Analysis of Lake Huron recreational fisheries data using models dealing with excessive zeros

Zhenming Su\textsuperscript{a, *}, Ji X. He\textsuperscript{b}

\textsuperscript{a} Institute for Fisheries Research, Michigan Department of Natural Resources and University of Michigan, 212 Museums Annex Building, 1105 N. University Avenue, Ann Arbor, MI 48109-1084, United States
\textsuperscript{b} Michigan Department of Natural Resources, Alpena Fisheries Research Station, 160 East Fletcher Street, Alpena, MI 49707, United States

\textbf{Abstract}

Excessive zeros in recreational catch data can cause problems for fish stock assessment and management. We evaluated a range of count regression models for analyzing the recreational catch data of walleye, Chinook salmon, and lake trout in Lake Huron. We also used modern predictive measures of effects to interpret the statistical results and extract year effects from the complex models. We found that models that account for both excessive zeros and overdispersion in recreational data, i.e., the zero-inflated negative binomial (ZINB) and hurdle negative binomial models, performed much better than those that cope with only one or none of the two common count data problems. Using the results from the best ZINB models, we identified important factors affecting catch rate of the three aforementioned species, and constructed standardized CPUE indices for each species.

© 2013 Published by Elsevier B.V.

\section{Introduction}

Timely monitoring, and assessing changes in the ecosystem and fisheries of the Great Lakes have important social, economic, and ecological values. This is exemplified by the drastic decline in prey fish abundance and Chinook salmon (\textit{Oncorhynchus tshawytscha}) fisheries in Lake Huron during the early and middle 2000s (Bence and Mohr, 2008; He et al., 2008; Riley et al., 2008). The decline in Chinook salmon fishery has cost Michigan major loss in Lake Huron fishing and other economic activities since 2004 (Dettmers et al., 2012).

Recreational fisheries in Michigan's Great Lakes waters are monitored through creel (on-site angler) surveys and charter-boat reporting systems (Su and Clapp, 2013). These monitoring programs provide total catch and fishing effort data that are essential for stock assessment and fisheries management. They also generate catch rate (catch-per-unit-of-effort, CPUE) data that are often used as an index of fish abundance.

Catch rates can be a distorted measure of trends in fish abundance, however, because they are affected by both fish abundance and factors that are unrelated to fish abundance, such as fishing site, season, target species sought by anglers, and changes in fishing techniques. Therefore, catch rates must be standardized to remove the effects of factors confounding fish abundance before they can be used as a reliable index of abundance (Maunder and Punt, 2004).

Catch-rate standardization refers to the process of isolating or removing effects of factors unrelated to trends in fish abundance. A historical method for standardizing catch and effort data involves selecting a "standard gear" and comparing the relative fishing power among all gears (Beverton and Holt, 1957). However, this method cannot account for the effects of those factors other than fishing gear. Modern approaches for standardizing catch and effort data utilize statistical models, such as generalized linear models (GLMs), to account for multiple factors that may confound the abundance trends in catch and effort data (Maunder and Punt, 2004; Quinn and Deriso, 1999). These models are built to relate catch or catch rates to a variable representing year effects and other explanatory variables that may cause variation in catch rates but are not related to abundance changes. The year effects can then be extracted from the fitted model and used as an index of abundance.

In Great Lakes waters, recreational catch data are characterized by an excessive number of zeros and small catch values. Many standard statistical models used for analyzing commercial or research-vessel catch data (Bishop et al., 2004; Deroba and Bence, 2009; Ye and Dennis, 2009) cannot deal with excessive zeros in the data (Levin et al., 2010; Minami et al., 2007; O'Neill and Faddy, 2003; Ortiz and Arocha, 2004). Hence, using these models to fit recreational catch data may lead to biased parameter estimates, erroneous uncertainty estimates (e.g., standard errors or confidence intervals), and incorrect assessment (Webley et al., 2011). For
such data, one should consider models that can handle excessive zeros.

