Investigating ecosystem-level effects of gillnet bycatch in Lake Erie: implications for commercial fisheries management

Yan Li

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science
In
Fisheries and Wildlife Sciences

Yan Jiao, Chair
Eric M. Hallerman, Member
Feng Guo, Member

July 22nd, 2010
Blacksburg, Virginia

Keywords: Gillnet, Lake Erie, Bycatch, Commercial fisheries, and Model analysis.

Investigating ecosystem-level effects of gillnet bycatch in Lake Erie: implications for commercial fisheries management

Yan Li

ABSTRACT

Lake Erie supports one of the world's largest freshwater commercial fisheries. Bycatch has become a concern in current fisheries management. This study focused on four species in Lake Erie that include two major commercial and recreational species, walleye (*Sander vitreus*) and yellow perch (*Perca flavescens*); an invasive species, white perch (*Morone americana*); and an endangered species, lake sturgeon (*Acipenser fulvescens*). The analyses were based on two datasets, the Partnership Index Fishing Survey (PIS) data and the commercial gillnet logbook data. The bycatch of walleye, yellow perch and white perch was predicted by a delta model developed on the PIS data. Discards were estimated as the difference between predicted bycatch and landed bycatch. Results highlighted bycatch and discard hotspots for these three species that have great management implications. Three classification tree models, a conditional inference tree and two exhaustive search-based trees, were constructed using the PIS data to estimate the probability of obtaining lake sturgeon bycach under specific environmental and gillnet fishing conditions. Lake sturgeon bycatch was most likely to be observed in the west basin of Lake Erie. The AdaBoost algorithm was applied in conjunction with the generalized linear/additive models to analyze catch rates of walleye, yellow perch and white perch. Three- and five-fold cross-validations were conducted to evaluate the performance of each candidate model. Results indicated that the Delta-AdaBoost model yielded the smallest training error and test error on average. I recommend the Delta-AdaBoost model for catch and bycatch analyses when data contain a high percentage of zeros.

ACKNOWLEDGEMENTS

I would like to acknowledge the professors, graduate students, staff, and technicians who helped me in the entire course of my master's research.

I express my deep gratitude to my major advisor Dr. Yan Jiao for providing me this opportunity to build up my career in the field of fishery modeling and management. Dr. Jiao provided great suggestions for my graduate life, study and research. She helped to construct this thesis project, validate results, and revise manuscripts and presentations. Dr. Jiao's expertise in the field of fish population dynamics and stock assessment and her personality as an advisor and a graduate student mentor impressed me deeply. I learned a lot from her, especially on methodology applied in stock assessment. I would not finish this thesis without her help and support.

I am highly indebted to my committee member Dr. Eric Hallerman for his generous help in revising my working plan, improving my technical writing skill, providing me valuable suggestions for this thesis. I also appreciate his valuable teaching of Conservation Genetics in that he extended my knowledge about fish population dynamics to the genetic level and brought me new insights on fisheries management incorporating genetic aspects.

I would like to thank my committee member Dr. Feng Guo for his assistance on my working plan and thesis. He gave me lots of valuable comments on my thesis project on the statistical view.

I extend my thanks to Kevin Reid at the Ontario Commercial Fisheries Association and Bob Sutherland at the Ontario Ministry of Natural Resources for their kind help on synthesizing the data and offering valuable information and comments for this thesis project.

Special acknowledgement goes to my teammates Michael Errigo, Dan Hua, Robert Leaf, Hao Yu, Qing He, and Joshua Hatch, for their generous help during my study and research at Virginia Tech.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1
Introduction

1.1. Biological characteristics of walleye, yellow perch, white perch and lake sturgeon

Lake Erie supports one of the world's largest freshwater commercial fisheries, where the landings are dominated by two highly valued species (Kinnunen 2003), walleye (*Sander vitreus*) and yellow perch (*Perca flavescens*). White perch (*Morone americana*) invaded Lake Erie in the 1950s and imposed considerable impacts on the lake ecosystem because of its interactions (competition and predation) with native species (Scott and Crossman 1973; Schaeffer and Margraf 1987; Parrish and Margraf 1990). Lake sturgeon (*Acipenser fulvescens*) has been listed as an endangered or threatened species under state legislation in 19 of the 20 states within its original range in the United States (GLLSCM 2002; Welsh 2004; GLLSCM 2010). This study focused on these four species with an emphasis on bycatch analysis and management. The major biological characteristics of these four species are summarized in Table 1-1.

Walleye (*Sander vitreus*), with a maximal size of 106.7 cm in length and 11.3 kg in weight, may live for three decades (Robins et al. 1991; FishBase 2009), but few older than 5 or 6 years of age are encountered in heavily fished populations. Walleye caught in Lake Erie average 33.0-63.5 cm in length and 0.5-2.3 kg in weight around age 3 (Scott and Crossman 1973). Adults migrate to tributary streams in late winter or early spring to spawn. Most commercial fisheries for walleye are situated in the Canadian waters of the Great Lakes (Ronald 2003).

Adult yellow perch (*Perca flavescens*) can grow to 50.8 cm in length and 1.9 kg in weight, and live for up to 11 years (Page et al. 1997; FishBase 2009). In Lake Erie, the landings are 17.8-25.4 cm in length and 0.2-0.5 kg in weight on average (Scott and Crossman 1973). Yellow perch reach sexual maturity at 1 to 3 years of age for males and 2 to 3 years of age for females. Spawning occurs between February and July.

White perch (*Morone americana*), native to Atlantic coastal regions, invaded Lake Erie in 1950s (Page et al. 1997; FishBase 2009). The maximum length of white perch is 49.5 cm and the maximum weight is 2.2 kg. The average size of white perch found in Lake Erie is 12.7-17.8 cm in length and 0.2 kg in weight (Scott and Crossman 1973). The males reach maturity at age 2 and the females at age 3 (Bur 1986). White perch is a prolific competitor with native species and

eats the eggs of walleye and white bass. They usually spawn in May through June (Sutton et al. 1996).

Lake sturgeon (*Acipenser fulvescens*), the largest freshwater fish in the Great Lakes, up to 274.3 cm in length and over 90.7 kg in weight, is distributed in North American temperate freshwater (Page et al. 1997). It can live for over 100 years and usually does not reach reproductively maturity until 8-12 years old for males and 25 years old for females (Bruch et al. 2001). The females spawn only once every 4-9 years (Bruch et al. 2001). Its population is highly vulnerable to fishing mortality.

Studies and discussions about interactions between walleye and yellow perch have been ongoing for many years and in many lakes, including Lake Erie. However, recent studies in Lake Erie did not indicate significant influence on walleye by predation from yellow perch (Parrish and Margraf 1990). White perch is considered as a prolific competitor of native fish species in that they prefer to feed on fish eggs, and they have hybridized with native white bass. Concerns have been raised on the decline of abundance of walleye and the introgression of gene pool of white bass caused by white perch invasion.

1.2. Population status of walleye, yellow perch, white perch and lake sturgeon in Lake Erie

Walleye and yellow perch have been the major recreational and commercial species in Lake Erie over the last 50 years. Their landings accounted for over 80% of the total recreational harvest in Michigan waters from 1999 to 2002 (Thomas and Haas 2005), and 44% of the total commercial harvest in Lake Erie in 2000 (Kinnunen 2003). After a productive period from the late 1940s to early 1950s, the population and commercial catch of walleye began to decline until the 1970s (Scott and Crossman 1973). A study in Michigan waters of Lake Erie showed that these two species have both experienced periods of low abundance in Lake Erie during the last 20 years. Walleye abundance from 1999 to 2002 (less than 25 million fish) decreased more than 50% from 1989. Yellow perch abundance was high in the late 1970s and early 1980s, and then declined because of high exploitation (25%-50%) from 1989-1994 (Boileau 1985; Thomas and Haas 2005).

White perch is native to Atlantic coastal regions and invaded Lake Erie in the 1950s (Page et al. 1997; FishBase 2009). Its populations expanded rapidly since the 1970s and have competitive interactions with native speices (Parrish and Margraf 1990).

Lake sturgeon (*Acipenser fulvescens*) was abundant in the Great Lakes during the late 1800s, but its populations have been reduced or extirpated dramatically due to overfishing, pollution, construction of dams and habitat loss (Scott and Crossman 1973; Birstein et al. 1997; Auer 1999; Bogue 2000). Nineteen of the 20 states within its original range in the United States have listed lake sturgeon as an endangered or threatened species (GLLSCM 2002; Welsh 2004; GLLSCM 2010).

1.3. Commercial fisheries of walleye, yellow perch and white perch in Lake Erie

Through the 1990s, there was a prominent harvest of walleye and yellow perch from Lake Erie. For example, in 2000, Lake Erie yielded 96.8% of the total commercial walleye harvest from the Great Lakes, and 84.1% of the total commercial yellow perch harvest (Kinnunen 2003). Of the total commercial harvest of all species from Lake Erie, walleye and yellow perch accounted for 27% and 17% by weight, respectively (Figure 1-1, Kinnunen 2003). A commercial fishery for white perch existed in the early 1990s, with the majority taken out of Canadian waters. In the last half of the 1990s, the white perch commercial fishery declined (Kinnunen 2003).

Since the mid-1970s, walleye and yellow perch have been managed lake-wide under an interagency quota system, with five management units for walleye and four for yellow perch (Kinnunen 2003). The annual harvest quotas for yellow perch and walleye are set largely based on the stock assessments by the Walleye Task Group and the Yellow Perch Task Group. A Coordinated Percid Management Strategy was adopted by the Lake Erie Committee, and greatly reduced quotas were instituted for 2001-2003 (Thomas and Haas 2005). As a result, the exploitation for walleye declined substantially during 2001-2003 (Thomas and Haas 2005).

1.4. Bycatch issues and bycatch management in the commercial gillnet fisheries in Lake Erie

Bycatch, usually defined as the unintentional capture of non-target species of fishes, turtles, birds and invertebrates, has become a well-recognized and unavoidable issue in fisheries management around the world, not only because valuable living resources are wasted, but also because bycatch has substantial ecological impacts (Hall et al. 2000; Hall and Mainprize 2005; Harrington et al. 2005; Kelleher 2005). Although some of the bycatch species may be retained for sale or use, most are discarded back into the water with a low survival rate (Hall et al. 2000; Harrington et al. 2005). Individual quota may be an important reason for discards. A global assessment provides an estimated fish discard of 27 million MT (Alverson et al. 1994). In 2002, the United States had one of the highest discard ratios (22%) in the world — 3.7 million tons of fish were landed and another 1.06 million tons of fish were discarded (Harrington et al. 2005). Among the discards, gillnet fisheries contributed 1.2% by weight (Harrington et al. 2005). In the Great Lakes, bycatch has been considered a problem that delays the population recovery of lake trout and contributes to the population decline of lake sturgeon in the Great Lakes (Johnson et al. 2004).

In general, there are three possible means of by-catch reduction: (1) modifying fishing methodology including gear, timing or location; (2) changing fishing gear or methods entirely; and (3) reducing fishing effort (Harrington et al. 2005). In addition, there are other means of converting discarded bycatch to landed catch, such as developing new markets and processing techniques, and changing regulatory limits and instituting requirements to land all catch (Harrington et al. 2005). Closures of a fishery during the spawning season or in the permanent or seasonal designated refuges, and minimum size restrictions are applied to keep a sustainable commercial fishery in the Great Lakes (Kinnunen 2003).

Gillnets are widely used in the commercial fisheries of walleye, yellow perch and white perch in Lake Erie. Because of the overlaps in biological characteristics (Table 1-1), overlaps in the fishing season, fishing gear (Table 1-2) and fishing location, the discrepancy in the market values (Table 1-2), and interactions among species such as competition and predation relationships (Table 1-1), bycatch and discards of these four species occur frequently in the commercial gillnet fisheries in Lake Erie, but have not been well documented (Scott and Crossman 1973; Hamley 1975; Kinnunen 2003; Johnson et al. 2004; K. Reid, Ontario Commercial Fisheries Association, personal communication). Failure to take account of bycatch

and discards into stock assessments may increase the bias when estimating fishing mortality, population abundance and available quotas, and may conceal the impacts of bycatch and discards on ecosystem stability (Johnson et al. 2004).

1.5. Bycatch and discard assessments

Bycatch data are characterized by a high number of low catches, a few high catches, and a high percentage of zeros (Ortiz et al. 2000; Maunder and Punt 2004). Three types of methods have been developed to assess bycatch. The first type of method, the ratio method, was employed to estimate the sea turtle bycatch from a sea scallop dredge fishery (Cochran 1977) and the monkfish bycatch from a gillnet fishery (Perez and Wahrlich 2005). In the ratio method we assume that the bycatch ratio should be a constant, but this assumption cannot always be met by the bycatch data (Murray 2004). The second type of method is to add a small constant to each zero observation, followed by a generalized linear or additive model analysis (Ortiz et al. 2000; Maunder and Punt 2004; Murray 2004; Shono 2008). However, the estimation results are sensitive to the choice of the constant (Ortiz et al. 2000; Maunder and Punt 2004). The third type of method is to use the delta model and the Tweedie distribution model. In the delta model, the positive values are fitted by a generalized linear or additive model, and the probability of observing zero values in the response variable are fitted by a generalized linear or additive model with an assumption of binomial distribution (Lo et al. 1992; Stefansson 1996; Ye et al. 2001; Maunder and Langley 2004). By contrast, the Tweedie distribution model handles zero values uniformly along with the positive values, where the Tweedie distribution is considered to be a Poisson-Gamma compound distribution when its power parameter is between 1 and 2 (Tweedie 1984; Shono 2008). The delta model has been applied to estimate the abundance of highly aggregated organisms, rare species and bycatch species (Pennington 1983; Lo et al. 1992; Stefansson 1996; Ortiz et al. 2000; Maunder and Punt 2004; Murray 2004). The Tweedie distribution model has been judged to outperform the generalized linear model with an additive constant and the delta model composed of two generalized linear models (Shono 2008).

The AdaBoost model can be used as an alternative to the generalized linear/additive model with an assumption of binomial distribution in a delta model to estimate the probability of obtaining non-zero captures. This model was originally used for classification problems. The

algorithm used for classification is called the classifier. The final strong classifier is obtained by successively applying a classification algorithm to reweighted data and then combining a sequence of weak classifiers which minimize the prediction error at each iteration (Freund and Schapire 1996; Friedman et al. 2000; Hastie et al. 2001; Kawakita et al. 2005). In a fishery context, zeros and positive captures can be converted into a categorical variable $\{-1,1\}$, indicating the events of no fish caught and the events of at least one fish caught, respectively, and then can be treated as a two-group classification problem (Kawakita et al. 2005). This method has been used to predict the occurrence of large silky shark bycatch in a tuna purse-seine fishery, and the results confirmed the superiority of AdaBoost model in bycatch analyses where data were skewed by zeros (Kawakita et al. 2005).

1.6. Objectives

This study was conducted based on two datasets. One dataset is from a fishery-independent survey, the Lake Erie Partnership Index Fishing Survey (PIS), which was primarily operated by the Ontario Ministry of Natural Resources (OMNR) and the Ontario Commercial Fisheries Association (OCFA) since 1989. Experimental gillnets with mesh size ranging from 32 to 152 mm were deployed across the Ontario waters of Lake Erie in the fall (August-November) annually, using commercial fishing vessels and commercial fishing crews (OCFA 2007). The fish captures, environmental factors and fishing effort were recorded in the PIS data. The PIS data contained a high percentage of zero observations. The other dataset is the commercial gillnet data. The commercial data recorded the landed bycatch and catch, environmental factors and fishing effort, but the information about bycatch and discards that actually occurred at lake was missing during my study period (1994-2007).

The objectives of this study were: (1) to select the models that can better deal with fishery data analyses having a high percentage of zero observations; (2) to develop applicable models to assess the bycatch and discards of the four key species (walleye, yellow perch, white perch and lake sturgeon) from the gillnet fisheries in Lake Erie, incorporating environmental and fishing factors; (3) to examine the influence of the environmental factors and the gillnet fisheries on the bycatch and discards of the four species, and (4) to generate applicable recommendations for bycatch management related to these four species.

Specifically, in Chapter 2, I develop a delta model composed by one generalized additive model and one AdaBoost model based on the PIS data and applied this delta model to the commercial data to predict the predicted bycatch and discards of three key species, walleye, yellow perch and white perch, from the commercial gillnet fisheries in Lake Erie in the fall (August to November) of 1994-2007. In Chapter 3, I construct a conditional inference classification tree and two exhaustive search based classification trees to estimate the probability of obtaining lake sturgeon bycatch under specific environmental and gillnet fishing conditions in Lake Erie based on the PIS data from 1989 to 2008. In Chapter 4, I apply the AdaBoost algorithm to the delta model to deal with data analyses having high percentage of zero observations, and compared this method with five candidate models based on the PIS data. Implications for fisheries management for the key commercial and recreational species (yellow perch and walleye), invasive species (white perch) and endangered species (lake sturgeon) are discussed in Chapters 2 and 3.

1.7. References

Alverson, D., M. Freeberg, J. Pope and S. Murawski. 1994. A global assessment of fisheries bycatch and discards. Food and Agriculture Organization of the United Nations, Rome.

Auer, N. A. 1999. Chapter 17, Lake sturgeon: a unique and imperiled species in the Great Lakes. In: Taylor, W. and C.P. Ferreri (eds), Great Lakes fisheries policy and management: a binational perspective. Michigan State University Press, East Lansing, MI.

Birstein, V., W. Bemis and J. Waldman. 1997. The threatened status of acipenseriform species: a summary. In: Birstein, V., J. R. Waldman and W. E. Bemis (eds), Sturgeon Biodiversity and Conservation. Springer, NY.

Bogue, M. 2000. Fishing the Great Lakes: an environmental history, 1783-1933. Univ of Wisconsin Press, Madison.

Boileau, M. 1985. The expansion of white perch, *Morone americana*, in the lower Great Lakes. Fisheries 10:6-10.

Bruch, R., T. Dick and A. Choudhury. 2001. A field guide for the identification of stages of gonad development in lake sturgeon, *Acipenser fulvescens*, with notes on lake sturgeon reproductive biology and management implications. Graphic Communications Center, Appleton, Wisconsin.

Bur, M. 1986. Maturity and fecundity of the white perch, *Morone americana*, in western Lake Erie. The Ohio Journal of Science 86:205-207.

Cochran, W. 1977. Sampling techniques. John Wiley and Sons, New York, NY.

FishBase. 2009. A global information system on fishes [online]. Available: www.fishbase.org. (March 2009).

Freund, Y. and R. Schapire. 1996. Experiments with a new boosting algorithm. In: Saitta, L. (Ed), Machine Learning: Proceedings of the Thirteenth International Conference, Bari, Italy.

Friedman, J., T. Hastie and R. Tibshirani. 2000. Additive logistic regression: A statistical view of boosting. The annals of statistics 28:337-374.

Great Lakes Lake Sturgeon Coordination Meeting (GLLSCM), 2002. Basin overview presentations of status and assessment activities. Presentations by N. Auer, H. Quinlan, M. Holtgren, R. Elliott, M. Thomas, E. Zollweg, A. Mathers, D. Carlson, Sault Ste, Marie, MI.

Great Lakes Lake Sturgeon Coordination Meeting (GLLSCM). 2010. Lake Sturgeon biology and population history in the Great Lakes. Available: http://www.fws.gov/midwest/sturgeon/biology.htm. (May 2010).

Hall, M., D. Alverson and K. Metuzals. 2000. By-catch: problems and solutions. Marine Pollution Bulletin 41:204-219.

Hall, S. and B. Mainprize. 2005. Managing by-catch and discards: how much progress are we making and how can we do better? Fish and Fisheries 6:134-155.

Hamley, J. 1975. Review of gillnet selectivity. Journal of the Fisheries Research Board of Canada 32:1944-1969.

Harrington, J., R. Myers and A. Rosenberg. 2005. Wasted fishery resources: discarded by-catch in the USA. Fish and Fisheries 6:350-361.

Hastie, T., R. Tibshirani and J. Friedman. 2001. The elements of statistical learning: data mining, inference and prediction, 2 nd edition. Springer, New York.

Johnson, J. E., J. L. Jonas and J. W. Peck. 2004. Management of Commercial Fisheries Bycatch, with Emphasis on Lake Trout Fisheries of the Upper Great Lakes. Fisheries Research Report, Michigan Department of Natural Resources, Lansing, Michigan.

Kawakita, M., M. Minami, S. Eguchi and C. Lennert-Cody. 2005. An introduction to the predictive technique AdaBoost with a comparison to generalized additive models. Fisheries Research 76:328-343.

Kelleher, K. 2005. Discards in the world's marine fisheries: an update. Food and Agriculture Organization of the United Nations, Rome.

Kinnunen, R. 2003. Great lakes commercial fisheries. Report from Michigan Sea Grant, Michigan Sea Grant Extension, East Lansing, Michigan.

Lo, N., L. Jacobson and J. Squire. 1992. Indices of relative abundance from fish spotter data based on delta-lognormal models. Canadian Journal of Fisheries and Aquatic Sciences 49:2515-2526.

Maunder, M. and A. Langley. 2004. Integrating the standardization of catch-per-unit-of-effort into stock assessment models: testing a population dynamics model and using multiple data types. Fisheries Research 70:389-395.

Maunder, M. and A. Punt. 2004. Standardizing catch and effort data: a review of recent approaches. Fisheries Research 70:141-159.

Murray, K. 2004. Magnitude and distribution of sea turtle bycatch in the sea scallop (*Placopecten magellanicus*) dredge fishery in two areas of the Northwestern Atlantic Ocean, 2001-2002. Fishery Bulletin 102:671-681.

OCFA (Ontario Commercial Fisheries Association). 2007. Lake Erie Partnership Index Fishing Survey [online]. Available: http: //www.ocfa.on.ca/Lake Erie Partnership Index Fishing Survey.htm. (September 2009).

Ortiz, M., C. Legault and N. Ehrhardt. 2000. An alternative method for estimating bycatch from the US shrimp trawl fishery in the Gulf of Mexico, 1972-1995. Fishery Bulletin 98:583-599.

Page, L., B. Burr and R. Peterson. 1997. A field guide to freshwater fishes: North America north of Mexico. Houghton Mifflin Harcourt, Boston.

Parrish, D. and F. Margraf. 1990. Interactions between white perch (*Morone americana*) and yellow perch (*Perca flavescens*) in Lake Erie as determined from feeding and growth. Canadian Journal of Fisheries and Aquatic Sciences 47:1779-1787.

Pennington, M. 1983. Efficient estimators of abundance, for fish and plankton surveys. Biometrics 39:281-286.

Perez, J. and R. Wahrlich. 2005. A bycatch assessment of the gillnet monkfish *Lophius gastrophysus* fishery off southern Brazil. Fisheries Research 72:81-95.

Robins, C., R. Bailey, C. Bond, J. Brooker, E. Lachner, R. Lea and W. Scott. 1991. Common and scientific names of fishes from the United States and Canada. American Fisheries Society Special Publication 20, Bethesda, MD.

Schaeffer, J. and F. Margraf. 1987. Predation on fish eggs by white perch, *Morone americana*, in western Lake Erie. Environmental Biology of Fishes 18:77-80.

Scott, W. B. and E. J. Crossman. 1973. Freshwater fishes of Canada. Fisheries Research Board of Canada Bulletin 184, Ontario.

Shono, H. 2008. Application of the Tweedie distribution to zero-catch data in CPUE analysis. Fisheries Research 93:154-162.

Stefansson, G. 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES Journal of Marine Science 53:577.

Sutton, C. C., J. C. O'Herron and R. T. Zappalorti. 1996. The scientific characterization of the Delaware Estuary. The Delaware Estuary Program (DRBC Project No. 321), Trenton, New Jersey.

Thomas, M. and R. Haas. 2005. Status of yellow perch and walleye populations in Michigan waters of Lake Erie, 1999–2003. Michigan Department of Natural Resources. Fisheries Research Report 2082:

Welsh, A. 2004. Factors influencing the effectiveness of local versus national protection of migratory species: a case study of lake sturgeon in the Great Lakes, North America. Environmental Science and Policy 7:315-328.

Ye, Y., M. Al-Husaini and A. Al-Baz. 2001. Use of generalized linear models to analyze catch rates having zero values: the Kuwait driftnet fishery. Fisheries Research 53:151-168.

TABLE 1-1.—Biological characteristics of walleye, yellow perch, white perch and lake sturgeon (adapted from Scott and Crossman 1973; Page et al. 1997; Bruch et al. 2001).

Species	Walleye	Yellow perch	White perch	Lake sturgeon
Scientific name	*Sander vitreus*	*Perca flavescens*	*Morone americana*	*Acipenser fulvescens*
Family	Percidae	Percidae	Moronidae	Acipenseridae
Max. length (cm)	106.68	50.8	49.53	274.32
Max. weight (kg)	11.34	1.91	2.22	125.01
Ave. length harvested in Lake Erie (cm)	33.02-63.5	17.78-25.4	12.7-17.78	91.44-149.86 juvenile 25.4-50.8
Ave. weight harvested in Lake Erie (kg)	0.45-2.27	0.17-0.45	0.23	4.54-36.29
Longevity/age of majority harvested in Lake Erie (yr)	29/3-4	11/4	16/2-3	152/-
Maturity (age)	male 2; female 3	male 2; female 2-3	male 1-2; female 2-3	male 8-12; female 25
Spawning	Mar-May; migrate to shallow and warmer waters	Feb-July	May-June; prolific	May-June; once every 4-9 years
Habitat	demersal; 29°C	benthopelagic; 0-30°C	demersal; temperate	demersal
Diet	shad, emerald shiner, alewife, round goby,	round goby, invertebrates	fish eggs (such as the eggs of	benthic organisms

	seasonally yellow perch and white perch	walleye, white bass), minnows

TABLE 1-2.—Commercial fisheries of walleye, yellow perch and white perch in the Great Lakes in 2000 (adapted from Kinnunen 2003).

Species	Walleye	Yellow perch	White perch
Fishing gear	gillnet, trap-net	gillnet, trap-net	gillnet, trap-net
Fishing season	spring, fall	spring, fall	all year, especially spring
Market price ($/kg)	3.06	4.65	1.15
% of total harvest	27	17	2

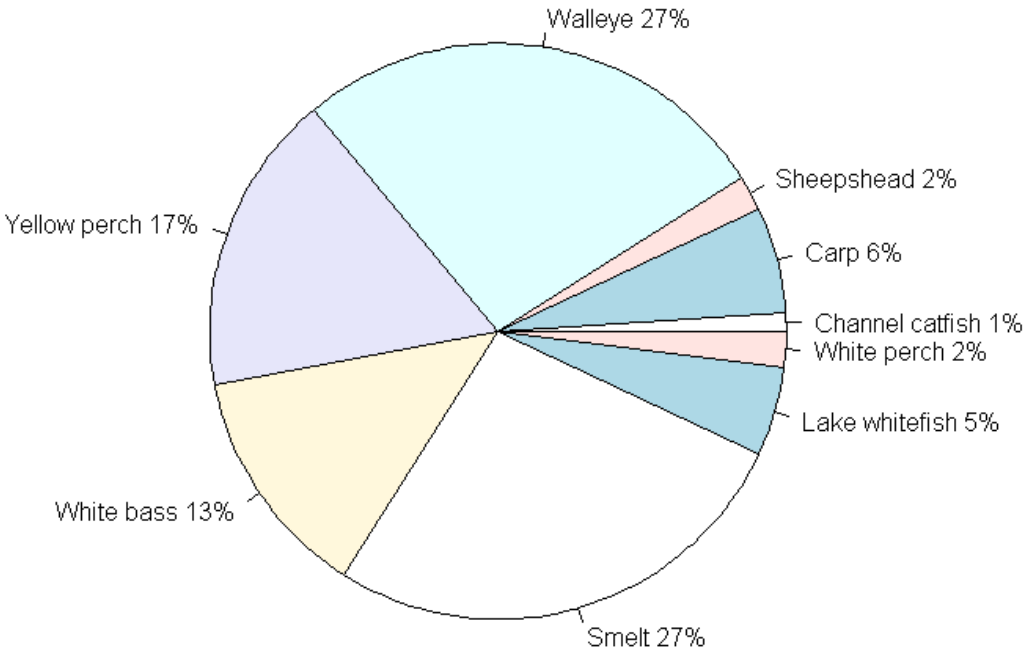

FIGURE 1-1.— The commercial harvest (by weight) by target species from Lake Erie in 2000 (adapted from Kinnunen 2003).

Chapter 2

Gillnet bycatch and discard assessments of walleye, yellow perch and white perch from Lake Erie commercial fisheries

2.1. Abstract

Bycatch (kg) of walleye (*Sander vitreus*), yellow perch (*Perca flavescens*), and white perch (*Morone americana*) from commercial gillnet fisheries in Lake Erie in fall (August to November) of 1994-2007 was predicted by a delta model developed on fishery-independent survey data (Lake Erie Partnership Index Fishing Survey, PIS). The delta model consisted of one generalized additive model and one AdaBoost model. The generalized additive model was used to predict positive captures of the bycatch species, and the AdaBoost model was used to predict the probability of non-zero captures. Discards (kg) for each species were estimated as the difference between the bycatch that was predicted from the delta model and the landed bycatch that was recorded in the commercial data. The predicted bycatch and the percentage of discards (the percentage of total estimated discards among the total landings of the species of interest; the total landings were a summation of total predicted bycatch and total landed catch of the species of interest) by major target species, month, basin and year were analyzed. The results indicated that more bycatch was obtained in the west basin in October for walleye, in the west central basin in November for yellow perch, and in the west central basin in October for white perch. More discards occurred in the west basin of Lake Erie during August to September for walleye, in the waters across the west central and east central basins in November for yellow perch, and in the west basin in August and November for white perch. Possible bycatch management strategies include restrictions on fishing season and hotspot areas for bycatch and discards, and a joint fishing license framework for target and bycatch species. An advanced fishery data recording system and observer program that can incorporate both bycatch and discard information into the commercial fisheries reports would be useful for further bycatch studies and fisheries management.

