} - \bar{y}_t)(y_{jt} - \bar{y}_t) / |B_t(g, h)|}{\sum_{i=1}^{n_t} (y_{it} - \bar{y}_t)^2 / n_t}, \quad (4.4.2)$$

where $B_t(g, h) = \{(i, j) : \| \mathbf{x}_{it} - \mathbf{x}_{jt} \| \in g \text{ and } |z_{it} - z_{jt}| \in h, i, j = 1, \dots, n_t\}$.

Cliff and Ord (1981) discuss (4.4.1) and (4.4.2). The same definition of (4.4.1) is given in Cressie (1993, eq. 2.3.21), where it is referred to as a correlogram. We only compute autocorrelations if $|A_t(g)|$ or $|B_t(g, h)| > 10$ pairs of observations.

4.4.1. Spatial autocorrelation results

We compute (4.4.1) for six distance classes. The results are presented in Figure 4.4a. Positive autocorrelations are usually estimated for catches separated by distances less than 100 kilometers. For the distance class “< 20 kilometers” the estimates are as high as 0.55 in 1987. In all years the autocorrelations were highest in this distance class (84, 85, 87, 91, 93, and 94) or the “20 – 50 kilometers” distance class (86, 88-90, 92). This indicates decreasing autocorrelations with distance. This is entirely reasonable because the factors that influence cod aggregations (temperature, salinity, prey aggregations, etc.) usually have positive spatial autocorrelations that decrease with distance. Note that the estimates are very low in 1986, 1989, 1991, and 1992. It is possible that the correlations in 1986, 1991, and 1992 are underestimated due to the presence of large anomalous catches (see Figures 4.1 a,b). We consider this in the next section.

Spatial autocorrelations in catches are small for tows separated by distances greater than 100 kilometers, with the exception of 1984 and 1994. In 1984 the estimated autocorrelation for distances between 200 and 400 kilometers is -0.16 . The estimated autocorrelation for distances greater than 400 kilometers is -0.15 in 1994. The negative sign indicates a potential spatial trend in abundance.

4.4.2. Spatial-temporal autocorrelation results

Spatial autocorrelations for several time classes are also presented in Figure 4.4a. These estimates are joined by lines, and should be compared with the purely spatial autocorrelations, which are plotted as points. Our interest is to determine if the timing of tows is important in addition to the location of tows. The temporal breakdown of spatial autocorrelations does not result in any consistent patterns. The spatial autocorrelations for tows less than 100 kilometers apart and taken within 12 hours of each other agree closely with the purely spatial autocorrelations (compare *'s and solid lines in Figure 4.4a). On the other hand, there is a close correspondence between the estimated spatial autocorrelation for tows separated by distances of at least 100 kilometers and times of 5 days or more, and the purely spatial autocorrelations.

To assess the differences in estimates of spatial and spatial-temporal autocorrelations we plot the estimates in the left-hand panels of Figure 4.4c, and indicate the number of times that spatial-temporal autocorrelations exceed corresponding spatial autocorrelations. Considerable discrepancy exists only when an estimate is based on a small number of observation pairs. For example, in 1990 the estimated spatial autocorrelation for tows between 20 and 50 kilometers apart is 0.32, while it is 0.92 for the same distance class and tows separated by 5 days or more. The increase is counter-intuitive; however, the latter estimate is based only on 52 observations pairs and has more estimation error associated with it than the former estimate, which is based on 502 observations pairs.

Warren (1997) considered a similar problem and found that the temporal components of space-time variograms for cod in NAFO divisions 2J, 3K, and 3L (see Figure 1.2) during 1985-1992 were at least as “significant” as spatial components; however, Warren (1997) analyzed log transformations of nonzero catches so his results are not directly comparable with those we have presented. Warren (1997) investigated a 25×30 array of space-time intervals. Sample sizes in 3NO prevent this scale of analysis; however, we repeated our analysis for the same distance classes and six time intervals (< 0.2 , $0.2 - 0.5$, $0.5 - 1$, $1 - 2$, $2 - 5$, > 5) and found essentially the same results, except with more variability in the space-time estimated autocorrelations, as expected because of the smaller bin sample sizes.

We also analyzed log transformed positive catches to see if spatial-temporal dependencies are more discernible in this scale. To save space we do not present the results; however, generally the estimated autocorrelations are higher after transforming the data. The correlations increase the most for years with a few large anomalous catches. Similar to the un-logged analysis, the spatial autocorrelations decrease with increasing distance between tows. Spatial-temporal correlations for pairs of tows taken within 5 days of each other increased for 65% of comparisons with corresponding spatial correlations, and this pattern is reversed for tows taken more than 5 days apart. The expected increase if no temporal autocorrelation is present is 50%. The difference in estimates are usually quite small and this analysis does not suggest that the timing of tows is of great importance beyond its association with tow location; however, the analysis also does not suggest that timing is not important either.

In the preceding analysis the relevance of the time between tows was assessed by comparing spatial autocorrelations with spatial-temporal autocorrelations. This relevance can also be assessed by comparing temporal autocorrelations with spatial-temporal autocorrelations. Some results for this purpose are presented in Figure 4.4b. Note that different

space and time bins are used in constructing these figures than in Figure 4.4a. The number of time bins was increased to better capture the nature of the temporal autocorrelations, while at the same time the number of space bins was decreased so that $|B_t(g, h)|$ was usually greater than ten. The estimated temporal autocorrelations, when large, increase for tows that are less than 50 kilometers apart, and decrease for tows that are greater than 200 kilometers apart. This is also evident in the right hand panels in Figure 4.4c. The temporal autocorrelations increased for 75% of the tows that are close together in space, but for only 29% of the tows that are far apart. Clearly the location of tows is very important in terms of temporal dependencies, and that correlation in trawl catches depends on more than the time between tows.

Occasionally pairs of tows occur that are close together in space but not in time. If temporal autocorrelation were a decreasing function of time then one would expect the autocorrelation for these tows to be less than the autocorrelation for tows that are close together in space and time. The relevant autocorrelations are presented in Table 4.1. Autocorrelations decreased in 7 of 10 years, and in 4 of 5 years when the autocorrelations are greater than 0.1; that is, 84, 85, 87, and 93. In some years the autocorrelations between tows separated by five days or more are based on only a small number of pairs, and as a result are poorly estimated. If we restrict our attention to the “data rich” years (85-87, 90-92) in which 30 or more pairs of tows exist that are close together in space but not time, then autocorrelations decrease in only 50% of the years. This is expected if no temporal autocorrelation is present. These results indicate again that some temporal autocorrelation in trawl catches may exist, although the evidence for this is far from convincing. The evidence of spatial autocorrelation is convincing.

The conclusion we reach from the results in Figures 4.4a-c and Table 4.1 is that the correlation in the survey data is primarily of a spatial nature, and that the time between tows is of secondary importance - but not to be discounted completely. A problem with

Table 4.1. Estimated autocorrelations in the 20-50 Km distance class. Time denotes the period between tows, in days.

Year	Autocorrelations		No. Pairs	
	Time < 0.5	Time > 5	$ A_t(h) $	$ B_t(h, g) $
84	0.287	0.068	175	12
85	0.115	0.092	335	34
86	0.016	0.098	397	46
87	0.150	0.033	419	72
88	0.198	-	312	7
89	0.068	0.011	394	16
90	0.266	0.916	302	52
91	0.018	0.017	415	40
92	0.002	0.633	330	35
93	0.151	-0.098	276	24
94	0.084	0.077	229	13

the analyses in this section is that we have not accounted for other covariates that could produce spurious spatial or temporal estimates of autocorrelation. It is difficult to account for such confounding among covariates, and this will not be dealt with until **Chapter 5**.

4.5. Covariate associations

Possible associations between trawl catches and covariates such as depth and temperature have been alluded to in previous sections. Exploratory analyses of potential relationships between covariates and 3NO research survey catches of cod are presented in this section. These analyses are primarily graphical. The covariates we consider are depth, bottom temperature, surface light conditions, sea state, and wind state. These covariates have been measured in the 3NO research survey continuously throughout 1984-94, and could affect trawl catches. The aim of this section is to decide if any of these covariates should

be included in our model for trawl catches and, if so, what is the functional form of the relationship.

We are not the first researchers to consider the importance of these covariates on trawl catches; therefore, it is useful to investigate how these covariates affect catches in other cod stocks, as suggested in the published literature. This may indicate ways to refine our analyses so that they are more relevant to our problem, and also assist in interpreting results. We restrict our attention to possible covariate associations with the distribution of cod abundance, and also covariate associations with trawl catchability. We will not cover the substantial amount of literature related to the influence of covariates on annual changes in stock abundance. For now our interest is merely in estimating the change in stock size and not the causes for change. The literature review is not essential to understanding the results in this section, and is therefore presented in an **Appendix (Section 4.9)**.

4.5.1. Depth analyses

In this section we investigate if (i) catches change systematically with tow depth, or (ii) different trends in annual average catches occur at different tow depths. The aim of (i) is to assess if depth has a consistent effect, or additive effect, on all catches. If so we will incorporate this into our model. The aim of (ii) is to assess the variability in trawl catches at different depth intervals, and also to assess the between-year temporal persistence in trawl catches at different depth intervals.

The square root of the numbers caught per tow are plotted versus tow depth in Figure 4.5a. The solid line in this figure represents a nonparametric regression of catch versus depth, and is included to assist in visualizing patterns in the data. The proportion of variance that the regression explains is indicated in the top right-hand corner of each panel. We use the “supersmooth” function in Splus with a cross-validation optimal bandwidth

for the nonparametric regression. The square root transformation is used for display purposes only, and not for the exploratory regression.

The change in the survey design with more tows at greater depths in the 1990s is apparent in this figure. Also apparent is a trend similar to the one indicated in Smith *et. al.* (1991; see the **Appendix, Section 4.9**) at depths less than 100 meters; that is, generally catches first increase with depth, and then decrease. However, the trend never accounts for a substantial amount of variance. In 1991, 1992, and 1994 the trend is not apparent. No trends are apparent at depths greater than 200 meters. Large catches of cod occur over a wide range of depths. Regressions for small cod, weighing less than 1.5 kilograms on average, and large cod followed basically the same patterns.

Differential trends in annual catches across depths are apparent in Figure 4.5b. The 25th, 50th, and 75th percentiles of catches have declined in most years at depths between 50 and 100 meters - where most of the tows occur. This indicates declining stock sizes in most areas sampled. Note that the mean catch per tow often exceeds the 75th catch percentiles, and this indicates the heavy-tailed nature of the data. Catches at depths greater than 200 meters have generally increased between 1987 and 1993. This is particularly evident at depths between 300 and 400 meters. This suggests that in the 1990s a greater proportion of the stock abundance occurred at deeper water depths than in the 1980s.

The results in Figures 4.5a,b suggest a decline in catches of cod in most areas sampled, as we have seen in previous analyses. The new result obtained from this analysis is evidence of a shift in 3NO cod distribution towards greater relative abundance at deeper depths between 1987-1993; that is, average catches have increased at deeper depths even though total average catches have declined. We have assumed that no systematic increase in research trawl catchability at deeper depths has occurred. This could also account for an increase in average catches at deeper depths, but not indicate a change in relative stock abundance. Whether this has occurred and could account for the trends in Figure 4.5b

requires further experimentation and investigation, but is beyond the scope of this thesis. We can do little but assume that the trends actually reflect trends in stock abundance. The results in this section suggest that tow depth is an important determinant of catches, and should be included in our model.

4.5.2. Temperature analyses

In this section we investigate the effect of water temperature on catches of cod. We focus on the same patterns as i-ii in the previous section.

Catches are plotted versus temperature in Figure 4.6a, along with a nonparametric regression of the catch data. Cod have rarely been caught at temperatures below -1°C , especially in the 1990s. During 1984-1990 catches peaked for tow temperatures between 0°C and 4°C . After 1991 greater catches of cod usually occurred in warmer water (2°C - 6°C). Large catches of cod have occurred throughout most of the sampled temperatures. Regressions for small cod, weighing less than 1.5 kilograms on average, and large cod followed basically the same patterns.

Time series of average catches in eight temperature intervals are presented in Figure 6b. The square root is used for plotting purposes. The 25th, 50th, and 75th catch percentiles have declined since 1987 in water with temperatures less than 4°C , and this indicates a stock decline in most areas with relatively cool bottom water. Catches in water with temperatures greater than 5°C generally increased between 1984 and 1991, declined in 1992, and then stabilized after 1992.

The results in Figures 4.6a,b suggest a shift in 3NO cod distribution throughout 1984-94 towards greater relative abundance at warmer water temperatures; however, this distributional shift may be confounded with a shift in depth distribution. This is because bottom water temperature is not independent of depth. Warm water from the Gulf Stream usually exists at deeper depths along the edge of the Grand Bank; hence, temperature

and depth associations are confounded. We need to assess the importance of temperature in determining trawl catches beyond the association of catches with depth. It may be that cod select for temperature rather than depth, and the depth association in the previous section is just an artifact caused by temperature selection; however, tow depth can be used for predictive purposes, whereas temperature can not. This is because we only know temperatures at the sample sites, and we can at best only use the component of a temperature association that is related to depth for inferences about stock size. This is why we are interested in temperature associations in catches after depth effects have been removed. This is considered in Chapter 5.

4.5.3. Surface light analyses

We investigated whether trawl catches are related to recorded surface light conditions in terms of i-ii above. To save space the results are not presented; however, we found no evidence of a general association. These analyses suggest that surface light is not a useful covariate to consider in a model for trawl catches of cod.

4.5.4. Wind strength analyses

In this section we investigate relationships between wind strength and trawl catches. Wind strength influences the ability to successfully tow a trawl on the ocean bottom, and therefore wind strength may possibly influence catches. Most likely wind strength affects trawl catchability rather than local abundance. If wind strength does influence catchability then this suggests that some tows, presumably those taken during high winds, are less informative about stock size than others. These tows should receive less weight in prediction of local stock density. Wind strength is measured on an ordinal scale with an anemometer. Wind scale codes range from 0-9, where 0 indicates no wind, and 9 indicates very high winds.

The distribution of catches for each level of the wind scale code is presented in Figure 4.7a. Average catches appear to be greater when wind scales are 0-3 than when wind scales are greater than 3 in 1984-1987; however the pattern is reversed in 1991-1994, where greater catches usually occur when wind scales are greater than 3. The medians also follow the same pattern, although with much less contrast.

Yearly ratios of catches when winds are low (wind scale ≤ 3) or high (wind scales > 3) are presented in Figure 4.7b. The numerator in the ratio is the average catch during low winds. It also seems likely that depth interactions exist so the ratios are computed for shallow (depth ≤ 100 meters) and deep (depth > 100 meters) tows. Throughout 1984-1994 trawl catches at tow depths ≤ 100 meters were approximately equal when winds were low or high; however, a consistent and strong trend exists in which relative catches have increased over time when winds are high. At depths greater than 100 meters considerably fewer fish have been caught throughout 1984-1994 when winds are light.

The main conclusion we draw from the analyses in this section is that there is evidence that the catchability of the trawl or the availability of cod to the trawl has changed over time, as a function of wind strength. Whatever the cause, wind strength appears to account for some of the variation in trawl catches. It could also be that this is an artifact, caused by the association of trawl catches with some other factor. Wind strength is also correlated with time between tows and spatial location because of the temporal persistence of winds. This could create a spurious relationship between wind strength and catches. Nonetheless it seems like a good idea to consider wind strength in our model. In doing so we will be better able to quantify the importance of the association, especially in conjunction with other covariates. If this relationship is significant then this suggests that some trawl catches are influenced by factors other than local fish density, and therefore their utility in terms of estimating stock abundance is reduced.

4.5.5. Sea state analyses

We also investigated the effect of sea state on trawl catches. These results are partly confounded with those for wind strength. This is because wind and sea states are confounded, although not completely. Sea states can remain rough long after winds decrease, and a change in wind direction can cause rough seas to calm. There is likely more measurement error in sea states than wind scales. Wind strength is measured with an anemometer, whereas sea states are determined visually which is more difficult at night.

The results of the sea state investigations are similar in nature to those in the previous section, but less consistent. Both wind strength and sea state are really just two measurements of the same “force” that affects trawl performance. It may be that wind strength affects trawl catches only through its influence on sea state, but because of the problems in measuring sea state we feel that wind strength is the better covariate to quantify the relationship with. For this reason sea state is not considered in our model.

4.6. Large catches

The purpose of this section is to investigate local subsets of the surveys that contain tows with large catches of cod. By local we mean the ten tows that precede and follow, in time, a tow with a large catch. A relatively small number of tows with large catches have great influence in our analyses. Such tows also have great influence in traditional stock assessments because of their influence on survey summary statistics which are an important component of the information basis used in assessments. Hence, we wish to understand in detail what factors influence the distribution of high density stock components. This information could be very useful in stock trend estimation if it assists in determining the spatial extent of these high density components.

To investigate large catches we construct multidimensional graphics for the five tows

with the largest numbers of cod caught throughout 1984–1994. We plot catches in terms of the spatial, depth, temperature, and temporal distances between tows with large catches and other local tows. The results are presented in Figure 4.8.

No consistent patterns are apparent in this figure. Four of the large tows occurred in areas with locally large temperature and depth gradients (refer to rows 5–6), but large tows in 1993 occurred in areas with locally small gradients (refer to column 4). The large tows occurred in water with temperatures that were locally warm or cold. The bottom temperatures during the large tows did not deviate systematically from average conditions within years (unreported results). The results in the first column demonstrate how much catches can change over short spatial and temporal distances within locally similar (depth and temperature) conditions. The results suggest that the factor(s) that cause large catches of cod are not highly correlated, at least locally, with the covariates we have considered.

4.7. Summary

The results in this chapter suggest that in 3NO research surveys during 1984–1994, cod catches are spatially correlated, and the correlation decreases with distance. Cod have been caught at a wide range of depths, and the distribution of catches has shifted over time with somewhat greater catches recently at deeper depths. Cod are not caught often in water with temperatures less than 0°C . The distribution of catches has also shifted over time with greater catches recently in relatively warm water; however, this is confounded with depth. In recent years relatively greater catches of cod have occurred during tows when wind strengths are relatively high for that year, although this trend may also be an artifact caused by the association of catches with some other covariate that is confounded with wind strength. We will investigate this further in **Chapter 5**, where we use high dimensional analytic variable selection methods to assist in deciding the relevant subset

of covariates to include in our model. The analyses in this chapter have also shown that catches are heavy-tailed, and large catches of cod occur apparently at random.

Throughout this chapter we have discussed the implications of covariate associations on stock size inferences. Dealing with these associations range from easy to practically impossible with the information that is available to us. The association of trawl catches with depth is easy to deal with. For accurate estimation of total trawlable abundance when catches vary with depth we must ensure that the estimation weight of catches reflects the proportion of the 3NO region with depths similar to the depths at the sampled tow sites. This reduces spurious signals in our estimates that are related to changes in the survey design and not changes in abundance. In our model we will use depth to predict local stock density at a large number of random tow stations (see Chapter 6) based on catches observed at the sampled tow stations. The prediction at one of these tow stations, say i , will be a weighted average of sampled catches, where the weights depend, in part, on differences between the water depth at tow station i and the depths at the sampled tow stations. The weights will be larger for tow stations with similar depths to station i . Because we use the sampled catches to predict local abundance at a large random sample of tow sites this weighting ensures that estimation weight of catches reflects the area of 3NO with depths similar to the depths at the sampled tow sites.

Temperature and wind associations are more difficult to deal with because we do not know the values of these covariates at non-sampled tow stations; however, we can attempt to predict these values, then follow the same procedure when dealing with depth associations with trawl catches.

We conclude that without more scientific information about the movement of cod in 3NO during April and May then we can do little to adjust for within-survey temporal autocorrelation. We have no reason to expect anything but random fish movements, which suggests we may under-count some schools of fish about as frequently as we over-count

others. This leads to greater variability in estimates of stock abundance, but not bias. In the next chapter we will further quantify the magnitude of the temporal dependencies in trawl catches, which at least gives an indication of the potential for under- and over-counting.

The results in this chapter suggest that the associations between trawl catches and the covariates we examined do not have a simple parametric form. The results suggest a nonparametric approach to regression modelling is appropriate. This will be the approach taken in **Chapters 5 and 6** in both the variable selection context and for stock size inferences. Our results suggest that apparently random large catches occasionally occur. It seems reasonable that the small, high-density aggregations of cod that give rise to these large catches exist even in years without large catches. This is because the probability of detecting these aggregations is small using stratified random sampling. We use a simple mixture model for this type of catch. The mixture model will divide catches into two types. The majority of catches will be of a type I. These are consistent with surrounding catches. By consistent we mean similar to other catches taken at approximately the same values of relevant covariates. Type II catches are not consistent with surrounding catches. For example, the large catch in 1986 in 3O seems to be of type II; however, the large catches in 1987 at the 3L Division boundary are not considered to be type II because they are consistent with surrounding catches. These are type I catches.

The mixture model will be specified exactly in **Chapter 5**, where nonparametric analytical methods will be used to measure the predictive capability of spatial location, tow depth, time between tows, temperature, and wind strength. These methods will accommodate confounding among these covariates. We will use these methods to determine which covariates to include in the mixture model. Robust methods are also used to determine which tows are type II catches. The potential advantage of this model is more precise inferences about trawlable stock abundance, and a reduction in the year to year

variability of abundance estimates, as will be shown in Chapter 6.

4.8. Appendix: Literature review of the effects of depth and hydrographic conditions on bottom trawl catches

This review covers stocks for which the processes causing covariate dependencies may be similar to those occurring in the 3NO cod stock. It is not exhaustive, but is presented merely to illustrate the types of dependencies that may exist, and methods that have been used to detect such dependencies.

For some time fisheries researchers have recognized that anomalous environmental conditions affect the availability of stocks to survey trawls, and in a relative sense introduce bias. Temperature is generally recognized as an influential environmental variable on availability (see Byrne *et. al.*, 1981). Temperature is an important influence in the dynamics of fish stocks in the Barents Sea (see Loeng, 1989). For example, the feeding distributions of cod in the North Sea are more easterly and northerly in warm years than in cold years. Rose *et. al.* (1994) suggest that cod in NAFO Divisions 2J-3L (2J3KL) exhibit more northerly distributions in years with relatively warm water temperatures, and more southerly distributions in cooler years. Atkinson *et. al.* (1997) confirm this distributional shift and also demonstrate an areal contraction of the stock range during this period. Kulka *et. al.* (1995) suggest that by 1989 cod in 2J3KL had shifted south, and to greater ocean depths, in distribution. Both Atkinson *et. al.* (1997) and Kulka *et. al.* (1995) note that the apparent southerly distributional shifts were coincident with anomalously cold ocean temperatures and heavy ice conditions. In opposition, Hutchings and Myers (1994a) found little association between annual average relative abundance in 2J3KL during the fall research surveys and anomalies in annual depth averaged water temperatures at a hydrographic station near St. John's, which they used as a proxy for

anomalies in average water temperatures in 2J3KL.

Stock migrations can be disrupted due to cold water temperatures. Hutchings and Myers (1994b) found that spring migrations of 3Ps cod were significantly related with average (January-June at 50 meters depth) spring water temperatures. They measured migration using the estimated ratio of biomass on the continental shelf and slope. However, to the north in 3L they found that, while migrations varied 10-fold, a relationship with water temperature was not apparent. Wroblewski *et. al.* (1995) found that along-shelf movements of large schools of cod was depth dependent but did not seem to be affected by temperature fluctuations in that the temperatures found throughout the study area appeared suitable for the presence of cod.

Disruptions of migration patterns because of large temperature gradients suggests that cod possess temperature preferences. Sinclair (1992) analyzed summer research survey results for cod in NAFO subdivisions 4Vs and 4W (see Figure 1.2). His results suggest that older fish are found at deeper depths and colder temperatures than younger fish. He concluded that cod “are not found ubiquitously with respect to temperature, but that they occur at the lower end of the observed temperature range”. Sinclair (1992) also suggested that cod are more selective for temperature than for depth. In contrast, Heessen and Daan (1994) observed that the highest densities of young cod in the North Sea are found at the lower end of the temperature range, whereas older cod had their highest densities at the upper end of the temperature range. However, Heessen and Daan (1994) note that these associations may be caused by some other process than active temperature selection. Georges Bank cod off the Northeast coast of the United States are widely distributed during the spring when water temperatures are between 4-6° C (Serchuk *et. al.*, 1994); however, in the fall when bottom water temperatures are greater than 10° C the highest densities of cod are found in the cooler, deeper waters. On a broader scale Serchuk *et. al.* (1994) found that for each degree increase in temperature

the average catch temperature of cod increased only 0.58°C , and these authors concluded that cod must change their spatial distribution to accomplish this.

Swain and Kramer (1995) analyzed annual temperature selection by age for cod in NAFO Division 4T. They compared the temperatures available to those in which cod were caught in September research surveys. They found that cod exhibited temperature preferences which varied from year to year; however, younger fish generally selected warmer water than older fish. Depth selection was evident as well. Cod in 4T tended to select colder water at higher levels of abundance. Swain and Kramer (1995) argue that such temperature selection is evidence for the hypothesis that cod distribution will switch from warm habitats at low levels of abundance to cold habitats at high levels of abundance because fish select habitats that maximize growth rate. As population abundance increases and food supplies decrease colder habitats become preferable because, without food, body mass decreases more slowly in colder conditions.

Water mass characteristics have also received attention in terms of their relationship with stock abundance. Smith *et. al.* (1991) examined fall and spring research survey catches for the Scotian shelf. Proportions of characteristic water masses near the ocean bottom were identified based on ranges of temperature and salinity data. Catches of four year-old cod occurred proportionately more in the cold intermediate layer (CIL). In addition, the largest catches from each survey were always associated with this water mass. Smith *et. al.* (1991) noted that tow depth had previously been shown to be important in predicting the catch of four year-old cod. They performed a regression analysis assuming extra-Poisson variation in numbers per tow, and used a log-link model involving depth, nested within water masses, with an intercept for each water mass and strata. They found that the estimated slopes for depth terms were significant only when nested within the CIL water mass. Generally they found that catches increased at first with depth, but then decreased for depths greater than 100 meters. Smith *et. al.* (1991) also observed

a relationship between estimates of age four cod abundance and the CIL proportion of bottom water. They found that “[d]iscrepancies between March estimates of abundance and those from the cohort analysis are consistent with changes in availability of cod to the trawl gear due to changes in the amount of intermediate layer water on the bottom”. The implication is that changes in survey estimates of abundance may be confounded with changes in water mass characteristics. The patterns were found to hold for all ages of cod caught in the survey.

Recently Smith and Page (1996) presented an analysis of 4Vs and 4W eastern Scotian shelf summer research survey catches in which the variance component related to hydrographic variables was removed, thereby removing “year effects” and “cleaning” the stock abundance signal. Using a variety of statistical analyses they found significant correlations in estimated annual abundance and the proportion of CIL for depths of 0-150 meters. Smith and Page (1996) found a clearer cohort signal after correcting for the CIL effects in the research survey time series. The authors hypothesize that when the proportion of bottom water of the CIL type is large then cod move closer to, and are more widespread along, the bottom. Consequently, cod are more available to the bottom trawl. The authors also recognize that an increase in catchability at lower water temperatures or a mechanism along the lines of Swain and Kramer (1995) could also explain the association between cod and the CIL. The authors end their paper with the cautionary note:

“fluctuations in estimated stock abundance generated by stock assessment procedures that do not consider [hydrographic] influences could potentially be misinterpreted as fluctuations in real stock abundance”.

The previous studies have involved relationships between hydrographic covariates and cod abundance and its availability to bottom trawl survey gear. Several of these papers mention that variations in availability are potentially confounded with variations in catchability. A relatively extensive amount of fisheries literature has been published on trawl

catchability. Remarkably, there are few ground-truth experiments, due no doubt to the difficulties in conducting such experiments. The catchability literature is not reviewed here in detail; instead, we summarize the discussions in Engås (1994), Godø (1994), and Walsh (1996). The main points in these papers is that bottom trawl catchability is dependent on:

1. fish size, age, and density, through schooling behaviors and the lack of avoidance reactions at high densities;
2. mesh size;
3. tow depth, through variations in trawl geometry, particularly wingspread;
4. bottom water temperature, through variations in the swimming speeds of fish;
5. bottom type, through herding due to sand clouds, etc.;
6. light conditions, through variations in fish behavior;
7. vessel type, and noise, through variations in fish behavior;
8. tow speed and direction;
9. current relative to the ship trajectory.

4.9. Appendix: Figures

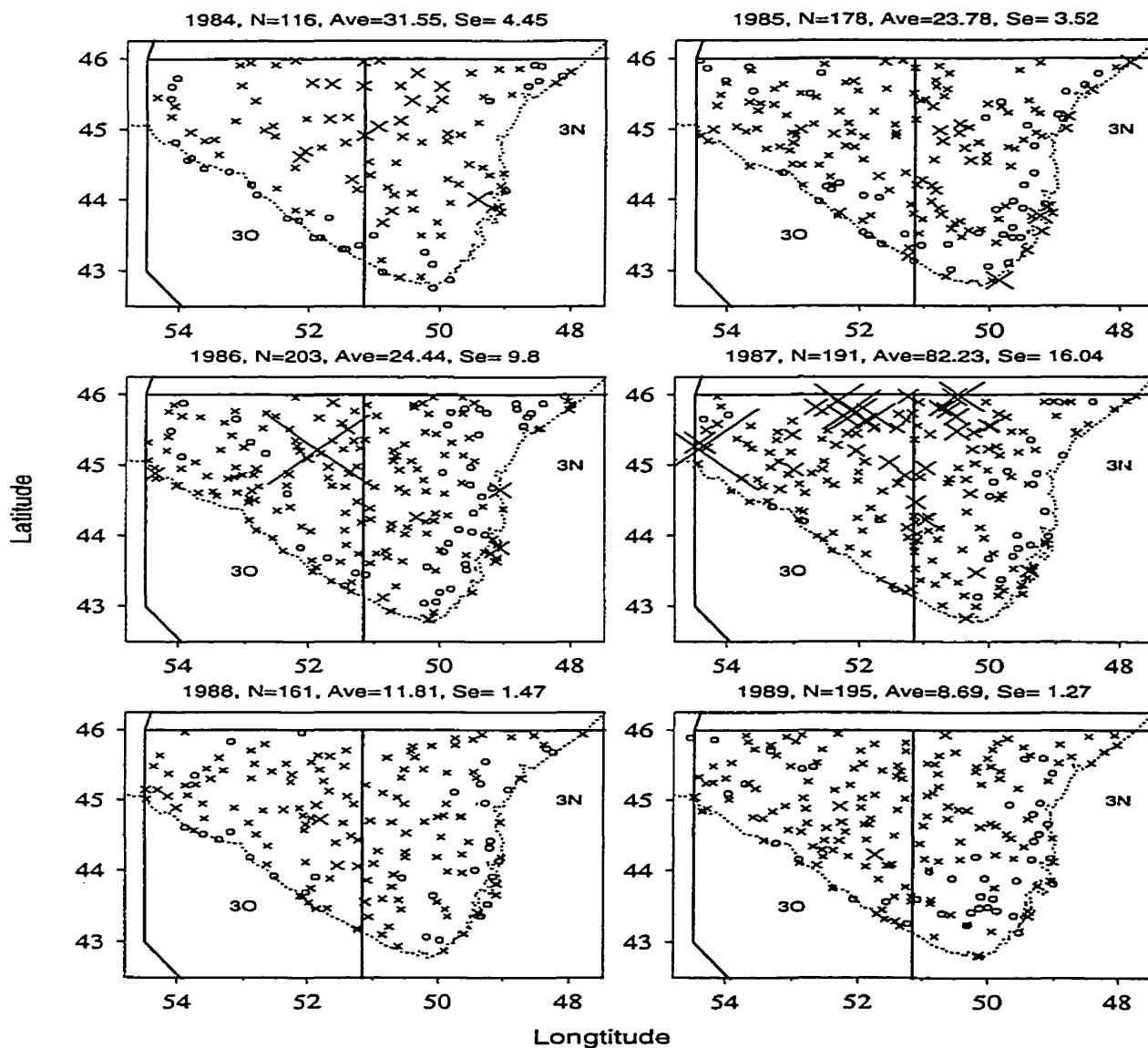


Figure 4.1a: Spring 3NO bottom trawl research survey catches of cod. Plotting symbols mark tow locations. A $\circ$ denotes no catch. The $\times$ size is proportional to the catch. The - - - line marks the 300 meter depth contour. N, Ave, and SE denote the number of tows, the sample mean catch, and the standard error.

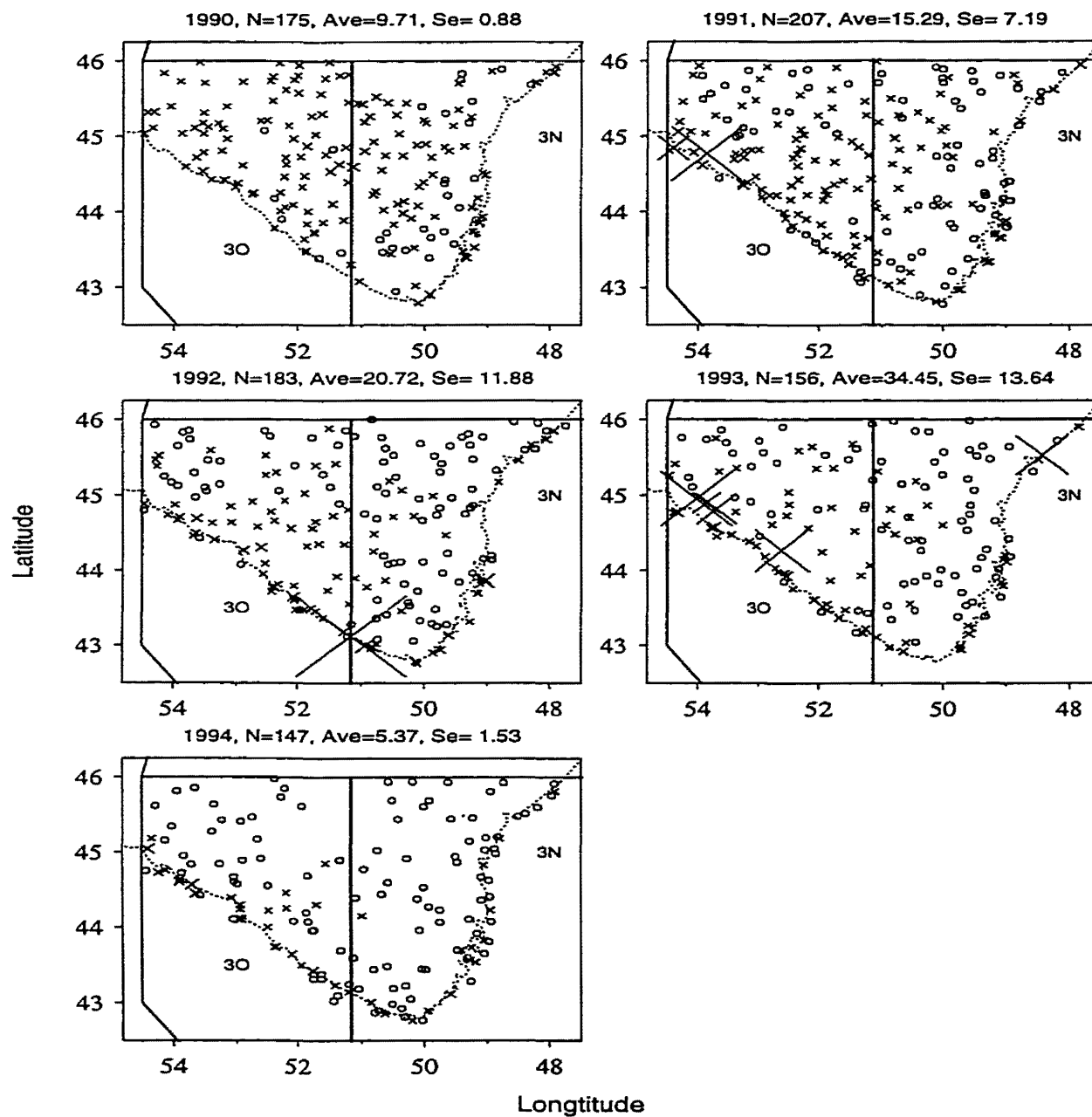


Figure 4.1b. Same as Figure 4.1a.

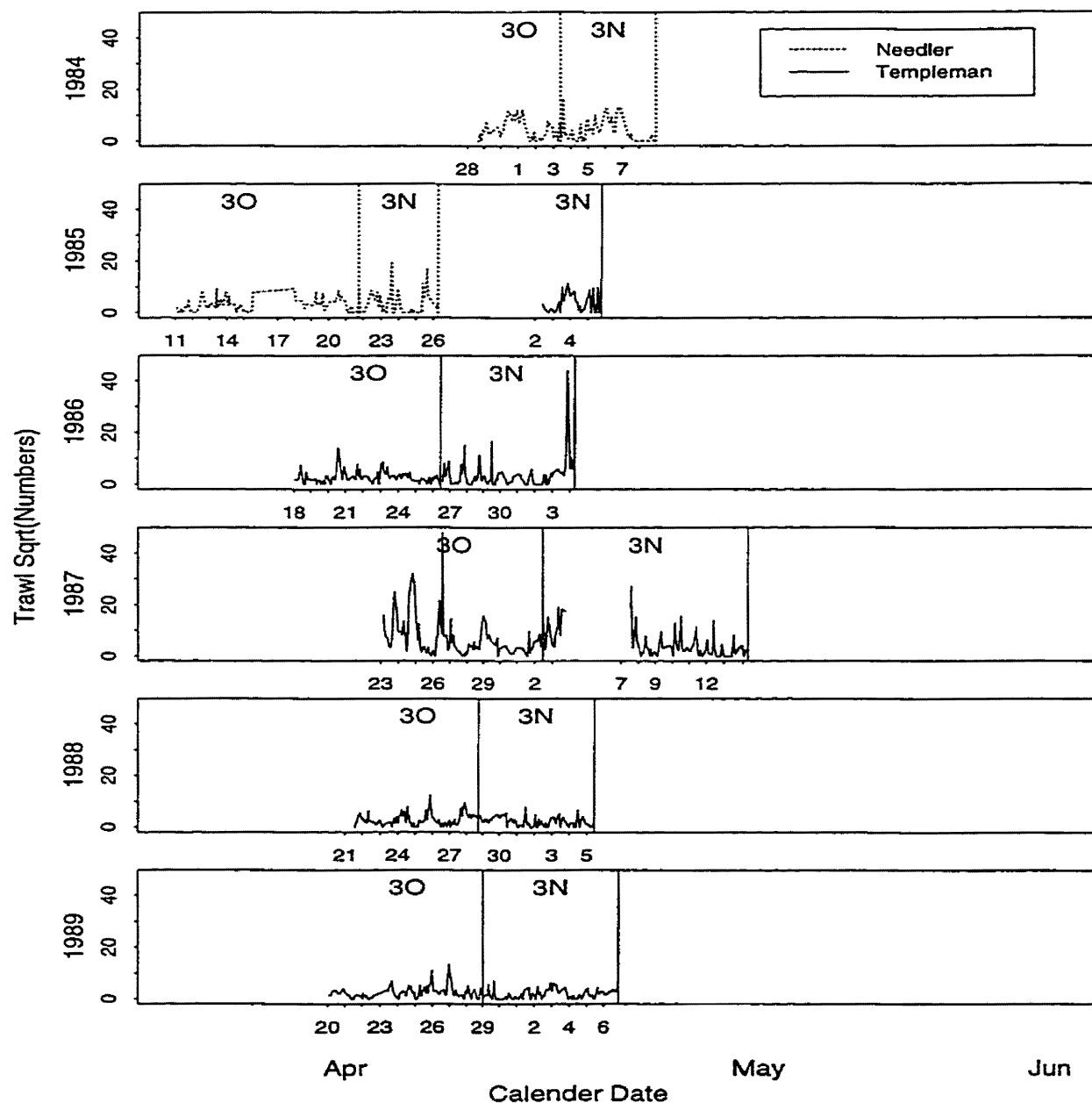


Figure 4.2a: Spring 3NO bottom trawl research survey time series of catches. The vertical lines mark when the survey vessel moves to a different NAFO division. Different line types are used for different vessels.

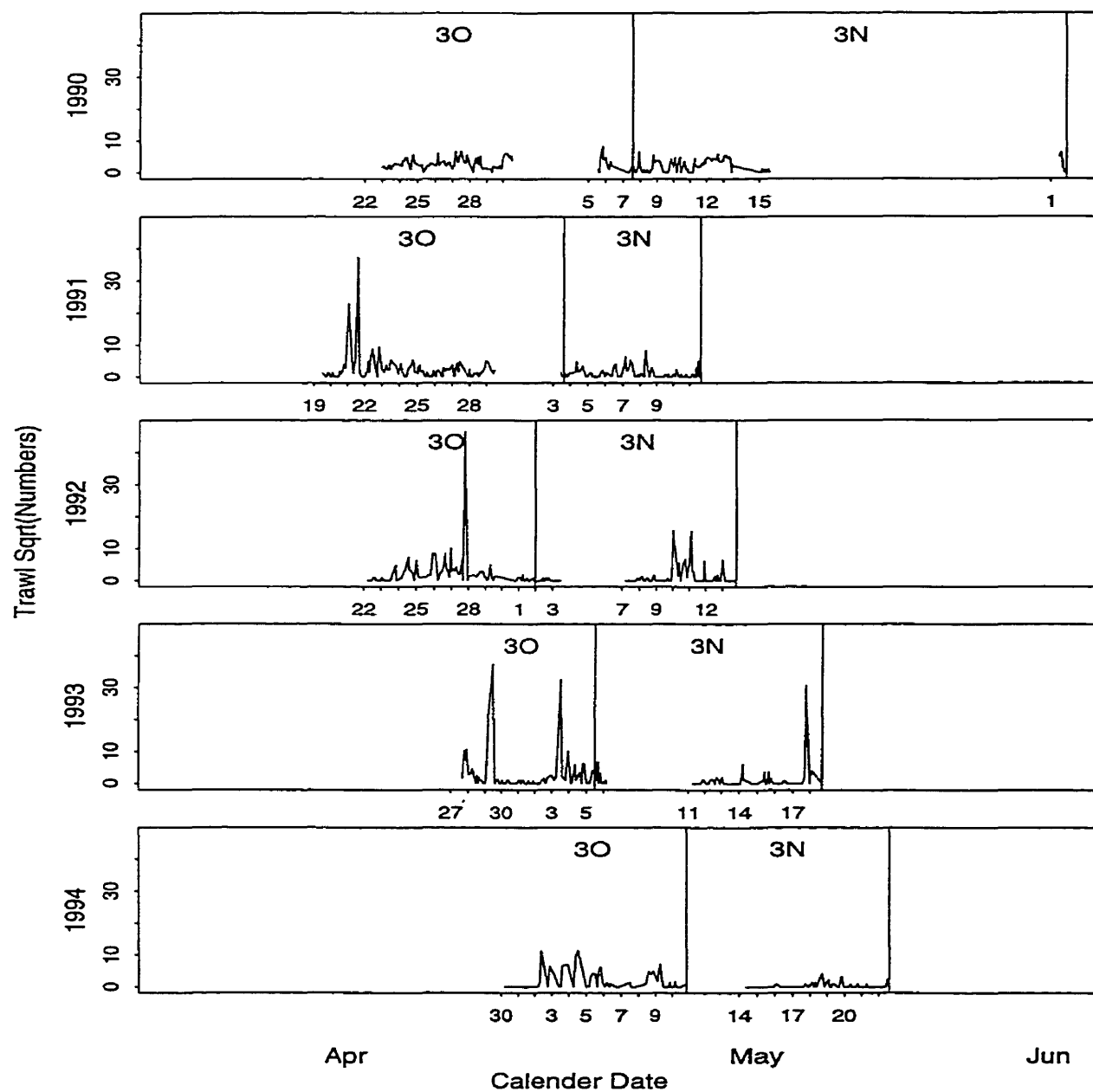


Figure 4.2b: Same as Figure 4.2a.

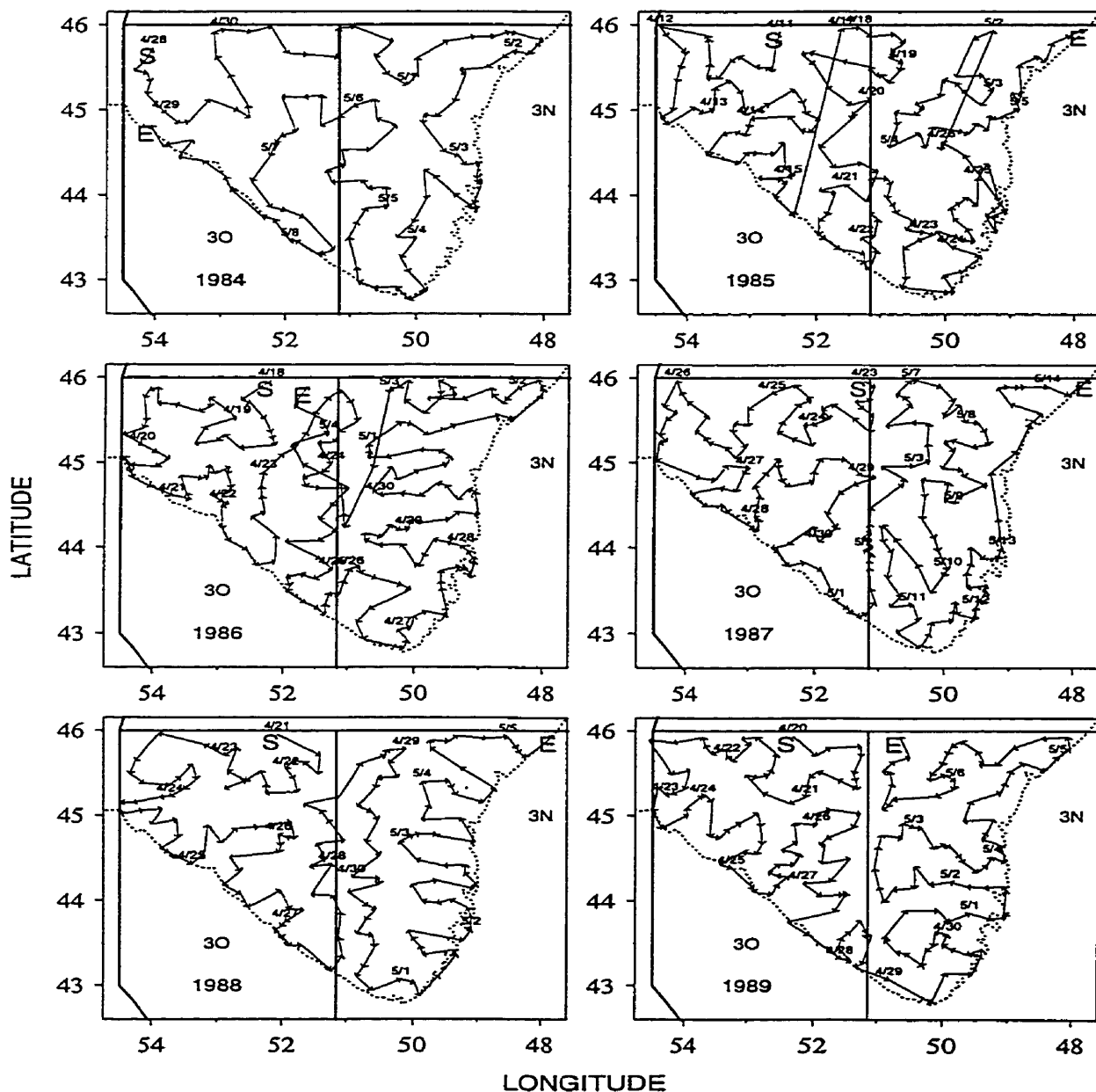


Figure 4.3a: Spring 3NO bottom trawl research survey trajectory. The >'s mark tow locations. The month/day is plotted near the first tow of a day. The - - - line marks the 300 meter depth contour. S and E mark the beginning and ending survey locations.

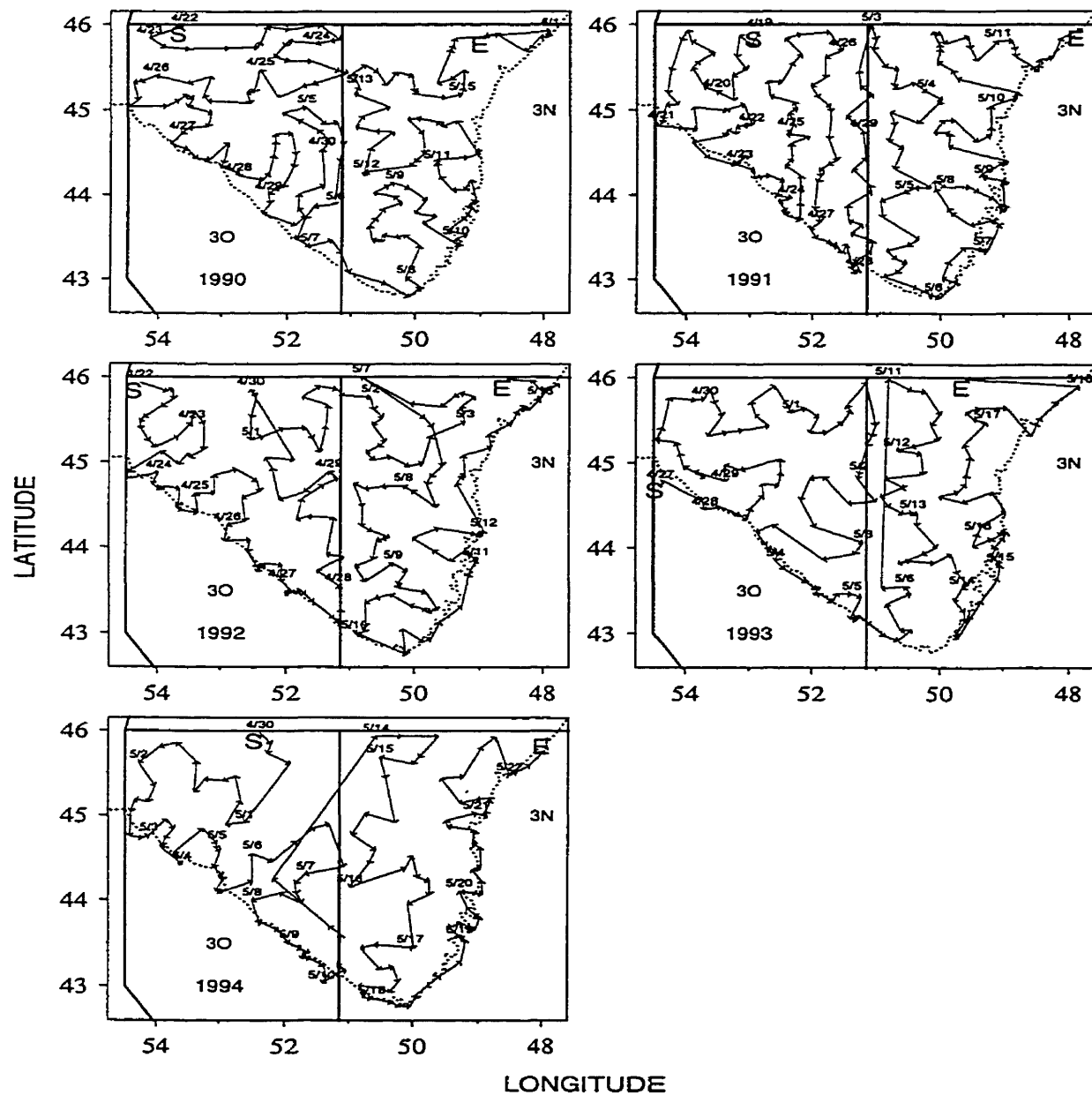


Figure 4.3b: Same as Figure 4.3a.