Two commonly used modeling techniques that can account for excessive zeros in count data are hurdle models (Mullahy, 1986) and zero-inflated models (Lambert, 1992). A hurdle model is a two-part model that handles zero and positive counts separately (Mullahy, 1986). A zero-inflated model is a mixture model of a standard count distribution (e.g., Poisson or negative binomial) and a degenerate distribution at zero (Lambert, 1992). In fisheries and ecological literature, these models have been used to analyze abundance of rare species (Martin et al., 2005; Welsh et al., 1996), bluefin tuna (Thunnus thynnus) trap catches (Lemos and Gomes, 2004), shark bycatch data (Minami et al., 2007), and species–environment relationships (Lewin et al., 2010). With a few exceptions (e.g., O’Neill and Faddy, 2003), models that can handle an excessive number of zeros have rarely been used to analyze recreational fisheries data.

In this paper, we applied two standard count regression models (Poisson and negative binomial), two hurdle, and two zero-inflated count regression models to recreational fisheries data collected from creel surveys in Michigan waters of Lake Huron. Our objectives are to: (1) find the probability distributions that best describe these recreational catch data; (2) identify factors affecting recreational catch rates; and (3) standardize recreational fisheries catch rates. We used the Akaike information criterion (AIC) for model selection. We also used modern predictive measures of effects, such as predictive margins and differences (Gelman and Hill, 2007), to interpret the statistical results and to extract year effects from the complex hurdle and zero-inflated models.

We applied the aforementioned statistical models to catch and effort data for three recreational fish species in Lake Huron: Chinook salmon, lake trout (Salvelinus namaycush), and walleye (Sander vitreus). Chinook salmon was the main target species in Lake Huron before 2004 (Bence and Mohr, 2008), but due to drastic changes in Lake Huron’s food web, Chinook salmon fisheries have declined substantially since 2004. Meanwhile, abundance and harvest rates of walleye and lake trout have increased. Walleye has become the major recreational species in Lake Huron. Thus, developing standardized catch rates for these three important recreational species and understanding major factors influencing the observed catch rates have important management implications.

2. Data and methods

2.1. Data

We used angler or boat trip data obtained from creel surveys conducted from 1987 to 2011 by the Michigan Department of Natural Resources (MDNR), Fisheries Division (Adlerstein et al., 2008; Su and Clapp, 2013). The data include fishing trip, species, and catch information obtained from interviews of anglers or angler parties who had finished their fishing trips (i.e., access or completed trips).

To demonstrate our methods, we focused our analysis on Chinook salmon and lake trout data from statistical district MH-2 and

![Fig. 1. Map of Lake Huron. MH-1 to MH-6 indicate statistical districts used for fisheries management purposes in Michigan.](image-url)
on walleye data from MH–4 (Fig. 1). MH–2 is a typical salmon and trout recreational fishing area, whereas MH–4 is the main walleye fishing area. We only analyzed trips made by boat angling in the peak fishing season (May–September) for the three species. Data for other fishing modes (e.g., shore and pier) and other months are very sparse and highly variable. For Chinook salmon and lake trout, we excluded trips in which anglers indicated targeting fish species other than salmon and trout. For all three species, we excluded trips in which anglers targeted fish species other than those interest herein. Because information on fish caught and released was not collected before 2000, we only used data on fish caught and kept (harvest), which was referred to as “catch” in the text. Table 1 lists variables used in the analysis. Fig. S1 (on-line Supplemental materials) and Fig. 2 (barplots) illustrate data for targeted species and the empirical distributions of recreational catch, respectively.

2.2. Methods

We evaluated six count regression models for recreational catch data from Lake Huron: (1) Poisson regression, which is a default or reference model for count data, (2) negative binomial (NB) regression, which is a more flexible model than the Poisson in that it can cope with overdispersion or extra unexplained variation in count data that cannot be accommodated by Poisson regression, and (3) four hurdle and zero-inflated models that can deal with excessive zeros in count data. The NB variant of the hurdle and zero-inflated models can also cope with overdispersion in the positive counts of the data (Zuur et al., 2009).