2.2. Introduction

Bycatch, particularly discarded bycatch, is of concern to conservation biology and fisheries management because of wasted living resources, threats to imperiled species, and impacts on ecosystem stability (Alverson et al. 1994; Crowder and Murawski 1998; Hall et al. 2000; Johnson et al. 2004; Harrington et al. 2005; Kelleher 2005). The major reasons for bycatch can be related to size limit, quota limit and fish market (Murawski 1996; Punt et al. 2006). It is essential to characterize and assess bycatch, especially discarded bycatch, which may have important implications for fisheries management. However, bycatch is not always measured or reported in many fisheries (Johnson et al. 2004; Borges et al. 2005; Punt et al. 2006). For example, in the Lake Erie commercial gillnet fisheries in this study, the bycatch and discards of walleye (*Sander vitreus*), yellow perch (*Perca flavescens*) and white perch (*Morone americana*) occurred, but the discards at lake were not recorded.

Walleye and yellow perch dominate the commercial gillnet fisheries in Lake Erie (Kinnunen 2003; Thomas and Haas 2005). White perch, as an invasive species, has imposed considerable impacts on the native fish communities and the lake ecosystem (Scott and Crossman 1973; Schaeffer and Margraf 1987; Parrish and Margraf 1990). Because of the overlap in biological characteristics between the target species and these three species, the overlap in fishing effort, low species selectivity and high mortality of gillnets, and the regulatory system for quota species, bycatch and discards of walleye, yellow perch and white perch occur frequently in the commercial gillnet fisheries in Lake Erie, but have not been well documented (Scott and Crossman 1973; Hamley 1975; Kinnunen 2003; Johnson et al. 2004; K. Reid, Ontario Commercial Fisheries Association, personal communication). Failure to take bycatch and discards into account in stock assessments may increase the bias when estimating fishing mortality, population abundance, and available quotas, and may conceal the impacts of bycatch and discards on ecosystem stability (Johnson et al. 2004).

In this study, the bycatch and discards of these three species were assessed based on both the commercial fishery data and the fishery-independent survey data. Since the commercial data recorded only the landed bycatch, and the information about bycatch and discards that actually occurred at lake has been missing during our study period (1994-1007), we developed a delta model using the fishery-independent survey data to predict bycatch from commercial fisheries. The survey program, the Lake Erie Partnership Index Fishing Survey

(PIS), was conducted by the Ontario Ministry of Natural Resources (OMNR) and the Ontario Commercial Fisheries Association (OCFA) since 1989. Experimental gillnets with fourteen mesh sizes, ranging from 32 to 152 mm, were set at sites distributed across the Ontario waters of Lake Erie in the fall (August to November) annually by commercial fishing vessels and commercial fishing crews (OCFA 2007).

The PIS data contained a high frequency of zero captures (75-77%). The presence of zeros may invalidate the assumptions of normality commonly used in fishery data analyses, and may cause computational difficulties. Ignorance of a considerable proportion of zero observations may result in a loss of information that may reflect the spatial or temporal distribution characteristics of the fish stock. The approaches to deal with zero values in previous fishery data analyses can be categorized into two types. One approach is to add a small constant to each observation of the response variable, followed by generalized linear or additive model analysis (Ortiz et al. 2000; Maunder and Punt 2004; Shono 2008). However, the estimation results are sensitive to the choice of this constant (Ortiz et al. 2000; Maunder and Punt 2004). The other approach to deal with zeros is to apply either the delta model or the Tweedie distribution model. In the delta model, positive values and zeros can be modeled by two sub-models separately (Lo et al. 1992; Stefansson 1996; Ye et al. 2001; Maunder and Langley 2004). The application of the delta model in fishery data analyses with zeros has been well documented (Lo et al. 1992; Stefansson 1996; Ortiz et al. 2000; Ye et al. 2001; Murray 2004). By contrast, the Tweedie distribution model can handle zero values uniformly along with the positive values; the Tweedie distribution is considered to be a Poisson-Gamma compound distribution when its power parameter is greater than 1 and less than 2 (Tweedie 1984; Shono 2008). The Tweedie distribution model was recommended in bycatch analysis when a high percentage of zero observations is involved (Shono 2008).

AdaBoost was originally used for classification problems. If the values of the response variable take either 1 or -1, the algorithm used to produce the value 1 or -1 is called a classifier. The final classifier is obtained by linearly combining a sequence of classifiers fit by the reweighted data at each iteration (Freund and Schapire 1996; Friedman et al. 2000; Hastie et al. 2001; Kawakita et al. 2005). In the fishery context, if the values 1 and -1 represent the presence and absence of the fish species in a fishing event, respectively, the problem can be treated as a

two-group classification problem. The AdaBoost model has been employed to predict the occurrence of large silky shark bycatch in a tuna purse-seine fishery, and yielded more accurate and stable predictions compared with the generalized additive models (Kawakita et al. 2005). We applied a AdaBoost algorithm to the delta model to analyze the PIS data that contained a high percentage of zeros. Specifically, we applied a real AdaBoost model to predict the probability of obtaining non-zero captures in the delta model instead of a generalized linear/additive model with an assumption of binomial distribution.

In the following analysis, bycatch referred to all the unintentional captures in the fisheries, including both landed and discarded bycatch (Crowder and Murawski 1998; Harrington et al. 2005). In this study, we aimed to: (1) develop a delta model that was composed by one generalized additive model and one AdaBoost model using the PIS data; (2) to predict the bycatch of these three species, walleye, yellow perch and white perch, from the commercial gillnet fisheries in Lake Erie in the fall (August to November) of 1994-2007, by applying the developed delta model to the commercial data; (3) to estimate the discards at lake for each species by comparing the bycatch that was predicted from the delta model and the landed bycatch that was recorded in the commercial data; (4) to analyze the predicted bycatch and the percentage of discards (the percentage of total estimated discards among the total landings of the species of interest) by major target species, month, basin and year; and (5) to suggest possible bycatch management strategies for these three species in the commercial gillnet fisheries in Lake Erie.

2.3. Methods

Data.—I conducted the gillnet bycatch and discard assessments for walleye, yellow perch, and white perch using two datasets: the PIS data from 1989 to 2008 and the commercial data from 1994 to 2007. Both datasets were provided by OCFA and contained information on fish captures, fishing effort and environmental factors. I developed the delta model using the PIS data, and applied the delta model to the commercial data to predict bycatch and to estimate the discards at lake from the commercial fisheries.

In the PIS data, 53,662 records were available for the model development. In the commercial data, 32,349 records were analyzed for walleye, 32,282 records for yellow perch

and 32,392 records for white perch. Each record in the PIS data and each record in the commercial fishery data from 1994-2001 referred to one net; each record in the commercial fishery data from 2002 to 2007 was based on the daily catch report that included 1-5 nets. These analyzed records included the nets that had a mesh size ranging from 51 to 140 mm, were distributed in the waters of 3 to 66 meters deep, and were soaked for 9 to 36 hours in the fall (August to November) from 1994 to 2007. Since I applied the delta model to the commercial data, I used the explanatory variables that were contained in both the PIS and commercial fishery data to construct the delta model. These explanatory variables included two continuous variables (site depth and soak time) and five categorical variables (basin, year, month, gear type, mesh size). The gear type referred to the canned or bottomed nets.

Delta model based on the PIS data and bycatch prediction.—Since the PIS data contained a high percentage of zero captures where the commonly used assumption of normality was violated, a delta model was developed to predict the bycatch of these three species from the commercial fisheries (Lo et al. 1992; Ortiz et al. 2000; Maunder and Punt 2004; Kawakita et al. 2005; Shono 2008). After log-transformation with a small additive constant of 0.001, the histograms showed that zeros were separated from positive values, which were assumed to follow a normal distribution (Figure 2-1, Stefansson 1996; Ye et al. 2001; Murray 2004; Damalas et al. 2007).

In the variable selection procedure, we examined the correlation coefficients among all seven explanatory variables to detect highly correlated variables. A preliminary stepwise selection based on Akaike Information Criterion (AIC, Akaike 1974) was conducted to eliminate one of the highly correlated pair of explanatory variables. The variable that yielded a larger AIC value was eliminated from the correlated pair detected in the correlation analysis. The remaining variables were selected through a stepwise procedure based on AIC (Akaike 1974; Burnham and Anderson 2002). The model with smaller AIC value was considered to fit the data better. Interaction terms were not included in the model because of the insignificant effects of interaction terms and the difficulties in model interpretation (Maunder and Punt 2004; Damalas et al. 2007).

The delta model to predict bycatch consisted of two components, the generalized additive model (GAM) to predict positive captures and the AdaBoost model to predict the probability of obtaining non-zero captures. The predicted bycatch from the delta model can be obtained by multiplying these two components (Lo et al. 1992; Pennington 1996; Stefansson 1996; Ortiz et al. 2000; Ye et al. 2001; Maunder and Punt 2004; Murray 2004):

$$\hat{Theoretical\ bycatch} = \hat{d} \times \hat{p},$$

where $\hat{Theoretical\ bycatch}$ is the predicted bycatch from the commercial gillnet fisheries, $\hat{d}$ is the predicted positive captures of the bycatch species, and $\hat{p}$ is the predicted probability of catching at least one fish of the bycatch species.

A generalized additive model (GAM) was constructed to predict the positive captures, $\hat{d}$. In the GAM, the effect of each variable can be modeled by a smooth function (Hastie and Tibshirani 1990). By assuming that the positive captures followed a log-normal distribution, the GAM can be written as:

$$\ln(\hat{d}) = \beta_0 + \sum_{j=1} f_j(X_j),$$

where $\hat{d}$ is the predicted positive captures of the bycatch species, β_0 is the intercept, f_j is a smooth function (a spline or a LOESS smoother) for the *j*th explanatory variable X_j.

To predict the probability of non-zero captures of the bycatch species, the AdaBoost model was applied as an alternative to the generalized linear/additive model assuming a binomial distribution. We denoted the vector of explanatory variables as X, and the response variable as $Y \in \{-1,1\}$. The value -1 represented the event of catching no fish and the value 1 represented the event of catching at least one fish. Then the problem was treated as a two-group classification problem (Kawakita et al. 2005). In the real AdaBoost algorithm, the classifier $g_t(x)$ returns a probability estimate at each iteration. The final classifier $F(x)$ was constructed as follows (Freund and Schapire 1996; Friedman et al. 2000; Hastie et al. 2001; Kawakita et al. 2005):

(1) Initialize the weights $w_i = 1/N$, $i = 1, 2, ..., N$,where *N* is the number of observations.

(2) For *t = 1* to *T*, where *T* is the number of iterations:

(a) Fit the classifier $g_t(x)$ using the data weighted by w_i and obtain a probability estimate $g_t(x_i) = \Pr(\hat{y}_i = 1 \mid x_i)$, i.e., the probability that the predicted value for y_i equals 1 given x_i.

(b) Set $h_t(x_i) = \frac{1}{2}\ln(g_t(x_i)/(1-g_t(x_i)))$, which indicates the contribution of the classifier $g_t(x)$ to the final classifier $F(x)$.

(c) Update the weights for the next iteration, $w_i = \frac{\exp(-y_i h_t(x_i))}{\sum_{i=1}^{N}\exp(-y_i h_t(x_i))}$,.

(3) Set $H(x_i) = \sum_{t=1}^{T} h_t(x_i)$. The final classifier for the *i*th observation,

$$F(x_i) = \begin{cases} 1, \textit{ if } H(x_i) > 0; \\ -1, \textit{ if } H(x_i) < 0. \end{cases}$$

(4) The probability of obtaining non-zero captures for the *i*th observation,

$$\hat{p}_i = \frac{e^{2H(x_i)}}{1+e^{2H(x_i)}}.$$

At iteration *t*, those observations that were misclassified at the previous iteration had their weights increased, whereas the weights were decreased for those classified correctly. As iterations proceeded, each classifier was forced to focus on those observations that were difficult to classify correctly. As a result of combining these classifiers, the final classifier provided accurate estimates, either the presence/absence of the fish species or the probability of obtaining non-zero captures.

Score plots generated from the AdaBoost model were utilized to detect the factors that had important influences on the response variable, i.e., on the probability of catching this species in our case (Kawakita et al. 2005). A higher score of an explanatory variable indicated that this variable would have higher influence on the probability of catching this species.

It is necessary to determine the optimal number of iterations, *T*, by balancing the training error and test error, because as iteration increases, training error will decrease monotonically whereas the test error may not (Kawakita et al. 2005). We split the PIS data into two sub-datasets with roughly equal size when determining *T*. We used one sub-dataset for model building, which was called the training data. The corresponding error from model

building was called the training error. The other sub-dataset was used for prediction, which was called the test data. When we applied the model built from the training data to the test data, the corresponding error from the test data was called test error. After examining the trends of training and test error in the AdaBoost model up to 1000 interactions, we determined T=200 as the optimal number of iterations in this study because both training and test error decreased dramatically and started to stabilize when the iteration proceeded to around 200.

To confirm the superiority of the delta model composed by a generalized additive model and an AdaBoost model (Delta-AdaBoost), five candidate models were constructed for comparison. The model comparison was conducted through the 5-fold cross-validation approach (Breiman et al. 1984; Tweedie 1984; Hastie et al. 2001; Damalas et al. 2007). The five candidate models included: (1) a delta model consisting of two generalized linear models (Delta-GLM), (2) a delta model consisting of two generalized linear models with polynomial terms up to degree 3 (Delta-GLM-Poly), (3) a delta model consisting of two generalized additive models (Delta-GAM), (4) a generalized linear model with Tweedie distribution (GLM-Tweedie), and (5) a generalized additive model with Tweedie distribution (GAM-Tweedie, Tweedie 1984; Shono 2008). In the Delta-GLM, the Delta-GLM-Poly, and the Delta-GAM models, the model to predict positive captures was assumed to follow a log-normal distribution, and the model to predict probability of non-zero captures was assumed to follow a binomial distribution.

To conduct the 5-fold cross-validation, the PIS data were split into five sub-datasets with roughly equal size. Each sub-dataset was used as test data for prediction, and the remaining four sub-datasets were combined as training data to build the model. The Delta-AdaBoost model and each candidate model were fit using the training data, and applied to the test data. In all the five candidate models, the explanatory variables were selected through the same procedure as the Delta-AdaBoost model. The training error and test error for each model from each pair of training and test datasets were calculated as follows (Hastie et al. 2001; Damalas et al. 2007):

$$Training(Test) \quad error = \frac{1}{N}\sum_{i=1}^{N} | y_i - \hat{y}_i |,$$

where N is the number of observations, y_i is the ith observation, and $\hat{y}_i$ is the estimated value (the predicted value) from the model for the ith observation. The model providing lower training error and test error was judged as the one with better performance (Hastie et al. 2001; Damalas et al. 2007; Shono 2008).

Discard estimation and bootstrap method.— Since the commercial data recorded only the landed catch and landed bycatch, the delta model developed on the PIS data was applied to the commercial fishery data estimate bycatch from the commercial gillnet fisheries. The discards (kg) at lake for each species were estimated as the difference between the bycatch that was predicted from the delta model and the landed bycatch that was recorded in the commercial data. For each species, the results of predicted bycatch were presented as the total predicted bycatch across all the analyzed records in the commercial fishery data, and the percentage composition of the total predicted bycatch by major target species, month, basin and year; the results of discards were presented as percentage of discards (the percentage of total estimated discards among total landings of the species of interest) across all the analyzed records in the commercial data, and the percentage of discards by major target species, month, basin and year. Total landings of the species of interest were calculated by summing up the total predicted bycatch of this species and the total landed catch of this species.

The uncertainties of the bycatch and the discards were analyzed through a nonparametric bootstrap approach. In the bootstrap approach, the data were re-generated by adding the fitted catch from the survey with the re-sampled residuals and then fitted by the delta model 1000 times. A joint distribution of the parameters in the delta model was obtained through the bootstrap approach. A 95% confidence interval was obtained for predicted bycatch and estimated discards based on the joint distribution of parameters. This analysis was programmed in R (Version 2.9.2).

2.4. Results

Correlation analysis did not detect any high correlation among the seven explanatory variables in the PIS data for walleye, yellow perch and white perch (Table 2-1). A stepwise selection was applied to these seven variables to construct the GAM and the AdaBoost model

for each bycatch species. The explanatory variables with significant effect (*P-value* <0.01) or considerable decreasing AIC values were successively selected into the model (Table 2-2). In the prediction of positive captures, the GAMs explained 20%, 31% and 22% of the deviance of PIS data for walleye, yellow perch and white perch, respectively. Results from the GAM (Table 2-2) indicated five factors that had significant impacts on the magnitude of the catch of these three species in Lake Erie, i.e., mesh size, year, basin, gear type and month. Soak time was identified to be a significant factor that influenced the magnitude of the catch of walleye and yellow perch; site depth was identified as an significant factor that influenced the magnitude of the catch of white perch. Score plots (Figure 2-2) from the AdaBoost model indicated gear type had the most important impact on the probability of catching yellow perch and white perch. The probability of catching walleye was mostly affected by basin.

Model comparison among the Delta-AdaBoost model and the five candidate models through 5-fold cross-validation showed that the Delta-AdaBoost model yielded the smallest training and test errors (Table 2-3). This result provided evidence that the delta model (Delta-AdaBoost) fit the PIS data better and produced more accurate estimation and prediction compared with these five candidate models.

Yellow perch, white bass (*Morone chrysops*), lake whitefish (*Coregonus clupeaformis*) and white perch were identified to be the major target species involving in walleye bycatch and discards (Figure 2-3a). Yellow perch bycatch and discards occurred frequently when targeting white perch, walleye, white bass and lake whitefish (Figure 2-4a). The major target species associated with white perch bycatch and discards was yellow perch (Figure 2-5a).

From the gillnet commercial fisheries in Lake Erie during the fall (August to November) of 1994-2007, with 32,349 records combined, the total predicted bycatch of walleye was predicted to be 1.29×10^6 kg on average with a 95% confidence interval (CI) of 1.14-1.46 $\times10^6$ kg (Table 2-4). Among the total predicted bycatch of walleye, 81% on average was observed when targeting yellow perch (Figure 2-3a), 33% on average when fishing in October (Figure 2-3b), 59% on average in the west basin (Figure 2-3c), and 74% on average in the years before 2000 (Figure 2-3d). Total bycatch of yellow perch across 32,282 records was predicted to be 0.042×10^6 kg on average with a 95% CI of 0.038-0.047 $\times10^6$ kg (Table 2-4). The yellow perch bycatch occurred most often when targeting white perch (72% on average,

Figure 2-4a), when fishing in November (40% on average, Figure 2-4b) or in the west central basin (63% on average, Figure 2-4c), and during years of 2001-2005 (71% on average, Figure 2-4d). A total of 1.31×10^6 kg white perch bycatch (95% CI= 0.92-1.74 $\times10^6$ kg) was predicted across 32,392 records (Table 2-4), among which 99% on average was obtained when targeting yellow perch (Figure 2-5a), 27% on average when fishing in October (Figure 2-5b), 53% on average when fishing in the west central basin (Figure 2-5c), and 48% on average in the years of 1994 and 2005-2006. (Figure 2-5d).

Total discards of walleye across 32,349 records from the commercial gillnet fisheries in Lake Erie during the fall of 1994-2007 were estimated as 0.75×10^6 kg on average with a 95%CI of $0.60\text{-}0.92\times10^6$ kg, which accounted for 10.9% (95%CI=8.9-13.0%) of the total walleye landings (Table 2-4). Total estimated discards of yellow perch across 32,282 records were 0.042×10^6 kg on average with a 95%CI of $0.038\text{-}0.047\times10^6$ kg, which accounted for 0.8% (95%CI=0.7-0.9%) of total yellow perch landings (Table 2-4). A total of 0.71×10^6 kg white perch was estimated to be discarded on average with a 95%CI of $0.31\text{-}1.14\times10^6$ kg across 32,392 records, which contributed 44.6% (95%CI=27.2-57.5%) of total white perch landings (Table 2-4). In terms of total discards across all analyzed records, white perch had the highest percentage of discards, followed by walleye and yellow perch.

The highest discard percentage of walleye occurred when targeting yellow perch (70% on average, Figure 2-6a), when fishing in the west basin (19% on average, Figure 2-6c), and in the years 1994 and 2003 (26% and 22% on average respectively, Figure 2-6d). The months when higher discard percentage of walleye was observed were August-September, but the difference among months was not significant (12% on average, Figure 2-6b). The higher percentage of discarded yellow perch was observed when targeting white perch and white bass (88% and 90% on average respectively, Figure 2-7a), when fishing in November (3% on average, Figure 2-7b) or in the waters across the west central and east central basins (0.9% on average, Figure 2-7c), and in the year 2001 (11% on average, Figure 2-7d). Among the total white perch landings when targeting yellow perch, 56% white perch was discarded on average (Figure 2-8a). An average of 55% and 57% of the total white perch landings was discarded when fishing in August and November, respectively (Figure 2-8b), an average of 64% when

fishing in the west basin (Figure 2-8c), and an average ranging from 33% to 83% during the years 1995-2006 (Figure 2-8d).

2.5. Discussion

The delta model was applied in this study to predict bycatch from Lake Erie commercial gillnet fisheries because the PIS data that were used for model development contained a high percentage of zeros (Figure 2-1). In the delta model that we developed, a generalized additive model was selected to capture the most likely nonlinear relationship between fish captures and environmental or operational factors (Bigelow et al. 1999; Damalas et al. 2007). In the prediction of the probability of catching the bycatch species, the AdaBoost model was applied as an alternative to the generalized linear/additive model assuming a binomial distribution (Kawakita et al. 2005). This is the first time that a delta-AdaBoost model was applied in fisheries. Results showed that a combination of a generalized additive model with an AdaBoost model increased model goodness-of-fit. Model comparison is often a problem when models are less comparable using information-based criteria such as AIC (Shono 2008). Here, *n*-fold cross-validation was suggested as a tool for model comparison to overcome the difficulties in comparing models that were constructed using different frameworks (Shono 2008).

I analyzed the commercial fishery records which involved gillnets with a mesh size ranging from 51 to 140 mm that were soaked in 3-66 meter deep waters for 9-36 hours during August to November in 1994-2007. I selected these records for analysis because the values of the explanatory variables in these records fell into the same range as in the PIS data, which ensured prediction accuracy when we applied the delta model developed based on the PIS data to the records in the commercial fishery data. Although the commercial fishery data contained lake-wide and year-round information about commercial gillnet fisheries in Lake Erie, it was inappropriate to predict the total lake-wide and/or annual discards beyond the scope of the PIS data on which the prediction model (Delta-AdaBoost) was built.

During August to November in 1994-2007, across all analyzed records, white perch had the highest discard percentage (44.6% on average), followed by walleye (10.9% on average) and yellow perch (0.8% on average, Table 2-4). Walleye, yellow perch and white perch were

easily caught as bycatch in gillnet fisheries in Lake Erie because of the overlap in size and spatial distribution with given target species (Scott and Crossman 1973; Jester 1977; Kinnunen 2003; Johnson et al. 2004). Occurrence of discards of walleye and yellow perch can be attributed to limited quota and restricted landing size since these two species are regulated through a quota system in the Great Lakes (Kinnunen 2003). Low marketing value and unmarketable size can be the major reasons for white perch discard on Lake Erie. Specific reasons for the discards of these three species from commercial gillnet fisheries in Lake Erie can be extrapolated by incorporating the information about age or length structure into bycatch and discard assessments in future studies.

The higher discard percentage of white perch compared with walleye and yellow perch can be attributed to the fewer landings, the higher predicted bycatch and the higher discards. Since the chances that fishermen targeted white perch in the commercial gillnet fisheries was low (only 1-2% records among all analyzed records), we had very limited white perch landings (1/23 of the total walleye landings and 1/18 of the total yellow perch landings), but a high chance of catching it as bycatch (Table 2-4). Meanwhile, the lower popularity of white perch led to the higher discards even though theoretically white perch is a non-quota species and the fishermen can keep all. Since the white perch was an invasive species and has imposed considerable influences on the lake ecosystem and fish communities through competition, more fishing effort on white perch may increase its fishing mortality and may mitigate the competition between white perch and the native species. Given the high discard percentage of white perch, developing a new market for white perch could be a potential solution to avoid wasting white perch resources.

The abundance of walleye in Lake Erie kept decreasing since the late 1980s and remained at a low level during 2000-2003; the abundance of yellow perch started to increase in the late 1990s (WTG 2010; YPTG 2010). To keep a sustainable percid fishery, the Lake Erie Committee launched the Lake Erie Coordinated Percid Management Strategy (the Strategy) in 2001. The total allowable catch (TAC) for both walleye and yellow perch was reduced to a conservative level and this conservative TAC was kept for a minimum of three years for walleye. Low walleye abundance since 2000 may reduce the chance of catching walleye as bycatch, which explained the lower bycatch of walleye in the years after 2000 (Figure 2-3d).

High level of yellow perch abundance during 2001-2005 may increase the chance of catching yellow perch as bycatch, which may lead to the higher bycatch of yellow perch in the years from 2001-2005 (Figure 2-4d). We observed a high level of walleye discards in 2003 (Figure 2-6d) because walleye abundance rebounded since 2003 but the TAC for walleye remained at a low level according to the Strategy. We obtained a high level of yellow perch discards in 2001 (Figure 2-7d), which can likely be attributed to high abundance of yellow perch and reduced TAC for yellow perch in 2001 according to the Strategy.

Several methods to reduce the bycatch and discards in the commercial fisheries in the Great Lakes have been employed, including reducing effort, modifying fishing gear, and using incentives and penalties in the quota system (Johnson et al. 2004). Minimum size restrictions and closures of a fishery during spawning season or in permanent/seasonal refuges have been applied to keep a sustainable commercial fishery in the Great Lakes (Kinnunen 2003). To improve the bycatch management in Lake Erie, my analysis highlighted the hotspots for bycatch and discards for these three species in the commercial gillnet fisheries. For example, more bycatch were obtained in the west basin in October for walleye, in the west central basin in November for yellow perch, and in the west central basin in October for white perch (Figures 2-3, 2-4 and 2-5). More discards may occur in the west basin of Lake Erie during August to September for walleye, in the waters across the west central and east central basin in November for yellow perch, and in the west basin in August and November for white perch (Figures 2-6, 2-7 and 2-8). Restricted fishing season and/or fishing location can be applied in these hotspots to reduce the probability of bycatch and discards for these three species.

Analyses of the percentage of discards by major target species in this study also had important implications for bycatch management. A joint license framework can be developed for these three species based on this study. For instance, the fishing license for yellow perch can be issued jointly with the license for walleye, white perch or white bass because when targeting yellow perch, more bycatch and discards of walleye (Figures 2-3a and 2-6a) and white perch (Figures 2-5a and 2-8a) may be observed, and targeting white perch (Figures 2-4a and 2-7a) or white bass (Figure 2-7a) may yield more bycatch and discards of yellow perch. Wasted fish resources in terms of discards can be reduced by converting discarded bycatch into landed catch through such a joint license framework. Quota allocation for joint-licensed

species in this framework can be quantified by analyzing the ratio of target catch to bycatch, which could become an extension of this study.

Advanced fishery data recording systems and observer programs, which can report both bycatch and discard information in the commercial gillnet fisheries in Lake Erie, can provide detailed information for more accurate analyses. For example, with additional discard information recorded in the commercial fishery data, a model to estimate bycatch and discards can be developed directly based on the commercial fishery data in this study.

The importance of fishery-independent surveys was emphasized in this study. Without the information from the Lake Erie Partnership Index Fishing Survey, it would have been difficult to assess the bycatch and discards that actually occurred at lake since this piece of information has not been recorded in the commercial fishery data during our study period. Limited information about the bycatch and discards on site is a problem commonly seen in bycatch analyses from commercial fisheries. Additional information from a fishery-independent survey can help to calibrate the bycatch and discard estimation that are conducted directly based on the commercial data. Thus, it is worthwhile to continue the fishery-independent surveys, such as PIS, for the key species.

When analyzing the target species-specific, temporal and spatial variations in percentage of discards, especially when analyzing the percentage of discards by major target species, we obtained some negative values. For example, I had a negative value of discard percentage for walleye when targeting white bass (Figure 2-6a). The negative values happened when the bycatch predicted for most of the records within the stratum was smaller than the landed bycatch that was recorded in the commercial fishery data. One possible reason is that the PIS data used for model development did not capture the characteristics of the commercial fisheries in terms of the co-existence of target species with bycatch species. For instance, the chance of catching walleye when fishing white bass in the PIS is smaller than in the commercial fisheries. The possible reasons can be investigated by analyzing the catch composition in the PIS data and in the commercial fishery data in future studies.

The values of discard percentage by major target species for walleye (Figure 2-6a) and yellow perch (Figure 2-7a) were much higher than those by month, basin or year. The higher values of discard percentage were obtained because we assumed only one species was targeted

for each record and excluded the records that targeted the species of interest when analyzing the discard percentage by major target species. The percentage of discards by major target species was actually the percentage of discards among the predicted bycatch. We did not present the results of the white perch discard percentage when targeting lake whitefish, white bass or walleye (Figure 2-8a) because the predicted bycatch of white perch predicted from the Delta-AdaBoost model was very low (Figure 2-5a) and we would not expect any discards from such low bycatch.