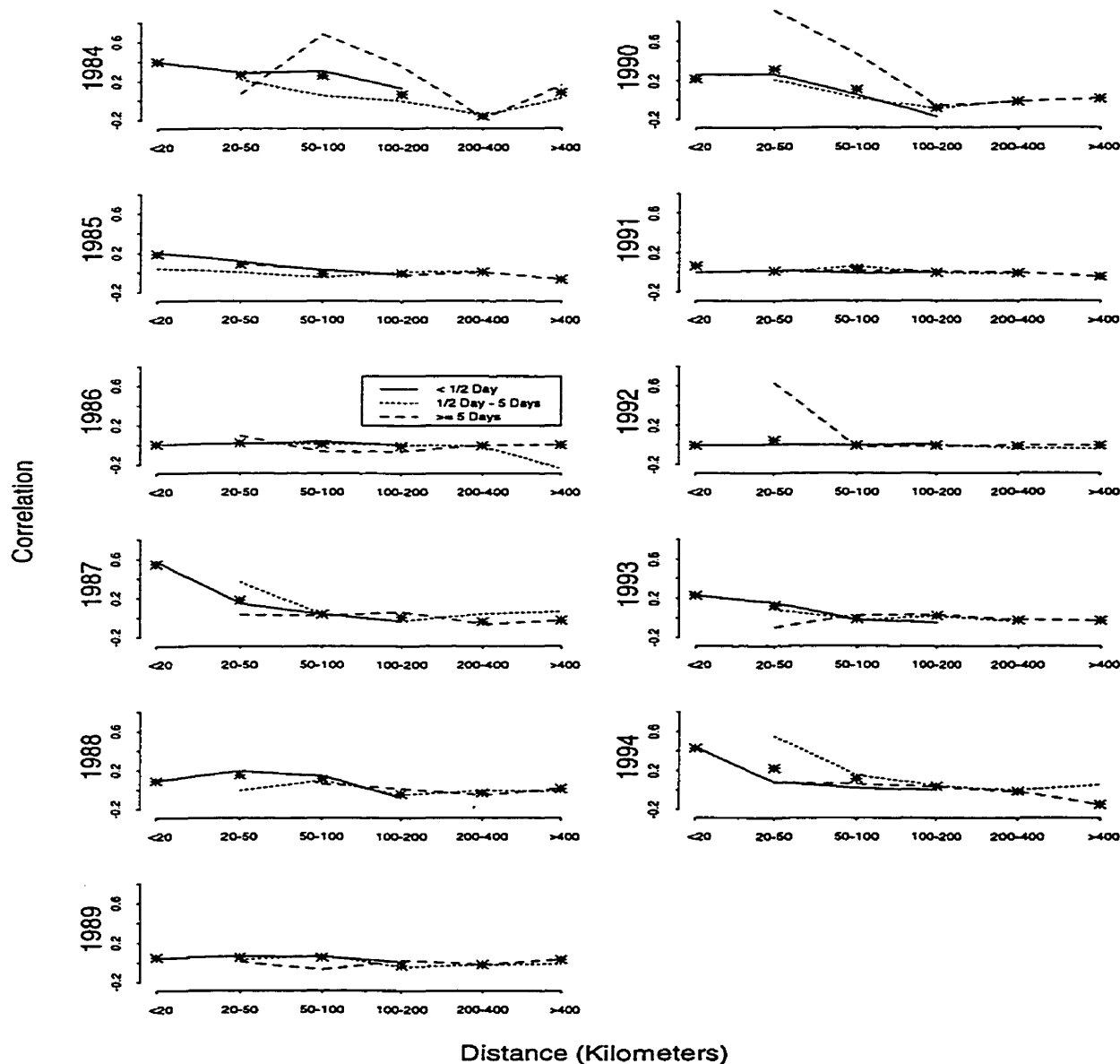


Figure 4.4a: Spring 3NO bottom trawl research survey spatial-temporal autocorrelations in catches. Spatial autocorrelations are plotted as *'s for six distance classes. Spatial-temporal autocorrelations are plotted as lines for the same distance classes and three time classes.

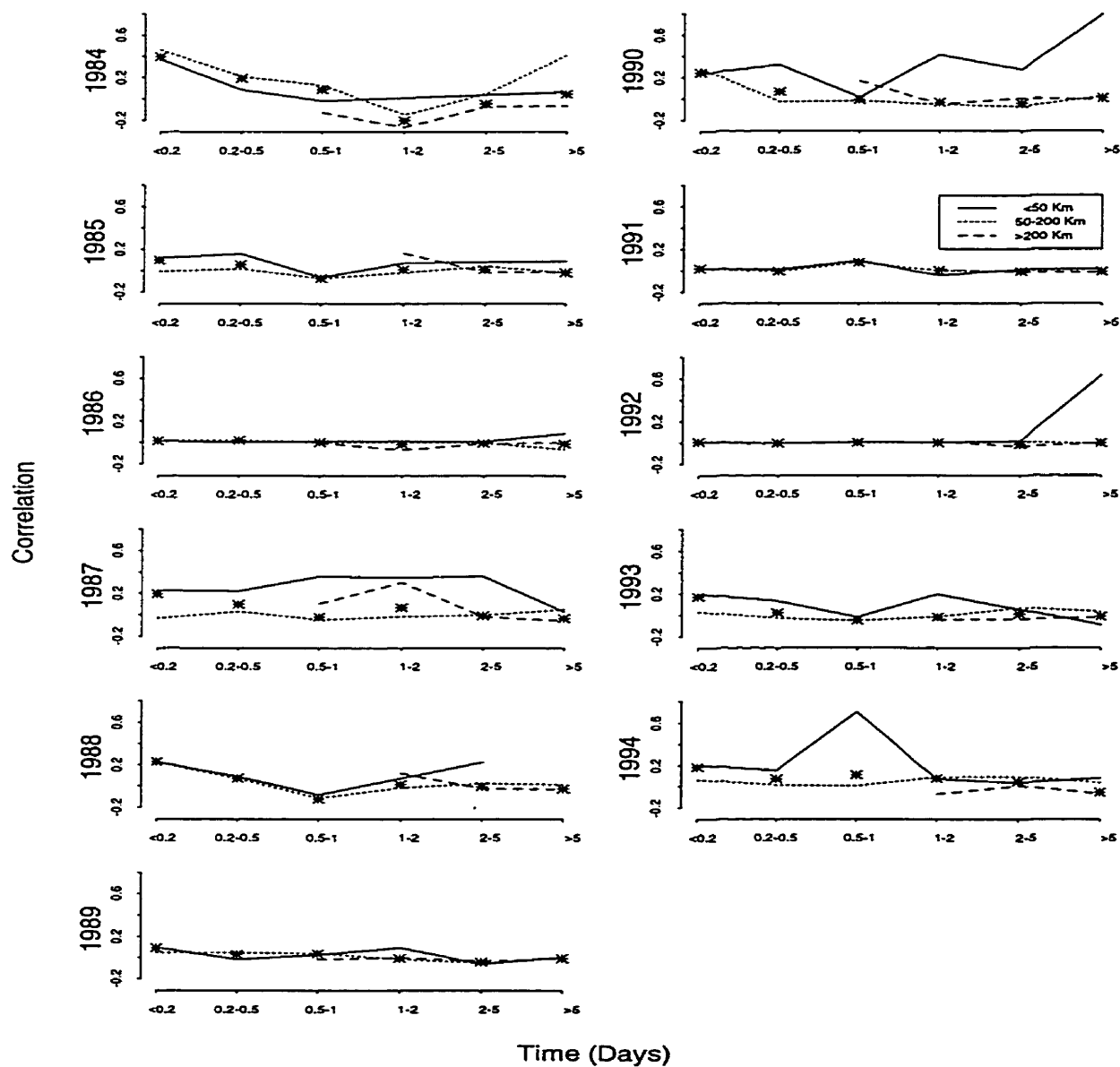


Figure 4.4b: Spring 3NO bottom trawl research survey spatial-temporal autocorrelations in catches. Temporal autocorrelations are plotted as *'s for six time classes. Spatial-temporal autocorrelations are plotted as lines for the same time classes and three distance classes.

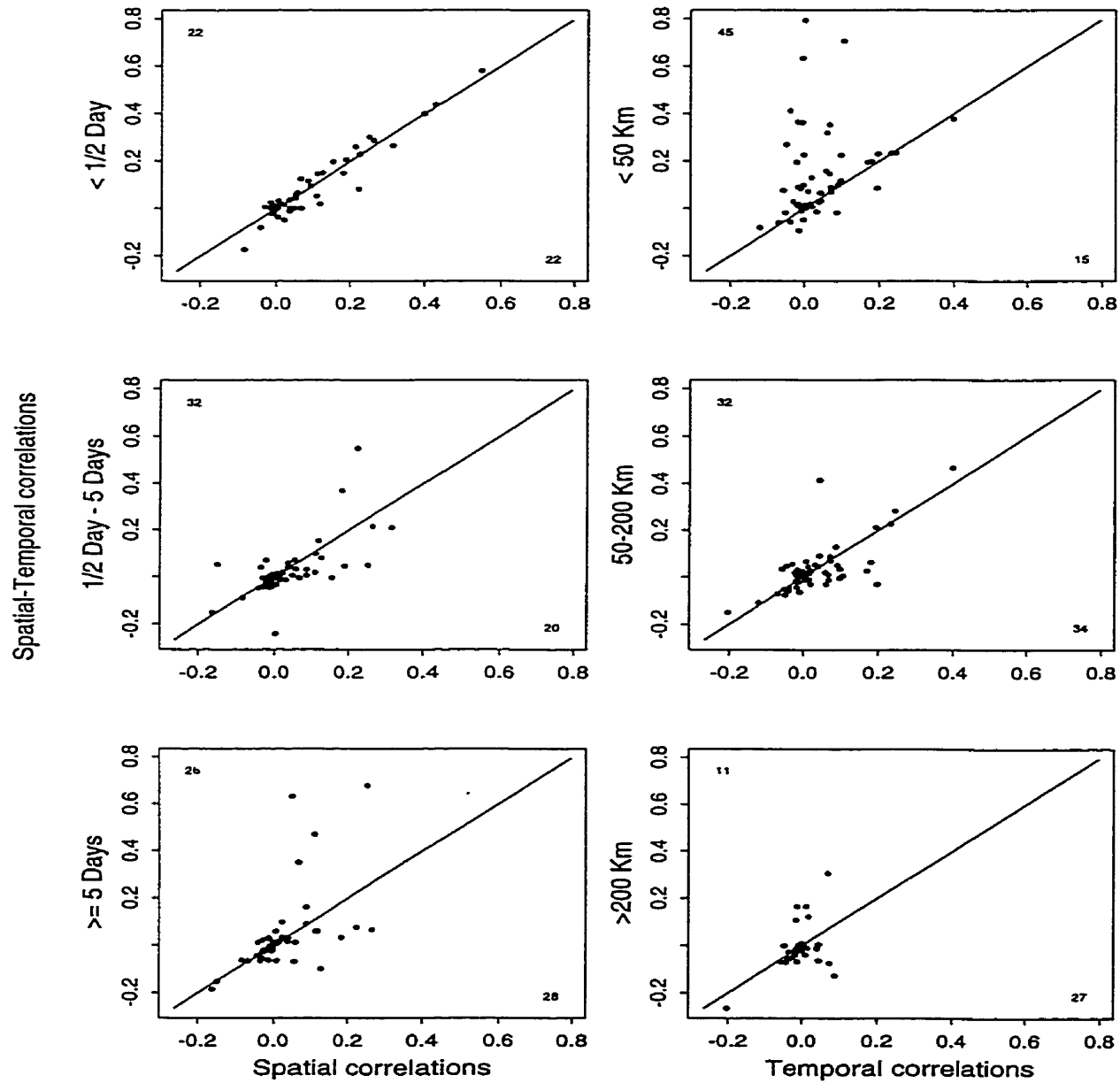


Figure 4.4c: Comparison of spatial and spatial-temporal autocorrelations in spring 3NO bottom trawl research survey catches. The number of pairs of correlations falling above and below the solid line are presented in the upper and lower corners of each panel.

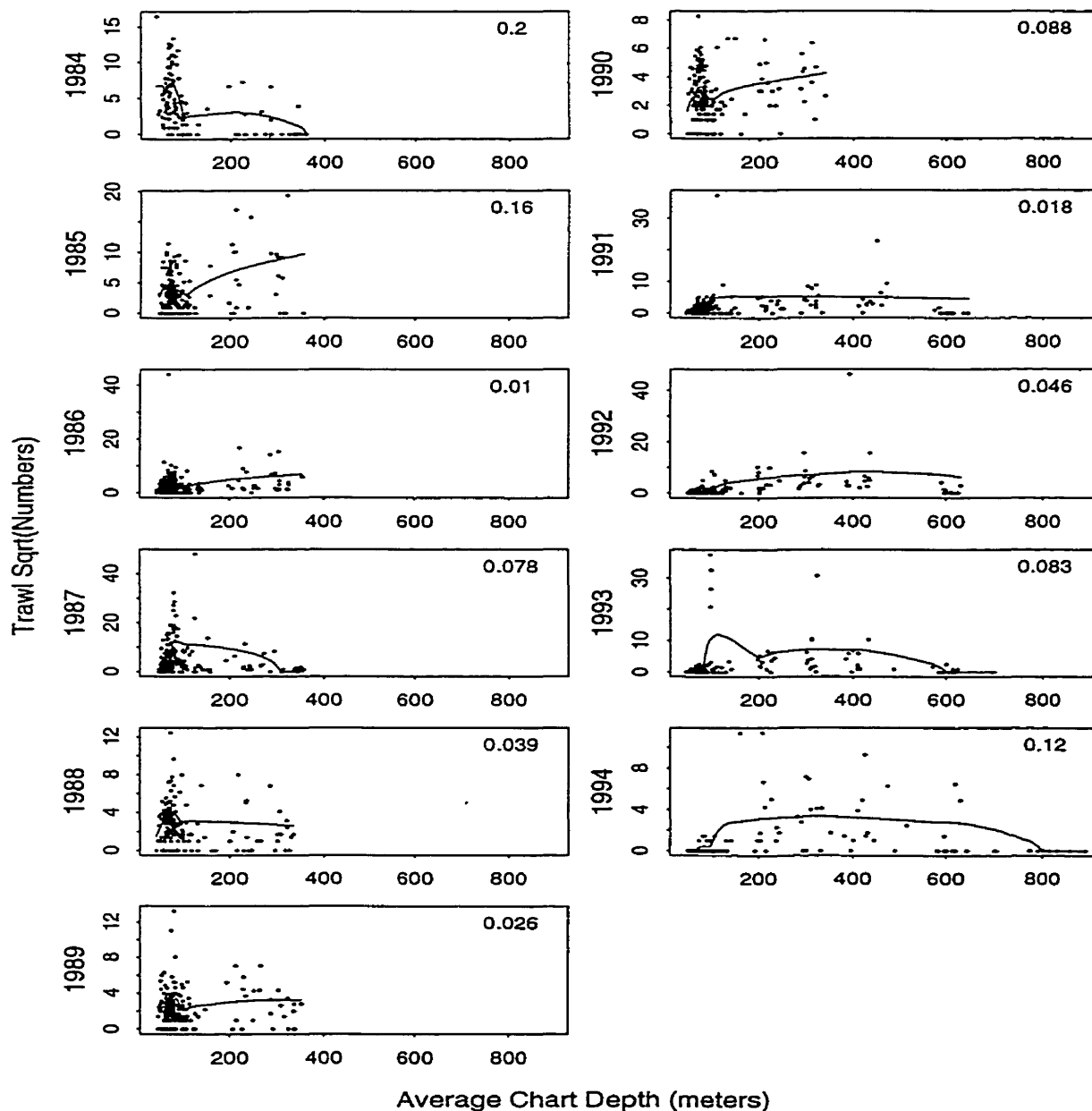


Figure 4.5a: Catches versus tow depth for the spring 3NO bottom trawl research survey. The line corresponds to a smooth between catch and depth. The R^2 of the smooth is presented in the upper corner of each panel.

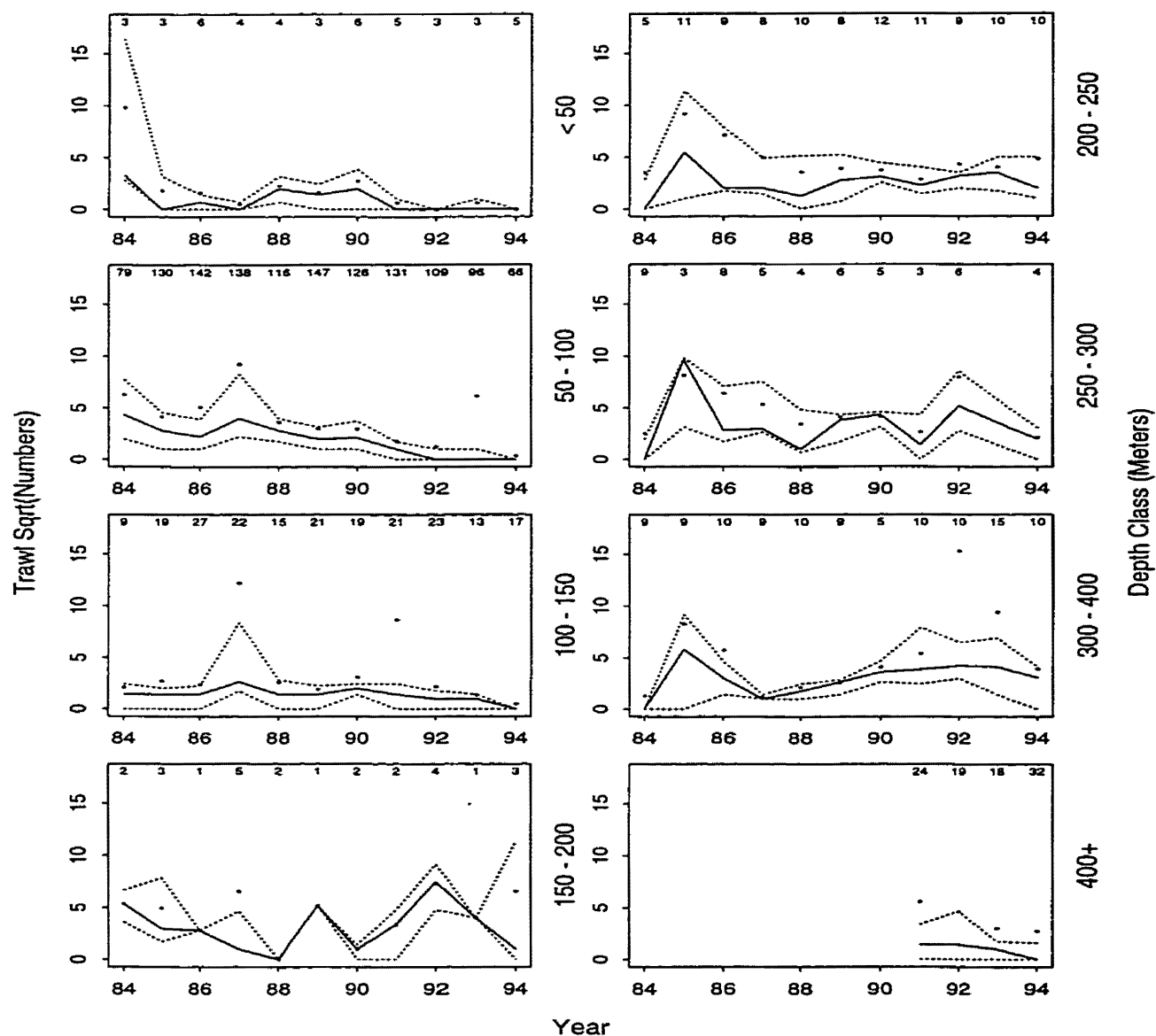


Figure 4.5b: Time series of catches at depth intervals for the spring 3NO bottom trawl research survey. The solid line connects the annual medians, and the dotted lines connect the 25th and 75th percentiles. Annual means are plotted as points. The number of tows in each year and depth interval is indicated at the top of the plot.

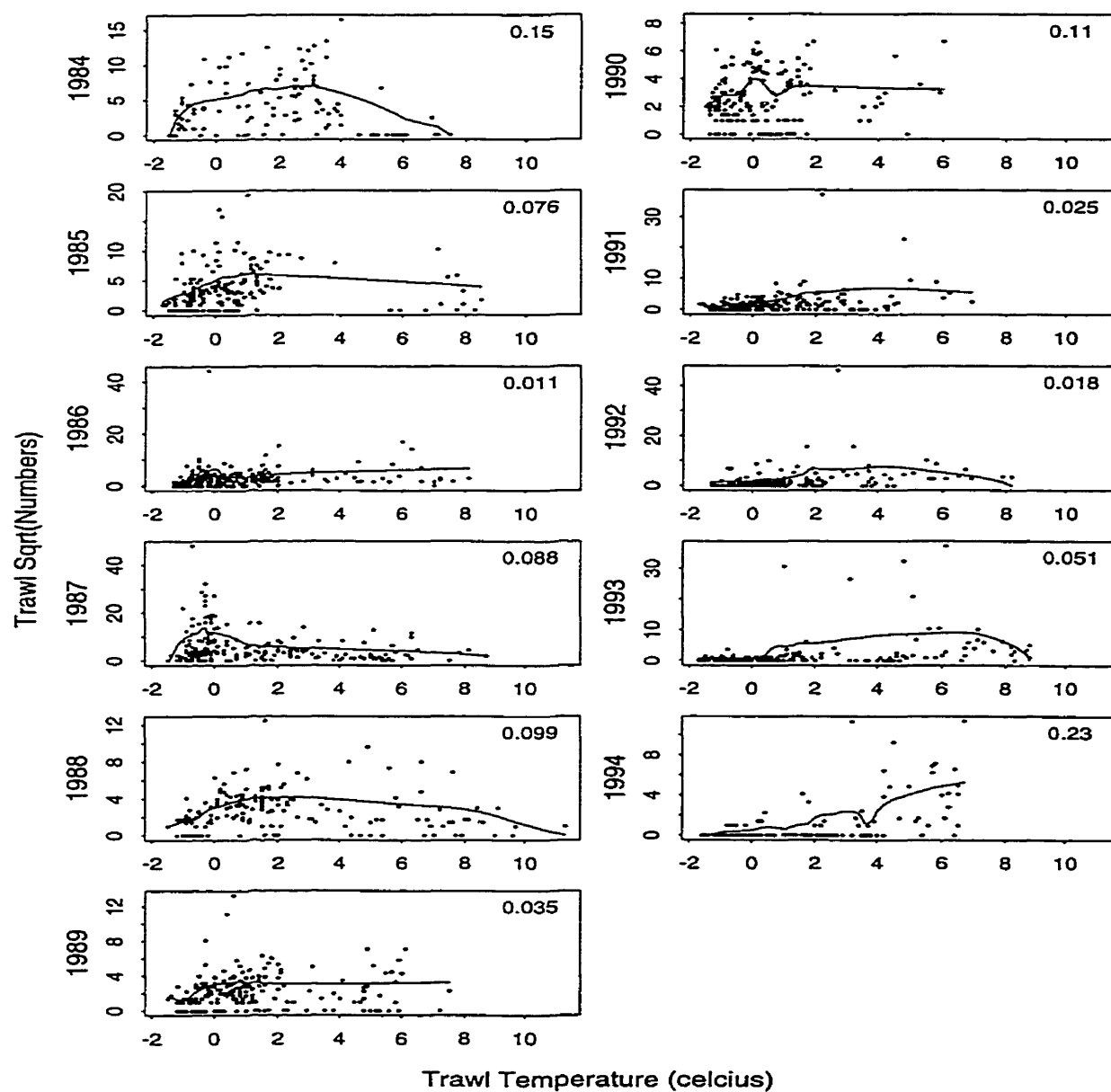


Figure 4.6a: Catches versus bottom temperature for the spring 3NO bottom trawl research survey. The line corresponds to a smooth between catches and temperature. The R^2 of the smooth is presented in the upper corner of each panel.

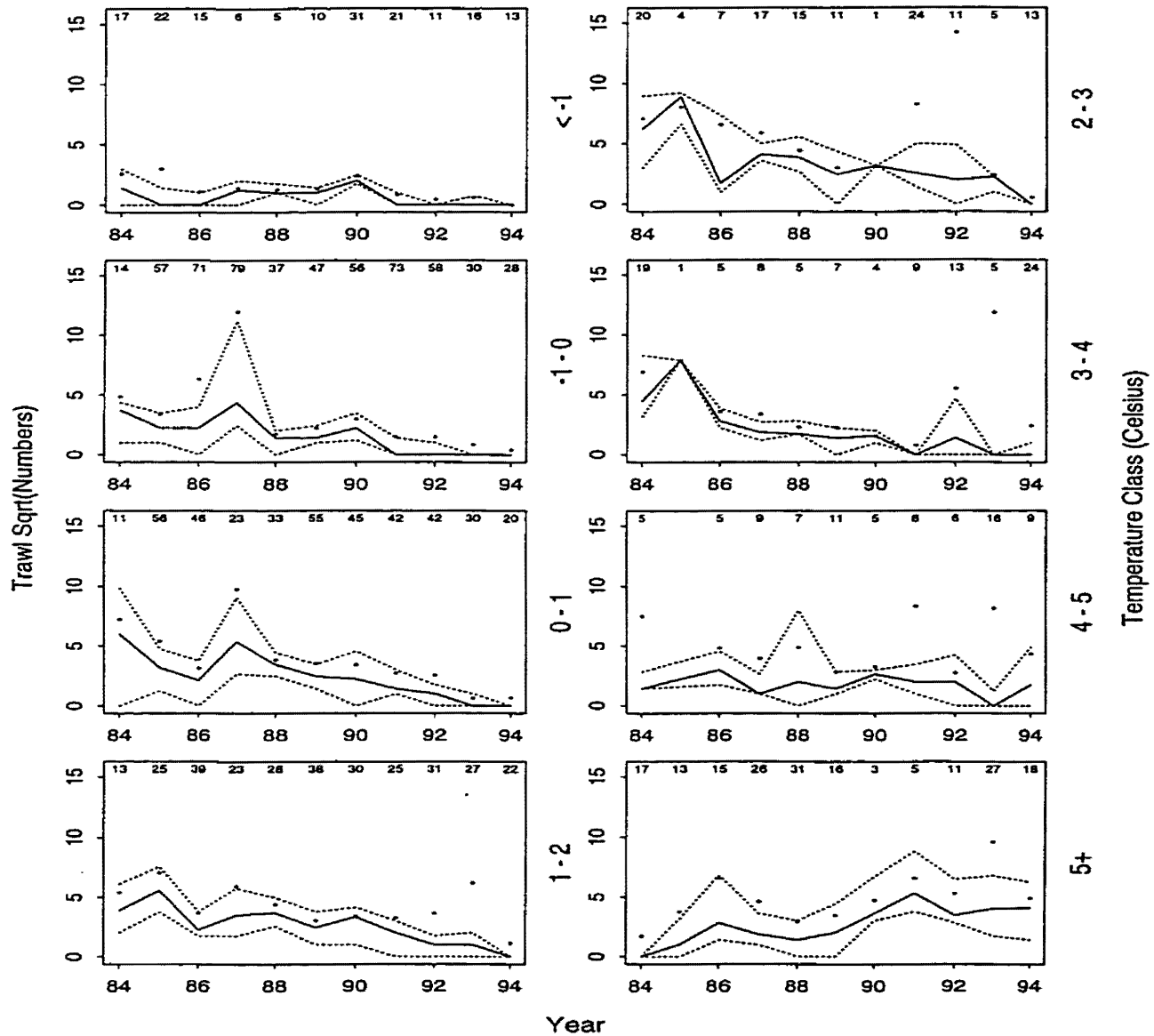


Figure 4.6b: Time series of catches at temperature intervals for the spring 3NO bottom trawl research survey. The solid line connects the annual medians, and the dotted lines connect the 25th and 75th percentiles. Annual means are plotted as points. The number of tows in each year and temperature interval is indicated at the top of the plot.

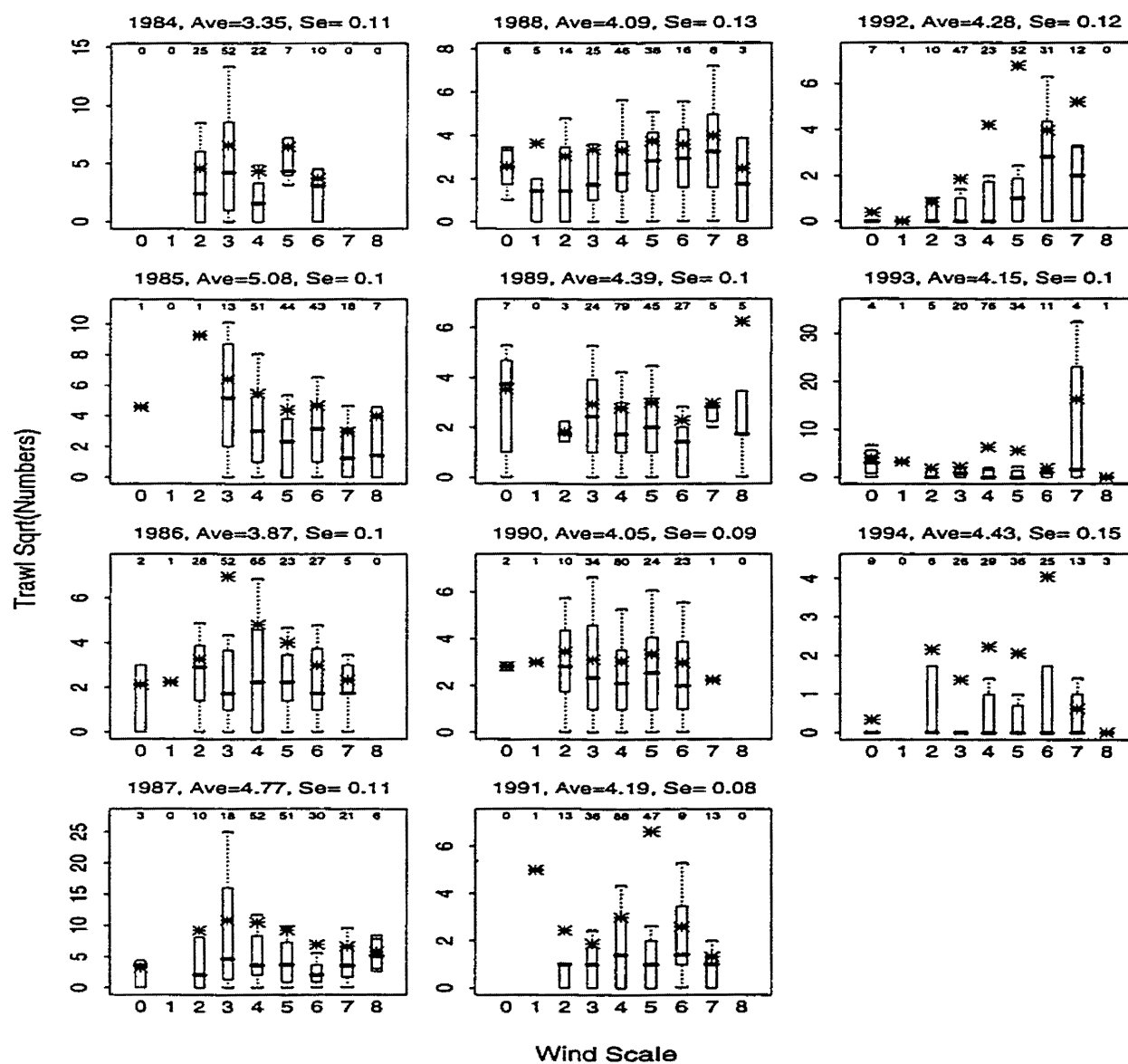


Figure 4.7a: Distributions of catches versus wind force for the spring 3NO bottom trawl research survey. The mean is plotted as an *, and the median is plotted as a heavy solid line. Box endpoints correspond to 25th and 75th percentiles. Dotted lines connect upper and lower “extremes” that do not include “outlying” observations (see Splus documentation). The average wind force and its standard error are printed at the top of each panel.

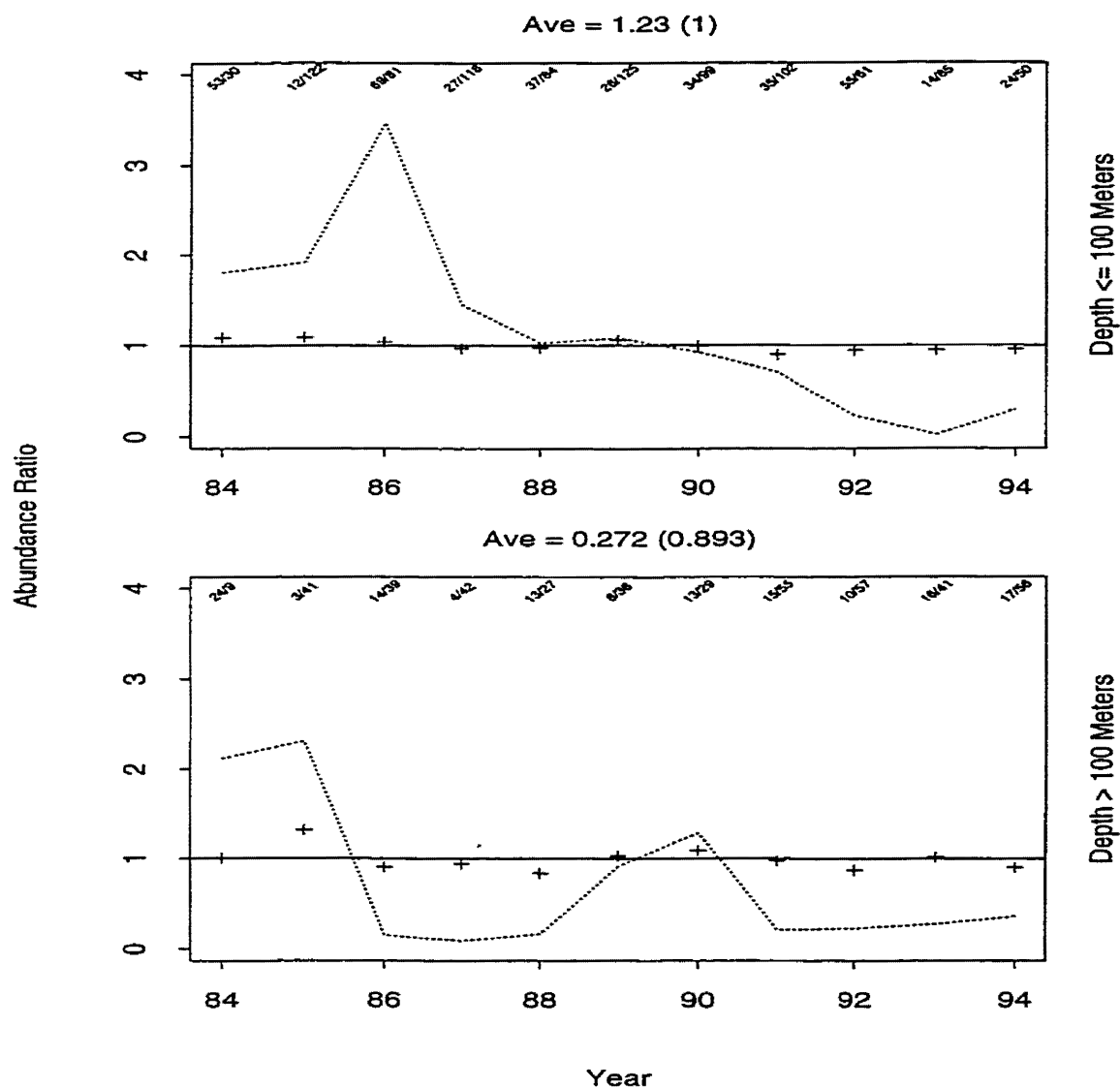


Figure 4.7b: Depth stratified annual ratios of average catches for low versus high winds (wind force scales 0-3 and 4-9). The combined average ratio for all years is indicated at the top of each panel, and is followed by the ratio of ranks. The number of tows in the numerator and denominator are indicated along the top of the first two panels.

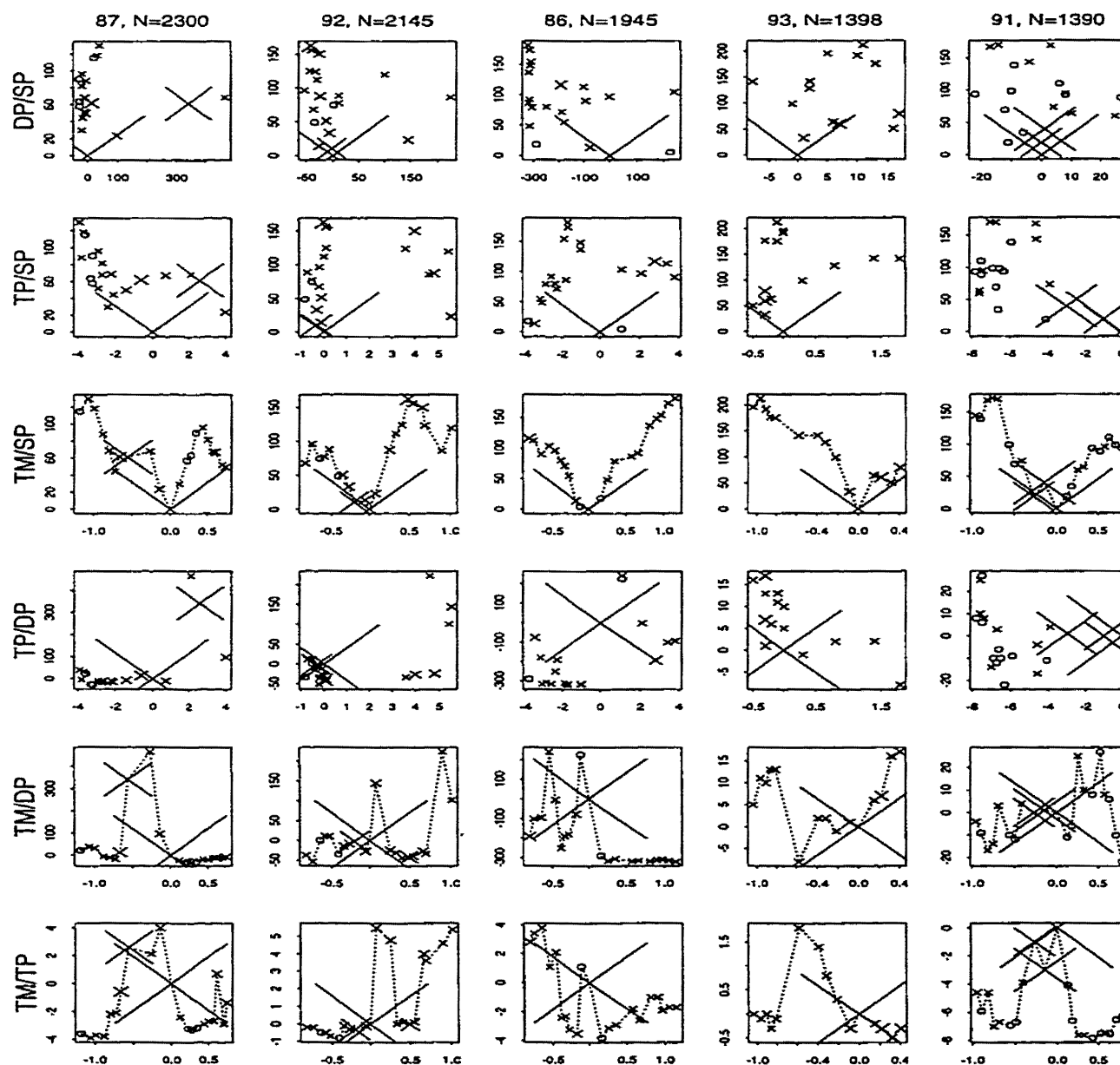


Figure 4.8: Synoptic analysis of the five tows with the largest catch of cod during 1984-1994. Ten tows taken before and after the large tow are included. The $\times$'s are proportional to catch. Spatial (SP), depth (DP), temperature (TP) and temporal (TM) distances are relative to the large tow. DP/SP indicates that depth is the horizontal axis, and Spatial location is the vertical axis, etc.

5. MODEL SPECIFICATION

5.1. Introduction

In this chapter we present all the analyses necessary to specify a defensible stochastic model for 3NO research survey catches of cod. Inference from this model is the subject of Chapter 6. The analyses in Chapter 4 suggest associations between trawl catches and spatial location, time of tow, and possibly other covariates. Consequently, a statistical model that utilizes these associations may improve the precision of our inferences about stock size. The graphical analyses also revealed that the covariate associations do not appear to have simple parametric forms, and that interactions between covariates may be important. To accommodate this, the model we use is nonparametric, and involves smoothing catches over neighborhoods defined by the covariates. This procedure can provide consistent estimation of local trawlable abundance for a wide range of covariate relationships. However, occasional large catches occur that are inconsistent with nearby catches, and cannot be described by the same model we use for the majority of the data. These large catches do not reflect fish abundance at nearby locations. They appear to occur randomly, and should not be used to predict stock density in regions with similar covariate values. This is not to say that we should ignore the large catches. In fact, they provide very important information about stock size, for that component of the stock that forms dense, apparently random aggregations. We referred to these large catches in Chapter 4 as type II catches, while the others were referred to as type I catches. A

simple mixture model with two components for the two types of catches was proposed in **Chapter 4**, and this model will be specified in more detail in this chapter.

We model type I and II catches as overdispersed Poisson random variables. The Poisson density parameter for type I catches is modelled as a smooth function of covariates using nonparametric regression techniques. In **Chapter 4** we were not able to specifically determine which covariates to use in the model, although we did eliminate several variables from future consideration. In this chapter we use nonparametric variable selection procedures to determine exactly which covariates to use in our smooth mean function model (SMFM) for type I catches. We estimate the incidence of type II catches separately over several large geographic regions; however, we assume that their density is constant from year to year. The difference in annual trends in median and average survey catches shown in Figure 5.4b in **Chapter 4** provides some evidence for this. Median catches have declined consistently at depths less than 150 meters (most of 3NO), while average catches have not. This suggests that in most areas sampled abundance has declined, but not in the small area that produces large catches. The small number of tows that have been observed from very high density cod aggregations do not suggest a change in density of these aggregations, so we assume this density is constant annually. A problem with the mixture model approach is that we do not know which catches are type I or II a priori. As a solution we use robust nonparametric regression methods to assist in detecting large catches that are inconsistent with nearby catches. These are then designated as type II catches.

The mixture-model has similar characteristics to theoretical population models for density-dependent habitat selection, in which population size and local density are important factors in determining the population distribution among patchy habitats (MacCall, 1990). Differential utilization of marginal habitats with changes in total population abundance has been observed in many species, including fish. Hutchings (1996) has sug-

gested that cod in NAFO Divisions 2J, 3K, and 3L (adjacent to 3NO; see Figure 1.3) exhibit density-dependent habitat selection. He proposed that cod in this region occur in three types of aggregations: low, medium, and high density. High-density aggregations maintain their size as population abundance declines because fish from lower density aggregations move in to more favorable high-density habitats as space becomes available. Thus, a decline in total population abundance is first associated with a decline in local low- and medium-densities, but not high-densities. High-density aggregations decline only after there are insufficient numbers of fish remaining in the population to replenish these aggregations. Hutchings (1996) refers to this process as “the nested aggregation model”, which provides the motivation for the mixture-model we use for 3NO cod. Hutchings (1996) also suggests that a similar process may have occurred for cod in the Southern Gulf of St. Lawrence (4TVn).

In this chapter we present all the analyses necessary to build a model for inferences about trawlable stock size, but these inferences are the subject of **Chapter 6**. In this chapter a kernel smoother is described in **Section 5.2** that is used to estimate the SMFM for type I catches. This involves locally weighted averages of trawl catches. The weights are functions of the covariates, and the size of the weights is determined using “bandwidth” parameters. Which covariates to use for smoothing is investigated in **Section 5.3**. We apply the nonparametric regression and variable selection procedures to the 3NO trawl survey data in **Section 5.4**. In **Section 5.5** we develop a robust nonparametric regression procedure suitable for trawl data. First we review robust M-estimation, and develop robust M-estimators for the mean of a Negative Binomial distribution, which we model trawl catches with. Simulations are conducted to examine some fixed sample properties of the robust estimators. These results are then used to develop robust smoothing procedures for trawl catches. These procedures are used for “robustified” variable selection in **Section 5.6** using the trawl survey data, and for identifying type II catches in **Section 5.7**.

In **Section 5.8** we finish specifying our model, having determined which covariates to include in our SMFM for type I catches and which catches are type II's. In **Section 5.9** we compare this model with the traditional stratum-effects model used to summarize trawl survey data. The comparison is based on overall fit and prediction performance. The main results in this chapter are summarized in **Section 5.10**.

In this chapter we conclude that spatial location and tow depth are the most important predictive covariates in a SMFM for trawl catches. Tow temperature, time between tows, and wind speed have less predictive power beyond their association with spatial location and tow depth, and do not seem useful for our model. We identify six type II catches for the period 1984-1994, and four of these occurred in 3O on the southwest slope of the Grand Bank. The nonparametric regression model we use for type I catches is superior to a stratum-effects model, both in terms of fitting the data and in a prediction sense.

5.2. Local likelihood smoother

In this section we describe the nonparametric regression estimator that we use to estimate the mean of type I catches. The essence of modelling trawl survey data sets such as what we have is to estimate the expected trawlable density of fish at every location in the survey region. Since there are no measurements at many locations this expectation has to be inferred from neighboring observations. For this inference a fundamental assumption has to be made, which is that fish abundance changes smoothly with spatial location, depth, etc. The amount of smoothness will be determined by the data. Some catches seem to conflict with this assumption, and we need to determine exactly which catches these are. We also need a mechanism to account for the component of the fish stock that these catches represent, so that we can estimate the size of this component and add this to the estimated size of the other stock component - the one that varies smoothly.

There are several methods discussed in the literature for using neighboring information.

Parametric models involving explicit functional relationships between expected trawlable density and covariates are very common, but do not seem useful for the 3NO data set, as was illustrated in **Chapter 4**. We restrict our attention to nonparametric models. These methods proceed by assuming a simple model for trawl catches within a local neighborhood. The most common model is a constant mean, which is what we use, but low order polynomial models are also used. In this section we discuss the local maximum likelihood method proposed by Tibshirani and Hastie (1987).

We begin describing this method by first assuming we have independent random variables $Y_1, \dots, Y_n$ with density function $f(y|\theta) = c(y) \exp\{\theta y - b(\theta)\}$. In this case $E(Y) = b'(\theta) = \mu$ and $Var(Y) = b''(\theta) = \sigma^2$. The log-likelihood function is given, except for a constant, by

$$l_n(\theta) = \sum_{i=1}^n \{\theta y_i - b(\theta)\}.$$

Maximizing this likelihood gives $\hat{\mu} = b'(\hat{\theta}) = \bar{y}$. As $n \rightarrow \infty$ $\hat{\mu}$ is weakly consistent for μ .

For our problem we assume that observations are independent, but the means vary; that is, Y_i has density $f(y_i|\theta_i)$. We cannot consistently estimate θ_i without further assumptions because we never have more than one observation at each trawl station. Also, without further assumptions we cannot estimate θ_i at stations that were not sampled. Let $\mathbf{x}_i$ be a vector of covariates observed at trawl station i . A common assumption that is reasonable in our case is that the θ_i vary smoothly as a function of $\mathbf{x}_i$. The statistical problem then is to estimate this function, or, equivalently, to estimate $\mu(\mathbf{x})$. We cannot directly proceed to estimate $\mu(\mathbf{x})$ via maximum likelihood without specifying this function more exactly, which we do not want to do in our analysis. Local averaging is a commonly used alternative to estimate $\mu(\mathbf{x})$.