The Poisson regression model is the simplest and standard statistical model for count data, such as catch (Cameron and Trivedi, 1998). There is a major deficiency with the Poisson distribution though, as it implicitly assumes that the variance equals the mean ("equidispersion") (Cameron and Trivedi, 1998). Such an assumption is untenable for most observed count data as they tend to be overdispersed, with a variance larger than the mean. For such data, the Poisson model often underestimates variance and overstates significances (Cameron and Trivedi, 1998).

Many alternative approaches have been suggested to relax the unrealistic equidispersion assumption of the Poisson distribution (Cameron and Trivedi, 1998; Hilbe, 2011). A common alternative is the negative binomial (NB) distribution, which extends the Poisson distribution by introducing extra unobserved heterogeneity into the Poisson model (Cameron and Trivedi, 1998). In this paper, we use the so-called NB2 form of the negative binomial distribution (Cameron and Trivedi, 1998), NB(yi | μi, θ), where yi denotes the number of fish caught (catch) by angling trip i and μi denotes the...
mean catch. The $\theta > 0$ represents a dispersion or scale parameter (Venables and Ripley, 2002). The conditional variance of the NB2 form is $\text{Var}(y_i|\mu_i, \theta) = \mu_i(1 + \mu_i/\theta) > \mu_i$. Since $\mu_i$ and $\theta$ are defined to be greater than zero, the NB distribution allows for overdispersion. For a given $\mu_i$, the overdispersion $(1 + \mu_i/\theta)$ is more severe with smaller $\theta$. When $\theta \to \infty$, the Poisson distribution is obtained.

In addition to overdispersion, many observed count data may also contain more zeros than expected under the Poisson or NB distribution (Mullahy, 1986; Zeileis et al., 2008). In the following, we describe two commonly used modeling techniques that can account for excessive zeros in count data.

2.2.1. Hurdle models

Hurdle models are one class of model capable of dealing with excessive zeros. A hurdle model is a two-part model consisting of a binary distribution used to determine whether a count is positive or zero and a zero-truncated count distribution modeling positive counts (Hilbe, 2011; Zorn, 1996).

Because our recreational catch data are characterized by excessive zeros and skewness, we specify a hurdle model in two stages. In the first stage, let $I_i$ represent an indicator variable for fishing success (positive or zero catch), and let $I_i = 1$ if $y_i > 0$ and $I_i = 0$ if $y_i = 0$. Then, $I_i$ can be modeled by logistic regression as follows:

$$I_i \sim \text{Bernoulli} (\pi_i)$$

where $\pi_i = Pr(y_i > 0) = 1 - P(y_i = 0)$, “Bernoulli” denotes the Bernoulli distribution, $\logit^{-1}(\pi_i)$ is the inverse logit function of $\pi_i$, and $\gamma$ in the vector of predictors that may affect the success of a fishing trip, and $\gamma$ is a vector of regression coefficients.

In the second stage, one option is to model the positive catch $y_i > 0$ using a zero-truncated Poisson distribution (i.e., a Poisson distribution with $y_i = 0$ excluded, denoted as TruncPois):

$$y_i \sim \text{TruncPois} (\mu_i = E_i R_i), \quad y_i = 1, 2, \ldots$$

where $E_i$ is the fishing effort (angler-hours) for trip $i$, and $\mathbf{x}_i$ is a vector of predictor variables. In Eqs. (2) and (3), the mean of the standard (un-truncated) Poisson distribution is $\mu_i = E_i R_i$, with $R_i$ representing the mean catch rate (catch per angler-hour). Eqs. (2) and (3) define a model for catch rate, with $\beta$ representing the effects of predictors on catch rate. Such a model is known as a rate model, with $\log(E_i)$ defined as the offset (Gelman and Hill, 2007).

An alternative option is to model the positive catch using a zero-truncated negative binomial distribution (a NB distribution with $y_i = 0$ excluded, denoted as TruncNB):

$$y_i \sim \text{TruncNB} (\mu_i = E_i R_i, \theta)$$

The expected count under the hurdle model defined above is:

$$E(y_i|\pi_i, \mu_i) = \frac{\pi_i \mu_i}{(1 - f(0|\mu_i))}$$

where $f$ denotes the probability mass function (PMF) of a standard count distribution (e.g., the Poisson or NB distribution). $\mu_i$ is the mean of the standard count distribution, and $f(0|\mu_i)$ is the probability of $y_i = 0$.

