Evaluating Methods for Optimizing Classification Success From Otolith Tracers for Spotted Seatrout (*Cynoscion nebulosus*) in the Chesapeake Bay

Stacy Kavita Beharry
*Old Dominion University*

Follow this and additional works at: [https://digitalcommons.odu.edu/oeas_etds](https://digitalcommons.odu.edu/oeas_etds)

Part of the Aquaculture and Fisheries Commons, Marine Biology Commons, and the Oceanography Commons

**Recommended Citation**
Beharry, Stacy K.. "Evaluating Methods for Optimizing Classification Success From Otolith Tracers for Spotted Seatrout (*Cynoscion nebulosus*) in the Chesapeake Bay" (2011). Doctor of Philosophy (PhD), Dissertation, Ocean & Earth Sciences, Old Dominion University, DOI: 10.25777/7kg5-8y61
[https://digitalcommons.odu.edu/oeas_etds/28](https://digitalcommons.odu.edu/oeas_etds/28)

This Dissertation is brought to you for free and open access by the Ocean & Earth Sciences at ODU Digital Commons. It has been accepted for inclusion in OES Theses and Dissertations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.
EVALUATING METHODS FOR OPTIMIZING CLASSIFICATION SUCCESS
FROM OTOLITH TRACERS FOR SPOTTED SEATROUT (Cynoscion
Nebulosus) IN THE CHESAPEAKE BAY

by

Stacy Kavita Beharry
B.S. May 2005, Morgan State University

A Dissertation Submitted to the Faculty of Old Dominion University in Partial
Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

OCEANOGRAPHY

OLD DOMINION UNIVERSITY
December 2011

Approved by:

Cynthia M. Jones (Director)

Dayanand Naik (Member)

Alexander Bochdansky (Member)
ABSTRACT

EVALUATING METHODS FOR OPTIMIZING CLASSIFICATION SUCCESS FROM OTOLITH TRACERS FOR SPOTTED SEATROUT (*Cynoscion nebulosus*) IN THE CHESAPEAKE BAY

Stacy Kavita Beharry
Old Dominion University, 2011
Director: Dr. Cynthia M. Jones

Identifying the natal sources of fish is an important step in understanding its population dynamics. Adult recruits are often sourced from multiple nursery areas, with good quality locations contributing disproportionately more fish to the adult stock. Because population persistence is strongly influenced by nursery habitat, methods that correctly identify the source of recruits are necessary for effective management. Within the last decade, otolith chemistry signatures have been increasingly used as a natural marker to delineate fish from a mixture of nursery sources. Despite the widespread use of otolith trace element and stable isotope ratios as habitat markers, the statistical approaches to handle these data have been slow to develop. Limited guidelines have been offered for constructing the discriminatory function in terms of the number of chemical variables used, the information conveyed by each variable, and the overall stability of important variables with time. Almost all studies argue that juvenile signatures must be collected anew each year at considerable expense. In this study, Rao’s test for additional information was used to identify the most useful discriminatory variables for identifying the nursery seagrass habitats for spotted seatrout (*Cynoscion nebulosus*) in Chesapeake Bay. Additionally, Akaike information criterion (AIC) and Bayes information criterion (BIC) were used to select the discriminant function analysis (DFA) model that minimized
the prediction error for determining provenance of adult fish. The AIC technique was also used to construct a short-term multi-year habitat tag for the Bay. Variable selection using Rao’s test show that classification accuracy was heavily dependent on the type and number of variables used in the model. Barium was the most important variable and it was the most stable variable over time. From the AIC model selection, adult fish were correctly assigned to nursery area with over 94% accuracy within-year, while the AIC multi-year tag accurately identified the source of historical collections of adult fish with over 80% classification accuracy. These results show that by using correct statistical approaches to construct the discriminatory model, the probability of misclassification for subsequent survivors is minimized. Additionally multi-year models can be developed, directing research for other species.
Dedicated to my mom, Grace Beharry, for her love, support, and encouragement
ACKNOWLEDGEMENTS

There are many people who contributed to the completion of this research. First I would like to thank the members of my dissertation committee: Dr. Cynthia M. Jones, Dr. Dayanand N. Naik, and Dr. Alexander Bochdansky. Also I would like to thank members of my advisory committee: Dr. Dayanand N. Naik and Dr. Fred Dobbs.

I am deeply grateful to my advisor, Dr. Cynthia M. Jones. I have been very fortunate to have an advisor that allowed me to explore questions on my own while supporting me throughout the course of my study. Her mentorship helped me develop strong research and critical thinking skills in the field of quantitative fisheries ecology. As a student in her lab, I had the opportunity develop teaching, managerial, and interpersonal skills that will serve me for many years into my career. Dr. Jones is a wonderful mentor, I have earned a diverse tool kit under her guidance, and for that I am very thankful.

I would also express sincere gratitude Dr. Dayanand N. Naik for the many hours he spent guiding me through the statistical analysis for this project. Dr. Naik formally introduced me to the world of statistics, and his insightful discussions helped me formulate ideas for this project. Without his patience, kind understanding, and enthusiasm for the field of statistics, this research could not be completed.

I would also like to thank Dr. Jason J. Schaffler, the most outstanding young scientist, instructor, SAS programmer, lab manager, field technician, chemical analysis technician, angler, and all around expert I know. Dr. Schaffler has guided me in almost every aspect of my research. His discussions helped me shape each analysis so it would be thoroughly executed. I thank him for his advice and his guidance in teaching me how
to be a good scientist.

I would also like to express my gratitude to Jacques van Montfrans from the Virginia Institute of Marine Science (VIMS). Captain Jacques collaborated on all our field collections and also went above his call of duty to collect fish by soliciting donations from anglers, inquiring about fish from commercial fishers, and supplying fish that he personally collected. I am particularly thankful for his instrumental role in informing recreational fishers of this project and for the opportunity to formally present my research to his angler club.

The faculty and staff of the Center for Quantitative Fisheries Ecology (CQFE) have also been instrumental in the completion of this research. Lab manager, Dr. Hongsheng Hank Liao managed all the collections of fish from the Virginia Marine Resources Commission (VMRC). I thank him for his support and scientific discussions. I also benefitted from countless hours of discussions, field support and advice from Dr. Joseph Ballenger and Reneé Hoover. I extend many thanks to Dr. Julian Ashford, Jason Ferguson, James Davies, and Dr. William Persons at CQFE for their support.

This work could not be completed without field collections from VMRC, many thanks to Joe Grist, Myra Thompson, and the rest of the staff at VMRC. I also thank Beth Burns of the North Carolina Department of Natural Resources (NC-DNR) for field collections and the staff of Sea Center Texas for supplying known age hatchery fish. I extend thanks to Dr. Robyn Hannigan at the University of Massachusetts for graciously allowing me to clean equipment when I literally had no place else left to go.