Define an ε neighborhood of observations around some given $\mathbf{x}$ as

$$A_{\mathbf{x}} = \{i : \|\mathbf{x}_i - \mathbf{x}\| \leq \varepsilon\}.$$

Assume that $\mu(\mathbf{x}_i) = \mu(\mathbf{x})$ for all $i \in A_{\mathbf{x}}$, which is approximately true when ε is small. In this case the log-likelihood for $\mu(\mathbf{x}) = \mu_{\mathbf{x}}$ is

$$l(\mu_{\mathbf{x}}) = \sum_{i \in A_{\mathbf{x}}} \{\mu_{\mathbf{x}} y_i - b(\mu_{\mathbf{x}})\},$$

and the maximum likelihood estimate (mle) for $\mu_{\mathbf{x}}$ is

$$\frac{\sum_{i \in A_{\mathbf{x}}} y_i}{|A_{\mathbf{x}}|},$$

where $|A|$ is the cardinality of the set A . Only y_i 's with $\mathbf{x}_i$ values within the distance ε of $\mathbf{x}$ have a full contribution to the likelihood and the estimate of $\mu_{\mathbf{x}}$. Other y_i 's have no contribution to the likelihood. Another approach is to use a weight function designed so that the contribution of y_i changes gradually according to the distance between $\mathbf{x}_i$ and $\mathbf{x}$. In local likelihood estimation this is achieved using a weighted local likelihood

$$l(\mu_{\mathbf{x}}) = \sum_{i=1}^n W \left\{ \frac{(\mathbf{x}_i - \mathbf{x})}{\mathbf{b}} \right\} \{\mu_{\mathbf{x}} y_i - b(\mu_{\mathbf{x}})\}.$$

The weight function measures the distance between the vectors $\mathbf{x}$ and $\mathbf{x}_i$. The size of the neighborhood is determined by the vector of bandwidths, $\mathbf{b}$. The division $(\mathbf{x}_i - \mathbf{x})/\mathbf{b}$ is elementwise. The maximum local likelihood estimate (lle) is

$$\hat{\mu}_{\mathbf{x}} = \frac{\sum_{i=1}^n W \left\{ \frac{(\mathbf{x}_i - \mathbf{x})}{\mathbf{b}} \right\} y_i}{\sum_{i=1}^n W \{(\mathbf{x}_i - \mathbf{x})/\mathbf{b}\}}.$$

The Gaussian weight function, $W(\mathbf{x}) \propto \exp(-\mathbf{x}'\mathbf{x}/2)$, is used here, although other functions are commonly used. The Gaussian weight function defines elliptical neighborhoods in terms of spatial location, tow depth, and other covariates. Different bandwidths are used for each covariate because the smoothness of $\mu_{\mathbf{x}}$ will not be the same over all covariate dimensions; for example, a 100 meter change in tow depth will potentially have a much greater effect on local cod abundance than a 100 meter change in the spatial location of a tow. As $\mathbf{b} \rightarrow \mathbf{0}$ the neighborhood includes just $\mathbf{x}_i$. As $\mathbf{b} \rightarrow \infty$ all observations are included in the neighborhood. We constrain the weights to sum to one for each i . For the sampled $\mathbf{x}$'s we denoted the weights as

$$w_{ij} = \frac{W\{(\mathbf{x}_i - \mathbf{x}_j)/\mathbf{b}\}}{\sum_{j=1}^n W\{(\mathbf{x}_i - \mathbf{x}_j)/\mathbf{b}\}}. \quad (5.2.1)$$

The choice of bandwidths is critical in smoothing. A bias-variance trade-off exists in determining the size of the bandwidths. When $\Sigma = \sigma^2 \mathbf{I}$ it is easy to show that

$$\begin{aligned} \text{as } \mathbf{b} \downarrow \mathbf{0} \quad & \text{bias}(\hat{\mu}_i) \rightarrow 0 \text{ and } \text{Var}(\hat{\mu}_i) \rightarrow \sigma^2, \text{ and} \\ \text{as } \mathbf{b} \uparrow \infty \quad & \text{bias}(\hat{\mu}_i) \rightarrow -(\mu_i - \bar{\mu}) \text{ and } \text{Var}(\hat{\mu}_i) \rightarrow \frac{\sigma^2}{n}. \end{aligned}$$

The choice of bandwidths is often based on minimizing the average (weighted) squared error

$$ASE(\mathbf{b}) = \sum_i \frac{(\hat{\mu}_i - \mu_i)^2}{n \text{Var}(Y_i)},$$

or $MSE(\mathbf{b}) = E\{ASE(\mathbf{b})\}$, and is a compromise between reducing bias and variance. Of course we do not know ASE or MSE ; however, we can estimate these quantities, and choose bandwidths that minimize these estimates. These ideas are closely linked with aspects of variable selection, which are discussed in the next section.

5.3. Variable selection

In this section we consider analytic methods for variable selection. We will use these methods to determine exactly which covariates to include in our SMFM for type I catches. Estimates of total trawlable stock abundance are based on estimates of local trawlable density, so it is important that local density be estimated as precisely as possible. Recall that part of our model for local density involves a smooth function of covariates. The other part involves a random contribution from type II stock components. In our application, and in many other applications, the initial proposed model contains more covariates than necessary, some of which have little relationship with the variable of interest. If all of the covariates are included in the model then the estimated smooth function may tend to overfit the data and have reduced predictive capabilities; hence, it is important to exclude unnecessary covariates from the model and restrict our attention only to important covariates. We will use our model to predict fish abundance at many tow stations, so this is important for us.

For some covariates it is easy to decide whether to include them or not. In the last chapter we concluded that light conditions and sea states during tows could be excluded from our model. Other covariates were harder to judge, largely due to confounding among the covariates; that is, on their own some covariates appear important, but this may only be through their association with other covariates. Several procedures are available for this type of problem when dealing with parametric models, but in our case we have to propose new procedures because of the nonparametric modelling approach we have taken. In this section we propose a nonparametric variable selection procedure to determine which covariates to include in our model. The algorithm is nonparametric in that the mean is assumed to be only a smooth function of covariates. A brief review of parametric variable selection theory is first presented, primarily for the normal linear regression model. This review provides some of the motivation for the nonparametric algorithm we propose.

5.3.1. Brief review of parametric variable selection

Most of the statistical literature on variable selection involves parametric linear models with additive errors:

$$Y = \mathbf{x}'\beta + \varepsilon.$$

In this model $\mu_{\mathbf{x}} = \mathbf{x}'\beta$, $\mathbf{x}$ is a vector of known regression covariates, β is a vector of regression parameters, and ε is the error term with $Var(\varepsilon) = \sigma^2$ constant. Thompson (1978) reviewed this situation. A common approach is to select covariates that minimize a measure of mean square prediction error:

$$MSPE = \frac{1}{n} \sum_{i=1}^n E \{Y_i^* - \hat{\mu}(\mathbf{x}_i)\}^2 \quad (5.3.1)$$

$$= \sigma^2 + \frac{1}{n} \sum_{i=1}^n E \{\hat{\mu}(\mathbf{x}_i) - \mu_i\}^2, \quad (5.3.2)$$

where Y^* is a new observation at $\mathbf{x}$. The idea in minimizing this quantity is to select the covariates that make $\hat{\mu}(\mathbf{x}_i)$ as close as possible to μ_i , on average. If we used $MSE = \sum_i E \{Y_i - \hat{\mu}(\mathbf{x}_i)\}^2 / n$ for variable selection then we would select all the covariates because adding a covariate to a regression analysis always results in a lower MSE . We are not interested in covariates that result in only small decreases in MSE ; therefore, selection criteria are required that penalize models with too many covariates. Many criteria have been proposed. Three common criteria, and the ones we consider later, are:

1. Cross validation (*CV*). This criterion is based on deleting an observation and using the remaining observations to predict the deleted one. The criterion is

$$CV = \frac{\sum_{i=1}^n (y_i - \hat{y}_{-i})^2}{\sigma^2 n},$$

where $\hat{y}_{-i}$ is the predicted value of Y_i based on a regression without y_i . The σ^2 is included so that CV is comparable with the other criteria, although its value is not required to minimize CV .

2. Generalized cross validation (GCV),

$$GCV = \frac{SSE_p}{\sigma^2 n(1 - p/n)^2},$$

where $SSE_p = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ is the error sum of squares based on a model with p covariates; that is, $\mathbf{x}$ is $p \times 1$. The σ^2 term is included for comparison purposes, its value is not required to minimize GCV .

3. C_p ,

$$C_p = \frac{SSE_p}{n\hat{\sigma}_f^2} + \frac{2p}{n},$$

where $\hat{\sigma}_f^2$ is an estimate of σ^2 based on the full model, with all covariates.

These criteria are all asymptotically equivalent, with expectations also asymptotically equivalent to $MSEP$. It is easy to show this for the CV criterion. Its expectation can be expressed as

$$E(CV) = 1 + \frac{\sum_{i=1}^n E\left\{\hat{Y}_{-i} - \mu_i\right\}^2}{n\sigma^2}.$$

If $\hat{Y}_{-i} \approx \hat{Y}_i$ then $E(CV)$ is asymptotically equivalent to $MSEP$ in (5.5.1). To show this we use

$$\hat{Y}_i - \hat{Y}_{-i} = \frac{h_{ii}(Y_i - \hat{Y}_i)}{1 - h_{ii}}.$$

where h_{ii} are the diagonal elements of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. If the correct covariates are selected then $\hat{Y}_i - \hat{Y}_{-i}$ has mean zero, and variance

$$\text{Var}(\hat{Y}_i - \hat{Y}_{-i}) = \frac{h_{ii}^2 \sigma^2}{1 - h_{ii}}.$$

If the covariates are such that $h_{ii} \rightarrow 0$ as $n \rightarrow \infty$ then $\hat{Y}_{-i} \approx \hat{Y}_i$, and $E(CV) \approx \text{MSEP}/\sigma^2$.

The asymptotic equivalence of CV and GCV can be shown using

$$CV = \frac{\sum_i (y_i - \hat{\mu}_{-i})^2}{n\sigma^2} = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{n\sigma^2(1 - h_{ii})^2}.$$

If we can approximate $h_{ii} \approx \text{trace}(\mathbf{H})/n = p/n$ then $CV \approx GCV$. If $\hat{\sigma}_f^2$ is consistent for σ^2 then the equivalence of GCV and GC_p can be shown using the approximation $1/(1 - x)^2 \approx 1 + 2x$, when x is small.

5.3.2. Nonparametric variable selection

In this section we extend the methods discussed in the previous section to nonparametric models with more arbitrary error structures, which includes the SMFM we use for type I catches. While nonparametric regression has recently become a popular technique, especially in exploratory analyses, the issue of variable selection has received little attention. As far as we know there is only one statistical publication concerning fully nonparametric variable selection for continuous covariates. Zhang (1991) considered variable selection using a simple kernel smoother. He assumed that covariates were pre-ranked in terms of importance, and dealt with the problem of determining the number of covariates to include in a model. Zhang (1991) used the same bandwidth b for all covariates. His algorithm is to select b that minimizes some measure of prediction error $PE(b, p)$ for a fixed p , such as $CV(b, p)$, then select p that minimizes $PE\{b(p), p\}$. Using the same bandwidth for all covariates does not seem like a good idea in our situation, so we do not pursue

Zhang's (1991) approach. Instead we simply modify the procedures discussed in **Section 5.3.1** to the nonparametric case.

The basic framework explored here is to select covariates that are useful for prediction purposes. The selected covariates are those that minimize some measure of PE over all possible subsets of covariates. All of the PE measures in **Section 5.3.1** can easily be adapted for a GLIM by substituting deviance (see McCullagh and Nelder, 1989) for SSE_p . The deviance is commonly used to measure goodness of fit in exponential family regression models, and is an important quantity in variable selection. Our variable selection procedure is essentially the same as in the previous section, and that is to select the subset of covariates that minimizes deviance, but with a penalty for the number of covariates in the model. First we show how to compute the deviance, and its expectation. Following this we present the generalized variable selection criteria.

The deviance is:

$$D(\hat{\mu}, y) = 2\{l(\mu_{\max}, y) - l(\hat{\mu}, y)\} = \sum_i D(\hat{\mu}_i, y_i),$$

where $y' = [y_1, \dots, y_n]$, $\mu' = [\mu_1, \dots, \mu_n]$ and $\mu_{\max}$ maximizes $l(\mu, y)$ over all μ . For our model $\mu_{\max} = y$. For the normal linear regression model the deviance is SSE/σ^2 . The expectation of $D(\hat{\mu}, Y)$ plays an important role in interpreting $D(\hat{\mu}, y)$ in this dissertation, and is commonly referred to as the error degrees of freedom (see Hastie and Tibshirani, 1990). For example, in the normal linear regression model $E\{D(\hat{\mu}, Y)\} = n - p$, where p is usually the number of covariates in the regression. This is the common definition of error degrees of freedom for this model. In general let $\epsilon = Y - \mu$ denote the vector of residuals, and let $\Sigma = \text{Diag}(\sigma_1^2, \dots, \sigma_n^2)$. The deviance can be written as

$$D(\hat{\mu}, Y) = \hat{\epsilon}' \Sigma^{-1} \hat{\epsilon} + o_p(\|\hat{\epsilon}\|),$$

where $\hat{\varepsilon} = (\mathbf{I} - \mathbf{W})\mathbf{Y}$. Using this approximation we can show that

$$E \{D(\hat{\boldsymbol{\mu}}, \mathbf{Y})\} \approx n - \text{tr}(2\mathbf{W} - \mathbf{W}'\Sigma^{-1}\mathbf{W}\Sigma) + \boldsymbol{\mu}'(\mathbf{I} - \mathbf{W})'\Sigma(\mathbf{I} - \mathbf{W})\boldsymbol{\mu},$$

where $\mathbf{W} = \{w_{ij}\}$ is the matrix of smoothing weights defined by (5.2.1). Ignoring the bias term,

$$df^{err} = n - \text{tr}(2\mathbf{W} - \mathbf{W}'\Sigma^{-1}\mathbf{W}\Sigma).$$

The generalized variable selection *PE* criteria we use are:

$$CV = \sum_i D(\hat{\mu}_{-i}, y_i)/n, \quad (5.3.3)$$

$$GCV = \frac{\sum_i D(\hat{\mu}_i, y_i)/n}{\{1 - \text{tr}(\mathbf{W})/n\}^2}, \text{ and} \quad (5.3.4)$$

$$GC_p = \sum_i D(\hat{\mu}_i, y_i)/n + 2\text{tr}(\mathbf{W})/n. \quad (5.3.5)$$

The deviance may involve overdispersion parameters, as in our case. We will estimate these separately, and use the same estimates in all our variable selection procedures. This is similar to the suggestion by Fitzmaurice (1997). These criteria are analogous to those for the normal linear regression model. For example, the correction term in GC_p can be obtained from a comparison of df^{err} and the generalized mean square prediction error (*MSPE*):

$$E \left\{ \sum_i D(\hat{\mu}_i, Y_i^*)/n \right\} = 1 + \text{tr}(\mathbf{W}'\Sigma^{-1}\mathbf{W}\Sigma)/n + O(n^{-1}).$$

Here Y_i^* denotes a new observation with covariates $\mathbf{x}_i$. Adding $2\text{tr}(\mathbf{W})/n$ makes

$$E(GC_p) \approx E \{D(\hat{\mu}, \mathbf{y}^{new})/n\} \approx MSPE.$$

This is the same rationale for the correction term in C_p presented in **Section 5.3.1**.

We also use these criteria for selecting bandwidths; that is, choose bandwidths that minimize one of these criteria. This is common practice (for example, see Hastie and Tibshirani, 1990). In fact, Li (1987) has shown for the additive homoscedastic errors case that these criteria are asymptotically optimal in the sense that

$$\frac{ASE(\mathbf{b}_{cv})}{\inf_{\mathbf{b} \in \mathbf{B}} ASE(\mathbf{b})} \rightarrow 1, \text{ in probability,}$$

where $\mathbf{b}_{cv}$ is the vector of bandwidths that minimize CV . The same optimality property extends to GC_p and GCV because of their asymptotic equivalence with CV .

Nonparametric variable selection algorithm

In this section we outline the nonparametric variable selection algorithm we use with the trawl survey data. A complication arises because we use the same criterion to select bandwidths and variables. Because bandwidths are selected to minimize a PE measure, the model with all covariates always has the lowest value of the PE measure. Hence, the algorithm is to

select the covariates that **minimize, or nearly minimize**, a PE measure.

The algorithm includes the option “nearly minimize” so that covariates that result in relatively small decreases in the criteria are not selected. By **nearly minimize** we mean covariate subsets with a $PE < \min PE \times \delta$, where $\delta > 1$. The $\min PE$ is for all covariates. This is because the same PE criterion is used to select bandwidths; that is, bandwidths

are also selected so that they minimize the PE criterion. It is impossible to get a lower PE criterion for a subset of covariates, because this solution would be found when minimizing the PE criterion for all covariates. This is different than in parametric variable selection, where the PE criterion is usually not a minimum for the model with all covariates. The difference is because in a parametric model adding a covariate increases the “effective” degrees of freedom by one, whereas in nonparametric regression the “effective” degrees of freedom can either increase or decrease when a covariate is added, depending on what bandwidths are used.

In linear regression, an F-test for added covariates takes the same form as our variable selection procedure. In this case one concludes a p_o dimensional subset of covariates is important if the SSE is less than the $\min SSE \times \delta$, where

$$\delta = 1 + \frac{p_o}{n - p} F_{p_o, n-p},$$

and $F_{p_o, n-p}$ is an F value for some appropriate α . Note that $\min SSE$ is obtained using all the p covariates. If $n = 150$, $p = 5$, and $p_o = 2$ then $\delta = 1.013$ and 1.025 , for $\alpha = 0.5$ and 0.05 . These values for n , p , and p_o are similar to those we shall consider, although we are not suggesting that the above form of δ is appropriate for our algorithm. Even in linear regression it is difficult to determine what value for α is appropriate, or even if the F distribution is appropriate; hence, the value for δ is somewhat arbitrary and also depends on a subjective choice for α . We used $\delta = 1.05$ in our analyses. This is also an arbitrary choice, but one that seems to give reasonable results.

The algorithm is simple to implement (on a fast computer). The algorithm is:

1. find bandwidths that minimize PE for all of the 2^p variable subsets, where p is the total number of covariates,

2. find the smallest subset of covariates such that the minimum PE for that subset is less than the minimum $PE \times \delta$ for all variables.

The rationale for the second step is that important covariates, when removed, should lead to a large increase in PE , whereas unimportant covariates should lead to only small increases in PE when removed.

We also analyze per observation deviances, or the squared deviance residuals (see McCullagh and Nelder, 1989), for selected model fits to assist with variable and bandwidth selection. Typically (but not always) a smoother with all covariates and min PE bandwidths explains more deviance than a smoother with some subset of covariates, and the corresponding min PE bandwidths. If the decrease in total deviance is greater than expected, using df^{err} , then we conclude that some of the deleted covariates are important. Conversely, if the decrease in deviance is less than the change in error degrees of freedom then we conclude that the deleted covariate subset is not important. This is somewhat analogous to the F-test for added covariates in the normal linear regression model, except that in the nonparametric case the “SSE” does not always have to increase when covariates are deleted. This depends on what bandwidths are used. Another use of examining the per observation change in deviances is to assess whether a covariate subset “explains” variation for much of the data, or only a small number of observations. If the latter is the case then robustness is a concern. We return to this point in Section 5.4.4. Finally, our variable selection results are sensitive to the choice of PE measure, and the per observation change in deviances are useful for deciding which measure is more useful for our data.

5.4. Analysis of the 3NO trawl survey data, I

In this section we apply the nonparametric regression and variable selection procedures to the 3NO trawl survey data. We use the Negative Binomial (NB) distribution to stochastically model trawl catches. The NB distribution is a flexible one that is appropriate for overdispersed count data. This distribution has been suggested for modelling trawl catches (see Gunderson, 1993; p17). Other distributions have also been suggested (e.g. delta-lognormal); however, the NB is simple and, we find, appropriate for our data. The mean of the NB distribution is modelled as a smooth function of spatial location (SP) and tow depth (DP), and possibly tow temperature (TP), tow time (TM), and wind speed (WD) during tows. That is, if m is the NB mean, then $m = m(SP, DP, TP, TM, WD) = m(\mathbf{x})$. The NB variance is $m(1 + m/k)$, where k is a Poisson overdispersion parameter. The discrete probability density function for this distribution is

$$\Pr(Y = y; m, k) = \frac{\Gamma(k + y)}{\Gamma(k)\Gamma(y + 1)} \left(\frac{m}{m + k} \right)^y \left(\frac{k}{m + k} \right)^k. \quad (5.4.1)$$

The distributional assumption plays an important role in bandwidth selection. Ignoring overdispersion can lead to small bandwidths and under-smoothing. We consider the estimation of k carefully, but first we must standardize the covariates. This involves two steps.

5.4.1. Standardizing covariates

In this section we address the problem of confounding amongst some covariates. Some of the covariates we consider are highly correlated; for example, time between tows and location of tows. This causes problems in bandwidth and variable selection. For example, if local abundance is only spatially correlated, then a temporal smoother can account for much of the variation in trawl catches because tows that are close together in space are

always close together in time. We considered this in some detail in **Chapter 4**. Confounding results in erratic and unreliable variable selection using the algorithm we outlined in the previous section. Our solution to this problem involves locally orthogonalizing covariates (see below), but first we detail some preliminary covariate standardizations we use.

Recall from **Section 5.2** that the nonparametric regression simply involves locally weighted averages. The estimate of $m(\mathbf{x})$ is

$$\hat{m}(\mathbf{x}) = \frac{\sum_{i=1}^n W \left\{ \frac{(\mathbf{x}_i - \mathbf{x})}{\mathbf{b}} \right\} y_i}{\sum_{i=1}^n W \left\{ \frac{(\mathbf{x}_i - \mathbf{x})}{\mathbf{b}} \right\}}.$$

The $W(\cdot)$ weights define distances between $\mathbf{x}$ and the $\mathbf{x}_i$'s. The size of the weights are determined by the bandwidths, $\mathbf{b}$. In order to make bandwidths comparable between variables we standardize covariates each year so that $\sum_i x_i = 0$ and $\sum_i x_i^2 = 1$. This is the first step.

Remark 2. *We see no reason why the weights should not be symmetric in latitude and longitude, which are the two components of spatial location, so we use the same bandwidth for both these covariates.*

The second step is critical, and involves removing the confounding among some of the covariates. In **Chapter 4** we concluded spatial location and tow depth were the most useful covariates for prediction purposes, because we know these covariates for the entire 3NO region. We only know tow temperatures and wind speeds at the sampled sites. Tow time is not a predictive covariate, but can be used to assess the importance of fish movements. Unfortunately tow temperature is correlated with spatial location and tow depth, wind speed is correlated with spatial location and time between tows, and tow time is correlated with spatial location. This confounding makes variable selection difficult; for

example, it is usually the case that the $\min CV$ for a smoother based on spatial location is about the same as the $\min CV$ for a smoother based on time between tows.

We assess the dependencies in trawl catches with temperature, time, and wind only after the dependencies in these covariates with spatial location and tow depth have been removed. For this purpose we also use kernel smoothing. The analogue to this approach in linear regression is orthogonalizing covariates. In this situation a decision must also be made about which covariates to transform. If x_1 and x_2 are two regression covariates then one way to make them orthogonal (uncorrelated) is to regress x_2 on x_1 and use the x_2 residuals (e_{x_2}) instead of x_2 in the regression with the y response; that is, regress y on x_1 and e_{x_2} . We do the same thing here, except that we use nonparametric regression instead of linear regression.

The decision to locally orthogonalize temperature, time, and wind is somewhat arbitrary. It is based on practicality grounds in that we want to use the model to predict local abundance throughout 3NO, and we only know spatial location and tow depth at non-sampled trawl stations. Note that this decision is not completely arbitrary. For example, if tow temperature is really important in determining local abundance then we could predict tow temperature at nonsampled tow stations using the spatial and depth persistence of water temperatures, and then use the predicted temperatures to predict local fish abundance at nonsampled tow stations from our observations of catches at sampled tow stations. This is possible because water temperature does not change at specific depths over fairly broad spatial scales. However, it is unlikely such a model would produce better predictions at nonsampled tow stations, compared to a model with just spatial location and tow depth. This is because of the nonparametric nature of our modelling. If $E(Y_x) = m_1(SP, TP)$ and $TP = m_2(SP, DP)$, then we can express $E(Y_x)$ as $m_3(SP, DP) = m_1\{SP, m_2(SP, DP)\}$, in which case we can estimate m_3 directly using lle's based on SP and DP .

We perform three nonparametric regressions to remove confounding among our covariates:

1. Regress tow time on spatial location.
2. Regress tow temperature on spatial location and tow depth.
3. Regress wind speed on spatial location and tow time.

The regressions are easy to implement. All we have to do is decide what bandwidths to use, then compute the kernel regression weights w_{ij} for all observations. If $\mathbf{W} = \{w_{ij}\}$ then the estimated responses are given by $\hat{\mathbf{y}} = \mathbf{W}\mathbf{y}$, where $\mathbf{y}$ is the $n \times 1$ vector of observations which, in this case, can either be tow time, temperature, or wind speed. We choose bandwidths that minimize the CV criterion,

$$CV = \frac{1}{n} \sum_i (y_i - \hat{y}_{-i})^2.$$

We feel the errors in the Y 's are additive and homoscedastic, which leads to the use of this form of CV . Bandwidths are chosen separately each year, and are presented in Table 5.1a. The df^{err} 's associated with these bandwidths are usually about 50% of the sample size each year. For regressions 1-3 above the average df^{err}/n 's across years is 0.46, 0.55, and 0.52, which is a reasonable amount of smoothing. The correlation coefficients, or the proportion of variance explained by the regression, are presented in Table 5.1b. The residual variation from the three regressions is small, especially for the time and temperature regressions. The correlation coefficients are large, which indicates the degree of confounding in the covariates. In the trawl catch regression analyses throughout the rest of this chapter we will use spatial location, tow depth, and the covariate regression residuals for temperature, time, and wind speed. These residual covariates should be useful for explaining potential variation in trawl catches that is not related to the spatial location or depth of tows.

Table 5.1a. Bandwidths that minimize the cross validation criterion for the covariate regression analyses. The bandwidths are for *SP*-spatial location, *DP*-depth, and time *TM*.

Year	Time	Temperature		Wind Speed	
	b_{SP}	b_{SP}	b_{DP}	b_{SP}	b_{TM}
84	0.012	0.017	0.005	0.016	0.005
85	0.003	0.040	0.018	8.175	0.018
86	0.015	0.013	0.005	7.939	0.005
87	0.010	0.012	0.030	0.008	0.030
88	0.011	0.011	0.045	11.208	0.045
89	0.001	0.011	0.054	0.109	0.054
90	0.010	0.019	0.037	0.021	0.037
91	0.007	0.013	0.008	3.960	0.008
92	0.007	0.014	0.019	9.866	0.019
93	0.012	0.013	0.031	7.586	0.031
94	0.012	0.027	0.007	0.006	0.007

Remark 3. For convenience we refer to these residuals simply as temperature, time, and wind speed.

5.4.2. Variable smoothing

Preliminary investigations have revealed that the smoothness assumptions required for the continental shelf region in 3NO (see Figure 1.4) are different than for the slope regions. Water depths and temperatures are relatively constant on the shelf, whereas on the slopes these variables can change rapidly over short spatial scales. This is why strata for the slopes are relatively small. This suggests that variable bandwidths are useful for modelling these data. A variety of procedures exist for selecting variable bandwidths; for examples, see Müller and Stadtmüller (1987) and Brockmann *et. al.* (1993). We choose a simple

Table 5.1b. Covariate regression correlation coefficients.

Year	Time	Temperature	Wind Speed
84	0.98	0.98	0.94
85	1.00	0.91	0.88
86	0.85	0.99	0.89
87	0.98	0.97	0.94
88	0.99	0.97	0.90
89	1.00	0.96	0.97
90	0.98	0.92	0.88
91	1.00	0.97	0.92
92	1.00	0.97	0.89
93	0.99	0.98	0.85
94	0.99	0.96	0.97

procedure in which we split our data into two groups representing the slope and shelf regions, and analyze the data separately. If the bandwidths are the same in both regions then splitting the data will result in fewer error degrees of freedom (i.e. df^{err} in Section 5.3.2). This is because if observations are ordered in y by region, then to achieve the splitting the smoothing matrix W is replaced by a block diagonal matrix corresponding to the two regions. The rows of this block diagonal matrix must still sum to one so its trace will be greater than $trace(W)$, resulting in fewer df^{err} 's. The advantage of the smoothing stratification is potentially lower deviances.

5.4.3. Choosing the Negative Binomial k parameter

The next step in our analysis is to estimate the NB overdispersion parameter k for use in the subset selection analyses. Similar to Fitzmaurice (1997), k is estimated only once for the full model, and this estimate is then used in computing PE measures and deviances in all submodels examined. We test the sensitivity of our results to the value of k we

choose by repeating our variable selection procedures using a different value of k .

First consider the homoscedastic additive errors nonparametric regression problem,

$$Y = \mu(\mathbf{x}) + \varepsilon,$$

where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. In this case estimating σ^2 is straightforward. The common estimate is just SSE divided by the appropriate df^{err} 's. Our case is not so simple. The k parameter must be estimated jointly with $m(\mathbf{x})$, which means that we must select bandwidths and estimate k simultaneously. One option for this problem is an iterative scheme in which bandwidths are selected based on a fixed value of k , then k estimated using maximum likelihood based on the $\hat{m}(\mathbf{x})$'s computed using these bandwidths, and the process iterated until convergence. Alternatively we can find the joint solution of the estimating equations associated with minimizing CV to get $\hat{m}(\mathbf{x})$, and the NB log-likelihood to get $\hat{k}$. In this approach two objective functions are used, one for finding bandwidths (CV) and one for finding k (log-likelihood). Our experience with both these approaches is that they produce two solutions, one with small bandwidths, few df^{err} 's, and a large value for $\hat{k}$, and the other with large bandwidths, many df^{err} 's, and a small value for $\hat{k}$. Neither of these solutions seems appropriate. If $\hat{k}$ is very small then we tend to under-fit catches, while if $\hat{k}$ is very large we tend to over-fit catches. The problem is that the likelihoods for these two types of solutions are similar.

The solution we propose for this problem is to maximize a type of predictive likelihood for both bandwidths and k . For fixed bandwidths we estimate $m(\mathbf{x}_i)$ without y_i using kernel smoothing, and denote this estimate as $\hat{m}_{-i}(\mathbf{x}_i)$. This estimate is a function of the

bandwidths $\mathbf{b}$. We estimate k as that value that, jointly with $\mathbf{b}$, minimizes

$$l_n(k, \mathbf{b}) = \sum_{i=1}^n \left[\Psi(k) - \Psi(k + y_i) - y_i \log \left\{ \frac{\hat{m}_{-i}(\mathbf{x}_i)}{k + \hat{m}_{-i}(\mathbf{x}_i)} \right\} - k \log \left\{ \frac{k}{k + \hat{m}_{-i}(\mathbf{x}_i)} \right\} \right] \quad (5.4.2)$$

where $\Psi(x) = \log \Gamma(x)$. Apart from a term depending only on the y_i 's, $l_n(k, \mathbf{b})$ is just the negative log-likelihood based on (5.4.1). If we used $\hat{m}(\mathbf{x}_i)$ instead of $\hat{m}_{-i}(\mathbf{x}_i)$ then minimizing $l_n(k, \mathbf{b})$ would yield small bandwidths which fit the data exactly for any k . In this case $\hat{k} = \infty$. If we used the predictive likelihood procedure in a linear regression model with p covariates and additive $N(0, \sigma^2)$ errors then the mple (maximum predictive likelihood estimate) for σ^2 has expectation $n\sigma^2/(n - p)$. For this model the mple overestimates error variance, whereas the mle underestimates error variance. This is easy to show using the results in **Section 5.3.1**. We suspect the mple of k has a small negative bias, whereas the mle of k has a very large positive bias, as indicated above. For this reason we chose to estimate k using (5.4.2).

We performed a simple simulation experiment to check the mple bias for k . We generated random NB data using a simple model and three values for k (e.g. $k = 1, 2$ and 3). We considered sample sizes of $n = 25, 50$, and 100 . The model we considered is $m(x) = 5 + 3\sin(x)$; hence, x is one-dimensional in our simulation. We generated NB responses for this model by

1. generating n random X 's from a Uniform $(0, 20)$ distribution,
2. computing $m(x)$ for the random X 's in 1, then
3. generate NB responses using the $m(X)$'s in 2.

For each simulation we computed the bandwidth and k that maximized (5.4.2). The simulated biases from mple estimates of k are presented in Table 5.2a. These biases are

negative, which is consistent with our previous arguments. We will show in Chapter 6 that our inferences are robust for the magnitude of biases shown in Table 5.2a , so we feel

Table 5.2a. Simulated NB mple bias for k .

k	n		
	25	50	100
1	-0.05	-0.11	-0.11
2	-0.11	-0.28	-0.25
3	-0.41	-0.50	-0.38

that equation (5.4.2) provides reasonable estimation of k for the trawl survey data.

Annual estimates of k are presented in Table 5.2b. For variable selection we assume that the value of k is the same in all years. The average estimated k in Table 5.2b for the shelf region is 0.99, while for the slope region this value is 0.55. To get a combined estimate of k we use the bandwidths in Table 5.2b to compute $\hat{m}_{-i}(\mathbf{x}_i)$ for all years and regions, and use these values to estimate a single k . The estimate is $\hat{k} = 0.65$. This is the value we use in subsequent variable selection analyses. We will test the sensitivity of our results to the choice of k by repeating our analyses with $k = 5$.

Note the large variation in the size of bandwidths for DP , TP , TM , and WD in Table 5.2b. There are several reasons for this. One reason is robustness. We will show later that small bandwidths for TP , TM , and WD may only provide better fits to a small number of catches. Another reason involves the variability in a covariate. If a covariate does not vary much then whether its bandwidth is large or small does not matter, because the neighborhood for this covariate includes nearly all observations no matter what the bandwidth is. This is why the DP bandwidths change widely for the shelf region from year to year. Finally, and perhaps most importantly, is that finding good bandwidths is difficult even with good data (see bandwidth simulations in Hastie and Tibshirani,

Table 5.2b. Estimates of the NB scale parameter k and bandwidths b that maximize a predictive likelihood function.

3NO shelf region						
Year	b_{SP}	b_{DP}	b_{TP}	b_{TM}	b_{WD}	$\hat{k}$
84	0.105	75.36	0.037	0.067	81.30	0.67
85	0.028	0.027	0.088	49.07	0.109	0.78
86	0.020	0.043	0.086	0.124	67.46	0.61
87	0.019	25.71	0.064	15.03	0.114	0.64
88	0.032	0.139	0.063	0.033	101.3	1.16
89	0.031	0.029	48.23	0.155	0.067	0.80
90	0.021	0.132	0.122	0.022	0.222	1.80
91	0.025	0.096	54.59	81.87	76.86	1.27
92	0.029	68.82	60.29	71.54	79.48	1.34
93	0.039	73.45	69.15	0.057	10.16	1.05
94	0.042	43.30	47.37	0.163	0.059	0.74

3NO slope region						
Year	b_{SP}	b_{DP}	b_{TP}	b_{TM}	b_{WD}	$\hat{k}$
84	0.111	0.036	0.406	52.99	0.064	0.29
85	0.023	0.043	99.94	0.107	0.184	0.61
86	0.033	0.087	0.152	0.018	0.160	0.55
87	0.027	0.060	0.274	0.124	103.6	0.39
88	0.044	0.059	34.53	81.72	0.115	0.57
89	0.051	0.038	91.81	0.186	89.73	0.99
90	0.113	0.029	67.16	133.0	59.76	1.16
91	0.042	0.031	0.102	89.07	0.097	0.39
92	0.023	0.029	0.068	75.53	0.279	0.44
93	0.065	0.015	0.199	0.134	62.89	0.25
94	0.087	0.010	68.70	84.72	78.31	0.40

Note: SP -spatial location, DP -depth, TM -time, TP -temperature, WD -wind speed.

1990, p 50). Even the bandwidths we finally select after robust considerations (see Table 5.19) still vary considerably from year to year. A potential solution to this problem is to constrain bandwidths to vary smoothly, but we do not pursue this here. Fortunately our inferences about total trawlable stock size are relatively insensitive to the choice of bandwidth (see Chapter 6).

5.4.4. Variable selection results

Having standardized some of the covariates and determined a value for k we can now proceed with variable selection. We only consider temperature, time, and wind speed for selection; that is, we assume that $m(\mathbf{x})$ is a function of spatial location and tow depth, and possibly some of the other covariates. This reduces the number of covariate subsets to analyze from 2^5 to 2^3 . In what follows we refer to the bandwidths that minimize a PE criterion as PE $b_{\min}$'s; for example, the bandwidths that minimize the CV criterion are referred to as the CV $b_{\min}$'s.

The bandwidths that minimize the CV , GCV , and GC_p criteria are presented in Tables 5.3a-c in the Appendix (Section 5.11). The near optimal bandwidths are based on the smallest subset of covariates whose estimated PE is close to the minimum PE for all covariates; that is, $PE \leq 1.05 \times \min PE$. In Table 5.3a tow temperature and wind speed have little predictive power in seven of eleven years, for both the shelf and slope regions. The time between tows is important in only three years. Using GCV , only spatial location and tow depth (see Table 5.3b) seem important. In some years minimizing GCV results in very small bandwidths and over-fitting; that is, almost no error degrees of freedom remain. These fits do not seem trustworthy for prediction purposes. Wind speed appears to have no predictive power using GC_p , and time between tows and tow temperature are important in only two years.

The CV results suggest that in most years spatial location and tow depth are the

most important predictive covariates, but that in some years the other covariates have some predictive capabilities as well. To further investigate this we examined the deviance residuals from some of the models in Tables 5.3a and 5.3c. To assess the improvement in fit resulting from $CV\ b_{\min}$'s with temperature, time, and wind speed we plot, in Figures 5.1a,b, the differences in the per observation deviance for the model with all covariates and the model with just spatial location and tow depth. The change in deviance and df^{err} 's are listed at the top of each panel. The change in df^{err} 's is that deviance expected to be explained by decreasing neighborhood sizes (bandwidths) alone. We expect that if tow temperature, time, or wind speed are important in determining trawl catches then the change in deviance should exceed the change in df^{err} 's. This occurs in five of eleven years, and most notably in 1993 where the difference in deviance is almost double the difference in df^{err} 's. Most of the differences are related to only a small number of tows. This suggests that time, temperature, and wind speed may be important for only a few observations. We return to this idea in the next section.

The differences in the per observation deviance for GC_p bandwidths are very small and are not presented. Changes in total deviance are usually small when temperature, time, and wind speed are not included in the model. These changes are approximately the same as the change in df^{err} 's. The GC_p results suggest that tow temperature, time, and wind speed are not important in determining mean trawl catches. The $GC_p\ b_{\min}$'s for spatial location and tow depth result in more df^{err} 's than the $CV\ b_{\min}$'s in every year except 1993. The total df^{err} 's for all years is 1340 for the $GC_p\ b_{\min}$'s, and 739 for the $CV\ b_{\min}$'s. This agrees with the general consensus in the smoothing literature that CV bandwidths tend to be too small, and consequently lead to too few df^{err} 's for inferences.

The above analyses were repeated with $k = 5$, and the results are presented in Table 5.4a-c. Overall the results suggest that the variable selection procedures (not bandwidth selection) are robust to reasonable values of k , and our conclusion that spatial location

and tow depth are the most important covariates remains valid when $k = 0.65$ or 5 . Usually the CV $b_{\min}$'s are similar for both values of k , and the amount of smoothing is about the same for both choices of k , i.e. the difference in df^{err} 's for all years is only 2. The CV estimates of prediction error are sensitive to the choice of k , and increase more than two-fold in most years. The GCV results are also similar, and most dissimilar in years when under-smoothing occurs. The GC_p results when $k = 0.65$ or $k = 5$ also suggest the same relative importance among covariates. The GC_p $b_{\min}$'s are considerably smaller when $k = 5$ than when $k = 0.65$, and the total df^{err} 's for all years changes from 1340 to 433. When $k = 5$ the GC_p criterion under-smooths relative to the CV criterion, which is opposite to that observed when $k = 0.65$.

Another aspect of robustness is the sensitivity of our conclusions to outlying observations. Data driven bandwidth selectors are very sensitive to outlying responses (see Boente *et. al.*, 1997). Also, the variables we select may be greatly influenced by a small number of observations; that is, a selected covariate may "explain" a relatively large amount of variation produced by a small number of observations. This seems to be the case for our CV results (see Figures 5.1a,b), but not so for GCV or GC_p . This could be either appropriate or inappropriate. It may be that most observations occur at about the same value of an important covariate (say X^*), and infrequently at different values of X^* . If the mean of the responses at the infrequent values of X^* are truly different from the mean of the responses at most X^* then it is appropriate that the infrequent observations have more influence in terms of selecting X^* . It could also be simply accidental that the means of the few responses at the infrequent values of X^* are different, and that X^* is not really related to the response. In this case selecting X^* is not appropriate. For observational data such as ours it will be difficult to discern between these two scenarios without additional data, or information from other studies. Our sampling design is random, so if there are infrequent levels of X^* in the sample then these levels are also likely to be infre-

quent in the population. This means that for most of the population X^* has a constant effect, and not including this covariate in our model should have little effect on inferences about total stock size. Hence, we do not want to select variables because they explain variability for only a small number of observations. In the following sections we propose robust local likelihood estimation procedures and robust variable selection procedures to investigate whether this is a problem in our analysis. The robust procedures are also used to assist in identifying type II catches.

5.5. Robust local likelihood estimation

In this section we develop a robust smoother for NB data. We use robust M-estimation for this purpose. The robust M-estimator we employ essentially downweights observations that are far from the mean of “most” of the data. The basic strategy we use is to develop an efficient robust M-estimator for the mean of a NB distribution, and then apply this estimator locally for robust smoothing. Efficiency is an important issue for robust smoothing, because the effective sample size within smoothing neighborhoods can be small, and an inefficient estimator can lead to erratic (nonsmooth) results. Our preliminary experience (unreported results) with a simple weighted median smoother demonstrated this. The robust M-estimator is new, so we consider its efficiency with finite sample sizes through simulations. We do this to demonstrate that robust estimation with NB data is feasible for the effective sample sizes that are used in our smoothing procedures. These are often quite small. We also consider the model robustness of the robust NB M-estimator, including how it performs with outlier contaminated data, and how it performs with Poisson data. We use the robust smoother for two purposes. The first is to “robustify” variable selection, which we outline in **Section 5.5.4**. We wish to select variables that have predictive importance for the majority of the data, and not just a few observations. The second purpose is to identify tows with catches that are inconsistent with a smooth mean

function model. We call these type II catches.

It is frequently the case that a small percentage of observations appear to come from a different process (or model) than the majority. For example, with trawl survey data occasional large catches occur that are inconsistent with nearby catches. Such an occurrence is not unique to fisheries data; in fact, Hampel *et. al.* (1986, p28) speculated that “1-10% gross errors in routine data seem to be more the rule than the exception”. In our analyses this is particularly problematic because one drawback with nonparametric regression compared to parametric regression is increased sensitivity to outliers because of the increased reliance on the data for the basic shape of the regression surface. Nevertheless it is desirable that our regression inferences not be influenced too much by outlying observations.

The specific approach we take is to use outlier-robust M-estimation to develop robust local likelihood methods for Negative Binomial (NB) nonparametric regression. M-estimation has some precedence in terms of kernel smoothing (see Section 5.5.3), and is desirable in that it has a non-zero breakdown point (proportion of outliers in a sample that do not grossly affect an estimator; see Hampel *et. al.*, 1986 for more details) and yet still produces reasonably smooth fits. This section begins with a brief review of robust M-estimation theory. An asymptotically efficient robust M-estimator for the NB location parameter is developed, and a small simulation experiment is presented to compare the small sample efficiency of the robust estimator to the mle. These methods are then used to develop robust local likelihood estimates (lle's) of a smooth mean for NB data.

5.5.1. Basic theory of robust M-estimation

M-estimation is a general procedure that encompasses mle's, quasi-likelihood estimates, and many other methods. An M-estimator is merely the minimizer of an objective function, or the root of an estimating equation (Carroll and Ruppert, 1988); that is, the

M-estimator of a $p \times 1$ dimensional parameter θ (denoted as $\bar{\theta}$) is the solution to

$$\sum_{i=1}^n \psi(y_i, \bar{\theta}) = 0,$$

where ψ is a p dimensional function, but not necessarily the gradient of an objective function. The lle's considered in **Section 5.2** are examples of M-estimates. If Y has distribution F then an almost necessary condition for $\bar{\theta}$ to be consistent is that $\int \psi(Y, t) dF(Y) = 0$ must have a unique solution $t = \theta$ (see Carroll and Ruppert, 1988). This assumption usually holds if ψ is the score function for F . Carroll and Ruppert (1988) show heuristically that, under standard conditions,

$$n^{1/2}(\bar{\theta} - \theta) \xrightarrow{D} N[0, V_{\psi}(\theta)],$$

where

$$\begin{aligned} V_{\psi}(\theta) &= B_{\theta}^{-1} A_{\theta} B_{\theta}^{-1}, \\ B_{\theta} &= E\{\dot{\psi}(Y, \theta)\} = E\left\{ \frac{\partial \psi(Y, t)}{\partial t} \Big|_{t=\theta} \right\}, \text{ and} \\ A_{\theta} &= E\{\dot{\psi}(Y, \theta) \dot{\psi}'(Y, \theta)\}. \end{aligned}$$

Note that B_{θ} and A_{θ} are $p \times p$ matrices. A rigorous proof is given in Huber (1967), and further details are given in Clarke (1983).

Robust M-estimation is a very flexible methodology that can be readily applied to the estimation of location and overdispersion parameters for discrete distributions. Robustness is obtained by choosing ψ so that $\hat{\theta}$ is insensitive to outliers, as described below. An optimally robust M-estimator is insensitive to outliers, and yet maximally efficient for estimating θ among all estimators with the same robust properties. The sensitivity of $\hat{\theta}$

to outliers is typically measured using the influence function

$$IC_\psi(y, \theta) = B_\theta^{-1} \psi(y, \theta),$$

which measures the p dimensional sensitivity of $\hat{\theta}$ to small changes in y . Let $\|\mathbf{x}\|_{\mathbf{V}} = \sqrt{\mathbf{x}'\mathbf{V}^{-1}\mathbf{x}}$. The scalar measure of sensitivity used here is the *self-standardized sensitivity* (Carroll and Ruppert, 1988):

$$\begin{aligned} \gamma_s(\theta) &= \sup_{y_i} \|IC_\psi(y_i, \theta)\|_{\mathbf{V}_\psi(\theta)} \\ &= \sup_{y_i} \left\{ \psi'(y_i, \theta) \mathbf{A}_\theta^{-1}(\theta) \psi(y_i, \theta) \right\}^{1/2}. \end{aligned}$$

Optimal robustness is obtained by choosing ψ so that γ_s is bounded by $C < \infty$, and also so that $\text{tr}\{\mathbf{V}_\psi(\theta)\}$ is minimized with respect to ψ . Let $\dot{l}(y, \theta) = \partial l(y, \theta) / \partial \theta$ denote the score function, where $l(y, \theta) = \log F(y, \theta)$. The optimal ψ is

$$\psi^{opt}(y, \theta) = \left\{ \dot{l}(y, \theta) - \alpha \right\} \min \left[1, \frac{C}{\left\| \dot{l}(y, \theta) - \alpha \right\|_{\mathbf{A}_\theta}} \right], \quad (5.5.1)$$

where α satisfies $E\{\psi^{opt}(Y, \theta)\} = 0$. Hampel *et. al.* (1986; p 245) show that this choice minimizes $\text{tr}\{\mathbf{V}_\psi(\theta) \mathbf{V}_{\psi^{opt}}^{-1}(\theta)\}$ over all ψ satisfying the consistency and boundedness requirement. Hampel *et. al.* (1986; p 244) referred to the M-estimator based on $\gamma_s(\theta)$ as the optimal B_s -robust estimator. While these estimators may not be most efficient among the class of all possible consistent estimators, they are most efficient among the class of consistent estimators that have a *self-standardized sensitivity* bounded by C .

In the following examples and in the next section we compare the efficiency of robust estimators to mle's using the asymptotic relative efficiency, $ARE = \{V_\psi(\mu) \mathcal{I}(\mu)\}^{-1}$, where $\mathcal{I}(\mu)$ is the Fisher information for μ .

Example 1

We illustrate these methods using a simple example involving the estimation of μ for Normal data. Consider $Y \sim N(\mu, 1)$. The score function for μ is $\dot{l}(\mu) = \varepsilon = y - \mu$. The optimal B_s -robust estimating equation is

$$\psi(\mu) = \varepsilon \min \left[1, C\sqrt{A}/|\varepsilon| \right]. \quad (5.5.2)$$

The α term in (5.5.1) is zero because $E(\psi) = 0$. The standardizing constant A is obtained as the solution to

$$A = \int \psi^2 d\Phi(\varepsilon) = \chi_3^2(d) + d \{1 - \chi_1^2(d)\},$$

where $d = C^2 A$ (see pg. 234 in Hampel *et. al.*, 1986). This choice of A ensures that the *self-standardized gross error sensitivity* (γ) is less than C . For the normal model $\mathcal{I}(\mu) = 1$. To compute $V_\psi(\mu)$ we need

$$\int \psi' d\Phi(\varepsilon) = 2\Phi(d) - 1 = \chi_1^2(d) = B;$$

hence, $RE = B^2/A$. Results for a range of C 's are presented in Table 5.5. The robust

Table 5.5. Standardizing constant (A) and asymptotic relative efficiency (ARE) of the bounded-influence estimator for μ in a $N(\mu, 1)$ model. The *self-standardized sensitivity* is $\leq C$.

	C						
	1.25	1.50	1.60	1.75	2.00	2.50	3.00
A	0.323	0.630	0.713	0.806	0.900	0.976	0.995
ARE	0.845	0.932	0.951	0.970	0.986	0.997	1.000

estimator when $C = 1.6$ is 95% as efficient as the mean. Note that for $C = 1.6$ the truncation point in (5.5.2) is $C\sqrt{A} = 1.6\sqrt{0.713} = 1.35$, which agrees with Hampel *et al.* (1986; Table 1).

