Statistical inference about the relative efficiency of a new survey protocol, based on paired-tow survey calibration data

Noel G. Cadigan (contact author)[1]

Jeff J. Dowden[2]

Email address for contact author: noel.cadigan@dfo-mpo.gc.ca

[1] Fisheries and Oceans Canada
Northwest Atlantic Fisheries Center
80 East White Hills Road
St. John's, NL, Canada A1C 5X1

[2] Research and Evaluation Department
Newfoundland and Labrador Centre for Health Information
28 Pippy Place
St. John's, NL, Canada A1B 3X4

Abstract—Paired-tow calibration studies provide information on changes in survey catchability that may occur because of some necessary change in protocols (e.g., change in vessel or vessel gear) in a fish stock survey. This information is important to ensure the continuity of annual time-series of survey indices of stock size that provide the basis for fish stock assessments. There are several statistical models used to analyze the paired-catch data from calibration studies. Our main contributions are results from simulation experiments designed to measure the accuracy of statistical inferences derived from some of these models. Our results show that a model commonly used to analyze calibration data can provide unreliable statistical results when there is between-tow spatial variation in the stock densities at each paired-tow site. However, a generalized linear mixed-effects model gave very reliable results over a wide range of spatial variations in densities and we recommend it for the analysis of paired-tow survey calibration data. This conclusion also applies if there is between-tow variation in catchability.

Manuscript submitted 20 April 2009.
Manuscript accepted 14 September 2009.
Fish. Bull. 108:15–29 (2010).

Surveys are an important source of information for most fish stock assessments. They provide indices of abundance often used in mathematical models of the population to estimate absolute stock size and to provide future catch advice or to evaluate catch options for fishery managers (Kimura and Somerton, 2006). Survey indices are measures that we expect to be proportional to, or to indicate, stock size. The expected value of a random index R_y available for year y is related to stock size (S_y) by the model $E(R_y)=qS_y$. The constant of proportionality, q, is usually referred to as the catchability of the index. Although we cannot directly infer stock size from a time series of indices $R_1, \ldots, R_Y$, we can infer trends in stock size when q is the same each year. The survey observation is commonly referred to as a *set* (as in *set* the gear), or a *tow* when a trawl is used. The average survey catch for all sets provides an index of stock size. If the same survey protocols are used from year to year then the catchability of the index should remain relatively constant. The catchability may depend on length or age of fish, and we consider such extensions later in this article.

There are many stock assessment models that are based on survey indices (e.g., see Quinn and Deriso, 1999) and information on fishing and natural mortality, to estimate absolute stock size. For most models it is necessary to have a fairly long time-series of survey indices, often 10 years or more. Over such time frames it may be necessary to change survey protocols. This could be due to a need to replace the survey vessel, or to change gears for new priority species, etc. When such changes occur, it is useful to have information about how the catchability of the new survey protocol compares to the old protocol.

Performing simultaneous paired-tow surveys using both protocols (e.g., old and new vessels, old and new fishing gears) provides direct data on how the catchabilities compare (e.g., Kimura and Zenger, 1997) Another approach is to simply fish side by side using both protocols and use the paired-catch data to estimate the ratio of catchabilities. We refer to this ratio as the relative efficiency,

$$\rho = \frac{q_c}{q_t}, \qquad (1)$$

where q_c and q_t = the catchabilities of the old (control, c) and new (test, t) survey protocols.

Notations are given in Table 1. If the fish densities entering both trawls are the same, or similar, and densities at different tow sites vary considerably, then for the same number of tows a paired-tow calibration study should produce better results than the

Table 1

Definitions of variables and acronyms for models used to estimate relative efficiency from comparative fishing data.

R_{ij}	Random variable for catches obtained at the i'th paired-tow station by survey protocol $j=c$ (control) or $j=t$ (test)
r_{ij}	Observation of R_{ij}
R_i	$R_{ic}+R_{it}$
R_{ijk}	Catches at the i'th tow station and k'th length class by survey protocol j
R_{ik}	Total catch (from both vessels) at length class k from set i, $R_{ick}+R_{itk}$
n	Total number of paired-tow stations
n_i	Number of length classes caught in the i'th pair of tows
n^*	Total number of sets and length classes, $n^* = \sum_i n_i$
λ_{ij}	Fish densities encountered at station i and tow j
δ_i	$\log(\lambda_{ic}/\lambda_{it})$
q_j	Probability an encountered fish is captured, $j=c, t$
ρ	Relative efficiency, $\rho=q_c/q_t$
β	$\log(\rho)$
p	Probability a captured fish was caught by the control protocol
D_{ij}	Tow duration at the i'th paired-tow station by vessel j
F_{ijl}	Subsampling fraction for length l fish
Z_{il}	Logit offset, $Z_{il}=\log(D_{ic}F_{icl}/D_{it}F_{itl})$
ϕ	Binomial over-dispersion
σ^2	Random effect variance
CIs	Confidence intervals
GLIM	Generalized linear model
MLE	Maximum likelihood estimation
GLMM	Generalized linear mixed model
PQLE	Penalized quasi-likelihood estimation
CV	Coefficient of variation
VO	Vessel-effect over-dispersed binomial model estimation
VM	Vessel-effect binomial model with random intercept for each set; marginal MLE
VP	Vessel-effect binomial model with random intercept for each set; PQLE
VLO	Vessel- and length-effects over-dispersed binomial model estimation
VLM_i	Vessel- and length-effects binomial model with random intercept for each set; marginal MLE
VLP_i	Vessel- and length-effects binomial model with random intercept for each set; PQLE
VLM_{is}	Vessel- and length-effects binomial model with random intercept and slope for each set; marginal MLE
VLP_{is}	Vessel- and length-effects binomial model with random intercept and slope for each set; PQLE

simultaneous survey approach. This is analogous to the common paired versus unpaired experiment situation (e.g., Devore, 1991). Pelletier (1998) reviewed estimation methods used in many vessel calibration experiments.

The basic data obtained from paired-tow calibration studies are the catches R_{ij} obtained at the ith paired-tow station (i=1, … ,n) by survey protocols $j=c$ (control) or $j=t$ (test). Let λ_{ij} denote the fish densities encountered at station i and tow j. These densities may be different because of small-scale spatial heterogeneity in stock densities. We assume that each tow catches fish with probabilities q_c and q_t which are the same from site to site (i.e., i), and that catches are Poisson random variables with means

$$E(R_{it}) = q_t\lambda_{it} = \mu_i, \text{and } E(R_{ic}) = q_c\lambda_{ic} = \rho\mu_i \exp(\delta_i), \qquad (2)$$

where $\delta_i = \log(\lambda_{ic}/\lambda_{it})$.

If both vessels encounter exactly the same stock densities at each tow station, then δ_i=0, i=1, … , n.

When there is no spatial heterogeneity in stock densities, ρ can be estimated by using a Poisson generalized linear model (GLIM; e.g., McCullagh and Nelder, 1989). This is essentially the approach used by Benoît and Swain (2003), although they adjusted for extra-Poisson variability in the catches. There are $2n$ observations that can be used to estimate the n density parameters (μ) and ρ. Pelletier (1998) used a similar approach, with a negative binomial mean-variance assumption, which is a type of Poisson over-dispersion. These approaches are complicated because the number of μ parameters can be large if many tow stations are sampled, and the situation is worse if there are length effects.

A better approach for inferences about ρ (see section 4.5 in Cox and Snell, 1989, and example 3.1 in Reid, 1995) when catches are Poisson random variables is to use the conditional distribution of $R_{i,c}$, given $R_i = R_{ic}+R_{it}$. Let r_i be the observed value of R_i. The conditional distribution is binomial with a probability mass function

$$prob(R_{ic} = x \mid R_i = r_i) = \binom{r_i}{x} p^x (1-p)^{r_i - x}, \quad (3)$$

where $p = \rho/(1+\rho)$ is the probability a captured fish is taken by the control vessel.

The only unknown parameter in this distribution is ρ. The n nuisance μ parameters are eliminated in Equation 3. There are n conditional observations that can be used to estimate ρ. For the binomial distribution $E(R_{ic})=r_i p$ and $\mathrm{Var}(R_{ic})=r_i p(1-p)$. This approach is commonly used in commercial fishing gear size-selectivity studies (e.g., Millar 1992).

Paired-tow experiments do not eliminate spatial heterogeneity between the stock densities fished by each vessel. This heterogeneity leads to Poisson over-dispersion which has to be properly accounted for to provide reliable statistical inferences. Similarly, the relative efficiency may vary somewhat from site to site and this must also be accounted for. It is well-known in fishing gear selectivity studies that not accounting for over-dispersion and correlation leads to confidence intervals that are too narrow and spurious statistical significance (Fryer, 1991; Millar et al., 2004).

An approach to deal with over-dispersion is to use quasi-likelihood (e.g., McCullagh and Nelder, 1989) with a Poisson over-dispersion parameter ϕ, $\mathrm{Var}(R_{ij})= \phi\, E(R_{ij})$, or a binomial over-dispersion parameter, $\mathrm{Var}(R_{ic} \mid R_i=r_i)= \phi\, r_i p(1-p)$. Confidence intervals (CIs) are adjusted based on an estimate of ϕ. This was the approach used by Benoît and Swain (2003) to account for extra-Poisson variation, and Lewy et. al. (2004) to account for extra-binomial variation. Benoît[1] observed that using an over-dispersion parameter did not completely account for the true variability in the data and too often led to the false statistical conclusion that $\rho \neq 1$. Benoît used randomization approaches to test for statistical significance of vessel effects. We consider this approach further in the *Discussion* section.