In conclusion, the bycatch and discard assessments for walleye, yellow perch and white perch conducted in this study had important implications for bycatch management for these species in the commercial gillnet fisheries on Lake Erie. The successful application of the AdaBoost algorithm combined with a delta model in this study indicated that the Delta-AdaBoost model can be considered as a candidate model when a high proportion of zero observations are included in the fishery data.

2.6. Acknowledgement

This research was supported by the Department of Fisheries and Wildlife Sciences at Virginia Polytechnic Institute and State University, the USDA Cooperative State Research, Education and Extension Service through Hatch Project #0210510, and the Ontario Commercial Fisheries Association to Y. Jiao.

2.7. References

Akaike, H. 1974. A new look at the statistical model identification. IEEE transactions on automatic control 19:716-723.

Alverson, D., M. Freeberg, J. Pope and S. Murawski. 1994. A global assessment of fisheries bycatch and discards. Food and Agriculture Organization of the United Nations, Rome.

Bigelow, K., C. Boggs and X. He. 1999. Environmental effects on swordfish and blue shark catch rates in the US North Pacific longline fishery. Fisheries Oceanography 8:178-198.

Borges, L., E. Rogan and R. Officer. 2005. Discarding by the demersal fishery in the waters around Ireland. Fisheries Research 76:1-13.

Breiman, L., J. Friedman, R. Olshen and C. Stone. 1984. Classification and regression trees. Wadsworth, Belmont.

Burnham, K. and D. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2 nd edition. Springer Verlag, New York.

Crowder, L. and S. Murawski. 1998. Fisheries bycatch: implications for management. Fisheries 23:8-17.

Damalas, D., P. Megalofonou and M. Apostolopoulou. 2007. Environmental, spatial, temporal and operational effects on swordfish (Xiphias gladius) catch rates of eastern Mediterranean Sea longline fisheries. Fisheries Research 84:233-246.

Freund, Y. and R. Schapire. 1996. Experiments with a new boosting algorithm. Page 148-156 in: Saitta, L. (ed), Machine Learning: Proceedings of the Thirteenth International Conference, Bari, Italy.

Friedman, J., T. Hastie and R. Tibshirani. 2000. Additive logistic regression: A statistical view of boosting. The annals of statistics 28:337-374.

Hall, M., D. Alverson and K. Metuzals. 2000. By-catch: problems and solutions. Marine Pollution Bulletin 41:204-219.

Hamley, J. 1975. Review of gillnet selectivity. Journal of the Fisheries Research Board of Canada 32:1944-1969.

Harrington, J., R. Myers and A. Rosenberg. 2005. Wasted fishery resources: discarded by-catch in the USA. Fish and Fisheries 6:350-361.

Hastie, T. and R. Tibshirani. 1990. Generalized additive models. Chapman and Hall, London, UK.

Hastie, T., R. Tibshirani and J. Friedman. 2001. The elements of statistical learning: data mining, inference and prediction, 2 nd edition. Springer, New York.

Jester, D. 1977. Effects of color, mesh size, fishing in seasonal concentrations, and baiting on catch rates of fishes in gill nets. Transactions of the American Fisheries Society 106:43-56.

Johnson, J. E., J. L. Jonas and J. W. Peck. 2004. Management of Commercial Fisheries Bycatch, with Emphasis on Lake Trout Fisheries of the Upper Great Lakes. Fisheries Research Report, State of Michigan Department of Natural Resources, Michigan.

Kawakita, M., M. Minami, S. Eguchi and C. Lennert-Cody. 2005. An introduction to the predictive technique AdaBoost with a comparison to generalized additive models. Fisheries Research 76:328-343.

Kelleher, K. 2005. Discards in the world's marine fisheries: an update. Food and Agriculture Organization of the United Nations, Rome.

Kinnunen, R. 2003. Great lakes commercial fisheries. Report from Michigan Sea Grant, Michigan Sea Grant Extension, Michigan.

Lo, N., L. Jacobson and J. Squire. 1992. Indices of relative abundance from fish spotter data based on delta-lognormal models. Canadian Journal of Fisheries and Aquatic Sciences 49:2515-2526.

Maunder, M. and A. Langley. 2004. Integrating the standardization of catch-per-unit-of-effort into stock assessment models: testing a population dynamics model and using multiple data types. Fisheries Research 70:389-395.

Maunder, M. and A. Punt. 2004. Standardizing catch and effort data: a review of recent approaches. Fisheries Research 70:141-159.

Murawski, S. 1996. Factors influencing by-catch and discard rates: analyses from multispecies/multifishery sea sampling. Journal of Northwest Atlantic Fishery Science 19:31-40.

Murray, K. 2004. Magnitude and distribution of sea turtle bycatch in the sea scallop (*Placopecten magellanicus*) dredge fishery in two areas of the Northwestern Atlantic Ocean, 2001-2002. Fishery Bulletin 102:671-681.

OCFA (Ontario Commercial Fisheries Association). 2007. Lake Erie Partnership Index Fishing Survey [online]. Available: http: //www.ocfa.on.ca/Lake Erie Partnership Index Fishing Survey.htm. (September 2009).

Ortiz, M., C. Legault and N. Ehrhardt. 2000. An alternative method for estimating bycatch from the US shrimp trawl fishery in the Gulf of Mexico, 1972-1995. Fishery Bulletin 98:583-599.

Parrish, D. and F. Margraf. 1990. Interactions between white perch (*Morone americana*) and yellow perch (*Perca flavescens*) in Lake Erie as determined from feeding and growth. Canadian Journal of Fisheries and Aquatic Sciences 47:1779-1787.

Pennington, M. 1996. Estimating the mean and variance from highly skewed marine data. Fishery Bulletin 94:498-505.

Punt, A., D. Smith, G. Tuck and R. Methot. 2006. Including discard data in fisheries stock assessments: two case studies from south-eastern Australia. Fisheries Research 79:239-250.

Schaeffer, J. and F. Margraf. 1987. Predation on fish eggs by white perch, *Morone americana*, in western Lake Erie. Environmental Biology of Fishes 18:77-80.

Scott, W. B. and E. J. Crossman. 1973. Freshwater fishes of Canada. Fisheries Research Board of Canada Bulletin 184, Ontario.

Shono, H. 2008. Application of the Tweedie distribution to zero-catch data in CPUE analysis. Fisheries Research 93:154-162.

Stefansson, G. 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES Journal of Marine Science 53:577.

Thomas, M. and R. Haas. 2005. Status of yellow perch and walleye populations in Michigan waters of Lake Erie, 1999–2003. Fisheries Research Report, 2082, Michigan Department of Natural Resources, MI.

Tweedie, M. 1984. An index which distinguishes between some important exponential families. Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference, Indian Statistical Institute, Calcutta.

WTG (Lake Erie Walleye Task Group). 2010. Lake Erie Walleye Task Group Annual Report. Available: http://www.glfc.org/lakecom/lec/WTG.htm (July 2010).

Ye, Y., M. Al-Husaini and A. Al-Baz. 2001. Use of generalized linear models to analyze catch rates having zero values: the Kuwait driftnet fishery. Fisheries Research 53:151-168.

YPTG (Lake Erie Yellow Perch Task Group). 2010. Lake Erie Yellow Perch Task Group Annual Report. Available: http://www.glfc.org/lakecom/lec/YPTG.htm (July 2010).

TABLE 2-1.—Spearman correlation coefficients among variables based on the data from the Lake Erie Partnership Index Fishing Survey (PIS) for walleye, yellow perch and white perch, 1989-2008.

Variables	Site depth	Soak time	Basin	Year	Month	Gear type	Mesh size
Site depth	1.00						
Soak time	-0.02	1.00					
Basin	0.48	-0.01	1.00				
Year	0.00	-0.10	0.02	1.00			
Month	-0.13	0.01	-0.49	-0.04	1.00		
Gear type	-0.09	0.01	-0.07	-0.06	0.02	1.00	
Mesh size	0.00	0.00	0.00	0.00	0.00	0.00	1.00

TABLE 2-2.—A stepwise generalized additive model (GAM) building to predict positive captures of walleye, yellow perch and white perch from commercial gillnet fisheries in Lake Erie. A log-normal distribution was assumed.

	Variables added	df	Deviance	AIC	*P-value* (χ^2)	Deviance decrement	Cumulative % of deviance explained
Walleye							
0	Null		46,888	41,248			
1	Mesh Size	13	42,881	40,448	$<2.2\times10^{-16}$	4,007	9.7
2	Year	19	41,567	40,198	$<2.2\times10^{-16}$	1,314	12.9
3	Basin	4	40,153	39,886	$<2.2\times10^{-16}$	1,414	16.3
4	Gear type	1	39,263	39,681	$<2.2\times10^{-16}$	890	18.5
5	Month	3	38,921	39,606	1.4×10^{-15}	342	19.3
6	f(Soak time)	1	38,815	39,583	5.1×10^{-7}	106	19.6
7	f(Site depth)	1	38,798	39,559	0.04	17	19.6
Yellow perch							
0	Null		368,175	81,816			
1	Mesh Size	13	342,416	80,879	$<2.2\times10^{-16}$	25,759	7.0
2	Year	19	311,055	79,641	$<2.2\times10^{-16}$	31,361	15.5
3	Gear type	1	282,739	78,376	$<2.2\times10^{-16}$	28,316	23.2
4	Basin	4	258,219	77,179	$<2.2\times10^{-16}$	24,520	29.9
5	f(Soak time)	1	256,444	77,089	$<2.2\times10^{-16}$	1,775	30.3
6	Month	3	255,430	76,917	2.2×10^{-11}	1,014	30.6
7	f(Site depth)	1	255,138	76,887	9.7×10^{-5}	292	30.7
White perch							
0	Null		70,984	58,918			
1	Year	19	67,067	58,219	$<2.0\times10^{-16}$	3,917	5.5
2	Gear type	1	64,534	57,721	$<2.2\times10^{-16}$	2,533	9.1
3	Mesh Size	13	60,588	56,927	$<2.2\times10^{-16}$	3,946	14.6
4	Month	3	58,653	56,512	$<2.2\times10^{-16}$	1,935	17.4

5	f(Site depth)	1	56,993	56,141	$<2.2\times10^{-16}$	1,660	19.7
6	Basin	4	55,521	55,412	$<2.0\times10^{-16}$	1,472	21.8
7	f(Soak time)	1	55,514	55,330	0.19	7	21.8

TABLE 2-3.—Training and test errors from 5-fold cross-validation based on the data from the Lake Erie Partnership Index Fishing Survey (PIS) for walleye, yellow perch and white perch, 1989-2008.

Model	Training error						Test error					
	1	2	3	4	5	Average	1	2	3	4	5	Average
Walleye												
Delta-GLM	0.476	0.474	0.480	0.482	0.476	0.478	0.482	0.485	0.475	0.468	0.481	0.478
Delta-GLM-Poly	0.472	0.469	0.475	0.477	0.471	0.473	0.478	0.480	0.471	0.463	0.476	0.473
Delta-GAM	0.472	0.469	0.475	0.477	0.471	0.473	0.478	0.480	0.471	0.464	0.476	0.474
GLM-Tweedie	0.487	0.484	0.490	0.492	0.487	0.488	0.491	0.495	0.485	0.479	0.494	0.489
GAM-Tweedie	0.482	0.480	0.485	0.487	0.482	0.483	0.487	0.489	0.481	0.474	0.489	0.484
Delta-AdaBoost	0.440	0.438	0.442	0.444	0.438	0.441	0.449	0.451	0.441	0.432	0.446	0.444
Yellow perch												
Delta-GLM	0.768	0.772	0.779	0.777	0.769	0.773	0.784	0.785	0.756	0.760	0.793	0.776
Delta-GLM-Poly	0.696	0.694	0.709	0.705	0.699	0.701	0.713	0.713	0.685	0.687	0.720	0.703
Delta-GAM	0.702	0.701	0.713	0.710	0.704	0.706	0.718	0.720	0.691	0.692	0.725	0.709
GLM-Tweedie	0.794	0.795	0.803	0.802	0.794	0.797	0.803	0.803	0.778	0.789	0.822	0.799
GAM-Tweedie	0.728	0.728	0.738	0.738	0.728	0.732	0.744	0.741	0.708	0.723	0.752	0.734
Delta-AdaBoost	0.672	0.673	0.684	0.682	0.674	0.677	0.687	0.691	0.669	0.666	0.694	0.682
White perch												
Delta-GLM	0.327	0.323	0.319	0.334	0.326	0.326	0.342	0.325	0.320	0.323	0.326	0.327
Delta-GLM-Poly	0.321	0.317	0.315	0.326	0.317	0.319	0.338	0.317	0.315	0.314	0.318	0.321
Delta-GAM	0.320	0.317	0.315	0.326	0.318	0.319	0.337	0.316	0.315	0.314	0.320	0.321
GLM-Tweedie	0.341	0.348	0.345	0.352	0.349	0.347	0.357	0.346	0.347	0.340	0.349	0.348
GAM-Tweedie	0.332	0.338	0.336	0.343	0.339	0.338	0.350	0.336	0.336	0.330	0.340	0.338
Delta-AdaBoost	0.311	0.317	0.314	0.323	0.323	0.317	0.316	0.293	0.296	0.292	0.299	0.299

TABLE 2-4.— Total bycatch (kg) predicted from the delta model, total discards (kg) estimated by comparing total predicted bycatch and total landed bycatch recorded in commercial data, and percentage of discards (%, percentage of total discards among total landings of the species of interest), across all analyzed records from commercial gillnet data in Lake Erie in the fall (August to November) of 1994-2007. A 95% confidence interval is indicated in parentheses.

Species	Walleye	Yellow perch	White perch
Total analyzed records[a]	32,349	32,282	32,392
Total predicted bycatch ($\times 10^6$ kg)	1.29 (1.14 to 1.46)	0.042 (0.038 to 0.047)	1.31 (0.92 to 1.74)
Total landed bycatch ($\times 10^6$ kg)	0.54	0.008	0.60
Total estimated discards ($\times 10^6$ kg)	0.75 (0.60 to 0.92)	0.035 (0.030 to 0.039)	0.71 (0.31 to 1.14)
Total landed catch ($\times 10^6$ kg)	5.61	4.32	0.24
Percentage of discards[b] (%)	10.9 (8.9 to 13.0)	0.8 (0.7 to 0.9)	44.6 (27.2 to 57.5)

a: In the commercial data of 1994-2001, each record represents one net; in the commercial data of 2002-2007, each record represents a daily report which includes 1-5 nets.

b: Percentage of discards (%)=discarded bycatch of the species of interest / (predicted bycatch of the species of interest + landed catch of the species of interest).

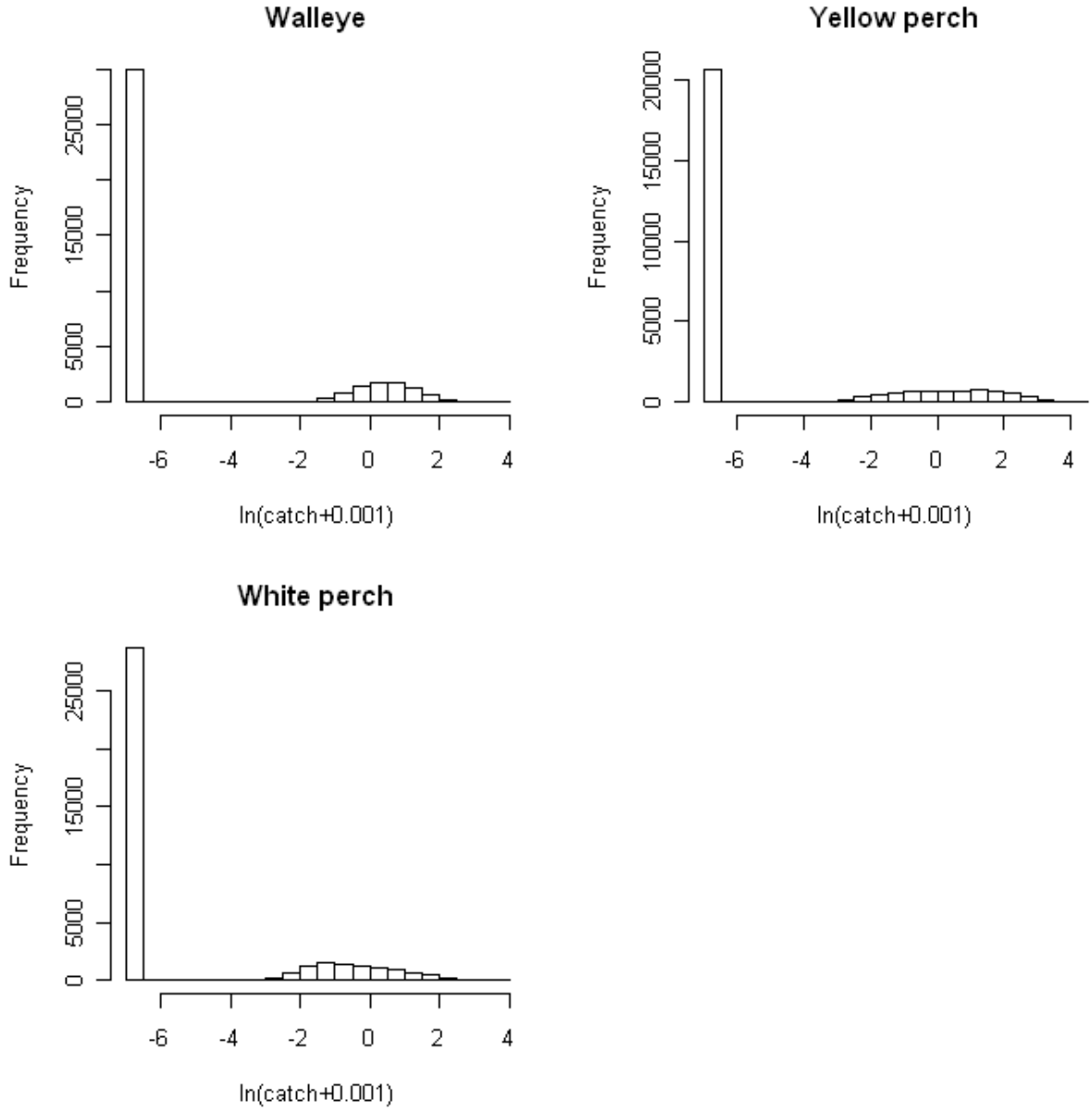

FIGURE 2-1.— Histograms of log-transformed catch data (kg) of walleye, yellow perch and white perch in Lake Erie, collected by the Lake Erie Partnership Index Fishing Survey, 1989-2008, ln(catch+0.001).

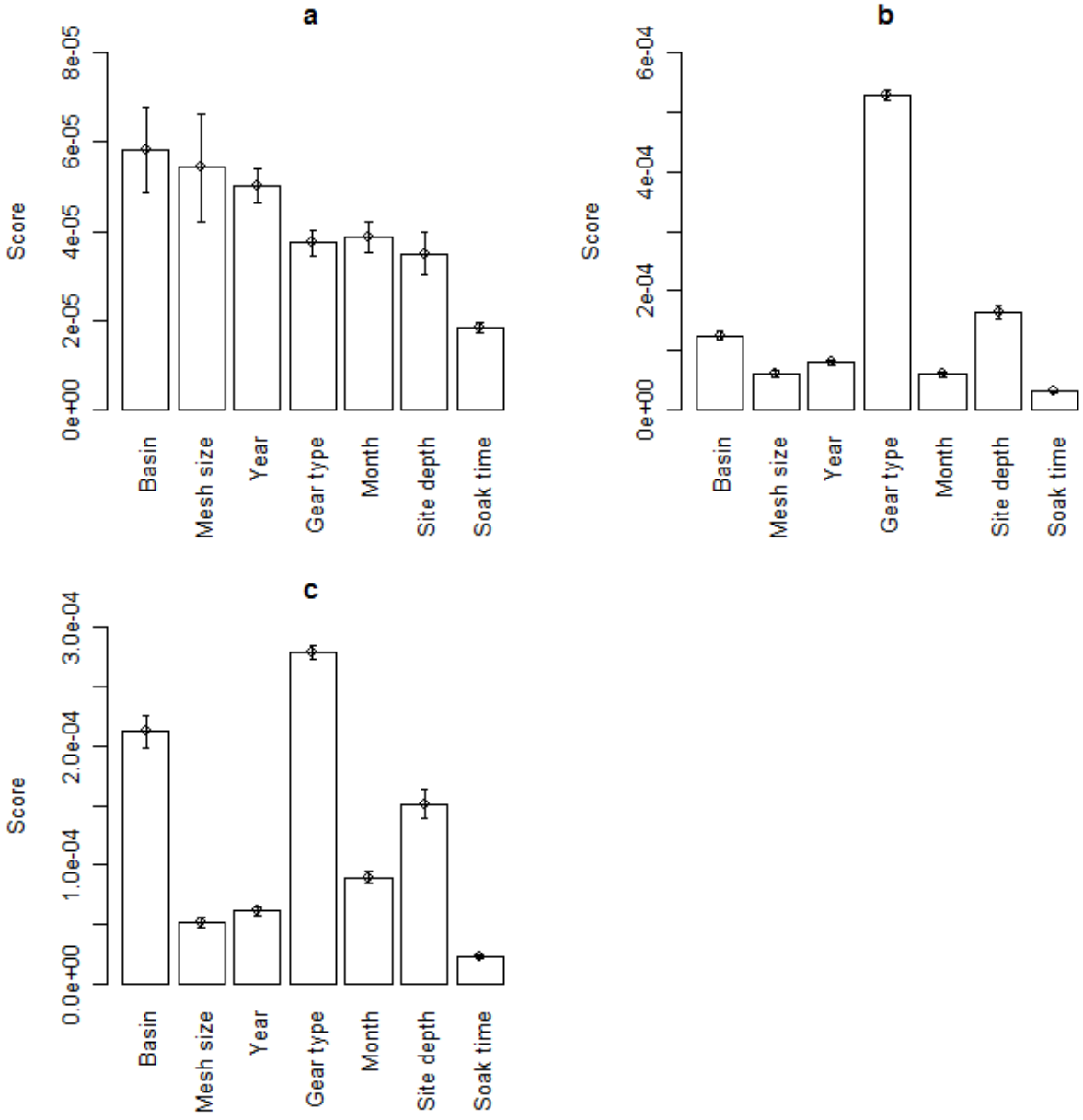

FIGURE 2-2.— Score plots for walleye, yellow perch and white perch (averaged over 1000 simulations), derived from an AdaBoost model to predict the probability of obtaining non-zero captures based on the data from Lake Erie Partnership Index Fishing Survey, 1989-2008. Error bars show the standard deviation.

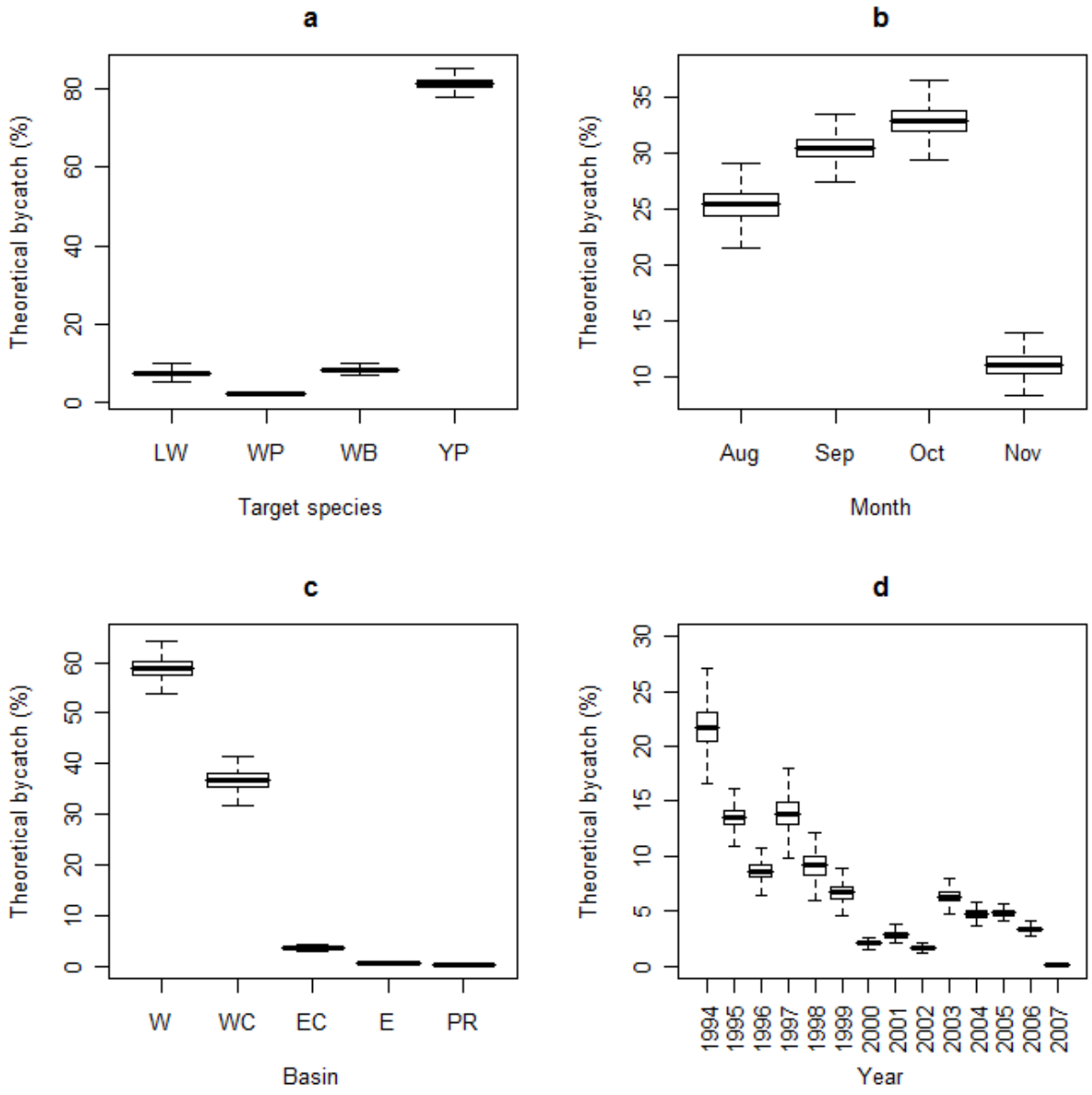

FIGURE 2-3.— Percentage composition of the predicted bycatch (%) of walleye by the major target species, year, month and basin from commercial gillnet fisheries in Lake Erie during August to November in 1994-2007. LW-lake whitefish, WP-white perch, WB-white bass, YP-yellow perch; Aug-August, Sep-September, Oct-October, Nov-November; W-west basin, WC-west central basin, EC-East central basin, E-east basin, PR-Pennsylvania Ridge.

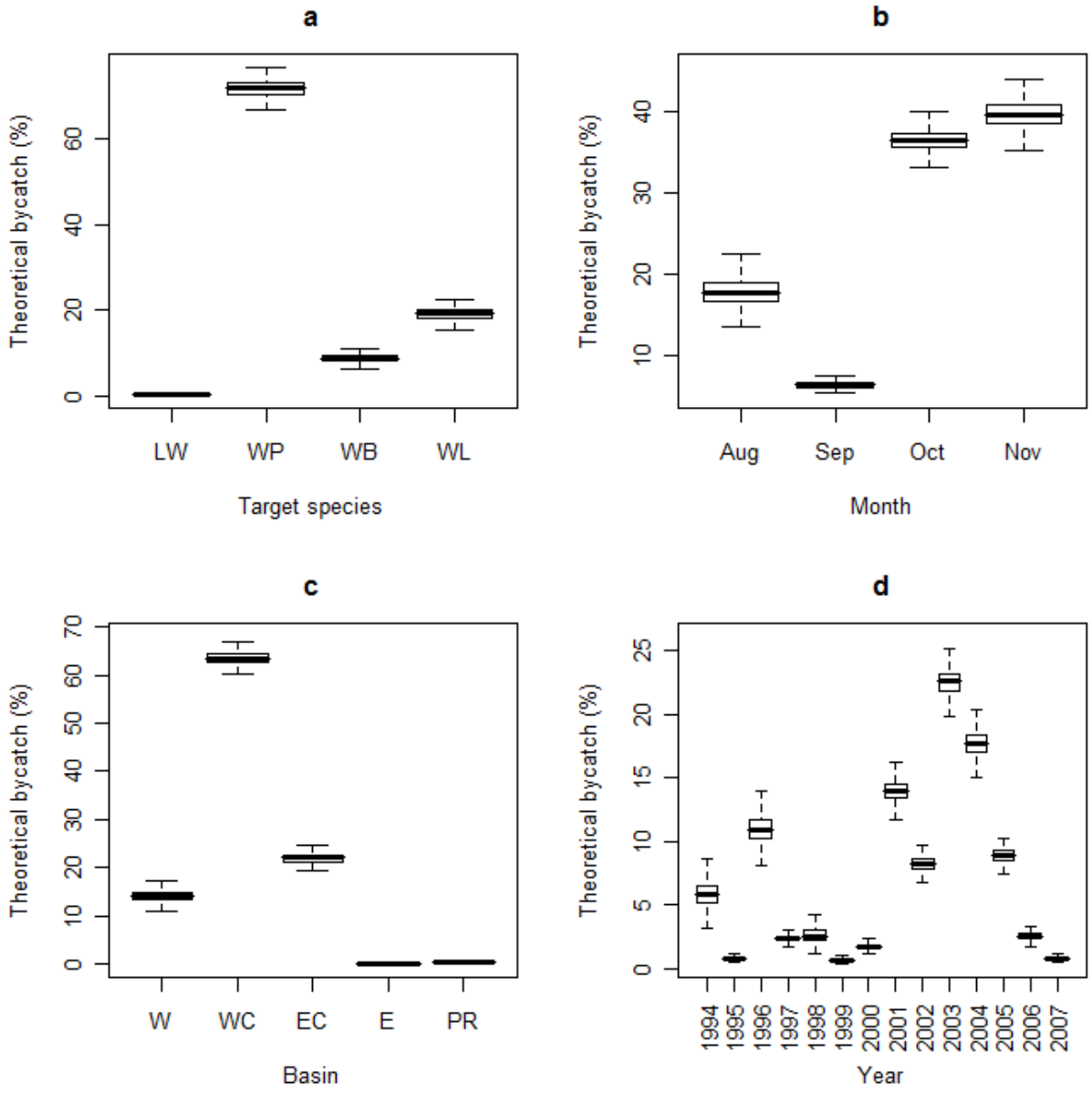

FIGURE 2-4.— Percentage composition of the predicted bycatch (%) of yellow perch by the major target species, year, month and basin from commercial gillnet fisheries in Lake Erie during August to November in 1994-2007. LW-lake whitefish, WP-white perch, WB-white bass, WL-walleye; Aug-August, Sep-September, Oct-October, Nov-November; W-west basin, WC-west central basin, EC-East central basin, E-east basin, PR-Pennsylvania Ridge.