I am thankful to Alana Beharry and Ruben Baez for their encouragement and support. Finally, I thank the Department of Ocean Earth and Atmospheric Sciences at Old
Dominion University for giving me the opportunity to complete my degree. This research was funded by a National Science Foundation Grant.
TABLE OF CONTENTS

LIST OF TABLES ......................................................................................................................... x

LIST OF FIGURES ......................................................................................................................... xi

Chapter

I. INTRODUCTION .................................................................................................................. 1
   ECOLOGICAL BACKGROUND ................................................................................................. 1
   VARIABLE SELECTION FOR DISCRIMINANT FUNCTION ..................................................... 5
   APPLYING A HABITAT TAG ..................................................................................................... 7
   OBJECTIVES ........................................................................................................................... 8
      SEAGRASSES AS NURSERY AREA ....................................................................................... 9
      MODEL SYSTEM ............................................................................................................... 10

II. VARIABLE SELECTION IN HIGH DIMENSIONAL DATA AND
APPLICATION TO IMPROVING CLASSIFICATION MODELS BASED
ON FISH-OTOLITH TRACERS ............................................................................................... 14
   INTRODUCTION ................................................................................................................... 14
   MATERIAL AND METHODS ................................................................................................. 18
      FISH COLLECTION .......................................................................................................... 18
      TRACE ELEMENT CHEMISTRY ....................................................................................... 18
      STABLE ISOTOPE CHEMISTRY ....................................................................................... 20
      AGE AND GROWTH .......................................................................................................... 21
      VARIABLE SELECTION ................................................................................................... 22
   RESULTS ............................................................................................................................... 24
   DISCUSSION ......................................................................................................................... 28
      MODEL COMPARISONS ..................................................................................................... 28
      VALUE OF MEASURING VARIABLE IMPORTANCE ......................................................... 30

III. OPTIMIZING CLASSIFICATION SUCCESS OF TEST DATA USING
VARIABLE SELECTION PROCEDURES IN CONSTRUCTING A
DISCRIMINANT FUNCTION ........................................................................................................ 33
   INTRODUCTION ................................................................................................................... 33
   MATERIAL AND METHODS ................................................................................................. 36
      FISH COLLECTION .......................................................................................................... 36
      TRACE ELEMENT CHEMISTRY ....................................................................................... 37
      STABLE ISOTOPES .......................................................................................................... 39
      MODEL SELECTION ......................................................................................................... 39
   RESULTS ............................................................................................................................... 41
   DISCUSSION ......................................................................................................................... 44
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV. DEVELOPING A MULTI-YEAR HABITAT TAG FOR DETERMINING</td>
<td></td>
</tr>
<tr>
<td>THE NATAL SOURCES OF FISH</td>
<td>49</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>49</td>
</tr>
<tr>
<td>MATERIAL AND METHODS</td>
<td>52</td>
</tr>
<tr>
<td>FISH COLLECTION</td>
<td>52</td>
</tr>
<tr>
<td>TRACE ELEMENT CHEMISTRY</td>
<td>53</td>
</tr>
<tr>
<td>STABLE ISOTOPE CHEMISTRY</td>
<td>53</td>
</tr>
<tr>
<td>STATISTICAL ANALYSIS</td>
<td>55</td>
</tr>
<tr>
<td>RESULTS</td>
<td>57</td>
</tr>
<tr>
<td>SIGNATURES FOR INDIVIDUAL YEARS</td>
<td>57</td>
</tr>
<tr>
<td>DEVELOPING A MULTIYEAR HABITAT TAG</td>
<td>61</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>63</td>
</tr>
<tr>
<td>IMPORTANT VARIABLES</td>
<td>63</td>
</tr>
<tr>
<td>MULTI-YEAR TAG</td>
<td>65</td>
</tr>
<tr>
<td>V. CONCLUSION</td>
<td>68</td>
</tr>
<tr>
<td>LITERATURE CITED</td>
<td>72</td>
</tr>
<tr>
<td>VITA</td>
<td>84</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Variables ordered by Mahalanobis distances $D^2$, and classification accuracies for each variable separately</td>
<td>24</td>
</tr>
<tr>
<td>2.</td>
<td>Rao's test starting with the most important variable and determining whether each addition in bold is significant ($\alpha = 0.05$)</td>
<td>25</td>
</tr>
<tr>
<td>3.</td>
<td>Comparison of classification percent accuracies and overall error rates for all models</td>
<td>28</td>
</tr>
<tr>
<td>4.</td>
<td>Mahalanobis distances ($D^2$) for all variables as a measure of variable importance</td>
<td>42</td>
</tr>
<tr>
<td>5.</td>
<td>Variable combination of each selection criteria ordered by importance and the overall error of misclassification of the juvenile dataset for each criterion</td>
<td>43</td>
</tr>
<tr>
<td>6.</td>
<td>Total number of fish collected for each habitat (ES and WS) and the ANOVA results of variables that show significant differences or no difference in mean concentrations between habitats for all years ($\alpha = 0.05$)</td>
<td>58</td>
</tr>
<tr>
<td>7.</td>
<td>Overall habitat effect for each year using MANOVA ($\alpha = 0.05$)</td>
<td>58</td>
</tr>
<tr>
<td>8.</td>
<td>The best model for each year determined by the AIC selection method and the correct classification probability for each shore</td>
<td>59</td>
</tr>
<tr>
<td>9.</td>
<td>Three most important variables for separating groups in each year with corresponding Mahalanobis distance ($D^2$) for ES and WS habitats</td>
<td>59</td>
</tr>
<tr>
<td>10.</td>
<td>Classification success using a multi-element discriminatory tag derived from fish spawned in 2006-2008 to delineate habitats across all years, including years where fish were not used to develop the tag (depicted in bold)</td>
<td>62</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Locations of sampling stations in the Chesapeake Bay: Western Shore (●) and Eastern Shore (★)</td>
<td>19</td>
</tr>
<tr>
<td>2.</td>
<td>Juvenile spotted seatrout growth on Eastern and Western Shore habitats for 2007: (a) raw data points for each habitat and (b) combined growth for all fish from each habitat using Gompertz growth model</td>
<td>27</td>
</tr>
<tr>
<td>3.</td>
<td>Locations of sampling stations in the Chesapeake Bay: Western Shore (●) and Eastern Shore (★)</td>
<td>38</td>
</tr>
<tr>
<td>4.</td>
<td>Overall error rate of simulated “unknown” fish to natal source using AIC (●), BIC (□), and minimum error (★) models for each combination of simulated “unknowns”</td>
<td>44</td>
</tr>
<tr>
<td>5.</td>
<td>Locations of sampling stations in the Chesapeake Bay: Western Shore (●) and Eastern Shore (★)</td>
<td>54</td>
</tr>
<tr>
<td>6.</td>
<td>Box plots of untransformed otolith elements for ES (dark bars) and WS (open bars) across a five year time period for the Chesapeake Bay.</td>
<td>60</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

Ecological background

Understanding the spatial dynamics of fish populations is fundamental to ensuring the productivity of stocks and the conservation of overfished or threatened species. Coastal and estuarine fish populations that were once thought of as open, with a broad dispersal of individuals across habitats, are now recognized as highly structured with individuals favoring specific habitats during distinct stages of development (Knutsen et al. 2003, Dorval et al. 2005b). For many species, a single population may have multiple nursery sources (Waples and Gaggiotti 2006, Wakefield et al. 2011). These nursery habitats have been shown to influence growth and survival through spatial heterogeneity. Therefore, specific nursery areas can strongly influence cohort strength by contributing disproportionately more fish to the adult stock (Beck et al. 2001, Minello et al. 2003). With population persistence highly dependent on these productive areas, identifying natal habitat of recruits is an important aspect in effectively managing the fishery.

Determining the natal origin of fish is a complex undertaking, particularly in the case of larval and juvenile fish (Levin 2006). During the early stages of development, many fish species experience very high mortality (Houde 1989, Shepherd et al. 1990, Campana 1996). Therefore, methods that track movement and dispersal of individuals from source locations using artificial tags are often unsuccessful. To ensure adequate recaptures using artificial tags, large numbers of juveniles must be tagged, making this

\footnote{The journal model is taken from Ecological Applications}
form of tagging an overwhelming task (Gillanders 2009). Other methods to determine movement have relied on differences in abundance, size, or age structure among groups (see Gillanders et al. 2003 for review). While these methods can provide useful information, juveniles are often under sampled leading to highly variable estimates. To accurately document the source of recruits, direct observation of dispersal from natal origin is preferred. Consequently, a tag that remains with the fish such as a permanent natural marker is better suited to unraveling provenance and subsequent philopatry, as all fish from a given source are marked and the tag remains with the fish until time of capture (Thorrold et al. 1997b, Thorrold et al. 2001, Campana 2005).