Example 2

Consider $Y \sim N(0, \sigma)$. The score function for σ is $\dot{l}(\sigma) = \sigma^{-1}(\varepsilon^2 - 1)/2$. The optimal B_s -robust estimating equation may be taken as

$$\psi(\sigma) = (\varepsilon^2 - \alpha) \min \left[1, C\sqrt{A}/|\varepsilon^2 - \alpha| \right], \quad (5.5.3)$$

where α is the solution to $E(\psi) = 0$ and $A = E(\psi^2)$. Let $d_1 = \alpha - C\sqrt{A}$ and $d_2 = \alpha + C\sqrt{A}$. The expectation of ψ is

$$\begin{aligned} E(\psi) &= \int_{d_1 < z \leq d_2} (z - \alpha) d\chi_1^2(z) + C\sqrt{A} \left(\int_{z > d_2} d\chi_1^2(z) + \int_{z \leq d_1} d\chi_1^2(z) \right) \\ &= \chi_3^2(d_2) - \chi_3^2(d_1) + d_2 \{1 + \chi_1^2(d_1) - \chi_1^2(d_2)\} - \alpha. \end{aligned}$$

The expectation of ψ^2 is

$$\begin{aligned} E(\psi^2) &= 3 \{ \chi_5^2(d_2) - \chi_5^2(d_1) \} - 2\alpha \{ \chi_3^2(d_2) - \chi_3^2(d_1) \} \\ &\quad + d_1 d_2 \{ \chi_1^2(d_2) - \chi_1^2(d_1) \} + C^2 A, \end{aligned}$$

and

$$E(\psi') = \frac{-2}{\sigma} \int_{d_1 < z \leq d_2} z d\chi_1^2(z) = \frac{-2}{\sigma} \{ \chi_3^2(d_2) - \chi_3^2(d_1) \} = \frac{-2}{\sigma} B.$$

Fisher's information is $2/\sigma^2$ so that $ARE = 2B^2/A$. Results for a range of C 's are presented in Table 5.6. Most notable in this table is the decrease in efficiency, for a fixed

Table 5.6. Standardizing constants (α and A) and asymptotic relative efficiency (ARE) of the bounded-influence estimator for σ in a $N(0, \sigma)$ model. The *self-standardized sensitivity* is $\leq C$.

	C						
	1.25	1.50	1.60	1.75	2.00	2.50	3.00
α	0.545	0.662	0.700	0.749	0.813	0.893	0.937
A	0.193	0.360	0.434	0.549	0.740	1.079	1.341
ARE	0.508	0.612	0.648	0.696	0.761	0.852	0.907

bound C , of the robust estimator for σ compared to the estimator for μ in **Example 1**. This indicates that the mle for σ is more sensitive to outliers than the mle for μ .

5.5.2. Optimal B_s -robust estimators of Poisson and Negative Binomial rate parameters

In this section we investigate robust estimators of the rate parameters for Poisson and NB random variables. This is a first step in developing a robust smoother for the trawl survey data. The Poisson results will be shown to be relevant in terms of identifying type II catches. Our interest is to:

1. Determine the ARE of robust estimators for a range of C 's, which provides a guide for the choice of C in robust smoothing.
2. Assess the finite-sample bias and relative efficiency of robust estimators of the mean for Poisson and NB samples. We use simulations for this purpose, where the simulated samples are designed to be relevant for the trawl survey data. This information is also useful when deciding what C to use in robust smoothing.

3. Assess the finite sample bias of the estimators in 2 when the data is contaminated with an outlier. This provides information about the suitability of the robust procedures for identifying type II catches.
4. Determine the consequences of using a Poisson-based robust estimator of m on NB data, and vice-versa. This is investigated because of the difficulty in differentiating between outlier-contaminated Poisson data and NB data. This information is also relevant for identifying type II catches.

Poisson location

Consider a Poisson random variable Y with mean m . The score function is $\dot{l}(m) = (y - m)/m$ and the Fisher information is $1/m$. The optimal B_s -robust estimating equation is

$$\psi(y; m) = \left(\frac{y - m}{m} - \alpha \right) \min \left[1, \frac{C\sqrt{A}}{|(y - m)/m - \alpha|} \right].$$

A and α are computed as the joint solutions to the equations

$$\begin{aligned} \sum_{y=0}^{\infty} \psi(y; m) p(y; m) - \alpha &= 0, \\ \sum_{y=0}^{\infty} \psi^2(y; m) p(y; m) - A &= 0, \end{aligned}$$

where $p(y; m)$ is the Poisson probability density function. We obtain these solutions numerically. For an i.i.d. sample of size n let $\tilde{m}_{poi}$ denote the root of this equation. Efficiency results for a range of C 's are presented in Table 5.7. If $m = 2$ and $C = 1.7$ then $A = 0.393$ and $RE = 0.947$. The truncation point in this case is $C\sqrt{A} = 1.7\sqrt{0.393} = 1.07$. This agrees with the result in Simpson *et. al.* (1987, Table 1). Note that in their table they report $1/RE$ and scale C by $1/\sqrt{m}$. Finite sample relative efficiencies for $\tilde{m}_{poi}$

Table 5.7. Standardizing constants (α and A) and asymptotic relative efficiency (ARE) for the bounded-influence estimator of m in a $Poi(m)$ model. The *self-standardized sensitivity* is $\leq C$.

		C						
m		1.25	1.50	1.60	1.75	2.00	2.50	3.00
15	α	-0.010	-0.007	-0.006	-0.004	-0.003	-0.001	-0.000
	A	0.022	0.042	0.047	0.054	0.060	0.065	0.066
	ARE	0.844	0.930	0.948	0.967	0.985	0.996	0.999
2	α	-0.063	-0.054	-0.049	-0.034	-0.023	-0.008	-0.003
	A	0.187	0.294	0.367	0.397	0.424	0.467	0.487
	ARE	0.831	0.917	0.935	0.951	0.971	0.990	0.996

are presented in Table 5.8. The results are based on 2000 random samples of size n drawn

Table 5.8. Finite-sample relative efficiency for the bounded-influence estimator of m in a $Poi(m)$ model. The *self-standardized sensitivity* is $\leq C$.

		C						
m	n	1.25	1.50	1.60	1.75	2.00	2.50	3.00
15	25	0.834	0.933	0.983	0.992	0.959	1.009	0.975
	50	0.811	0.954	0.952	0.944	0.974	1.010	0.954
	100	0.843	0.972	0.952	0.934	1.042	1.023	1.002
2	25	0.840	0.859	0.976	0.933	1.003	1.047	0.970
	50	0.832	0.859	0.954	0.981	0.976	0.971	0.947
	100	0.824	0.893	0.960	1.007	0.921	0.977	0.941

from a $Poi(m)$ distribution. They agree fairly well with the results in Table 5.7. They suggest that reasonably efficient ($RE > 85\%$) robust estimators of m can be constructed for sample sizes as small as $n = 25$ when $C \geq 1.5$. Note that the largest simulated bias was 0.034 (unreported results) which occurred when $n = 25$, $m = 15$, and $C = 1.25$. This

bias is negligible for our purposes.

Outlier sensitivity In this section we explore the bias in $\hat{m}$ and $\tilde{m}_{poi}$ caused when Poisson samples are contaminated with a single large outlier. This characterizes an important robust property of $\tilde{m}_{poi}$, and will also be useful when assessing NB results developed in the next section. We simulate data from a Poisson distribution and add an outlier with a value of 500. We choose this value to make the simulations relevant to the 3NO survey data because this value roughly seems to be a lower bound on catches that appear to be type II's in Figures 4.1a,b. The results are presented in Table 5.9. These biases are

Table 5.9. Outlier-induced bias in the the bounded-influence influence estimator of m in a $Poi(m)$ model.

m	n	C						
		1.25	1.50	1.60	1.75	2.00	2.50	3.00
15	25	0.26	0.29	0.26	0.30	0.34	0.44	0.50
	50	0.11	0.12	0.10	0.16	0.15	0.17	0.25
	100	0.05	0.06	0.05	0.07	0.08	0.09	0.10
2	25	0.08	0.09	0.11	0.12	0.12	0.15	0.19
	50	0.05	0.05	0.05	0.05	0.06	0.08	0.08
	100	0.03	0.02	0.03	0.03	0.03	0.04	0.04

negligible for our purposes, especially when compared to the bias in the mle for m , which is $(500 - m)/n$. For $m = 15$ the biases in $\hat{m}$ are 19.40, 9.70, 4.86 for $n = 25, 50, 100$, and for $m = 2$ the biases are 19.92, 9.96, 4.98.

Negative Binomial location

Consider a NB random variables Y with probability density function given by (5.4.1). The first two central moments of Y are m and $v = m(1 + m/k)$. The parameter k

represents the amount of over-dispersion in Y relative to the Poisson distribution, and when k is large the distribution of Y becomes Poisson. When k is small then $Var(Y)$ is much greater than m , the Poisson variance. The score function is

$$\dot{l}(m, k) = \begin{bmatrix} (y - m)/v \\ \Psi'(k + y) - \Psi'(k) + \log\left(\frac{k}{k+m}\right) - \frac{y-m}{k+m} \end{bmatrix},$$

where $\Psi(x) = \log \Gamma(x)$. The Fisher information matrix is $Diag(\mathcal{I}_m, \mathcal{I}_k)$, where $\mathcal{I}_m = 1/v$ and $\mathcal{I}_k = \Psi''(k) - m^2/vk^2 - E[\Psi''(Y + k)]$. The optimal B_s -robust estimating equation is obtained by plugging in the NB score function in (5.5.1). In this case α is a 2×1 vector and A is a 2×2 matrix. However, we will not pursue robust estimation of both m and k .

In Section 5.4 we discussed a problem when estimating k and selecting bandwidths; that is, k is somewhat confounded with bandwidths. In that section we found that the likelihoods for large values of k and small bandwidths tend to be similar to the likelihoods for small values of k and large bandwidths. We proposed a predictive likelihood for estimating k and selecting bandwidths. We will not pursue robust estimation of k in this dissertation. For simplicity we will assume k is known. However, we will investigate the bias caused by assuming small values for k when the data is actually Poisson distributed (i.e. $k \rightarrow \infty$), and this will provide an upper bound on the bias in estimating m caused by mis-specifying k . We will show that this bias is not too large, and can be much smaller than that caused by assuming k is large when in fact it is small. Hence, in our smoothing situation where estimating k is difficult, choosing a small value for k seems safer than choosing a large value.

For fixed k and an i.i.d. sample of size n let $\tilde{m}_{nb}$ denote the optimal B_s -robust estimator of m using $\dot{l}(m) = (y - m)/v$. A and α are computed in the same manner as for the Poisson case. The ARE's for $\tilde{m}_{nb}$ are presented in Table 5.10 for $m = 2, 15$ and $k = 1, 5$. The values of m were chosen to roughly bound the usual range of average

trawl catches within strata, while the values for k represent large and small amounts of Poisson over-dispersion. Generally, as k decreases larger bounds (C) on the *sensitivity* are

Table 5.10. Asymptotic relative efficiency (ARE) of the bounded-influence estimator of m and in a $NB(m, k)$ model.
The *self-standardized sensitivity* is $\leq C$.

k	m		C						
			1.25	1.50	1.60	1.75	2.00	2.50	3.00
5	15	α	-0.016	-0.012	-0.010	-0.008	-0.005	-0.002	-0.001
		A	0.005	0.009	0.011	0.012	0.013	0.015	0.016
		ARE	0.801	0.888	0.909	0.931	0.956	0.982	0.992
1	15	α	-0.017	-0.012	-0.010	-0.008	-0.006	-0.003	-0.002
		A	0.001	0.001	0.002	0.002	0.002	0.003	0.003
		ARE	0.648	0.743	0.772	0.810	0.860	0.922	0.956
5	2	α	-0.082	-0.070	-0.052	-0.044	-0.027	-0.011	-0.005
		A	0.104	0.206	0.232	0.246	0.279	0.317	0.336
		ARE	0.791	0.876	0.896	0.919	0.947	0.977	0.990
1	2	α	-0.112	-0.078	-0.065	-0.054	-0.038	-0.019	-0.011
		A	0.035	0.055	0.064	0.074	0.091	0.118	0.135
		ARE	0.643	0.737	0.767	0.806	0.856	0.920	0.955

required to achieve the same ARE . Finite sample relative efficiencies for $\tilde{m}$ are presented in Table 5.11. These results are based on 1000 random samples from an NB distribution. They agree fairly well with the results in Table 5.10. The results suggest that reasonably efficient ($RE > 85\%$) robust estimators of m can be constructed for sample sizes as small as $n = 25$ when $C \geq 2$. Note that the largest simulated bias was 0.33 (unreported results), which occurred when $n = 25$, $m = 15$, and $C = 1.5$. This bias is negligible for our purposes.

Outlier sensitivity In this section we repeat, for the NB distribution, the outlier simulation presented with the Poisson results; that is, we generated random samples from

Table 5.11. Relative efficiency (RE) of the bounded-influence estimator of m in a $NB(m, k)$ model. The *self-standardized sensitivity* is $\leq C$.

k	m	n	C						
			1.25	1.50	1.60	1.75	2.00	2.50	3.00
5	15	25	0.773	0.944	0.910	0.881	0.942	0.973	1.041
		50	0.750	0.887	0.853	0.946	0.933	0.906	0.961
		100	0.805	0.883	0.918	0.922	0.976	0.981	1.073
1	15	25	0.691	0.746	0.779	0.841	0.854	0.869	1.008
		50	0.610	0.760	0.779	0.795	0.804	0.868	0.959
		100	0.674	0.711	0.757	0.784	0.786	0.982	1.055
5	2	25	0.805	0.894	0.961	0.880	0.934	0.910	0.972
		50	0.756	0.897	0.897	0.871	0.958	0.958	0.943
		100	0.757	0.870	0.901	0.926	0.947	1.016	1.014
1	2	25	0.633	0.712	0.774	0.768	0.885	0.979	0.859
		50	0.646	0.710	0.779	0.781	0.845	0.946	0.925
		100	0.636	0.682	0.803	0.758	0.849	0.942	0.905

NB distributions, add an outlier with a value of 500, and compute $\bar{m}_{nb}$. We use the same values of k when computing $\bar{m}_{nb}$ as used for generating the random samples. The outlier-induced biases in $\bar{m}_{nb}$ are presented in Table 5.12. While the biases are larger than in the Poisson case, especially when $k = 1$, they are still small enough for us to ignore. These biases are much smaller than the biases for the mle $\hat{m}_{nb}$, which are identical to the biases for $\hat{m}_{poi}$ presented previously. This is because $\hat{m}_{nb} = \hat{m}_{poi}$.

The robust estimator defined by (5.5.1) is just a weighted score function, with an adjustment for bias because the weights depend on the Y 's. The weights are one's for all y 's that are close to m , and the weights are less than one for y 's that are not close to m . The weights decrease as $|y - m|$ increases. The robust weights are commonly used for outlier detection. Observations with very small robust weights, relative to the majority, are often identified as outliers. The reduction in bias in $\bar{m}_{nb}$ in our outlier simulations

Table 5.12. Outlier-induced bias in the bounded-influence estimator of m in a $NB(m, k)$ model.

k	m	n	C						
			1.25	1.50	1.60	1.75	2.00	2.50	3.00
5	15	25	0.35	0.53	0.58	0.59	0.60	0.83	1.09
		50	0.29	0.26	0.34	0.38	0.35	0.46	0.53
		100	0.12	0.19	0.15	0.14	0.14	0.20	0.21
1	15	25	1.18	1.21	1.24	1.27	1.44	1.74	2.02
		50	0.73	0.62	0.70	0.76	0.72	0.89	1.07
		100	0.35	0.35	0.26	0.33	0.37	0.46	0.44
5	2	25	0.11	0.12	0.15	0.15	0.15	0.20	0.20
		50	0.06	0.06	0.06	0.06	0.08	0.10	0.11
		100	0.02	0.02	0.03	0.03	0.03	0.05	0.05
1	2	25	0.22	0.21	0.24	0.23	0.25	0.29	0.34
		50	0.10	0.10	0.12	0.12	0.11	0.16	0.17
		100	0.04	0.05	0.04	0.06	0.06	0.06	0.09

occurs because the robust weight for the outlier is usually very small; hence, the results suggest that the B_s -robust estimator produces robust weights that are useful for detecting outliers. These procedures should also be useful for identifying type II catches, because these catches are just outliers in terms of the smooth mean function model.

Bias when the model is wrong In this section we explore the potential biases in robust M-estimators that arise when the data are assumed be Poisson distributed, but are actually NB distributed, and vice-versa. The rationale for this is as follows. At the beginning of the chapter we proposed the NB distribution to model trawl catches, and the estimates of k in Table 5.2b support this because the $\hat{k}$'s are small which indicates considerable Poisson overdispersion. However, it may be that the estimates are small because what we really have is Poisson data "contaminated" with type II catches. Since

the robust procedures will be used to identify type II catches it is useful to investigate the biases in the robust estimators arising from a mis-specification of the distribution of the type I catches. This provides information about which of $\tilde{m}_{poi}$ or $\tilde{m}_{nb}$ (or whether k should be large or small) is more suitable for identifying type II catches.

In Table 5.13a we present the biases in $\tilde{m}_{poi}$ based on the NB simulation experiment.

Table 5.13a. Simulated biases in the Poisson bounded-influence estimator of m applied to $NB(m, k)$ data.

k	m	n	C						
			1.25	1.50	1.60	1.75	2.00	2.50	3.00
5	15	25	-0.86	-0.90	-0.87	-0.85	-0.71	-0.55	-0.46
		50	-0.95	-0.82	-0.82	-0.89	-0.80	-0.62	-0.50
		100	-0.94	-0.88	-0.86	-0.89	-0.79	-0.61	-0.49
1	15	25	-4.16	-4.17	-4.24	-4.15	-4.12	-3.84	-3.34
		50	-4.37	-4.31	-4.45	-4.36	-4.19	-3.93	-3.52
		100	-4.48	-4.54	-4.41	-4.38	-4.27	-3.98	-3.64
5	2	25	-0.10	-0.10	-0.11	-0.07	-0.06	-0.02	-0.01
		50	-0.11	-0.10	-0.10	-0.07	-0.04	-0.03	-0.02
		100	-0.11	-0.11	-0.10	-0.08	-0.05	-0.03	-0.02
1	2	25	-0.53	-0.50	-0.46	-0.41	-0.35	-0.25	-0.17
		50	-0.56	-0.50	-0.48	-0.44	-0.35	-0.26	-0.20
		100	-0.57	-0.52	-0.49	-0.43	-0.37	-0.25	-0.18

For NB data $\tilde{m}_{poi}$ is not distributionally robust, particularly when k is small. The biases are large even for large C . The biases in $\tilde{m}_{nb}$ based on the Poisson simulation experiment are presented in Table 5.13b. When C is small these biases are large, particularly when k is small. For small values of C the biases are slightly larger than in Table 5.13a; however, when $C \geq 2$ the biases in Table 5.13b are much smaller than in Table 5.13a. This suggests that for $C \geq 2$ relatively more bias is introduced by using the Poisson-robust M-estimator when the data is NB distributed than by using the NB-robust M-estimator

Table 5.13b. Simulated biases in the NB bounded-influence estimator of m applied to $Poi(m)$ data.

k	m	n	C						
			1.25	1.50	1.60	1.75	2.00	2.50	3.00
5	15	25	0.98	0.74	0.61	0.46	0.33	0.10	0.05
		50	0.92	0.69	0.63	0.50	0.31	0.15	0.05
		100	0.95	0.72	0.61	0.47	0.31	0.14	0.05
1	15	25	5.40	3.45	2.95	2.26	1.56	0.78	0.40
		50	5.37	3.48	2.88	2.28	1.55	0.80	0.39
		100	5.34	3.47	2.90	2.27	1.52	0.74	0.39
5	2	25	0.13	0.11	0.09	0.07	0.06	0.03	0.02
		50	0.13	0.12	0.10	0.08	0.06	0.03	0.01
		100	0.13	0.11	0.10	0.08	0.05	0.03	0.02
1	2	25	0.77	0.52	0.41	0.32	0.23	0.12	0.08
		50	0.77	0.52	0.44	0.34	0.24	0.13	0.07
		100	0.76	0.51	0.43	0.35	0.24	0.12	0.06

for fixed k when the data is Poisson distributed. This suggests that $\tilde{m}_{nb}$ is more useful for detecting type II catches because if catches are actually Poisson distributed then the proportion of type II catches not selected using $\tilde{m}_{nb}$ will be less than the proportion of catches incorrectly selected using $\tilde{m}_{poi}$ when the data are NB distributed. The latter error is more grievous in that the separation of catches as type I and II is only a hypothesis at this stage, and we wish to be conservative in assigning catches as type II.

If k is robustly along with m then the differences in biases are even greater. Unreported results suggest that $\tilde{k}$'s are usually very large for Poisson data; hence, there is little to be lost by using $\tilde{m}_{nb}$ on Poisson data, especially if k is estimated, but substantial biases exist in $\tilde{m}_{poi}$ used with NB data. Unfortunately in robust smoothing k is difficult to estimate in conjunction with bandwidth selection, so these methods are not pursued further.

5.5.3. Local likelihood estimates

In this section we present an algorithm for robust smoothing of NB count data. We first review some literature on robust kernel smoothing which suggests that robust M-estimation retains the same optimality properties as in the i.i.d. case. We then present the algorithm we use to analyze the 3NO trawl survey data in terms of robust variable selection, and to identify type II catches.

It is a fairly easy matter to develop robust local likelihood regression estimates of $\mu(\mathbf{x})$ for NB data using $\tilde{m}_{nb}$. All we have to do is weight (5.5.1) with local (kernel) weights for the sample. Hence, two sets of weights are used for estimation. The kernel weights are used to weight observations according to their distance from $\mathbf{x}$, and robust weights are used to weight observations according to their distance from $\mu(\mathbf{x})$. To estimate a robust smooth mean for the entire sample involves n robust fits, each of which involves an iterative minimization to get $\tilde{\mu}_{nb}(\mathbf{x}_i)$, $i = 1, \dots, n$. This is computationally laborious, but straight-forward to implement.

Background

M-estimators have been considered by Härdle and Gasser (1984) for robust nonparametric univariate regression based on an additive errors model. Their robust nonparametric estimate of μ_x , denoted as $\tilde{\mu}_x$, is the solution to

$$\sum_i \gamma_i(x) \psi(y_i - \tilde{\mu}_x) = 0,$$

where $\gamma_i(x)$ are integral-type kernel (local) weights. Härdle and Gasser (1984) show that the asymptotic variance of $\tilde{\mu}_x$ factors into two parts, one related to the smoothing and another related to the robust trimming of the estimating equation. The traditional optimality theory of robust estimation applies in a straightforward manner to kernel

smoothing because of this factorization; that is, the optimal ψ function is of the form (5.5.1) in Section 5.5.1.

Hall and Jones (1990) studied kernel M-estimates under the “random design” case. They advocated selecting the bandwidth and C using CV or GCV , and showed that if the additive errors are i.i.d. with finite moments then

$$\frac{ASE(b_{cv}, C_{cv})}{\inf_{b \in B, 0 \leq C \leq \infty} ASE(b, C)} \rightarrow 1 \text{ almost surely,}$$

where the average squared error is computed over a central range of x_i 's. We will not pursue this approach to select C . We use the simulations presented in Sections 5.5.2 to choose a value for C that is reasonably efficient for the i.i.d. problem.

Algorithm

In this section we present two algorithms we use to compute bounded influence lle's for variable selection purposes, and to identify type II catches. The first algorithm is straightforward, and uses robust and local weights. It is essentially a local version of the algorithm given in Carroll and Ruppert (1988) for computing robust M-estimates. For a simple kernel smoother and a known k parameter this involves estimating only one parameter for each of the n sets of local weights; hence, α and A_θ in (5.5.1) are scalars. For each observation, α_i and A_i , and robust weights, must be computed and the procedure iterated until convergence. If a robust estimate of k is also required then we suggest an additional step in the iterative process in which an M-estimate of k is computed based on the robust estimates of the $\mu(x_i)$'s. The score function for m is

$$s(y, m) = \dot{l}(y, m) = \frac{y - m}{m(1 + m/k)}.$$

To simplify notation let

$$\omega(y, m, \alpha, A) = \min \left[1, \frac{CA^{1/2}}{|s(y, m) - \alpha|} \right].$$

The steps in the algorithm are:

1. Compute the kernel weights,

$$w_{ij} = \frac{W\{(\mathbf{x}_i - \mathbf{x}_j)/\mathbf{b}\}}{\sum_{j=1}^n W\{(\mathbf{x}_i - \mathbf{x}_j)/\mathbf{b}\}}.$$

Recall that the $\mathbf{x}_i$'s are the covariates.

2. Supply starting values for $\bar{\mu}(\mathbf{x}_i)$'s, α_i , and A_i .
3. For each observation i compute the n robust weights

$$w_{ij}^r = \omega\{y_j, \bar{\mu}(\mathbf{x}_i), \alpha_i, A_i\}, \quad j = 1, \dots, n,$$

based on the values of α_i , A_i , and the estimate of $\mu(\mathbf{x}_i)$, i.e. $\bar{\mu}(\mathbf{x}_i)$, obtained from the previous iteration.

4. Update $\bar{\mu}(\mathbf{x}_i)$ by solving for m in

$$\frac{\sum_{j=1}^n \{s(y_j, m) - \alpha_i\} w_{ij}^r w_{ij}}{\sum_{j=1}^n w_{ij}^r w_{ij}} = 0.$$

5. Using the $\bar{\mu}(\mathbf{x}_i)$'s, α_i , and A_i from steps 3 and 4, update α_i and A_i , $i = 1, \dots, n$,

using

$$\begin{aligned}\alpha^{\text{new}} &= \sum_{y=0}^{\infty} s\{y, \bar{\mu}(\mathbf{x}_i)\} \omega\{y, \bar{\mu}(\mathbf{x}_i), \alpha_i, A_i\} p\{y; \bar{\mu}(\mathbf{x}_i), k\}, \\ A^{\text{new}} &= \sum_{y=0}^{\infty} [s\{y, \bar{\mu}(\mathbf{x}_i)\} - \alpha_i]^2 \omega^2\{y, \bar{\mu}(\mathbf{x}_i), \alpha_i, A_i\} p\{y; \bar{\mu}(\mathbf{x}_i), k\}.\end{aligned}$$

The NB pdf $p(y; m, k)$ is given in (5.4.1).

6. Iterate steps 3-5 until convergence.

Steps 3-6 are repeated for each observation. The algorithm is slow, but practical for the samples sizes we have in the 3NO trawl survey. Convergence requires reasonable starting values for α_i , A_i , and $\mu(\mathbf{x}_i)$.

We used this algorithm extensively with our survey data. We performed many robust smooths using different kernel weights for various subsets of covariates, and various bandwidth values. Each time we performed robust smooths 22 times for the 11 years in our data set, and two smoothing regions we employed. Too frequently to ignore, the above algorithm failed to converge. For a few $\mathbf{x}_i$'s, M-estimates of $\mu(\mathbf{x}_i)$ would oscillate between two values. These values were usually similar, but occasionally they were quite different, especially if the smoothing neighborhood was small. In such cases we used another algorithm.

The second algorithm we employed involved direct M-estimation for each local neighborhood. It is computationally more difficult, but converged every time we used it. We use the same notation as above. The steps in the algorithm are:

1. Compute the kernel weights, w_{ij} .

2. For each $\mathbf{x}_i$, solve simultaneously for m , α , and A in

$$\begin{aligned} \sum_{j=1}^n w_{ij} \{s(y_j, m) - \alpha\} \omega(y_j, m, \alpha, A) &= 0, \\ \alpha - \sum_{y=0}^{\infty} s(y, m) \omega(y, m, \alpha, A) p(y; m, k) &= 0, \\ A - \sum_{y=0}^{\infty} \{s(y, m) - \alpha\}^2 \omega^2(y, m, \alpha, A) p(y; m, k) &= 0. \end{aligned}$$

The robust lle for $\mu(\mathbf{x}_i)$ is the m solution.

We used numeric root-finding procedures, so starting values for m , α , and A were still required. We used the appropriate values in Table 5.10, where m was selected to be close to the kernel-weighted average for each $\mathbf{x}_i$. This procedure, although slow, worked very well in practice.

The robust smoother appeared to work well for the trawl survey data, unless bandwidths were very small. In this case the size of smoothing neighborhoods became very small, with many local neighborhoods having less than seven observations. Robust estimation with very small sample sizes such as these is difficult. For reasonably large bandwidths (such that smoothing neighborhoods usually contained >10 observations) the robust smoother performed as expected. Observations that we thought *a priori* were outliers tended to be downweighted considerably, whereas no observations that did not appear to be outliers received very small robust weights.

5.5.4. Robust variable selection algorithm

In this section we present a “robustified” nonparametric bandwidth and variable selection algorithm. This algorithm has three stages. The first stage involves nonrobust smoothing with all variables, which will usually be conducted anyway. This stage is used to provide a

pilot choice of bandwidths for the second stage, which is robust smoothing using the pilot bandwidths. These bandwidths should be larger than those found in the first stage. This is because the nonrobust bandwidth selection used in the first stage will usually result in under-smoothing if the observations are “contaminated” with outliers. The second stage is performed only to assist in identifying the contaminated observations. Such observations usually have small robust weights, and are excluded from the third stage. Recall that observations with small robust weights are those that have relatively great influence in a nonrobust analysis. Such observations are called discordant. In stage three the variable selection procedures in Section 5.3.2 are applied directly to the “cleaned” data in which the discordant observations have been removed.

Our procedure is not equivalent to robust parametric variable selection *per se*, although in practice many people deal with parametric variable selection and outliers in a similar manner. Hampel *et. al.* (1986), Ronchetti and Staudte (1994), and Ronchetti *et. al.* (1997), among others, discuss robust parametric variable selection. In these approaches robust prediction error (PE) measures are used. For the related problem of bandwidth selection, Boente *et. al.* (1997) have presented extensive simulations that document the sensitivity of $CV\ b_{\min}$ ’s to outliers. To adapt these approaches for robust nonparametric variable selection we require a robust smoother, and a robust method for bandwidth selection. While robust PE measures have been developed, another level of iteration is required to find “optimal” robust bandwidths for the robust smoother. This is on top of the n optimizations required to compute the robust estimates of the $\mu(\mathbf{x}_i)$ ’s. This is a computationally daunting task that is also complicated because measures of PE are not differentiable everywhere in terms of $\mathbf{b}$. This problem was also noted in Hastie and Tibshirani (1990, pg. 237). Attempts to find robust optimal bandwidths have been unsuccessful. This is why we propose the three stage approach above, which is greatly simplified in that only one robust smooth is required, after which bandwidths are selected

in the usual manner.

Unfortunately there are two somewhat arbitrary decisions required in this algorithm. The first is the pilot bandwidth choice, and the second is which observations to remove before stage 3. We choose to remove observations whose robust weights are much smaller than the majority of weights. Sometimes this is easy to decide because a wide gap appears between the small robust weights and those close to one. The choice of pilot bandwidths is less problematic because, in our analyses at least, essentially the same observations are removed for a range of bandwidths greater than those found in stage one. If this variable selection algorithm is to be used routinely then objective criteria for determining when robust weights are small need to be developed. This is not a trivial task. The criteria will probably depend on C , n , and other factors. Our interest in robust variable selection is only to test the sensitivity of earlier “nonrobust” results to a small number of observations; therefore, we will not pursue this here.

In smoothing, each y_j produces n robust weights according to its discrepancy with the $\bar{\mu}(\mathbf{x}_i)$'s. We want a single summary of these weights to assess the discordancy of y_j . We do not really care if the robust weight for an observation at $\mathbf{x}_j$ is large for estimating $\mu(\mathbf{x}_i)$ when $\|\mathbf{x}_i - \mathbf{x}_j\|$ is large, because the kernel weight for such observations will be small. The kernel weighted average of robust weights for y_j

$$\bar{w}_j = \frac{\sum_{i=1}^n w_{ij} w_{ij}^r}{\sum_{i=1}^n w_{ij}},$$

is reasonable because it is usually insensitive to the robust weights for pairs of observations that are far apart in the covariate space; that is, if $\|\mathbf{x}_i - \mathbf{x}_j\|$ is large then w_{ij} will be close to zero, and the robust weight for y_i in estimating $\mu(\mathbf{x}_j)$ will not be part of the summary $\bar{w}_j$. This is what we use for the analyses of trawl survey data, and in the following example.

Example

We illustrate our procedure for robust bandwidth selection using a simple measurement error problem. We describe in detail many of the computations, which are similar to those in robust smoothing of the trawl survey data, so this example should help fix ideas for the trawl survey data as well. In this problem we compute a robust nonparametric regression function, and also a robust estimate of scale for the observations. This example involves acoustic measurements of the distance between the trawl wings (see Figure 1.2), commonly referred to as the wing spread. The measurements are taken periodically during a tow using two acoustic devices mounted on the wings. The measurements are plotted as circles in the top panel of Figure 5.2 in the **Appendix (Section 5.12)**. The meaning of the circle diameters is described below. Occasionally the acoustic devices malfunction, usually because they are somehow pointed away from each other, in which case they measure something other than wing spread.

We assume a normal regression model for wing spread measurements; that is, if Y_t denotes the measurement at time t , then

$$Y_t = \mu_t + \varepsilon_t, \quad t = 1, \dots, T$$

where $\varepsilon_1, \dots, \varepsilon_T$ are i.i.d. $N(0, \sigma^2)$ random variables. We use the *CV* measure presented in Section 5.3.1 for bandwidth (b) selection. We compute robust estimates of the μ_t 's and σ^2 separately; that is, we use (5.5.2) and (5.5.3), instead of estimating these parameters jointly. The algorithm we use is to assume starting values for σ^2 and the μ_t 's, which we denote as σ^{*2} as μ_t^* 's, then

1. compute kernel weights $w_{ti} = W\left(\frac{t-i}{b}\right)$

2. Compute robust weights for μ_t^* ,

$$w_{ti}^\mu = \min \left[1, \left(\frac{1.35}{|y_i - \mu_t^*|/\sigma^*} \right) \right].$$

3. Update values for μ_t^* as

$$\mu_t^* = \frac{\sum_{i=1}^T w_{ti}^\mu y_i}{\sum_{i=1}^T w_{ti}^\mu}, \quad t = 1, \dots, T.$$

Let $e_t^* = y_t - \mu_t^*$.

4. Compute robust weights for σ^{*2} ,

$$w_t^\sigma = \min \left[1, \left(\frac{1.05}{|(e_t^*/\sigma^*)^2 - 0.7|} \right) \right].$$

5. Update the value of σ^* as

$$\sigma^{*2} = \frac{\sum_{t=1}^T w_t^\sigma e_t^{*2}}{0.7 \sum_{t=1}^T w_t^\sigma}.$$

6. Iterate steps 2-5 until convergence.

The converged values are denoted as $\tilde{\mu}_t$, $t = 1, \dots, T$, and $\tilde{\sigma}^2$. The value of $C\sqrt{A} = 1.35$ that we use in w_{ti}^μ yields a robust estimator that is asymptotically 95% as efficient as the mle (see Table 5.5). The value of $C\sqrt{A} = 1.05$ that we use in w_t^σ yields a robust estimator that is asymptotically 65% as efficient as the mle (see Table 5.6). The value of α for this choice of C is 0.7.

A kernel smooth of the wing spread measurements, and a robust kernel smooth of these measurements, are plotted in the top panel of Figure 5.2. These are based on $CV \, b_{\min} = 0.25$. The inverse of the locally-averaged robust weights, $\bar{w}_i$'s, are used to

define the diameter of the circles. A large circle corresponds to a discordant observation. Estimated wing spreads are very erratic, and some fishing gear experts have suggested that it is almost impossible for a trawl to behave as the estimates suggest. The robust smoother is even slightly more erratic than the nonrobust smoother, which is surprising. One might expect the robust smoother to down-weight extreme points and result in a smoother fit than the nonrobust smoother. The cause for this is the small bandwidth, or small neighborhoods, used for local averaging. Consider the data in Table 5.14. The

Table 5.14. Wing spread data, with kernel (w_t) and robust (w_t^r) weights

Time (t)	3.0	3.25	3.5	3.75	4.0
Y_t	17.2	17.9	16.4	25.1	28.9
w_t	0.0002	0.0056	0.0736	0.3444	0.5761
w_t^r	0.220	0.238	0.203	1.00	0.904
$w_t^r \times w_t$	0.0000	0.0015	0.0169	0.3910	0.5908

observation at $t = 4$ minutes is the largest in the data set. The kernel estimate of μ_t at $t = 4$ minutes is a weighted sum of the y_t 's, where the weights are the w_t 's. This sum is 26.6, which is the number plotted in the top panel of Figure 5.2 for the kernel smoother estimate of trawl width at $t = 4$. Using this value for μ_t , the robust weights are also presented in Table 5.14. The robust lle of μ_t is a weighted sum of the y_t 's and standardized product of the kernel and robust weights, standardized so that the sum of the products is one. These weights are presented in the last row in Table 5.14, and give $\tilde{\mu}_t = 27.2$. This is greater than $\hat{\mu}_t$ because of the pattern in the kernel weights. Had the kernel weights all been equal, as in bin smoothing, then $\tilde{\mu}_t$ would be less than $\hat{\mu}_t$. What we have described is the first iteration involved in computing the robust lle for μ_t . After 40 iterations $\tilde{\mu}_t = 28.44$ with a relative error $\leq 10^{-6}$. This is the value plotted in the top panel of Figure 5.2 for the robust kernel smoother estimate of trawl width at $t = 4$.

We chose a pilot bandwidth of $b = 1$ based on the above analysis. The nonrobust and robust smooths for this bandwidth are plotted in the middle panel of Figure 5.2. The kernel smooth is still considerably influenced by a small number of observations. This analysis is only used to identify which observations to remove from further analysis. We removed 14 observations with robust weights less than 0.8. These weights are, in time sequence,

0.19, 0.14, 0.21, 0.49, 0.20, 0.20, 0.33,
0.35, 0.41, 0.52, 0.55, 0.34, 0.25, and 0.25.

For the remaining observations, 77% had robust weights ≥ 0.95 .

In the bottom panel in Figure 5.2 the “cleaned” data is plotted along with a robust and nonrobust smooth. The $CV\ b_{\min} = 0.89$, and is much larger than the bandwidth we selected for the full sample (see panel a in Figure 5.2). The estimates of wing spread in this panel appear reasonable in terms of the trawl behavior. The robust and nonrobust fits are nearly identical which suggests there are no discordant observations in the “cleaned” data.

Remark 4. *The three panels in Figure 5.2 represent the three-stage procedure used in the following section for “robustified” nonparametric variable selection.*

This example succinctly demonstrates how our “robustified” bandwidth selection procedure works. For the trawl survey data we use the same procedure, except that we smooth over multidimensional covariates. The wing spread data is a simple example that demonstrates that the robust smoothing procedures we have developed work reasonably in practice.

As an aside, recall in **Chapter 3** that we said the nominal wing spread assumed by DFO in swept-area calculations for the survey trawl was 45 feet, or 13.7 meters. The wing

spread measurements in Figure 5.2 are much greater than this, and are usually around 17 meters; however, these are for the new survey gear used by DFO since 1995, which is a shrimp trawl that is better at sampling small fish. The data we analyze are for the old survey trawl. Note that the nominal swept area for the shrimp trawl is 55 feet, or 16.8 meters. This is quite close to the data presented in Figure 5.2. This particular wing spread data set was randomly chosen from thousands that are available.

5.6. Analysis of the 3NO trawl survey data, II

In this section we apply the “robustified” variable selection procedures to the 3NO data set. The intent of this analysis is to choose variables for our smooth mean function model for type I trawl catches. The variables chosen should result in estimates of $\mu(\mathbf{x}_i)$ that are close to the majority but not necessarily all of the $\mu(\mathbf{x}_i)$ ’s. The outline of the analysis is as follows. First, we choose pilot bandwidths based on the results in Tables 5.3a-c, and perform a robust Negative Binomial (NB) nonparametric regression. We again assume the NB k parameter is 0.65. We remove a small number of discordant observations and then run the variable selection algorithm outlined in Section 5.3.2 on the “cleaned” data.

Based on the results in Table 5.3a-c we used only spatial location and tow depth for robust smoothing. For the slope region we choose $b_{SP} = 0.2$ and $b_{DP} = 0.1$ for pilot bandwidths in all years. For the shelf region we choose $b_{SP} = 0.1$ and $b_{DP} = 1$. The robust weights were computed using $C = 2$ based on the efficiency and robust properties of $\tilde{m}_{nb}$ discussed in Section 5.5.2. Tows with robust weights < 0.4 were identified as discordant and removed from the next step. This resulted in the removal of 91 tows for all years, which is approximately 5% of the data. The maximum number of tows removed (15) was in 1987. The maximum percentage of tows removed (9.1%) was in 1994. The results from applying the variable selection procedures to the “cleaned” data are presented

in Tables 5.15a-c. On the shelf region only one of the “cleaned” tows had a positive catch, so CV variable selection could not be applied in this year.

Spatial location and tow depth again are the variables that have the greatest predictive ability, as measured by all criteria. Tow temperature has importance in terms of CV in six of eleven years, in two years for GCV , and little importance in any year for GC_p . Time between tows appears about as important as temperature. Wind speed appears less important than temperature and time. The “robustified” bandwidths usually are similar to those in Tables 5.3a-c; however, substantial differences sometimes exist. For example, in 1991 the CV $b_{SP} = 0.039$ for the full model, while this value is 0.193 in Table 5.15a. The CV results appear more sensitive to the removal of discordant observations than the GCV or GC_p results.

The CV smoother goodness-of-fit using all covariates for the “cleaned” data is generally slightly better than for all the data. The total deviance for all years based on the “cleaned” data is 681 with 748 df^{err} ’s, while for all the data the total deviance was 746 with 739 df^{err} ’s. With 9 more df^{err} ’s the CV smoother for the “cleaned” data can explain 65 more units of deviance. This does not seem like a great improvement, which might indicate that there are no outliers in the original data, or that the CV bandwidths are too small and lead to over-fitting, which would tend to mask outliers in the smooth mean function model. The plots in Chapter 4 suggest that outliers do exist, so the suggestion of over-fitting seems more reasonable. The data “cleaning” procedure had a greater impact on the GC_p results. For the smoother using only spatial location and tow depth, the total deviance for all the data is 1305. For the “cleaned” data the deviance is 1251. This is not a large reduction, but the smoother for the “cleaned” data has 165 more df^{err} ’s, so this magnitude of deviance reduction occurs with substantially fewer “parameters”, which suggests the robust procedures we employed has removed some outliers. This also suggests that the bandwidths selected using GC_p do not under-smooth, at least as much

as for bandwidths selected using CV .

The per observation change in deviance based on CV $b_{\min}$'s caused by removing temperature, time, and wind speed from the model are plotted in Figure 5.3a,b. Total deviance increased more than expected in only three years, and substantially so only in 1985. The total change in deviance for all years is 237, while the total change in df^{err} 's is 282. This suggests that removing temperature, time, and wind speed from our model does not change the goodness-of-fit any more than what is expected due to the change in neighborhood size resulting from the different bandwidths. Note that the deviances from fitted models with just spatial location and tow depth usually exceed the deviances from a constant mean model by a factor of more than twice the difference in df^{err} 's (unreported results). This suggests that spatial location and tow depth account for "significant" amounts of deviance. The model goodness-of-fit for GC_p bandwidths with and without (unreported results) temperature, time, and wind speed are nearly identical.

We conclude in this section that GC_p is a better criterion for bandwidth and variable selection for our data because it appears that bandwidths and variables selected using CV result in under-smoothing. The GCV criterion leads to gross under-smoothing in some years, and this criterion does not appear suitable for modelling trawl catches. The GC_p results in Table 5.15c suggest that only spatial location and tow depth are important predictive covariates for trawl catches throughout 1984-94. The only occasion where any of the other covariates appear at all to have predictive importance is in 1994 on the shelf region, but as already mentioned these results are suspect because only one of the "cleaned" tows encountered cod in this region and year. The full sample results for this year and region (see Table 5.3c) indicate that only space and depth are important.

5.7. Identification of Type II catches

In this section we use the robust procedures and results previously developed to identify type II catches in the trawl survey data. Recall that these catches are inconsistent, or discordant, with a Negative Binomial (NB) smooth mean function model, where the mean is assumed to be a smooth function of spatial location and tow depth. We first select bandwidths and the NB k parameter from the “cleaned data” in the previous section, and use these values in our robust nonparametric regression analysis to produce robust weights and residuals to assist in identifying type II catches. Recall that the NB k parameter represents Poisson overdispersion; that is, if Y is a NB random variable with mean m , then $Var(Y) = m(1 + m/k)$. The reason for using the “cleaned” data is that if we select bandwidths and the k parameter using all the data then the presence of large type II catches would cause us to select smaller bandwidths, and estimate a small value for k . This, in turn, would make it more difficult to identify type II catches. This is because with smaller smoothing neighborhoods and a small value for k , a wide range of observations appear plausible in terms of the NB distribution. Hence, type II catches may “mask” themselves if we do not use robust procedures to select bandwidths and estimate k . While our variable selection procedures were, by and large, insensitive to the value of k , we do not feel that the identification of type II catches will share this sensitivity.

We use the same procedure for estimating k as in Section 5.4. As usual let b_{SP} and b_{DP} denote the smoother bandwidths for spatial location and tow depth. We simultaneously minimize the predictive likelihood defined in (5.4.2) for b_{SP} , b_{DP} , and k . The results are presented in Table 5.16a. Again the $\hat{k}$ ’s are smaller on the slope than on the shelf. The k ’s do not seem to vary greatly among years however. For simplicity we use the same value for k in every year. The average k in Table 5.16a is 1.47 for the shelf and 0.56 for the slope region.

We concluded in the previous section that the CV bandwidths seemed to small and

Table 5.16a. Estimates of the NB scale parameter k and bandwidths b that maximize a predictive likelihood function.

Year	3NO shelf region			3NO slope region		
	b_{SP}	b_{DP}	$\hat{k}$	b_{SP}	b_{DP}	$\hat{k}$
84	0.0203	0.1299	1.09	0.1041	0.0510	0.31
85	0.0268	0.0189	0.69	0.0604	0.0607	0.34
86	0.0201	0.0435	1.19	0.0735	0.0721	0.76
87	0.0194	0.0234	0.81	0.0448	0.0882	0.57
88	0.0289	0.0947	1.14	0.1194	0.0461	0.47
89	0.0227	0.0541	1.53	0.0500	0.0316	0.98
90	0.0207	0.1733	1.51	13.3520	0.1248	0.85
91	0.0262	0.1174	1.87	0.0370	0.0809	0.49
92	0.0325	17.2187	3.56	0.0344	0.0327	0.61
93	0.0457	0.0814	1.44	0.0858	0.0284	0.72
94	.	.	.	19.5741	0.0245	0.39

Note: SP -spatial location, DP -depth.

led to under-smoothing. We decided that among the three measures of prediction error we studied (CV , GCV , GC_p), GC_p appeared most suitable for our data. The predictive likelihood defined in (5.4.2) is similar to CV , which suggests that the bandwidths in Table 5.16a may also be too small, which would result in an overestimate of k . The analogous criterion to GC_p for bandwidth selection and the estimation of k is

$$AIC = -2l + 2 \{trace(\mathbf{W}) + 1\},$$

where l is the NB log-likelihood (eq. 5.4.1). The number of “parameters” in the model is $trace(\mathbf{W}) + 1$, where $trace(\mathbf{W})$ is the number of smoothing parameters, and the $+1$ accounts for k . The results of minimizing AIC for b_{SP} , b_{DP} , and k are presented in Table 5.16b. This criteria in many years favors large values of k and very small bandwidths,

Table 5.16b. Estimates of the NB scale parameter k and bandwidths b that minimize AIC.

Year	3NO shelf region			3NO slope region		
	b_{SP}	b_{DP}	$\hat{k}$	b_{SP}	b_{DP}	$\hat{k}$
84	0.0010	1.0101	18280	0.9132	26.9261	0.17
85	0.0029	2.2614	3374.0	0.0005	1.0000	361.70
86	0.0022	0.0798	2259.8	0.0002	18.1923	211.97
87	0.0002	6.3361	19026	0.0049	0.0598	8223.6
88	0.0012	0.0767	7521.1	0.0201	0.0171	5639.0
89	0.0077	0.0133	4673.1	43.2036	0.0644	0.63
90	0.0013	0.6726	1671.5	0.0593	0.0016	8540.4
91	0.0160	13.6298	496.05	0.0349	26.8025	0.66
92	0.0368	49.1488	462.81	0.0432	17.4913	0.44
93	0.0576	17.1989	3.45	0.0752	11.3883	0.48
94	39.1307	0.3441	87.53	0.0020	5.4148	52.30

Note: SP -spatial location, DP -depth.

which leave very few df^{err} 's and results in gross under-smoothing. We noted the sensitivity of GC_p b_{min} 's to the choice of k in Section 5.4.1, and this is another illustration of this. For our model the AIC criterion does not appear suitable for bandwidth selection and k estimation. Another popular criterion is

$$BIC = -2l + \log(n) \{trace(W) + 1\},$$

where n is the sample size. The results of minimizing BIC are presented in Table 5.16c. The bandwidths and $\hat{k}$'s in this table are comparable with those in Table 5.16a. The average k in Table 5.16c is 1.39 for the shelf and 0.49 for the slope region. Note that the value for $\hat{k}$ in 1994 for the shelf region was not used when computing the annual average. This is because only one of the "cleaned" tows had a catch of cod, and we feel that the estimation of k in this situation is not reliable.

Table 5.16c. Estimates of the NB scale parameter k and bandwidths b that minimize BIC.

Year	3NO shelf region			3NO slope region		
	b_{SP}	b_{DP}	$\hat{k}$	b_{SP}	b_{DP}	$\hat{k}$
84	0.0904	23.7825	0.80	83.9398	26.2269	0.17
85	0.0657	24.6909	0.52	23.2131	0.0851	0.30
86	0.0397	100.799	0.94	0.0967	0.1280	1.07
87	0.0399	25.7787	0.74	44.1112	0.1001	0.52
88	0.0725	12.0240	0.98	42.2628	42.7806	0.41
89	0.0379	14.2083	1.52	47.0630	173.346	0.54
90	0.0365	17.2239	1.40	34.4499	0.1482	0.93
91	0.0489	22.1306	1.97	30.1808	31.9556	0.45
92	0.0626	36.9332	3.46	2.6983	9.9051	0.32
93	0.0846	20.7650	1.58	30.7104	24.0171	0.38
94	34.6503	18.8284	45.11	25.7775	21.0921	0.18

Note: SP -spatial location, DP -depth.