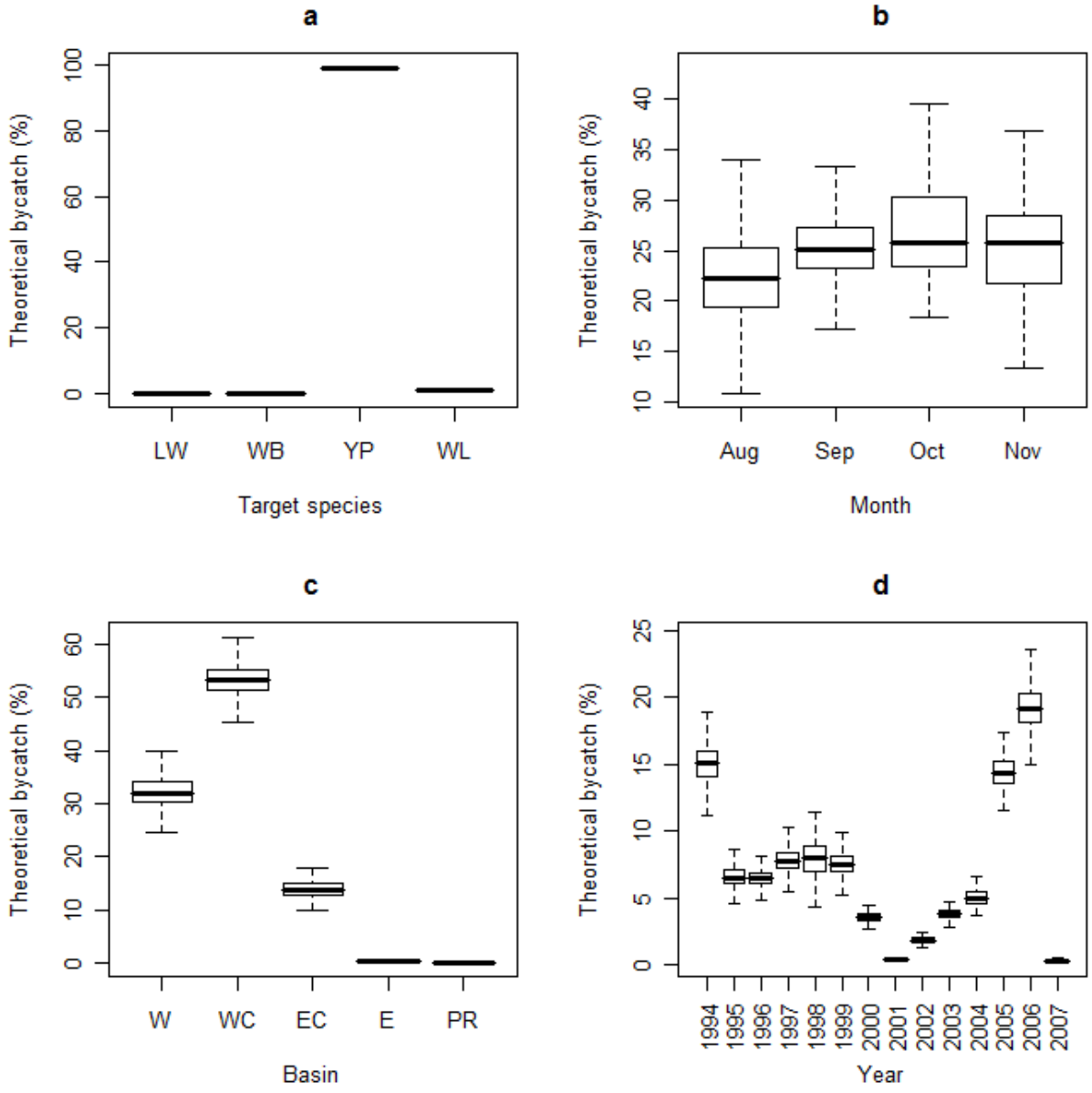

FIGURE 2-5.— Percentage composition of the predicted bycatch (%) of white perch by the major target species, year, month and basin from commercial gillnet fisheries in Lake Erie during August to November in 1994-2007. LW-lake whitefish, WB-white bass, YP-yellow perch, WL-walleye; Aug-August, Sep-September, Oct-October, Nov-November; W-west basin, WC-west central basin, EC-East central basin, E-east basin, PR-Pennsylvania Ridge.

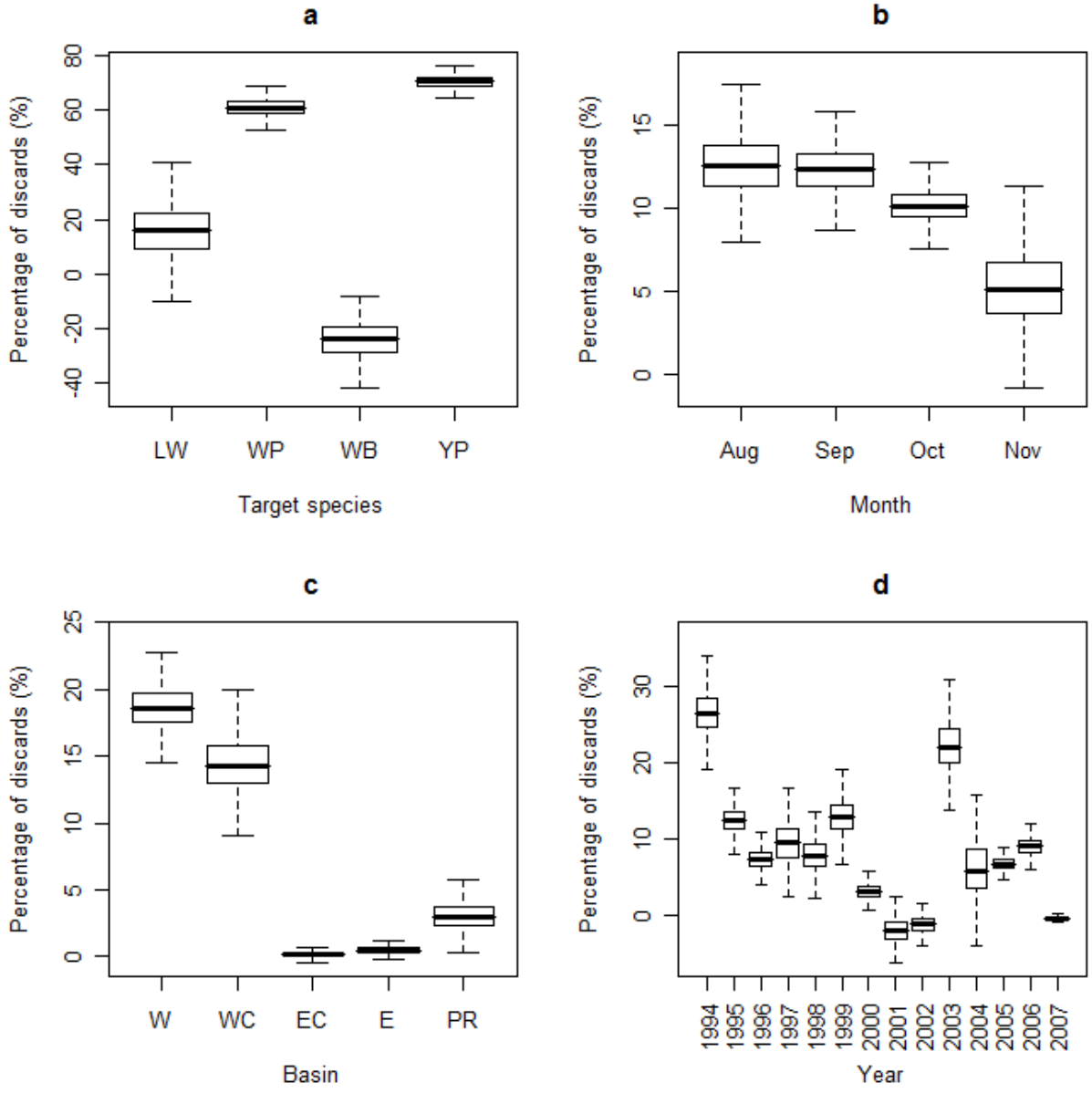

FIGURE 2-6.— Percentage of discards (%) of walleye by the major target species, year, month and basin from commercial gillnet fisheries in Lake Erie during August to November in 1994-2007. LW-lake whitefish, WP-white perch, WB-white bass, YP-yellow perch; Aug-August, Sep-September, Oct-October, Nov-November; W-west basin, WC-west central basin, EC-East central basin, E-east basin, PR-Pennsylvania Ridge.

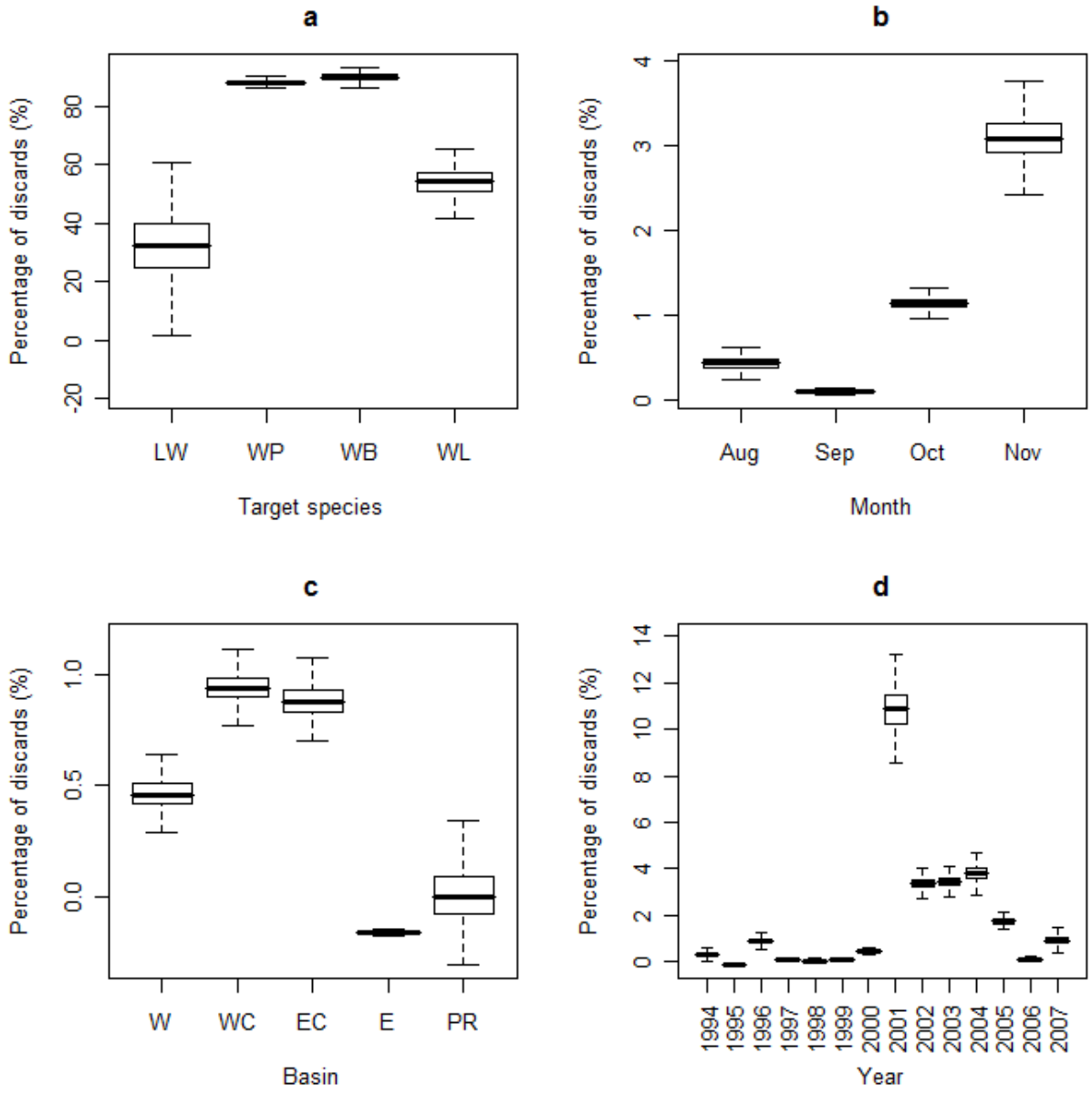

FIGURE 2-7.— Percentage of discards (%) of yellow perch by the major target species, year, month and basin from commercial gillnet fisheries in Lake Erie during August to November in 1994-2007. LW-lake whitefish, WP-white perch, WB-white bass, WL-walleye; Aug-August, Sep-September, Oct-October, Nov-November; W-west basin, WC-west central basin, EC-East central basin, E-east basin, PR-Pennsylvania Ridge.

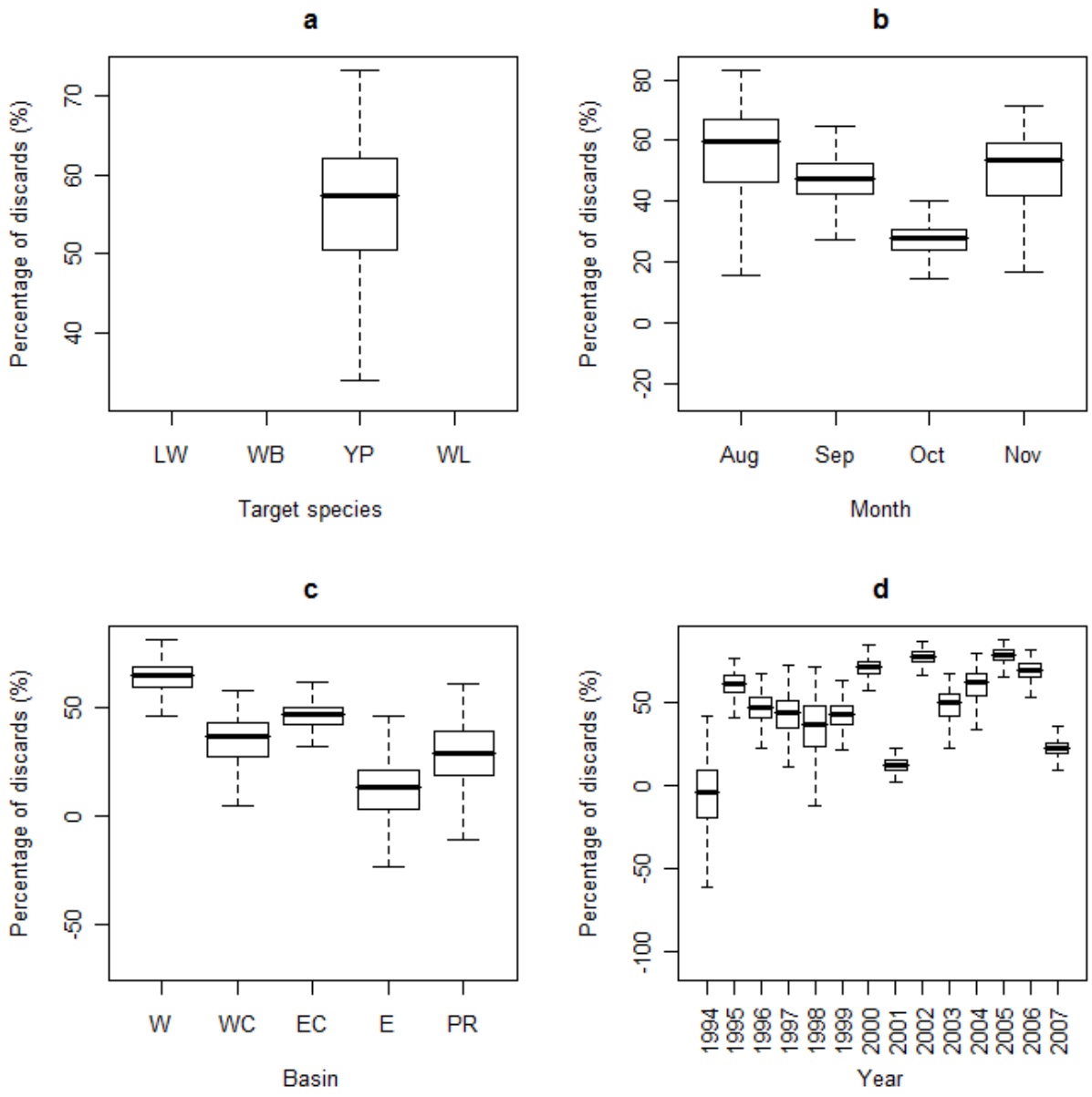

FIGURE 2-8.— Percentage of discards (%) of white perch by the major target species, year, month and basin from commercial gillnet fisheries in Lake Erie during August to November in 1994-2007. LW-lake whitefish, WB-white bass, YP-yellow perch, WL-walleye; Aug-August, Sep-September, Oct-October, Nov-November; W-west basin, WC-west central basin, EC-East central basin, E-east basin, PR-Pennsylvania Ridge.

Chapter 3

Influences of gillnet fishing on lake sturgeon bycatch in Lake Erie and implications for conservation

3.1. Abstract

Three classification tree models were constructed to estimate the probability of obtaining lake sturgeon (*Acipenser fulvescens*) bycatch under specific environmental and gillnet fishing conditions in Lake Erie. This analysis was based on data collected from 1989 to 2008 from a fishery-independent survey, the Lake Erie Partnership Index Fishing Survey (PIS). The three classification tree models included the conditional inference classification tree generated by the package 'party' in R, and two exhaustive search based classification trees generated by the packages 'tree' and 'rpart' in R, respectively. The discriminative performance of these three trees was evaluated by the Receiver Operating Characteristic (ROC) curve and the area under curve (AUC) through a jackknife procedure. I found that most lake sturgeon captured by gillnets in the PIS were juvenile fish. All three trees indicated that the lake sturgeon bycatch was most likely to be observed in the west basin during the years 1994 and 2007 although the structures of the three trees were different. Model comparison based on ROC and AUC analyses revealed that the 'tree' tree performed best on the training data, whereas the 'party' tree performed best on the test data. The gillnet fisheries in Lake Erie may potentially increase the mortality of juvenile lake sturgeon and thereby impede the recovery of the lake sturgeon population. The west basin of Lake Erie could be a hotspot for lake sturgeon bycatch in the gillnet fisheries, and more attention should be given to gillnet fisheries management in the west basin with an emphasis on lake sturgeon conservation.

3.2. Introduction

Lake sturgeon (*Acipenser fulvescens*) was abundant in the Great Lakes during the late 1800s, contributing to both recreational and commercial fisheries (Auer 1999). The demographic characteristics of lake sturgeon, such as high longevity, late maturity, and intermittent spawning, make its populations vulnerable to heavy exploitation and dramatic environmental changes (Noakes et al. 1999; Welsh 2004). Within the last century, populations

of lake sturgeon have been reduced dramatically or extirpated due to overfishing, habitat degradation, anthropogenic impacts on nursery and spawning areas, damming that impedes its migration, and water pollution (Scott and Crossman 1973; Birstein et al. 1997; Auer 1999; Bogue 2000). Although the lake sturgeon is not listed under the United States Endangered Species Act (ESA, Public Law 93-205), it has been listed as an endangered or threatened species in 19 of the 20 states within its original range in the United States (GLLSCM 2002; Welsh 2004; GLLSCM 2010). The Canadian province of Ontario shares the waters of four Great Lakes with the United States. In the province of Ontario, although the recreational or commercial harvest of lake sturgeon is permitted, it has been considered to be a sensitive species and its harvest is regulated by a quota system with increasing concerns of fishery management and fish conservation (CESCC 2001; Welsh 2004;GLLSCM 2005).

Lake sturgeon migrate to rivers throughout the lake basins to spawn and spend the remaining time in the open waters of the Great Lakes. Lake Erie supports several non-spawning and few spawning populations of lake sturgeon (Welsh 2004). Some rivers off of Lake Erie have been identified to be the major spawning areas of lake sturgeon, including the Detroit River that connects Lake Erie and the St. Clair system (Lake St. Clair and the St. Clair River) and the Niagara River (Welsh 2004; Thomas and Haas 1999; GLLSCM 2002; Thomas and Haas 2002; Boase 2003; Caswell 2003). Lake Erie also supports large gillnet fisheries of some key fish species, such as walleye *Sander vitreus* and yellow perch *Perca flavescens* (Kinnunen 2003; Thomas and Haas 2005). Because of the overlap in life history traits between (juvenile) lake sturgeon and the major target species in the gillnet fisheries, the low species-specific selectivity and higher mortality of gillnets, and the protected status of lake sturgeon, gillnet bycatch could be considered as an additional threat to lake sturgeon populations in Lake Erie (Scott and Crossman 1973; Hamley 1975; GLLSCM 2002; Kinnunen 2003; Johnson et al. 2004).

Previous studies have focused on the demographic characteristics and stock structure of lake sturgeon populations residing in the river systems through genetic tools and experimental surveys (Threader and Broussaeu 1986; Chiasson et al. 1997; Ferguson and Duckworth 1997; McKinley et al. 1998; Noakes et al. 1999; Thomas and Haas 1999; Thomas and Haas 2002). However, the impact of gillnet fisheries on bycatch of lake sturgeon in Lake Erie has been less

well documented. Failure to take account of the influence of gillnet fisheries in lake sturgeon conservation and fishery management may conceal the potential threats from gillnet fisheries, and may reduce the efficiency of recovering its populations.

I analyzed the potential influence of gillnet fisheries on lake sturgeon bycatch using classification tree approaches. The data analyzed in this study were collected by a fishery independent survey program, the Lake Erie Partnership Index Fishing Survey (PIS), which was mainly conducted by the Ontario Ministry of Natural Resources (OMNR) and the Ontario Commercial Fisheries Association (OCFA) since 1989. Experimental gillnets with fourteen mesh sizes, ranging from 32 to 152 mm, were set at the bottom (bottomed net) or suspended (canned net) at sites distributed across the Ontario waters of Lake Erie in the fall (August to November) annually, using commercial fishing vessels and commercial fishing crews (OCFA 2007).

The PIS data contained a high percentage of zero captures (31 individuals captured in 53,562 nets), which is typically encountered in catch or bycatch analyses of rare species (Maunder and Punt 2004). The presence of zeros may invalidate the assumptions of normality commonly used in fishery analyses, and may cause computational difficulties (Maunder and Punt 2004). Elimination of a considerable proportion of zeros may result in a loss of the information on spatial or temporal distribution characteristics (Maunder and Punt 2004). The methods to deal with the fishery data having zeros in previous studies can be categorized into two types. One method is to add a small constant to each observation in the generalized linear/additive model analysis (Ortiz et al. 2000; Maunder and Punt 2004; Shono 2008). The other method is to use the delta model or Tweedie distribution model. The delta model deals with the positive values and zeros using two sub-models separately (Lo et al. 1992; Stefansson 1996; Ye et al. 2001; Maunder and Langley 2004), whereas the Tweedie distribution model handles zeros uniformly along with the positive values (Tweedie 1984; Shono 2008). However, based on a preliminary analysis, it was difficult to capture quantitatively the complexity of the data and the underlying relationship between the lake sturgeon bycatch and the predictor variables by applying these commonly-used methods because of the extremely high percentage of zero observations (> 99%) in the PIS data and the interactions stemming from the biological and ecological characteristics of lake sturgeon populations.

The classification tree model approach was applied to capture the complex interactions in the PIS data and to analyze the relationship between the lake sturgeon bycatch and the predictors given a dataset having more than 99% zeros. The tree-based approach was introduced by Breiman et al. (1984) from a statistical perspective and has been used widely for classification problems in clinical, genetic and ecological studies because this method allows for easy interpretation of the model, graphical visualization of model structure, and identification of complex interactions in the data (Ribic and Ainley 1997; Fielding 1999; Harrell 2001; Gansky 2003; Austin 2008; Cutler et al. 2009). In the fishery context, if the values of the response variable take 1 or 0, indicating the presence or absence of lake sturgeon bycatch, respectively, the PIS data can be treated as a two-group classification problem, and the goal is to use the predictor variables to classify the observations with the label 1 or 0.

A tree is constructed starting from selecting a single predictor variable at the root node that contains all the observations. A split on the predictor is determined to partition this node and send the observations in this node to one of the two descendant nodes (Breiman et al. 1984; Cutler et al. 2009). The same partitioning procedure is applied to each descendant node recursively until a stopping criterion is met. The non-partitioned nodes in a tree that stops growing are called terminal nodes. The classification outcome for each observation is obtained by computing either class proportions or the class of the majority from the observations at a terminal node (Breiman et al. 1984; Cutler et al. 2009). There are two types of classification tree methods that are commonly used, the exhaustive search based tree and the conditional inference tree. In the exhaustive search-based tree, the predictor variable at a node and the split on this predictor variable are selected simultaneously through an exhaustive search over all possible splits on all predictor variables that satisfy the splitting criterion (Hothorn et al. 2006; Cutler et al. 2009). Compared with the exhaustive search-based tree, the conditional inference tree employs the well-defined statistical test framework and separates the procedure of variable selection at a node and the procedure of splitting so that the over-fitting and biased variable selection problems induced by the exhaustive search for partitioning are solved (Hothorn et al. 2006). The discriminative performance of the conditional inference tree has been proven equivalent to the optimally pruned trees that were derived by exhaustive search (Hothorn et al. 2006).

In this study, a conditional inference classification tree and two exhaustive search based classification trees were constructed to estimate the probability of obtaining lake sturgeon bycatch under specific environmental and gillnet fishing conditions in Lake Erie. The potential influences of gillnet fisheries on lake sturgeon conservation and management in Lake Erie were examined.

3.3. Methods

Data.—I developed a conditional inference classification tree and two exhaustive search based classification trees for lake sturgeon bycatch based on the PIS data that were collected in the fall (August-November) from 1989 to 2008 and provided by OCFA. In the PIS data, results of catches from 53,562 nets were available for analysis, in which 31 lake sturgeon were captured. The predictor variables (Table 3-1) included six continuous variables (site depth, gear depth, secchi depth, gear temperature, dissolved oxygen and soak time) and five categorical variables (basin, year, month, gear type, mesh size).

The conditional inference classification tree and the exhaustive search based classification tree.—Results of a preliminary analysis showed that it was difficult to model and interpret quantitatively the complex interactions embedded in such data by employing the commonly used generalized-linear/additive-based models because the PIS data contained an extremely high percentage of zero captures (>99%) and very few positive captures (31 individuals). Three tree-based models, i.e., a conditional inference classification tree and two exhaustive search based classification trees, were applied in order to estimate the probability of obtaining lake sturgeon bycatch under specific environmental and gillnet fishing conditions in Lake Erie.

A conditional inference classification tree is a unified framework in which the recursive partitioning of the tree is conducted on the basis of well-defined statistical tests (Hothorn et al. 2006). We denoted the response variable as $Y \in \{0,1\}$, where the value 0 represented the event of capturing no lake sturgeon and the value 1 represented the event of capturing at least one lake sturgeon. The *m* predictor variables were denoted as $X = (X_1,...X_m)$. Then the problem was converted into a two-group classification problem. The goal was to classify each

observation with the label 0 or 1 given a complex of predictor variables. The procedure of growing a conditional inference classification tree was described as follows (Breiman et al. 1984; Fielding 1999; Hothorn et al. 2006; Cutler et al. 2009):

(1) Test the global null hypothesis of independence between the response variable and any of the *m* predictor variables. Stop if this global null hypothesis cannot be rejected at a pre-specified significance level α based on the test statistic or *p*-value. Otherwise (i.e., at least one of the *m* predictor variables is associated with the response variable significantly), the association between Y and each of the *m* predictor variables is measured by the test statistic or *p*-value, and the variable X_j with the strongest association to Y is selected to split a node, where $j \in \{1,...,m\}$.

(2) Determine a set $A \subset X_j$ to split X_j into two disjoint sets A and $\bar{A}$, where $A \cap \bar{A} = \varnothing$ and $A \cup \bar{A} = X_j$. The set A was determined by measuring the discrepancy between all possible A and the corresponding $\bar{A}$ based on the test statistic. The A that maximizes the discrepancy is selected to determine the split value at this node. The observations at this node are assigned to the "left" if $X_{j,i} \in A$ or to the "right" if $X_{j,i} \in \bar{A}$ in order to form two descendant nodes, where $i = 1,...,n$, n is the number of observations at this node and for the root node *n* is the total number of observations.

(3) Recursively repeat steps (1) and (2) in order to select a predictor variable at a node and to determine the split value for splitting the node and for forming the descendant nodes.

(4) When the tree stops growing, the observations at a terminal node are used to compute the proportion of each class that gives the probability of obtaining each class, or to compute the class based on the most frequent class that the observations have at this terminal node.