There are a variety of natural tags that can be used to identify natal origin of recruits. Extrinsic tags such as natural parasites have long been used as biological markers in fish (Sindermann 1961). Although natural parasites are often unique to a natal area, they may change as the fish grows making it unsuitable as a long term marker and all fish from a natal area may not carry the parasite. Another marker for natal habitat can come from biological processes such as growth (Guido et al. 2004). Differences in growth have been used to identify stocks but it is rarely applied as marker for movement. However, if growth differs significantly between nursery areas it can be a useful tagging method. A more powerful approach to distinguish natal origin is to use an intrinsic tag, such as the chemistry found in calcified structures such as the otolith, or ear stone in fish.

Since the 1980s, researchers have been developing the use of chemical tracers derived from the otolith as a natural tag for fish origin (Campana 1999). The otolith is a calcium carbonate structure found in the head of the fish, and among the structures used to understand movement, it has become a standard marker to identify natal habitats for a
variety of taxa and systems (Gillanders and Kingsford 2000, Thorrold et al. 2001, Campana 2005, Dorval et al. 2005b, Brown 2006b, Rooker et al. 2008a, Standish et al. 2008, Walther and Thorrold 2010). As the fish grows, the otolith incrementally accretes calcium carbonate layers on a proteinaceous matrix (Degens et al. 1969). As these daily layers form, depending on environment and metabolic processes, elements such as barium (Ba), strontium (Sr), and magnesium (Mg) replace the calcium (Ca) ion in the crystallization process. Additionally, rare earth elements such as lanthanum (La) and rubidium (Rb) can also be trapped in the crystalline structure as the otolith grows. The elemental chemistry of the otolith functions as a natural tag, and can be vital in distinguishing fish that have experienced different physical and biological environments (Campana 1999, Dorval et al. 2005b). As the otolith is acellular and deposited material is not mobilized with time, which is especially important for a permanent tag, the chemical composition of the otolith thereby provides a chronological record of birth place and movement.

Trace element-to-calcium ratios are used to separate habitat, as the incorporation of these ratios not only reflect ambient concentration for some elements but also track physical parameters such as temperature and salinity which are often specific to an area. Bath et al. (2000) showed that although otolith strontium (Sr) and barium (Ba) concentrations in otoliths of spot (Leiostomus xanthurus) reflected the relative proportion of these elements in ambient waters, temperature significantly affected incorporation of Sr. Similarly, Collingsworth et al. (2010) found that the incorporation of Sr, Ba, and manganese (Mn) were all influenced by either water temperature or the interaction between temperature and ambient concentration.
In addition to trace element ratios, the stable isotope ratios of oxygen (\( ^{18}\text{O}/^{16}\text{O} \)) defined as \( \delta^{18}\text{O} \), and carbon (\( ^{13}\text{C}/^{12}\text{C} \)) defined as \( \delta^{13}\text{C} \), are also used for determining birth place based on past environmental history (Edmonds and Fletcher 1997, Schwarcz et al. 1998, Rooker et al. 2008b, Newman et al. 2010). The oxygen isotope ratio is incorporated into otoliths in equilibrium with ambient water (Thorrold et al. 1997a, Høie et al. 2004) and is useful in separating fish along latitudinal clines (Ashford and Jones 2007). Unlike oxygen, carbon is an indicator of both abiotic and biotic processes as the incorporation of carbon is connected to dissolved inorganic carbon (DIC) and metabolic activity (Kalish 1991b, Solomon et al. 2006, Tohse and Mugiya 2008). These ratios can delineate the natal habitat of fish, and when combined with trace element ratios they may better describe the habitat.

Since the otolith acts as a chronological environmental recorder, information held in its chemical composition can be used as a habitat tag as long as fish remain on the nursery grounds such that sufficient information is recorded in their otoliths. By matching the otolith nuclei chemistry of adult fish to the juvenile otolith chemistry, juvenile production and the contribution of different habitat sources to the adult population can be identified (Gillanders et al. 2003). Over the last decade, this methodology has been applied to a variety of systems and taxa to understand habitat use (Thorrold et al. 1998b, Thorrold et al. 2001, Gillanders 2005a, Mateo et al. 2010). Miller and Shanks (2004), using microchemistry of black rockfish (\( Sebastes melanops \)) otoliths, show that movement from origin was highly restricted contrary to what was previously believed. Brown (2006b) demonstrates significant philopatry in English sole (\( Pleuronectes vetulus \)) to estuarine habitat on the West Coast of the United States.
Otolith chemical markers show considerable promise in delineating natal sources. However, a major limitation of this methodology has been the statistical approaches used in constructing the best classification model for the system. Discriminant function analysis (DFA) is a commonly used statistical approach for studying the differences among groups. Otolith chemical variables are used in the DFA model to determine a classification function that best describes each nursery area in the system which is then subsequently applied to classify unknown-origin adults to their natal habitat. However, there are several aspects relating to the construction of the DFA that is overlooked by practitioners. Otolith chemical variables often describe different aspects of the habitat. Therefore, to build the most parsimonious classification model for use in subsequent studies of movement and philopatry, the number and type of variables used should be first assessed. Additionally by choosing a best model, temporal variation in parameters can be determined and multi-year habitat tags can be developed. These are important issues in classification that will be further explored in this dissertation.

Variable selection for discriminant function

The most commonly used method to determine group membership of fish has been discriminant function analysis (DFA). This method allows researchers to separate mixtures by grouping individuals that are similar with regard to several discriminatory variables (Klecka 1980). In fisheries ecology, most studies have focused on the predictive form of classification, whereby the main goal of the classification function is to identify group membership. In a literature review from 2000 – 2010, more than 75% of all studies using otolith tracers to assign fish to habitat used DFA as the classification technique.
The correct application of DFA has been hampered by a variety of theoretical issues (James and McCulloch 1990). To correctly use this statistical tool, a few assumptions should be made with respect to the equality of variance-covariance matrices (Williams 1983). When groups share a common variance-covariance matrix then linear discriminant function analysis (LDFA) can be employed. This is the simplest form of DFA, as a linear combination of variables can be used in the discriminant function (Klecka 1980). There is also an assumption that each group is drawn from a population that is multivariate normal. When variance-covariance matrices are not equal, then quadratic discriminant function analysis can be employed (QDFA). This technique also requires that data must be multivariate normal to correctly determine group membership probabilities. If assumptions are not met then the resulting prediction error from the DFA will be incorrect. Although LDFA and QDFA are parametric methods that are widely used in ecological research these assumptions are often disregarded in fisheries applications.

Once assumptions are met, researchers must select variables to construct the classification model. Discriminatory variables often carry disproportionate amounts of information relevant in separating groups. Some variables may not express any relevant information, while others may convey similar information and including these variables can reduce classification success (Van Ness and Simpson 1976). Variable or model selection methods to choose the most parsimonious model from the suite of potential variable combinations are necessary. As noted in the statistical literature, variable importance and model construction are important in optimizing the classification function (Van Ness and Simpson 1976, Klecka 1980, Huberty 1984, Johnson and Omland 2004).
Methods to assess variable importance and to construct a multi-year natal signature are necessary in minimizing the prediction error for fish from several nursery sources. If the classification model does not adequately describe the habitat, any subsequent classification of unknown fish can result in large misclassification error and estimates on the proportion of individuals from each source would be incorrect (White and Ruttenberg 2007). Issues in constructing a model such as variable importance or number of variables needed are rarely addressed in fisheries applications of DFA but form the foundation for correctly assigning fish to natal areas, and in turn, deriving the source locations of the adult stock. Therefore, variable importance and model building techniques is an area of research that requires further exploration.