Based on the results in Tables 5.16a,c we use $k = 1$ on the shelf and $k = 0.5$ on the slopes. The GC $b_{\min}$'s for these values of k are presented in Table 5.17 based on the "cleaned" data. Compared to the results in Table 5.15c, where $k = 0.65$ for both the shelf and slopes, the shelf bandwidths (with $k = 1$) tend to be smaller, and the slope bandwidths (with $k = 0.5$) tend to be larger. We use the bandwidths in Tables 5.17 for the robust smoothing used to assist in identifying type II catches.

We use robust weights to assist in identifying type II catches. Recall that the robust weights are given by

$$w_{ij}^r = \min \left[1, CA^{1/2} \left| \left(\frac{y_i - \bar{m}_j}{\bar{m}_j(1 + \bar{m}_j/k)} \right) - \alpha \right|^{-1} \right].$$

This weight will be small if $|y_i - \bar{m}_j|$ is large when weighted by the NB variance. In robust smoothing each observation has n robust weights associated with the estimation

Table 5.17. $GC_p b_{min}$'s with $k = 1$ on the shelf, and $k = 0.5$ on the slopes.

Year	3NO shelf region		3NO slope region	
	b_{SP}	b_{DP}	b_{SP}	b_{DP}
84	0.0410	4.9117	0.1060	0.1037
85	0.0257	0.0278	0.0395	0.0401
86	0.0250	0.0611	0.0966	0.1193
87	0.0178	14.1816	0.1137	0.0967
88	0.0332	12.0395	0.1399	0.0770
89	0.0281	17.8503	27.3479	0.0714
90	0.0267	28.5571	14.5688	0.1408
91	0.0396	16.3686	33.8927	0.0316
92	0.0462	23.8616	0.0552	0.0381
93	0.0628	19.9095	0.2029	0.0441
94	4.9736	0.3617	13.6359	0.0234

Note: SP -spatial location, DP -depth.

of the mean for the n observations in the sample. As discussed in Section 5.5.4, we summarize these robust weights by their kernel weighted average. The robust weights we report essentially describe the average variance-weighted discrepancy between y_i and $m_j = m(\mathbf{x}_j)$, where the average is basically over all j such that $\|\mathbf{x}_i - \mathbf{x}_j\|$ is small. Observations with small robust weights are “outliers” with respect to the smooth mean function model, and are type II candidates.

We choose $C = 5$ to identify type II catches because we wish to select only those catches that are clearly discordant with the SMFM. Inverse average robust weights are plotted in Figures 5.4a,b. For simplicity we do not plot weights > 0.9 . Two discordant catches in 1986 and 1991 are apparent. Another type II candidate is the large catch along the 3O/3PS boundary in 1987 (see Figure 4.1a), which has a robust weight of 0.34. The robust standardized residual for this catch is 8.70, which is large; hence, we feel this catch is discordant. In 1993 the large catch in the northern part of 3N had a robust weight of

0.52, which was considerably smaller than the next lowest weight of 0.82. We conclude that this catch is also discordant.

Isolated large catches in 1992, and in 3N in 1993 (see Figure 4.1b), did not have relatively small robust weights. The large catch in 1992 (2145 fish) was the second highest during 1984-1994. The problem here is related to the small number of samples in the region of this tow. The slope region around the large catch in 1992 is very restricted, and essentially only five tows along the 300 meter depth contour in this year can be used to determine if the large catch is an outlier. It is difficult to reliably detect outliers with such a small sample, but because of the magnitude of this catch, we conclude (subjectively) that it is type II. The large catches in 3O in 1993 are difficult to deal with as well. The robust weights for these catches are not small, at least in relation to weights observed for type II catches in other years. We assume that the cluster of three large catches in 3O represents a large component of cod that was not present in other years. We feel it is less reasonable than otherwise that such a concentration of cod existed in other years but was missed due to the random selection of tow stations. The catch to the southeast of the cluster of large catches does seem consistent with a type II catch, and we identify it as such. The robust weights for this catch are large because of the influence of the three large catches to the northwest.

The catches we have identified as type II are tabulated in Table 5.18 and plotted in Figure 5.5. The average catch for the type II tows is 1629 with a standard deviation of 580. Most of the type II catches have occurred along the southwest slope of the Grand Bank in 3O. The general notion of cod distribution in this region is that they occur in relatively dense aggregations in warm water along the slopes of the Grand Bank where they spawn in the winter, and then they migrate and disperse to the shelf region to feed in late spring and early summer (Templeman, 1962). The annual research surveys occur during this migration stage, and it is possible in any year that not all of the stock has

Table 5.18. Type II catches, Y . Depth is in meters. $\epsilon = (Y - \mu)Var(Y)^{-1/2}$.

Div.	Year	Y	$\hat{Y}$	$\tilde{Y}$	$\hat{\epsilon}$	$\tilde{\epsilon}$	w^r	Lat.	Long.	Depth
3O	87	2300	249.17	172.23	5.81	8.70	0.34	45.23	54.43	127
3O	92	2145	638.85	665.76	1.67	1.57	0.77	43.12	51.16	390
3O	86	1945	279.70	123.06	5.94	14.75	0.18	45.22	51.86	67
3O	91	1390	59.42	14.22	15.77	67.23	0.06	44.77	53.85	108
3O	93	1056	239.05	249.13	2.41	2.29	0.82	44.27	52.61	97
3N	93	938	168.85	175.97	3.22	3.06	0.52	45.53	48.44	323

spread out along the shelf of the bank. Consequently, the probability of getting a type II catch may be lower on the shelf than on the slopes where cod are suggested to aggregate more densely, because it is presumably the dense aggregations that give rise to type II catches. This suggests that the model for type II catches should be modified to allow for different mixture components on the slopes and the shelf of the bank. The data in Figure 5.5 support this in that four of the six type II catches occurred along the southwestern slope of the bank, while only one type II catch was identified along the southeast slope in 1993 and one the shelf in 1986.

An estimator of the probability of a tow resulting in a type II catch is the number of type II catches divided by the total number of tows in a region. Let τ_{II} denote this probability. We use the survey stratification scheme to define three mixture regions:

1. the Grand Bank shelf;
2. The southwestern Grand Bank slope in 3O;
3. The southeastern Grand Bank slope in 3N.

The shelf is defined as the area contained within strata 330-331, 338-340, 351-353, 360-362, 373-376 and 383 (see Figure 1.5). These are the strata where depths are less than 91

meters, except strata 339 in the interior of Division 3O where depths range between 92 and 183 meters (see Tables 3.1a,b). The southwest and southeast slope regions are defined as the remaining strata in 3O and 3N, respectively. We assume the probability of a type II catch is different in these three regions, but constant within a region. The total number of survey trawl sets taken on the southwestern slope during 1984-1994 is 385; hence, $\tau_{II}^{sw} = 4/385 = 0.0104$. Similarly, for the southeastern slope $\tau_{II}^{se} = 1/257 = 0.0039$ and for the shelf region $\tau_{II}^{sh} = 1/1270 = 0.00079$. These estimates suggest that it is approximately 13 times more likely that a catch is type II on the southwestern slope compared to the shelf region, and 5 times more likely on the southeastern slope. We recognize that these mixing proportions are based on very small sample sizes. In Chapter 6 we will investigate the significance of our inferences about these proportions.

The identification of type II catches is very much a subjective procedure even with the aid of the robust smoothing procedures. We could have subjectively identified type II catches without the robust results; however, these results have provided valuable information about those catches that are clearly discordant and those that are not, and this information was not readily apparent in the graphical analyses in Chapter 4. We could have also used a simple subjective “cutoff” criteria to identify type II catches (this was attempted in preliminary investigations); however, this approach would have selected some of the large catches in 1987 along the northern 3L boundary (see Figure 4.1a), and the large 3O catches in 1993 close to the 3Ps boundary (see Figure 4.1b). Such clusters of large catches, especially those in 1987, likely represent large concentrations of cod that we cannot assume to have existed every year and were missed due to the random allocation of tow stations. The robust smoothing procedures did not indicate that these clusters of catches were clearly discordant. In fact the cod in these catches may not be from the 3NO stock; however, the solution to this problem likely involves examining additional biological data measured from the catches, and this is beyond the scope of this thesis.

5.8. Smooth mean function model for type I catches

In the previous section we identified type II catches in the trawl survey data. As well, we estimated the mean of the type II catches, and the probability that a random catch is type II. This probability varies over three large spatial areas in 3NO. This is essentially all we require to model the type II catches. In this section we specify the smooth mean function model that we use for type I catches.

We follow the same procedures used in the previous section to select values for the Negative Binomial k parameter and bandwidths, b . The values of these parameters that we used to identify type II catches were based on “cleaned” data, which excluded type II catches, but also other catches as well. It seems appropriate now to compute these values again for all of the type I catches. The first step is to estimate k . We use the same methods as in Section 5.7. The results are similar to those in Tables 5.16a-c, and are not reported to save space. The across-years average estimate of k that minimizes (5.4.2) in Section 5.4 is 0.917 for the shelf region, and 0.466 for the slope regions. The values obtained using BIC are 0.863 and 0.404. The AIC average estimates of k are, again, large. Based on these results we use $k = 0.89$ for the shelf region, and $k = 0.44$ for the slope regions. As expected, these values are somewhat smaller than those obtained for the “cleaned” data in Section 5.7.

The GC $b_{\min}$ ’s based on these values for k are presented in Table 5.19. In many years the bandwidths are similar to those obtained for the “cleaned” data; however, substantial differences do exist in some years. The depth bandwidths on the shelf in 1984-85 and 1994 are quite different in Table 5.19 compared to Table 5.17. This is not much of a concern however, because depth does not vary greatly on the shelf, and a wide range of bandwidths produce similar fits. The spatial location bandwidths are quite different in 1994. The bandwidths in Table 5.19 for 1994 seem more appropriate in that they are more consistent with bandwidths in other years than those in Table 5.17. The values in

Table 5.19. $GC_p b_{min}$'s for type I catches
with $k = 0.89$ on the shelf,
and $k = 0.44$ on the slopes.

Year	3NO shelf region		3NO slope region	
	b_{SP}	b_{DP}	b_{SP}	b_{DP}
84	0.0466	0.0562	0.0734	0.0764
85	0.0220	17.1197	0.0380	0.0412
86	0.0228	0.0579	0.0525	0.1495
87	0.0156	15.0246	0.0570	0.0951
88	0.0362	16.8408	0.1234	0.0709
89	0.0242	10.0000	33.5390	0.0758
90	0.0276	25.7851	8.2627	0.1964
91	0.0334	12.9909	0.2421	0.0257
92	0.0414	24.6757	0.0583	0.0345
93	0.0514	28.4258	0.0347	0.0696
94	0.0915	26.9131	0.0955	0.0149

Note: SP -spatial location, DP -depth.

Table 5.17 for 1994 are suspect for reasons mentioned in Section 5.7. On the whole the results in Table 5.19 seem reasonable. These are the values that we use for constructing inferences about stock size in Chapter 6.

Predicted type I trawl catches, $\hat{y}_i$'s, at the sampled tow stations are presented in Figures 5.6a-b. The average of the $\hat{y}_i$'s are very similar to the average of all the y_i 's for years in which no type II catches have been identified (1982-3, 1988-90, 1994; compare Figures 5.6a,b and Figures 4.1a,b). The main difference is that the sample variances of the $\hat{y}_i$'s is less than the sample variance of the y_i 's, as expected. The averages are quite different in years with type II catches, and this demonstrates the effect these catches have on inferences. Standardized residuals are plotted in Figures 5.7a,b. The total deviance of the model fit to the type I catches is listed at the top of each panel, for each year, along with the deviance error degrees of freedom (df^{err} 's). The deviances exceed the df^{err} 's as

often as not, which indicates that the Negative Binomial model has adequately accounted for overdispersion in the data. Note that the deviances tend to exceed df^{err} 's in the earlier years in the time series, and this indicates a possible trend in k with time. This is not explored further. The number of standardized residuals that exceed two is usually less than 5%, which also indicates that the model fits the data reasonably well. In the next section we compare the nonparametric model we have developed with a parametric model that is commonly used to analyze trawl survey data.

5.9. Comparison with the stratum-effects model

A common model-based method used to analyze trawl survey data is to assume that catches are i.i.d. within strata, but that the means vary between strata. This is merely a fixed-effects NB regression model, where each stratum is an effect; that is,

$$E(Y_i|\mathbf{x}_i) = \mu_j \text{ if } \mathbf{x}_i \in \text{stratum } j.$$

We refer to this as the stratum-effects model. In this section we compare this approach for modelling type I catches with the nonparametric approach we have used. We focus our comparison on deviance, error degrees of freedom (df^{err} 's), and measures of prediction error.

Model deviances and prediction error measures are presented in Table 5.20. Several points are apparent:

1. The bandwidths we chose lead to more df^{err} 's compared to the stratum-effects model in most years. This essentially means that the nonparametric model has fewer "parameters" than the stratum-effect model. For all years our nonparametric model has 32 more df^{err} 's than the stratum-effects model.

Table 5.20. Deviances and prediction error measures for Type I catches.
 Smoother minus stratum-effects estimates are in parentheses.
 The NB scale parameter is $k = 0.89$ on the shelf, and 0.44
 on the slopes.

Year	df^{err}	Deviance	CV	GCV	GC_p
1984	77.8 (-3.16)	87.72 (-8.11)	1.65 (-45.1)	1.68 (-0.01)	1.41 (-0.02)
1985	120.2 (-22.8)	148.34 (-58.5)	1.51 (-35.1)	1.83 (0.03)	1.48 (-0.07)
1986	155.5 (-11.5)	145.39 (-42.2)	1.14 (-3.31)	1.21 (-0.14)	1.18 (-0.10)
1987	128.7 (-26.3)	149.21 (-100)	1.72 (-3.05)	1.71 (-0.26)	1.43 (-0.25)
1988	136.6 (10.58)	148.03 (11.39)	1.16 (-10.9)	1.28 (-0.11)	1.22 (-0.06)
1989	168.8 (8.77)	160.07 (0.61)	1.19 (-18.1)	1.10 (-0.12)	1.09 (-0.09)
1990	152.6 (11.60)	125.96 (-18.6)	0.85 (-7.82)	0.95 (-0.33)	0.98 (-0.24)
1991	182.8 (23.80)	160.18 (26.05)	0.98 (-14.8)	0.99 (-0.11)	1.00 (-0.10)
1992	146.7 (11.68)	101.39 (13.06)	0.81 (-16.8)	0.86 (-0.02)	0.95 (-0.06)
1993	119.1 (10.07)	101.38 (24.99)	0.94 (-3.05)	1.10 (0.11)	1.11 (0.03)
1994	114.0 (18.97)	70.15 (22.38)	0.66 (-33.9)	0.79 (0.02)	0.93 (-0.11)
total	1502.68 (31.68)	1502.7 (-24.54)	1.14 (-16.2)	1.18 (-0.17)	1.16 (-0.10)

2. The nonparametric model accounts for more Negative Binomial variation than the stratum-effects model in five of eleven years. The total deviance over all years for the stratum-effects model exceeds that for the nonparametric model by almost 25.
3. The nonparametric model has greater predictive capability for most years when measured using CV , GCV and GC_p . The CV 's for the stratum-effects model are always substantially greater than for the smoother.

This last point about CV 's is related to boundary estimates of μ ; that is, in some strata only one tow encountered cod, and the mle of μ for the strata is zero when this tow is deleted. In this case, when computing CV , the deviance for the tow with a cod catch is infinity, because $\hat{\mu} = 0$. In our analyses we replaced the zero mle with a small number; however, the deviances are still very large, and this leads to the large increases in CV 's.

The results in Table 5.20 show that over all years the nonparametric model explains more of the variation in trawl catches than a simple stratum-effects model, and uses fewer df^{err} 's to do so. The differences are not enormous however, and this suggests that the stratification scheme is related to variations in cod abundance.

5.10. Summary

The results in this chapter suggest that spatial location and tow depth are the most important predictive covariates in a smooth mean function model for research trawl type I catches of cod throughout 1984-94. The importance of these two covariates is different on the shelf of the Grand Bank than on the slopes. Beyond their association with spatial location and depth, tow temperature and within-survey timing, and wind speed during tows, are of much lesser importance in explaining variations in trawl catches. These covariates do not appear useful for modelling trawl catches. Hence, the nonparametric variable selection procedures we have developed in this chapter have reduced the number of covariates to consider for modelling purposes from five to two. This results in a more parsimonious model, and a model more useful for prediction purposes. The nonparametric model we have constructed performs better in most years than the commonly used stratum-effects model, both in terms of explaining variation in type I trawl catches and in predicting catches.

We have identified six type II catches throughout 1984-94. The average of these catches is much greater than the average of the type I catches. The analyses in this chapter suggest a further refinement to the mixture model we use for trawl catches. The spatial location of the type II catches suggests that the probability that a random catch is type II varies geographically. We choose to model this probability separately for three broad geographical areas. This assumption is tentative however, because of the small number of type II catches involved. In our inferential analyses in the next chapter we will

investigate this assumption further. Beyond the varying mixing probabilities, we find no evidence that the distribution of type II catches is not homogeneous; hence, we assume they are i.i.d.

On a methodological level we find that asymptotically equivalent measures of prediction error can produce substantially different small sample results. We find GC_p most useful for nonparametric variable selection with the trawl catch data. Bandwidths selected using CV resulted in under-smoothing relative to bandwidths that minimized the GC_p criterion. The under-smoothing did not result in “significant” decreases in deviance, and as such was undesirable. Variable selection using CV resulted in less parsimonious models compared to GC_p . The GCV criterion behaved erratically, and in some years resulted in gross under-smoothing with almost no error degrees of freedom for inferences.

Robust M-estimation for Negative Binomial data is quite feasible in that outlier resistant estimation methods for the mean of Negative Binomial samples can be constructed that have reasonable efficiency relative to maximum likelihood estimators. Robust local likelihood estimators of a smooth mean function model are relatively simple to construct using robust M-estimation theory. The methods worked fairly well for the trawl survey data, although at times difficulties arose with their use due to the small sample sizes within smoothing “neighborhoods”. Methods for robust bandwidth selection are necessary for contaminated data, although bandwidth selection for a robust smoother is computationally difficult for several or more smoothing parameters.

In this chapter we have specified the details of a stochastic mixture model for two types of trawl catches that involves a smooth mean function model for type I trawl catches and a simple model for type II catches. In the next chapter we show how to use this model to estimate trawlable stock size. Procedures will be also developed to give confidence intervals for stock size.

5.11. Appendix: Tables

Table 5.3a. Cross-validation (CV) variable selection results. The Negative Binomial scale parameter is $k = 0.65$ for all subsets. The bandwidths are for SP -spatial location, DP -depth, TP -temperature, TM -time, WD -wind speed.

Shelf - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.038	0.025	0.020	0.019	0.032	0.031	0.022	0.027	0.029	0.039	0.041
b_{DP}	0.033	0.045	0.043	1.258	0.143	0.034	0.195	0.097	24.56	23.61	1.323
b_{TP}	0.308	0.092	0.087	0.064	0.061	47.32	0.135	16.28	23.69	23.64	1.732
b_{TM}	0.134	0.095	0.124	32.07	0.033	28.75	0.023	17.54	17.63	0.056	0.323
b_{WD}	0.031	0.094	24.41	0.118	31.49	8.510	0.159	15.40	0.084	22.24	0.013
CV	1.122	0.996	1.098	1.207	0.759	0.953	0.642	0.707	0.539	0.580	0.144
Shelf - Optimal variable subset with $\min(CV)/CV \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.038	0.028	0.020	0.020	0.034	0.031	0.022	0.027	0.030	0.035	0.040
b_{DP}	0.038	0.024	0.044	0.103	0.109	0.034	0.306	0.097	23.92	17.72	12.78
b_{TP}	.	0.101	.	0.072	.	.	.	.	.	.	.
b_{TM}	.	.	0.142	.	0.032	.	0.023	.	.	.	.
b_{WD}	0.031	.	.	.	.	.	.	.	.	.	0.018
CV	1.181	1.012	1.143	1.241	0.790	0.953	0.664	0.707	0.543	0.605	0.146
Slope - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.113	0.024	0.038	0.026	0.046	0.051	0.114	0.039	0.022	0.060	0.086
b_{DP}	0.035	0.046	0.085	0.059	0.047	0.038	0.029	0.029	0.028	0.014	0.010
b_{TP}	38.13	49.73	0.152	0.222	0.591	39.74	39.00	0.089	0.064	0.191	25.65
b_{TM}	0.723	3.843	0.019	0.118	38.36	0.183	46.01	30.10	37.97	0.121	7.878
b_{WD}	0.073	0.259	2.153	48.30	0.091	39.65	0.972	0.102	0.178	33.43	7.238
CV	1.136	1.212	1.355	1.813	1.092	0.787	0.725	1.649	1.370	2.693	1.124
Slope - Optimal variable subset with $\min(CV)/CV \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.113	0.024	0.038	0.028	0.046	0.050	0.114	0.039	0.023	0.061	0.086
b_{DP}	0.035	0.049	0.086	0.062	0.048	0.031	0.029	0.029	0.029	0.014	0.010
b_{TP}	.	.	.	.	.	.	.	0.089	0.066	0.114	.
b_{TM}	.	.	0.019	.	.	.	.	.	.	.	.
b_{WD}	0.073	.	.	.	0.092	.	.	0.102	.	.	.
CV	1.138	1.279	1.367	1.836	1.094	0.809	0.725	1.649	1.399	2.755	1.124

Table 5.3b. Same as Table 5.3a, except for generalized cross-validation (*GCV*).

Shelf - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.058	0.030	0.024	0.020	0.042	0.024	0.026	0.030	0.028	0.040	0.009
b_{DP}	0.145	31.07	0.059	31.28	22.84	18.41	25.10	31.36	23.99	23.32	0.070
b_{TP}	41.82	0.092	38.88	32.19	0.267	22.41	27.19	38.84	23.91	23.60	0.091
b_{TM}	41.74	32.95	0.183	31.58	21.54	19.38	0.622	34.01	25.35	0.242	0.003
b_{WD}	42.69	31.49	38.78	31.45	23.16	21.00	23.90	31.84	24.07	23.60	0.130
<i>GCV</i>	1.231	1.320	1.120	1.318	0.949	0.931	0.856	0.778	0.564	0.652	0.000
Shelf - Optimal variable subset with $\min(GCV)/GCV \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.058	0.027	0.024	0.020	0.040	0.024	0.026	0.030	0.028	0.041	0.003
b_{DP}	0.145	24.53	0.059	23.79	23.73	14.19	25.21	10.95	24.39	23.47	0.070
b_{TP}	.	.	.	.	.	.	.	.	.	.	.
b_{TM}	.	.	0.183	.	.	.	.	.	.	.	.
b_{WD}	.	.	.	.	.	.	.	.	.	.	0.130
<i>GCV</i>	1.231	1.340	1.120	1.318	0.959	0.931	0.857	0.778	0.564	0.662	0.000
Slope - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.012	0.003	0.045	0.033	28.70	40.94	0.172	0.402	0.005	53.59	0.037
b_{DP}	0.007	0.012	0.117	52.46	28.67	0.067	0.034	0.024	0.009	0.031	56.47
b_{TP}	0.001	1.000	55.67	51.37	0.140	0.130	33.76	0.103	0.009	0.090	57.96
b_{TM}	1.000	0.018	60.33	0.114	32.37	40.87	43.01	36.94	0.050	0.039	57.66
b_{WD}	0.020	0.080	57.02	52.51	0.153	41.82	40.29	36.33	0.049	56.09	0.275
<i>GCV</i>	0.000	0.003	1.276	2.375	1.787	1.266	0.958	1.493	0.000	2.761	2.028
Slope - Optimal variable subset with $\min(GCV)/GCV \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.008	0.001	0.045	0.032	7.589	31.15	0.172	0.402	0.007	54.56	0.036
b_{DP}	0.007	0.000	0.117	54.28	7.356	0.070	0.034	0.024	0.006	0.031	48.90
b_{TP}	0.003	.	.	.	.	.	.	0.103	0.004	0.090	.
b_{TM}	.	0.069	.	.	.	.	.	.	.	0.039	.
b_{WD}	.	.	.	.	.	.	.	.	.	.	.
<i>GCV</i>	0.000	0.000	1.276	2.444	1.866	1.312	0.958	1.493	0.000	2.761	2.099

Table 5.3c. Same as Table 5.3a, except for generalized C_p (GC_p).

Shelf - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.056	0.029	0.026	0.019	0.047	0.028	0.031	0.036	0.044	0.054	31.41
b_{DP}	0.141	23.75	0.070	23.63	24.45	22.10	23.46	23.96	21.84	30.24	31.49
b_{TP}	31.69	0.090	32.39	24.53	0.286	17.84	23.29	23.97	21.21	30.96	0.173
b_{TM}	31.19	24.97	0.187	23.84	26.57	16.98	1.595	24.26	20.26	0.503	0.047
b_{WD}	32.40	23.90	32.73	23.80	23.96	18.44	23.38	24.42	23.26	30.98	31.14
GC_p	1.189	1.244	1.124	1.241	0.969	0.971	0.922	0.843	0.683	0.746	0.387
Shelf - Optimal variable subset with $\min(GC_p)/GC_p \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.056	0.025	0.026	0.019	0.045	0.028	0.031	0.036	0.044	0.054	0.093
b_{DP}	0.141	20.16	0.070	23.91	19.75	17.56	23.56	18.14	22.17	22.95	30.80
b_{TP}	.	.	.	.	.	.	.	.	.	.	.
b_{TM}	.	.	0.187	.	.	.	.	.	.	.	.
b_{WD}	.	.	.	.	.	.	.	.	.	.	.
GC_p	1.189	1.261	1.124	1.241	0.976	0.971	0.922	0.843	0.683	0.747	0.395
Slope - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.075	0.030	0.046	0.029	0.069	40.06	16.75	27.67	0.051	0.101	0.086
b_{DP}	0.032	0.038	0.122	0.090	0.054	0.064	0.147	0.010	0.032	0.017	0.011
b_{TP}	40.49	45.17	42.32	31.57	43.57	0.117	0.311	0.096	0.246	0.083	44.10
b_{TM}	42.92	42.31	45.36	30.95	42.66	40.11	18.60	31.05	0.165	0.038	43.61
b_{WD}	42.66	42.28	43.03	27.37	42.43	40.78	16.34	27.02	42.09	50.91	40.64
GC_p	1.605	1.611	1.218	1.734	1.561	1.226	0.984	1.373	1.454	1.749	1.528
Slope - Optimal variable subset with $\min(GC_p)/GC_p \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.075	0.030	0.046	0.029	0.069	30.47	23.32	29.97	0.044	0.101	0.086
b_{DP}	0.032	0.038	0.122	0.090	0.054	0.063	0.150	0.010	0.022	0.017	0.011
b_{TP}	.	.	.	.	.	.	.	0.096	.	0.083	.
b_{TM}	.	.	.	.	.	.	.	.	.	0.038	.
b_{WD}	.	.	.	.	.	.	.	.	.	.	.
GC_p	1.605	1.611	1.218	1.734	1.561	1.281	1.004	1.373	1.510	1.749	1.528

Table 5.4a. Cross-validation (CV) variable selection results. The Negative Binomial scale parameter is $k = 5$ for all subsets. The bandwidths are for SP -spatial location, DP -depth, TP -temperature, TM -time, WD -wind speed.

Shelf - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.037	0.025	0.021	0.020	0.032	0.037	0.020	0.023	0.037	0.041	0.043
b_{DP}	0.031	0.056	0.045	0.060	0.145	0.044	0.296	0.092	0.054	31.39	1.043
b_{TP}	0.262	0.085	0.102	0.069	0.065	0.035	0.136	0.483	32.88	31.27	1.497
b_{TM}	0.136	0.093	0.129	54.30	0.034	34.89	0.132	32.84	33.60	0.059	0.422
b_{WD}	0.031	0.085	41.39	0.129	43.24	33.48	0.179	33.40	32.21	31.82	0.013
CV	6.278	3.861	5.360	6.422	2.985	3.800	2.005	1.591	1.000	0.977	0.211
Shelf - Optimal variable subset with $\min(CV)/CV \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.038	0.027	0.022	0.021	0.034	0.032	0.018	0.024	0.037	0.041	0.042
b_{DP}	0.038	0.042	0.047	0.056	0.117	0.032	15.65	0.091	0.054	23.58	16.44
b_{TP}	.	0.093	.	0.074	.	.	.	.	.	.	.
b_{TM}	.	.	0.142	.	0.034	.	0.050	.	.	0.059	.
b_{WD}	0.033	.	.	.	.	.	.	.	.	.	0.019
CV	6.667	4.070	5.461	6.557	3.162	3.852	2.064	1.592	1.000	0.977	0.212
Slope - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.109	0.024	0.024	0.024	0.039	0.052	0.110	0.156	0.059	0.049	0.084
b_{DP}	0.036	0.045	0.031	0.053	0.051	0.039	0.028	0.010	0.033	0.012	0.010
b_{TP}	53.61	77.93	39.11	0.148	0.416	53.29	53.92	0.062	0.122	0.191	0.117
b_{TM}	1.054	68.54	0.025	0.098	52.83	0.211	64.57	0.084	0.059	0.104	56.64
b_{WD}	0.068	0.237	34.71	78.75	0.106	53.20	0.482	0.182	0.050	63.44	51.54
CV	3.312	5.441	7.005	9.631	4.061	2.985	3.010	7.536	7.156	14.07	3.826
Slope - Optimal variable subset with $\min(CV)/CV \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.109	0.025	0.024	0.027	0.039	0.050	0.109	0.137	0.022	0.048	0.084
b_{DP}	0.036	0.049	0.031	0.059	0.051	0.032	0.028	0.010	0.030	0.012	0.011
b_{TP}	.	.	.	.	.	.	.	0.062	0.066	0.099	.
b_{TM}	.	.	0.025	.	.	.	.	.	.	.	.
b_{WD}	0.068	.	.	.	0.106	.	.	.	.	.	.
CV	3.314	5.733	7.005	9.990	4.085	2.996	3.013	7.869	7.537	14.58	4.040

Table 5.4b. Same as Table 5.4a, except for generalized cross-validation (*GCV*).

Shelf - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.043	0.027	0.023	0.016	0.036	0.020	0.024	0.019	0.026	0.039	0.002
b_{DP}	0.056	34.70	0.055	42.78	18.86	23.71	29.91	23.49	14.47	23.69	0.070
b_{TP}	0.781	0.091	38.89	52.41	0.148	23.56	0.314	23.77	27.24	23.62	0.099
b_{TM}	0.086	34.10	0.165	51.22	21.55	23.13	0.254	24.03	16.43	0.224	0.059
b_{WD}	73.33	32.03	39.17	49.24	13.39	24.00	28.60	22.53	13.57	23.69	1.000
<i>GCV</i>	5.976	4.624	3.732	6.337	3.595	3.161	2.542	1.635	0.953	1.036	0.000
Shelf - Optimal variable subset with $\min(GCV)/GCV \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.048	0.024	0.023	0.016	0.035	0.020	0.024	0.019	0.026	0.041	0.003
b_{DP}	0.068	32.86	0.055	32.62	16.28	24.70	29.69	24.29	26.45	23.48	0.070
b_{TP}	.	.	.	.	.	.	.	.	.	.	.
b_{TM}	.	.	0.165	.	.	.	.	.	.	.	0.060
b_{WD}	.	.	.	.	.	.	.	.	.	.	.
<i>GCV</i>	6.014	4.821	3.732	6.337	3.710	3.161	2.620	1.635	0.953	1.071	0.000
Slope - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.070	0.014	0.044	0.029	67.66	0.065	0.142	0.135	0.057	0.494	0.090
b_{DP}	0.025	0.013	0.104	51.73	71.04	0.049	0.031	0.020	0.034	0.023	0.012
b_{TP}	0.046	0.060	69.76	31.06	0.044	73.75	51.46	0.090	0.089	0.089	57.52
b_{TM}	0.262	0.017	78.68	0.089	69.81	79.33	57.31	73.82	0.141	0.039	56.34
b_{WD}	68.33	0.011	73.72	58.60	0.177	72.46	50.44	74.04	97.52	99.86	30.85
<i>GCV</i>	5.859	0.017	5.729	11.20	6.295	4.142	3.498	5.735	6.311	12.31	5.731
Slope - Optimal variable subset with $\min(GCV)/GCV \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.101	0.000	0.044	0.030	48.61	0.065	0.142	0.135	0.057	0.494	0.090
b_{DP}	0.045	0.011	0.104	61.19	42.08	0.049	0.031	0.020	0.034	0.023	0.012
b_{TP}	.	.	.	.	0.029	.	.	0.090	0.089	0.089	.
b_{TM}	.	0.002	.	.	.	.	.	.	0.141	0.039	.
b_{WD}	.	.	.	.	.	.	.	.	.	.	.
<i>GCV</i>	5.940	0.000	5.729	11.74	6.488	4.142	3.498	5.735	6.311	12.31	5.731

Table 5.4c. Same as Table 5.4a, except for Generalized C_p (GC_p).

Shelf - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.020	0.024	0.021	0.007	0.015	0.008	0.014	0.017	0.031	0.043	0.087
b_{DP}	0.009	0.010	0.016	0.055	0.035	8.989	17.71	17.59	22.99	22.72	31.72
b_{TP}	0.174	12.28	0.113	0.009	13.38	8.773	17.23	18.44	21.52	23.38	31.92
b_{TM}	0.145	0.166	0.150	0.163	0.008	1.759	0.062	19.78	23.90	0.251	30.82
b_{WD}	42.87	0.005	0.035	24.72	12.64	8.837	0.432	15.78	23.62	23.46	32.00
GC_p	1.894	1.938	1.907	1.958	1.874	1.797	1.770	1.394	1.005	1.053	0.472
Shelf - Optimal variable subset with $\min(GC_p)/GC_p \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.020	0.017	0.002	0.000	0.009	0.008	0.014	0.017	0.031	0.043	0.087
b_{DP}	0.002	0.006	0.031	0.097	0.009	10.46	1.000	1.000	23.23	22.91	24.26
b_{TP}	.	.	.	.	.	.	.	.	.	.	.
b_{TM}	.	.	.	.	.	.	0.062	.	.	.	.
b_{WD}	.	.	.	.	.	.	.	.	.	.	.
GC_p	1.921	2.040	1.999	2.000	1.941	1.798	1.772	1.394	1.005	1.073	0.472
Slope - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.061	0.018	0.020	0.007	0.046	0.029	0.050	0.015	0.010	0.030	0.013
b_{DP}	0.019	0.013	0.015	12.47	0.019	0.027	0.010	0.015	0.016	0.018	0.009
b_{TP}	0.010	31.77	0.056	0.034	0.081	0.102	43.36	0.096	33.17	0.022	34.13
b_{TM}	0.268	0.734	17.06	9.920	29.98	41.13	43.52	0.081	0.050	0.332	33.44
b_{WD}	32.12	0.263	19.02	0.110	0.097	32.78	0.541	45.75	29.81	43.04	33.45
GC_p	1.837	1.805	1.939	1.964	1.928	1.849	1.806	1.900	1.876	1.875	1.895
Slope - Optimal variable subset with $\min(GC_p)/GC_p \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.044	0.017	0.004	0.001	0.017	0.020	0.053	0.004	0.009	0.030	0.013
b_{DP}	0.020	0.009	0.022	1.000	0.036	0.026	0.008	0.021	0.015	0.017	0.009
b_{TP}	0.010	.	.	.	.	.	.	.	.	0.022	.
b_{TM}	.	.	.	.	.	.	.	.	.	.	.
b_{WD}	.	.	.	.	.	.	.	.	.	.	.
GC_p	1.872	1.809	1.979	2.000	1.949	1.933	1.807	1.967	1.919	1.889	1.895

Table 5.15a. Cross-validation (CV) variable selection results. The Negative Binomial scale parameter is $k = 0.65$ for all subsets. The bandwidths are for SP -spatial location, DP -depth, TP -temperature, TM -time, WD -wind speed.

Shelf - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.038	0.028	0.017	0.019	0.034	0.023	0.022	0.030	0.032	0.046	.
b_{DP}	0.031	0.025	0.040	0.021	0.077	0.054	0.240	0.126	18.95	13.51	.
b_{TP}	0.303	0.091	0.079	0.105	32.98	32.74	0.156	24.63	27.43	0.135	.
b_{TM}	0.126	0.104	0.083	18.22	0.069	0.814	0.112	26.42	0.557	0.058	.
b_{WD}	0.030	0.125	0.070	0.122	0.189	23.45	0.197	0.535	0.088	0.059	.
CV	0.649	1.005	0.630	0.904	0.728	0.656	0.655	0.661	0.512	0.474	.
Shelf - Optimal variable subset with $\min(CV)/CV \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.038	0.028	0.019	0.018	0.034	0.023	0.022	0.030	0.033	0.048	.
b_{DP}	0.031	0.024	0.041	0.038	0.081	0.053	0.519	0.120	23.68	0.099	.
b_{TP}	0.303	0.101	0.088	.	.	.	.	.	.	.	.
b_{TM}	0.126	.	.	.	0.068	.	0.098	.	.	0.081	.
b_{WD}	0.030	.	.	.	.	.	.	.	.	.	.
CV	0.649	1.025	0.657	0.948	0.732	0.656	0.665	0.662	0.517	0.494	.
Slope - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.105	0.070	0.080	0.030	0.054	0.051	0.125	0.193	0.042	0.087	0.538
b_{DP}	0.062	0.015	0.070	0.085	0.052	0.038	0.032	0.033	0.032	0.026	0.025
b_{TP}	87.40	36.30	0.365	0.199	53.13	39.74	25.47	0.069	0.064	0.240	32.05
b_{TM}	70.64	0.066	0.135	27.86	45.31	0.183	32.88	0.117	0.097	0.655	0.213
b_{WD}	0.298	0.063	36.50	0.156	0.106	39.65	33.18	32.23	0.513	44.61	37.81
CV	1.036	1.154	0.967	1.168	1.146	0.787	0.838	1.111	0.902	0.990	0.995
Slope - Optimal variable subset with $\min(CV)/CV \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.107	0.070	0.073	0.044	0.054	0.050	0.125	14.94	0.042	0.045	0.617
b_{DP}	0.055	0.015	0.072	0.090	0.052	0.031	0.032	0.033	0.032	0.045	0.025
b_{TP}	.	.	.	0.209	.	.	.	.	0.065	.	.
b_{TM}	.	0.066	.	.	.	.	.	.	0.097	.	.
b_{WD}	.	0.063	.	.	0.106	.	.	.	.	.	.
CV	1.044	1.154	0.970	1.189	1.146	0.809	0.838	1.134	0.903	1.024	1.001

Table 5.15b. Same as Table 5.15a, except for generalized cross-validation (*GCV*).

Shelf - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.055	0.032	0.025	0.023	0.040	0.027	0.026	0.036	0.032	0.051	0.029
b_{DP}	30.86	30.40	0.062	19.34	9.674	13.80	19.00	15.14	14.75	12.08	0.049
b_{TP}	26.86	0.090	20.61	36.61	0.314	36.18	28.03	29.44	16.47	0.302	0.109
b_{TM}	30.25	32.36	0.828	43.10	17.41	8.718	0.589	40.26	0.353	0.129	0.089
b_{WD}	30.34	30.71	20.96	23.55	23.59	30.80	36.30	24.59	32.95	34.35	0.101
<i>GCV</i>	1.098	1.322	0.967	1.154	0.892	0.777	0.859	0.720	0.541	0.525	0.000
Shelf - Optimal variable subset with $\min(GCV)/GCV \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.055	0.028	0.025	0.023	0.037	0.027	0.026	0.036	0.031	0.051	0.032
b_{DP}	17.65	24.15	0.062	13.96	18.33	18.32	23.28	8.164	26.01	12.91	0.038
b_{TP}	.	.	.	.	.	.	.	.	.	.	.
b_{TM}	.	.	.	.	.	.	.	.	.	0.136	0.081
b_{WD}	.	.	.	.	.	.	.	.	.	.	.
<i>GCV</i>	1.098	1.347	0.967	1.154	0.899	0.777	0.860	0.720	0.542	0.528	0.000
Slope - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.007	0.004	0.075	67.03	0.312	40.94	0.276	33.67	0.007	0.173	34.29
b_{DP}	0.007	0.010	0.097	0.020	14.16	0.067	0.068	0.031	0.009	0.043	0.022
b_{TP}	0.004	0.280	36.96	0.508	10.02	0.130	0.340	0.418	0.005	56.49	38.94
b_{TM}	1.333	0.004	45.23	0.202	14.06	40.87	27.41	38.95	1.000	53.83	29.40
b_{WD}	0.048	0.130	42.82	56.47	16.72	41.82	1.799	32.20	1.000	56.21	39.10
<i>GCV</i>	0.000	0.001	0.984	1.309	1.598	1.266	0.889	1.173	0.000	1.311	1.033
Slope - Optimal variable subset with $\min(GCV)/GCV \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.008	0.004	0.075	55.41	0.153	31.15	0.187	37.44	0.007	0.173	30.77
b_{DP}	0.007	0.010	0.097	0.020	0.083	0.070	0.037	0.031	0.009	0.043	0.022
b_{TP}	0.003	.	.	.	.	.	.	.	0.005	.	.
b_{TM}	.	0.004	.	0.202	.	.	.	.	.	.	.
b_{WD}	.	.	.	.	.	.	.	.	.	.	.
<i>GCV</i>	0.000	0.010	0.984	1.315	1.606	1.312	0.913	1.184	0.000	1.311	1.033

Table 5.15c. Same as Table 5.15a, except for Generalized C_p (GC_p).

Shelf - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.058	0.031	0.027	0.024	0.049	0.033	0.031	0.043	0.050	0.073	22.84
b_{DP}	31.52	23.19	18.74	21.93	16.49	8.437	19.85	15.57	15.70	20.67	22.32
b_{TP}	28.74	0.087	20.95	34.81	0.324	27.44	32.65	29.45	26.32	25.10	22.32
b_{TM}	30.99	24.79	21.09	38.42	0.718	37.43	1.314	39.45	21.88	0.170	0.195
b_{WD}	32.13	23.54	18.23	26.50	32.41	18.75	37.68	26.64	38.99	34.14	23.05
GC_p	1.092	1.246	1.010	1.137	0.923	0.841	0.924	0.779	0.649	0.633	0.079
Shelf - Optimal variable subset with $\min(GC_p)/GC_p \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.058	0.026	0.027	0.024	0.047	0.033	0.031	0.043	0.050	0.066	23.16
b_{DP}	17.95	19.08	19.09	20.55	21.40	18.60	25.41	21.45	23.25	19.90	22.81
b_{TP}	.	.	.	.	.	.	.	.	.	.	.
b_{TM}	.	.	.	.	.	.	.	.	.	.	0.195
b_{WD}	.	.	.	.	.	.	.	.	.	.	.
GC_p	1.092	1.268	1.010	1.137	0.929	0.841	0.924	0.779	0.649	0.646	0.079
Slope - All Variables											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.099	0.035	0.087	0.113	0.108	40.06	22.36	33.60	0.050	0.149	30.03
b_{DP}	0.119	0.038	0.107	0.098	0.062	0.064	0.118	0.031	0.035	0.043	0.023
b_{TP}	0.424	49.05	37.91	0.356	42.33	0.117	0.303	0.407	37.40	45.27	32.06
b_{TM}	55.04	50.44	44.69	0.132	40.92	40.11	44.30	35.07	31.90	44.00	25.99
b_{WD}	26.13	26.42	42.81	41.68	41.97	40.78	0.426	30.32	38.84	43.77	34.66
GC_p	1.192	1.597	1.025	1.373	1.408	1.226	0.912	1.154	1.367	1.241	1.037
Slope - Optimal variable subset with $\min(GC_p)/GC_p \geq 95\%$											
Year	84	85	86	87	88	89	90	91	92	93	94
b_{SP}	0.104	0.035	0.087	0.100	0.108	30.47	33.78	35.81	0.050	0.149	31.88
b_{DP}	0.094	0.038	0.107	0.089	0.062	0.063	0.120	0.031	0.035	0.043	0.023
b_{TP}	.	.	.	.	.	.	.	.	.	.	.
b_{TM}	.	.	.	.	.	.	.	.	.	.	.
b_{WD}	.	.	.	.	.	.	.	.	.	.	.
GC_p	1.199	1.597	1.025	1.432	1.408	1.281	0.947	1.164	1.367	1.241	1.037

5.12. Appendix: Figures

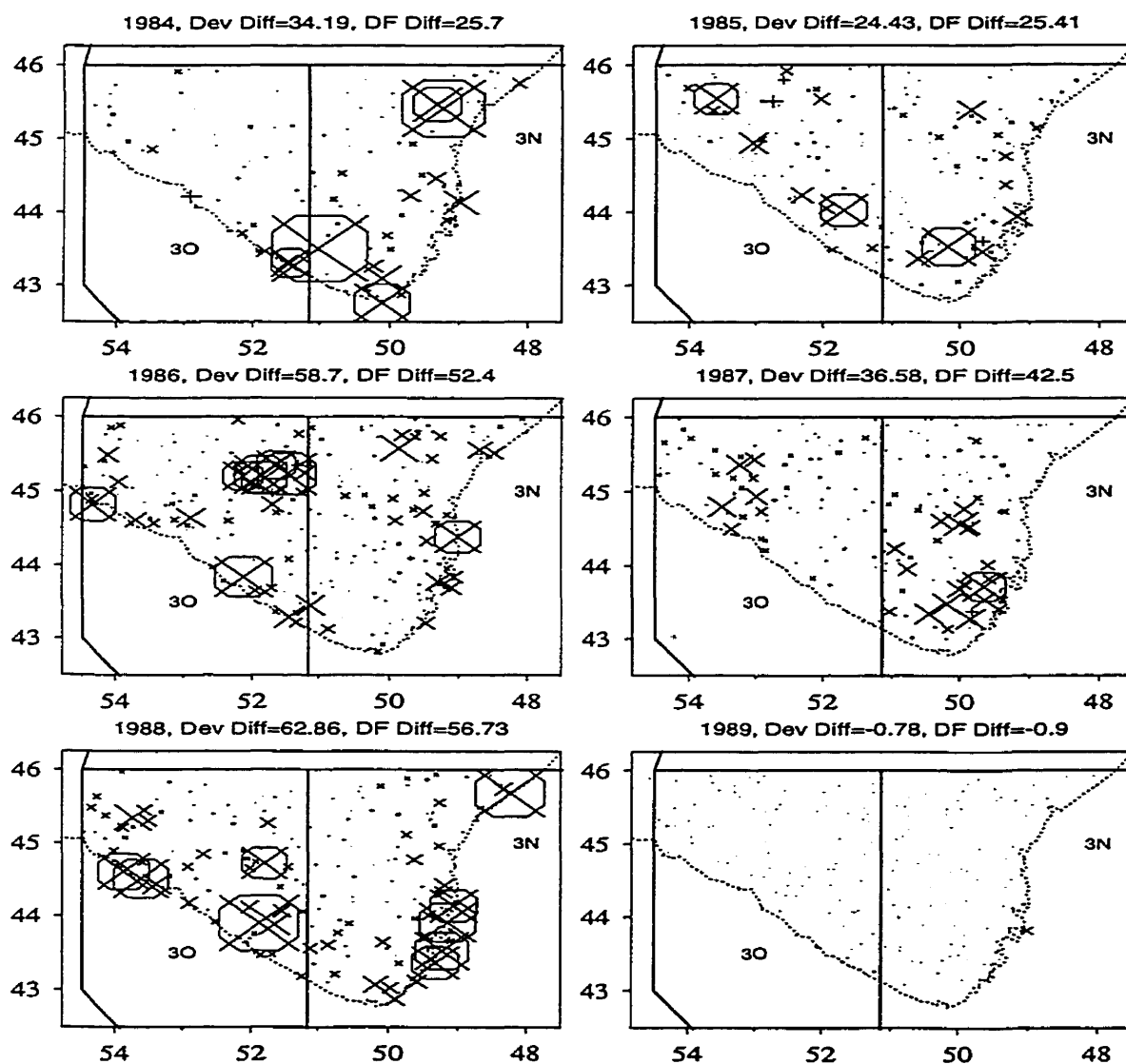


Figure 5.1a: Difference between the $CV b_{\min}$ per observation deviance for the model with all covariates and the model with spatial location and tow depth.