In this study, the package 'party' in R (Version 2.9.2) was utilized to construct the conditional inference classification tree, and this tree was referred as the 'party' tree in the following analysis.

Two optimally pruned classification trees obtained by exhaustive search methods were also applied to the PIS data in order to estimate the probability of obtaining lake sturgeon bycatch in Lake Erie. The exhaustive based tree was grown in a way similar to that described above, except that with step (1) and (2) combined, the procedure of selecting the split variable at a node and the procedure of determining the split value were conducted simultaneously in a different partition framework (Hothorn et al. 2006). That is, the particular split to partition a node was selected by exhaustively searching for every possible split on every predictor variable (Breiman et al. 1984; Harrell 2001; Hothorn et al. 2006). The predictor and split combination that satisfied the splitting criterion best was selected to partition the node (Breiman et al. 1984; Cutler et al. 2009). The two exhaustive search based classification trees were programmed using the packages 'tree' and 'rpart' in R, respectively, and were referred to as the 'tree' tree and the 'rpart' tree in the following analysis.

Comparison among the three classification tree models.— The discriminative performance of each classification tree was evaluated by plotting the receiver operating characteristic (ROC) curve and calculating the area under the ROC curve (AUC). The ROC curve has long been used to visualize the performance of a classification algorithm, which plots the true positive rate $\Pr(\hat{Y} = \oplus \mid Y = \oplus)$ against the false positive rate $\Pr(\hat{Y} = \oplus \mid Y = \ominus)$ (Bradley 1997; Austin 2008). The classification algorithm is considered to exhibit better discriminative performance if it yields a ROC curve shooting to the top-left corner of the plot and extending towards the top-right corner which gives a larger AUC (Bradley 1997; Austin 2008).

The jackknife re-sampling approach was employed to conduct the uncertainty analysis of AUC for each classification tree (Miller 1974; Efron 1979; Meyer et al. 1986). In the jackknife framework, we left each observation out at one time. The one observation left out comprised the test data and the rest of the data comprised the training data. At each run of jackknifing, the classification tree was developed using the training data and the corresponding AUC was calculated. The classification tree built based on the training data was applied to the test data for prediction. The discriminative performance of each classification tree on the training data was evaluated by the average AUC with its 95% confidence interval generated

from the jackknifing. The discriminative performance on the test data was evaluated through a ROC curve that was generated by combining the predicted value of each test data from each run of jackknifing across the whole dataset.

3.4. Results

A total of 31 lake sturgeon were observed as bycatch in the Lake Erie Partnership Index Fishing Survey conducted each fall from 1989 to 2008 (Table 3-2). The lake sturgeon bycatch from the PIS had a total length of 570 mm on average with a 95% confidence interval from 229 to 857 mm (Figure 3-1). Of the 31 lake sturgeon, 26 (83.9%) were captured in the west basin, 25 (80.6%) were captured in the waters where the temperature at the gear set depth was less than 22.6 °C, 30 (96.8%) were captured in the waters where the dissolved oxygen level was greater than 6.9 mg/L, and 16 (51.6%) were captured in the waters with a depth less than 8.5 m.

The conditional inference classification tree was developed to estimate the probability of obtaining lake sturgeon bycatch under specific environmental and gillnet fishing conditions in (Figure 3-2). The tree was "grown" starting from basin at the root node, and then selected year, site depth, gear temperature and dissolved oxygen at the descendent internal nodes. This result indicated that basin (p<0.001) was the factor most strongly associated with the presence/absence of lake sturgeon bycatch in Lake Erie, and year (p<0.001), site depth (p<0.001), gear temperature (p<0.001), and dissolved oxygen (p=0.002) also had important impacts on the lake sturgeon bycatch. Seven terminal nodes were identified in the well-constructed tree. Overall, lake sturgeon bycatch was most likely (35.3% chance) to be observed in the waters in the west basin during the years 1994, 1998 and 2007, where the water temperature at gear set depth was less than 22.6 °C and the dissolved oxygen level was greater than 6.9 mg/L.

The two exhaustive search based classification trees showed different tree structures. The 'tree' tree (Figure 3-3) was grown from basin at the root node, whereas the 'rpart' tree (Figure 3-4) was rooted at site depth. Both trees contained ten terminal nodes, and selected year, mesh size, soak time and site depth at the internal nodes. Gear type was also included in the 'tree' tree at an internal node, and basin was also included in the 'rpart' tree.

The ‘tree’ tree (Figure 3-3) indicated a 100% chance to observe lake sturgeon bycatch when fishing by a bottomed gillnet with soak time greater than 22 hours in the waters deeper than 9.1 m in the west basin during the years 1994, 1998 and 2007. If gillnet fishing happened in waters shallower than 9.1 m in the west basin during the same years, and the gillnets with a mesh size of 70, 114, 140 and 152 mm were set at the bottom for more than 22 hours, there was still a 29% chance to get lake sturgeon bycatch.

Based on the ‘rpart’ tree (Figure 3-4), there was a 100% chance of obtaining lake sturgeon bycatch in the waters in the west basin where the water depth was greater than 9.1 m and when gillnets were soaked for more than 22.5 hours during the years 1994 and 2007. When fishing using a gillnet with a mesh size of 114 mm and deploying the net for more than 22.9 hours in the waters shallower than 8.3 m during the year 2002, there was still a 60% chance of catching lake sturgeon as bycatch.

The discriminative performance of the three classification trees on the training data was evaluated by calculating the AUC from a jackknife procedure (Table 3-3). The ‘tree’ tree produced the largest mean value of AUC (0.9745±0.0252), followed by the ‘rpart’ tree (0.9014±0.0039) and the ‘party’ tree (0.8908±0.0003). This result indicated that ‘tree’ tree had greater discriminative performance on the training data compared with the ‘party’ tree and the ‘rpart’ tree. The ROC curves and AUC values generated by combining the predicted value of each test data from each run of jackknifing across the whole dataset (Figure 3-5) implied that the ‘party’ tree (the conditional inference classification tree) performed best on the test data (AUC=0.77).

3.5. Discussion

In this study, I estimated the probability of catching lake sturgeon as bycatch under specific environmental and gillnet fishing conditions in Lake Erie using three classification tree models. All three tree models revealed that the west basin was a hotspot for lake sturgeon bycatch from gillnet fisheries, and most lake sturgeon caught by the PIS were juvenile fish with an average total length of 570 mm. Incidental capture of juvenile lake sturgeon in the west basin of Lake Erie has been documented (GLLSCM 2002). Relatively higher abundance of juvenile lake sturgeon in the west basin of Lake Erie likely can be attributed to better food

availability and more suitable habitat. Availability of prey largely determines the spatial and temporal distribution and abundance of juvenile lake sturgeon (Chiasson et al. 1997). Juvenile lake sturgeon feed primarily on benthic macro-invertebrates in substrate dominated by sand and clay, such as burrowing mayflies (Ephemeridae: *Hexagenia*) (McKinley et al. 1993; Chiasson et al. 1997; Beamish et al. 1998; McCabe et al. 2006). The west basin is the most productive basin in Lake Erie and supports a variety of benthic macro-invertebrates (Krieger et al. 1996).

The spawning migration of lake sturgeon during the summer and fall between Lake Huron, the St. Clair system, the Detroit River and Lake Erie has been documented (Thomas and Haas 2002). The St. Clair system and the Detroit River are known to support relatively abundant spawning populations of lake sturgeon (Thomas and Haas 1999; GLLSCM 2002; Thomas and Haas 2002; Caswell 2003). Tag and telemetry studies have shown upstream and downstream movement patterns where the lake sturgeon migrate up to Lake Huron or down to Lake Erie from their natal spawning sites located in the St. Clair system and the Detroit River (Thomas and Haas 2002; Caswell 2003; Boase 2003). However, the sources of the juvenile lake sturgeon caught in PIS are unclear without further research efforts.

None of the three classification trees detected the seasonal distribution characteristics of lake sturgeon bycatch in Lake Erie due to the data limitation in our study, although its seasonal migration patterns have been well documented (Thomas and Haas 2002; Caswell 2003). A downstream migration of lake sturgeon into Lake Erie during the summer and fall was observed in the study by Caswell (2003), and high complexity in lake sturgeon migration patterns has been suggested (Knights et al. 2002; Thomas and Haas 2002). With the PIS data collected in the fall rather than year-round data, it was impossible to characterize the seasonal distribution of lake sturgeon bycatch in this study.

I developed three classification tree models (Figures 3-2, 3-3 and 3-4) which exhibited different tree structures. The conditional inference tree (the 'party' tree) tended to select fewer variables at the internal nodes than the two exhaustive search based trees (the 'tree' tree and the 'rpart' tree), whereas the split values for the common variables were similar. For example, all three trees split the predictor basin into west basin and other basins, and split site depth at 8.5 m (the 'party' tree), 8.6 m (the 'tree' tree) or 8.3 m (the 'rpart' tree). We might prefer to

make inference regarding the importance of predictor variables based on the conditional inference tree rather than on the two exhaustive search based trees because the conditional inference tree selected variables and split nodes separately in a framework of a well-defined statistical hypothesis test that examined the association between the response variable and the predictor variables. This approach solved the overfitting and biased variable selection problems embedded in the exhaustive search based trees (Hothorn et al. 2006). Hothorn (2006) suggested that the conditional inference tree performed equivalently to the well-established exhaustive search based tree. However, the conditional inference classification tree showed less reliability on the training data based on AUC (Table 3-3) in this study. Different data structure and different judgment criteria used for making conclusions may have led to this disagreement. Although we recommended 'tree' tree on the training data and 'party' tree on the test data, model selection should be conducted on a case-by-case basis rather than by following a generalized rule.

Tree-based models receive a mixed reception from different research fields mostly because they capture the complex interactions in the data and the results can be interpreted easily with a visual diagram (Ribic and Ainley 1997; Fielding 1999; Harrell 2001; Gansky 2003; Austin 2008; Cutler et al. 2009). However, the largest drawback of tree-based models is the less accuracy compared with the commonly used generalized linear/additive based model approaches (Gansky 2003; Cutler et al. 2009). More accurate results can be obtained by combining a variety of suitably chosen trees, which is a method known as tree-based ensembles incorporating bagging and boosting techniques (Breiman 1996; Friedman et al. 2000; Hastie et al. 2001; Cutler et al. 2009). This approach could become a topic for further research about application of tree-based methods to fishery data analysis.

The potential influences of gillnet fisheries on the lake sturgeon bycatch in Lake Erie posed important implications for lake sturgeon conservation and management. The rare capture (31 individuals from 53,562 gillnets) of lake sturgeon indicated its low abundance in Lake Erie, and the capture of juvenile lake sturgeon may limit recruitment and impede population recovery. The influence of gillnet fisheries on the lake sturgeon recruitment should be taken into account in the lake sturgeon conservation and management.

The west basin of Lake Erie could be a hotspot for lake sturgeon bycatch. More attention should be drawn to gillnet fisheries management in the west basin with an emphasis on lake sturgeon conservation. Further research and monitoring programs of spawning and residency areas of lake sturgeon can provide information for more accurate analyses. Since lake sturgeon migrates between its spawning sites and residency sites, the inter-country and inter-agency cooperation would be essential to restore lake sturgeon populations and to improve lake sturgeon management.

3.6. Acknowledgement

This research was supported by the Department of Fisheries and Wildlife Sciences at Virginia Polytechnic Institute and State University, the USDA Cooperative State Research, Education and Extension Service through Hatch Project #0210510, and the Ontario Commercial Fisheries Association to Y. Jiao.

3.7. References

Auer, N. A. 1999. Chapter 17, Lake sturgeon: a unique and imperiled species in the Great Lakes. Page 515–536 in Great Lakes fisheries policy and management: a binational perspective, East Lansing, MI.

Austin, P. 2008. R and S-PLUS produced different classification trees for predicting patient mortality. Journal of Clinical Epidemiology 61:1222-1226.

Birstein, V., W. Bemis and J. Waldman. 1997. The threatened status of acipenseriform species: a summary. Page 427-435 in: Birstein, V., J. R. Waldman and W. E. Bemis (eds), Sturgeon Biodiversity and Conservation, Springer, NY.

Boase, J. C. 2003. Integrating sonic tracking and GIS to determine habitat selection and benthic prey distribution of adult lake sturgeon in Lake St. Clair. Master's thesis. University of Michigan, Ann Arbor.

Bogue, M. 2000. Fishing the Great Lakes: an environmental history, 1783-1933. Univ of Wisconsin Press, Madison.

Bradley, A. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30:1145-1159.

Breiman, L., J. Friedman, R. Olshen and C. Stone. 1984. Classification and regression trees. Wadsworth Publishing Company, Belmont, CA.

Breiman, L. 1996. Bagging predictors. Machine learning 24:123-140.

Caswell, N. M. 2003. Population characteristics, spawning sites, and movements of lake sturgeon (*Acipenser fulvescens*) in the Detroit River. Master's thesis. Central Michigan University, Mt. Pleasant, MI.

Canadian Engangered Species Conservation Council (CESCC), 2001. Wild Species 2000: The General Status of Species in Canada. Minister of Public Works and Government Services, Ottawa, Canada.

Chiasson, W., D. Noakes and F. Beamish. 1997. Habitat, benthic prey, and distribution of juvenile lake sturgeon (*Acipenser fulvescens*) in northern Ontario rivers. Canadian Journal of Fisheries and Aquatic Sciences 54:2866-2871.

Cutler, A., D. Cutler and J. Stevens. 2009. High-Dimensional Data Analysis in Cancer Research. Springer, New York.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. The Annals of Statistics 7:1-26.

Ferguson, M. and G. Duckworth. 1997. The status and distribution of lake sturgeon, *Acipenser fulvescens*, in the Canadian provinces of Manitoba, Ontario and Quebec: a genetic perspective. Environmental Biology of Fishes 48:299-309.

Fielding, A. H. 1999. Machine learning methods for ecological applications. Kluwer Academic Publisher, Norwell, Massachusetts.

Friedman, J., T. Hastie and R. Tibshirani. 2000. Additive logistic regression: A statistical view of boosting. The Annals of Statistics 28:337-374.

Gansky, S. 2003. Dental data mining: potential pitfalls and practical issues. Advances in Dental Research 17:109.

Great Lakes Lake Sturgeon Coordination Meeting (GLLSCM), 2002. Basin overview presentations of status and assessment activities. Presentations by N. Auer, H. Quinlan, M. Holtgren, R. Elliott, M. Thomas, E. Zollweg, A. Mathers, D. Carlson, Sault Ste, Marie, MI.

Great Lakes Lake Sturgeon Coordination Meeting (GLLSCM), 2005. Proceedings of the second (2004) Great Lakes Lake Sturgeon Coordination Meeting, Sault Ste. Marie, MI.

Great Lakes Lake Sturgeon Coordination Meeting (GLLSCM). 2010. Lake Sturgeon biology and population history in the Great Lakes. Available: http://www.fws.gov/midwest/sturgeon/biology.htm. (May 2010).

Hamley, J. 1975. Review of gillnet selectivity. Journal of the Fisheries Research Board of Canada 32:1944-1969.

Harrell, F. E. 2001. Regression modeling strategies. Springer, New York.

Hastie, T., R. Tibshirani and J. Friedman. 2001. The elements of statistical learning: data mining, inference and prediction, 2 nd edition. Springer, New York.

Hothorn, T., K. Hornik and A. Zeileis. 2006. Unbiased recursive partitioning. Journal of Computational and Graphical Statistics 15:651-674.

Johnson, J. E., J. L. Jonas and J. W. Peck. 2004. Management of Commercial Fisheries Bycatch, with Emphasis on Lake Trout Fisheries of the Upper Great Lakes. Fisheries Research Report, State of Michigan Department of Natural Resources, Michigan.

Kinnunen, R. 2003. Great lakes commercial fisheries. Report from Michigan Sea Grant, Michigan Sea Grant Extension, East Lansing, Michigan.

Knights, B., J. Vallazza, S. Zigler and M. Dewey. 2002. Habitat and movement of lake sturgeon in the upper Mississippi River system, USA. Transactions of the American Fisheries Society 131:507-522.

Krieger, K., D. Schloesser, B. Manny, C. Trisler, S. Heady, J. Ciborowski and K. Muth. 1996. Recovery of burrowing mayflies (Ephemeroptera: Ephemeridae: Hexagenia) in western Lake Erie. Journal of Great Lakes Research 22:254-263.

Lo, N., L. Jacobson and J. Squire. 1992. Indices of relative abundance from fish spotter data based on delta-lognormal models. Canadian Journal of Fisheries and Aquatic Sciences 49:2515-2526.

Maunder, M. and A. Langley. 2004. Integrating the standardization of catch-per-unit-of-effort into stock assessment models: testing a population dynamics model and using multiple data types. Fisheries Research 70:389-395.

Maunder, M. and A. Punt. 2004. Standardizing catch and effort data: a review of recent approaches. Fisheries Research 70:141-159.

McKinley, S., G. Van Der Kraak and G. Power. 1998. Seasonal migrations and reproductive patterns in the lake sturgeon, *Acipenser fulvescens*, in the vicinity of hydroelectric stations in northern Ontario. Environmental Biology of Fishes 51:245-256.

Meyer, J., C. Ingersoll, L. McDonald and M. Boyce. 1986. Estimating uncertainty in population growth rates: jackknife vs. bootstrap techniques. Ecology 67:1156-1166.

Miller, R. 1974. The jackknife--a review. Biometrika 61:1-15.

Noakes, D., F. Beamish and A. Rossiter. 1999. Conservation implications of behaviour and growth of the lake sturgeon, *Acipenser fulvescens*, in northern Ontario. Environmental Biology of Fishes 55:135-144.

OCFA (Ontario Commercial Fisheries Association). 2007. Lake Erie Partnership Index Fishing Survey [online]. Available: http: //www.ocfa.on.ca/Lake Erie Partnership Index Fishing Survey.htm. (September 2009).

Ortiz, M., C. Legault and N. Ehrhardt. 2000. An alternative method for estimating bycatch from the US shrimp trawl fishery in the Gulf of Mexico, 1972-1995. Fishery Bulletin 98:583-599.

Ribic, C. and D. Ainley. 1997. The relationships of seabird assemblages to physical habitat features in Pacific equatorial waters during spring 1984-1991. ICES Journal of Marine Science 54:593.

Scott, W. B. and E. J. Crossman. 1973. Freshwater fishes of Canada. Fisheries Research Board of Canada Bulletin 184, Ottawa, Ontario.

Shono, H. 2008. Application of the Tweedie distribution to zero-catch data in CPUE analysis. Fisheries Research 93:154-162.

Stefansson, G. 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES Journal of Marine Science 53:577.

Thomas, M. and R. Haas. 1999. Capture of lake sturgeon with setlines in the St. Clair River, Michigan. North American Journal of Fisheries Management 19:610-612.

Thomas, M. and R. Haas. 2002. Abundance, age structure, and spatial distribution of lake sturgeon, *Acipenser fulvescens*, in the St. Clair System. Journal of Applied Ichthyology 18:495-501.

Thomas, M. and R. Haas. 2005. Status of yellow perch and walleye populations in Michigan waters of Lake Erie, 1999–2003. Fisheries Research Report, 2082, Michigan Department of Natural Resources, Lansing, MI

Threader, R. W. and C. S. Broussaeu. 1986. Biology and management of the lake sturgeon in the Moose River, Ontario. North American Journal of Fisheries Management 6:383-390.

Welsh, A. 2004. Factors influencing the effectiveness of local versus national protection of migratory species: a case study of lake sturgeon in the Great Lakes, North America. Environmental Science and Policy 7:315-328.

Ye, Y., M. Al-Husaini and A. Al-Baz. 2001. Use of generalized linear models to analyze catch rates having zero values: the Kuwait driftnet fishery. Fisheries Research 53:151-168.

TABLE 3-1.—Predictor variables included in the classification tree models for lake sturgeon bycatch in Lake Erie. The data were collected from the Lake Erie Partnership Index Fishing Survey (PIS), 1989-2008.

Predictor variable	Type	Mean	Range
Site depth	Continuous	21.0 m	3.3-65.5 m
Gear depth[a]	Continuous	14.4 m	0.9-65.2 m
Secchi depth	Continuous	3.3 m	0.2-11.0 m
Gear temperature[b]	Continuous	16.8 °C	2.7-26.2 °C
Dissolved oxygen	Continuous	8.8 mg/L	0.2-21.4 mg/L
Soak time	Continuous	22.2 hour	9.5-35.9 hour
Basin	Categorical	west, west central, east central, east, Pennsylvania Ridge	
Year	Categorical	1989-2008	
Month	Categorical	August-November	
Gear type	Categorical	canned or bottomed	
Mesh size	Categorical	32-152 mm	

[a]: the water depth at which the gillnet was set.

[b]: the water temperature at the depth where the gillnet was set.

TABLE 3-2.—Observed lake stugeon bycatch (number) by basin, gear temperature, dissolved oxygen and site depth from the Lake Erie Partnership Index Fishing Survey (PIS), 1989-2008.

Gear temperature (C°)	Dissolved oxygen (mg/L)	West		West central		East central		East		Pennsylvania Ridge		Total in number
		<=8.5[a]	>8.5	<=8.5	>8.5	<=8.5	>8.5	<=8.5	>8.5	<=8.5	>8.5	
<=22.6	<=6.9	0	1	0	0	0	0	0	0	0	0	1
	>6.9	14	5	0	3	0	2	0	0	0	0	24
>22.6	<=6.9	0	0	0	0	0	0	0	0	0	0	0
	>6.9	2	4	0	0	0	0	0	0	0	0	6
Total		16	10	0	3	0	2	0	0	0	0	31

[a]: site depth (m).

TABLE 3-3.—The mean area under curve (AUC) with its standard deviation (sd) obtained from a jackknife procedure for each classification tree: party-the conditional inference classification tree generated by the R-package 'party'; tree-the exhaustive search based tree generated by the R-package 'tree'; rpart- the exhaustive search based tree generated by the R-package 'rpart'.

Model	Mean (sd)
party	0.8908 (0.0003)
tree	0.9745 (0.0252)
rpart	0.9014 (0.0039)

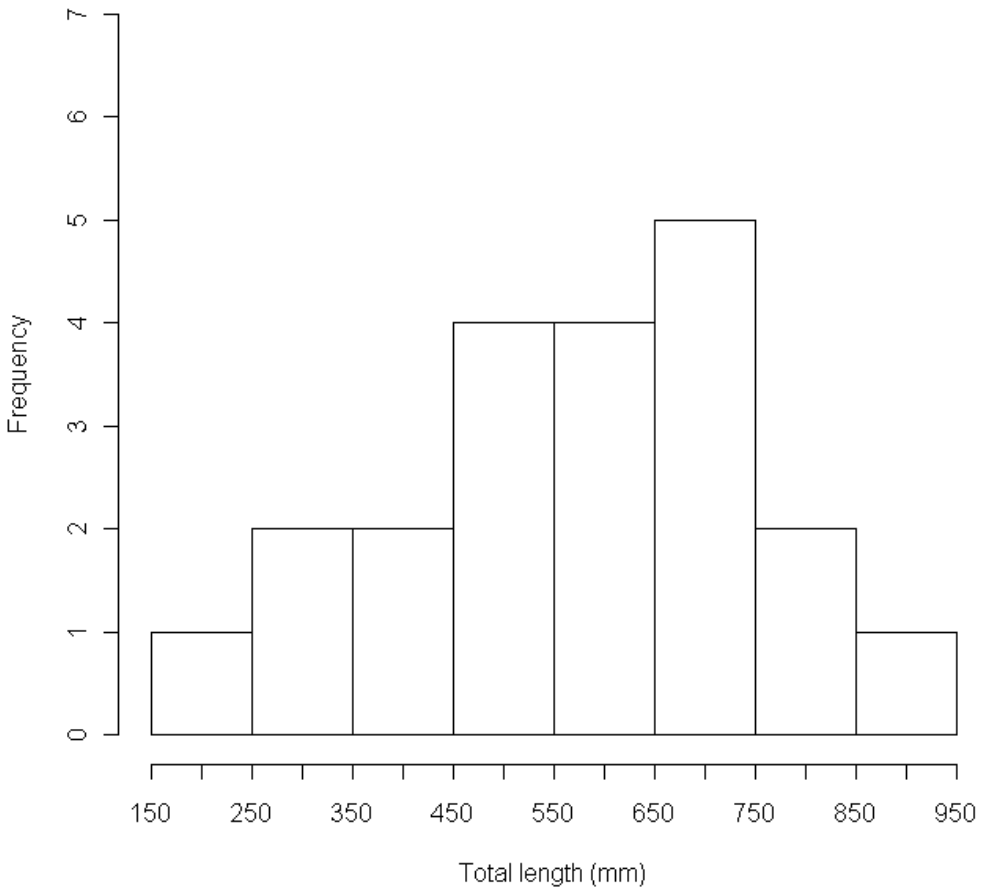

FIGURE 3-1.— Histogram of the total length (mm) of lake sturgeon bycatch collected from the Lake Erie Partnership Index Fishing Survey (PIS), 1989-2008. It was generated from 21 fish samples.

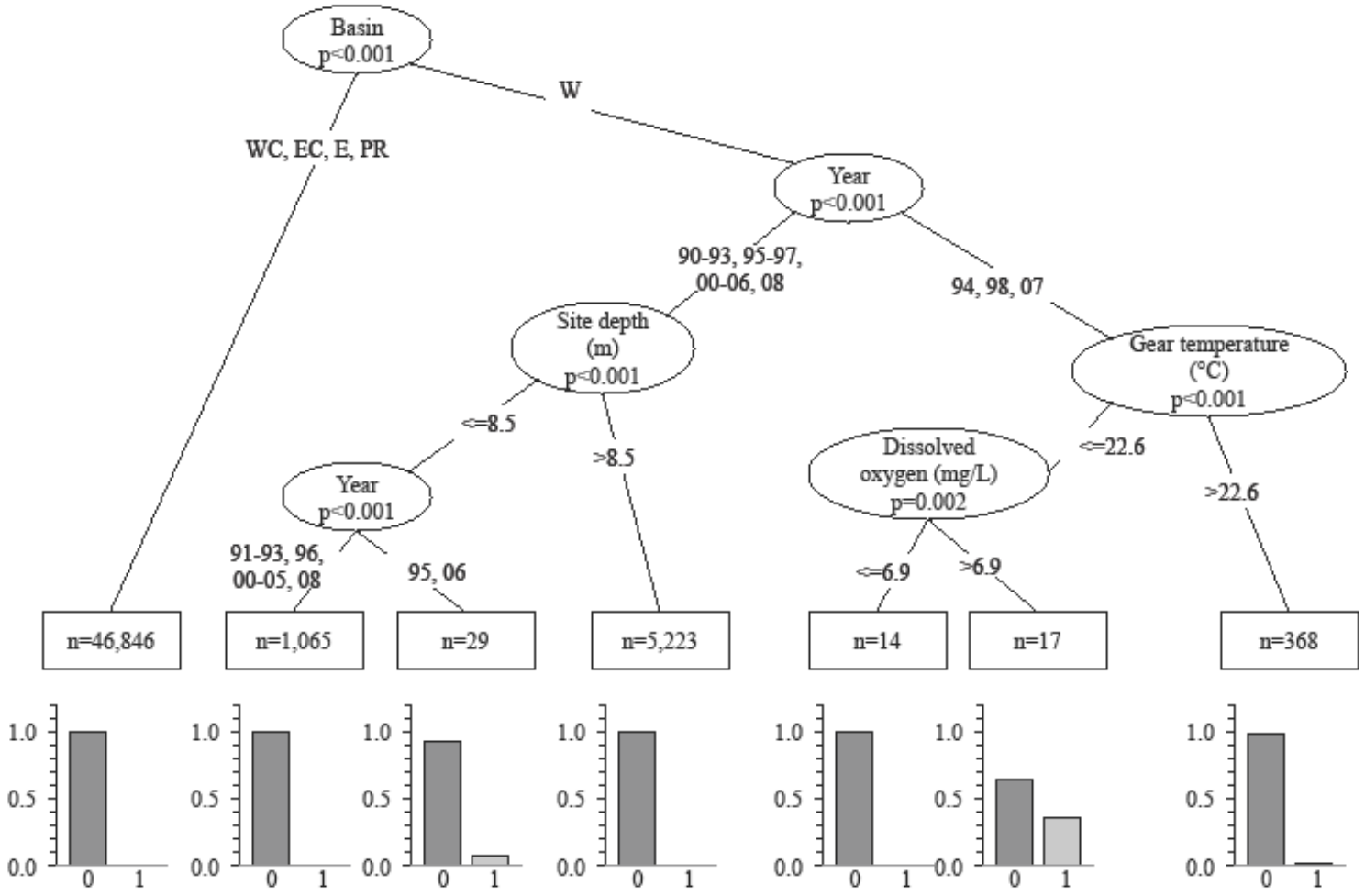
Basin
p<0.001
W
WC, EC, E, PR
Year
p<0.001
90-93, 95-97,
00-06, 08
94, 98, 07
Site depth
(m)
p<0.001
Gear temperature
(°C)
p<0.001
<=8.5
>8.5
<=22.6
>22.6
Year
p<0.001
Dissolved
oxygen (mg/L)
p=0.002
91-93, 96,
00-05, 08
95, 06
<=6.9
>6.9
n=46,846
n=1,065
n=29
n=5,223
n=14
n=17
n=368
1.0
0.5
0.0
0
1

FIGURE 3-2.—The conditional inference classification tree for lake sturgeon bycatch generated by the R-package 'party' using data from the Lake Erie Partnership Index Fishing Survey (PIS), 1989-2008. The internal nodes are denoted by ovals, and the terminal nodes are denoted by rectangles. At each terminal node, the number of observations (*n*) falling into this node is indicated in the rectangle, and the probability of obtaining lake sturgeon bycatch (1) or not (0) is presented in a bar chart. W-west basin, WC-west central basin, EC-East central basin, E-east basin, PR-Pennsylvania Ridge.