A reasonable assumption for spatial heterogeneity in stock densities is that λ_{ic} and λ_{it} in Equation 2 are independent and identically distributed gamma random variables with means λ_i and variances $\tau\lambda_i^2$. If $R_{ic} \mid \lambda_{ic}$ and $R_{it} \mid \lambda_{it}$ are Poisson distributed, then the marginal distributions of R_{ic} and R_{it} are negative binomials (e.g., see Cameron and Trivedi, 1998). This distribution is often suggested to be appropriate for modeling the variability of measurement error in trawl survey catches (e.g., Gunderson, 1993). This implicitly provides a rationale for assuming stock densities are gamma distributed. Dowden[2] showed that τ=0.049, 0.223, 0.372 corresponds to $\mathrm{Var}(\delta_i)$=0.1, 0.5, and 0.9, and that the distribution of δ (see Eq. 2) is well approximated by a normal distribution with $\sigma^2=\mathrm{Var}(\delta_i)$. In this case a generalized linear mixed-effects model (GLMM; e.g., McCulloch and Searle, 2001) can be used to estimate ρ and account for spatial heterogeneity in stock densities. GLMMs contain both fixed and random effects, and usually the random effects are assumed to have normal distributions. We refer to models with no random effects as fixed effects models (e.g., GLIMs).

GLMMs are frequently used to account for heterogeneity in fishing gear size-selectivity data (e.g., Fryer, 1991; Fryer et al., 2003; Millar et al., 2004). Fryer et al. (2003), Cadigan et al.[3] and Holst and Revill (2008) used GLMMs with paired-tow calibration data. Cadigan et al.[3] compared models with and without random effects for vessel calibration data for seven species, and suggested that GLMMs provided results that were more reliable. Cadigan et al.[3] concluded that vessel effects were not significantly different from zero; however, different conclusions could be drawn from some of their GLIM results.

There are a variety of approaches available for fitting GLMMs. A common approach is marginal maximum-likelihood estimation (MLE), which is limited in the complexity of random effects that can be accommodated. A more flexible approach is penalized quasi-likelihood estimation (PQLE). Bolker et al. (2009) provided some advantages and disadvantages of these methods. They also provided many references, including some for software packages. In some situations, PQLE is known to produce biased estimates of fixed-effect parameters like ρ.

In this article we extend the analyses for one of the stocks considered by Cadigan et al.[3]. By means of simulation studies we examine which of the approaches—the GLIM, GLMM with marginal MLE, or GLMM with PQLE—provides more reliable statistical inferences about ρ. We focus on the bias in estimates of ρ, on the accuracy of CIs, and on the power to detect if $\rho \neq 1$ (i.e., a true difference in catchabilities between vessels). Our purpose is to recommend the most reliable approach, at least for paired-tow survey calibration studies similar to those in Cadigan et al.[3]. We focus on methods to accommodate within-pair variations in stock densities, but our methods are also applicable when there is between-set variations in relative efficiency.

[1] Benoît, H. P. 2006. Standardizing the southern Gulf of St. Lawrence bottom trawl survey time series: Results of the 2004–2005 comparative fishing experiments and other recommendations for the analysis of the survey data. DFO Can. Sci. Advisory Secretariat Res. Doc. 2006/008. [Available from http://www.dfo-mpo.gc.ca/csas/csas/publications/resdocs-docrech/2006/2006_008_e.htm, accessed April 2009.]

[2] Dowden, J. J. Generalized linear mixed effects models with application to fishery data. M.A.S. practicum report, 128 p. Memorial Univ. Newfoundland. St. John's, Newfoundland and Labrador, Canada.

[3] Cadigan, N. G., S. J. Walsh, and W. Brodie. Relative efficiency of the Wilfred Templeman and Alfred Needler research vessels using a Campelen 1800 shrimp trawl in NAFO Subdivision 3Ps and Divisions 3LN. DFO Can. Sci. Advisory Secretariat Res. Doc. 2006/085. [Available from http://www.dfo-mpo.gc.ca/csas/Csas/Publications/ResDocs-DocRech/2006/2006_085_e.htm, accessed April 2009.]

Materials and methods

We focus on statistical inferences for ρ (i.e., Eq. 1) based on data obtained from paired-tow vessel calibration studies like those described in Cadigan et al.[3]. Briefly, in their study, data from paired-tows were collected to quantify potential differences in the catchabilities of two research survey vessels fishing with the same trawl and other protocols. Ranges of catch sizes, fish sizes in the catch, and tow depths were sought for the distributions of the species likely to be encountered. Tow stations were selected randomly as part of research surveys. High density aggregations were not specifically targeted because information was required on differences in catchability when stock densities were both high and low—a variability in densities that typically occurs in research surveys. The full details of this calibration study are given in Cadigan et al.[3]. We use their data on witch flounder (*Glyptocephalus cynoglossus*) as a case study to illustrate methods.

The focus in Cadigan et al.[3] was on the relative efficiency of two vessels fishing with otherwise identical protocols (gears, speed, tow duration, etc.). Hence, in this article we refer to vessel effects, but more generally the effects relate to differences in fishing protocols.

The first step in analyzing calibration data is to examine whether there is an effect on total catch per set. In the next section we describe a model for this purpose. Effects on the length compositions of the catches are considered later in this article.

Vessel effect

A common approach used for analyzing comparative fishing data is binomial regression with an adjustment for over-dispersion. This is one of the options we considered. In the conditional binomial model defined by Equation 3, the logit function of the binomial probability (p) is

$$\log\left(\frac{p}{1-p}\right)=\log(\rho)=\beta, \tag{4}$$

and β can be estimated as the intercept with a logit link function by using software for binomial regression. The range of β is $(-\infty, \infty)$. We derived CIs for ρ by exponentiating intervals for β and therefore CIs for ρ should have better coverage properties, and they at least would never include infeasible values. We used version 9.1.3 of SAS/STAT (SAS, Cary, NC.) PROC GENMOD software to estimate this model, and we used the option (dscale) that estimates ϕ as the deviance divided by the degrees of freedom. We also selected the option (lrci) that provides two-sided CIs based on the profile likelihood function. We refer to this GLIM model and estimation approach as the VO (vessel-effect binomial model with over-dispersion) approach.

If there is spatial heterogeneity in stock densities, then the model for the logit proportion of catch taken by the control vessel at station i is

$$\log\left(\frac{p_i}{1-p_i}\right)=\beta+\delta_i, \tag{5}$$

Usually it is reasonable to assume that the ratio of stock densities varies randomly from site to site. Earlier we claimed it was reasonable to assume $\delta_i \sim N(0, \sigma^2)$, $i=1, \ldots n$. In this case equation 5 defines a standard GLMM and there are many approaches and software packages available to estimate β and σ^2 (e.g., Bolker et al., 2009). We examined the robustness of statistical inferences about ρ to the normal approximation for δ (see *Simulations* section) when δ is actually a log ratio of gamma random variables.

We used two different packages to estimate the GLMMs. The first was SAS/STAT PROC NLMIXED, which fits nonlinear mixed models, including binomial logistic regression, using marginal MLE. We refer to this model and estimation procedure as the VM (vessel effect and random set-effects binomial model with marginal MLE) approach. The second was the more flexible SAS/STAT PROC GLIMMIX, which fits GLMMs using PQLE. We refer to this as the VP (vessel effect and random set-effects binomial model with PQL estimation) approach. We used the default estimation method in PROC GLIMMIX, which is a restricted pseudolikelihood estimation with an expansion around the current estimate of the best linear unbiased predictors of the random effects. Both packages provide Wald-type CIs for fixed-effect parameters such as β.

Vessel and fish-length effects

Length effects are expected if there is a change in the survey trawl, but they could also occur with only a change in the survey vessel. Length-based models for paired-tow comparative fishing data are straightforward extensions of the models in the previous section. The data are extended to include the paired catches at length, R_{ijk}, $i=1, \ldots, n$; $j=1,2$; $k=1, \ldots n_i$, where n_i is the number of length classes caught in the i'th pair of tows.

If it is reasonable to assume that there are no between-pair differences in the length distributions of fish encountered by both vessels, then binomial logistic regression models are appropriate. Usually the effect of length will be such that relative efficiency changes monotonically with length, l. If the change is linear in $\beta=\log(\rho)$, then a binomial GLIM with a logistic link can be used to estimate the intercept and slope; that is, $\beta \equiv \beta(l) = \beta_0+\beta_1 l$ in Equation 4, where β is taken to be a function of length, $\beta(l)$. If the length effect is more complicated, then alternative models may be required (see Fryer et al., 2003; Holst and Revill, 2008); however, in this article we focus only on linear models.