The combination of the logistic part (Eq. (1)) and a Poisson (Eqs. (2) and (3)) or NB (Eqs. (4) and (5)) part constitutes a hurdle-Poisson regression model (HP) or a hurdle-NB regression model (HNB). Hurdle-Poisson models can deal with excessive zeros but cannot deal with over- or under-dispersion, whereas hurdle-NB regression models can deal with both over- or under-dispersion and excessive zeros (Lewin et al., 2010).

2.2.2. Zero-inflated models

An alternative approach for handling excessive zeros in count data is zero-inflated (ZI) count models (Lambert, 1992). A zero-inflated model assumes that counts are generated from a mixture of two distributions: one is a degenerate distribution at zero (i.e., a point mass at zero) with probability $\phi_i$, and the other is a standard count distribution $f$, such as the Poisson or NB distribution, with probability $1 - \phi_i$ (Min and Agresti, 2005). Thus, $y_i$ in a zero-inflated model follows the PMF:

$$Pr(y_i|\phi_i, \mu_i) = \begin{cases} \phi_i + (1 - \phi_i)f(0|\mu_i), & y_i = 0 \\ (1 - \phi_i)f(y_i|\mu_i), & y_i \geq 1 \end{cases}$$

Based on the PMF, zeros in a ZI model may come from the degenerate distribution or the count distribution. The mean of $y_i$ in a zero-inflated distribution is:

$$E(y_i|\phi_i, \mu_i) = (1 - \phi_i)\mu_i$$

Similar to the probability parameter, $\pi_i$, in a hurdle model, we link $\phi_i$ to a set of predictors, $\mathbf{z}_i$, by a logistic component:

$$\phi_i = \logit^{-1}(\mathbf{z}_i' \beta)$$

where $\gamma$ is a vector of regression coefficients.

The expected catch rate, $R_i$, in a zero-inflated model is related to predictors ($\mathbf{z}_i$) by either a Poisson model or a NB model such that:

$$\log(E_i R_i) = \log(E_i) + \mathbf{z}_i' \beta$$

The combination of the logistic regression (Eq. (9)) with the Poisson or NB regression (Eq. (10)) forms a zero-inflated Poisson (ZIP) or zero-inflated NB model (ZINB). ZINB models can deal with excessive zeros but cannot deal with overdispersion, while ZINB models can deal with both overdispersion and excessive zeros (Lewin et al., 2010).

2.3. Model estimation and selection

We fitted six regression models, i.e., Poisson (POIS), negative binomial (NB), hurdle-Poisson (HP), hurdle-negative binomial (HNB), zero-inflated Poisson (ZIP), and zero-inflated negative binomial (ZINB), to walleye, Chinook salmon, and lake trout recreational catch data from Lake Huron. We performed the Poisson regression using the glm() function (Chambers and Hastie, 1992), the NB regression using glm.nb() from the R MASS package (Venables and Ripley, 2002), the ZIP and ZINB regressions using zeroInfl() from the R pscl package (Jackman, 2012), and the HP and HNB using hurdle() from pscl (on-line Supplemental materials). In the pscl package, the logistic component of the hurdle models describes the probability of observing a positive count (see Eq. (1)) (Zeileis et al., 2008).

The explanatory variables examined were year, month, day type (weekend or weekday day), time of fishing in a day (morning or afternoon), trip length (hours fished in a fishing trip), target species, party size (number of anglers in a fishing party), and catch of another related species (Table 1). Fishing effort was always treated as an offset in the regression.