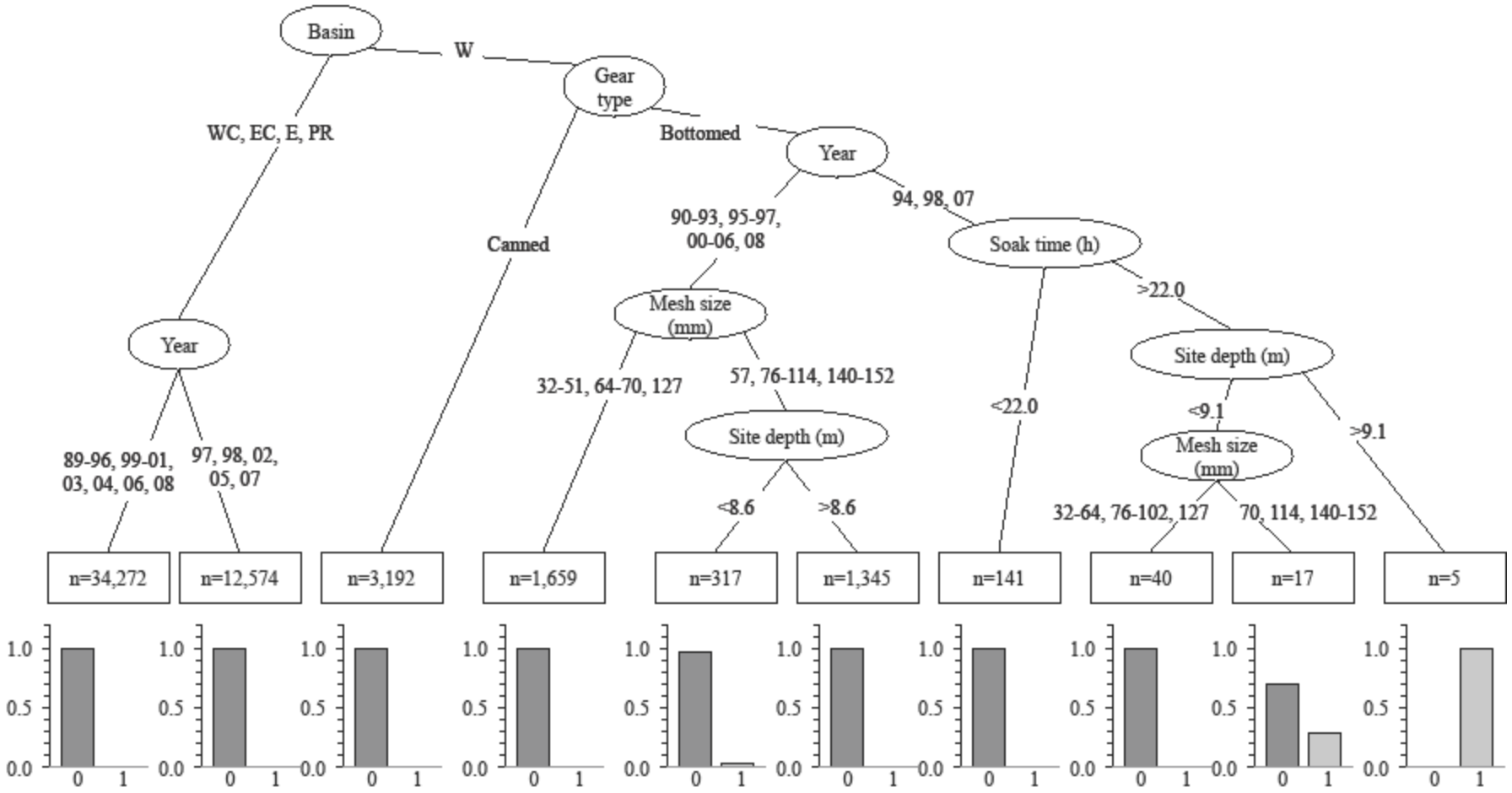
Basin
W
WC, EC, E, PR
Gear type
Canned
Bottomed
Year
90-93, 95-97, 00-06, 08
94, 98, 07
Mesh size (mm)
Soak time (h)
Year
32-51, 64-70, 127
57, 76-114, 140-152
<22.0
>22.0
Site depth (m)
89-96, 99-01, 03, 04, 06, 08
97, 98, 02, 05, 07
Site depth (m)
<9.1
>9.1
Mesh size (mm)
<8.6
>8.6
32-64, 76-102, 127
70, 114, 140-152
n=34,272
n=12,574
n=3,192
n=1,659
n=317
n=1,345
n=141
n=40
n=17
n=5
1.0
0.5
0.0
0
1

FIGURE 3-3.— The exhaustive search-based classification tree for lake sturgeon bycatch generated by the R-package 'tree' using the data from the Lake Erie Partnership Index Fishing Survey (PIS), 1989-2008. See Figure 3-2 for the explanations of internal and terminal nodes and the abbrevations for basins.

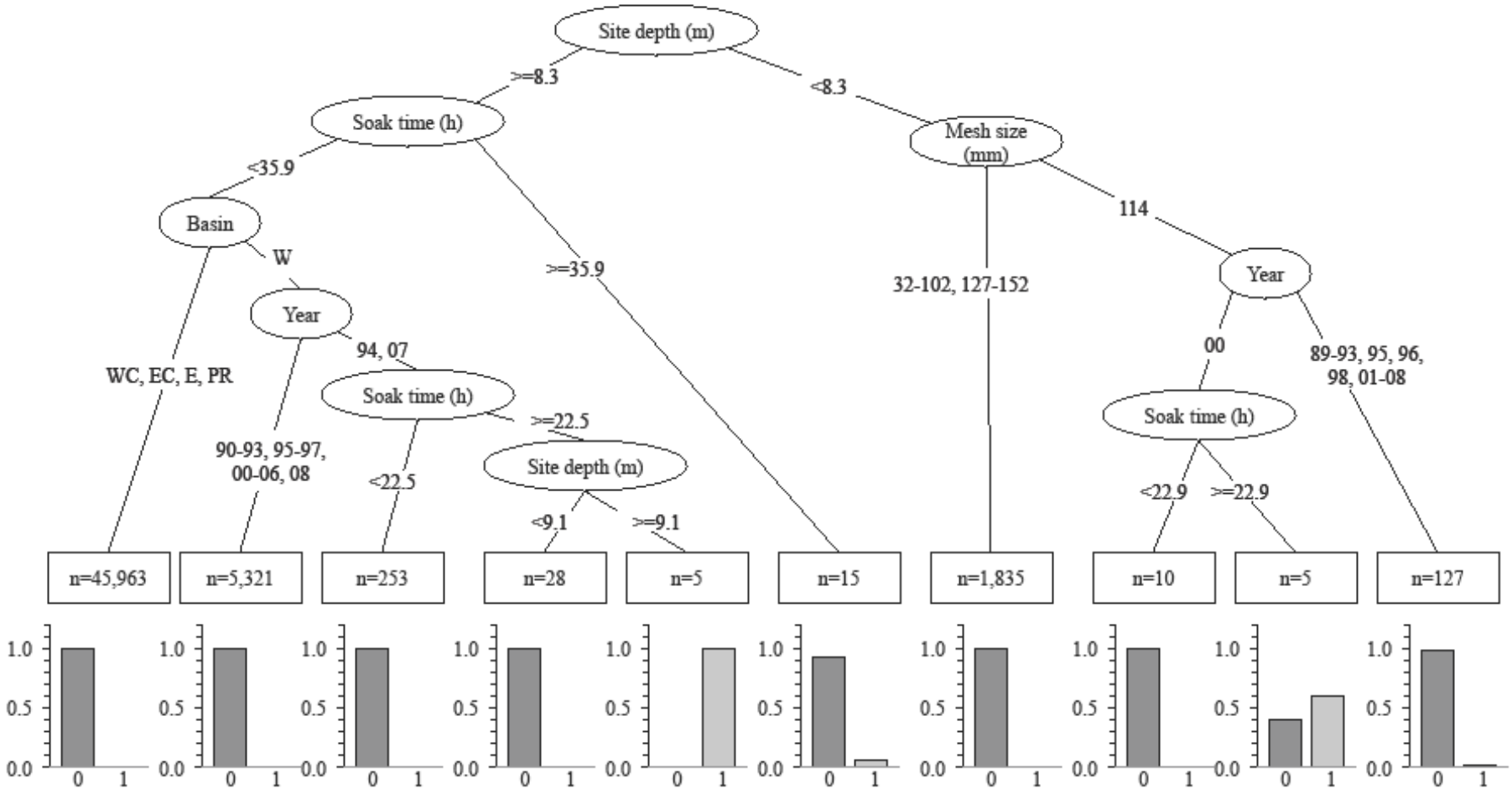
Site depth (m)
>=8.3
<8.3
Soak time (h)
Mesh size (mm)
<35.9
>=35.9
32-102, 127-152
114
Basin
W
Year
Year
WC, EC, E, PR
94, 07
00
89-93, 95, 96, 98, 01-08
Soak time (h)
Soak time (h)
90-93, 95-97, 00-06, 08
>=22.5
<22.5
Site depth (m)
<22.9
>=22.9
<9.1
>=9.1
n=45,963
n=5,321
n=253
n=28
n=5
n=15
n=1,835
n=10
n=5
n=127
1.0
0.5
0.0
0
1

FIGURE 3-4.— The exhaustive search-based classification tree for lake sturgeon bycath generated by the R-package ‘rpart’ using the data from the Lake Erie Partnership Index Fishing Survey (PIS), 1989-2008. See Figure 3-2 for the explanations of internal and terminal nodes and the abbrevations for basins.

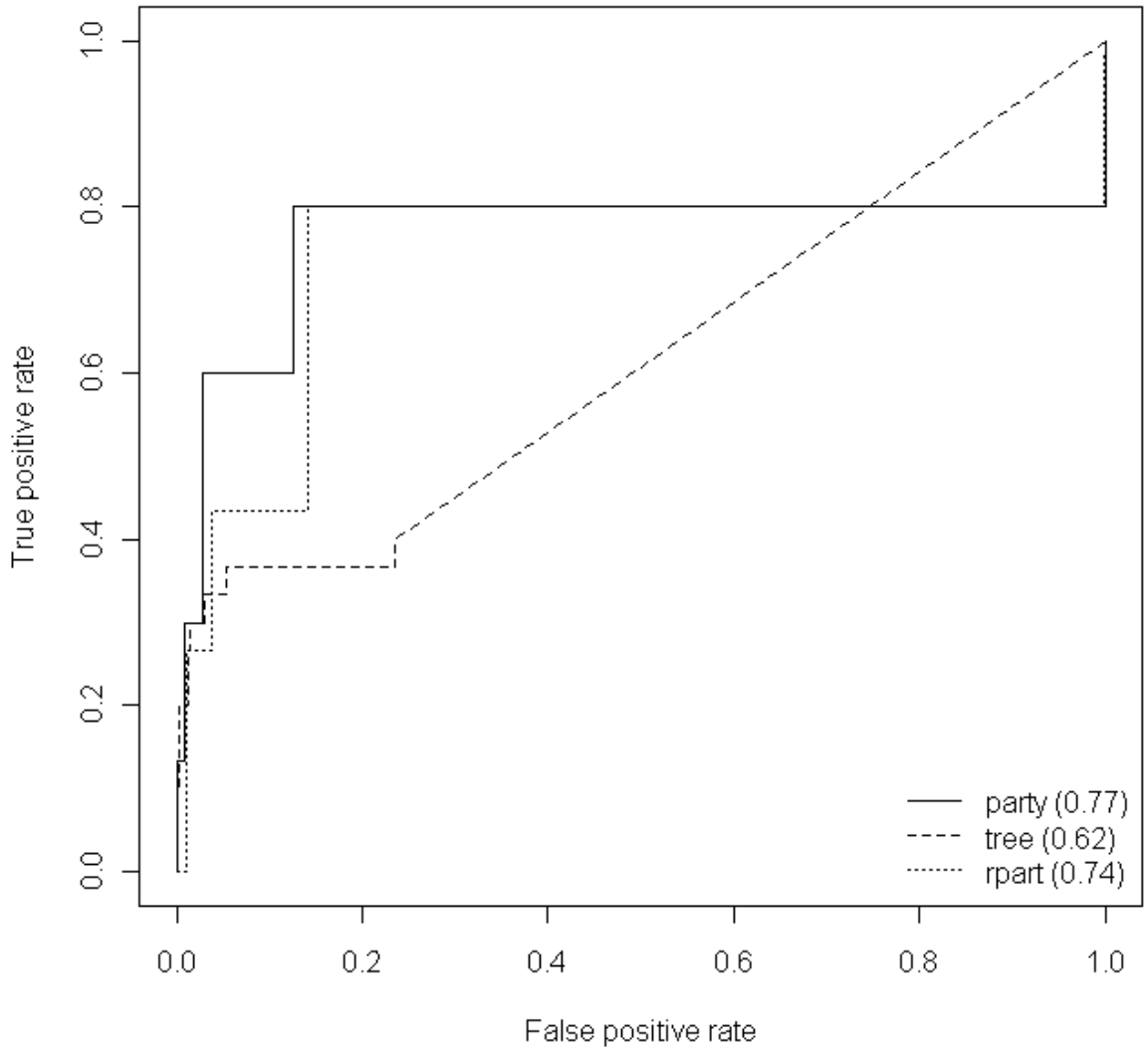

FIGURE 3-5.— Receiver Operating Characteristic (ROC) curves generated by combining the predicted value of each observation from each run of jackknifing across the whole dataset. The corresponding area under curve (AUC) is indicated in parenthesis: party-the conditional inference classification tree generated by the R-package 'party'; tree-the exhaustive search-based tree generated by the R-package 'tree'; rpart- the exhaustive search-based tree generated by the R-package 'rpart'.

Chapter 4

Decreasing uncertainty in catch rate analyses using Delta-AdaBoost: an alternative approach in catch and bycatch analyses with high percentage of zeros

4.1. Abstract

Gillnet data for walleye (*Sander vitreus*), yellow perch (*Perca flavescens*), and white perch (*Morone americana*) collected in a fishery-independent survey (the Lake Erie Partnership Index Fishing Survey, PIS) from 1989 to 2008 contained 75-83% of zero observations. The AdaBoost algorithm was applied to model analyses of fishery data for each species, and three- and five-fold cross-validation was conducted to evaluate the performance of each candidate model. The performance of the delta model consisting of one generalized additive model and one AdaBoost model (Delta-AdaBoost) was compared with five candidate models. The five candidate models included the delta model consisting of two generalized linear models (Delta-GLM), the delta model consisting of two generalized linear models with polynomial terms up to degree 3 (Delta-GLM-Poly), the delta model consisting of two generalized additive models (Delta-GAM), the generalized linear model with Tweedie distribution (GLM-Tweedie), and the generalized additive model with Tweedie distribution (GAM-Tweedie). To predict the presence/absence of fish species, the performance of the AdaBoost model was compared in terms of error rate with conventional generalized linear and additive models assuming a binomial distribution. Results from 3- and 5-fold cross-validation indicated that the Delta-AdaBoost model yielded the smallest training error (0.431-0.433 for walleye, 0.528-0.519 for yellow perch and 0.251 for white perch) and test error (0.435-0.436 for walleye, 0.524 for yellow perch and 0.254-0.255 for white perch) on average, followed by the Delta-GLM-Poly model for yellow perch and white perch, and the Delta-GAM model for walleye. In the prediction of the presence/absence of fish species, the AdaBoost model had the lowest error rate, compared with generalized linear and additive models. We suggested AdaBoost algorithm to be an alternative to deal with the high percentage of zero observations in catch and bycatch analyses in fisheries studies.

4.2. Introduction

Catch and bycatch rate estimations play an indispensable role in fish stock assessment and management (Gunderson 1993; Helser and Hayes 1995; Maunder and Punt 2004). Various methods have been developed to estimate the catch and bycatch rates for a specific fishery. The commonly used methods include the ratio method, which determines catch rates relative to a standard value (Beverton and Holt 1957); the generalized linear model, which incorporates multiple variables to describe environmental and operational effects on catch/bycatch (Gavaris 1980; Kimura 1981); and the generalized additive model, which demonstrates a nonlinear relationship between catch/bycatch rate and explanatory variables through a smooth function (Bigelow et al. 1999; Damalas et al. 2007). However, these methods have difficulties in dealing with highly skewed data with a high percentage of zero values. Such data are frequently encountered in fisheries of less abundant species and in bycatch analyses (Ortiz et al. 2000; Maunder and Punt 2004). The presence of zeros may invalidate the assumptions of normality underlying many statistical tests, and may cause computational difficulties.

Ignorance of a high proportion of zero observations may result in a loss of information that may reflect the spatial or temporal distribution characteristics of fish stocks. Two types of approaches have been applied in previous studies to deal with zero values in fishery data analyses. One approach is to add a small constant to each zero observation, followed by a generalized linear or additive model analysis (Ortiz et al. 2000; Maunder and Punt 2004; Shono 2008). However, the estimation results are sensitive to the choice of the constant (Ortiz et al. 2000; Maunder and Punt 2004). The other approach is to utilize the delta model and the Tweedie distribution model. In the delta model, the positive values are fitted by a generalized linear or additive model, and the probability of observing zero values of the response variable are fitted by a generalized linear or additive model with an assumption of binomial distribution (Lo et al. 1992; Stefansson 1996; Ye et al. 2001; Maunder and Langley 2004). Although combining two sub-models complicates model interpretation where the explanatory variables may differ, the delta model has been widely used to estimate bycatch (Ortiz et al. 2000; Murray 2004), catch rate, and abundance index (Lo et al. 1992; Stefansson 1996; Ye et al. 2001). By contrast, the Tweedie distribution model handles zero data uniformly along with the positive data, where the Tweedie distribution is considered to be a Poisson-Gamma compound distribution when its power parameter p is greater than 1 and less than 2 (Tweedie 1984; Shono 2008). The Tweedie

distribution model has been judged to outperform the generalized linear model with an additive constant and the delta model composed of two generalized linear models (Shono 2008).

AdaBoost is a typical boosting algorithm that was originally used for classification problems (Freund and Schapire 1996). The algorithm used for classification is called classifier. The final strong classifier is obtained by successively applying a classification algorithm to reweighted data and then combining a sequence of weak classifiers that minimize the prediction error at each iteration (Freund and Schapire 1996; Friedman et al. 2000; Hastie et al. 2001; Kawakita et al. 2005). In a fishery context, zeros and positive captures can be converted into a categorical variable $\{-1,1\}$, indicating the events of no fish caught and the events of at least one fish caught, respectively, and the analysis then can be treated as a two-group classification problem (Kawakita et al. 2005). This method has been used to predict the occurrence of large silky shark bycatch in a tuna purse-seine fishery, and the results confirmed the superiority of AdaBoost model in bycatch analyses where data were skewed by zeros (Kawakita et al. 2005).

The present study was conducted based on the gillnet data that were collected from a fishery-independent survey, the Lake Erie Partnership Index Fishing Survey (PIS). The PIS was primarily operated by the Ontario Ministry of Natural Resources (OMNR) and Ontario Commercial Fisheries Association (OCFA) since 1989. The experimental gillnets with mesh sizes ranging from 32 to 152 mm were deployed across the Ontario waters of Lake Erie in the fall (August-November) annually, using commercial fishing vessels and commercial fishing crews (OCFA 2007).

I focused on three species, walleye (*Sander vitreus*), yellow perch (*Perca flavescens*), and white perch (*Morone americana*). Walleye and yellow perch dominate the commercial gillnet fisheries in Lake Erie (Kinnunen 2003; Thomas and Haas 2005), and white perch has imposed considerable impacts on fish communities and the lake ecosystem as an invasive species (Scott and Crossman 1973; Schaeffer and Margraf 1987; Parrish and Margraf 1990).

In this study, the delta model composed of one generalized additive model and one AdaBoost model (Delta-AdaBoost) was developed to estimate the catch rates of walleye, yellow perch and white perch based on the PIS data from 1989 to 2008. The performance of the Delta-AdaBoost model was compared with five candidate models, including: the delta model consisting of two generalized linear models (Delta-GLM), the delta model consisting of two

generalized linear models with polynomial terms (Delta-GLM-Poly), the delta model consisting of two generalized additive models (Delta-GAM), the generalized linear model with Tweedie distribution (GLM-Tweedie), and the generalized additive model with Tweedie distribution (GAM-Tweedie). The performance of the AdaBoost model to predict the presence/absence of fish species was compared in terms of error rate with the generalized linear and additive model assuming a binomial distribution. Each model was evaluated through 3- and 5-fold cross-validation.

4.3. Methods

Data and variables.—I estimated the catch rates of walleye, yellow perch, and white perch using PIS data from 1989 to 2008 provided by OCFA. In total, 53,662 records were available for analysis and the catch rate was expressed as catch in weight (kg) per net (30.5m long $\times$ 1.8m deep). The PIS data included a high frequency of zero observations (75-83%) and as a result, the commonly used normal or lognormal distribution was violated (Ortiz et al. 2000).

In total, fourteen explanatory variables were available, including nine continuous variables (site depth, gear depth, secchi depth, gear temperature, dissolved oxygen, soak time, site temperature, longitude, and latitude) and five categorical variables (basin, year, month, gear type, and mesh size). Site temperature is the water surface temperature. Gear temperature means the water temperature at the gear set depth. Gear type refers to canned or bottomed gillnets. The correlation coefficients among all explanatory variables were examined first to detect those that were highly correlated. A preliminary stepwise selection based on Akaike Information Criterion (AIC, Akaike 1974) was conducted to eliminate one of the correlated pair of variables, i.e., the variable that yielded a larger AIC value was eliminated from the correlated pair. The remaining variables were selected through a stepwise procedure based on AIC (Akaike 1974; Burnham and Anderson 2002). The model with smaller AIC was considered to fit the data better. Interaction terms were not included in the regression model to avoid additional multicollinearity problems and difficulties in model interpretation (Maunder and Punt 2004; Damalas et al. 2007).

Delta model and Delta-Adaboost Model.— A delta model usually consists of two components, one model to fit the positive values and the other to estimate the probability of obtaining non-zero captures. Estimates of the catch rate from a delta model can be obtained by

multiplying these two components (Lo et al. 1992; Pennington 1996; Stefansson 1996; Ortiz et al. 2000; Ye et al. 2001; Maunder and Punt 2004; Murray 2004):

$$\hat{Catch\ rate} = \hat{d} \times \hat{q} ,$$

where $\hat{Catch\ rate}$ is the estimate of catch rate, $\hat{d}$ is the estimate of catch rate when only positive values of the response variable are analyzed, and $\hat{q}$ is the estimate of the probability of obtaining non-zero captures.

In a delta model, the model to fit the positive values could be a generalized linear model (Eq. 1), a generalized linear model with polynomial terms up to degree 3 (Eq. 2), or a generalized additive model (Eq. 3), which were built by assuming a lognormal distribution as follows:

$$\ln(\hat{d}) = \beta_0 + \sum_{j=1} \beta_j X_j , \qquad \text{Eq. 1}$$

$$\ln(\hat{d}) = \beta_0 + \sum_{j=1} (\beta_j^{'} X_j + \beta_j^{''} X_j^2 + \beta_j^{'''} X_j^3) , \qquad \text{Eq. 2}$$

$$\ln(\hat{d}) = \beta_0 + \sum_{j=1} f_j(X_j) , \qquad \text{Eq. 3}$$

where $\hat{d}$ is the estimate of catch rate when only positive values of the response variable are analyzed, β_0 is the intercept, β_j is the parameter for the *j*th explanatory variable X_j, and f_j is a smooth function (a spline or a loess smoother) for the *j*th explanatory variable X_j. The explanatory variables selected in each model could be different.

To estimate the probability of non-zero captures, values of 0 (no fish of interest caught) or 1 (at least one fish of interest caught) were treated as independent measurements from a Binary variable with a probability *q* of catching at least one fish. Similarly, the model to estimate the probability *q* could be a generalized linear model (Eq. 4), a generalized linear model with polynomial terms up to degree 3 (Eq. 5), or a generalized additive model (Eq. 6), which were constructed by assuming a binomial distribution as follows:

$$\ln(\frac{\hat{q}}{1-\hat{q}}) = \alpha_0 + \sum_{j=1} \alpha_j X_j , \qquad \text{Eq. 4}$$

$$\ln(\frac{\hat{q}}{1-\hat{q}}) = \alpha_0 + \sum_{j=1} (\alpha_j^{'} X_j + \alpha_j^{''} X_j^2 + \alpha_j^{'''} X_j^3) , \qquad \text{Eq. 5}$$

$$\ln(\frac{\hat{q}}{1-\hat{q}}) = \alpha_0 + \sum_{j=1} s_j(X_j), \qquad \text{Eq. 6}$$

where $\hat{q}$ is the estimate of the probability of obtaining non-zero captures, α_0 is the intercept, α_j is the parameter for the *j*th explanatory variable X_j, and s_j is a smooth function for the *j*th explanatory variable X_j.

In the delta model, the AdaBoost model can be applied as an alternative to the generalized linear or additive model to estimate the probability of obtaining non-zero captures. The vector of explanatory variables was denoted as *X*, and the input response variable as $Y \in \{-1,1\}$ where the value -1 represented zero captures and the value 1 represented non-zero captures. In the real AdaBoost algorithm, the classifier $g_t(x)$ returns a probability estimate at each iteration. The final classifier $F(x)$ was constructed as follows (Freund and Schapire 1996; Friedman et al. 2000; Hastie et al. 2001; Kawakita et al. 2005):

A. Initialize the weights $w_i = 1/N$, $i = 1,2,...,N$, where *N* is the number of observations.

B. For *t = 1* to *T*, where *T* is the number of iterations:

   (a) Fit the classifier $g_t(x)$ using the data weighted by w_i and obtain a probability estimate, $g_t(x_i) = \Pr(\hat{y}_i = 1 \mid x_i)$, i.e., the probability that the predicted value for y_i equals 1 given x_i.

   (b) Set $h_t(x_i) = \frac{1}{2}\ln(g_t(x_i)/(1-g_t(x_i)))$, which indicates the contribution of the classifier $g_t(x)$ to the final classifier $F(x)$.

   (c) Update the weights for the next iteration, $w_i = \frac{\exp(-y_i h_t(x_i))}{\sum_{i=1}^{N}\exp(-y_i h_t(x_i))}$,.

C. Set $H(x_i) = \sum_{t=1}^{T} h_t(x_i)$. The final classifier for the *i*th observation,

$$F(x_i) = \begin{cases} 1, \; if \;\; H(x_i) > 0; \\ -1, \; if \;\; H(x_i) < 0. \end{cases}$$

D. The probability of obtaining non-zero captures for the *i*th observation $\hat{q}_i = \frac{e^{2H(x_i)}}{1+e^{2H(x_i)}}$.

At iteration *t*, those observations that were misclassified at the previous iteration had their weights increased, whereas the weights were decreased for those classified correctly. As iterations proceeded, each classifier was forced to focus on those observations that were difficult to classify correctly. As a result of combining these classifiers, the final classifier provided accurate strong estimates, either the presence/absence of the fish species or the probability of obtaining non-zero captures. All the explanatory variables selected by the correlation analysis were included in AdaBoost model.

In this study, the Delta-AdaBoost model was constructed by combining a generalized additive model (Eq. 3) to fit the positive values and an AdaBoost model (T=1000) to estimate the probability of obtaining non-zero captures. Three other delta models were developed for comparison as follows: the delta model consisting of two generalized linear models (Delta-GLM) was constructed by combining Eq. 1 and Eq. 4; the delta model consisting of two generalized linear models with polynomial terms up to degree 3 (Delta-GLM-Poly) was constructed by combining Eq. 2 and Eq. 5; and the delta model consisting of two generalized additive models (Delta-GAM) was constructed by combining Eq. 3 and Eq. 6.

Tweedie distribution model.—In this study, two Tweedie distribution models were developed by assuming a Tweedie distribution in a generalized linear model (GLM-Tweedie, Eq. 7) and a generalized additive model (GAM-Tweedie, Eq. 8), respectively (Tweedie 1984; Shono 2008):

$$\widehat{Catch\ rate} = \gamma_0 + \sum_{j=1} \gamma_j X_j , \qquad \text{Eq. 7}$$

$$\widehat{Catch\ rate} = \gamma_0 + \sum_{j=1} m_j (X_j) , \qquad \text{Eq. 8}$$

where $\widehat{Catch\ rate}$ is the estimate of catch rate, γ_0 is the intercept, γ_j is the parameter for the *j*th explanatory variable X_j, and m_j is a smooth function for the *j*th explanatory variable X_j.

Tweedie distribution has been applied to handle zero values uniformly with positive values in the fishery data analysis, instead of separating zero values from positive values in delta models (Tweedie 1984; Shono 2008). The probability density function of Tweedie distribution is expressed as follows (Tweedie 1984; Shono 2008):

$$f(y:\mu,\sigma^2,p) = a(y:\sigma^2,p)\exp\{-\frac{1}{2\sigma^2} d(y:\mu,p)\},$$

where μ is the location parameter, σ^2 is the diffusion parameter, and p is the power parameter. When $1 < p < 2$, the Tweedie distribution actually can be treated as a Poisson-Gamma compound distribution, and can be expressed more explicitly as:

$$d(y:\mu,p) = 2\{\frac{\max(y,0)^{2-p}}{(1-p)(2-p)} - \frac{y\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p}\},$$

$$a(y:\sigma^2,p) = \{\frac{\sigma^{2(\alpha+1)}y^{\alpha}}{(1-p)^{\alpha}(2-p)}\}^{n} \frac{1}{n!\Gamma(n\alpha)y},$$

where $\alpha = (2-p)/(p-1)$, and $n = 1,2,...$

In the Tweedie distribution models, the power parameter p was determined by maximizing the log-likelihood in its likelihood profile (Shono 2008). In this study, the power parameter p was determined to be 1.37 for walleye and 1.45 for yellow perch and white perch (Figure 4-1).