If there is spatial heterogeneity in stock densities, then the situation is more complicated. If the heterogeneity is such that one vessel encounters more fish than the other, but otherwise the length distributions are the same, then the use of a random intercept binomial GLMM is appropriate:

$$\log\left(\frac{p_{ik}}{1-p_{ik}}\right)=\beta_0+\delta_i+\beta_1 l_{ik}, \tag{6}$$

$$\delta_i \overset{iid}{\sim} N(0,\sigma^2),\ i=1,...,n;\ k=1,...,n_i,$$

where p_{ik} = the probability that at site i a length l_{ik} captured fish came from the control vessel.

Holst and Revill (2008) used a random intercepts model, although their models for fixed lengths effects were more complicated than what we consider. However, if there are differences in the length distributions encountered by both vessels then Equation 6 will not be appropriate. The differences will usually be such that $\delta=\log(\lambda_c/\lambda_t)$ varies smoothly with length. Several hypothetical examples are also shown in Figure 1. This type of spatial heterogeneity can be approximated by linear functions, whose slopes (δ_1) and intercepts (δ_0) vary randomly from set to set. A GLMM for this model is

$$\log\left(\frac{p_{ik}}{1-p_{ik}}\right)=\beta_0+\delta_{i0}+(\beta_1+\delta_{i1})l_{ik},$$
$$\delta_{ij} \overset{iid}{\sim} N(0,\sigma_j{}^2),\ i=1,...,n; \tag{7}$$
$$j=0,1;\ k=1,...,n_i.$$

This is a common GLMM used in fishing gear selectivity studies. If the means of the densities are the same and the only difference is the height of the distributions, then the δ log ratio would be a horizontal line in Figure 1, which is the type of effect accounted for in Equation 6.

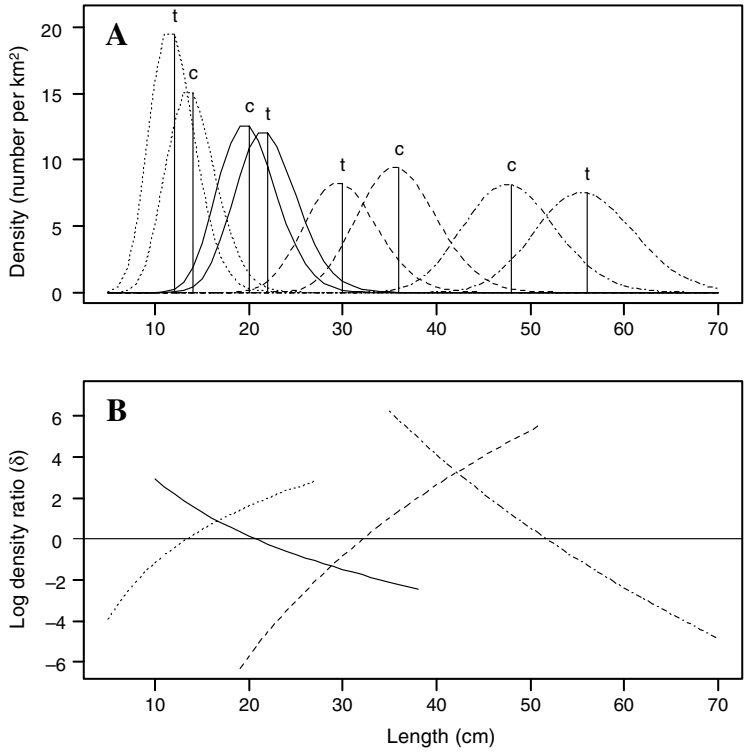

Figure 1

(**A**) Hypothetical length distributions of fish encountered by the control (c) and test (t) fishing protocols from paired-tow fishing experiments, to illustrate the impact of within-pair variations in stock densities on the relative efficiency of the test compared to the control protocol. Line types correspond to pairs of tows. Thin vertical lines indicate mean length per tow and coefficients of variation equal to 0.5. (**B**) Corresponding log density ratios (i.e., δ's) are shown for ranges of lengths where the average density (in A) was greater than 0.01.

We used the same SAS software to estimate the length-based GLIMs and GLMMs. We denote the length-based model with no random effects as VLO. Mixed-effects random intercept models are denoted as VL_i, (i.e., Eq. 6), and models with random intercepts and slopes are denoted as VL_{is}. Models and estimation methods are denoted as VLM_i, VLP_i, VLM_{is}, and VLP_{is} (see Table 1) depending on whether marginal MLE or PQLE is used.

Standard errors for $\beta(l)=\beta_0+\beta_1 l=\log\{\rho(l)\}$ can be constructed from the estimates of β_0 and β_1, and their estimated covariances:

$$SE\left\{\hat{\beta}(l)\right\}=\left\{\mathrm{var}(\hat{\beta}_0)+2\mathrm{cov}(\hat{\beta}_0,\hat{\beta}_1)l+\mathrm{var}(\hat{\beta}_1)l^2\right\}^{1/2}. \tag{8}$$

These standard errors can be used to produce approximate 95% pointwise CIs for $\rho(l)$:

$$CI=\exp\left[\hat{\beta}(l)\pm 1.96\times SE\left\{\hat{\beta}(l)\right\}\right]. \tag{9}$$

Occasionally in comparative fishing the duration (D_{ij}) of the tows may differ somewhat between vessels. Also, because of operational time constraints the catches may have to be subsampled for some species. The subsampling fraction (F_{ijl}) may depend on size as well. To account for these effects we added an offset (Z) to Equations 6 and 7, $Z_{il}=\log(D_{ic}F_{icl}/D_{it}F_{itl})$. For length-pooled analyses of vessel effects we added the offset $Z_i=\log(D_{ic}F_{ic}/D_{it}F_{it})$ to Equation 5, where F_{ij} is the total subsampling fraction.

Simulations

Vessel effects The design of the simulation experiments mimicked the design for the data analyzed by Cadigan et al.[3], that is, we simulated data for the seven species that they considered, and for the same number of tows and total catches for both paired-tows (R). This design is summarized in Table 2. Therefore, for example, in

Table 2

Paired-tow fishing simulation parameters based on the species in Cadigan et al.[3] n is the total number of paired-sets, n^* is the total number of sets and length classes, $R=R_t+R_c$ is the total catch by the test (t) and control (c) vessels, and R_t/R_c is the ratio of catches from each vessel. The mean, median and coefficient of variation (CV) are for total catch (R). R25 and R75 are the lower and upper 25th percentiles of R. See Table 1 for definitions of notations.

Species	n	n^*	R	R_t/R_c	mean	median	R25	R75	CV
American plaice (*Hippoglossoides platessoides)*	105	2035	11,494	1.051	109.5	40.0	9	169	126
Atlantic cod (*Gadus morhua)*	91	1132	3926	1.067	43.1	4.0	2	31	217
Deepwater redfish (*Sebastes mentella*)	63	1030	12,069	1.207	191.6	88.0	14	379	104
Greenland halibut (*Reinhardtius hippoglossoides*)	56	585	1359	1.243	24.3	13.5	4	36	120
Thorny skate (*Raja radiata*)	79	990	2394	1.124	30.3	9.0	3	20	226
Witch flounder (*Glyptocephalus cynoglossus*)	57	970	5046	1.334	88.5	52.0	6	151	108
Yellowtail flounder (*Limanda ferruginea*)	24	536	5795	1.250	241.5	159.0	5	503	102

the simulation based on American plaice (*Hippoglossoides platessoides)*, data were generated for 105 paired-tows with an average R of 109.5. Twenty-five percent of the sets had $R \leq 9$, and 25% of sets had $R \geq 169$. The number of paired-tows in the seven sets of simulations varied from 25 to 105 which is a practical range consistent with many comparative fishing studies (e.g., Table 2 in Pelletier, 1998). The catches ranged from low (Atlantic cod, *Gadus morhua*) to high (yellowtail flounder, *Limanda ferruginea*). Some of the stocks had very skewed catches; for example, Atlantic cod and thorny skate (*Raja radiata*) had mean catches that exceeded their 75th percentiles.

The simulations were similar to a parametric bootstrap procedure; however we varied β and σ^2 to examine how the accuracy of statistical inferences varied with changes in these parameters. The values of β ranged from 0 to 2 by increments of 0.25, with $\beta=0$, or $\rho=1$, representing no vessel effect, and $\beta=2$, or $\rho=7.4$, representing a test vessel catchability that was 14% of the control vessel. Note that this range in β is much larger than the results in Cadigan et al. (Table 7 in Cadigan et al.[3]). Their largest absolute estimate was 0.08; however, these simulations were designed to examine the accuracy of statistical inferences for small and large vessel effects. The levels of $\sigma^2=0$, 0.1, 0.5, and 0.9 represented no to high spatial heterogeneity and broadly reflected the range of estimates in Cadigan et al. (Table 7 in Cadigan et al.[3]). The lowest estimate of σ^2 in Cadigan et al.[3] was 0.10 for Greenland halibut (*Reinhardtius hippoglossoides*), and the highest estimate was 0.99 for Atlantic cod.

Simulated catches for the control vessel (R_c) were generated as binomial random numbers; the number of trials was equal to the observed R in Cadigan et al.[3] and probability was based on Equation 5. The δs were generated randomly from a normal distribution with mean zero and variance σ^2. The simulated test vessel catch was $R_t=R-R_c$. Note that the total catches for each paired-tow, $R, \ldots, R_n$, were the same in each simulation; hence, our simulation results were conditioned on these values.