Model selection in our analysis involved two aspects: one was to choose suitable probability distributions (i.e., POIS, NB, HP, HNB, ZIP, and ZINB) for the dependent variable (i.e., catch), and the other was to select predictors for inclusion into a model. The latter aspect, also called variable or subset selection, is the process of selecting the best subsets of predictor variables (including linear, quadratic, and interaction terms) from a larger set. Both aspects can affect the fit and predictive power of the models.
We performed variable selection using `stepAIC()` from the R MASS package for each of the six regression models of the three species. To run this procedure, we first defined the range of models to be searched. The minimal model in the range only included the YEAR variable, which was always included in the model, and the maximal model included all the explanatory variables considered along with their two-way interactions. The offset term, log(EFFORT), was always kept in the models. We deemed the best subset returned from `stepAIC()` for a regression model to be the one with the smallest Akaike information criterion (AIC) value (online Supplemental materials).

The six regression models with the subset of relevant predictors selected from the previous step were further compared, and the regression model with the smallest AIC was treated as the best regression model for a species and used for further analyses. We also considered the goodness of fit of the six regression models to the frequencies of the counts (i.e., catch).

We encountered problems when performing estimation and variable selection using the aforementioned R functions. First, due to the large size of the data sets (~50,000 records each), the design matrix for a model with interaction terms among some discrete variables, such as YEAR and MONTH, became quite large, and exceeded memory-size limits of the 32-bit builds of R. This was especially acute for the zero-inflated and hurdle models because they had two sets of design matrices. In this case, the dimension of the design matrix of a ZINB model approached 50,000 × 400. The memory problem was overcome by performing the analyses with 64-bit R running on a 64-bit Ubuntu Linux operating system. Second, the data were quite unbalanced in terms of YEAR, MONTH, and target species, especially for lake trout and Chinook salmon (Fig. 2). For some combinations (interaction levels) of these variables, the number of observations (trips) could be zero or very low, and that led to estimation problems for the parameters (coefficients) for these regression terms. For the zero-inflated and hurdle models, we removed predictors whose coefficients were poorly estimated, as indicated by their large or non-finite standard errors due to lack of data. These sparse data problems can be dealt with in the future by treating the interaction terms as random effects (Gelman and Hill, 2007).

### 2.4. Interpreting the effects of factors affecting catch rate

One purpose of this study was to understand the effects of various factors on recreational catch rate and to obtain year effects that may be used as an index of fish abundance. In a linear regression model without interactions, the effect of an explanatory or input variable of interest, \(x_k\), on the outcome variable \(y\) can be interpreted directly based on its coefficient \(\beta_k\): a unit change in \(x_k\) is associated with an expected change in \(y\) by \(\beta_k\) units while being identical in all the other inputs (Gelman and Hill, 2007). However, the effect of an input variable in nonlinear regression models, including the count models considered in this paper, can be difficult to interpret based on the regression coefficient alone.

Various measures have been suggested in the literature to interpret the effects of input variables on the outcome variable from nonlinear models (Gelman and Hill, 2007; Greene, 2012; Long, 1997). We used predictive margins (Graubard and Korn, 1999) and average predictive differences (Gelman and Hill, 2007) to describe how the response variable (e.g., catch rate) varies as one or more input variables change. We define the predictive margin for an input variable, \(x_k\), at a given value \(r\) as:

\[
\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} \tilde{y}_i(x_k = r, x_{-k,i}, \hat{\beta})
\]

where \(n\) is the number of observation units, \(\tilde{y}_i(x_k = r, x_{-k,i}, \hat{\beta})\) is the predicted or fitted value of \(y_i\) for the \(i\)th unit evaluated at \(x_k = r\), the observed values of other input variables \(x_{-k,i}\), and the parameter estimates \(\hat{\beta}\). Predictive margins are also called margins of responses, adjusted predictions, recycled predictions, and partial dependencies (Hastie et al., 2001; StataCorp, 2009). They are known as estimated marginal means or least-squares means for balanced data. Predictive margins can be plotted over a range of \(x_k\) values to visualize the effects of \(x_k\) on \(y\) (Hastie et al., 2001). The resulting plots are called profile plots, or partial dependence plots (Hastie et al., 2001).