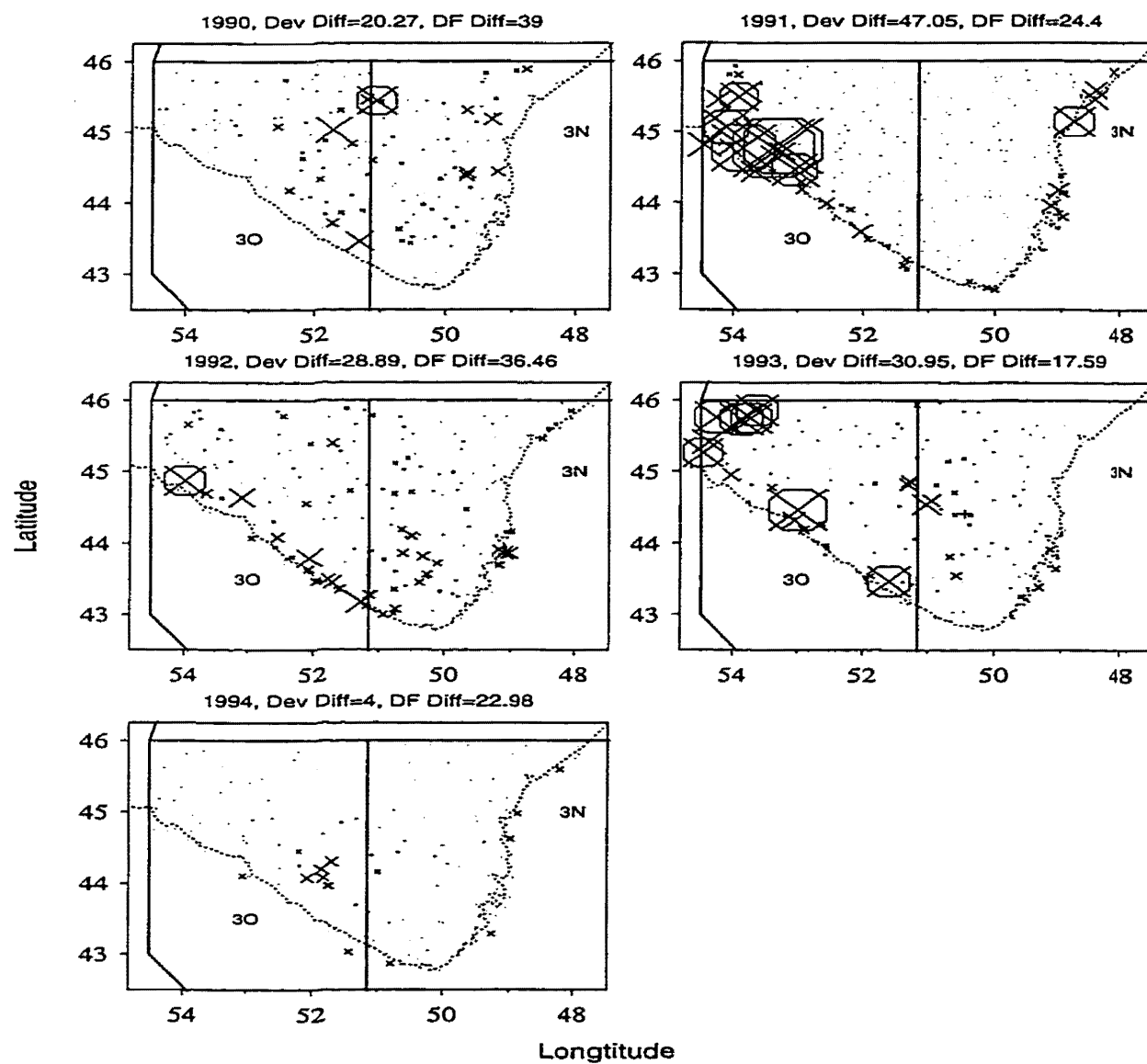


Figure 5.1b: Same as Figure 5.1a.

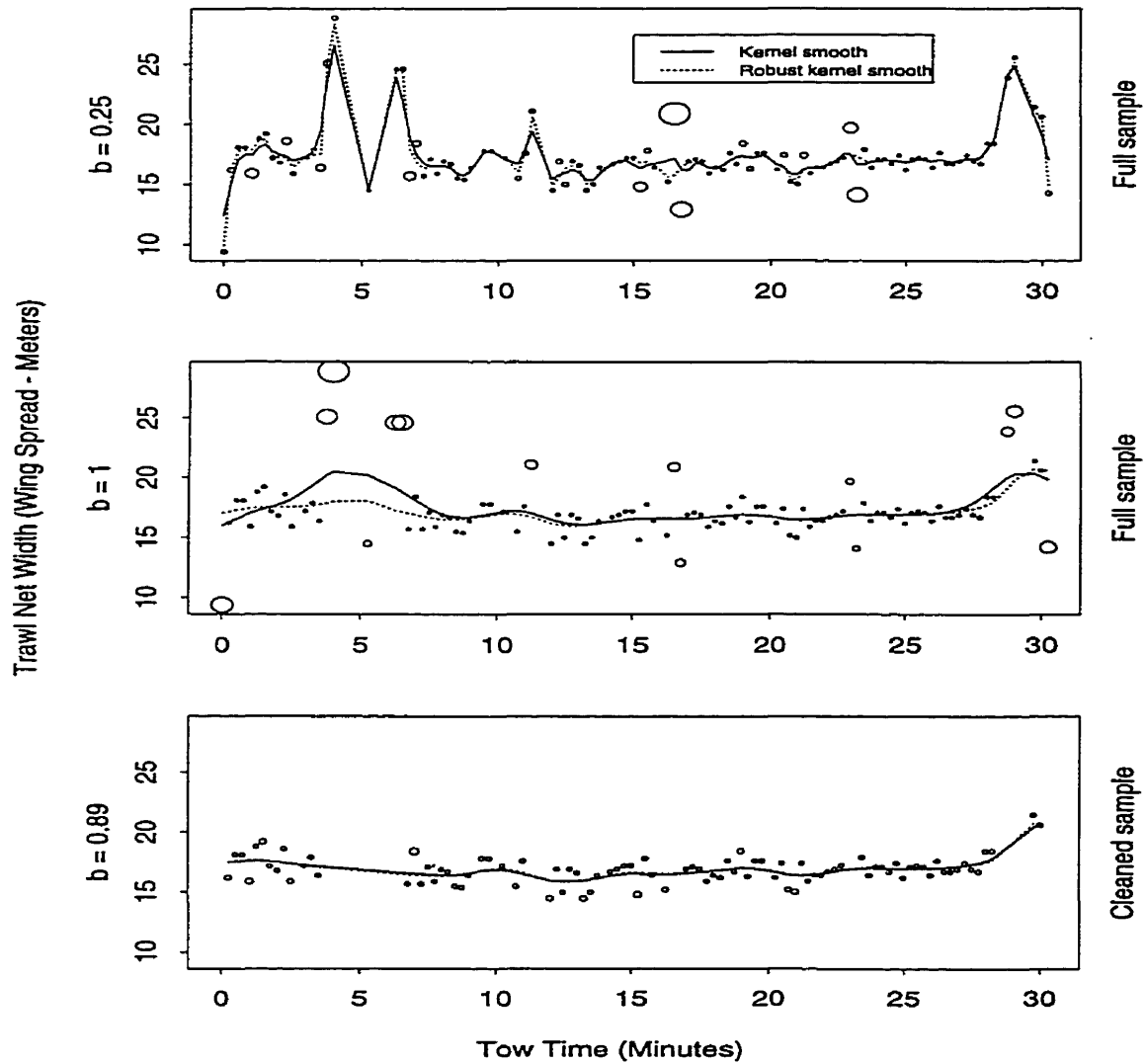


Figure 5.2: "Robustified" smoothing analysis of trawl wing-spread data. See text for details.

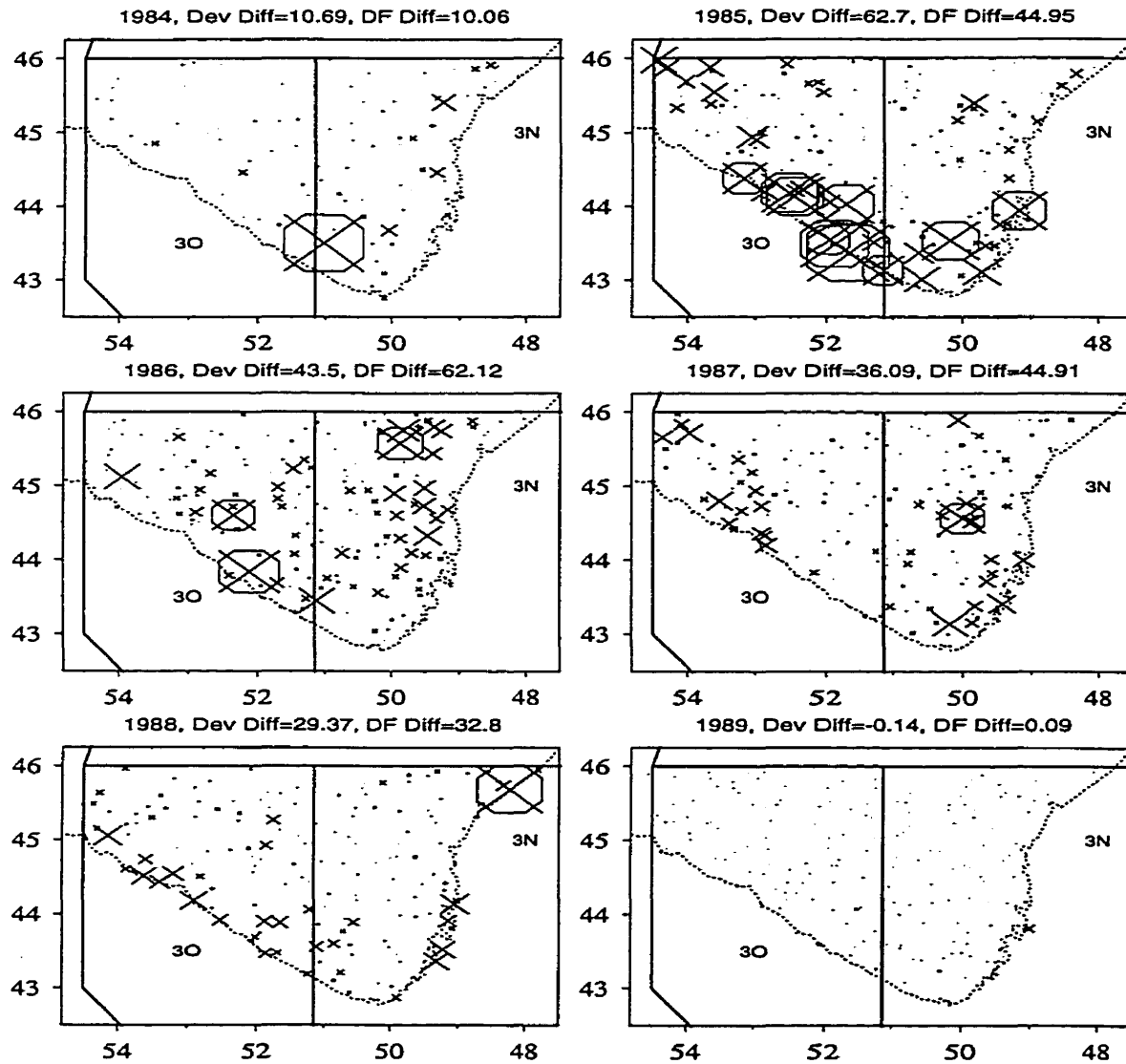


Figure 5.3a: Difference between the $CV b_{\min}$ per observation “cleaned” deviance for the model with all covariates and the model with spatial location and tow depth.

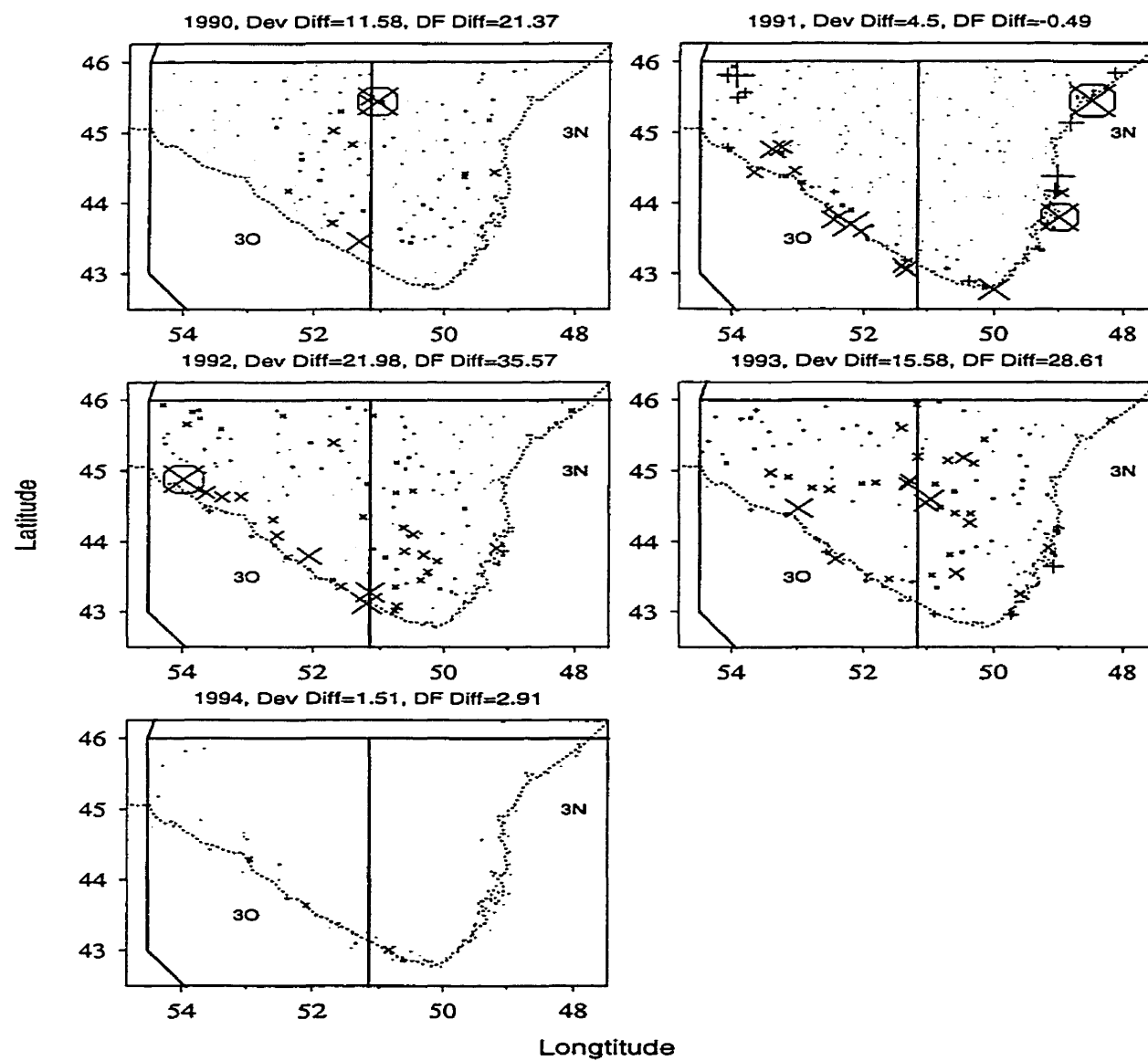


Figure 5.3b: Same as Figure 5.3a.

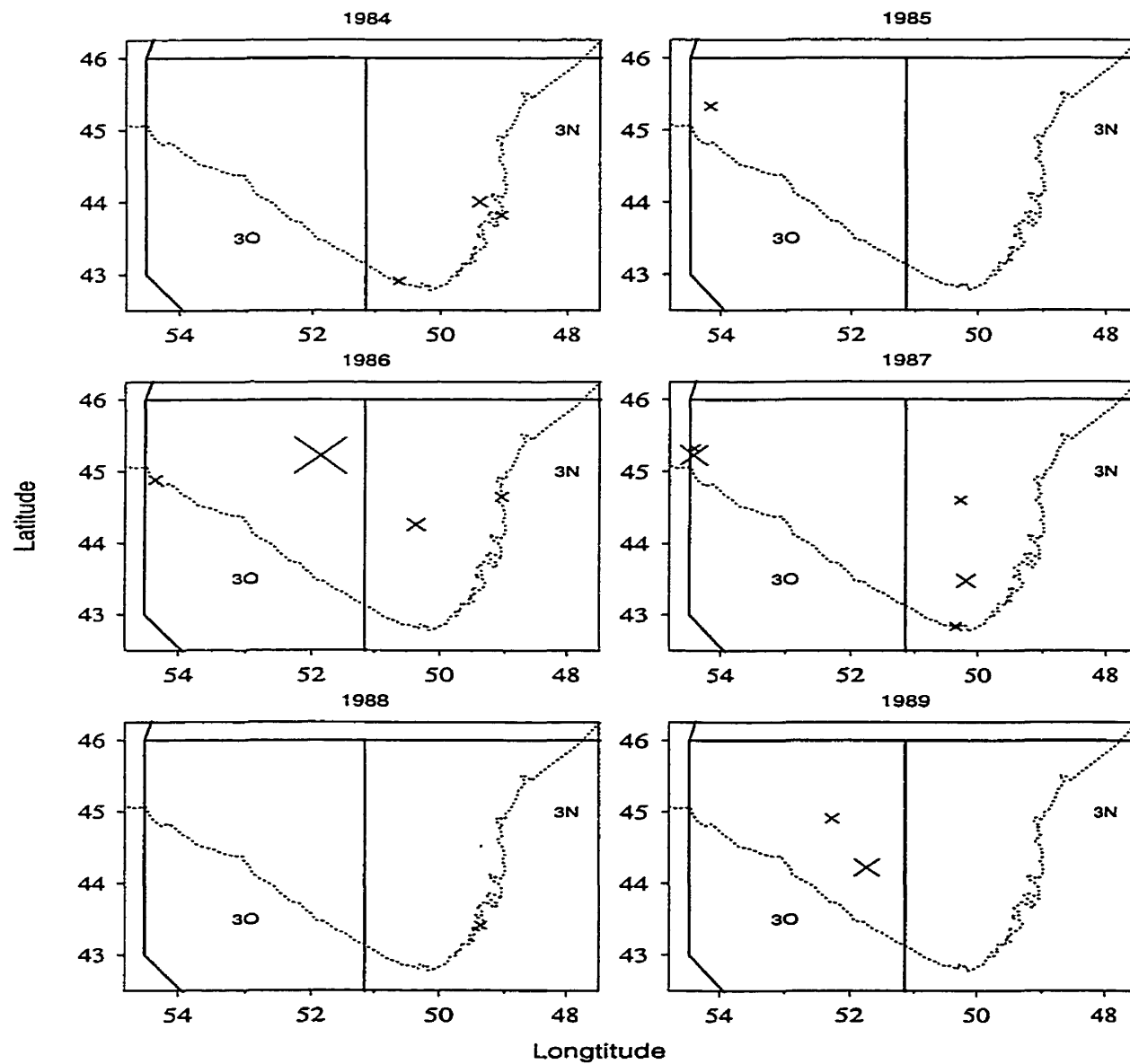


Figure 5.4a: Inverse robust weights for a smoother based on spatial location and tow depth. The gross-error sensitivity is 5.

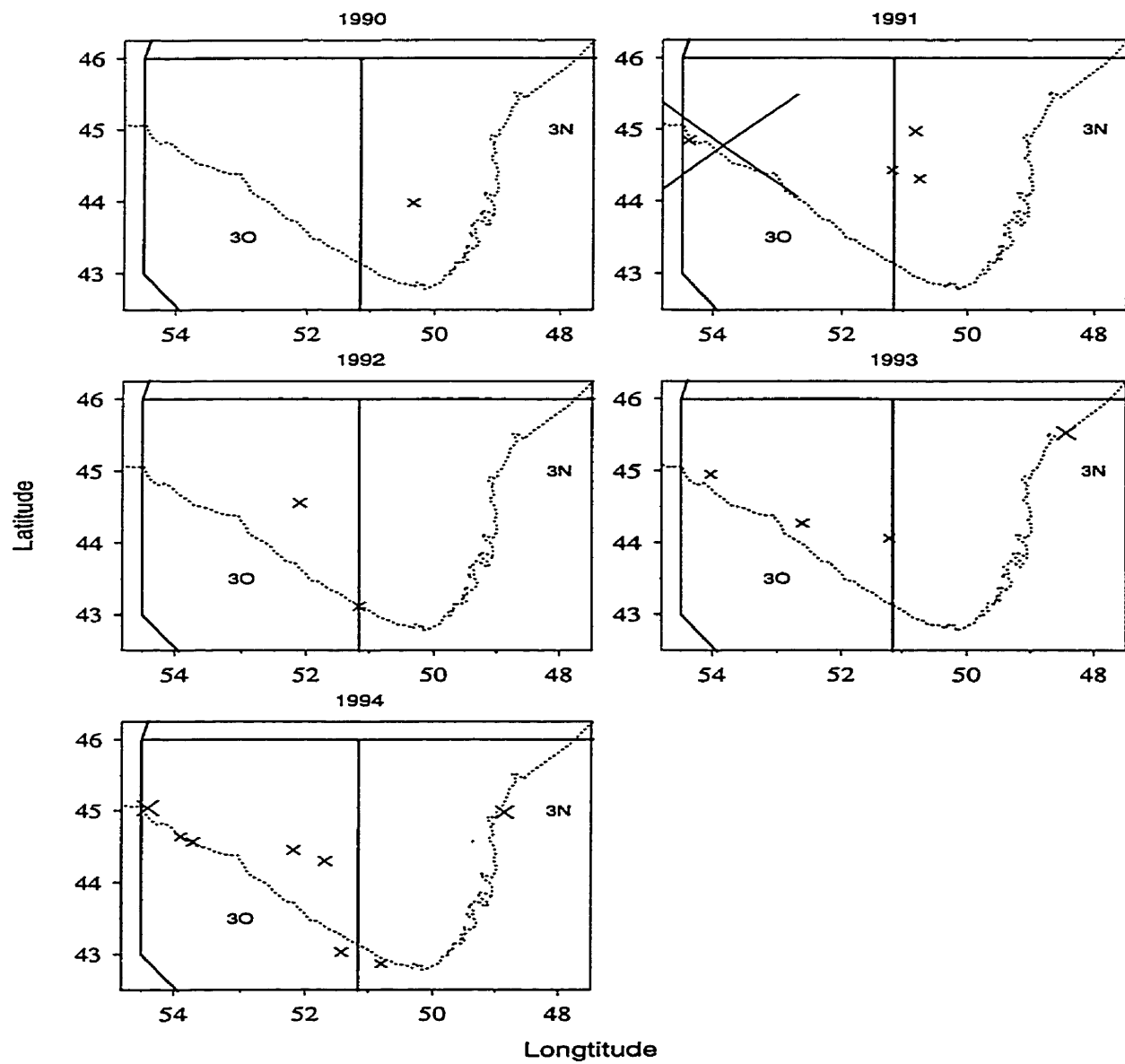


Figure 5.4b: Same as Figure 5.4a.

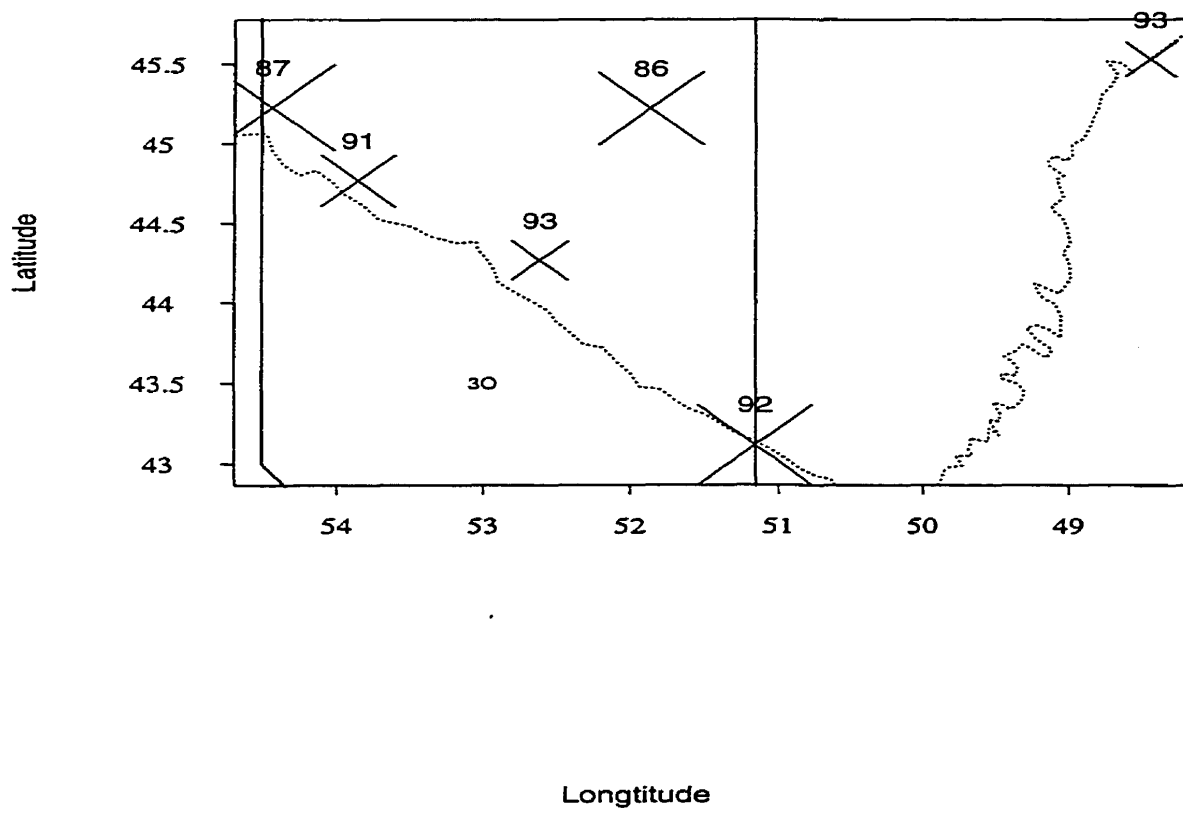


Figure 5.5: Year and location of type II catches.

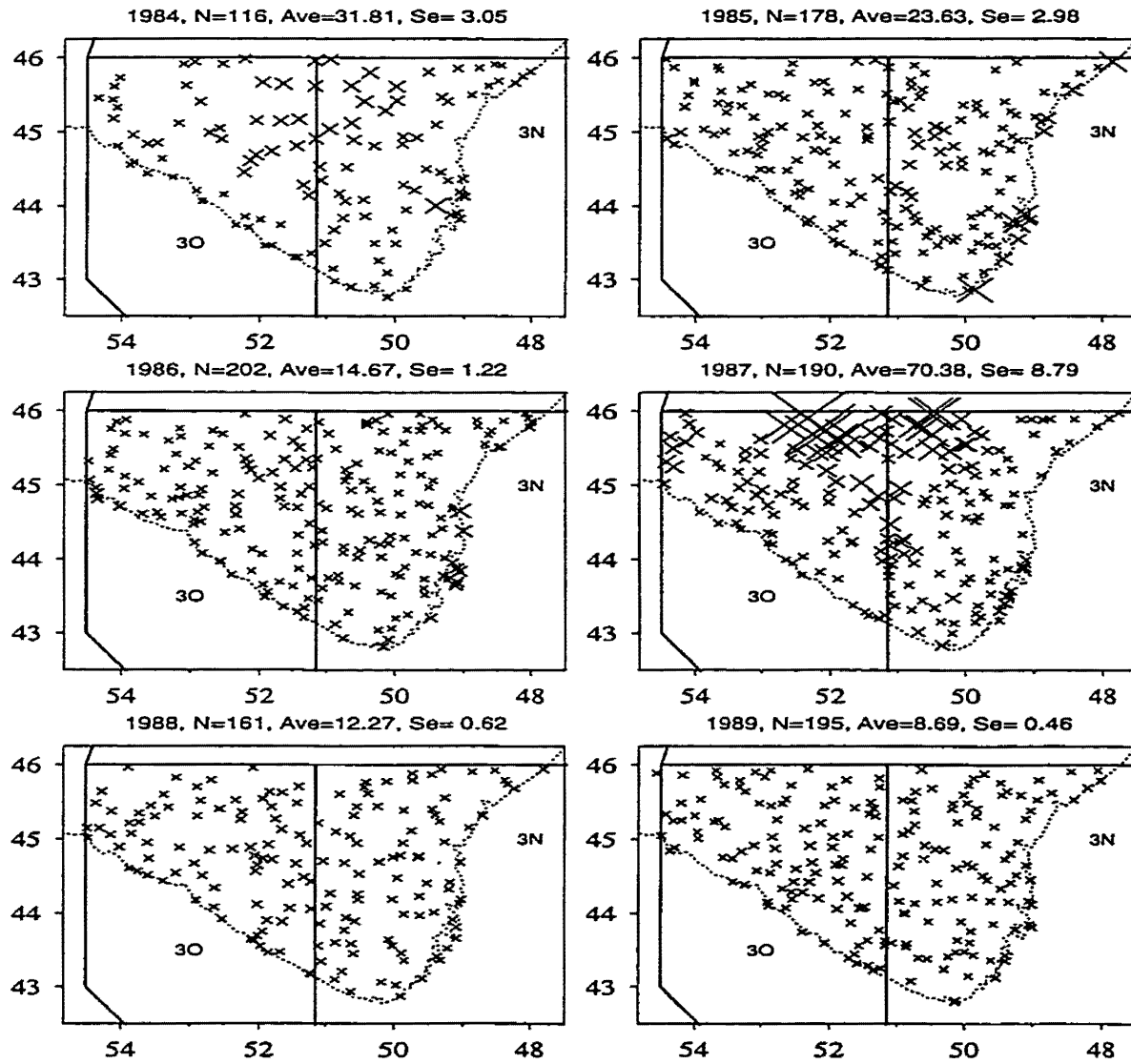


Figure 5.6a: Smoother predicted type I cod catches. The kernel smoother is based on spatial location and depth of tow.

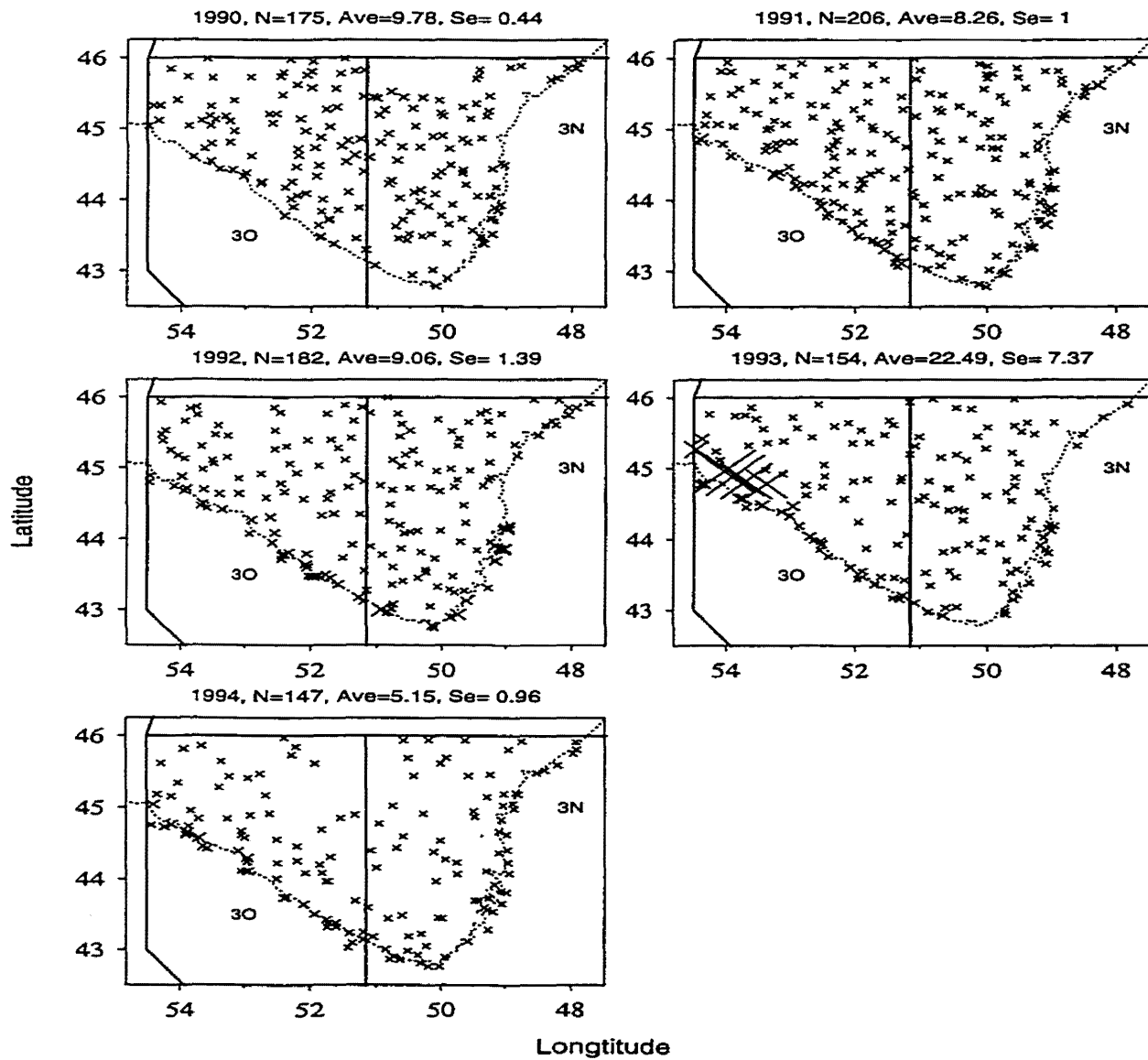


Figure 5.6b: Same as Figure 5.6a.

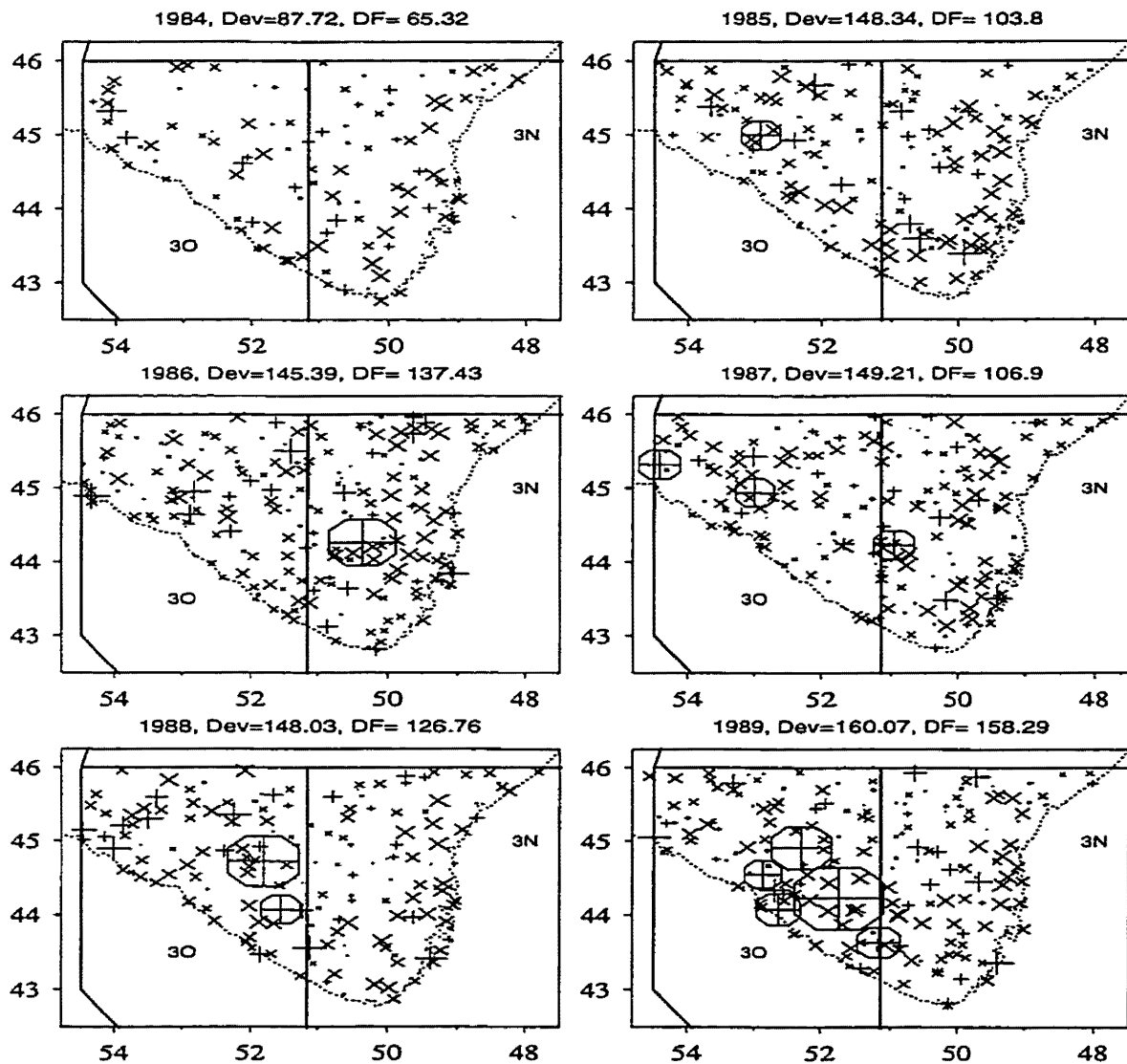


Figure 5.7a: Smoother standardized type I residuals. The kernel smoother is based spatial location and depth of tow. Positive and negative residuals are plotted as + and -'s, respectively. Circles indicate residuals greater than ± 2 .

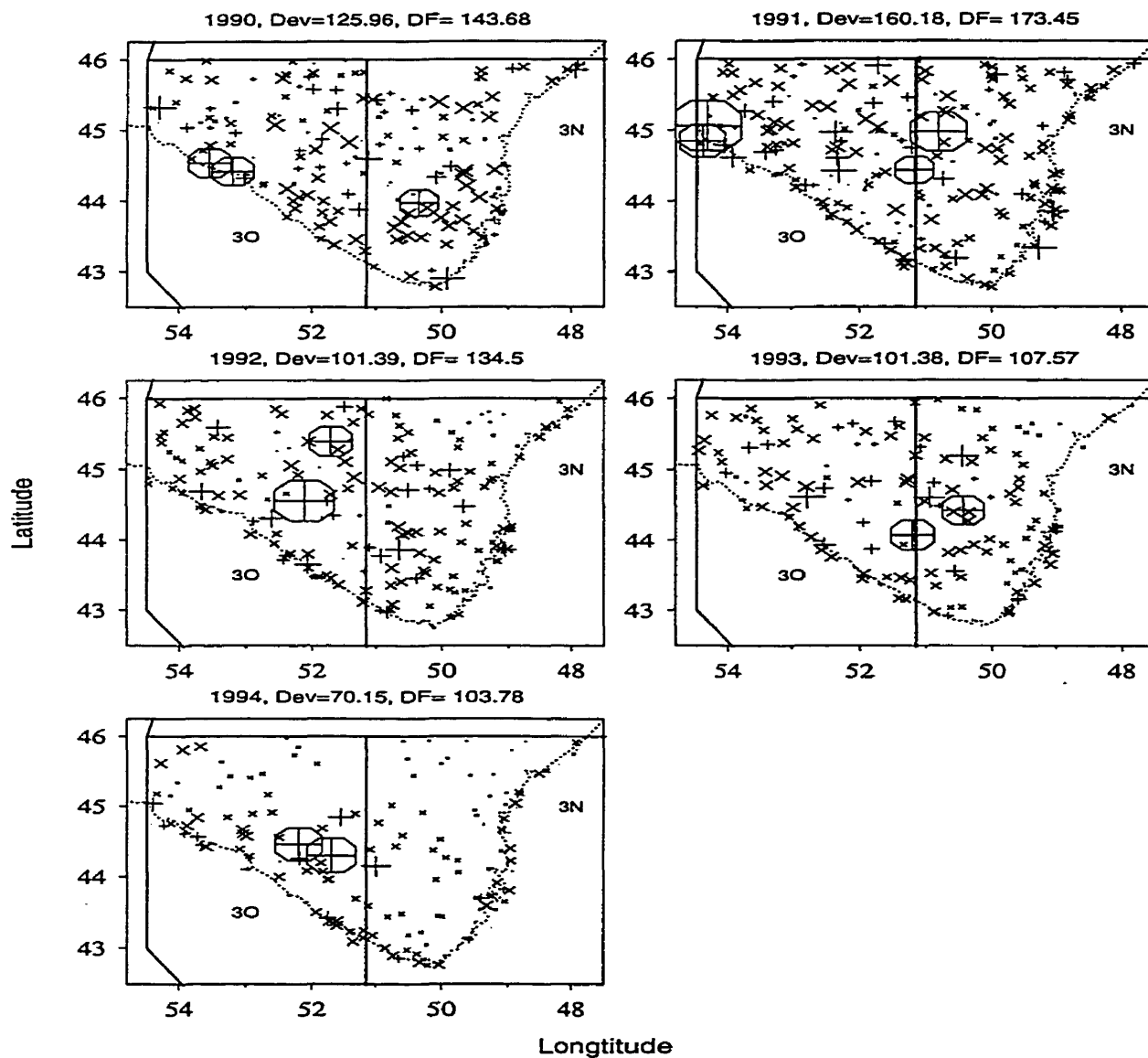


Figure 5.7b: Same as Figure 5.7a.

6. INFERENCES

6.1. Introduction

In this chapter we produce confidence intervals for 3NO trawlable stock size based on the mixture model. The basis for inference is as follows. We assume that the trawl catch at tow station i in year t can be adequately described by a random variable (rv) Y_{ti} . Let μ_{ti} be the trawlable abundance of fish in the water column above tow station i . The distribution of Y_{ti} involves μ_{ti} and other parameters. Our aim is to construct inferences about $\mu_{t.} = \sum_i \mu_{ti}$ based on a sample of the Y_{ti} 's. The sum is over all trawling stations in 3NO.

The analyses in **Chapter 5** dealt with specifying the details of a stochastic model for trawl catches that can be used for stock size inferences. This model is a mixture of two random components. One component represents trawl catches that are locally dependent; that is, catches that vary smoothly as a function of covariates. We refer to these as type I catches. In **Chapter 5** we concluded that spatial location and tow depth were the most useful covariates to include in a model for the mean of the random variable we use to stochastically model type I catches. Based on these covariates we developed a nonparametric regression model for type I catches. This model essentially estimates the mean function using locally weighted averages. The weights were specified in **Chapter 5**. The other component of the mixture model represents occasional large catches that are inconsistent with our model for type I catches. We call these inconsistent catches type

II's. We assumed that the stock components producing type II catches are present in the 3NO survey region every year, but sometimes are not detected because the areal extent of these components is quite small. We postulated a simple model for type II catches in which the mean is constant from year to year, and also spatially constant. The mixture proportions were assumed to vary over three large spatial regions. In Chapter 5 we identified six type II catches in the 3NO research survey series from 1984-1994.

We develop inferences for type I and type II stock components separately, and then combining results to get confidence intervals for μ_t . If we let μ_{ti}^I denote local type I trawlable stock abundance, then we first consider asymptotic inferences for μ_{ti}^I and then μ_t^I . This involves considering the large sample distribution of standardized estimators of μ_{ti}^I and μ_t^I . Next we consider asymptotic methods for inferences about μ^{II} - local type II trawlable abundance. We use profile likelihood methods to combine results for inferences about μ_t , which is a weighted sum of μ_t^I and μ_t^{II} .

The outline of this chapter is as follows. In Section 6.2 we use the mixture model to estimate total trawlable stock abundance in 3NO. In Section 6.3 we develop methods for constructing confidence intervals for total trawlable stock abundance, and apply these methods in Section 6.3.4. We test some of the confidence interval theory in Section 6.4 using simulations on a known population. Conclusions are presented in Section 6.5.

6.2. Mixture model estimation

In Chapter's 4 and 5 we concluded that a sensible approach to modelling trawl survey data is to assume that Y_{ti} , the random variable for the trawl catch at tow station i in year t , is a mixture of two rv's representing type I and type II catches. Recall that most catches are type I's, while occasional large catches that are inconsistent with nearby catches are

type II's. The mixture model is

$$Y_{ti} = (1 - I_i)Y_{ti}^I + I_iY_{ti}^{II} \quad (6.2.1)$$

where Y_{ti}^I and Y_{ti}^{II} are rv's for the two stock components. The mixture indicator random variables I_i are all independent with expectations p_i . The mixture proportions, p_i 's, are assumed to be equal throughout three large geographical areas in 3NO.

$$p_i = \begin{cases} p^1, & \text{if } i \in \text{southwestern Grand Bank slope,} \\ p^2, & \text{if } i \in \text{southeastern Grand Bank slope,} \\ p^3, & \text{if } i \in \text{Grand Bank shelf.} \end{cases}$$

For simplicity we assume the three p^i 's are known exactly, although we estimated these in Chapter 5. We will not consider uncertainty in our inferences arising from the unknown mixture proportions; however, in the next section we will consider the sensitivity of our inferences to other reasonable mixture proportions.

Type I catches are denoted as Y^I 's, where $Y_{ti}^I \stackrel{i.i.d.}{\sim} NB(\mu_{ti}^I, k^I)$, and $\mu_{ti}^I = \mu_t^I(\mathbf{x}_i)$ is a smooth function of covariates $\mathbf{x}_i$'s (latitude, longitude, and depth of the tow). NB denotes the Negative Binomial distribution (see Section 5.4 in Chapter 5). Type II catches are denoted as Y^{II} 's, where $Y_{ti}^{II} \stackrel{i.i.d.}{\sim} NB(\mu^{II}, k^{II})$; that is, the mean of Y_{ti}^{II} does not depend on location or year. This assumption is based on habitat selection concepts, where fish move into favorable habitats as space becomes available, thereby maintaining their density in small amounts of highly favorable habitat as overall density declines. This assumption was discussed in more detail at the beginning of Chapter 5. We assume that the Y_{ti}^I 's and Y_{ti}^{II} 's are also independent. The expected catch using (6.2.1) is

$$E(Y_{ti}) = \mu_{ti} = (1 - p_{ti})\mu_{ti}^I + p_{ti}\mu^{II}.$$

6.2.1. Estimates of 3NO trawlable stock size, μ_t .

We can estimate trawlable stock size by estimating each μ_{ti} and summing i over the roughly 2.5 million trawl stations in 3NO. If we let $r = 1, 2, 3$ index the three geographical areas mentioned above, then the estimator of μ_t is

$$\hat{\mu}_t = \sum_r (1 - p^r) \hat{\mu}_{tr}^I + \left(\sum_r p^r N_r \right) \hat{\mu}^{II}, \quad (6.2.2)$$

where $\hat{\mu}_{tr}^I$ is the estimator of type I stock size in region r and year t , N_r is the total number of trawl stations in region r , and $\sum_r p^r N_r$ is the total number of trawl stations with type II stock components. Unfortunately we do not know the covariates required to estimate $\mu_t^I(\mathbf{x}_i)$ everywhere, so this approach is not possible. We do know the covariates for a large number of tow stations randomly sampled during many surveys however, and we will use this information to assist in estimating μ_t .

We can pool covariates from annual research surveys to provide a large sample of covariates for the population. This sample is much larger than the number of trawl stations sampled each year. This is possible because the covariates we consider (spatial location and tow depth) do not change over time. In addition, many tow stations are sampled for experimental purposes, apart from the regular survey design, and we can also add these to our pool of covariates. It seems reasonable to assume that this large sample of covariates are random with strata, but not between strata. The “density” of covariates varies between strata; that is, if N_h is the total number of tow stations in stratum h and M_h is the total number of random stations in this strata that we have covariate values for, then M_h/N_h varies with h . We must account for this in estimation. Let H_r denote the number of strata in region r . The estimator we use is

$$\hat{\mu}_{tr}^I = \sum_{h=1}^{H_r} \frac{N_h}{M_h} \sum_{i=1}^{M_h} \hat{\mu}_t^I(\mathbf{x}_i) = \sum_{h=1}^{H_r} N_h \hat{\mu}_{th},$$

where $\hat{\mu}_t(\mathbf{x}_i)$ is the average predicted trawlable stock size for the random tow stations selected in stratum h . Note that $\hat{\mu}_{th}$ is not a stratum average of catches; in fact, the estimate for each station is based, in part, on locally weighted averages of type I survey catches for all strata in year y .

We demonstrate using simulations in Section 6.4 that estimation of a population total based on a large sample of covariates is almost as good as estimation using the covariates for the entire population. The basic idea here is to weight the catches in our sample by the representativeness of the corresponding covariates in terms of the covariates for the whole population. For example, if a catch occurs at a depth that is infrequent in the population, then that observation should receive less weight for estimation than catches that occur at depths that are frequent in the population.

We restrict our random covariate samples to the 300-series strata (see Tables 3.3a-k) because these are the strata sampled continuously throughout 1984-1994 and we do not wish to extrapolate densities in deep-water strata in years in which these strata were not sampled. Because of the increased precision in measurements of spatial location and depths in the 1990's we base our covariate sample on the 1636 tows DFO has conducted in 3NO since 1990. Most of these tows are not included in the data set we have analyzed so far because they were taken after 1994, or at different times of the year, or were not survey tows, etc. The density of covariate samples varies considerably between strata, and this is why we use a strata-size weighted total. For example, the density ranges from 47 stations in strata 332 (0.05%) to 24 in stratum 335 (0.55%). The average density is 0.14%.

The estimator of $\mu_t^I(\mathbf{x}_i)$ proposed in Chapter 5 is a simple locally weighted average (smoother) of trawl catches. The bandwidths we use for smoothing are those in Table 5.19. The values of the NB k parameter we use for inferences are also listed in Table 5.19. We used this model in Chapter 5 to predict local abundance at all the trawl stations sampled

each year. In this section we merely use the model to predict local trawlable abundance at the nonsampled tow stations we have selected. These predictions are simply weighted averages of type I catches each year in the survey. Estimates of annual type I trawlable abundance for each region ($\hat{\mu}_{tr}^I$) are presented in the middle three columns in Table 6.2, and will be discussed shortly.

The type II catches identified in Chapter 5 are listed in Table 5.16. Based on these data, $\hat{\mu}^{II} = 1629$, and $\hat{k}^{II} = 8.86$. The deviance for the type II catches in 6.11, which is only slightly larger than $df^{err} = 5$. This indicates that the NB model fits these data well. The total number of type II catches identified for the three regions outlined in Section 5.7 in Chapter 5, and the p^r , are presented in Table 6.1. In this table $\hat{\mu}_r^{II} = p^r N_r \hat{\mu}^{II}$ is

Table 6.1. Summary of type II catches. N_r is the number of trawlable units in a region. $\hat{\mu}_r^{II}$ is the total type II trawlable abundance. in 1000s.

Region (r)	Number type II	Total number sets examined	p^r	N_r	$\hat{\mu}_r^{II}$
SW slope in 3O	4	385	0.0104	357756	6054.9
SE slope in 3N	1	257	0.0039	157305	997.1
3NO shelf	1	1270	0.00079	2075525	2662.2

the estimate of total type II trawlable stock size for region r . The sum of type II trawlable abundance for the three regions is 9.714 million, which suggests that approximately 20000 tonnes of cod aggregate in spatially-small dense schools each year, based on an average weight per fish of a little over 2 Kg.