We examined the robustness of the GLMM results to the assumption of a normal distribution for the random effects, δ, when in fact δ was the log of a ratio of two independent and identically distributed gamma random variables with $\mathrm{Var}(\delta)=\sigma^2$.

Estimates and 95% CIs for β were obtained from 2000 simulations. We approximated the estimation bias as the median β from the simulations minus the true simulation value. Bias results based on means were very similar. The coverage accuracy of the CIs was measured as the proportion of simulations in which the CI contained the true value of β. If the 95% CIs are accurate, then the simulation proportion should be close to 0.95. We also computed the proportion of simulations in which β was less than the lower CI, and the proportion in which β was greater than the upper CI. If the CIs are two-sided accurate, then these proportions should both equal 0.025.

We performed other simulations using a much finer scale for β to examine the power of detecting a vessel effect based on the proportion of simulations whose CIs did not cover zero when the true β was greater than zero.

Vessel and fish-length effects These simulations were similar to those described in the last section. We simulated data for the seven species and with the same number of tows and total catches-at-length for both

paired-tows (R_l). However, there was an additional simulation factor for the slope of the length effect. We standardized lengths in the data from Cadigan et al.[3]:

$$l_{std} = \frac{l - l_{50}}{l_{75} - l_{25}}, \tag{10}$$

where l_α = the αth percentile of the lengths from all sets, weighted by total catch, for each species.

This standardization allowed us to use the same slopes in simulations for different species; we considered β_1=0, 0.5, 1.0 to represent no, medium, and large length effects. This scale increased the number of simulations three-fold. The length-based models were also slower to estimate because of the larger size of the data sets (see n^* in Table 2), and to save time we reduced the number of simulations to 1000.

We simulated data from Equation 7. We set $\sigma_1^2=\sigma_2^2=\sigma^2$ and used the same values as before, σ^2=0, 0.1, 0.5, and 0.9. However, we fitted both Equation 6 and Equation 7 to the simulated data. This procedure allowed us to examine the accuracy of statistical inferences from the random intercept model, which is a common mixed effects model, when slopes were random as well. We summarized the simulations for $\beta(l_{std})=\beta_0+\beta_1 l_{std}$ at three points, l_{std}=0, 0.5, and 1.0, which reflects relative efficiency at median to large lengths.

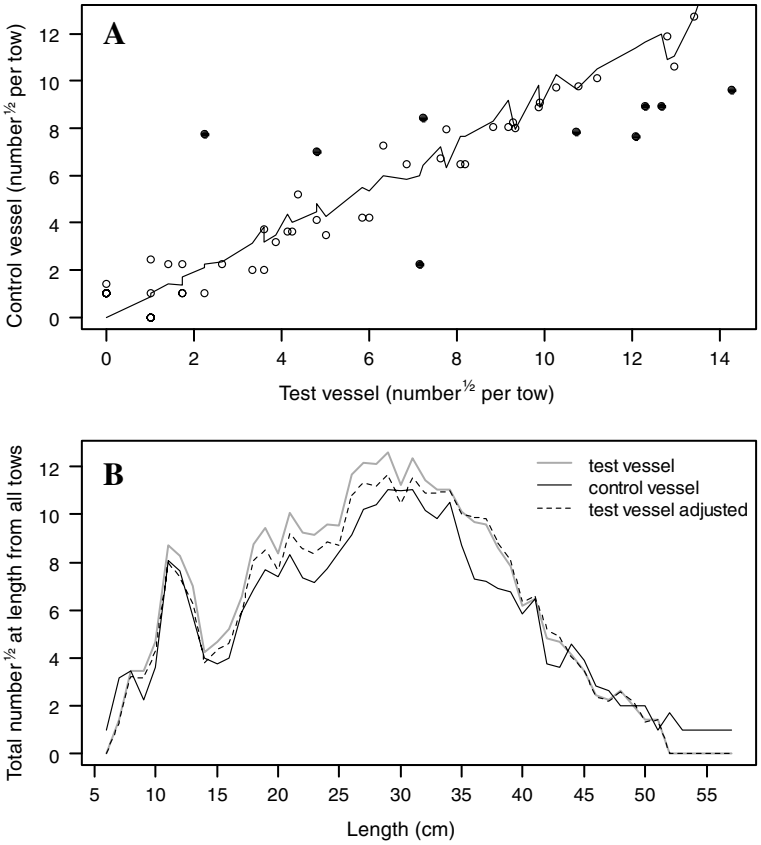

Figure 2

Comparative fishing survey results for witch flounder (*Glyptocephalus cynoglossus*). (**A**) Catches from the test and control vessels with 57 paired-tows. Thick line segments connect the offset adjustments for differences in tow duration and subsampling of catches (see text for further detail) and indicate equal catchability (i.e., ρ=1); that is, if ρ=1, the scatter of points should be centered around the offset adjustments. Solid circles indicate paired-sets with substantially different catches. (**B**) Total catch per 1-cm length classes over all sets, for the test vessel, the test vessel adjusted for tow duration and subsampling, and the control vessel.

Results

Case study

We illustrate methods using data for witch flounder from Cadigan et al.[3]. There is some evidence that the catchability of the control vessel was less than the test vessel. For example, there were five sets (solid circles in Fig. 2 where the test vessel caught more than 100 fish but the control vessel caught fewer than 100 fish. Rarely were catches by the control vessel much larger than the test vessel. However, in most paired-tows the catches by both vessels, when adjusted for tow distance and subsampling, were similar. The length distributions over all sets (Fig. 2, bottom panel) did not indicate that potential differences in catchabilities were length dependent because catch differences were approximately equally distributed over a broad range of sizes.

The GLIM estimate of $\beta=\log(\rho)$ (Table 3; VO model) was significantly less than zero indicating that the control vessel had a catchability that was significantly less than the test vessel. However, both GLMM estimates (VM, VP) were somewhat larger and not significant, indicating that the test and control vessels catchabilities were not significantly different. The VM and VP estimates of β and σ^2 were very similar. The PQL software (PROC GLIMMIX) we used did not provide standard errors for the estimate of σ^2, but the marginal MLE software (PROC NLMIXED) did.

The length-based models provided similar results, with the GLIM model (VLO) producing significant differences whereas the mixed models did not. Note that the length-based estimates are very different from those in Cadigan et al.[3] because we standardized lengths (i.e.,

Table 3

Parameter estimates, standard errors (SE), and lower and upper 95% confidence interval limits from various models for paired-tow comparative fishing data. See Table 1 for definitions of model acronyms and parameters. The v and l parameter subscripts indicate a vessel or fish-length effect.

Model	Parameter	Estimate	SE	Lower	Upper
VO	β_v	−0.153	0.069	−0.290	−0.018
	ϕ	5.888	—	—	—
VP	β_v	−0.094	0.093	−0.279	0.091
	σ_v^2	0.301	0.088	—	—
VM	β_v	−0.097	0.093	−0.283	0.089
	σ_v^2	0.302	0.089	0.124	0.480
VLO	β_v	−0.166	0.037	−0.239	−0.094
	β_l	−0.053	0.052	−0.154	0.048
	ϕ	1.628	—	—	—
VLP_i	β_v	−0.102	0.092	−0.286	0.082
	β_l	0.058	0.052	−0.045	0.160
	σ_v^2	0.296	0.086	—	—
VLM_i	β_v	−0.105	0.092	−0.290	0.079
	β_l	0.058	0.052	−0.047	0.164
	σ_v^2	0.295	0.086	0.121	0.468
VLP_{is}	β_v	−0.099	0.094	−0.288	0.090
	β_l	0.059	0.080	−0.098	0.215
	σ_ϖ^2	0.299	0.088	—	—
	σ_l^2	0.119	0.061	—	—
VLM_{is}	β_v	−0.103	0.095	−0.294	0.088
	β_l	0.061	0.081	−0.101	0.224
	σ_ϖ^2	0.303	0.091	0.121	0.484
	σ_l^2	0.122	0.065	−0.008	0.252

Eq. 10) but Cadigan et al.[3] did not. The 25th, 50th, and 75th length percentiles were 22, 29, and 34 cm. CIs for $\rho(l)$ based on Equation 9 were derived for l_{std}=0, … , 2 (Fig. 3). The VLO model suggested $\rho(l)$ decreased with l and was significantly different from one over the range of lengths. The four mixed models all indicated a slight increase in $\rho(l)$ with l but were not significantly different from one for any length. The CIs for the random intercept model (Eq. 6) were shorter than those for the random intercept and slope models, especially when l_{std}>1. The marginal MLE CIs (VLM_i, VLM_{is}) were slightly wider than PQLE CIs (VLP_i, VLP_{is}).