**Applying a habitat tag**

To successfully apply otolith chemical markers as a habitat fingerprint, there should be substantial differences in otolith elemental chemistry among fish residing in each area and these differences should be both spatially and temporally consistent (Campana et al. 2000). To determine whether spatial differences exist, it is first necessary to collect juvenile fish from each nursery, even if these areas are separated by large geographical distances. Proctor et al. (1995) could not resolve differences in spawning areas of southern bluefin tuna (*Thunnus maccoyii*) even though fish were sampled across a large spatial scale (>1000 km). Conversely, Dorval et al. (2005b) was able to apply otolith chemical tags to identify juvenile spotted seatrout (*Cynoscion nebulosus*) from nursery habitats less than 10 km apart. Areas that maintain differences in biogeochemical signatures regardless may be successfully delineated using otolith chemistry even across
small spatial scales.

Similarly, temporal variation in otolith markers can also limit the application of this methodology. The chemical composition of the otolith is regulated by a variety of environmental and physiological factors that can often change with time. Hence, is important to determine the chemical variables that best assign fish to nursery areas and the temporal scale in which this tag can be applied (Brown 2006a). Numerous studies have found significant interannual or monthly variability in chemical tags among locations (Gillanders and Kingsford 2000, Milton and Chenery 2001, Hamer et al. 2003, Patterson et al. 2004). This variation in chemical signatures limits the application of the habitat tag as juveniles must be sampled each year in order to construct a habitat marker for subsequent adults. However even in very dynamic systems such as estuaries, where several processes such as river discharge, tidal flow, lithology etc. contribute unique components to the overall chemical signature, the magnitude of these contributions can remain stable over time (Smith 1977, Shumilin et al. 1993, Hannigan et al. 2010). In these instances, using model selection criterion to elucidate important variables from the system can lead to the development of a stable tag that can delineate habitats across longer time scales.

Objectives

This dissertation research aims to develop the statistical tools relevant for correctly selecting variables for discriminant function analysis, and developing and evaluating the effectiveness of a multi-year classification model for juvenile fish. The three main questions that will be addressed are:
1) Can classification models for linear discriminant function analysis be optimized using the most informative variables?

2) Can model selection techniques determine the most parsimonious model for identifying adults using quadratic discriminant function analysis?

3) Can a multi-year habitat tag be identified for the seagrass nursery habitats in Chesapeake Bay?

To answer these questions, I sampled juvenile spotted seatrout (Cynoscion nebulosus) on their nursery seagrass habitats using a model system of the Chesapeake Bay.

**Seagrasses as nursery area**

Seagrasses are flowering marine plants that play an important role in stabilizing sediments, improving water quality, sequestering carbon, and most importantly, serving as nursery habitat to a rich assemblage of marine and estuarine species. These areas are crucial to fish population persistence as they offer juveniles refuge from predation and shelter from strong currents. Moreover, seagrasses provide an abundant supply of food as these habitats are used extensively by numerous invertebrates and other small fish species (Beck et al. 2001, Orth et al. 2006). Many important commercial and recreational fish species depend on seagrasses as primary nursery habitat during the juvenile stage. In a literature review, Heck et al. (2003) showed that growth and survival of fish greatly increased on seagrass beds. Consequently, these areas are now considered essential fish habitat (EFH). The importance of nursery habitat was also noted by Congress, which mandated that all EFH be identified and conserved in the 1996 reauthorization of the Magnuson-Stevens Fishery Conservation and Management Act.
In the Chesapeake Bay, the second largest estuary in the world, there has been a sustained decline in seagrass acreage over the last few decades (Orth and Moore 1984). Over 70% of the historic seagrass acreage has been lost in the Bay and despite intensive efforts to restore many of the seagrass beds, there has been mixed success. Frequently, restoration efforts focus on replanting seagrasses from seeds, bare shoots, and patches of shoots still held together by sediments (Goshorn 2006). The Chesapeake Bay Program restoration project has been replanting seagrasses since the 1990’s. Their untiring effort has led to seagrass growth in several areas, but overall, seagrasses continue to decline the estuary.

It should be noted that all seagrass beds are not equal in terms of the biological and physical parameters operating within the grass bed. This spatial heterogeneity can affect both growth and survival of post-settlement fish, thereby allowing for disproportionate survivorship among grass beds. Simply put, some seagrass beds can confer a survival advantage to fish that settle there. Smith et al. (2008) found that spotted seatrout (Cynoscion nebulosus) grew faster depending on which seagrass bed juveniles settled and Sogard (1997) makes the point that faster growth increases the likelihood of survival. Therefore, it is important to accurately identify the grass beds that contribute more fish to the adult stock.

**Model system**

Spotted seatrout (Cynoscion nebulosus) are a well-studied estuarine dependent species that is resident to estuaries along the eastern coast of the United States from the Chesapeake Bay to the Gulf of Mexico (Bortone 2003). This fish is dependent on
seagrass habitat during the juvenile stage, as juveniles prefer seagrass habitat as nursery to other areas (Powell et al. 2004). When seagrass beds are not available, seatrout will settle on other structurally complex nursery areas such as oyster beds or rocky reefs. In estuaries, once juvenile settle on a seagrass bed they seldom move, and tend to remain there throughout the juvenile period. Research has shown that these fish settle onto seagrass beds within a few weeks of hatching and form schools within weeks of settlement (McMichael and Peters 1989, Rutherford et al. 1989). Consequently, the dependency that seatrout exhibit to their natal seagrass beds make them an important monitor for the condition of the estuary.

Across the species range, seatrout are divided into unique populations with populations residing in estuaries along the Atlantic Ocean being genetically unique from those occurring in the Gulf of Mexico (Gold et al. 1999, Ward et al. 2007). The genetic structuring of these populations follows an isolation-by-distance hypothesis as individuals typically remain year round residents in their respective estuary (Weinstein and Yerger 1976, Peebles and Tolley 1988). In the Chesapeake Bay, seatrout are at the northern extent of the species range. These fish are genetically distinct from all other populations and unlike the species congener, weakfish (*Cynoscion regalis*), this marginal population has not extended its range further north (Wiley and Chapman 2003).

The growth, maximum age, and spawning season of seatrout also varies geographically. In the Chesapeake Bay seatrout are long lived when compared to all other populations, as fish can live up to 12 years (Brown 1981). Moreover, these fish maintain a much faster growth rate when compared to other populations. Smith et al. (2008) demonstrated that juvenile growth in the Virginia far exceeded growth elsewhere with a
growth rate of 1.44 mm per day during the summer months. Unlike southerly populations with protracted growing seasons, Bay fish grow rapidly as water temperatures decline early in the fall.

Unlike populations further south, the population in Chesapeake Bay undergo a unique annual migration (Dorval et al. 2005b). During the spring months seatrout spawn and juveniles settle onto nursery seagrass beds and remain there for the duration of the spring and summer months. However, as water temperatures decline with the oncoming winter, these juvenile seatrout migrate to the warmer waters of the coastal ocean and return to the Bay following spring as adults to spawn. Chesapeake Bay has the only speckled trout population that exhibits a migratory pattern as all other populations remain localized within their respective bays (Dorval et al. 2005b). As temperatures in the Bay during the winter months are below the tolerance limit for these fish, they maintain an offshore movement in winter and inshore in the spring.