The estimates of mixture proportions are based on small sample sizes, because the total number of type II catches is very small. The total number of trawl stations with type II stock components is $\sum_r p^r N_r = 5963$. If we assume that the mixture proportions are constant throughout 3NO then $p = 6/1912$, which is the total number of type II

catches divided by the total number of tows during 1984-94. Using this value the total number of trawl stations with type II stock components is $pN = 8129$, which is 36% greater than 5963. This gives a total type II trawlable stock biomass of about 27000 tonnes, and has an appreciable effect on estimates of trawlable stock abundance. This demonstrates that our assumptions about the incidence of type II stock components has influence on our estimates.

As a simple test of whether the mixture proportions differ between the three geographic regions we assume within each region that the number of type II catches identified is $\text{Bin}(p^r, n_r)$, where n_r is the total number of samples collected for region r in the 3NO research survey during 1984-94. Fisher's exact test for

$$H_o : p^1 = p^2 = p^3 \text{ versus } H_a : \text{not so}$$

gives a p-value = 0.015, which is strong evidence against the null hypothesis. An equivalent test for just the two slope regions gives a p-value = 0.653. This suggests that the incidence of type II catches is not significantly different for the two slope regions. If we make this assumption then the total number of trawl stations with type II stock components is 5646, which is nearly identical to the value we use (5963), and this assumption would result in only small changes in our estimates of stock abundance. We choose not to combine estimates for these two slope regions because of the different nature of the environment between these regions. We conclude that although our estimates of the mixture proportions are based on small sample sizes we feel that our estimates will not change too much for other mixture proportions that are reasonably consistent with the incidence of type II catches in our data.

Estimates of annual type I trawlable abundance and total abundance for both stock components are presented in Table 6.2. In this table $\hat{\mu}_t$ is the sum of the type I abundance from each region plus 9714.2, which is the total type II abundance in Table 6.1. These

Table 6.2. Mixture model estimates of annual trawable stock abundance. $\hat{\mu}_{tr}^I$ is total type I trawable abundance in each region. $\hat{\mu}_t$ is total trawable abundance. Numbers are in 1000's.

Year	SW slope	$\hat{\mu}_{tr}^I$ SE slope	3NO shelf	$\hat{\mu}_t$
1984	1430.3	1515.6	92688.6	105348.8
1985	3896.5	9541.1	37556.6	60708.4
1986	3146.7	5908.4	23475.0	42244.3
1987	16144.3	4069.5	159608.6	189536.5
1988	4287.6	1216.0	27341.2	42559.0
1989	2469.9	1456.2	18158.8	31799.1
1990	3567.8	1751.9	18799.6	33833.5
1991	4610.1	2690.2	5072.8	22087.3
1992	7490.4	5478.6	2197.3	24880.4
1993	35969.5	1442.9	1511.2	48637.8
1994	8645.0	780.1	236.3	19375.6

values are also plotted in Figure 6.2 in Section 6.3.4, along with the stratum area-weighted averages used by DFO (see Table 3.4a). The mixture model confidence intervals presented in this figure are described in the next section. The stratum-means confidence intervals are based on ± 2 standard errors, and these are described in Chapter 3 and presented in Table 3.4a. The mixture model suggests much different trends in 1992-94, with population abundance apparently increasing 3-fold in 1993 then decreasing 3-fold, whereas stratified-means suggest a 10-fold increase in 1993, and a 20-fold decrease in 1994. The mixture model has lead to some smoothing, with lower peaks and higher valleys in Figure 6.2 than the stratified-means.

6.3. Inferences

In this section we develop confidence intervals for annual trawlable stock size, μ_t . The basic procedure we use is to develop an asymptotically pivotal statistic for μ_t^I , and combine this with the distribution of $\hat{\mu}_t^{II}$, which is proportional to a NB random variable. Recall that our estimator of μ_t is a weighted sum of $\hat{\mu}_t^I$ and $\hat{\mu}_t^{II}$. We use the profile likelihood procedure (see Barndorff-Nielsen and Cox, 1994) to develop confidence intervals for μ_t based on likelihoods for μ_t^I and μ_t^{II} . We approximate the distribution of the profile likelihood statistic as chi-square with one degree-of-freedom. We first consider the consistency of $\hat{\mu}_t^I(\mathbf{x})$ for $\mu_t^I(\mathbf{x})$, and then consider the consistency and asymptotic distribution of $\hat{\mu}_t^I = \sum_{h=1}^H \frac{N_h}{M_h} \sum_{i=1}^{M_h} \hat{\mu}_t^I(\mathbf{x}_i)$.

We assume throughout this section that survey trawl catches are independently distributed as either $NB\{\mu_t^I(\mathbf{x}_i), k\}$ for type I catches, or $NB(\mu_t^{II}, k^{II})$ for type two catches. We assume that the k 's are known because of the confounding between bandwidths and k , as discussed in Chapter 5. We investigate the sensitivity of this assumption using simulations. For type I catches we take $k = 0.89$ for the 3NO shelf region, and $k = 0.44$ for the slope region. These values were derived in Chapter 5.

6.3.1. Consistency of $\hat{\mu}_t^I(\mathbf{x})$

In this section we show that $\hat{\mu}_t^I(\mathbf{x})$ is consistent for local type I trawlable abundance, $\mu_t^I(\mathbf{x})$, under some reasonable conditions on the bandwidths and local weights. We treat the local weights as fixed; that is, we do not explicitly consider that the bandwidths are chosen to minimize a criterion based on the responses; however, we do consider the impact that the method for choosing bandwidths may have on asymptotic inferences.

Remark 5. *For simplicity we drop the t and I notation in this section.*

The abundance estimator is based on survey catches, $(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)$, where the y 's are the trawl catches and the $\mathbf{x}$'s are $p \times 1$ vectors of covariates. In this chapter $p = 3$, because the only covariates we consider are latitude, longitude, and tow depth. The estimator is simply a weighted average of the y 's, where the weights are all positive and sum to one. The estimator is

$$\hat{\mu}(\mathbf{x}) = \sum_{j=1}^n W_j(\mathbf{x}) Y_j, \quad (6.3.1)$$

where

$$W_j(\mathbf{x}) = \frac{\exp \left\{ - \sum_{k=1}^p (x_{jk} - x_k)^2 / 2b_k^2 \right\}}{\sum_{i=1}^n \exp \left\{ - \sum_{k=1}^p (x_{ik} - x_k)^2 / 2b_k^2 \right\}},$$

and x_k is an element of $\mathbf{x}$. We have dropped the bandwidth notation in $W(\cdot)$.

Locally weighted average-type estimators like (6.3.1) have been extensively studied in the statistical literature. A common approach when investigating the distributional properties of (6.3.1) is to assume both Y_i and $\mathbf{x}_i$ are random (for example, see Härdle, 1991). Stone (1977) considered conditions for the weights so that (6.3.1) is consistent for $\mu(\mathbf{x})$. In keeping with the standard practice of conditioning on the covariates in a regression analysis, we assume that the covariates are not random. Ruppert and Wand (1994) also assumed the same. Stone's (1997) results can be easily modified for this setting. We also assume that the survey design is such that as the sample size increases the sampled sites spread out evenly; that is, we do not sample specific areas much more frequently than other areas. For the 3NO survey design this is true.

The expected value of the estimator of local stock abundance is

$$E[\hat{\mu}(\mathbf{x})] = \sum_{i=1}^n W_i(\mathbf{x}) \mu(\mathbf{x}_i),$$

which can be written as

$$E[\hat{\mu}(\mathbf{x})] = \sum_{i=1}^n W_i(\mathbf{x}) \mu(\mathbf{x}_i) I_{\{\|\mathbf{x}-\mathbf{x}_i\| \leq a\}} + \sum_{i=1}^n W_i(\mathbf{x}) \mu(\mathbf{x}_i) I_{\{\|\mathbf{x}-\mathbf{x}_i\| > a\}},$$

for some constant $a > 0$. We assume the Lipsitz condition for local stock abundance,

$$|\mu(\mathbf{x}) - \mu(\mathbf{x} + \mathbf{d})| \leq c\|\mathbf{d}\|,$$

where c is also a constant. Let $b_i(\mathbf{x}) = \mu(\mathbf{x}_i) - \mu(\mathbf{x})$. Using this assumption we have that

$$\begin{aligned} \sum_{i=1}^n W_i(\mathbf{x}) |b_i(\mathbf{x})| &= \sum_{i=1}^n W_i(\mathbf{x}) |b_i(\mathbf{x})| I_{\{\|\mathbf{x}-\mathbf{x}_i\| \leq a\}} + \sum_{i=1}^n W_i(\mathbf{x}) |b_i(\mathbf{x})| I_{\{\|\mathbf{x}-\mathbf{x}_i\| > a\}}, \\ &\leq ca \sum_{i=1}^n W_i(\mathbf{x}) I_{\{\|\mathbf{x}-\mathbf{x}_i\| \leq a\}} + \sum_i W_i(\mathbf{x}) \{b_i(\mathbf{x})\} I_{\{\|\mathbf{x}-\mathbf{x}_i\| > a\}}, \\ &\leq ca + \max_i |b_i(\mathbf{x})| \sum_{i=1}^n W_i(\mathbf{x}) I_{\{\|\mathbf{x}-\mathbf{x}_i\| > a\}}. \end{aligned}$$

The following condition establishes the asymptotic unbiasedness of $\hat{\mu}(\mathbf{x})$.

$$\text{As } n \rightarrow \infty, \sum_{i=1}^n W_i(\mathbf{x}) I_{\{\|\mathbf{x}-\mathbf{x}_i\| > a\}} \rightarrow 0 \text{ for all } a > 0. \quad (6.3.2)$$

Because $\max_i b_i(\mathbf{x}) = \max_i |\mu(\mathbf{x}_i) - \mu(\mathbf{x})|$ is bounded the first term on the right-hand side of the inequality goes to zero using condition (6.3.2). Since a can be made arbitrarily small we have that

$$\sum_{i=1}^n W_i(\mathbf{x}) \mu(\mathbf{x}_i) \rightarrow \mu(\mathbf{x}) \text{ as } n \rightarrow \infty;$$

hence, $\hat{\mu}(\mathbf{x})$ is asymptotically unbiased for $\mu(\mathbf{x})$.

Condition (6.3.2) follows if the bandwidths decrease as n gets large, and if the covariates spread out evenly. Roughly speaking for the covariates we mean that their density remains bounded away from zero; that is, for any covariate region A ,

$$\frac{\sum_{i=1}^n I(\mathbf{x}_i \in A)}{n|A|} > 0, \text{ as } n \rightarrow \infty. \quad (6.3.3)$$

Data-driven methods of bandwidth selection (e.g. GC_p) have the property that bandwidths decrease as n increases. We demonstrate that (6.3.2) holds with the assumptions about bandwidth size and covariate density using a simple example at the end of this section.

We can easily show using Chebyshov's inequality that $\hat{\mu}(\mathbf{x}) \xrightarrow{P} \mu(\mathbf{x})$ if, in addition to being asymptotically unbiased, $Var[\hat{\mu}(\mathbf{x})] \rightarrow 0$ as $n \rightarrow \infty$. The variance for $\hat{\mu}(\mathbf{x})$ is

$$\begin{aligned} Var[\hat{\mu}(\mathbf{x})] &= \sum_{i=1}^n W_i^2(\mathbf{x}) Var(Y_i), \\ &\leq \left\{ \max_i Var(Y_i) \right\} \sum_{i=1}^n W_i^2(\mathbf{x}). \end{aligned}$$

For our problem it is reasonable to assume that $0 < \varepsilon \leq Var(Y_i) \leq M < \infty$ for all i , where ε and M are constants. Hence, $Var[\hat{\mu}(\mathbf{x})] \rightarrow 0$ if

$$\sum_{i=1}^n W_i^2(\mathbf{x}) \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (6.3.4)$$

Because the weights are constrained to sum to one (6.3.4) is equivalent to $\max_i W_i(\mathbf{x}) \rightarrow 0$, and this implies that the bandwidths must not decrease too slowly with n . Hence, if conditions (6.3.2) and (6.3.4) hold then $\hat{\mu}(\mathbf{x})$ is consistent for $\mu(\mathbf{x})$.

Example

We illustrate, using a simple example, that (6.3.2) and (6.3.4) hold if, as $n \rightarrow \infty$, $b_n \rightarrow 0$ and $nb_n^p \rightarrow \infty$. These are the basic assumptions often made in nonparametric kernel regression. Consider the case when $p = 1$, $x = 0$, and the x_i all have densities that are uniform on $(-\frac{1}{2}, \frac{1}{2})$. We can rewrite $\sum_{i=1}^n W_i(\mathbf{x}) I_{\{\|\mathbf{x}-\mathbf{x}_i\|>a\}}$ in (6.3.2) as

$$\sum_{i=1}^n W_i(\mathbf{x}) I_{\{\|\mathbf{x}-\mathbf{x}_i\|>a\}} = \frac{\sum_{i=1}^n (2\pi)^{-1/2} (nb_n)^{-1} \exp(-x_i^2/2b_n^2) I_{\{|x_i|>a\}}}{\sum_{i=1}^n (2\pi)^{-1/2} (nb_n)^{-1} \exp(-x_i^2/2b_n^2)}$$

Taking limits, for the denominator we have

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sum_{i=1}^n (2\pi)^{-1/2} (nb_n)^{-1} \exp\left(\frac{-x_i^2}{2b_n^2}\right), \\ &= \lim_{n \rightarrow \infty} \int_{-1/2b_n}^{1/2b_n} (2\pi)^{-1/2} \exp(-z^2/2) dz, \\ &= \lim_{n \rightarrow \infty} 1 - 2\Phi(-1/2b_n). \end{aligned}$$

This limit is one if $b_n \rightarrow 0$ as $n \rightarrow \infty$. For the numerator, if $a > 0$ we have

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sum_{i=1}^n (2\pi)^{-1/2} (nb_n)^{-1} \exp\left(\frac{-x_i^2}{2b_n^2}\right) I_{\{|x_i|>a\}}, \\ &= \lim_{n \rightarrow \infty} \int_{a/b_n}^{1/2b_n} (2\pi)^{-1/2} \exp(-z^2/2) dz, \\ &= \lim_{n \rightarrow \infty} \Phi(1/2b_n) - \Phi(a/b_n). \end{aligned}$$

This limit is zero with our assumption about b_n ; hence, if $b_n \rightarrow 0$ as $n \rightarrow \infty$ then (6.3.2) is holds.

In a similar manner we can rewrite $\sum_{i=1}^n W_i^2(\mathbf{x})$ in (6.3.4) as

$$\sum_{i=1}^n W_i^2(\mathbf{x}) = \frac{(2\pi)^{-1/2}(nb_n)^{-1} \left[\sum_{i=1}^n (2\pi)^{-1/2}(nb_n)^{-1} \exp(-x_i^2/b_n^2) \right]}{\left[\sum_{i=1}^n (2\pi)^{-1/2}(nb_n)^{-1} \exp(-x_i^2/2b_n^2) \right]^2}.$$

The limit as $n \rightarrow \infty$ of the denominator is again one. The limit of $[\cdot]$ in the numerator is $2^{-1/2}$, hence, if $nb_n \rightarrow \infty$ as $n \rightarrow \infty$ then the limit of the numerator is zero, and (6.3.4) is satisfied. Note that if x is $p \times 1$ and the bandwidths are the same for each element in x then we require only that $b_n^p \rightarrow 0$ as $n \rightarrow \infty$ and $nb_n^p \rightarrow \infty$ as $n \rightarrow \infty$. These results make more precise how the bandwidths must decrease as n increases.

6.3.2. Consistency and asymptotic normality of $\hat{\mu}_t^I$

In this section we show the consistency of $\hat{\mu}_t^I$ for μ_t^I , and asymptotic normality of $\hat{\mu}_t^I$, for some reasonable conditions on bandwidths and local weights. In this section we again drop the t and I notation for simplicity; however, we will re-introduce this notation later. We use the asymptotic normality result to develop approximate confidence intervals for μ_t^I . For simplicity we assume for now that all the population is composed of type I stock components. We first deal with the case when the covariates are known for the entire population; that is, $\mathbf{x}_1, \dots, \mathbf{x}_N$ are known, then we consider the case where the covariates are known only for a large stratified simple random sample of tow stations. This sample includes the stations sampled in our trawl survey data set for 1990-94, and many more (see Section 6.2.1 for details).

Remark 6. When referring to covariates that are observed each year in surveys we use the i subscript (i.e. $\mathbf{x}_i$), while we use the j subscript when referring to covariates in our “population” sample (i.e. $\mathbf{x}_j$).

In the following analyses we develop asymptotics by imbedding our finite population of tow stations into an increasing sequence of finite populations of size $N(\nu)$, where

$N(\nu) \xrightarrow{\nu} \infty$ and $n(\nu) \xrightarrow{\nu} \infty$.

Covariates known for the entire population

The estimate of $\mu_{\cdot}$ is just a sum of locally-weighted averages that we use to estimate the stock abundance at random trawl stations. That is,

$$\hat{\mu}_{\cdot} = \sum_{j=1}^N \hat{\mu}(\mathbf{x}_j),$$

where $\hat{\mu}(\mathbf{x}_j)$ is a nonparametric estimate of local trawlable abundance at tow station j , and N is the total number of tow stations in the population. Note that the $\mathbf{x}_j$'s are the covariate values for the entire population, and not just the sampled trawl sites. We can also express $\hat{\mu}_{\cdot}$ as

$$\hat{\mu}_{\cdot} = \sum_{i=1}^n f_i^* y_i,$$

where $f_i^* = \sum_{j=1}^N W_i(\mathbf{x}_j)$. The f_i^* provide a relative measure of how representative the sampled $\mathbf{x}_i$'s are in terms of the population $\mathbf{x}_1, \dots, \mathbf{x}_N$. Note that $\sum_{i=1}^n f_i^* = N$. In a sense $\hat{\mu}_{\cdot}$ is similar to the expansion estimator of $T = \sum_{i=1}^N y_i$; that is, $\hat{T} = \sum_{i=1}^n \pi_i^{-1} y_i$ (see Section 3.3.2). Whereas the π weights give pre-data collection measures of how representative the sampled $\mathbf{x}$'s are in terms of the population, the f^* weights give a post-data collection measure of the same thing (see Chambers *et. al.*, 1993).

Remark 7. In what follows let $f_i = f_i^*/N = \sum_{j=1}^N W_i(\mathbf{x}_j)/N$, which is an estimate of the density of $\mathbf{x}_i$ among $\mathbf{x}_1, \dots, \mathbf{x}_N$.

Note that f_i 's behave like densities in that $\sum_{i=1}^n f_i = 1$.

Consistency It is convenient to first deal with

$$\bar{\mu}_N = \frac{1}{N} \sum_{i=1}^N \mu(\mathbf{x}_i) = \frac{\mu_{\cdot}}{N}.$$

We have shown in **Section 6.3.1** that when (6.3.2) holds $\text{bias}\{\hat{\mu}(\mathbf{x}_i)\} \rightarrow 0$ as $n \rightarrow \infty$. This is implied when the bandwidths decrease as n increases and (6.3.3) holds; that is, the covariate density remains bounded away from zero. Because of the continuity of $\mu(\mathbf{x})$ we can also show using the bounded convergence theorem that when $b_n \rightarrow 0$ as $n \rightarrow \infty$ and (6.3.3) holds uniformly for all x_i ,

$$E(\hat{\mu}_N) - \bar{\mu}_N = \frac{1}{N} \sum_{i=1}^N [E\{\hat{\mu}(\mathbf{x}_i)\} - \mu(\mathbf{x}_i)] \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (6.3.5)$$

In many situations the biases for the $\hat{\mu}(\mathbf{x}_i)$'s may cancel, so $\hat{\mu}_N$ may be asymptotically unbiased for $\bar{\mu}_N$ even if (6.3.2) doesn't hold. If, in addition, we have that

$$\max_i f_i \rightarrow 0 \text{ as } n \rightarrow \infty, \quad (6.3.6)$$

then we can show using Chebyshov's inequality that $\hat{\mu}_N$ is consistent for $\bar{\mu}_N$. Condition (6.3.6) essentially requires most sampled covariates to be representative of population values.

Asymptotic normality We establish the asymptotic normality of $\hat{\mu}_n$ in two steps. First consider the distribution of $\hat{\mu}_n - E(\hat{\mu}_n)$. Using Liapunov's central limit theorem,

$$\frac{\hat{\mu}_n - E(\hat{\mu}_n)}{\{\sum_{i=1}^n f_i^2 \sigma_i^2\}^{1/2}} \Rightarrow N(0, 1)$$

if $0 < \varepsilon_1 \leq \sigma_i^2 \leq M_1 < \infty$ and $0 < \varepsilon_2 \leq E|Y_i - \mu(\mathbf{x}_i)|^3 \leq M_2 < \infty$ for all i , and

$$\lim_{n \rightarrow \infty} \frac{\{\sum_{i=1}^n f_i^3\}^{1/3}}{\{\sum_{i=1}^n f_i^2\}^{1/2}} = 0. \quad (6.3.7)$$

In our application samples are collected, approximately, using stratified random sampling with proportional allocation, so the f_i 's are approximately equal. In this case the skewness factor in (6.3.7) is $O(n^{-1/6})$, and the limit is zero.

So far we have shown that a suitably scaled version of $\hat{\mu}_n - E(\hat{\mu}_n)$ has an asymptotic standard normal distribution. For inferences it is desirable if $\hat{\mu}_n - \bar{\mu}_N$ also has such an asymptotic distribution, because this result can directly be used for inferences. Otherwise we have to estimate the bias, or employ some other correction for this term. The standardized bias is

$$\frac{E(\hat{\mu}_n) - \bar{\mu}_N}{\{\sum_{i=1}^n f_i^2 \sigma_i^2\}^{1/2}}.$$

The bias is not asymptotically negligible in general. A common practice is to use a high-order kernel to reduce this bias, or to use the bootstrap to estimate the bias. Simulation results for a population similar to that in our application suggest that this bias is not large, so we choose simply to ignore it.

Conditions (6.3.5) through (6.3.7), in addition with the above bias argument, imply

$$\frac{\hat{\mu}_n - \bar{\mu}_N}{\{Var(\hat{\mu}_n)\}^{1/2}} = \frac{\hat{\mu}_n - \bar{\mu}_N}{\{\sum_{i=1}^n f_i^2 \sigma_i^2\}^{1/2}} \Rightarrow N(0, 1). \quad (6.3.8)$$

Variance estimation Unfortunately the denominator in (6.3.8) involves unknown σ_i^2 's, so we cannot directly use (6.3.8) for inferences about $\bar{\mu}_N$. We propose a modification for

Negative Binomial data when k is known, or assumed to be known. In this case

$$\sigma_i^2 = \mu(x_i) + \mu^2(x_i)/k.$$

First consider when k is the same for the entire sample. Let

$$\begin{aligned}\sigma_N^2 &= \bar{\mu}_N + \bar{\mu}_N^2/k, \text{ and} \\ v_n &= \sum_{i=1}^n f_i^2.\end{aligned}$$

If $Var(\hat{\mu}_n)$ was actually $v_n \sigma_N^2$ then we would not have to estimate it, because (6.3.8) would give an asymptotic pivotal statistic for $\bar{\mu}_N$ which could directly be used for inferences. The modification we propose uses this idea. We construct confidence intervals using the approximate pivotal statistic

$$\frac{\hat{\mu}_n - \bar{\mu}_N}{v_n^{1/2} [\sigma_N^2 + d_n]^{1/2}} \approx N(0, 1), \text{ where } d_n = \left\{ \sum_{i=1}^n f_i^2 \hat{\sigma}_i^2 - \hat{\sigma}_N^2 \right\} / v_n. \quad (6.3.9)$$

Note that $\sigma_N^2 + d_n$ is a difference estimate of $\sum_{i=1}^n f_i^2 \sigma_i^2$. If there is little variation in the $\mu(x_i)$'s then d_n will be close to zero. In this case variance estimation for Negative Binomial data is not really necessary, and (6.3.9) is the preferred statistic for inferences about $\bar{\mu}_N$. If $d_n = 0$ then confidence intervals based on (6.3.9) will not cover negative values because the asymptotic pivot $\rightarrow \infty$ as $\bar{\mu}_N \rightarrow 0$. This is important when $\bar{\mu}_N$ is close to zero. The resulting interval will also be skewed, which is often more realistic than a symmetric interval.

To apply this procedure to our data a slight modification is required, because we have assumed two different values of k for the slope and shelf regions. We use an average value of these two k 's for inferences. As usual let s denote our sample. Let $s_{sh} = \{i : x_i \text{ is on}$

the shelf}, and let $s_{sl} = s \setminus s_{sh}$. The average value for k that we use is

$$\bar{k} = \frac{k_{sl}k_{sh}}{k_{sl}q^{sh} + k_{sh}(1 - q^{sh})}, \quad q^{sh} = \frac{\sum_{i \in s_{sh}} f_i^2}{\sum_{i=1}^n f_i^2}.$$

If the $\mu(\mathbf{x}_i)$'s are all equal to $\bar{\mu}_N$ then this value for $\bar{k}$ would be the correct choice for inferences; hence it is the appropriate value to use when there is little variation in the $\mu(\mathbf{x}_i)$'s. We construct confidence intervals for total trawlable stock size using (6.3.9) with $\bar{k}$ in σ_N^2 . Note that we do not use $\bar{k}$ in d_n ; the appropriate k are used to compute d_n .

We can obtain confidence intervals for $\mu_{\cdot}$ as $N \times$ the intervals we obtain for $\bar{\mu}_N$ using (6.3.9).

Covariates known for a large random sample of the population

In the preceding section we developed methods for inference about $\mu_{\cdot}$ assuming that the covariates were known for the entire population. In this section we develop inferences for $\mu_{\cdot}$ assuming only that the covariates are known for a large stratified simple random sample of tow stations. Let M_h denote the number of tow stations in stratum h that we use for inferences about $\mu_{\cdot}$. In this section we make the same assumptions as in the previous section, plus we assume that $M = \sum_h M_h \gg n$.

The estimate of $\mu_{\cdot}$ is just a sum of locally-weighted averages that we use to estimate the stock abundance at random trawl stations. That is,

$$\hat{\mu}_{\cdot} = \sum_{h=1}^H \frac{N_h}{M_h} \sum_{j=1}^{M_h} \hat{\mu}(\mathbf{x}_j),$$

where $\hat{\mu}(\mathbf{x}_i)$ is a nonparametric estimate of local trawlable abundance at tow station i , N_h is the total number of tow stations in the stratum, and H is the total number of strata.

Using the same notation as in the previous section, we can write

$$\hat{\mu}_i = N \sum_{i=1}^n \hat{f}_i y_i,$$

where $\hat{f}_i = \sum_{h=1}^H \frac{N_h}{M_h} \sum_{j=1}^{M_h} W(\mathbf{x}_j - \mathbf{x}_i)/N$. It is easy to verify that $\sum_{i=1}^n \hat{f}_i = 1$. The $\hat{f}_i$'s are estimates of the f_i 's - the population densities of the sampled $\mathbf{x}_i$'s. These estimates are based on a large sample $\mathbf{x}_1, \dots, \mathbf{x}_M$. We use over 1600 tows stations for M , so we feel the $\hat{f}_i$'s are accurate for the f_i 's. If (6.3.5) holds, in addition to the bias argument presented in the previous section, then (6.3.9) can be used to construct confidence intervals for μ_i .

Comparing the f -weights with the design-based π -weights reveals the difference in how catches affect our estimate of type I trawlable stock abundance using these two approaches. In Figures 6.1a,b in the **Appendix (Section 6.6)** we plot the $f - \pi$ weights. The π -weights are for type I catches only, and have been standardized to sum to one. Several conclusions are apparent:

1. Tows in the deep-water strata (added to the survey after 1990) receive more f -weight than π -weight. This is because the deep-water strata tows are not included in the design-based estimates of type I trawlable stock size. On the other hand these f -weights are not too large because the regression estimate of stock size is based only on the shallow water region (300 series strata), and observations outside this region receive less weight than those within this region. This seems reasonable as well. We would be uncomfortable if the difference between model and design-based estimates was based on catches from deep water strata that were never sampled prior to 1991. On the other hand we cannot ignore catches in the deep water strata either. The approach we have taken weights the deep water catches reasonably.
2. Tows near the 3L boundary usually have smaller f -weights than π -weights. In light of the stock mixing concerns identified in **Chapter 2** for this area, smaller weights

for these catches seems appropriate.

3. Clusters of tows usually have smaller f -weights than π -weights. In Figures 6.1a,b this is evidenced by the higher frequency of spatial clusters of $\times$'s than $+$'s. This is also intuitively reasonable. Catches at tow sites very close together provide less information about total stock abundance than catches at tow sites that are spread throughout the survey region.
4. Tows that surround a large sampling "gap" have higher f -weights than π -weights. This is because these tows are more informative about trawlable abundance within the "gap" region that was not sampled. The design-based estimator has no "knowledge" of sampling gaps, whereas the model-based estimator does, and it uses this "knowledge" effectively.

The results in Figures 6.1a,b suggest that the model-based estimator of type I trawlable abundance weights the sampled catches in a reasonable manner relative to the design-based stratum-weighted average estimator.

6.3.3. Confidence intervals for μ^{II}

Recall that type II catches are assumed to be i.i.d. $NB(\mu^{II}, k^{II})$ random variables. It is easy to show that the sufficient statistic for μ^{II} is $Y^{II} = \sum_i Y_i^{II}$ if k is known. If the total number of type II catches is n^{II} , then $Y^{II}|n^{II} \sim NB(n^{II}\mu^{II}, n^{II}k^{II})$. Following standard statistical practice, we base inferences involving μ^{II} on $\sum_i Y_i^{II}$. If k^{II} is known, which we assume is the case, then $Y^{II}|n^{II}$ is pivotal for μ^{II} , and we can use it's distribution to construct confidence intervals for μ^{II} .

6.3.4. Confidence intervals for μ_t .

In this section we use the profile likelihood method to construct confidence intervals for

$$\mu_t = \mu_t^I + \mu_t^{II}, = \sum_r (1 - p^r) \mu_{tr}^I + N^{II} \mu_t^{II},$$

where $N^{II} = \sum_r p^r N_r$. The method is as follows. Let $l(\mu_t^I)$ and $l(\mu_t^{II})$ denote the loglikelihoods for μ_t^I and μ_t^{II} . These are found using the asymptotic normality result for $\hat{\mu}_t^I$ in **Section 6.3.2** and the distribution of $Y_t^{II}|n^{II}$ given in the previous section. We write these loglikelihoods explicitly below. If we fix μ_t as some constant c , the distributional approximation we use for confidence intervals is

$$pl(c) = -2 [\{l(c - \bar{\mu}_t^{II}) + l(\bar{\mu}_t^{II})\} - \{l(\hat{\mu}_t^I) + l(\hat{\mu}_t^{II})\}] \sim \chi_1.$$

Here $\bar{\mu}_t^{II}$ is the value for μ_t^{II} that maximizes $l(c - \mu_t^{II}) + l(\mu_t^{II})$. A $(1 - \alpha)100\%$ confidence interval for μ_t is given as the smallest (U_t) and largest (L_t) values of c such that $pl(U_t)$ and $pl(L_t) \geq \chi_{1,1-\alpha}$.

More specifically, the loglikelihoods are

$$l(\mu_t^I; \hat{\mu}_t^I) = \frac{1}{2} \left[\log(2\pi) + \log \{v_n(\sigma_N^2 + d_n)\} + \frac{(\hat{\mu}_t^I - \mu_t^I)^2}{v_n(\sigma_N^2 + d_n)} \right],$$

and

$$\begin{aligned} l(\mu_t^{II}; y_t^{II}) &= y_t^{II} \log \left\{ \frac{\mu_t^{II}}{\mu_t^{II} + N^{II} k^{II}} \right\} + n^{II} k^{II} \log \left\{ \frac{N^{II} k^{II}}{\mu_t^{II} + N^{II} k^{II}} \right\} \\ &\quad + \Psi(y_t^{II} + n^{II} k^{II}) - \Psi(n^{II} k^{II}) - \Psi(y_t^{II} + 1). \end{aligned}$$

Recall that $\Psi(\cdot) = \log\{\Gamma(\cdot)\}$.

The f_i 's are not quite the same as in **Section 6.3.2**. In that section we assumed there

were no type II catches (i.e. $p = 0$), whereas in this section we have these catches. The appropriate f 's are

$$f_i = \frac{\sum_{r=1}^3 (1 - p_r) \sum_{h=1}^{H_r} \frac{N_h}{M_h} \sum_{j=1}^{M_h} W(\mathbf{x}_j - \mathbf{x}_i)}{\sum_{r=1}^3 (1 - p_r) N_r}$$

The profile likelihood confidence intervals are presented in Table 6.3 and plotted in Fig-

Table 6.3. Mixture model estimates of annual trawlable stock abundance, $\hat{\mu}_t$, with 95% confidence intervals (L,U). Numbers are in 1000's.

year	L	$\hat{\mu}_t$	U	year	L	$\hat{\mu}_t$	U
84	80198	105349	137994	90	28580	33834	40096
85	47237	60708	76588	91	17627	22087	27222
86	34420	42244	51352	92	17493	24880	33082
87	137871	189537	249352	93	9246	48638	90648
88	35304	42559	51412	94	10180	19376	29590
89	26946	31799	37525				

ure 6.2. The difference in the mixture-model estimates and the stratum-weighted averages has been discussed previously. In Figure 6.2 we see that the mixture-model confidence intervals are substantially smaller in many years compared to the design-based intervals. In some years the mixture-model confidence intervals are slightly larger. This is because of the added uncertainty from type II catches in those years. The confidence intervals are slightly asymmetric, with the upper interval width ranging from 3-12% greater than the lower interval width.

In Section 6.4 we test the asymptotic procedures we develop for inferences about the population total type I abundance. We demonstrate that the inference procedures work very well. We feel the chi-square approximation for the profile likelihood statistic is good. This is because the normal approximation for $\hat{\mu}_t^I$ works well, as demonstrated

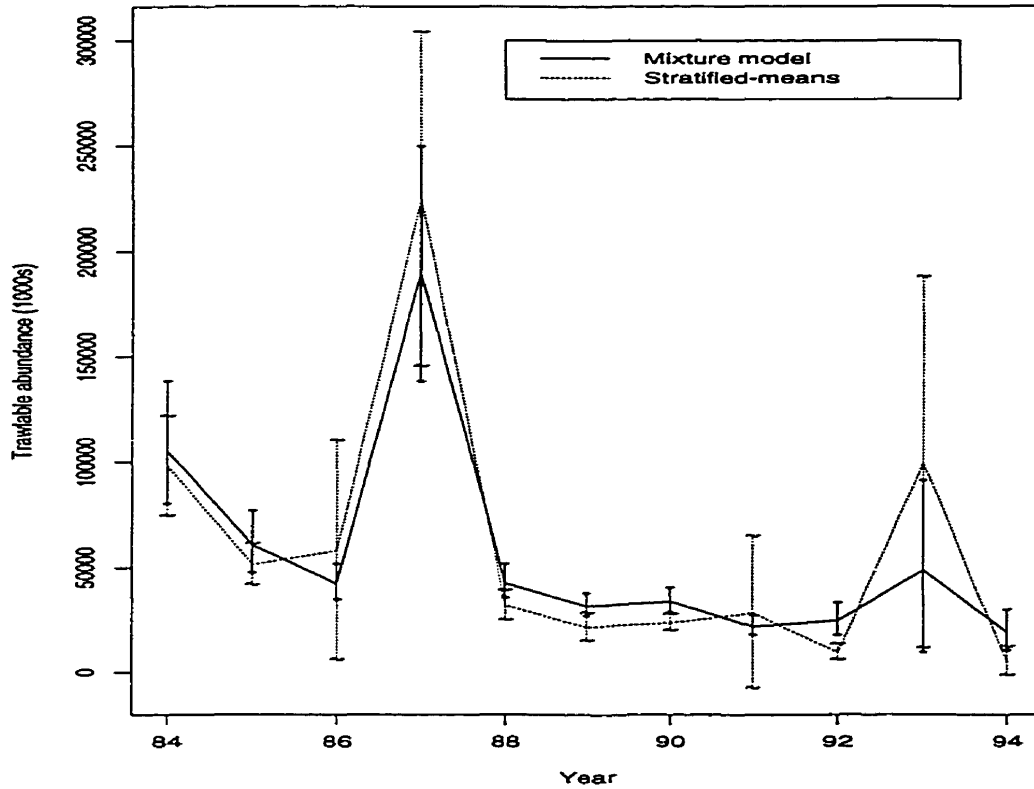


Figure 6.2: Annual estimates of trawlable stock abundance from the mixture model, and stratum area-weighted averages.

in the simulations. Also, $\hat{\mu}^{II}$ has an approximately normal distribution because it is just the average of type II catches. Although the number of type II catches is too small to draw on the central limit theorem for the approximate normality of $\hat{\mu}^{II}$, we can utilize the fact that if $Y \sim NB(m, k)$ then $Y \Rightarrow N\{m, m(1 + m/k)\}$ as $m \rightarrow \infty$. The mean of the type II catches is large, and these catches are approximately normally distributed. Hence, $\hat{\mu}^{II}$ is just an average of six approximately normal rv's, and is itself approximately normally distributed. Our inferences may be overly optimistic in years when the total type I abundance is small, because we do not account for uncertainty in the mixture proportions. This is hard to do with the small number of type II catches we have identified.

We could easily profile with the mixture proportions to get confidence intervals for μ_t , but the hard part is specifying the distribution of the mixture indicator random variables. In some years total type I abundance is large (1984-88, 1993), and the unaccounted mixture proportion uncertainties will probably have little effect on inferences, because total type II abundance is less than 25% of total abundance.

6.4. Simulation

We consider a simple population to test the accuracy of the confidence intervals used in Section 6.3. We feel that our inferences concerning type II catches are sound, and that the profile likelihood method in Section 6.3 is reliable if (6.3.8) is a good approximation. This approximation will not be good if substantial bias is present, or if we have chosen the wrong values for k . To test this we created a known population consisting of 10 000 values of densities μ 's and covariates X_1 and X_2 . The population is plotted in Figure 6.3.

This represents a hypothetical population of type I trawl densities. The average density for this population is $\bar{\mu} = 6.145$. Our simulation consists of drawing random samples of size n from the population of 10000 $\{X_1, X_2, \mu(X_1, X_2)\}$'s, then randomly generating $NB(\mu, k)$ observations. We use $n = 100$ and 200 , and $k = 2$. The values of n bound the sample sizes in the 3NO survey. We do this using 2000 simulations for each n and k .

We compute confidence intervals for $\bar{\mu}$ and observe whether $\bar{\mu}$ exceeds these intervals for each simulation. We use $k = 1, 2$ or 3 when computing these intervals to assess the impact mis-specifying k has on inferences. For convenience we chose a systematic sample of 1600 $\{X_1, X_2\}$'s for smoother predictions. In each simulation we find the $GC_p b_{\min}$'s, then compute confidence intervals using these bandwidths. The profile likelihood procedure does not have to be used because we are considering only type I stock size, and we can use (6.3.9) directly to compute confidence intervals for this quantity. These can be

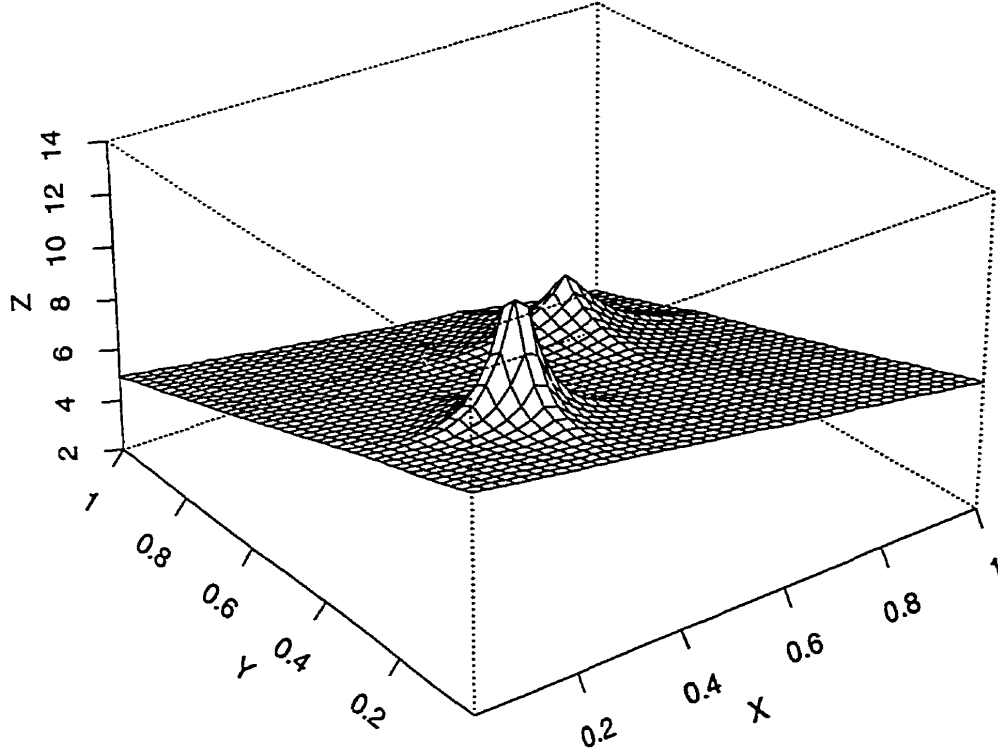


Figure 6.3: Simulation population.

solved analytically. If

$$\beta = \frac{\hat{\mu} + \frac{1}{2}v_n\chi_{1,1-\alpha}^2}{1 - \frac{1}{k}v_n\chi_{1,1-\alpha}^2} \text{ and } \gamma = \frac{v_n\chi_{1,1-\alpha}^2 d_n - \hat{\mu}^2}{1 - \frac{1}{k}v_n\chi_{1,1-\alpha}^2}$$

then a $(1 - \alpha)100\%$ confidence interval for $\bar{\mu}$ is obtained as $\beta \pm \sqrt{\beta^2 + \alpha}$.

The simulated exceedances are presented in Table 6.4. We count exceedances for $\alpha = 0.2, 0.1, 0.05$, and 0.01 . The results demonstrate that the confidence intervals work very well, especially when $\alpha > 0.01$. The actual coverage probabilities are very close to their nominal values, and are robust to the mis-specifications of k we considered. The reason for this is the confounding between bandwidths and k . If we assume to large a

Table 6.4. Simulated exceedances of $\bar{\mu}$ outside $(1 - \alpha)100\%$ confidence intervals. The proportion of times $\bar{\mu}$ exceeds the lower and upper bounds are presented in parantheses. The standardized bias in the simulated estimates is also presented. $\bar{\mu} = 6.14$

			α			
			0.2	0.1	0.05	0.01
n	k	bias				
100	1	-0.176	(0.089,0.123)	(0.050,0.058)	(0.023,0.025)	(0.008,0.005)
	2	-0.167	(0.092,0.123)	(0.051,0.062)	(0.027,0.028)	(0.005,0.005)
	3	-0.160	(0.077,0.101)	(0.035,0.052)	(0.016,0.028)	(0.004,0.005)
200	1	-0.118	(0.088,0.111)	(0.040,0.052)	(0.021,0.021)	(0.003,0.006)
	2	-0.158	(0.085,0.116)	(0.041,0.053)	(0.020,0.026)	(0.006,0.005)
	3	-0.142	(0.088,0.121)	(0.047,0.062)	(0.023,0.026)	(0.005,0.006)

value for k then σ_N^2 in (6.3.9) will be too small. At the same time v_n will also be too large, so that our estimate of $Var(\bar{\mu}_n)$ may still be reasonable. The simulation results suggest that this is the case.

We also computed the standardized bias

$$\frac{E(\hat{\mu}_N) - \bar{\mu}_N}{\{\sum_{i=1}^n f_i^2 \sigma_i^2\}^{1/2}},$$

where the expectation is the average over the simulated data sets. This incorporates both randomness in the data and randomness in the site selections. The bias is small, but not completely inconsequential. The sign and magnitude of the bias is consistent with the lower than nominal coverage probabilities for the lower tail, and the higher than nominal coverage probabilities for the upper tail. These differences are minor however, and bias correction, if possible, does not seem necessary.

We also analyzed the decrease in confidence interval widths caused by increasing the sample size. The widths were 30% smaller for all values of k and α .

We performed another simulation to assess the improvements gained by knowing the covariates for the entire population. In this simulation we used the smoother to make predictions at all the 10000 covariates in the population. This is computationally more difficult, so we only performed 1000 simulations for each value of n and k . The results are presented in Table 6.5. The standardized bias is smaller than in Table 6.5. The coverage

Table 6.5. Simulated exceedances of $\bar{\mu}$ outside $(1 - \alpha)100\%$ confidence intervals. The proportion of times $\bar{\mu}$ exceeds the lower and upper bounds are presented in parantheses. The standardized bias in the simulated estimates is also presented. $\bar{\mu} = 6.14$

			α			
			0.2	0.1	0.05	0.01
n	k	bias				
100	1	-0.095	(0.102,0.111)	(0.061,0.061)	(0.033,0.025)	(0.009,0 .005)
	2	-0.064	(0.115,0.101)	(0.062,0.045)	(0.024,0.016)	(0.007,0 .002)
	3	-0.120	(0.105,0.113)	(0.053,0.054)	(0.026,0.028)	(0.007,0.003)
200	1	-0.019	(0.103,0.093)	(0.058,0.057)	(0.028,0.032)	(0.006,0.006)
	2	0.024	(0.114,0.082)	(0.061,0.033)	(0.033,0.016)	(0.004,0.003)
	3	-0.019	(0.112,0.097)	(0.055,0.047)	(0.027,0.021)	(0.009,0.004)

probabilities for the two tails are more similar, due to the smaller bias. The improvements are not substantial however. This suggests that knowledge of covariate values for a large sample of the population is almost as good as knowledge of the covariates for the entire population. In our application we choose roughly 1600 covariates for prediction purposes, but many more are available. Our simulation results suggest that including more covariates for predictions may not improve our inferences much.

The confidence intervals we developed in Section 6.3.3 are conditional on the sampled covariates. Our simulation was based on random covariates. It demonstrates that the asymptotic bias will tend not to be a problem for a random sampling design; however, there are many samples in which the bias could be a problem. All one has to do is imagine

the bias that results if none of the samples sites fall on one of the “bumps” in Figure 6.3. Fortunately the 3NO research survey design is robust against such samples because of the three-stage random sampling design employed. This ensures tow sites are spread out.

6.5. Conclusions

In this chapter we have pulled together all of the results presented in Chapters 4-5 to produce a model-based estimate of trawlable stock size in NAFO Divisions 3NO. This estimate appears to be more consistent with the visual impression of changes in abundance that one gets from examining Figures 4.1a,b in Chapter 4, and in this sense the estimates are more realistic than stratum area-weighted averages. This is especially true for the 1992-1994 survey results. The trends from the design-based estimates of trawlable abundance in these years seem implausible based on stock dynamics, and are inconsistent with trends in catches in most areas sampled. The interannual variability in the mixture model estimates is generally less than for the stratified averages, which seems more reasonable. We have also considered the asymptotic properties of the nonparametric estimator of type I abundance, and we have developed methods for constructing confidence intervals for type I and type II abundances, and type I + type II abundance. The mixture model confidence intervals suggest much greater certainty about trawlable stock size in some years.

Our analyses suggest that our model for the incidence of type II catches is reasonable; that is, the incidence of these catches seems to vary spatially. This conclusion is statistically significant, although our sample size is quite small and one can legitimately question the power of our test. We have presented some results that suggest that our estimate of type II trawlable stock size is robust to the mixture probabilities if these probabilities are reasonably consistent with the location of the type II catches we have observed.

The simulation experiment was designed to test the accuracy of our inference results for

type I stock size. The population we considered had two inferentially difficult features: i) a spatial trend in density, and ii) two large density “bumps”. We feel the annual population distribution of type I catches in 3NO are no more difficult than the one used in our simulation. The simulation results are encouraging because the coverage probabilities of the confidence intervals were very close to their nominal values, even when k is misspecified. Bias associated with the nonparametric approach was not a problem in the simulation. This is because covariates were randomly sampled in each iteration. This simulation feature is relevant to our problem, because random sampling is used each year to allocate tow sites for surveying. The simulation results do not suggest any reason why the coverage probabilities for the mixture-model confidence intervals in Figure 6.2 might, in every year, consistently depart from their nominal values. The simulation results also demonstrate the benefit of using a random sampling design in the 3NO cod survey; this is reliable inference, at least for type I stock size.

The concept of design-robustness in model-based inference is not new. Särndal *et al.* concluded that linear regression estimators of population totals are approximately design-unbiased no matter what the shape of the finite population point scatter is. Firth and Bennett (1998) have shown that a wider class of regression estimators also have this property. The importance of choosing the correct model is in terms of precision; it is critical that the correct model is chosen to achieve small variance. The point is that a good probability sampling design is a prerequisite for this design-robustness property. The estimator of type I stock size we have chosen is not design-unbiased, for several reasons. Firth and Bennett (1998) show that this generally holds for nonparametric local likelihood models. We have not investigated design-consistency, but a basic requirement would be that our $N \times f$ -weights are asymptotically equivalent to the π -weights, which does not appear reasonable. The design-robustness property that we invoke, and that we think is really useful, is that the standardized bias is small under repeated sampling. It is difficult

to see how this can be achieved without a sampling design that ensures that, on average, $\bar{\mu}_n$ is close to $\bar{\mu}_N$ relative to the variance. This does not ensure design-consistency, but it does ensure reasonable inferences, as our simulation suggests.