Simulations

Vessel effect Simulation results were very similar for the seven species. We present the best and worst cases in Figures 4 and 5. The VO model performed poorly even when there was small spatial heterogeneity in stock densities (i.e., σ^2=0.1). The likelihood ratio CIs had poor coverage properties and the probability that they contained the true value of the vessel effect (β) was much less than the 95% nominal value. When σ^2 and β were large this method produced biased estimates of β and very inaccurate CIs. Note that to facilitate comparison of the methods the y-axis was fixed to be less than the range of some of the GLIM results, particularly in Figure 5. The VP CIs were more accurate, except when $\sigma^2 \geq 0.5$ for the thorny skate simulation (Fig. 5). For larger values of β the CIs from this method covered less than 95%, about 80% for σ^2=0.9 and β =2. The bias was negative which meant that the lower and upper bounds were too small. The VM CIs were quite reliable across the range of values for σ^2 and β, and for all seven simulation scenarios (i.e., species). The log gamma ratio simulation results were almost identical to the normal simulations results and are not presented.

We performed simulations at a finer scale of β to determine the size of a vessel effect that could be detected with a power of 0.8 or 0.95, based on the VM model. The power was computed from the proportion of CIs that did not cover zero. The results are shown in Table 4, expressed in terms of percent change, $100\times(\rho-1)$. For example, when σ^2=0.5 there was a 95% chance of detecting a 44% increase in catchability with data like that for American plaice.

Vessel and length effects The VLO model performed poorly compared to the mixed models and those results are not presented. We examined statistical inferences for $\beta(l_{std})$ based on the VLM_i, VLP_i, VLM_{is}, and VLP_{is} models. Note that simulated data were generated by using Equation 7 but fitted with both Equation 6 and Equation 7; hence, the results based on Equation 6, i.e. VLM_i and VLP_i, will reflect model mis-specification biases. Results were similar for each simulation scenario (i.e., species). We present results for $\beta(l_{std})$ only for the Atlantic cod scenario, small and medium spatial variability (σ =0.1, 0.5), no or large lengths effects (β_1=0, 1), and at the center of the length distribution (l_{std}=0; Fig. 6 or at a larger value (l_{std}=1; Fig. 7).

The random intercept models gave unreliable results especially when l_{std}=1. The total random effect variance based on Equation 7 increased with length and this was not accounted for by the VLM_i or VLP_i models. The poor performance of CIs for $\beta(l_{std})$ derived from the VLM_i and VLP_i models was caused by both bias in estimates of $\beta(l_{std})$ and bias in estimates of the variance of the estimator for $\beta(l_{std})$. These biases are a complicated function of β_0, β_1, σ^2 and l_{std}.

The PQL estimation bias was similar to the results of the vessel effects only simulation (not shown for Atlantic cod) when l_{std}=0, which is not surprising because the conditional distributions based on Equation 5 and Equation 7 are essentially the same in this case. However, when l_{std}=1 the VLP_{is} model gave less accurate results compared to those when l_{std}=0. The VLM_{is} model gave reliable results in all simulation settings.

The worst case VLM_{is} result for the lower 95% CI coverage was for the deepwater redfish (*Sebastes mentella*) scenario in which the simulation coverage was 0.057 (nominal value is 0.025) when β_0=0.5, β_1=1, σ^2=0.5, and l_{std}=0. The worst case result for the upper interval

Table 4

The size of a vessel effect (i.e., change in relative efficiency, ρ–1, in %) that can be detected with power=0.8 or 0.95. Columns are for values of σ^2 (i.e., the random effect variance) and rows are for species simulation scenarios.

	Power=0.8 /σ^2				Power=0.95 /σ^2			
Species	0.0	0.1	0.5	0.9	0.0	0.1	0.5	0.9
American plaice (*Hippoglossoides platessoides)*	5	13	23	32	9	24	44	64
Atlantic cod (*Gadus morhua)*	10	18	33	42	17	33	68	92
deepwater redfish (*Sebastes mentella*)	5	15	30	43	7	27	60	99
Greenland halibut (*Reinhardtius hippoglossoides*)	16	23	39	54	28	44	86	134
thorny skate (*Raja radiata*)	13	20	34	45	23	38	71	101
witch flounder (*Glyptocephalus cynoglossus*)	8	17	34	49	13	31	72	113
yellowtail flounder (*Limanda ferruginea*)	8	25	62	101	13	50	164	372

was 0.055 when β_0=2, β_1=0, σ^2=0.9, and l_{std}=0 for Atlantic cod. In 95% of the simulations for all species the absolute error for the lower CI was less than 0.013, and for the upper interval it was less than 0.016. This demonstrates that CIs from the VLM_{is} model were almost always very accurate.

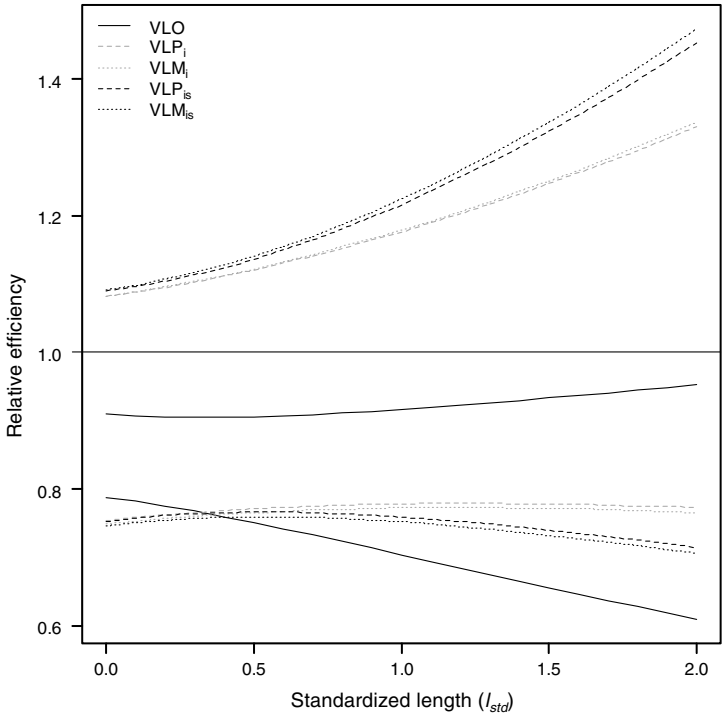

Figure 3

95% confidence intervals for the relative efficiency of the test vessel compared to the control vessel for catches (in numbers) of witch flounder (*Glyptocephalus cynoglossus*). Relative efficiency was modeled as a function of length l, $\rho(l)=\exp(\beta_0+\beta_1 l)$, and length was standardized, $l_{std}=(l-l_{50})/(l_{75}-l_{25})$, where l_α was the $\alpha\times100$% percentile of the lengths caught in all sets. The five models indicated by different line patterns and shading are described in Table 1. Two lines of the same pattern and shading are plotted for the lower and upper confidence interval endpoints. The thin horizontal line represents equal catchability, $\rho(l)$=1.

Discussion

Our simulation results demonstrated that the commonly used over-dispersed binomial logistic regression model did not provide accurate statistical inferences for paired-trawl calibration data when there was spatial variation in stock densities. In practice, such variations will occur and therefore this approach is not recommended. Fortunately, our simulations showed that a binomial logistic regression model that included random site effects in addition to fixed vessel effects did provide accurate inferences for a wide range of spatial variations in stock densities. This conclusion also applied to pooled or length-based analyses. We recommend this binomial generalized linear mixed-effects model (GLMM) for analyzing comparative fishing data. When assessing for length effects, we recommend using a binomial GLMM with between-site random variation in both the vessel and length effects.

Between-set variability in catchability is commonly observed in covered-codend experiments (e.g., Fryer 1991, Millar et. al. 2004) that directly measure catchability. This will also produce between-site variability in ρ. Trenkel and Skaug (2005) assumed that the Poisson density for fish abundance was spatially constant on a small scale (~1000 km^2), and that between-haul variation in catchability caused all additional Poisson over-dispersion in bottom-trawl survey catches. Cadigan et al.[2] demonstrated that this type of