This species strong dependence on nursery habitat makes it an ideal candidate for habitat classification studies. Additionally, previous research by Dorval et al. (2005a) on Chesapeake Bay fish has shown that once juveniles settle onto grass beds and remain there throughout the summer months, their otolith chemistry becomes a habitat fingerprint. Even seagrass beds that are separated by small distances, maintain differences in water chemistry through the unique circulation pattern in the Bay (Dorval et al. 2007). Seagrass habitats found on the Western Shore are predominantly influenced riverine discharge while grass beds on the Eastern Shore are mainly influenced by oceanic waters moving north due to Coriolis effect. This tightly coupled relationship with its nursery habitat and the unique hydrology of the Bay makes this spotted seatrout
population a model species for evaluations using otolith tracers in habitat-classification studies. The ability to use otolith-habitat fingerprints to track subsequent adult survival depends on the development of parsimonious and predictive habitat classification rules. This dissertation is the first-time use of statistically correct development of parsimonious classification rules for otolith chemistry as a natural tag. It is a pivotal development in the use of otolith-chemistry natural tag to follow differential adult survival that can be attributed to nursery habitat use.

In this dissertation I address each of the three questions outlined in the objectives in Chapters II, III, and IV respectively using statistical approaches that have not been applied to otolith tracers. Methods are described in detail for each chapter. In chapter V, the major findings of the research are summarized with a discussion of the significance to population conservation and assessment.
CHAPTER II

VARIABLE SELECTION IN HIGH DIMENSIONAL DATA AND APPLICATION TO IMPROVING CLASSIFICATION MODELS BASED ON FISH-OTOLITH TRACERS

Introduction

Fisheries ecologists have long recognized the importance of employing accurate methods for classifying individuals to natal habitats, discriminating populations, and delineating migratory routes. There have been a growing number of studies that have used linear discriminant function analysis (LDFA) to classify individuals, which is a good approach when groups share a common variance-covariance matrix (see Williams 1983 for review). With the burgeoning use of otolith chemistry as a natural tag to determine connectivity (Thorrold et al. 2001, Campana 2005, Elsdon et al. 2008), LDFA is an attractive tool to classify individuals of unknown origin using a discriminatory function based on otolith variables collected from fish of known origin. However, selecting variables to use in LDFA is challenging and guidance is limited, especially for ecological data, such as otolith tracers, which are often high dimensional.

A discriminatory model that uses a larger variable list may contain uninformative variables, or variables that carry similar information, while one that uses too few informative variables may not describe the groups adequately. Adding variables to a model that are uninformative, or collinear with existing variables can reduce classification success (Williams 1983). To build a function that maximizes classification
and achieves parsimony, it is first important to determine the classification power of each
discriminatory variable. Evaluating discriminatory power allows each variable to be
ranked and provides insight into the drivers of the system. Second, it is necessary to
determine the discriminatory power of each variable when combined with others (Van
Ness and Simpson 1976). This is an important issue as variables that are strong
discriminators independently may lose some of their explanatory power when combined
with others. By assessing the discriminatory power of all variables collected from known
fish in the first step, and using the most important variables to build the classification
function in the second step, greatly reduces the subsequent misclassification of unknown
fish in the final step.

In otolith chemistry studies, the choice of explanatory variables has often been ad
hoc: because they can be easily obtained, or because they have been widely used by other
investigators. Only recently have discriminatory variables been selected because their
mean concentrations differ among groups as determined by analysis of variance
(ANOVA) (Schaffler and Winkelman 2008). An ANOVA approach guides workers in
removing uninformative variables from the dataset, but any redundancy in information
conveyed by collinear variables will not be identified. Stepwise discriminant analysis is
also widely used as a selection method, but its use in LDFA has been strongly
contradicted in the literature (Huberty 1984, James and McCulloch 1990). Other methods
that classify groups, such as machine learning approaches, result in models which may
not show the contribution of each variable to the classification (Mercier et al. 2011).

I highlight the importance of variable selection in building a classification model
using high-dimensional tracers from the otolith, such as trace element concentrations
(Proctor et al. 1995, Campana et al. 2000, Gillanders 2005b), stable isotope ratios (Edmonds and Fletcher 1997, Schwarcz et al. 1998, Rooker et al. 2008b), and growth (Quinn et al. 1999, Quinn et al. 2006). In the case of the trace elements, as many as 18 elemental ratios (Mercier et al. 2011) to as few as one (Yamashita et al. 2000) have been used as discriminatory variables. Trace element concentrations take advantage of geochemical differences between habitats (Gillanders and Kingsford 2000, Ashford et al. 2005, Brown 2006b) and provide a complete environmental record of movement (Campana 1999, Dorval et al. 2005b). Element-to-calcium ratios of magnesium (Mg), manganese (Mn), strontium (Sr), and barium (Ba) are routinely used in classification (Thorrold et al. 1998b, 2001, Patterson et al. 2004, Schaffler and Winkelman 2008, Reichert et al. 2010). As advances in technology have allowed for a wider suite of trace element ratios to be collected, caution must be used when building the discriminatory function, as this increase in variables may not translate to an increase in accurate classification.

Similar to trace elements, stable isotope ratios of carbon and oxygen are routinely used to separate groups (Devereux 1967, Gao et al. 2001, Jones and Campana 2009). Otolith $^{18}$O accumulates close to equilibrium with ambient water (Thorrold et al. 1997a, Weidman and Millner 2000, Høie et al. 2004) while $^{13}$C is linked to dissolved inorganic carbon and to metabolic activity (Kalish 1991b, Smith and Jones 2006, Solomon et al. 2006). Because of the relationship to both environmental and metabolic processes the stable isotope composition of the otolith is useful in discriminating fish from different locations, particularly on large scales (Thorrold et al. 2001, Ashford and Jones 2007). Although, commonly used in concert both ratios may not confer the same amount of
information and it is unclear whether one, both, or none is necessary in classification, particularly when combined with trace element concentrations.

Growth has been used to identify populations and determine life history events (see Jones 1992 for review, Smith et al. 2006). However, daily growth patterns are also a natural tag, as increment deposition is correlated with physical parameters such as temperature and salinity (Brett and Groves 1979), as well as several important early life history events such as hatch dates that may be unique to an area (Miller and Storck 1984). Smith et al. (2008) found that growth in juvenile spotted seatrout, *Cynoscion nebulosus*, within the Chesapeake Bay differed between seagrass habitats even on the microhabitat scale. Therefore, growth rates can also be used to track origin or delineate animals from multiple populations.

In this study, I introduce Rao’s (1965) test for additional information, which can be used to determine the change in Mahalanobis distance that occurs by the addition of a new variable to the model. The application of this method is new to fish tracers, and unlike some of the commonly used methods for evaluating variable importance, this technique measures the contribution of each variable to the classification and guides workers in selecting variables that carry unique information. As the assumptions for this test is equivalent to those required by LDFA this method is a valuable procedure to determine variable importance for LDFA classification models. To illustrate the technique, I classified spotted seatrout (*Cynoscion nebulosus*) of known origin to their natal seagrass habitats in Chesapeake Bay by reducing high-dimensional variables to those selected by Rao’s test.
Material and Methods

Fish collection

I collected juvenile spotted seatrout from two major habitat zones in the lower Chesapeake Bay: the Eastern Shore (ES) and Western Shore (WS). These habitat zones ranged across the entire lower Bay with sampling locations nested within each zone (Fig. 1). Seatrout were sampled at each location from July to October in 2007 using an 8m otter trawl with 0.64 cm stretched mesh. Each location was sampled every two weeks with six 2 minute tows on each seagrass bed. The net was towed at a speed of 1.5 ms\(^{-1}\). Tow distances and locations were measured using a GPS navigation system. All fish were stored in Whirl-Pak bags and kept on ice until returned to the laboratory and subsequently frozen. Total and standard lengths (mm) were measured and fish were weighed (g) before the otoliths were extracted (see Dorval et al. 2005b for further details).