6.6. Appendix: Figures

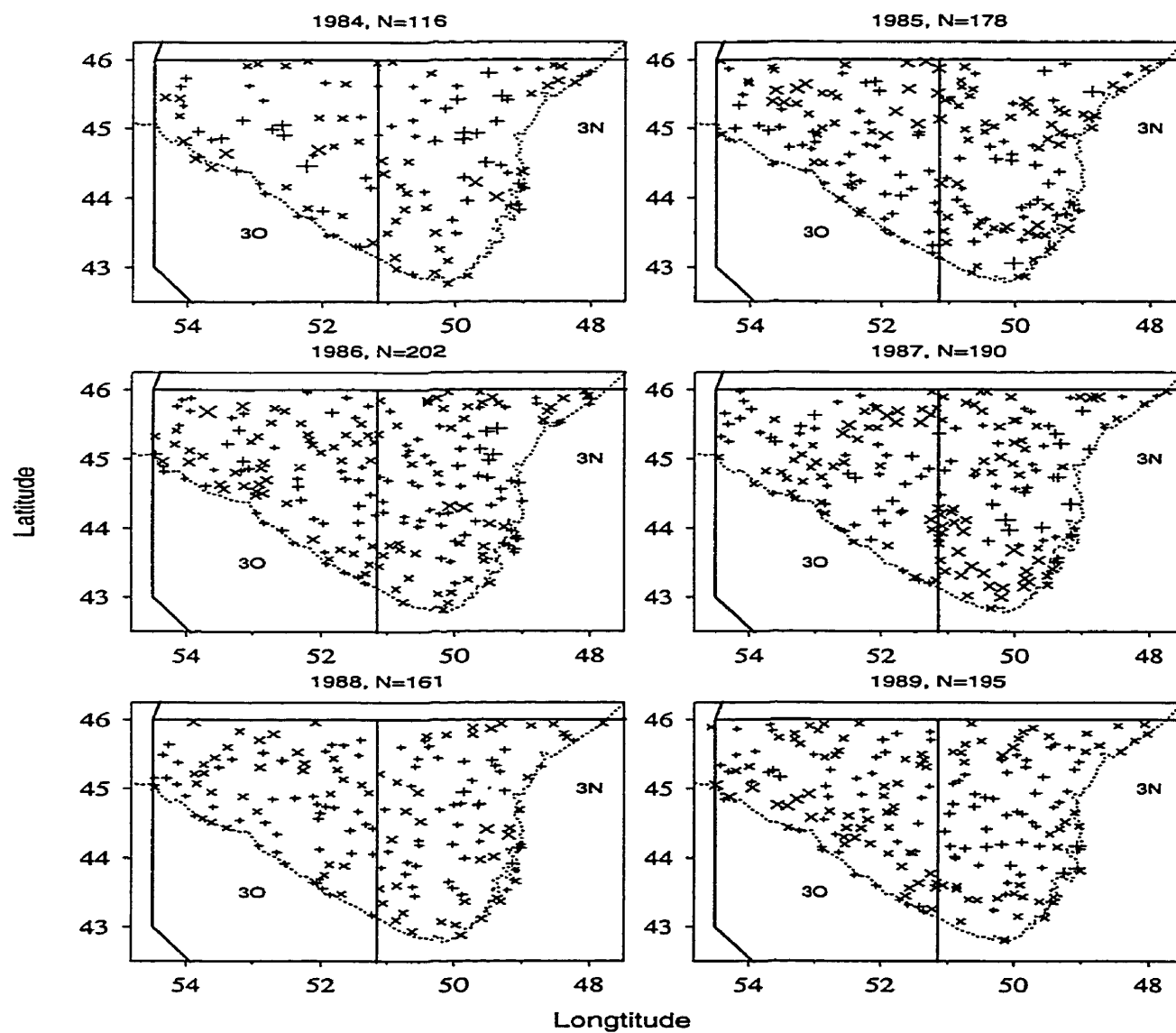


Figure 4.2a: The difference between model f -weights and design π -weights for estimates of type I trawlable stock abundance. An f -weight that is greater than a π -weight is plotted as a +.

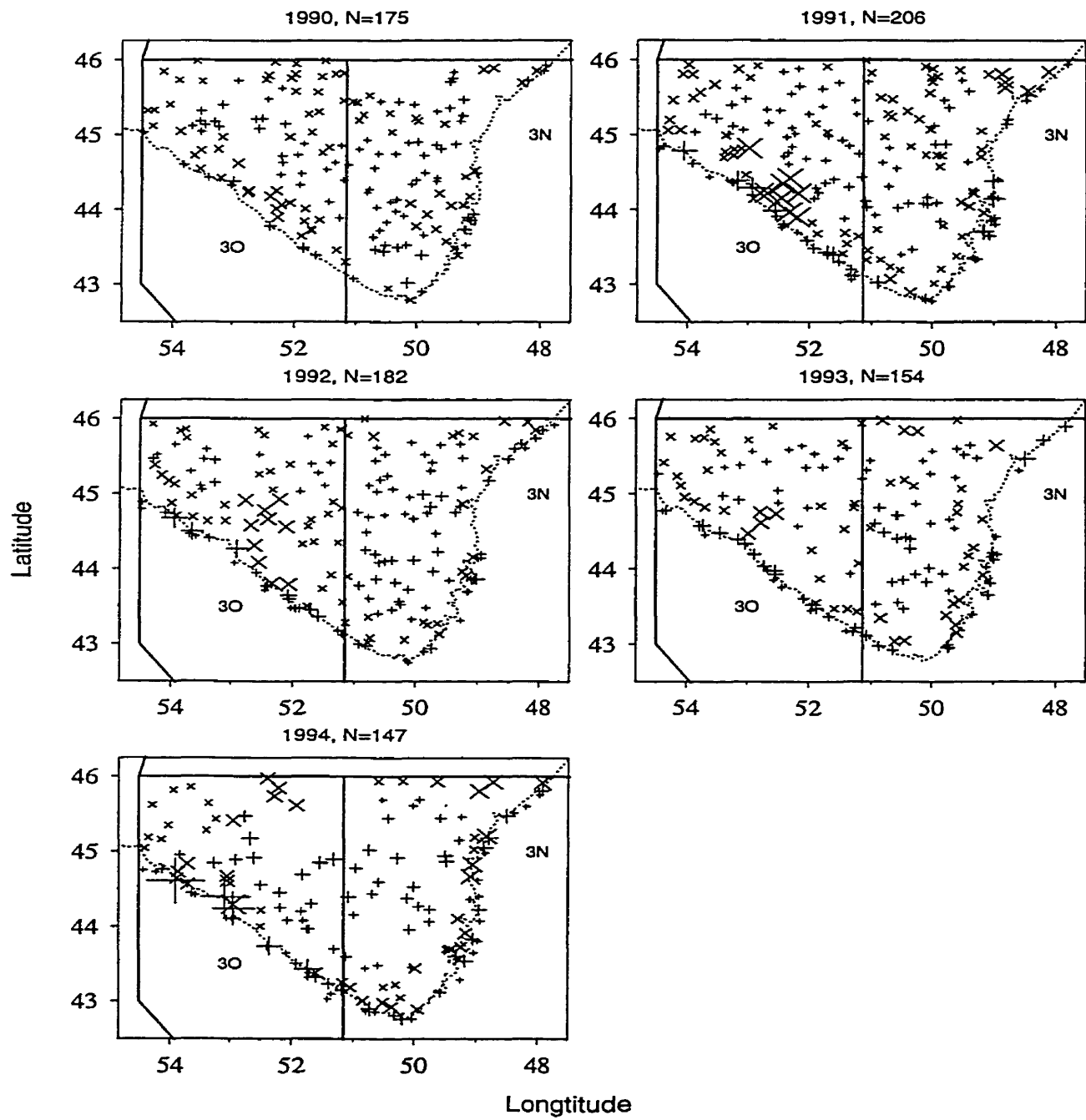


Figure 6.1b. Same as Figure 6.1a.

7. SUMMARY, AND FUTURE RESEARCH

7.1. Summary

The 3NO research surveys provide information about the abundance of cod in NAFO Divisions 3N and 3O, for the component of the population that is available to be sampled by the fishing gear. This component is referred to as the trawlable population. This population depends on the gear type used when surveying; for example, if a survey trawl with a large mesh size is used then the trawlable population will be the large sized part of the total population. In this dissertation we considered how to efficiently summarize the information about annual trawlable abundance from the 3NO research survey. We did this using estimates and confidence intervals. The methods we used are applicable to other cod stocks.

Estimates of trawlable abundance are an important source of information for the management of many commercially exploited fish stocks. In **Chapter 2** we described common models used to translate information on trawlable abundance to absolute stock abundance through the use of production models and age-structured sequential population models. These models use other sources of information on stock abundance, most importantly time series of commercial catches. A critical assumption with these methods is that the survey region covers only one cod stock. This allows us to translate relative information about the size of fish within the survey region to the absolute size of the stock. The assumption that the survey region covers only one stock is never strictly true for highly mobile species

like cod, because there are no physical boundaries in the ocean that correspond exactly to the boundaries of the survey region. In Chapter 2 we described scientific studies that suggested that the cod within 3NO are a stock. However, the 3NO research survey information we analyzed suggested that incursions from other stocks may occur. For example, this may be why average survey catches were so high in 1987.

Remark 8. Additional biological information is required to assess the stock origin of trawl catches. Age and length compositions samples are routinely collected during the surveys and are useful for this purpose.

The stock concept defines a closed population with respect to reproduction. One reason the stock concept is important for management is that it allows us to project what effect a harvesting strategy has on the population. Information on the stock reproduction capability is critical in medium to long term stock size projections.

The methods currently used by many fisheries agencies, including DFO, to summarize survey catches are based on the survey design. We considered the 3NO survey design in Chapter 3. There are subjective elements to the design, so it is not strictly a probability sampling design. The survey design is stratified three-stage cluster sampling, with one sample per cluster. Design-based methods for inference are complicated for the 3NO research survey because of i) the mobility of the population, ii) measurement errors in the catches, iii) the inability to sample in some locations, and iv) the one-per-cluster sampling scheme. These difficulties also affect other fisheries surveys. We find these to be compelling reasons to pursue a model-based approach for inferences. This is not to say that the survey design is irrelevant. In fact, random allocation of sampling in the survey design is important for reliable model-based inferences. We also conclude that the stratification scheme in 3NO (see Figure 1.4) is related to cod abundance, and it is not a trivial task to build a model that can account for more variation in trawl catches than the 3NO strata can. The smoothing approach we used, with different bandwidths for the

slope and shelf regions, did account for more variation, but we were not able to do this using the same bandwidths for both regions.

We approached model-building ingenuously. We did not assume which covariates were relevant a priori, as is often the case in fisheries. We had compelling reasons for this. Most notably was the research by Smith and Page (1996) for cod on the Scotian Shelf in 4W and 4Vs (see Figure 1.3). They showed the potential for hydrographic signal in research survey catches that is unrelated to fish abundance. This is caused by changes in the catchability of fish by the survey gear, or changes in the trawlable fraction of the stock available to the gear. Another important paper is by Warren (1997). He showed for cod in 2J, 3K, and 3L that the time between tows can account for as much variation in catches as the spatial distance between tows. The confounding between distance and time between tows also exists in 2J3KL cod, so we feel Warren's conclusions probably can be questioned, but nonetheless they raise the possibilities of under-counting and double-counting of fish. These are important concepts to understand in order to produce precise information about abundance from research surveys. Failure to account for fish movements means that our inferences can be overly optimistic. It is not difficult to imagine how even small movements of fish could substantially change estimates of abundance, which reflects the uncertainty caused by the mobility of the population. Unfortunately it is difficult to assess the existence and importance of these processes using survey data alone. It is also difficult to adjust inferences to accommodate these processes.

Remark 9. *Additional information is usually required to adjust for sources of variation unrelated to abundance. Tagging information could be valuable to address the mobility issue.*

In Chapter 4 we used graphical techniques to explore for covariate dependencies in trawl catches. We concluded that some of the covariates routinely measured in the survey of 3NO did not seem to be related to trawl catches of cod. We also concluded

that six covariates were potentially related to trawl catches. The relationships were not simple however, which suggested a nonparametric approach to modelling was suitable. A problem with the graphical approach is that we could not decide if covariates were related to trawl catches, or to other covariates that were truly related to trawl catches. We called this confounding. We employed high-dimensional analytic nonparametric variable selection techniques to assist with this problem. Our graphical analyses also revealed that the distribution of survey catches is very skewed. It often occurs that average catches exceed the 75th catch percentiles. This is related to a small number of very large catches which have great influence on inferences. We concluded that these should be treated separately. Our graphical analyses indicated that these catches occurred apparently at random in terms of covariates. Based on these results we proposed a separate model for large catches based on simple habitat-selection theory. A mixture-model was proposed to combine the nonparametric model for most catches (type I's) and the model for extremely large catches (type II's).

In Chapter 5 we collected all the analyses we felt were necessary to construct a defensible model for trawl catches. This included the identification of covariates to include in our nonparametric model for type I catches, and the identification of type II catches. We used robust nonparametric smoothing techniques to assist in identifying type II catches, and for robust variable selection. We assumed a Negative Binomial stochastic model for trawl catches, which has been suggested in the fisheries literature. For robust smoothing we developed a method for robust estimation of the mean for i.i.d. Negative Binomial samples. We also tested the robust methods on simulated data to assess their performance with sample sizes relevant to our problem. These results were used to develop a robust Negative Binomial smoother. We also considered the amount of smoothing in Chapter 5. We found bandwidths that resulted in reasonably smooth model fits but also accounted for reasonable amounts of variation in trawl catches. We used a combination of predictive

and “significance” methods for bandwidth selection.

We used smoothing techniques to locally “orthogonalize” covariates to deal with confounding between some of them. A subjective decision had to be made about which covariates to orthogonalize. This decision has great influence on which covariates are included in the model via variable selection routines. This is a difficult issue that we discussed at length in Chapter’s 4 and 5. In the end we decided on practicality grounds to orthogonalize those covariates that were less useful for predictive purposes.

We concluded in Chapter 5 that spatial location and tow depth were the most important covariates to include in our nonparametric smoother for type I catches. We also selected bandwidths for our smoother. We compared this model with one based on the survey design, in which a separate mean was estimated for each stratum. The nonparametric model performed better, in terms of explaining more variation with more error degrees of freedom, and in a predictive sense. We also concluded that six of the 1912 tows in the survey time series we examined were type II catches. The spatial incidence of type II catches suggested a modification to our mixture-model to allow for different mixing proportions for three large geographic regions in 3NO. This modification was consistent with life-history information about cod in this region, because dense spawning aggregations of cod in 3NO tended to form in warm water on the continental slopes.

Remark 10. Additional information about patch size in 3NO could be used to refine the mixture-model, and make it more realistic. Acoustics can supply this information.

Having constructed an informative model for trawl catches, the remaining task was to construct confidence intervals. This was the subject of Chapter 6. We used our mixture-model to estimate trawlable abundance in 3NO by predicting local trawlable density at a large number of random tow stations, and adding these predictions to estimate total trawlable abundance. The mixture suggested that approximately 20000 tonnes of cod annually form dense type II aggregations in 3NO. The mixture-model estimates of

abundance appeared more consistent with the visual impression one gets from examining plots of survey catches than did stratum area-weighted averages.

We developed asymptotic inferential theory for total type I abundance, and combined this with survey information about type II abundance using likelihood methods to produce approximate confidence intervals for total trawlable abundance. We used the profile-likelihood method for this purpose. The mixture-model confidence intervals were usually considerably smaller than $\pm 2 \times$ design-based standard errors. A small simulation experiment indicated that confidence intervals for type I stock size work well. A difficult issue we faced when building our Negative Binomial smoother was estimating the Negative Binomial overdispersion parameter k . Fortunately, the simulation experiment also indicated that reasonable mis-specification of k had little effect on coverage properties of our confidence intervals. This is a somewhat surprising result, but is related to confounding between bandwidths and k . This is discussed more fully in Chapter 6.

Our nonparametric estimator of average type I trawlable density can be formulated as a simple weighted average. We compared these weights with the corresponding design weights based on sample inclusion probabilities. The model-based approach weights catches in a more reasonable manner than the design-based approach for our data. Boundary observations and clusters of observations tended to receive less weight, while catches in regions sparsely sampled tended to receive more weight.

The issue of whether to use a design- or model-based approach for the 3NO survey, apart from the mobility of the population, seems to rest with sampling intensity. The sampling intensities currently used by DFO are insufficient for cod abundance to be estimated with much precision. Increasing sample size does not seem to be an option either. It is not clear than even twice the sampling intensity would improve this situation substantially, whereas the costs for this increase would be prohibitive. In the model-based approach we add information to increase precision - at the possible loss of accuracy. The

model-based approach allows us to channel “common scientific sense”, or scientific information obtained from experiments and experience, into inferences about stock size. There is a degree of subjectivity in doing this, and herein lies the drawback of the model-based approach. For example, with our mixture-model we assumed that the type II model component does not change annually. But what if it has, and how can we detect if it has? The conundrum is that to be robust to model mis-specification involves a loss in precision. We hope that the mixture-model provides a reasonable compromise for 3NO cod.

7.2. Future research

In our summary we have remarked on several occasions how additional information can be incorporated to improve the accuracy and precision of stock size inferences based on research survey data. This is one direction we propose for future research. For the results of this dissertation to be used in stock assessments, through estimation of SPA's, two things must occur. The first is that our estimates of trawlable stock size must be disaggregated by length and age. Length distributions are sampled from the catch of some survey tows. These are not selected entirely at random, because we do not know a priori what tows will catch cod. Typically most of the length samples are taken early in the survey. Population length distributions are currently estimated using catch-weighted averages of the length distribution samples. A natural direction for future research is to use the mixture-model for this weighting. The population age distribution is estimated using the length distribution estimate and an age-length key. This key is based on length-stratified age samples. The second thing we must do is develop a relationship between the trawlable stock components for the two types of trawls used in the 3NO survey during 1984-98. In 1995 the survey gear was changed to a shrimp trawl. The tow duration was also changed from 30 to 15 minutes. The shrimp trawl has a much smaller mesh size and is more effective at catching small fish. It also has been suggested that it is less effective

at catching larger fish than the trawl used till 1995. The effect of this is that the trawlable portion of the stock has changed with the change in gear type and tow duration. It is desirable to understand this relationship to use average survey catch rates for estimating SPA parameters.

The type II part of the mixture-model provides some temporal persistence in our estimates. Further research in this direction may be fruitful. Consider Figure 7.1. This is a plot of average trawl catches for each stratum in two consecutive years. There is considerable temporal persistence in catch rates between years. Type II catches are included in this figure, which partially accounts for the low temporal correlation's in some years. Figure 7.1 suggests further improvements in the precision of inferences about trawlable stock size may be possible using the temporal persistence in type I stock components.

The results in this dissertation can also be applied to other stocks, with considerable simplifications. In other applications it may not be necessary to use the analytic methods to identify type II catches. The analyses we conducted have shown that the potential for these analytic methods may be limited by sample size, and in the end a subjective decision about type II catches may be required. In this case one can make this decision from inspecting the data, and dispense with the analytic procedures without great loss. While variable selection is important, we have shown that it is difficult for trawl research surveys because of confounding. Again one could proceed with a model using spatial location and depth, at least with less peril than before our research. The same can be said for bandwidth selection and estimating the Negative Binomial overdispersion parameter, k . Our results suggest specific procedures that are useful (e.g. GC_p for bandwidths), and this could be used directly without comparisons with other methods (e.g. CV , GCV).

The cod stock to the west of 3NO in NAFO Division 3Ps (see Figure 1.3) is a good candidate for the mixture-model methods. This stock currently has the largest direct commercial cod fishery off the east-coast of Canada, with a catch quota of 20000 tonnes

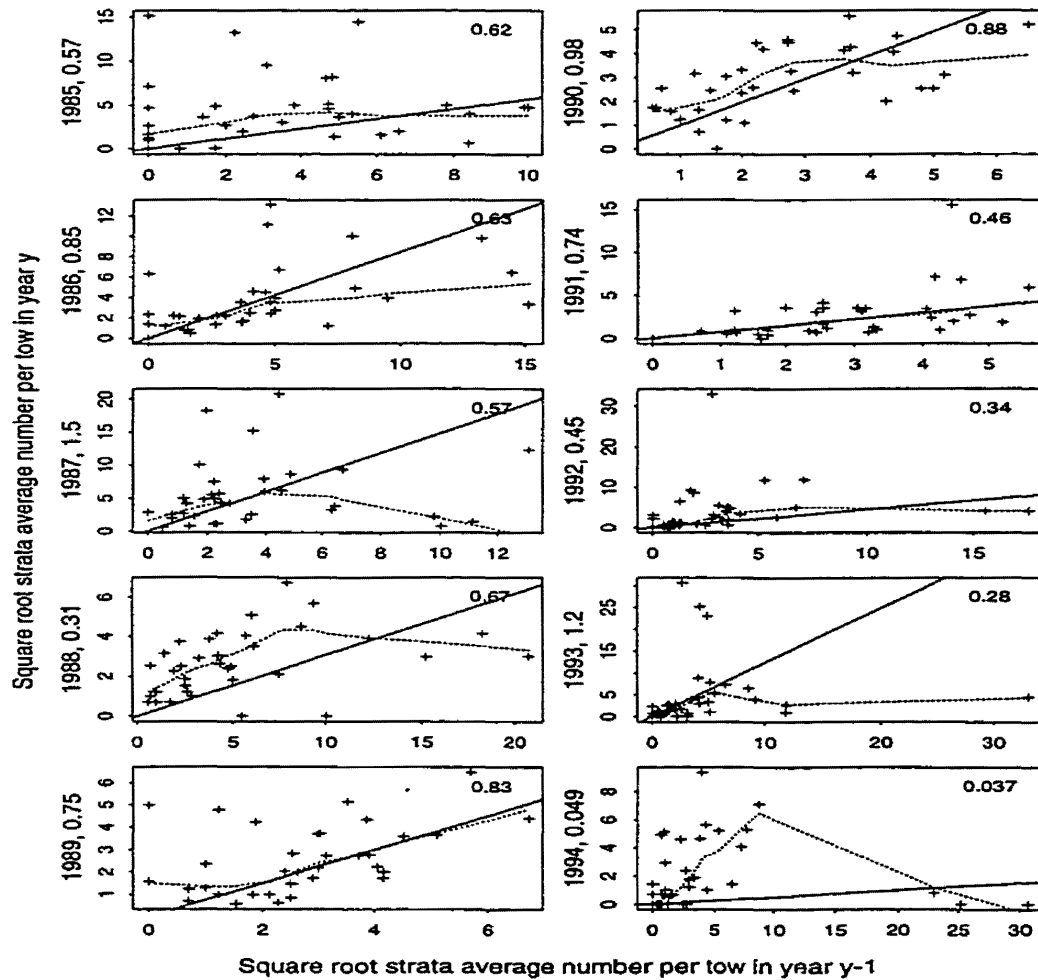


Figure 7.1: Between year temporal persistence in stratum-average survey catches. The correlation coefficient is listed in the upper right-hand corner of each panel. The slope of the regression line $Y = aX + \varepsilon$ is listed following the year. The solid line is a least-squares linear fit, the dotted line is an Splus *lowess* fit.

in 1997. In 1995 the catch in one survey tow was 80% of the total survey catch for that year. To understand the effect this catch has on our inferences about stock size, consider Figure 7.2. In this figure we plot survey catch rates and SPA model predictions. The SPA was based on the offshore component of 3Ps cod. Recall from Chapter 2

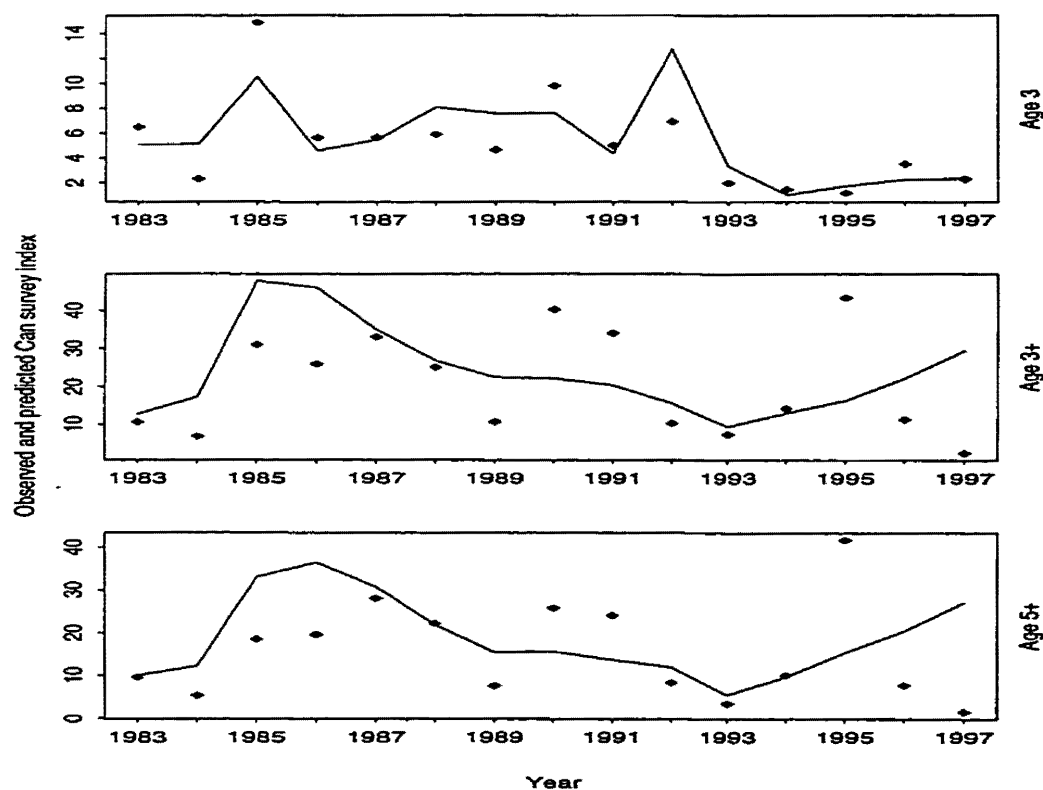


Figure 7.2: Observed (points) and predicted (lines) Canadian 3PS survey indices.

that the basic model used to estimate SPA parameters is $E(R_{ay}) = q_a N_{ay} = \mu_{ay}$, where q_a is the catchability coefficient and R_{ay} is an index of stock size. In this case R_{ay} is the traditional stratum area-weighted average catch-at-age. Quasi-likelihood methods, assuming $Var(R_{ay}) = \phi \mu_{ay}^2$, were used to estimate parameters, and for inferences. The key feature we wish to point out is the large discrepancy between the observed and predicted index in 1995. This is largely related to the effect of one extremely large tow in

this year. This leads to an over-inflated estimate of ϕ . It is easy to see that the average trawlable stock size estimator we propose would be more consistent with the SPA in 1995.

A potential implication of this large tow is lost yield of fish to the fishery. The reason for this is as follows. One approach in deciding catch quotas for the coming year is to assess the effect of quota options on projected stock size. The projection is based on an SPA and a recruitment model. If the probability is “large” that our stock size projection is less than some desired reference level then the quota option the projection is based on is judged too risky, and a lower option is considered. The most recent projected risk analysis for 3Ps is presented in Figure 7.3. The risk curves are very flat compared to other North Atlantic groundfish stocks. Part of the reason for this is the large catch in 1995 which results in a large coefficient of variation (ϕ) for the risk analysis. Catch quota’s must be lowered because of the large tow to maintain a small probability of not exceeding the target, which in this case is fishing mortality $< F_{0.1}$. This is lost yield to the fishery. Whether this is appropriate is debatable. The mixture-model suggests it is not reasonable. At the same time we should note that there are many sources of uncertainty that the SPA and projection methods do not explicitly consider, which may imply higher risks than indicated in Figure 7.3. The point is that lower projection precision means lost yield to the commercial fishery. This idea is well known in industrial settings, but is only now getting recognition in fisheries science and management. If we do not incorporate all of our certainties, as well as our uncertainties, into our risk analysis then we may be guilty of causing overly conservative harvesting strategies. This is a socio-economic realization of the impact that methods for inference can have.

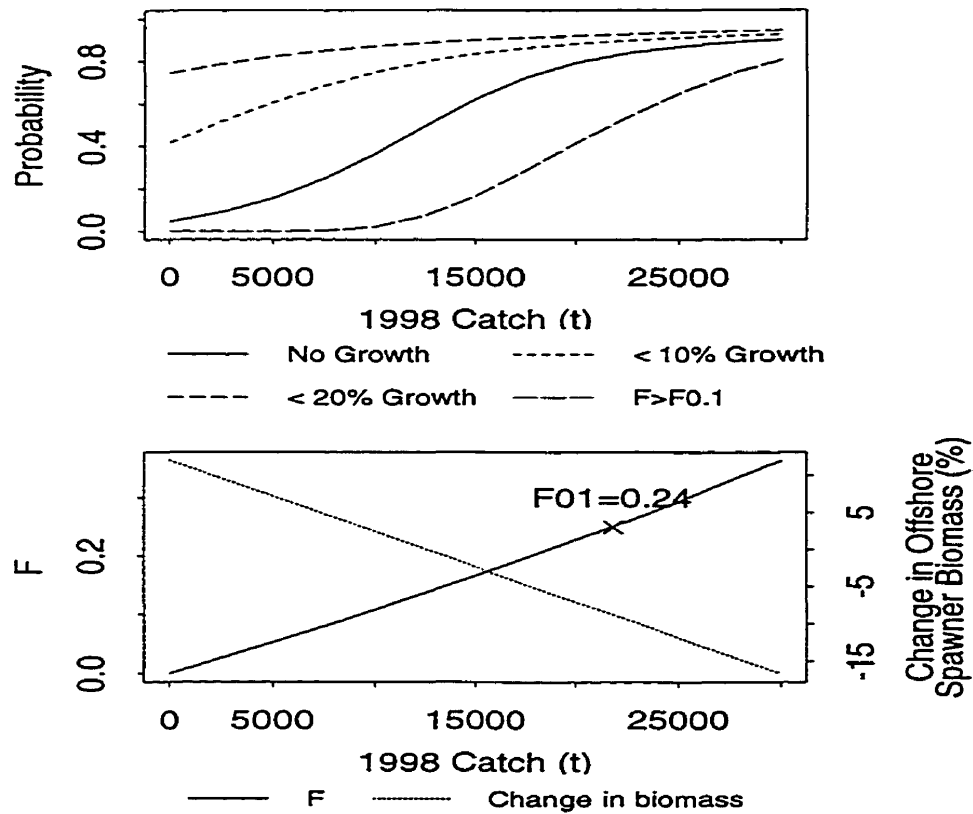


Figure 7.3: Profile quasi-likelihood based stochastic descriptions of the change in offshore 3PS spawner biomass (SB) and fishing mortality (F) for 7 TAC options. The top panel presents the probabilities that 1999 SB projections will not grow by more than 0, 10, and 20% of 1998 SB, along with the probability of exceeding $F_{0.1}$, which is taken as 0.24 for this stock. The bottom panel presents the estimated change in SB projection, and F , for the various TAC's.

8. BIBLIOGRAPHY

- Arreguín-Sánchez, F. (1996) Catchability: a key parameter for fish stock assessment. *Rev. Fish Biol. and Fish.*, **6**, 221-242.
- Atkinson, D. B., Rose, G. A., Murphy, E. F., and Bishop, C. A. (1997) Distributional changes and abundance of northern cod (*Gadus morhua*), 1981-1993. *Can. J. Fish. Aquat. Sci.*, **54**(Suppl. 1), 132-138.
- Barndorff-Nielsen, O. E., and Cox, D. R. (1994) *Inference and asymptotics*. London: Chapman and Hall.
- Bishir, J., and Lancia, R. A. (1996) On catch/effort methods of estimating animal abundance. *Biometrics*, **52**, 1457-1466.
- Bishop, C. A. (1994) *Revisions and additions to stratification schemes used during research vessel surveys in NAFO Subareas 2 and 3*. NAFO SCR Doc 94/43, Serial No. N2413.
- Boente, G., Fraiman, R., and Meloche, J. (1997) Robust plug-in bandwidth estimators in nonparametric regression. *J. Statist. Plan. Inf.*, **57**, 109-142.
- Booth, J. G., Butler, R. W., and Hall, P. (1994) Bootstrap methods for finite populations. *J. Am. Statist. Assoc.*, **89**, 1282-1289.
- Breckling, J. U., Chambers, R. L., Dorfman, A. H., Tam, S. M., and Welsh, A. H. (1994) Maximum likelihood inference from sample survey data. *Int. Statist. Rev.*, **62**, 349-363.
- Brockmann, M., Gasser, T., and Hermann, E. (1993) Locally adaptive bandwidth choice

- for kernel regression estimators. *J. Am. Statist. Assoc.*, **88**, 1302-1309.
- Byrne, C. J., Azarovitz, T. R., and Sissenwine, M. P. (1981) Factors affecting variability of research vessel trawl surveys, 258-273 in Doubleday, W. G., and Rivard, D. (eds.) Bottom trawl surveys. *Can. Spec. Publ. Fish. Aquat. Sci.*, No. 58.
- Carroll, R. J., and Ruppert, D. (1988) *Transformation and weighting in regression*. London: Chapman and Hall.
- Chambers, R. L., M., Dorfman, A. H., and Wehrly, T. E. (1993) Bias robust estimation in finite populations using nonparametric calibration. *J. Am. Statist. Assoc.*, **88**, 268-277.
- Chaudhuri, P., and Dewanji, A. (1995) On a likelihood-based approach in nonparametric smoothing and cross-validation. *Statist. Prob. Lett.*, **22**, 7-15.
- Clarke, B. R. (1983) Uniqueness and Frechet differentiability of functional solutions to maximum likelihood type equations. *Ann. Statist.*, **4**, 1196-1205.
- Cliff, A. D., and Ord, J. K. (1981) Spatial and temporal analysis: autocorrelation in space and time, 104-110 in Wrigley, N., and Bennett, R. J. (eds.) *Quantitative geography: a British view*. London: Routledge & Kegan Paul.
- Collie, J. S., and Sissenwine, M. P. (1983) Estimating population size with relative abundance data measured with error. *Can. J. Fish. Aquat. Sci.*, **40**, 1871-1879.
- Cressie, N. A. C. (1993) *Statistics for spatial data*. Rev. Ed. New York: John Wiley & Sons.
- Cushing, D. H. (1981) *Fisheries biology: A study in population dynamics*. Madison: The University of Wisconsin Press.
- Cushing, D. H. (1988) *The provident sea*. Cambridge: Cambridge University Press.
- de la Mare, W. K. (1984) On the power of catch per unit effort series to detect declines in whale stocks. *Rep. Int. Whal. Commn.*, **34**, 655-661.

- Deriso, R. B., Quinn, T. J., and Neal, P. R. (1985) Catch-age analysis with auxiliary information. *Can. J. Fish. Aquat. Sci.*, **42**, 815-824.
- Doubleday, W. G. (1976) A least squares approach to analysing catch at age data. *ICNAF Res. Bull.*, **12**, 69-81.
- Doubleday, W. G. (ed.) (1981) Manual on groundfish surveys in the Northwest Atlantic. *NAFO Sci. Coun. Stud.*, **2**, 7-55.
- Efron, B., and Tibshirani, R. J. (1993) *An introduction to the bootstrap*. London: Chapman and Hall.
- Engås, A. (1994) The effects of trawl performance and fish behavior on the catching efficiency of demersal sampling trawls, 45-68 in Fernö, A., and Olsen, S. (eds.) *Marine fish behavior in capture and abundance estimation*. Oxford: Fishing News Books.
- Firth, D., and Bennet, K. E.. (1998) Robust models in probability sampling. *J. R. Statist. Soc. B*, **60**, 3-21.
- Fitzmaurice, G. M. (1997) Model selection with overdispersed data. *Statist.*, **46**, 81-91.
- Fleming, A. M. (1960) Age, growth and sexual maturity of cod (*Gadus morhua* L.) in the Newfoundland area, 1947-1950. *J. Fish. Res. Board. Can.*, **17**, 775-809.
- Fournier, D. A., and Archibald, C. P. (1982) A general theory for analyzing catch at age data. *Can. J. Fish. Aquat. Sci.*, **39**, 1195-1207.
- Fournier, D. A., and Doonan, I. J. (1987) A length-based stock assessment utilizing a generalized delay-difference model. *Can. J. Fish. Aquat. Sci.*, **44**, 422-437.
- Freeman, S. N., and Kirkwood, G. P. (1995) On a structural time series method for estimating stock biomass and recruitment from catch and effort data. *Fish. Res.*, **22**, 77-98.
- Gavaris, S. (1980) Use of a multiplicative model to estimate catch rate and effort from

- commercial data. *Can. J. Fish. Aquat. Sci.*, **37**, 2272-2275.
- Gavaris, S. (1988) *An adaptive framework for the estimation of population size*. CAFSAC Res. Doc 88/29.
- Gavaris, S. (1988) Abundance indices from commercial fishing, 3-15 in Rivard, D. (ed.) *Collected papers on stock assessment models*. CAFSAC Res. Doc 88/61.
- Godambe, V. P. (1996) Estimation of parameters in survey sampling. *Can. J. Statist.*, **23**, 227-243.
- Godø, A. R. (1994) Factors affecting the reliability of groundfish abundance estimates from bottom trawl surveys, 166-199 in Fernö, A., and Olsen, S. (eds.) *Marine fish behavior in capture and abundance estimation*. Oxford: Fishing News Books.
- Gomes, M. (1993) *Predictions under uncertainty: Fish assemblages and food webs on the Grand Banks of Newfoundland*. ISER, Memorial University of Newfoundland.
- Gudmundsson, G. (1994) Time series analysis of catch-at-age observations. *Appl. Stat.*, **43**, 117-126.
- Gulland, J. A. (1983) *Fish stock assessment: A manual of basic methods*. Chichester: John Wiley & Sons.
- Gunderson, D. R. (1993) *Surveys of fisheries resources*. New York: John Wiley.
- Hall, P., and Jones, M. C. (1990) Adaptive M-estimation in nonparametric regression. *Ann. Statist.*, **18**, 1712-1728.
- Halliday, R. G., and Pinhorn, A. T. (1990) The delimitation of fishing areas in the North-west Atlantic. *J. Northw. Atl. Fish. Sci.*, **10**, 1-51.
- Hampel, F. R., Ronchetti, E., Rousseeuw, P. J., and Stahel, W. (1986) *Robust statistics: the approach based on influence functions*. New York: John Wiley.
- Härdle, W. (1991) *Smoothing techniques*. Springer-Verlag: New York.
- Härdle, W., and Gasser, T. (1984) Robust non-parametric function fitting. *J. R. Statist.*

- Soc. B*, 46, 42-51.
- Hastie, T. J., and Tibshirani, R. J. (1990) *Generalized additive models*. London: Chapman and Hall.
- Heessen, H. J. L., and Daan, N. (1994) Cod distribution and temperature in the North Sea. *ICES Mar. Sci. Symp.* 198, 244-253.
- Hilborn, R., and Walters, C. J. (1992) *Quantitative fisheries stock assessment: Choice, dynamics and uncertainty*. New York: Chapman and Hall.
- Huber, P. J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.*, 1, 221-233.
- Hutchings, J. A., and Myers, R. A. (1994a) What can be learned about the collapse of a renewable resource? Atlantic cod, *Gadus morhua*, of Newfoundland and Labrador. *Can. J. Fish. Aquat. Sci.*, 51, 2126-2146.
- Hutchings, J. A., and Myers, R. A. (1994b) Timing of cod reproduction: interannual variability and the influence of temperature. *Mar. Ecol. Prog. Ser.*, 108, 21-31.
- Hutchings, J. A. (1996) Spatial and temporal variation in the density of northern cod and a review of hypotheses for the stocks collapse. *Can. J. Fish. Aquat. Sci.*, 53, 943-962.
- Innis, H. A. (1954) *The cod fisheries: The history of an international economy*. Toronto: University of Toronto Press.
- Koeller, H. (1991) Approaches to improving groundfish survey abundance estimates by controlling the variability of survey gear geometry and performance. *J. Northw. Atl. Fish. Sci.*, 11, 51-58.
- Krewski, D., and Rao, J. N. K. (1981) Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Ann. Statist.*, 9, 1010-1019.

- Kulka, D. W., Wroblewski, J. S., and Narayanan, S. (1995) Recent changes in the winter distribution and movements of northern Atlantic cod (*Gadus morhua* Linnaeus, 1758) on the Newfoundland-Labrador shelf. *ICES J. Mar. Sci.*, **52**, 889-902.
- Li, K. (1987) Asymptotic optimality for C_p , C_l , cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, **15**, 958-975.
- Loeng, H. (1989) The influence of temperature on some fish population parameters in the Barents sea. *J. Northw. Atl. Fish. Sci.*, **9**, 103-113.
- MacCall, A. D. (1990) *Dynamic geography of marine fish populations*. Seattle: University of Washington Press.
- McCullagh, P., and Nelder, J. A. (1989) *Generalized linear models*. 2nd ed. London: Chapman and Hall.
- Mendelsohn, R. (1988) Some problems in estimating population sizes from catch-at-age data. *Fish. Bull.*, **86**, 617-630.
- Methot, R. D. (1989) Synthetic estimates of historical abundance and mortality for northern anchovy. *Amer. Fish. Soc. Symp.*, **6**, 66-82.
- Müller, H.-G., and Stadtmüller, U. (1987) Variable bandwidth kernel estimators of regression curves. *Ann. Statist.*, **15**, 182-201.
- Myers, R. A., and Cadigan, N. G. (1995a) Statistical analysis of catch-at-age data with correlated errors. *Can. J. Fish. Aquat. Sci.*, **52**, 1265-1273.
- Myers, R. A., and Cadigan, N. G. (1995b) Was an increase in natural mortality responsible for the collapse of northern cod. *Can. J. Fish. Aquat. Sci.*, **52**, 1274-1285.
- Myers, R. A., Hutchings, J. A., and Barrowman, N. J. (1997) Why do fish stocks collapse? The example of cod in Atlantic Canada. *Ecol. Appl.*, **7**, 91-106.
- Paloheimo, J. E. (1980) Estimation of mortality rates in fish populations. *Trans. Amer. Fish. Soc.*, **109**, 378-386.

- Paloheimo, J. E., and Dickie, L. M. (1964) Abundance and fishing success. *Rapp. Proc. Réunions, Con. Perm. Int. l'Exp. Mer*, **155**, 152-163.
- Pfeffermann, D. (1993) The role of sampling weights when modelling survey data. *Int. Statist. Rev.*, **61**, 317-337.
- Pinhorn, A. T. (1976) Living marine resources of Newfoundland~Labrador: Status and potential. *Bull. Fish. Res. Board. Can.*, **194**, 64 p.
- Pinhorn, A. T., and Halliday, R. G. (1990) Canadian versus international regulation of northwest Atlantic fisheries: management practises, fishery yields, and resource trends, 1960-1986. *N. Am. J. Fish. Manag.*, **10**, 154-174.
- Pitt, T. K., Wells, R., and McKone, W. D. (1981) A critique of research vessel otter trawl surveys by the St. John's Research and Resources Services, 42-61 in Doubleday, W. G., and Rivard, D. (eds.) Bottom trawl surveys. *Can. Spec. Publ. Fish. Aquat. Sci.*, No. 58.
- Pope, J. G. (1988) Collecting fisheries assessment data, 63-82 in Gulland, J. A. (ed.) *Fish population dynamics*, 2nd Ed. Chichester: John Wiley & Sons.
- Pope, J. G., and Shepherd, J. G. (1982) A simple model for the consistent interpretation of catch-at-age data. *J. Cons. int. Explor. Mer.*, **40**, 176-184.
- Rao, J. N. K., and Wu, C. F. J. (1988) Resampling inference with complex survey data. *J. Am. Statist. Assoc.*, **83**, 231-241.
- Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992) Some recent work on resampling methods for complex surveys. *Surv. Meth.*, **18**, 209-217.
- Reed, W. J. (1986) Analyzing catch-effort data allowing for randomness in the catching process. *Can. J. Fish. Aquat. Sci.*, **53**, 2157-2166.
- Reed, W. J., and Simons, C. M. (1996) Analyzing catch-effort data by means of the Kalman filter. *Can. J. Fish. Aquat. Sci.*, **43**, 174-186.

- Reed, W. J., and Simons, C. M. (1996) A contagion model of a fishery and its use in analyzing catch-effort data. *Ecol. Model.*, **92**, 179-191.
- Richards, L. J., and Schnute, J. T. (1992) Statistical models for estimating CPUE from catch and effort data. *Can. J. Fish. Aquat. Sci.*, **49**, 1315-1327.
- Ronchetti, E., and Staudte, R. (1994) A robust version of Mallow's C_p . *J. Am. Statist. Assoc.*, **89**, 550-559.
- Ronchetti, E., Field, C., and Blanchard, W. (1997) Robust linear model selection by cross-validation. *J. Am. Statist. Assoc.*, **92**, 1017-1023.
- Rose, G. A., Atkinson, D. B., Baird, J., Bishop, C. A., and Kulka, D. W. (1994) Changes in distribution of Atlantic cod and thermal variations in Atlantic waters, 1980-1992. *ICES Mar. Sci. Symp.* **198**, 542-552.
- Rosenberg, A. A., and Beddington, J. R. (1988) Length-based methods of fish stock assessment, 83-103 in Gulland, J. A. (ed.) *Fish population dynamics*, 2nd Ed. Chichester: John Wiley & Sons.
- Royall, R. M., and Cumberland, W. G. (1981) An empirical study of the ratio estimator and estimators of its variance. *J. Am. Statist. Assoc.*, **76**, 66-77.
- Ruppert, D., and Wand, M. P. (1994) Multivariate locally weighted least squares regression. *Ann. Statist.*, **22**, 1346-1370.
- Särndal, C., Swensson, B., and Wretman, J. (1992) *Model assisted survey sampling*. New York: Springer-Verlag.
- Scott, W. B., and Scott, M. G. (1988) Atlantic fishes of Canada. *Can. Bull. Fish. Aquat. Sci.*, **213**, 731 p.
- Serchuk, F. M., Grosslein, M. D., Lough, R. G., Mountain, D. G., and O'Brien, L. (1994) Fishery and environmental factors affecting trends and fluctuations in the Georges Bank and Gulf of Maine Atlantic cod stocks: an overview. *ICES Mar. Sci. Symp.*

198, 77-109.

- Shepherd, J. G. (1988) Fish stock assessments and their data requirements, 35-62 in Gulland, J. A. (ed.) *Fish population dynamics*, 2nd Ed. Chichester: John Wiley & Sons.
- Shepherd, J. G., and Nicholson, M. D. (1986) Use and abuse of multiplicative models in the analysis of fish catch-at-age data. *Statistician*, **35**, 221-227.
- Simpson, D. G., Carroll, R. J., and Ruppert, D. (1987) M-estimation for discrete data: asymptotic distribution theory and implications. *Ann. Statist.*, **15**, 657-669.
- Sinclair, A., Gascon, D., O'Boyle, R., Rivard, D., and Gavaris, S. (1991). Consistency of some northwest Atlantic groundfish stock assessments. *NAFO Sci. Coun. Stud.*, **16**: 59-77.
- Sitter, R. R. (1992) A resampling procedure for complex survey data. *J. Am. Statist. Assoc.*, **87**, 755-765.
- Smith, S. J. (1990) Use of statistical models for the estimation of abundance from groundfish survey data. *Can. J. Fish. Aquat. Sci.*, **47**, 894-903.
- Smith, S. J., Perry, R. I., and Fanning, L. P. (1991) Relationships between water mass characteristics and estimates of fish population abundance from trawl surveys. *Env. Monit. Assess.*, **53**, 227-245.
- Smith, S. J., and Page, F. H. (1996) Associations between Atlantic cod (*Gadus morhua*) and hydrographic variables: implications for the management of the 4VsW cod stock. *ICES J. Mar. Sci.*, **53**, 597-614.
- Stansbury, D., Bishop, C. A., Murphy, E. F., and Davis, M. B. (1995) *An assessment of the cod stock in NAFO Div. 3NO*. NAFO SCR Doc 95/70, Serial No. N2585.
- Stone, C. J. (1977) Consistent nonparametric regression. *Ann. Statist.*, **5**, 595-645.
- Sullivan, P. J. (1992) A Kalman filter approach to catch-at-length analysis. *Biometrics*

- 48, 237-257.
- Swain, D. P., and Kramer, D. L. (1995) Annual variation in temperature selection by Atlantic cod *Gadus morhua* in the southern Gulf of St. Lawrence, Canada, and its relation to population size. *Mar. Ecol. Prog. Ser.*, **116**, 11-23.
- Templeman, W. (1962) Division of cod stocks in the northwest Atlantic. *ICNAF Redbook Part III*, 79-123.
- Templeman, W. (1966) Marine resources of Newfoundland. *Bull. Fish. Res. Board. Can.*, **154**, 170 p.
- Templeman, W. (1974) Migrations and intermingling of Atlantic cod (*Gadus morhua*) stocks of the Newfoundland area. *J. Fish. Res. Board. Can.*, **31**, 1073-1092.
- Templeman, W. (1979) Migration and intermingling of stocks of Atlantic cod, *Gadus morhua*, of the Newfoundland and adjacent areas from tagging in 1962-66. *ICNAF Res. Bull.*, **14**, 5-50.
- Templeman, W. (1981) Vertebral numbers in Atlantic cod, *Gadus morhua*, of the Newfoundland and adjacent areas, 1947-71, and their use for delineating cod stocks. *J. Northw. Atl. Fish. Sci.*, **2**, 21-45.
- Thompson, M. L. (1978) Selection of variables in multiple regression: Part I. A review and evaluation. *Int. Statist. Rev.*, **46**, 1-19.
- Tibshirani, R, and Hastie, T. (1987) Local likelihood estimation. *J. Am. Statist. Assoc.*, **82**, 559-567.
- Walsh, S. J. (1996) Efficiency of bottom sampling trawls in deriving survey abundance indices. *NAFO Sci. Coun. Studies*, **28**, 9-24.
- Warren, W. G. (1997) Changes in the within-survey spatio-temporal structure of the northern cod (*Gadus morhua*) population, 1985-1992. *Can. J. Fish. Aquat. Sci.*, **54**, 139-148.

- Wolter, K. M. (1985) *An introduction to variance estimation*. New York: Springer-Verlag.
- Wroblewski, J. S., Kulka, D. W., Narayanan, S., Oake, A. M., Collier, A. G., and McGrath, B. D. (1995) Winter distribution and movement of northern Atlantic cod (*Gadus morhua*) along the Newfoundland-Labrador continental shelf edge derived from observations on commercial trawlers. *Fish. Oceanogr.*, 4, 128-146.
- Zhang, P. (1991) Variable selection in nonparametric regression with continuous covariates. *Ann. Statist.*, 19, 1869-1882.