Average predictive difference (APD) is the difference between the predictive margins calculated at two levels of \(x_k\) (1 and \(h\)):

\[
\text{APD}_{kh} = \hat{m}_{kh=1} - \hat{m}_{kh=-1}
\]

Predictive differences are also known as first differences (King et al., 2000), discrete changes (Long, 1997), or partial effects (Hilbe, 2011). APD compares the average of the differences in predicted response values at two levels of an input variable while controlling for the effects of other variables. Because the computation of predictive margins and average predictive differences relies only on model predictions, they are often used to interpret complex predictive models (Hastie et al., 2001).

We further adopted a simulation-based approach of King et al. (2000) and Gelman and Hill (2007) to quantify the uncertainty of the predictive margin and average predictive difference. In this approach, we first generated \(M\) (e.g., \(M = 1000\)) sets of simulated parameter values from an approximate multivariate normal distribution of the parameters, with the covariance matrix set to the estimated covariance matrix obtained from the regression procedures. Predictive margin and average predictive difference were then calculated for each set of the simulated parameter estimates. The standard errors and confidence intervals (CIs) of the predictive margins and average predictive differences were calculated over the M simulations and used to quantify their uncertainty.

In our analysis, the average predictive difference of a binary input variable is calculated at the two levels (0 or 1) of the variable. For a continuous input variable, the average predictive difference was calculated at the 10% and 90% quantiles of the variable. For example, the average predictive difference for the continuous variable, ANGCNT, was computed at ANGCNT = 1 and ANGCNT = 4 (Table 1). For a categorical variable, such as MONTH or YEAR, predictive margins were calculated at each level of the variable and plotted against its levels to demonstrate the effects of the variable (see the on-line Supplemental materials).

### 3. Results

#### 3.1. Model comparison

The empirical distributions of recreational catch of Chinook salmon, lake trout, and walleye in Lake Huron are all highly skewed to the right and show extremely high frequencies of zero catches (Fig. 2, barplots). For these data, our model selection indicates that the zero-inflated negative binomial (ZINB) model is the most preferred model (with the lowest AIC) among the six regression models (POIS, NB, HP, HNB, ZIP, and ZINB) for all three species (Fig. S2, online Supplemental materials). The ranking order from best to worst for the six models in terms of AIC is ZINB, HNB, NB, ZIP, HP, and POIS for walleye and lake trout, and ZINB, HNB, ZIP, NB, HP, and POIS for Chinook salmon (Fig. S2). The top five models have much smaller AIC values than that of Poisson (POIS), indicating that Poisson is not an appropriate distribution for these catch data (Fig. S2).

Comparisons between the predicted and observed frequencies of catch per angling trip also reveal the inadequacy of NB, ZIP,
HP, and POIS (Fig. 2). The Poisson model (POIS) seriously underestimates the number of zero-catch trips, and overestimates the number of small positive-catch trips for all three species (Fig. 2). The hurdle-Poisson (HP) and zero-inflated Poisson (ZIP) models fit the observed frequencies of zero-catch trips, but do not fit well other observed frequencies. The negative binomial (NB) model has difficulties fitting the observed frequencies of zero- and positive-catch trips. In contrast, both the hurdle NB (HNB) and zero-inflated NB (ZINB) models fit the observed frequencies of zero- and positive-catch trips adequately (Fig. 2).

In the on-line Supplemental materials, we provide coefficient estimates and their estimated standard errors, and model-fitting measures obtained from the six regression models that included only main effects (i.e., linear terms) as predictors (Table S1 and Fig. S3). They provide similar conclusions as those described in the previous paragraphs.

The ZINB models with the best subset of predictors obtained from subset selection are given in Table 2. These models are used for further analyses as presented in the following sections.