Trace element chemistry

Juvenile spotted seatrout otoliths were extracted in a positive flow class 100 clean room using acid washed glassware. Otoliths were placed on a glass slide and visible tissue was removed using a glass probe and Milli-Q water. To remove fine organics, I soaked the otolith in ultrapure hydrogen peroxide (30\% \text{H}_2\text{O}_2) for 5 minutes followed by a 10 minute sonication in Milli-Q water. Otoliths were allowed to air-dry under a positive flow hood and one sagittal otolith from each fish was randomly selected for trace and minor element assay while the other was used for ageing and stable isotope analysis (Dorval et al. 2005b, Dorval et al. 2007). Otoliths chosen for trace element assay
FIG. 1. Locations of sampling stations in the Chesapeake Bay: Western Shore (○) and Eastern Shore (●).
were subsequently weighed to the nearest 10 µg and dissolved for analysis by directly pippetting 20 µl of ultrapure nitric acid (70% HNO₃) and increasing the sample volume to 1 ml with 1% ultrapure HNO₃ spiked with 2 ppb indium (internal standard).

To measure elemental chemistry, I analyzed otolith solutions using a Finnigan MAT Element 2 Inductively Coupled Plasma Mass Spectrometer (ICPMS). Samples from each location were randomized within trays to account for the effects of machine drift. I measured $^7$Li, $^{25}$Mg, $^{48}$Ca, $^{55}$Mn, $^{86}$Rb, $^{89}$Y, $^{138}$Ba, and $^{139}$La in low resolution and $^{48}$Ca, and $^{88}$Sr in medium resolution mode (Chen and Jones 2006). The concentration of each element was calibrated using the NRC otolith standard (Sturgeon et al. 2005) and multi-element standards prepared from ultrapure stock solutions which cover the breadth of elemental concentrations found in otoliths (Dorval et al. 2005b, Schaffler and Winkelman 2008). Additionally, instrument blank solutions of 2% HNO₃ were run every six samples and linearly interpolated between sample runs. Trace element to calcium ratios were calculated for all samples. The limits of detection (LOD) were calculated as the mean blank values plus three standard deviations (Thorrold et al. 1997b, Wells et al. 2003). When 25% of samples were below LOD, I excluded these elements from the analysis. As otoliths are primarily composed of calcium, all elements were normalized to Ca and expressed as element-to-calcium ratios.

**Stable isotope chemistry**

The remaining sagittal otolith was used for both stable isotope analysis and ageing. A 0.3 mm transverse section was removed for ageing and the remaining halves were analyzed for carbon ($\delta^{13}$C) and oxygen ($\delta^{18}$O) concentrations using an automated
Isoprime Micromass carbonate analyzer. Otolith halves were homogenized and approximately 60-100 μg of otolith powder was used for the analysis. Otolith samples and limestone standards were spiked with 102% phosphoric acid at 90°C with five standards for every ten otolith samples. Isotope concentrations were measured based on the dual inlet method and coldfinger mode (Dorval et al. 2005b). Data were corrected and stated in PDB (Peedee Belemnite).

Age and growth

In addition to field collections, I obtained known-age fish to validate the periodicity of increment formation and to ensure accuracy and precision between readers. Otoliths were mounted on a Loctite slip and fixed to a glass slide with crystalbond. The otolith was then covered in 100% silicone and allowed to set overnight. Transverse sections 0.3 mm thick were sectioned from the otolith using an isomet saw. Each section was ground to the core using lapping film and polished using aluminum slurry (Smith et al. 2008). Polished sections were examined at ×120 – 400 magnification using a compound microscope equipped with a digital camera and image analysis software.

Daily growth increments were counted by two readers and measured along the sulcal groove using ImagePro version 6.0 (Media Cybernetics Inc., Bethesda, MD, U.S.A.). Total increment counts between the hatch ring and capture was used as the total age in days. Final age was determined as the mean of the ages when counts by the two readers were within 10% of one another. If counts between readers varied greater than 10% then increments were recounted. Otoliths that could not be resolved to within 10% were excluded from the analysis. Daily increments were validated using known age
hatchery fish and the overall growth rate was modeled using a Gompertz model:

\[ Y_t = \beta_0 e^{(-e^{-\beta_1 (t-\beta_2)}}) + \varepsilon, \]

where \( Y_t = \) standard length (mm), \( t = \) fish age (days), \( \beta_0 = \) asymptotic length, \( \beta_1 = \) growth rate, \( \beta_2 = \) inflection point and \( \varepsilon = \) error.

**Variable Selection**

Variables were first tested for multivariate normality using Mardia’s test for multivariate skewness and kurtosis (Khattree and Naik 1999). Data that did not meet multivariate normality were transformed using a Box-Cox transformation until residuals conformed to normality. To reduce the high dimensionality of the dataset, I used multivariate analysis of variance (MANOVA) to determine overall habitat effect and analysis of variance (ANOVA) to determine whether single elements were distinguishable between ES and WS habitats with respect to their means. To evaluate the statistical importance of each variable to the classification model, I started with trace element, stable isotope, and growth variables that were significantly different between ES and WS. For growth, I included the \( \beta_i \) parameters from the Gompertz model as independent variables. Due to scale differences between variables, I standardized these data by removing the variable mean and dividing by the standard deviation for all observations (Thorrold et al. 1998b). As homogeneity of variance-covariance matrices was met (\( \chi^2 = 96.75, \text{df} = 78, P = 0.52 \)) I classified the data using jackknifed LDFA (Khattree and Naik 2000).

I tested the value added by the inclusion of each independent variable from all tracers to the classification model using Rao’s (1965) test for additional information. This
method determines the Mahalanobis distance \((D^2)\) between groups and tests whether the addition of another variable into the model increases this distance (McLachlan 1980).

Where \(D^2\) can be calculated using the following:

\[
D^2 = (\mu_{ES} - \mu_{WS})' (\Sigma_{pooled})^{-1} (\mu_{ES} - \mu_{WS})
\]

If \(p\) variables are used to classify fish then I can test whether the inclusion of \(q\) variables to the existing model increases the separation distance. The information gained by the inclusion of \(q\) is determined using the ratio \(U_{q,p}\) given as:

\[
U_{q,p} = \frac{1 + \frac{N_1 N_2}{(N_1 + N_2)(N_1 + N_2 - 2)} D^2_{p+q}}{1 + \frac{N_1 N_2}{(N_1 + N_2)(N_1 + N_2 - 2)} D^2_p} - 1
\]

Here \(N_1\) and \(N_2\) are the sample size for ES and WS. Under the null hypothesis of no information gained, Rao’s test statistic defined as:

\[
U_{q,p} \frac{N_1 + N_2 - q - p - 1}{q}
\]

has an F distribution with \(q\) and \((N_1 + N_2 - q - p - 1)\) degrees of freedom (see Rao 1965 for details).