Model evaluation.—Catch rate standardization was conducted for each species by these six models: Delta-GLM, Delta-GLM-Poly, Delta-GAM, GLM-Tweedie, GAM-Tweedie, and Delta-AdaBoost. The year effect was extracted relative to the means (for continuous variables) or weighted means (for categorical variables) of the other explanatory variables (Maunder and Punt 2004).

The performance of each model was evaluated using k-fold cross-validation (Hastie et al. 2001; Damalas et al. 2007; Shono 2008). The whole dataset was divided randomly into k sub-datasets with roughly equal size. Each subset was used as test data to predict from the model, and the remaining k-1 subsets were combined as training data to fit the model. The training error and test error for each model based on each pair of training and test data was calculated as:

$$Training(Test) \quad error = \frac{1}{N}\sum_{i=1}^{N} | y_i - \hat{y}_i |,$$

where N is the number of observations, y_i is the ith observation, and $\hat{y}_i$ is the estimated value (the predicted value) from the model for the ith observation.

I performed 3-fold (k=3) and 5-fold (k=5) cross-validation for Delta-AdaBoost model and each candidate model. The model providing lower training error and test error was judged as the one with better performance (Hastie et al. 2001; Damalas et al. 2007; Shono 2008).

To test the superiority of the AdaBoost model in estimating and predicting the presence/absence of fish species when a high proportion of zero captures was obtained, the performance of the AdaBoost model (T=1000) was compared with three candidate models, including the generalized linear model, GLM (Eq. 4), the generalized linear model with polynomial terms up to degree 3, GLM-Poly (Eq. 5), and the generalized additive model, GAM (Eq. 6). A binomial distribution was assumed in these three candidate models. Each model was evaluated through 3-fold and 5-fold cross-validations. The training and test errors in terms of error rates were defined as the proportion of the mis-classified or mis-predicted observations among the total observations in the dataset. All analyses in this study were programmed in R (Version 2.9.2).

4.4. Results

Correlation analysis (Table 4-1) among all the explanatory variables detected high correlations between site temperature and month (-0.84), longitude and latitude (-0.85), longitude and basin (-0.86), and latitude and basin (0.85) for all three species. Preliminary stepwise selection revealed that the model including month or basin yielded smaller AIC values than the model including site temperature, longitude, or latitude. Site temperature, longitude, and latitude were eliminated before a stepwise selection because of their high correlation with month or basin, and less prediction power. A stepwise selection was applied to the remaining eleven variables to construct the delta models and the Tweedie distribution models. The explanatory variables with significant effects (P<0.01) and a decreasing AIC value were successively selected into model (Table 4-2).

In the three delta models, the model to estimate the catch rates when only positive values were analyzed explained 19.6-21.4% of the deviance of PIS data for walleye (Table 4-2), 38.9-42.2% for yellow perch and 28.6-34.2% for white perch. The model to estimate the probability of non-zero captures explained 23.1-23.6% of the deviance of PIS data for walleye, 51.1-57.2% for yellow perch and 41.9-43.2% for white perch. The two Tweedie distribution models (GLM-Tweedie and GAM-Tweedie) explained 31.3-33.2% of the deviance of PIS data for walleye, 68.0-72.7% for yellow perch, and 58.3-61.0% for white perch (Table 4-2). Among these five candidate models, the Delta-GAM yielded the smallest AIC for yellow perch and white perch

when modeling the positive values and estimating the probability of non-zero captures. The Delta-GAM for walleye also gave the smallest AIC when modeling the positive values, while the Delta-GLM-Poly had the smallest AIC when estimating the probability of non-zeros. Catch rate standardization was conducted by the Delta-AdaBoost model and each candidate model based on the PIS data. Similar trends for standardized catch rates over time were observed among the Delta-AdaBoost model and the five candidate models (Figures 4-2, 4-3 and 4-4).

Model comparison was conducted between the Delta-AdaBoost model and the five candidate models through 3- and 5-fold cross-validation (Tables 4-3 and 4-4). The Delta-AdaBoost model yielded the smallest training error (0.431 by 3-fold cross-validation and 0.433 by 5-fold cross-validation for walleye, 0.528 by 3-fold cross-validation and 0.519 by 5-fold cross validation for yellow perch, and 0.251 by 3- and 5- fold cross-validations for white perch) and test error (0.435 by 3-fold cross-validation and 0.436 by 5-fold cross-validation for walleye, 0.524 by 3- and 5-fold cross-validations for yellow perch, and 0.254 by 3-fold cross-validation and 0.255 by 5-fold cross-validation for white perch) on average, followed by the Delta-GLM-Poly model for yellow perch and white perch, and the Delta-GAM model for walleye. These results indicated that the Delta-AdaBoost model provided more accurate estimation and prediction than the currently used generalized-linear/additive-based delta models in fisheries catch and bycatch data analyses when the percentage of zero observations was high.

The superiority of the AdaBoost model in estimating and predicting the presence/absence of fish species when the data contained a high percentage of zeros was further confirmed by comparing the error rates from the AdaBoost model and the three candidate models, i.e., GLM, GLM-Poly, and GAM (Figures 4-5, 4-6 and 4-7). As the number of iterations increased, the training and test error rates from the AdaBoost model declined monotonically for each species in both 3- and 5-fold cross-validation. After 1000 iterations, the AdaBoost model ended up with the lowest error rate, compared with the generalized linear and additive models.

4.5. Discussion

Our analyses provided evidence that the AdaBoost algorithm can be applied as an alternative approach to deal with fishery data containing a high frequency of zero observations. One of the favorable properties of the AdaBoost method in fishery data analyses is that it can

yield stable and accurate estimation and prediction without extra effort to remove the highly correlated explanatory variables (Kawakita et al. 2005). High correlation among explanatory variables may cause imprecision and inaccuracy in the conventional model analysis, such as generalized linear or additive models. We examined the correlation among all the available explanatory variables and eliminated one of the correlated pair of variables through a preliminary selection before constructing the models. Thus, the AdaBoost model applied here did not include any highly correlated variables. The advantage of the AdaBoost method regarding correlated variables can be tested through a sensitivity analysis in future studies, i.e., to compare the performance of the AdaBoost model both with and without correlated variables.

In the AdaBoost method, over-fitting may decrease prediction ability. As the number of iterations increases, training error decreases monotonically whereas the test error may behave differently. The optimal number of iterations can be obtained by continuously increasing the number of iterations until the test errors start to increase. We split the PIS data into two sub-datasets (i.e., the training data and the test data) with roughly equal size to determine the number of iterations for this study. We presented the results from the Delta-AdaBoost model and the AdaBoost model with 1000 iterations. We did not exhaustively search for the number of iterations where the test errors started to increase because: (1) the computing ability was limited, (2) both training and test errors decreased dramatically within the first 200 iterations and then stabilized when approaching 1000 iteratons (Figure 4-8), and (3) the prediction accuracy is not overly sensitive to the number of iterations if AdaBoost model proceeds to around the optimal iteration number (Kawakita et al. 2005).

Developing models to deal with data having a high frequency of zeros has become a methodological concern in fishery studies, especially in catch and bycatch analyses of rare species. Shono (2008) compared the Tweedie distribution model and the delta model consisting of two generalized linear models and concluded that the Tweedie distribution model performed the best when the proportion of zeros in the data was greater than 80%. However, the results from 3- and 5-fold cross-validation for each species in this study indicated that the Tweedie distribution model (GLM-Tweedie and GAM-Tweedie) did not performed as well (Tables 4-3 and 4-4) as the delta model composed of two generalized linear models (Delta-GLM) with the PIS data, although the proportion of zeros in the data was similar to that in Shono (2008), around

75-83%. This outcome indicated that the percentage of zeros in the data was not the only factor that affected the selection of models. A sensitivity analysis to test the influence of the datasets with different percentages of zeros and different structures on the selection of models would be an extension of this study.

Generalized additive models are often preferred in fishery analyses because of their superiority in describing the relationship between fish captures and environmental factors, which is most likely to be nonlinear in a biological context (Damalas et al. 2007). We observed that the generalized linear model with polynomial terms up to degree 3 (Delta-GLM-Poly) performed as well as the generalized additive model (Delta-GAM) for walleye, and slightly better for yellow perch and white perch (Tables 4-3 and 4-4). This result indicated that when the computation of generalized additive models became complicated, generalized linear models with polynomial terms could be an alternative.

In conclusion, when the percentage of zero observations in the data is high, the delta model consisting of a generalized additive model and an AdaBoost model (Delta-AdaBoost) can be an alternative due to its high performance in the 3- and 5-fold cross-validation. The AdaBoost algorithm to estimate and predict the presence/absence of fish species yielded more accurate and stable results, and thus it is an alternative to generalized linear and additive models with an assumption of binomial distribution. Model selection should be conducted on a case-by-case basis since it can be confounded with several factors including data structure.

4.6. Acknowledgement

This research was supported by the Department of Fisheries and Wildlife Sciences at Virginia Polytechnic Institute and State University, the USDA Cooperative State Research, Education and Extension Service through Hatch Project #0210510, and the Ontario Commercial Fisheries Association to Y. Jiao. I would like to acknowledge Qing He for reviewing the method session.

4.7. References

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716-723.

Beverton, R. and S. Holt. 1957. On the dynamics of exploited fish populations. Fishery Investigations, Series II. Marine Fisheries, Great Britain Ministry of Agriculture, Fisheries and Food 19:533.

Bigelow, K., C. Boggs and X. He. 1999. Environmental effects on swordfish and blue shark catch rates in the US North Pacific longline fishery. Fisheries Oceanography 8:178-198.

Burnham, K. and D. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2 nd edition. Springer Verlag, New York.

Damalas, D., P. Megalofonou and M. Apostolopoulou. 2007. Environmental, spatial, temporal and operational effects on swordfish (*Xiphias gladius*) catch rates of eastern Mediterranean Sea longline fisheries. Fisheries Research 84:233-246.

Freund, Y. and R. Schapire. 1996. Experiments with a new boosting algorithm. Page 148-156 in: Saitta, L. (ed), Machine Learning: Proceedings of the Thirteenth International Conference, Bari, Italy.

Friedman, J., T. Hastie and R. Tibshirani. 2000. Additive logistic regression: A statistical view of boosting. The Annals of Statistics 28:337-374.

Gavaris, S. 1980. Use of a multiplicative model to estimate catch rate and effort from commercial data. Canadian Journal of Fisheries and Aquatic Sciences 37:2272-2275.

Gunderson, D. 1993. Surveys of fisheries resources. John Wiley and Sons Inc, New York.

Hastie, T., R. Tibshirani and J. Friedman. 2001. The elements of statistical learning: data mining, inference and prediction., 2nd edition. Springer, New York.

Helser, T. and D. Hayes. 1995. Providing quantitative management advice from stock abundance indices based on research surveys. Fishery Bulletin 93:290-298.

Kawakita, M., M. Minami, S. Eguchi and C. Lennert-Cody. 2005. An introduction to the predictive technique AdaBoost with a comparison to generalized additive models. Fisheries Research 76:328-343.

Kimura, D. 1981. Standardized measures of relative abundance based on modelling log (cpue), and their application to Pacific ocean perch (*Sebastes alutus*). ICES Journal of Marine Science 39:211.

Kinnunen, R. 2003. Great lakes commercial fisheries. Report from Michigan Sea Grant, Michigan Sea Grant Extension, East Lansing, Michigan.

Lo, N., L. Jacobson and J. Squire. 1992. Indices of relative abundance from fish spotter data based on delta-lognormal models. Canadian Journal of Fisheries and Aquatic Sciences 49:2515-2526.

Maunder, M. and A. Langley. 2004. Integrating the standardization of catch-per-unit-of-effort into stock assessment models: testing a population dynamics model and using multiple data types. Fisheries Research 70:389-395.

Maunder, M. and A. Punt. 2004. Standardizing catch and effort data: a review of recent approaches. Fisheries Research 70:141-159.

Murray, K. 2004. Magnitude and distribution of sea turtle bycatch in the sea scallop (*Placopecten magellanicus*) dredge fishery in two areas of the Northwestern Atlantic Ocean, 2001-2002. Fishery Bulletin 102:671-681.

OCFA (Ontario Commercial Fisheries Association). 2007. Lake Erie Partnership Index Fishing Survey [online]. Available: http: //www.ocfa.on.ca/Lake Erie Partnership Index Fishing Survey.htm. (September 2009).

Ortiz, M., C. Legault and N. Ehrhardt. 2000. An alternative method for estimating bycatch from the US shrimp trawl fishery in the Gulf of Mexico, 1972-1995. Fishery Bulletin 98:583-599.

Parrish, D. and F. Margraf. 1990. Interactions between white perch (*Morone americana*) and yellow perch (*Perca flavescens*) in Lake Erie as determined from feeding and growth. Canadian Journal of Fisheries and Aquatic Sciences 47:1779-1787.

Pennington, M. 1996. Estimating the mean and variance from highly skewed marine data. Fishery Bulletin 94:498-505.

Schaeffer, J. and F. Margraf. 1987. Predation on fish eggs by white perch, *Morone americana*, in western Lake Erie. Environmental Biology of Fishes 18:77-80.

Scott, W. B. and E. J. Crossman. 1973. Freshwater fishes of Canada. Fisheries Research Board of Canada Bulletin 184, Ottawa, Ontario.

Shono, H. 2008. Application of the Tweedie distribution to zero-catch data in CPUE analysis. Fisheries Research 93:154-162.

Stefansson, G. 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES Journal of Marine Science 53:577.

Thomas, M. and R. Haas. 2005. Status of yellow perch and walleye populations in Michigan waters of Lake Erie, 1999–2003. Fisheries Research Report, 2082, Michigan Department of Natural Resources, Lansing, MI.

Ye, Y., M. Al-Husaini and A. Al-Baz. 2001. Use of generalized linear models to analyze catch rates having zero values: the Kuwait driftnet fishery. Fisheries Research 53:151-168.

TABLE 4-1.—Spearman correlation coefficients among the explanatory variables based on the data from the Lake Erie Partnership Index Fishing Survey (PIS), 1989-2003. 1-site depth, 2-gear depth, 3-secchi depth, 4-gear temperature, 5-dissolved oxygen, 6-site temperature, 7-longitude, 8-latitude, 9-soak time, 10- basin, 11-year, 12-month, 13-gear type, 14-mesh size.

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1.00													
2	0.57	1.00												
3	0.42	0.29	1.00											
4	-0.25	-0.35	0.07	1.00										
5	0.01	-0.08	-0.21	-0.35	1.00									
6	0.12	0.08	0.44	0.58	-0.48	1.00								
7	-0.48	-0.31	-0.54	0.13	0.11	-0.25	1.00							
8	0.28	0.21	0.51	-0.12	-0.12	0.21	-0.85	1.00						
9	-0.02	-0.04	-0.02	0	0.04	-0.03	0.03	-0.05	1.00					
10	0.48	0.33	0.59	-0.10	-0.13	0.33	-0.86	0.85	-0.01	1.00				
11	0	0.07	-0.13	0.06	0.05	0.15	-0.02	-0.02	-0.10	0.03	1.00			
12	-0.13	-0.09	-0.47	-0.43	0.44	-0.84	0.38	-0.35	0.01	-0.49	-0.04	1.00		
13	-0.09	0.59	-0.04	-0.20	-0.04	-0.03	0.07	-0.06	0.01	-0.07	-0.06	0.02	1.00	
14	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00

TABLE 4-2.—Stepwise model building based on Akaike Information Criterion (AIC) for the five candidate models fitted to the data from the Lake Erie Partnership Index Survey (PIS), 1989-2008. The five candidate models included: the delta model consisting of two generalized linear models (Delta-GLM), the delta model consisting of two generalized linear models with polynomial terms up to degree 3 (Delta-GLM-Poly), the delta model consisting of two generalized additive models (Delta-GAM), the generalized linear model with Tweedie distribution (GLM-Tweedie), and the generalized additive model with Tweedie distribution (GAM-Tweedie). In the delta models, a lognormal distribution was assumed when estimating the catch rates with only positive values analyzed, and a binomial distribution was assumed when estimating the probability of obtaining non-zero captures. Models contained the explanatory variables marked with '√'.

Variables	Delta-GLM		Delta-GLM-Poly		Delta-GAM		GLM-Tweedie	GAM-Tweedie
	POS[a]	PROB[b]	POS	PROB	POS	PROB		
Walleye								
Site depth	√		√	√	√		√	√
Gear depth		√	√	√		√	√	√
Secchi depth	√	√	√	√	√	√	√	√
Gear temperature	√	√	√	√	√	√	√	√
Dissolved oxygen				√			√	√
Soak time	√	√	√	√	√	√	√	√
Basin	√	√	√	√	√	√	√	√
Year	√	√	√	√	√	√	√	√
Month	√	√	√	√	√	√	√	√
Gear type	√	√	√	√	√	√	√	√
Mesh size	√	√	√	√	√	√	√	√
AIC	39,323	38,029	39,219	37,816	39,132	37,821	83,924	83,488
% explained[c]	19.6	23.1	21.4	23.6	19.6	23.1	31.3	33.2
Yellow perch								

Site depth	√	√	√	√	√	√	√	√
Gear depth	√	√	√	√	√	√	√	√
Secchi depth	√	√	√	√	√	√	√	√
Gear temperature	√	√	√	√	√	√	√	√
Dissolved oxygen		√	√	√		√	√	√
Soak time	√	√	√	√	√	√	√	√
Basin	√	√	√	√	√	√	√	√
Year	√	√	√	√	√	√	√	√
Month	√	√	√	√	√	√	√	√
Gear type	√	√	√	√	√	√	√	√
Mesh size	√	√	√	√	√	√	√	√
AIC	70,858	29,487	70,144	25,831	70,118	25,639	88,108	84,783
% explained[c]	38.9	51.1	42.2	57.2	38.9	51.1	68.0	72.7
White perch								
Site depth	√	√	√	√	√	√	√	√
Gear depth	√	√	√	√	√	√	√	√
Secchi depth	√	√	√	√	√	√	√	√
Gear temperature	√	√	√	√	√	√	√	√
Dissolved oxygen	√		√	√	√		√	√
Soak time	√		√	√	√		√	√
Basin	√	√	√	√	√	√	√	√
Year	√	√	√	√	√	√	√	√
Month	√	√	√	√	√	√	√	√
Gear type	√	√	√	√	√	√	√	√
Mesh size	√	√	√	√	√	√	√	√
AIC	53,422	34,621	52,377	33,849	51,624	33,566	71,072	69,642
% explained[c]	28.6	41.9	34.2	43.2	28.6	41.9	58.3	61.0

[a]: the sub-model in the delta model to estimate the catch rates when only positive values of the response variable were analyzed.

[b]: the sub-model in the delta model to estimate the probability of obtaining non-zero captures.

[c]: cumulative % of deviance explained by the model.

TABLE 4-3.—Training and test errors from the Delta-AdaBoost model and the five candidate models by 5-fold cross-validation. The Delta-AdaBoost model consisted of one generalized additive model and one AdaBoost model (Delta-AdaBoost). See Table 4-2 for the explanation of models.

	Training error						Test error					
Model	1	2	3	4	5	Average	1	2	3	4	5	Average
Walleye												
Delta-GLM	0.467	0.465	0.459	0.465	0.466	0.464	0.460	0.463	0.475	0.467	0.462	0.465
Delta-GLM-Poly	0.463	0.461	0.454	0.460	0.461	0.460	0.456	0.458	0.469	0.464	0.460	0.461
Delta-GAM	0.461	0.459	0.452	0.459	0.459	0.458	0.455	0.457	0.468	0.461	0.457	0.460
GLM-Tweedie	0.478	0.476	0.471	0.478	0.477	0.476	0.473	0.474	0.486	0.478	0.474	0.477
GAM-Tweedie	0.470	0.468	0.463	0.470	0.469	0.468	0.464	0.465	0.477	0.472	0.469	0.469
Delta-AdaBoost	0.437	0.435	0.426	0.434	0.435	0.433	0.431	0.435	0.445	0.435	0.435	0.436
Yellow perch												
Delta-GLM	0.564	0.560	0.567	0.565	0.554	0.562	0.554	0.594	0.540	0.548	0.591	0.565
Delta-GLM-Poly	0.526	0.517	0.526	0.526	0.513	0.522	0.519	0.546	0.504	0.508	0.550	0.526
Delta-GAM	0.530	0.523	0.530	0.531	0.518	0.527	0.525	0.552	0.511	0.511	0.557	0.531
GLM-Tweedie	0.598	0.591	0.601	0.602	0.589	0.596	0.579	0.620	0.582	0.582	0.625	0.598
GAM-Tweedie	0.558	0.551	0.561	0.563	0.550	0.557	0.548	0.575	0.545	0.540	0.586	0.559
Delta-AdaBoost	0.522	0.516	0.523	0.523	0.511	0.519	0.517	0.545	0.505	0.503	0.550	0.524
White perch												
Delta-GLM	0.280	0.284	0.281	0.278	0.277	0.280	0.292	0.295	0.272	0.283	0.272	0.283
Delta-GLM-Poly	0.258	0.257	0.256	0.253	0.251	0.255	0.268	0.266	0.245	0.256	0.253	0.258
Delta-GAM	0.259	0.261	0.260	0.258	0.256	0.259	0.269	0.268	0.250	0.263	0.258	0.262
GLM-Tweedie	0.298	0.299	0.305	0.301	0.302	0.301	0.309	0.311	0.289	0.306	0.296	0.302
GAM-Tweedie	0.284	0.285	0.290	0.286	0.287	0.286	0.294	0.296	0.273	0.294	0.282	0.288

Delta-AdaBoost	0.252	0.253	0.253	0.250	0.249	0.251	0.261	0.261	0.246	0.257	0.251	0.255

TABLE 4-4.—Training and test errors from the Delta-AdaBoost model and the five candidate models by 3-fold cross-validation. The Delta-AdaBoost model consisted of one generalized additive model and one AdaBoost model (Delta-AdaBoost). See Table 4-2 for the explanation of models.

	Training error				Test error			
Model	1	2	3	Average	1	2	3	Average
Walleye								
Delta-GLM	0.458	0.465	0.470	0.464	0.475	0.464	0.458	0.466
Delta-GLM-Poly	0.453	0.460	0.465	0.459	0.470	0.461	0.452	0.461
Delta-GAM	0.452	0.458	0.464	0.458	0.469	0.459	0.452	0.460
GLM-Tweedie	0.469	0.478	0.481	0.476	0.487	0.476	0.469	0.477
GAM-Tweedie	0.461	0.469	0.473	0.468	0.478	0.468	0.461	0.469
Delta-AdaBoost	0.425	0.430	0.439	0.431	0.445	0.431	0.429	0.435
Yellow perch								
Delta-GLM	0.566	0.564	0.555	0.562	0.565	0.560	0.569	0.565
Delta-GLM-Poly	0.523	0.525	0.516	0.521	0.525	0.523	0.529	0.525
Delta-GAM	0.529	0.528	0.521	0.526	0.529	0.527	0.537	0.531
GLM-Tweedie	0.599	0.599	0.591	0.596	0.594	0.597	0.602	0.598
GAM-Tweedie	0.560	0.558	0.552	0.556	0.553	0.562	0.562	0.559
Delta-AdaBoost	0.520	0.520	0.514	0.518	0.522	0.520	0.530	0.524
White perch								
Delta-GLM	0.283	0.278	0.278	0.280	0.281	0.282	0.283	0.282
Delta-GLM-Poly	0.258	0.243	0.253	0.252	0.262	0.246	0.259	0.256
Delta-GAM	0.261	0.256	0.258	0.258	0.263	0.258	0.261	0.261
GLM-Tweedie	0.304	0.302	0.298	0.301	0.302	0.303	0.301	0.302
GAM-Tweedie	0.288	0.287	0.284	0.287	0.288	0.288	0.287	0.288
Delta-AdaBoost	0.253	0.249	0.250	0.251	0.255	0.253	0.254	0.254

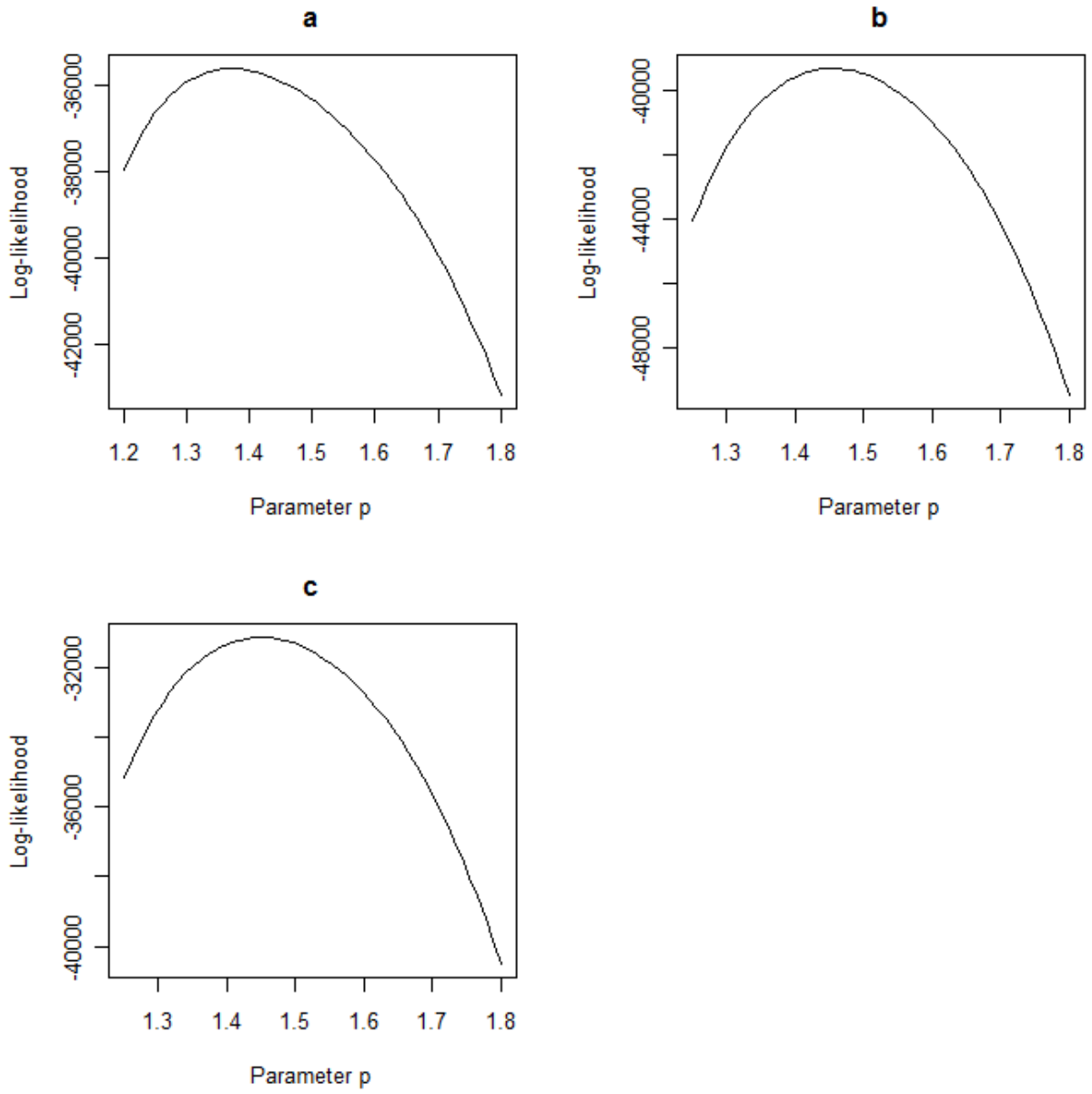

FIGURE 4-1.— Log-likelihood profiles demonstrating the corresponding log-likelihoods given different power prameter *p* in the Tweedie distribuion models for walleye (a), yellow perch (b) and white perch (c).