3.2. Effects of explanatory variables

Using average predictive differences and predictive margins calculated from the ZINB models (Table 2), we examined the effects of the most important input variables on catch rates of the three species. Several points are observed for the effects of factors affecting the catch rates of the three species. First, trips that targeted walleye had a catch rate about 0.19 fish per angler-hour higher on average than those targeting other species (Fig. 3a). Targeted fishing also had a large positive effect on the catch rate of lake trout and Chinook salmon (Fig. 3b and c). For all three species, ANGCNT (angler-party size) had a negative effect on catch rate, that is, angling parties with more anglers had lower catch rate than those with fewer anglers (Fig. 3). Furthermore, trips with high catch rates of either Chinook salmon or lake trout had a positive effect on the catch rates of the other species. In contrast, TOD, DAYTYPE, and TRIPLEN had little or no effect on catch rates for the three species.

Catch rates of the three species show seasonal patterns as indicated by the predictive margins by month (Fig. 4). Catch rates of walleye peaked in June and July, and were the lowest in September. Those of lake trout peaked in May, and then decreased to a low level in August. The peak months for catch rates of Chinook salmon were July and August.

3.3. Standardized CPUE indices

Predictive margins of catch rate along with their 95% confidence intervals (Fig. 5) were calculated for each year for the three species based on the ZINB models (Table 2) and were treated as standardized CPUE indices. Wide 95% confidence intervals were found in some estimates of predictive margins of the catch rates in the lake trout and Chinook salmon data.

Walleye CPUE fluctuated cyclically at a low level of about 0.10 fish per angler-hour before 2006 (Fig. 5a). Then the CPUE jumped to a higher level of about 0.22 in 2006, and peaked at 0.32 in 2008. The CPUE dropped to an intermediate level of 0.16 and 0.19 in 2010 and 2011.

Lake trout CPUE experienced three periods of changes (Fig. 5b). From 1987 to 1995, lake trout CPUE was at a very low level (<0.03 fish per angler-hour). Then the CPUE jumped to a higher level of about 0.09 in 1996 and 1997. However, the CPUE declined rapidly to a low level of 0.03 in 2003. The CPUE has jumped again to 0.08 since 2004.

Chinook salmon CPUE varied around 0.13 fish per angler-hour before 2005, then declined rapidly to a very low level of merely 0.014 in 2009 (Fig. 5c). The Chinook salmon CPUE was still very low in 2010 and 2011 (Fig. 5c).

4. Discussion

We evaluated a range of count regression models for analyzing recreational catch data of walleye, Chinook salmon, and lake trout in Lake Huron. We found that models (i.e., ZINB, HNB) that account for both excessive zeros and overdispersion in the recreational catch data performed better than those that can cope with only one or none of these data problems. Using the results from the best ZINB models, we identified important factors affecting catch rate and catch of the three species, and constructed standardized CPUE indices for these species. We found that the standardized recreational CPUE indices of walleye, Chinook salmon, and lake trout in Lake Huron show distinct seasonal (monthly) variations and annual trends. These trends are consistent with the major changes that occurred in the lake. We also found that information on species targeted, party size, catch of related species, time of fishing in a day, and type of day all had various effects on catch rates of the three species.
Table 2
Predictors of the count and zero (logistic) components of the zero-inflated negative binomial (ZINB) regression model obtained from subset selection for walleye, Chinook salmon, and lake trout.

<table>
<thead>
<tr>
<th>Species</th>
<th>Count</th>
<th>Zero (logistic)</th>
</tr>
</thead>
</table>

We tackled two difficult problems in using complex statistical models to model catch and effort data and obtain standardized CPUE indices: interpreting model results, and extracting year effects. These difficulties are due to the complexity of the models and the large amount of data involved. We adopted modern predictive measures of effects (i.e., predictive margins and average predictive differences) to interpret the results from the ZINB models, which involves a mixture distribution of a degenerate distribution of zeros and a NB distribution in the stochastic component and contains interaction terms in the regression component. These effect measures directly answer the most important question of a regression analysis: what is the expected change in the outcome for a unit change in one of the input variables?