Starting with the variable that has the largest distance, I subsequently added each variable in descending order of magnitude and tested whether a significant increase in distance occurred. In addition to Rao’s additional information test, I determined the overall error rates from three other widely-used methods. I developed a classification model using: 1) commonly selected variables in otolith classification studies, 2) using all variables that were different between areas, and 3) variables that were collected from the otolith. Finally, I compared the overall error rates among all methods to determine the best model.
Results

I analyzed a total of 60 juvenile spotted seatrout (40 from ES, 20 from WS). Data were multivariate normal after Box-Cox transformation for all variables. Of the suite of trace elements sampled, lanthanum (La) was not above detection limits and was excluded from the analysis. I found overall significant differences between ES and WS habitats (Pillai’s trace = 0.55; $F_{12,47} = 4.83; P < 0.0001$) using all variables from each tracer. Analysis of variance (ANOVA) showed that mean concentration for all trace element variables except rubidium ($Rb, P = 0.95$) were significantly different between ES and WS. For the stable isotope ratios, only carbon showed significant differences between habitats while all growth parameters were significant.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$D^2$</th>
<th>ES</th>
<th>WS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ba</td>
<td>3.55</td>
<td>87.5</td>
<td>70.0</td>
</tr>
<tr>
<td>Sr</td>
<td>2.39</td>
<td>87.5</td>
<td>70.0</td>
</tr>
<tr>
<td>$\delta^{13}C$</td>
<td>2.20</td>
<td>92.5</td>
<td>70.0</td>
</tr>
<tr>
<td>Li</td>
<td>0.96</td>
<td>85.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Mg</td>
<td>0.94</td>
<td>87.5</td>
<td>35.0</td>
</tr>
<tr>
<td>Mn</td>
<td>0.76</td>
<td>85.0</td>
<td>40.0</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>0.63</td>
<td>95.0</td>
<td>20.0</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.62</td>
<td>85.0</td>
<td>30.0</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.61</td>
<td>85.0</td>
<td>25.0</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.30</td>
<td>90.0</td>
<td>25.0</td>
</tr>
</tbody>
</table>
Many of the trace element concentrations (n = 10) were significantly different between ES and WS, indicating that these elements could be used to discriminate habitat. However, all variables were not equal in their classification power (Table 1). When Mahalanobis distances were calculated, Ba had the largest distance between groups ($D_{Ba}^2 = 3.55$) and contributed heavily to my ability to separate fish. Classification using Ba as the only independent variable gave 87.5% accuracy on the ES and 70% on the WS. For stable isotopes, only $\delta^{13}C$ was significantly different between ES and WS ($P < 0.001$). Carbon was the third most important variable and showed good discriminatory power ($D_C^2 = 2.20$). When $\delta^{13}C$ was added to the model containing Ba there was a significant increase in distance (Table 2) and classification success increased.

Classification using growth had the highest overall error rate of all three tracer types even though juvenile seatrout growth varied significantly between ES and WS seagrass beds. Despite clear differences, this tracer was less accurate in classifying fish

Table 2. Rao’s test starting with the most important variable and determining whether each addition in bold is significant ($\alpha = 0.05$). Best model for data indicated by asterisk.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Classification %</th>
<th>Rao's Additional Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES</td>
<td>WS</td>
</tr>
<tr>
<td>Ba Sr</td>
<td>90.0</td>
<td>70.0</td>
</tr>
<tr>
<td>Ba $\delta^{13}C^*$</td>
<td>87.5</td>
<td>85.0</td>
</tr>
<tr>
<td>Ba C Li</td>
<td>87.5</td>
<td>85.0</td>
</tr>
</tbody>
</table>
to habitat (Fig. 2). Since I obtained three growth variables from the Gompertz model that were highly correlated, I could have chosen only one of the three to reduce redundancy in the data set. However, based on $D^2$ growth was not useful as a discriminatory tracer and was not used in the final model. Using any one of the three growth variables, classifications ranged between 85 – 95% ES and only 20 – 30% WS.

In building the model, based on comparison of statistical distance, I started with Ba and then added Sr, the next most important variable. However, Sr conveyed no additional information beyond that contained in Ba, based on Rao’s additional information test $P > 0.05$ (Table 2). Both Ba and Sr are highly correlated (Pearson’s Correlation coefficient 0.77). When $\delta^{13}C$ was added to the model containing Ba, I found that it added significantly more information $P < 0.05$. I also found that $\delta^{13}C$ increased accuracy on the WS by 15% thereby reducing the overall error to 13.8%. Adding more variables to the model did not result in significant information gained (Table 2). The most parsimonious model used only Ba and $\delta^{13}C$.

In comparing models, I found that the highest classification accuracy based on minimal overall error was found using the variables that were selected using Rao’s test of additional information. The model that used Ba and $\delta^{13}C$ correctly identified fish from ES with 87.5%, WS with 85.0% and produced an overall error rate of 13.8% (Table 3). Similarly, using commonly selected variables for trace elements and stable isotopes also proved to have lower success in classifying fish with an overall error of 17.5%. The model using all elements that were significantly different between areas (ANOVA approach) also produced an overall error of 18.8%. The weakest model contained all the variables collected from the otolith. Overall this model had a 30% error rate.
Fig. 2. Juvenile spotted seatrout growth on Eastern and Western Shore habitats for 2007: (a) raw data points for each habitat and (b) combined growth for all fish from each habitat using Gompertz growth model.
TABLE 3. Comparison of classification percent accuracies and overall error rates for all models.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Classification %</th>
<th>Overall Error %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ES</td>
</tr>
<tr>
<td>Rao’s selection</td>
<td>Ba $\delta^{13}C$</td>
<td>87.5</td>
</tr>
<tr>
<td>Commonly used ANOVA</td>
<td>Mg Mn Sr Ba $\delta^{13}C \delta^{18}O$</td>
<td>85.0</td>
</tr>
<tr>
<td>All collected</td>
<td>Li Mg Mn Sr Y Ba $\delta^{13}C \beta_0 \beta_1 \beta_2$</td>
<td>87.5</td>
</tr>
<tr>
<td></td>
<td>Li Mg Mn Rb Sr Y Ba $\delta^{13}C \delta^{18}O \beta_0 \beta_1 \beta_2$</td>
<td>90.0</td>
</tr>
</tbody>
</table>

Discussion

Model comparisons

Selecting variables with Rao’s test from high dimensional data resulted in minimal prediction error and this variable selection approach surpassed the performance of the other methods. All variables from a tracer do not convey information relevant to delineating groups. I obtained maximum classification success using two variables from a suite of 12. These results identify two issues in developing appropriate classification models. Firstly I found that tracers from multiple categories (trace elements, stable isotopes etc.) were required from maximal separation, which is likely to be the case for most ecological data (Higgins et al. 2010). In a review of the fisheries literature where fish were classified into groups, the majority of studies did not employ multiple tracers in the analysis. Using multiple tracers, workers can better elucidate differences in the environment and ontogenetic processes between habitats. Secondly, I note that only a few variables from each tracer may be important for optimal classification. By using several data sets, Mercier et al. (2011) successfully classified fish with anywhere from 3
trace element variables to as many as 9 variables depending on the system and taxa.

In this study, I found that choosing only a few easily detectable variables (in this example, Mg, Mn, Sr, Ba, $\delta^{13}$C, and $\delta^{18}$O), or all variables, resulted in models that did not produce minimal prediction error. Similarly, the ANOVA approach was useful in excluding uninformative variables, however even variables that were different among areas did not contribute significantly to the model. Strontium had a large distance score however it conveyed no additional information above what was already conveyed by Ba. This is a surprising result as both variables are commonly used in many habitat evaluations. Similarly, Mg and Mn did not add any information, although these variables were determined to be significantly different between areas and are often used in concert in fish classification studies (Thorrold et al. 1998b, Patterson et al. 2004, Dorval et al. 2005b). I show that variables that differ between locations can in fact add no new information to a classification model. This method assesses the magnitude of contribution to the classification function and does not require lengthy simulation time (Mercier et al 2010).