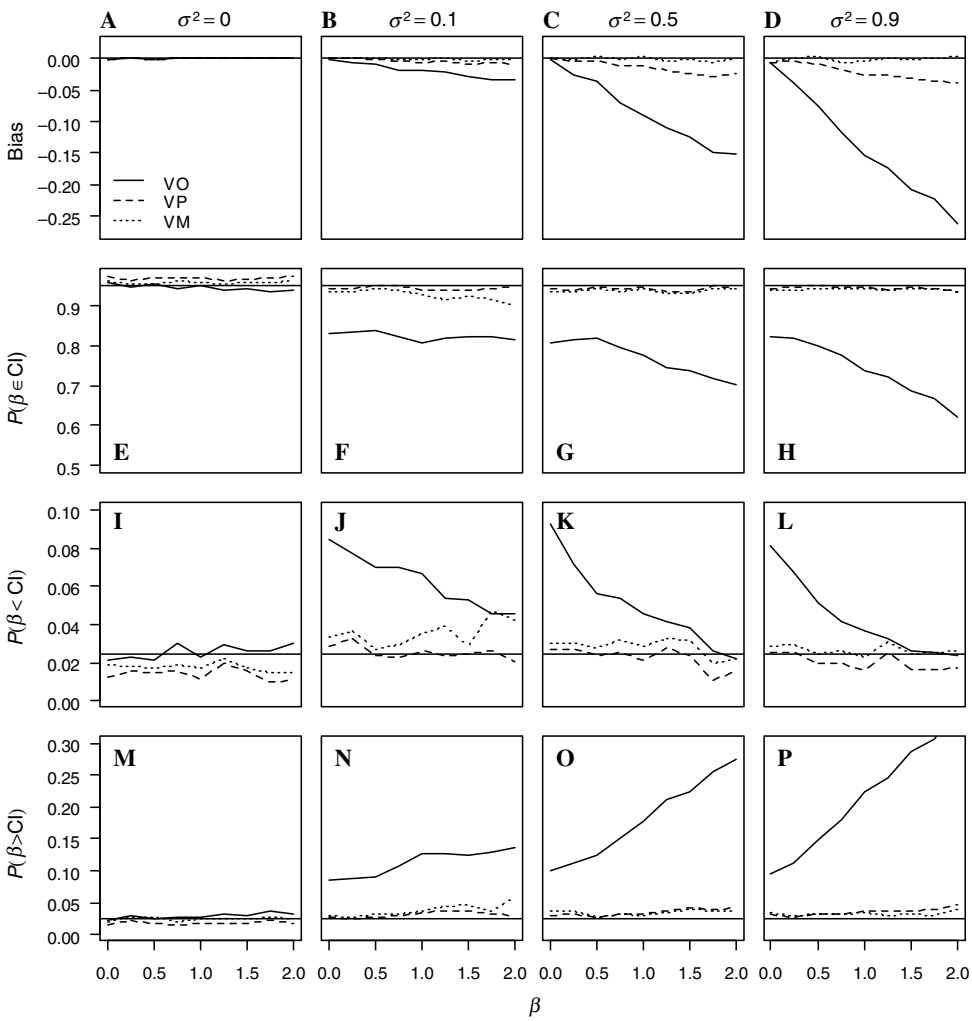

Figure 4

Paired-tow comparative fishing simulation results for log relative efficiency (β) of the test vessel compared to the control vessel. Simulations were based on the yellowtail flounder (*Limanda ferruginea*) scenario and data were generated for different assumed values of β (i.e., β_0) and spatial heterogeneity (σ^2) in fish densities encountered in each tow. Three models of spatial heterogeneity, described in Table 1, were used to estimate β, and three line patterns are used to show the results from each model. Panel columns are for levels of σ^2 (i.e., σ^2=0 in **A**, **E**, **I**, and **M**, etc.) and the x-axis of each panel are for levels of β. Bias (**A–D**) is the simulation median estimate of β minus β_0. CI indicates confidence interval, and P($\beta \in$ CI) indicates the probability the CI contains β, etc. References lines (solid) are shown in each panel, at zero (**A–D**), 0.95 (**E–H**), and 0.025 (**I–P**).

variability has a similar effect to that of spatial heterogeneity in stock densities encountered by both vessels at a trawl station. Hence, these two sources of variability are confounded in paired-trawl experiments and the random effects represent the cumulative impacts of both types of variability. For reasons outlined in the previous paragraph, we recommend the GLMM approach when there is between-set variability in catchability.

Our power analyses indicated that 50% changes in catchability could not be detected with high probability (i.e., 0.95) for some species. For example, with data like that obtained for Atlantic cod (see Table 2, and Cadigan et al.[3]) the power was fairly low. For this stock, the estimate of σ^2 was 0.99 and our power analysis indicated that we could detect only large changes in catchability (>90%) in this situation. Estimates of σ^2 were closer to 0.5 for most other species, in which case the power to detect a 50% change in catchability would be between 0.8 and 0.95. The exception was for yellowtail flounder which would have even lower power when σ^2=0.5 because of the smaller number of positive sets (n=24) for this species. Changes in catchability between 20%

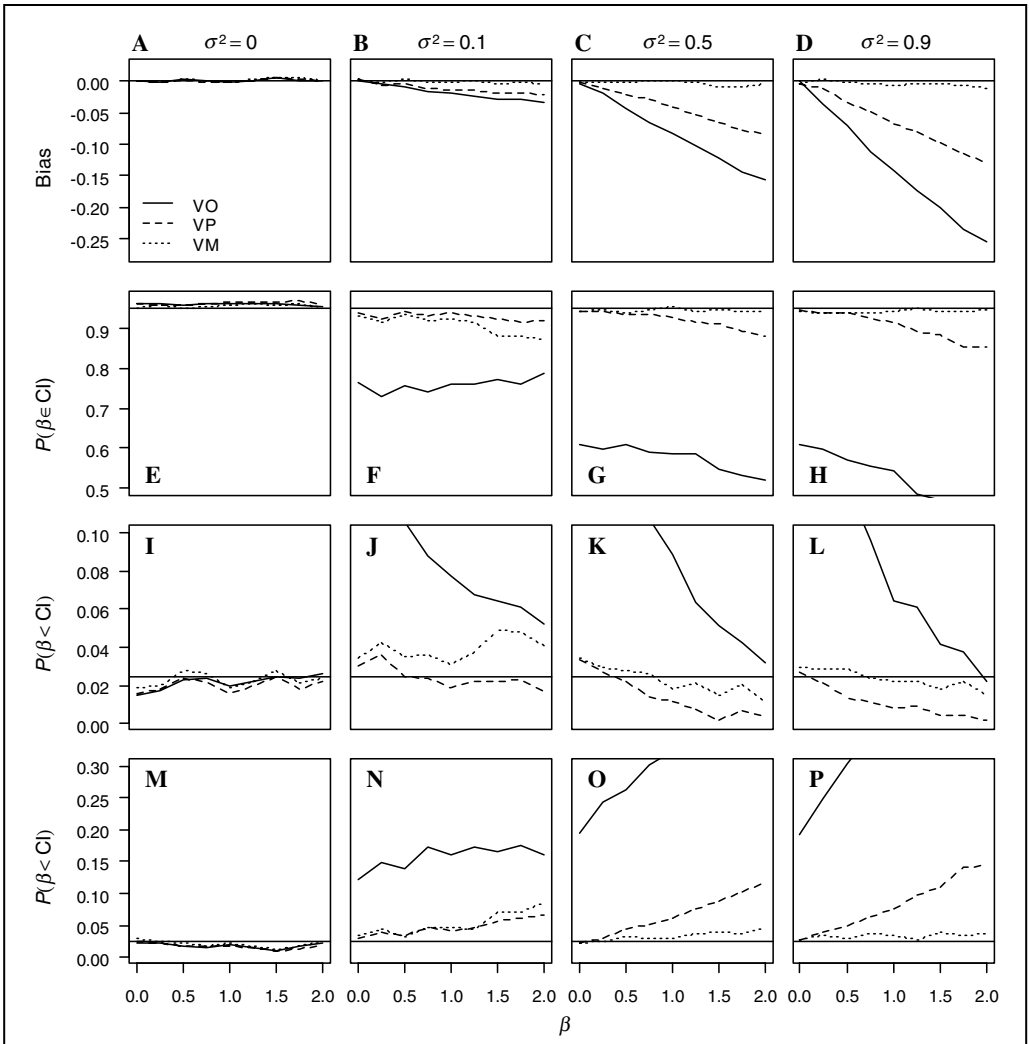

Figure 5

Paired-tow comparative fishing simulation results for log relative efficiency (β) of the test vessel compared to the control vessel. Simulations were based on the thorny skate (*Raja radiata*) scenario and data were generated for different assumed values of β (i.e., β_0) and spatial heterogeneity (σ^2) in fish densities encountered in each tow. Three models of spatial heterogeneity, described in Table 1, were used to estimate β, and three line patterns are used to show the results from each model. Panel columns are for levels of σ^2 (i.e., $\sigma^2=0$ in **A**, **E**, **I**, and **M**, etc.) and the x-axis of each panel are for levels of β. Bias (**A–D**) is the simulation median estimate of β minus β_0. CI indicates confidence interval, and P($\beta \in$CI) indicates the probability the CI contains β, etc. References lines (solid) are shown in each panel, at zero (**A–D**), 0.95 (**E–H**), and 0.025 (**I–P**).

and 50% would be important in stock assessment, and our simulation results indicated that more sets would be necessary to detect such changes in a comparison of paired-tow fishing data when the amount of spatial heterogeneity is similar to the levels in Cadigan et al.[3]. If spatial heterogeneity could somehow be removed or kept low, then 50% changes in catchability could be detected with high power.

Another common approach to analyze comparative fishing data is to log transform catches and use normal linear models for analysis; however, this approach does not often adequately account for the stochastic nature of the data (e.g., counts) and involves arbitrary choices to deal with zero catches. However, the lognormal approach may be reasonable and appropriate in some situations, or when the focus is on catch weights (e.g., Kingsley et al., 2008).