The major changes shown in our standardized recreation CPUE indices were consistent with the major changes that occurred in the lake. The sharp decline in Chinook salmon catch rate after 2004 was due to the collapse of the alewife population in 2003 (Riley et al., 2008) and the subsequent recruitment failure of hatchery-stocked Chinook salmon (Johnson and Gonder, 2012). In contrast, recruitment of naturally-reproduced walleye increased substantially immediately after the collapse of alewives (Fielder et al., 2007; Fielder and Thomas, in press). Nevertheless, lake trout abundance did not show any apparent response to the substantial changes in prey fish abundance because they have a much longer life span than both Chinook salmon and walleye, and have adapted rapidly to the new and evolving prey fish community (He et al., 2008, 2012).

In Lake Huron, annual fishery-independent surveys have been conducted for lake trout in spring (late April through early June, gillnetting) since 1975 (He et al., 2012) and for walleye (September, bottom trawling) since 1970 (Fielder and Thomas, in press). In contrast, fishery-independent surveys are not available for Chinook salmon. Our standardized CPUE index for lake trout broadly resembles the adult CPUE index obtained from the gillnet surveys (see Fig. 3b in He et al., 2012). Lake trout caught in the recreational fishery are typically age-4 and older, and the adult lake trout in He et al. (2012) were defined as those fish with total length greater than 533 mm, corresponding to fish about age-5 and older, thus the two indices are comparable. Both indices show a period of low abundance before 1996 and increased abundance after 1996. The increase in abundance is related to the change in age and size composition of the lake trout stock. Before 1996, the lake trout stock was mainly composed of small and young lake trout, and the adult abundance increased to a substantially higher level after 1996 (He et al., 2012), leading to higher recreational catch rates.

Our walleye recreational CPUE index is very similar to the fishery-independent trawl index, but the former shows a big jump in catch rate in 2006, rather than in 2003 as in the trawl CPUE index (Fig. 5a; Fielder and Thomas, in press). Fish caught in the trawl surveys included all age groups (>age-0). In contrast, the recreational fishery was subject to a minimum length limit (38.1 cm, about age-3). Beginning in 2003 and through at least 2005, there was an unprecedented surge of walleye reproductive success (Fielder et al.,

![Fig. 4. Predictive margins of catch rates by month for walleye, lake trout, and Chinook salmon calculated from the zero-inflated negative binomial (ZINB) model with the best subset of predictors.](image-url)
It took three years for these strong year classes to recruit into the fishery, causing a three-year difference in the two CPUE indices.

Catch and catch rates in consecutive periods (e.g., days or weeks) may be correlated and clustered. Such a cluster effect can be incorporated in zero-inflated generalized linear mixed-effects models (ZI-GLMMs), which are a class of models that incorporate aspects of generalized linear models, mixed models, and zero-inflated models (Bolker et al., 2009). However, GLMMs are an active area of research. Although several software packages are available for fitting these models, they are very computational expensive and are not readily comparable with each other. Therefore, we will consider the application of GLMMs to the recreational data in future studies.

Compared to the fishery-independent survey data, catch rates of recreational fisheries were obtained as a by-product of regular creel surveys and fishing reports whose primary purpose is to provide total catch and effort estimates (Su and Clapp, 2013). Thus, the information of recreational catch rate is available at no extra cost. Recreational fisheries data also have greater coverage in time and space than the fishery-independent surveys. As such, they may provide information of fish abundance and distribution not available from the fishery-independent surveys. On the other hand, the fishery-independent surveys provide abundance information for all age groups and their catchability may be better controlled than that of the recreational fisheries. Therefore, these two kinds of data are complementary and both are valuable. Still, fishery-independent data are not available for species such as Chinook salmon in Lake Huron. For these species, recreational catch rates provide the only source of relative abundance information.

Acknowledgements

This study was funded by Federal Aid in Sport Fish Restoration Grant F-80 (Study 230568) to the Michigan Department of Natural Resources, Fisheries Division. Additional funding was provided through the Game and Fish Protection Fund, Michigan Department of Natural Resources. We thank Jim Breck, Damon Krueger, and two anonymous reviewers for many helpful comments on this manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fishres.2013.08.012.
References


Jackman, S., 2012. pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University. Department of Political Science, Stanford University, Stanford, CA. R package version 0.90 http://pscl.stanford.edu