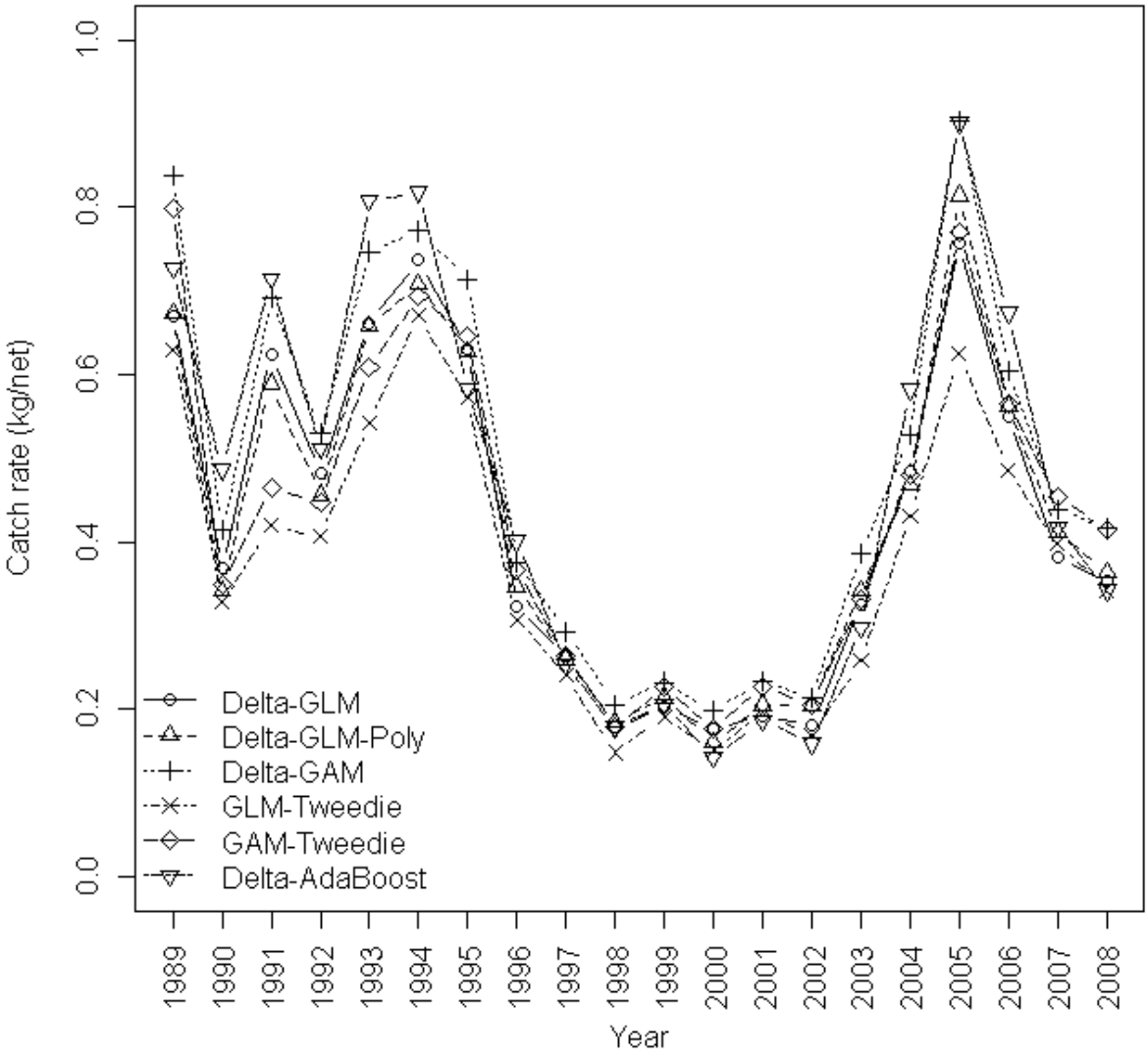

FIGURE 4-2.— Trends of the standardized catch rates (kg/net, 30.5m long × 1.8m deep) for walleye over time generated by the Delta-AdaBoost model and the five candidate models. The Delta-AdaBoost model consisted of one generalized additive model and one AdaBoost model (Delta-AdaBoost). See Table 4-2 for the explanation of models.

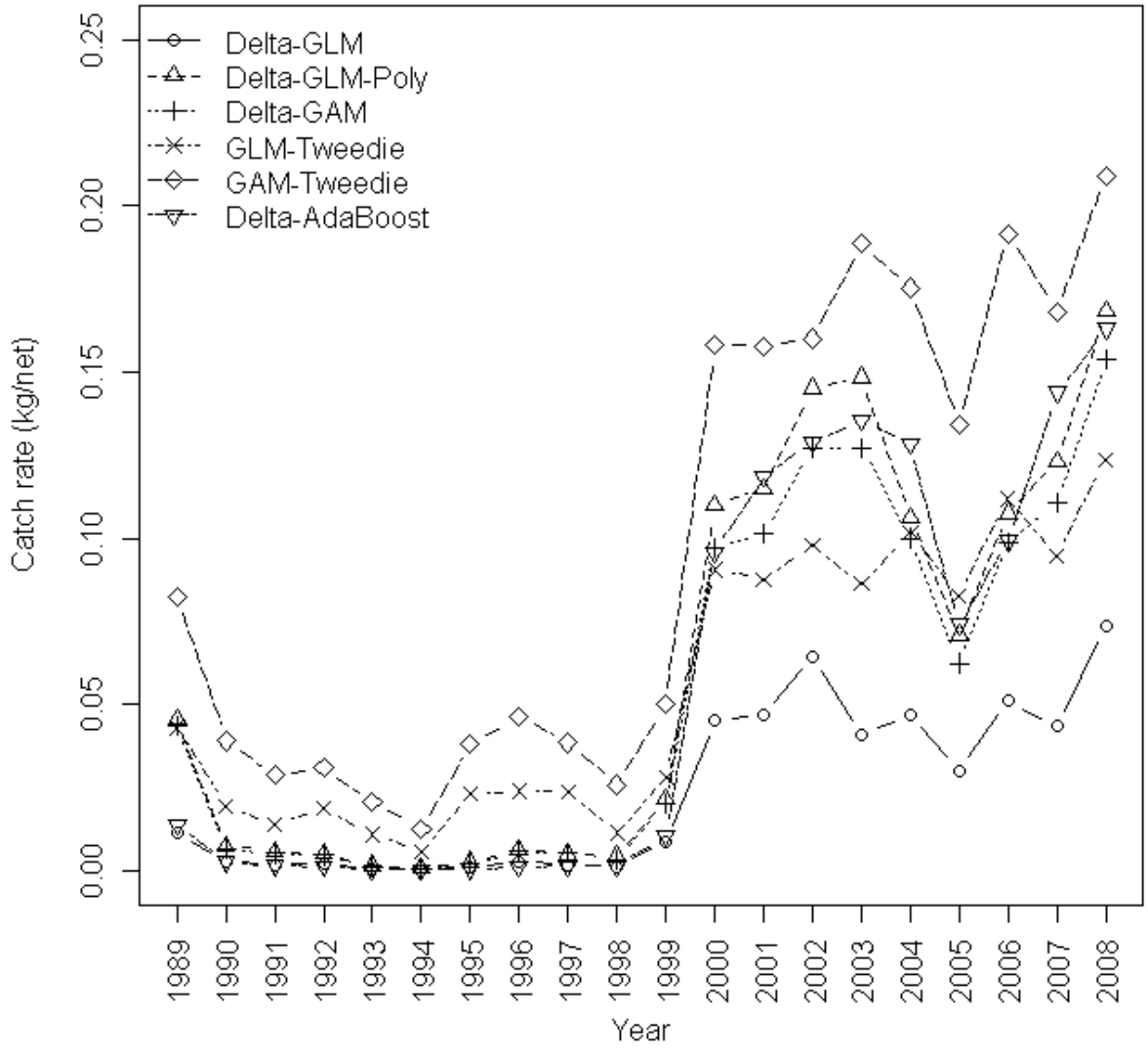

FIGURE 4-3.— Trends of standardized catch rates (kg/net, 30.5m long × 1.8m deep) for yellow perch over time generated by the Delta-AdaBoost model and the five candidate models. The Delta-AdaBoost model consisted of one generalized additive model and one AdaBoost model (Delta-AdaBoost). See Table 4-2 for the explanation of models.

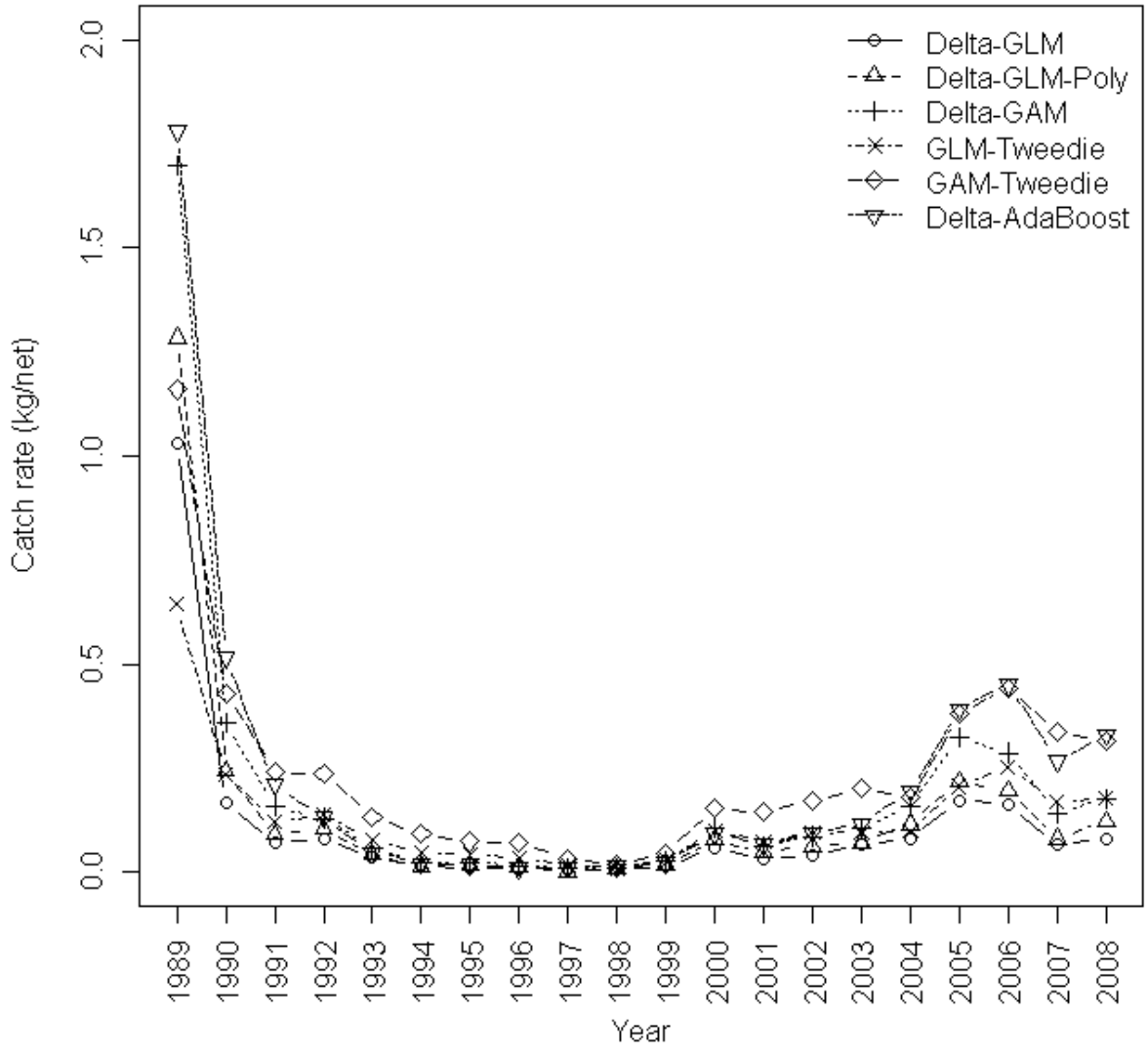

FIGURE 4-4.— Trends of standardized catch rates (kg/net, 30.5m long × 1.8m deep) for white perch over time generated by the Delta-AdaBoost model and the five candidate models. The Delta-AdaBoost model consisted of one generalized additive model and one AdaBoost model (Delta-AdaBoost). See Table 4-2 for the explanation of models.

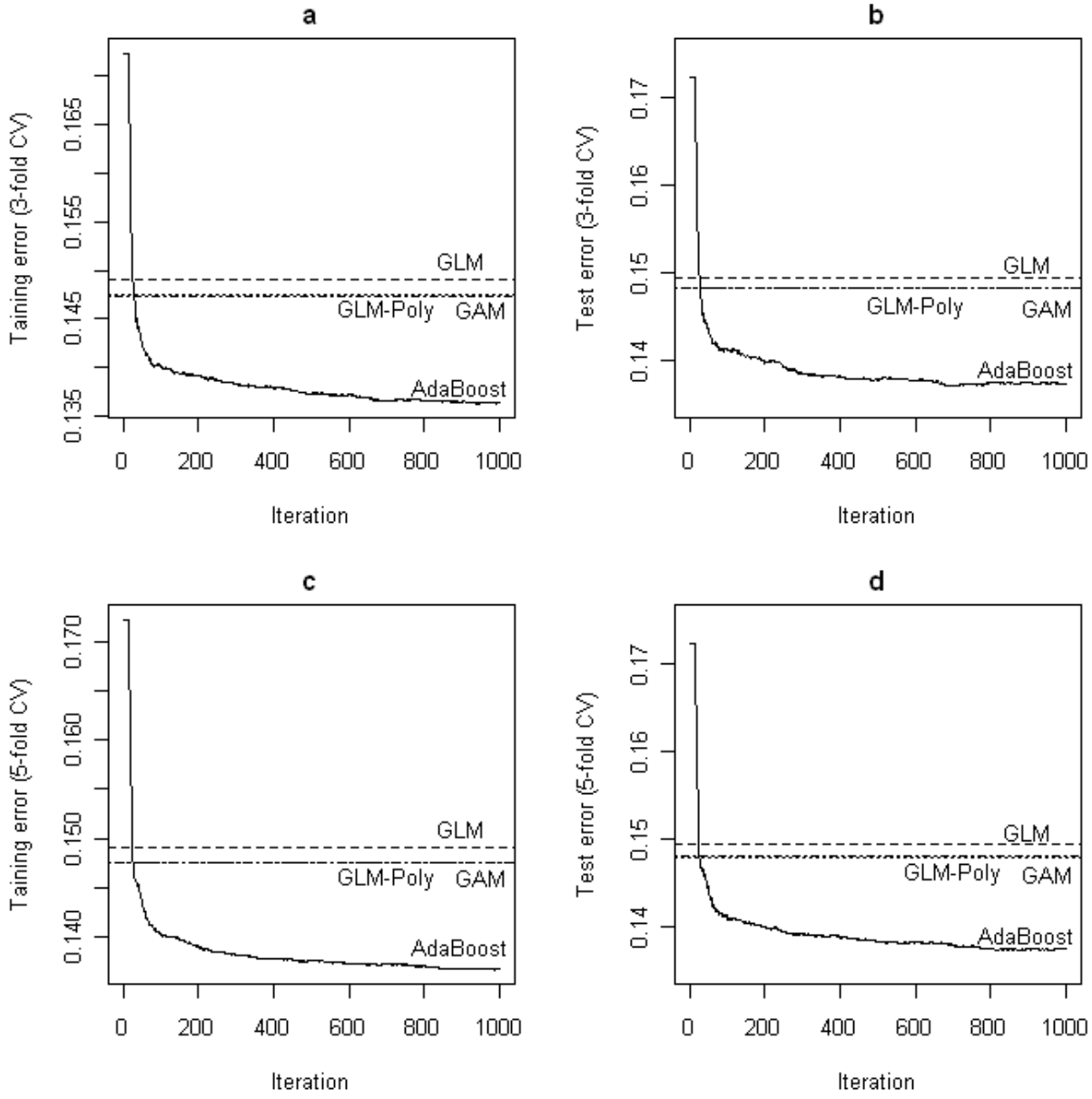

FIGURE 4-5.— Training and test error rates in analyzing the presence/absence of the fish species from the AdaBoost model and the three candidate models for walleye by 3- and 5-fold cross-validation(CV). The three candidate models included the generalized linear model assuming a binomial distribution (GLM), a generalized linear model with polynomial terms up to degree 3 assuming a binomial distribution (GLM-Poly), and the generalized additive model

assuming a binomial distribution (GAM). Error rates were averaged from the 3- or 5- fold cross-validation.

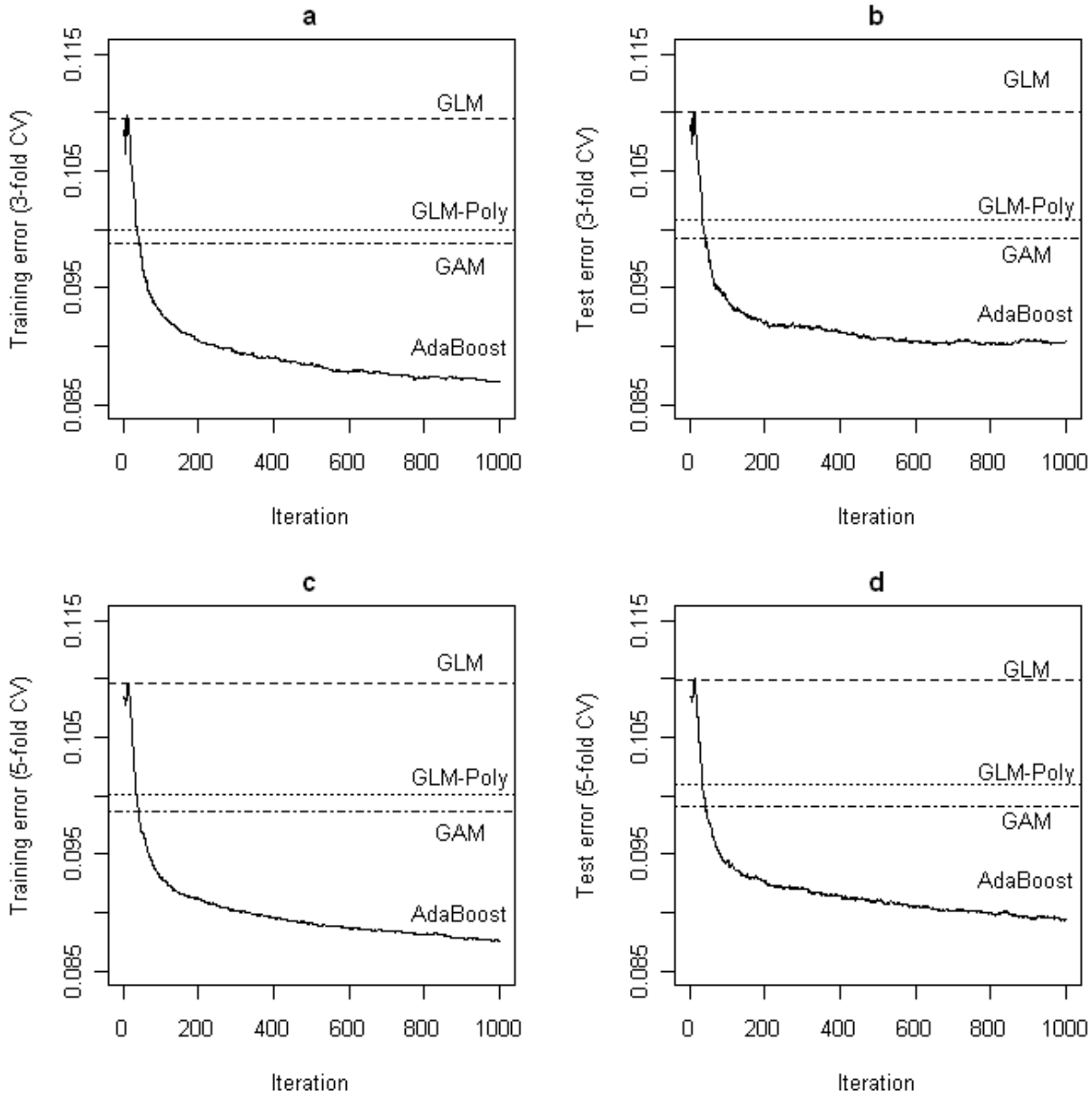

FIGURE 4-6.— Training and test error rates in analyzing the presence/absence of the fish species from the AdaBoost model and the three candidate models for yellow perch by 3- and 5-fold cross-validation (CV). See Figure 4-5 for the explanation of models.

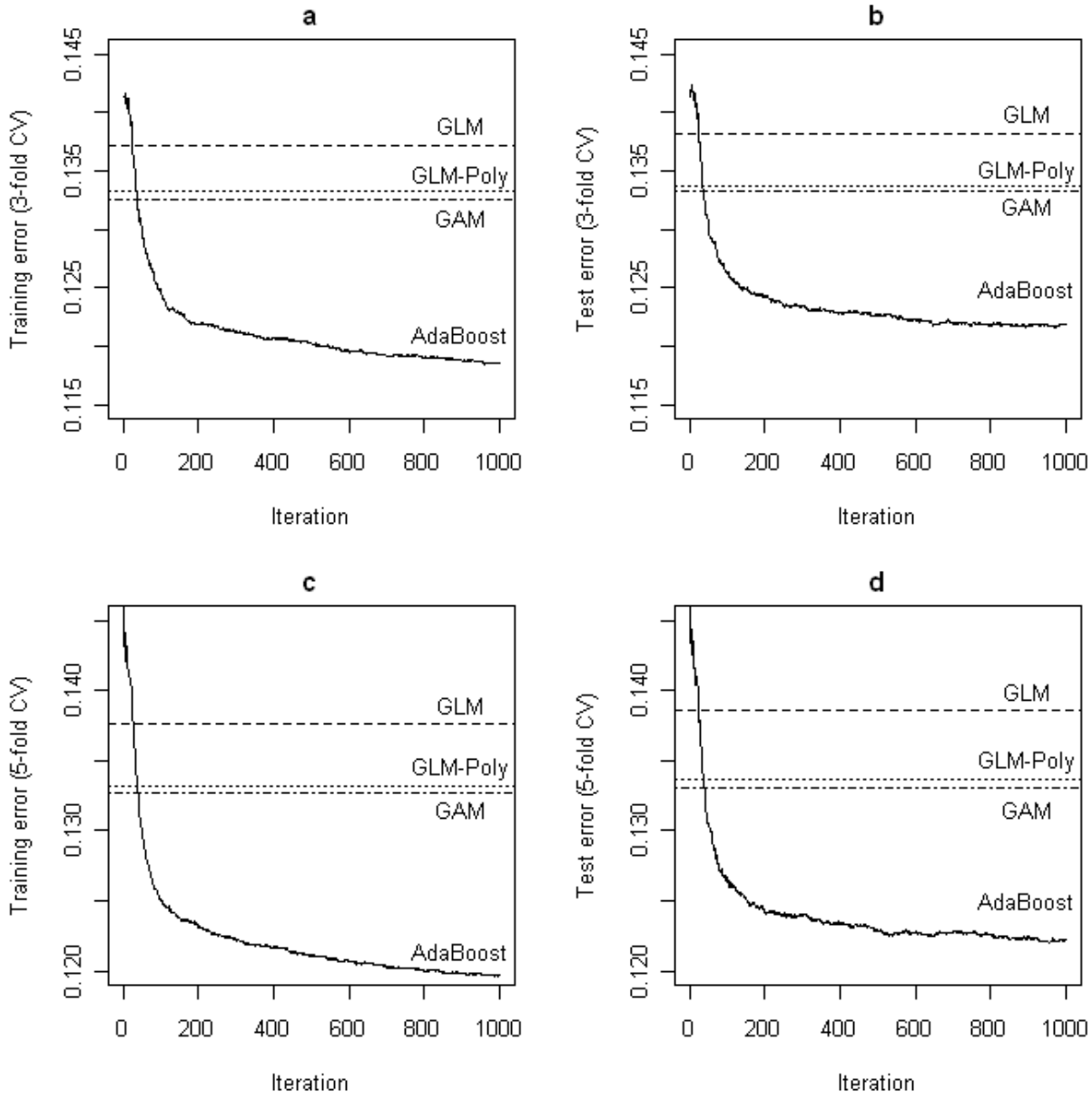

FIGURE 4-7.— Training and test error rates in analyzing the presence/absence of the fish species from the AdaBoost model and the three candidate models for white perch by 3- and 5-fold cross-validation (CV). See Figure 4-5 for the explanation of models.

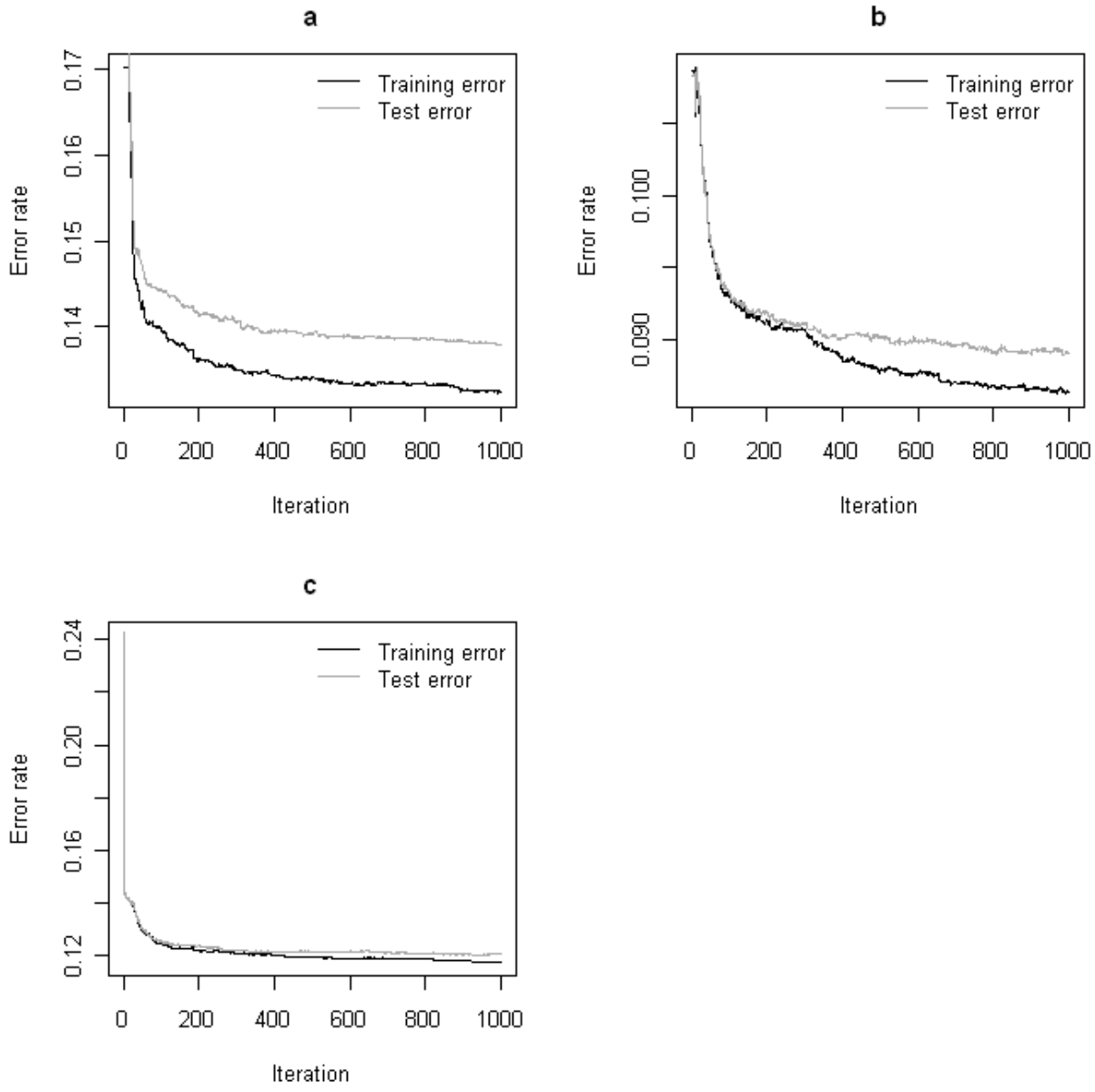

FIGURE 4-8.— Training and test error rates from the AdaBoost models for walleye (a), yellow perch (b) and white perch (c). The PIS data were split into two sub-datasets with roughly equal size, the training dataset and the test dataset.

Chapter 5
Conclusions

This study explored the application of the AdaBoost algorithm in analyzing fishery data with high percentage of zeros and examined the impacts of environmental factors and gillnet fishing on bycatch and discards for three key species (walleye, yellow perch and white perch) and one threatened/endangered species (lake sturgeon) and the implications for fishery management and conservation in Lake Erie.

The bycatch and discard analyses for walleye, yellow perch and white perch posed important implications for bycatch management in the commercial gillnet fisheries in Lake Erie. The gillnet fisheries had potential influences on lake sturgeon bycatch and further on its population conservation. The application of the AdaBoost algorithm combined with a delta model indicated that the Delta-AdaBoost model can be considered as a candidate model when a high proportion of zero observations are included in the fishery data.

The gillnet bycatch and discard analyses of walleye, yellow perch and white perch indicated the hotspots for bycatch and discards in the commercial gillnet fisheries in Lake Erie, i.e., more bycatch may be obtained in the west basin in October for walleye, in the west central basin in November for yellow perch, and in the west central basin in October for white perch; more discards can be observed in the west basin of Lake Erie during August to September for walleye, in the waters across the west central and east central basin in November for yellow perch, and in the west basin in August and November for white perch. Possible bycatch management strategies include restrictions on fishing season and hotspot areas for bycatch and discards, and a joint fishing license framework for target and bycatch species. More fishing efforts on the invasive species white perch may increase its fishing mortality and mitigate the competition between white perch and the native species. An advanced fishery data recording system and observer program that can incorporate both bycatch and discard information into the commercial fishery reports would be useful for further bycatch studies and management.

The classification tree model can be considered when the generalized linear/additive-based models showed poor performance because of the extremely high percentage of zero observations and the complicated data structure. The bycatch analysis of lake sturgeon through

the classification tree model approach indicated that the gillnet fisheries in Lake Erie may potentially increase the mortality of juvenile lake sturgeon and impede the recovery of lake sturgeon populations. The west basin of Lake Erie could be a hotspot for lake sturgeon bycatch in the gillnet fisheries, and more attention should be drawn to gillnet fisheries management in the west basin with an emphasis on lake sturgeon conservation.

Data with high percentage of zero observations are more likely to be encountered in fishery analyses for rare species or bycatch species. When the percentage of zero observations in the data is high, the AdaBoost algorithm can be applied to estimate the probability of obtaining the non-zero captures in the delta model as an alternative to the generalized linear/additive model with an assumption of binomial distribution. Model selection should be conducted on a case-by-case basis, since this procedure can be confounded with several factors including data structure.