We studied two methods to estimate GLMMs. One was based on maximizing the marginal likelihood, integrated over the random effects. The other approach was penalized quasi-likelihood estimation based on a linearization of the model and a double optimization

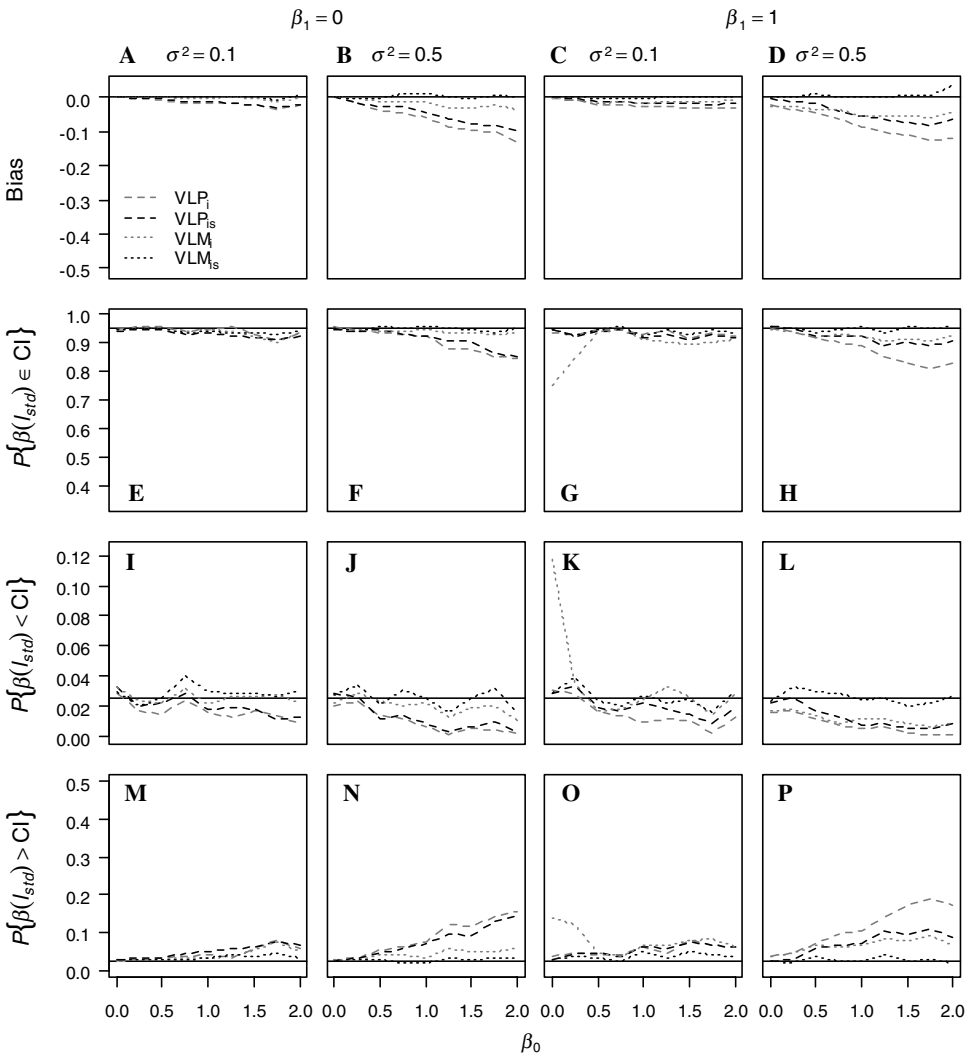

Figure 6

Paired-tow fishing simulation results for log relative efficiency (β) of the test vessel compared to the control vessel. Simulations were based on the Atlantic cod (*Gadus morhua*) scenario and fish-length–based data were generated for different assumed values of $\beta(l)=\beta_0+\beta_1 l$ and spatial heterogeneity (σ^2) in fish densities encountered in each tow. Lengths were standardized, $l_{std}=(l - l_{50})/(l_{75}-l_{25})$, where l_α was the $\alpha\times100\%$ percentile of the lengths caught in all sets. Four models of spatial heterogeneity, described in Table 1, were used to estimate $\beta(l)$, and four line patterns and shadings are used to show the results from each model at $l_{std}=0$. Panel columns are for levels of σ^2 and β_1 (i.e., $\sigma^2=0$ and $\beta_1=0$ in **A**, **E**, **I**, and **M**, etc.) and the x-axis of each panel are for levels of β_0. Bias (**A–D**) is the simulation median estimate of β minus β_0. CI indicates confidence interval, and $P\{\beta(l_{std})\in CI\}$ indicates the probability the CI contains $\beta(l_{std})$, etc. References lines (solid) are shown in each panel, at zero (**A–D**), 0.95 (**E–H**), and 0.025 (**I–P**).

procedure. The advantage of the latter approach was its ability to accommodate more complicated types of random effects like those with autocorrelation. However, our simulations results indicated that estimates and CIs from the linearization method were less reliable than those from the marginal likelihood approach. We recommend the marginal approach to estimate GLMMs for comparative fishing data.

We demonstrated that statistical inferences from GLMMs based on normal distribution random effects were equally as accurate when the random effects were actually the log of a ratio of two independent and identically distributed gamma random variables, which we hypothesize is a real and important source of overdispersion in vessel calibration studies. This is good because otherwise we could not recommend the stan-

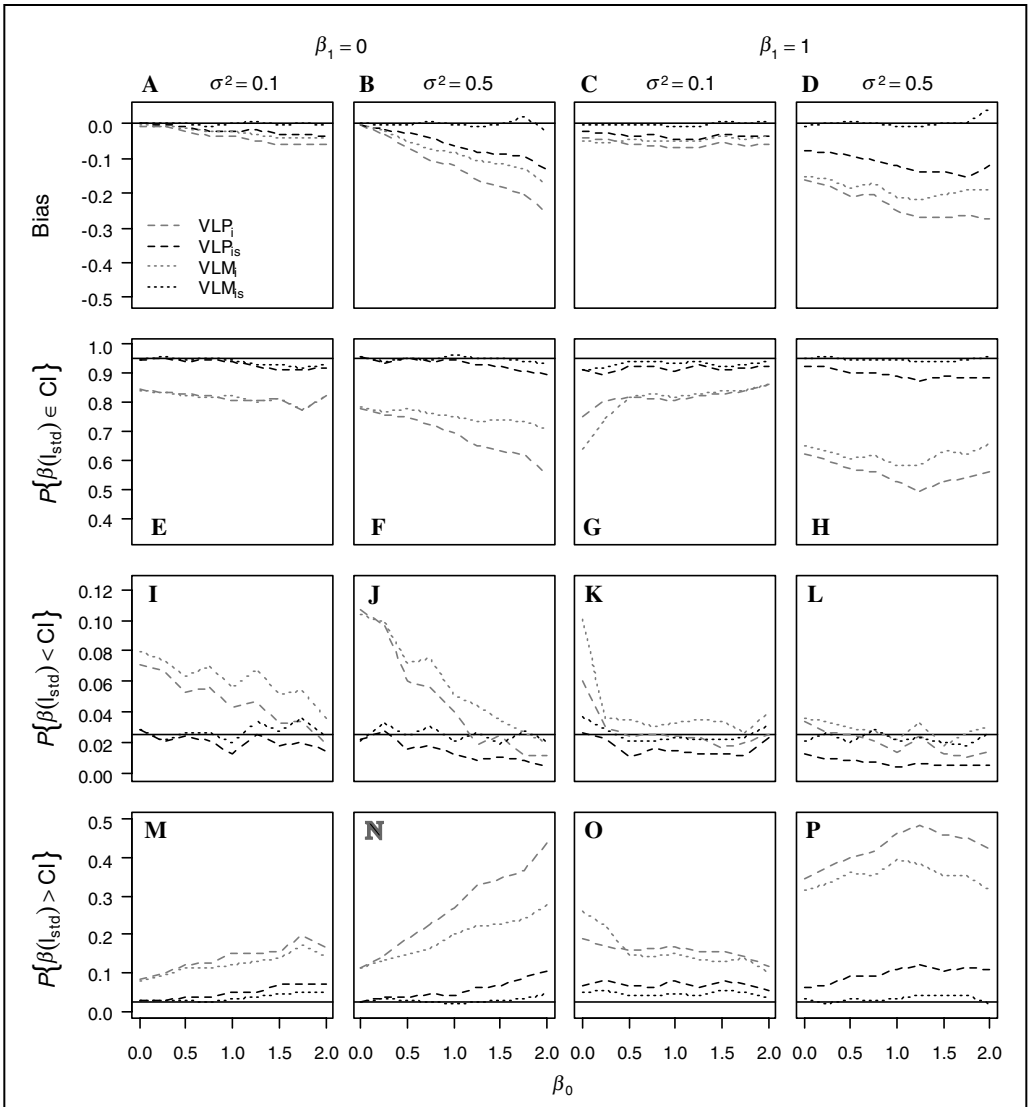

Figure 7

Paired-tow fishing simulation results for log relative efficiency (β) of the test vessel compared to the control vessel. Simulations were based on the Atlantic cod (*Gadus morhua*) scenario and length based data were generated for different assumed values of $\beta(l)=\beta_0+\beta_1 l$ and spatial heterogeneity (σ^2) in fish densities encountered in each tow. Lengths were standardized, $l_{std}=(1 - l_{50})/(l_{75}-l_{25})$, where l_α was the $\alpha\times100\%$ percentile of the lengths caught in all sets. Four models of spatial heterogeneity, described in Table 1, were used to estimate $\beta(l)$, and four line patterns and shadings are used to show the results from each model at l_{std} =1. Panel columns are for levels of σ^2 and β_1 (i.e., σ^2=0 and β_1=0 in **A**, **E**, **I**, and **M**, etc.) and the x-axis of each panel are for levels of β_0. Bias (**A–D**) is the simulation median estimate of β minus $\beta_0+\beta_1$. CI indicates confidence interval, and P{$\beta(l_{std})\in$CI} indicates the probability the CI contains $\beta(l_{std})$, etc. References lines (solid) are shown in each panel, at zero (**A–D**), 0.95 (**E–H**), and 0.025 (**I–P**).

dard GLMM approach. However, this does not mean that GLMMs are always robust to mis-specifications of random effects (e.g., Heagerty and Kurland, 2001). We demonstrated this in our length-based simulations.

We demonstrated that the random intercept model gave inaccurate statistical inferences (bias and CIs) if there was between-set random variations in length effects. This is a type of model mis-specification. However, we did not fully examine the other side of this result, which is bias caused by assuming between-set random variation in length effects when in fact none exists. Results from simulations with σ^2=0 for both vessel and length effects showed that the random intercept and slope model (i.e., VLM$_{is}$) provided reliable statistical

inferences when no random effects exist, and we speculate that the same result will hold when only random intercept effects exist. Hence, we recommend the VLM_{is} approach when there is either within-pair variation in the density of fish or in their length compositions.

A type of model we did not consider involves large-scale within-pair random variations in the densities of fish encountered by both vessels, and smaller-scale length-specific random variations in length compositions. When plotted like Figure 1, log density ratios would appear as a scatter of points about horizontal lines. The intercepts of the horizontal lines would depend on the large-scale random variation in the fish densities, and the variation in the scatter about these horizontal lines would depend on the smaller-scale length-specific random variations in length compositions. These types of effects can be modeled with hierarchical random effects (i.e., set, and length within set). This procedure would be fairly straightforward with a PQL approach but more difficult with marginal MLE. The reliability of estimates (relative efficiency and variance parameters) is uncertain. Even more complicated hierarchical random-effect models for both vessel and catch length-composition effects could be considered. This was beyond our scope but important to understand for reliable statistical inference. In reality, random effects in log density ratios may also have non-normal and skewed distributions, and it would be helpful to understand the robustness of VLM_{is} to these types of model mis-specifications.

Another sensible simulation procedure is to specify the stock densities fished by both vessels from site to site, and generate random catches for both vessels. The stock densities could be specified by using a variety of spatial models, as long as the within-pair and between-pair variations in densities are reasonably consistent with what one might expect in practice. However, this would not be a conditional simulation because the total catches by both vessels (and for all sets) would vary from simulation to simulation. It would still be useful to examine if the conditional CIs are accurate in this more general setting. However, the results from our conditional simulations were very similar for each of the seven species scenarios we considered and we anticipate they would also be accurate in the more general setting.

Cadigan et al. (2006) used a random effect for $\delta \equiv \delta(l)$ that was autocorrelated in length l, $\text{corr}\{\delta(l),\delta(l')\} = \gamma^{|l-l'|}$. This is a strategy to model smooth functions (e.g., Brown and de Jong, 2001). We used a simpler approach based on the assumption that $\delta(l)$ varied linearly with l. A GLMM in which $\delta(l)$ is autocorrelated can be fitted by using PQL software, but is time consuming to simulate and therefore we decided to focus on Equation 7 which is much easier to estimate.

Another approach to estimate relative efficiency (i.e., Pelletier, 1998) is maximum likelihood based on the negative binomial (NB) distribution. We have pursued this approach; however, there are complications in estimating the NB over-dispersion parameter based on the joint likelihood of both trawl catches and this problem affects the accuracy of CIs. The conditional approach is also more complicated for the NB distribution. We will report on this elsewhere.

Fryer et al. (2003) showed how to use spline methods for smooth, but otherwise nonparametric, estimates of relative efficiency. This is a useful estimation approach, especially to check the adequacy of a parametric model. The simple logistic-linear model we considered may be sufficient to test whether there is a significant length-dependent vessel effect but the logistic-linear model may not be sufficiently flexible for reliable estimation of relative efficiency over all lengths.

Lewy et. al. (2004) advocated paired-trawls along the same trawl track-line to avoid complications due to spatial variations in stock densities. However, such trawling introduces a different complication which involves disturbance of the fish densities encountered by the second vessel because of the fishing activity of the first vessel.

Another potential advantage of GLMMs is less sensitivity to outliers. Figures 14 and 15 of Cadigan et al.[2] indicated that GLMM estimates of β were less sensitive to outliers than GLIM estimates. This lack of sensitivity is a considerable advantage because identifying outliers is time consuming in practice when conversion factors are estimated for many species. In addition, standard errors may be too small when observations are incorrectly deemed to be outliers and removed from the analysis. Atlantic cod and thorny skate had mean total catches (for both vessels) that exceeded the 75th percentile (Table 2), indicating that there were a few large catches that may have undue influence on estimates. The GLMMs seem better suited for this type of data. It would be useful to test these methods through simulation to assess robustness to outliers. The randomization approach used by Benoît[1] is another appropriate and robust procedure (e.g., Cox and Hinkley, 1974, p. 180–181) to test for the statistical significance of vessel effects, and we recommend this approach in addition to the use of GLMMs. However, it does not provide robust estimates of vessel effects and the approach cannot replace a GLMM.

Acknowledgments

The authors are grateful for the expertise, assistance, and many discussions with S. Walsh and B. Brodie, of the Northwest Atlantic Fisheries Center, Fisheries and Oceans Canada (DFO). M. Koen-Alonso of DFO also provided comments on earlier draft.

Literature cited

Benoît, H.P., and D. P. Swain.
2003. Accounting for length- and depth-dependent diel variation in catchability of fish and invertebrates in an annual bottom-trawl survey. ICES J. Mar. Sci. 60:1298–1317.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens and J.-S. S. White.
2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24:127–135.

Brown, P. E., and P. de Jong.
2001. Nonparametric smoothing using state space techniques. Can. J. Stat. 29:37–50.

Cameron, A. C., and P. K. Trivedi.
1998. Regression analysis of count data, 411 p. Cambridge Univ. Press, Cambridge.

Cox, D. R., and D. V. Hinkley.
1974. Theoretical statistics, 511 p. Chapman and Hall, London.

Cox, D. R., and E. J. Snell.
1989. Analysis of binary data, 2nd ed., 236 p. Chapman and Hall, London.

Devore, J. L.
1991. Probability and statistics for engineering and sciences, 3rd ed., 716 p. Brooks/Cole Publ. Co., Pacific Grove, CA.

Fryer, R.J.
1991. A model of between-haul variability in selectivity. ICES J. Mar. Sci. 48:281–290.

Fryer, R.J., A.F. Zuur, and N. Graham.
2003. Using mixed models to combine smooth size-selection and catch-comparison curves over hauls. Can. J. Fish. Aquat. Sci. 60:448–459.

Gunderson, D. R.
1993. Surveys of fisheries resources, 248 p. John Wiley, New York.

Heagerty, P. J., and B. F. Kurland.
2001. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika 88:973–985.

Holst, R., and A. Revill.
1997. A simple statistical method for catch comparison studies. Fish. Res. 95:254–259.

Kimura, D. K., and D. A. Somerton.
2008. Review of statistical aspects of survey sampling for marine fisheries. Rev. Fish. Sci. 14:245–283.

Kimura, D. K., and H. H. Zenger.
2006. Standardizing sablefish (*Anoplopoma fimbria*) long-line survey abundance indices by modeling the log-ratio of paired comparative fishing CPUES. ICES J. Mar. Sci. 54:48–59.

Kingsley, M. C. S., K. Wieland, B. Bergström, and M. Rosing.
2008. Calibration of bottom trawls for northern shrimp. ICES J. Mar. Sci. 65: 873–881.

Lewy, P., J. R. Nielsen, and H. Hovgård.
2004. Survey gear calibration independent of spatial fish distribution. Can. J. Fish. Aquat. Sci. 61:636–647.

McCullagh, P., and J. A. Nelder.
1989. Generalized linear models, 2nd ed., 511 p. Chapman and Hall, London.

McCulloch, C. E., and S. R. Searle.
2001. Generalized, linear, and mixed models, 325 p. John Wiley & Sons, New York.

Millar, R. B.
1992. Estimating the size-selectivity of fishing gear by conditioning on the total catch. J. Am. Stat. Assoc. 87:962–968.

Millar, R. B, M. K. Broadhurst, and W. G. Macbeth.
2994. Modelling between-haul variability in the size-selectivity of trawls. Fish. Res. 67:171–181.

Pelletier, D.
1998. Intercalibration of research survey vessels in fisheries: a review and an application. Can. J. Fish. Aquat. Sci. 55:2672–2690.

Quinn, T. J., and R. B. Deriso.
1999. Quantitative fish dynamics, 542 p. Oxford Univ. Press, New York.

Reid, N.
1995. The roles of conditioning in inference. Stat. Sci. 10:138–199.

Trenkel, V. M., and H. J. Skaug.
2005. Disentangling the effects of capture efficiency and population abundance on catch data using random effects models. ICES J. Mar. Sci. 62:1543–1555.