In fish classification studies, the main objective is to develop a classification rule that creates the largest distance between disparate groups with a few variables. This study highlights the importance of choosing only the most useful variables to maximize classification. Few investigators have evaluated the choice of variables statistically and to my knowledge none have done so parametrically. I introduce Rao’s test for variable selection when groups are multivariate normal and have the same covariance matrix. As the assumptions for both Rao’s test and LDFA are the same, this calculation is a simple yet elegant method for choosing the important variables for classification models in the
linear case. However, when data do not conform to the assumptions for LDFA, the variable selection step should not be ignored. It has been recommended to determine the prediction error of all combinations of variables and select the model from this list that contained the lowest prediction error. The disadvantage of this approach is that the model selected would not be parsimonious and could not be generalized to larger datasets. Additionally, the variables driving the separation would not be determined and the relevant processes structuring the system would remain unknown. Therefore, selection methods for datasets with unequal variance-covariance matrices should be explored.

Value of measuring variable importance

By determining the most important variables, the unique factors that influence each variable can be identified. I found that Ba and $\delta^{13}$C conveyed sufficient information to accurately classify fish with over 80% accuracy, with Ba contributing most heavily. Previous workers have widely used Ba to classify fish in conjunction with other elements. The clear separation using Ba observed in this study can be attributed to variation in the water chemistry occupied by fish (Dorval et al. 2005a, Gillanders 2005a). Coffey et al. (1997) showed that Ba release in estuaries is regulated by factors such as salinity and suspended particulate matter from riverine discharge. In the Chesapeake Bay, seagrass beds located on the ES are predominantly influenced by marine waters while the WS is strongly influenced by several major rivers. As Ba reflects ambient water chemistry (Bath et al. 2000, Wells et al. 2003, Gibson-Reinemer et al. 2009) and is regulated by differences in salinity, (Dorval et al. 2005a, Dorval et al. 2007) it is not surprising that Ba was a strong discriminator.
For the stable isotope variables, $\delta^{13}C$ and $\delta^{18}O$ are commonly used in concert, yet I show that they can differ in their discriminatory power. Carbon was more valuable in predicting habitat than $\delta^{18}O$. The precipitation of $\delta^{13}C$ in the aragonite matrix of the otolith is due to both metabolic and environmental processes (Degens et al. 1969, Radtke et al. 1996, Solomon et al. 2006). However, it has been shown that environmental processes are largely responsible for $\delta^{13}C$ composition in the otolith (Thorrold et al. 1997a, Solomon et al. 2006). Although not significant in this instance, $\delta^{18}O$ can be very important in studies that cover larger spatial scales as there is a latitudinal gradient observed in the incorporation of $\delta^{18}O$ (Ashford and Jones 2007). These results demonstrate the importance of identifying the informative variables in any classification model.

The success of combining trace element and stable isotope variables reflect similar studies that have separated groups using these tracers. Thorrold et al (1998) clearly discriminated juvenile weakfish (*Cynoscion regalis*) from three rivers within the Chesapeake Bay. Similarly, Dorval et al. (2005b) found that by including stable isotopes to trace element chemistry as a natural tag, the natal seagrass habitat for spotted seatrout in Chesapeake Bay could be accurately identified. In a more recent study, Comyns et al. (2008) determined the major nursery habitats for seatrout in Mississippi by using both tracers with high classification success. Combining trace element and stable isotope ratios for a habitat fingerprint is becoming more prevalent in classification studies (Thorrold et al. 2001, Clarke et al. 2009). However, unlike previous studies I show that using all trace element and stable isotope variables simultaneously is not necessary or desirable in characterizing a habitat and that all variables differ in their classification power. I
anticipate that with more technological advances, data will become more dimensional and the importance of variable selection greater.

For growth, I found that even though this tracer varied between both habitats it did not significantly improve the classification. Young fish in both areas grew at a similar rate until 25 days after which WS fish grew faster. If these collections had included many more fish older than 25 days then growth may have been much more important. The difference in fish growth between ES and WS, although significant, included substantial numbers of small fish whose growth rates had not yet diverged resulting in a Mahalanobis distance that was not significantly different between zones. However, this should not be taken as a general rule. In studies where a population may range across a variety of habitats, growth could be a valuable tracer (Guido et al. 2004).

The methods presented here offer a guide to variable selection in linear datasets. In ecological studies LDFA is commonly used to discriminate among groups (James and McCulloch 1990), and as Rao’s test is a direct measure of variable inclusion for LDFA, these methods are widely applicable. By determining the most important variables for a given system over time, workers can better predict the drivers of ecological systems. For example, if Ba is always important in this system over time I can assign fish to each shore with a certain probability. Assessing variable importance not only allows for a true natural tag to be developed but also provide ecological insight into the drivers of the system. As researchers move to more complicated ecological questions, knowledge on underlying processes can allow for more effective management of our resources.
Introduction

The application of statistical tools that can correctly distinguish groups from a mixture of objects has greatly increased in over the past decade. In particular, multivariate approaches that recognizes group structuring, such as discriminant function analysis (DFA) has been employed in a variety of applications. For example, DFA has been used to identify diet and trophic position of marine predators (Herman et al. 2005), to delineate fish populations (Cadrin 2000), and understand patterns of movement in fish (Schaffler et al. 2009). In fisheries science, a popular use of DFA is to identify natal origin of adult recruits. Many fish species often exhibit heterogeneity in nursery habitat occupancy (Beck et al. 2001). For these highly structured populations, a single harvest can consist of a mixture of recruits from several source locations (Knutsen et al. 2003). As DFA is ideally suited to classify fish of unknown group membership (test data) based on information collected from known origin groups (training data), it is widely used to understand connectivity among habitats, and the relative contribution of adults from different nursery sources (White and Ruttenberg 2007). However, despite the widespread use of this technique, the correct application of DFA has been hampered by several theoretical issues concerning the discriminatory function build from test data.
In building the discriminatory model researchers have overlooked the importance of selecting a single best model from the suite of discriminatory markers collected (but see Wood et al. 2005). Model selection procedures ranging from theoretical methods to graphical techniques have been widely explored in the statistical literature (McLachlan 2004). They play a central role in balancing model fit and the concomitant loss of information as the ratio of parameters to samples decline (Posada and Buckley 2004). Although studies that identify group membership of fish using DFA have been numerous, model selection procedures are rarely employed. However, these procedures can offer significant advantage in elucidating important discriminatory variables, and building a discriminatory function that is parsimonious.

In studies that identify provenance of fish, markers such as otolith trace element composition (Mg, Ba, Sr etc.) and stable isotope ratios ($\delta^{13}C$, $\delta^{18}O$) are commonly used to delineate fish from multiple sources (Dorval et al. 2005b, Brown 2006b, Rooker et al. 2008b, White et al. 2008). In addition to these metrics, many studies have extended to collecting multiple marker types such as genetic parameters, morphometric data etc. to define groups (Higgins et al. 2010). Many of these variables, particularly in the case of otolith chemistry, can be affected by factors such as temperature, salinity, and physiological conditions (Bath et al. 2000, Walther and Thorrold 2006, Dorval et al. 2007). Consequently, markers may convey disproportionate amounts of information relevant in delineating mixtures (See Chapter II). Some variables may not express any relevant information, while others may convey similar information and including these variables can reduce classification success (Van Ness and Simpson 1976). Therefore, model selection procedures that can identify the best model from a suite of variables can
be an important step in determining the most informative variables for the system.

Presently there are several approaches used to determine the "best" model from training data. The discriminatory model which maximizes known origin juvenile fish classification is frequently used to determine the natal source of subsequent adult recruits. A main assumption of this approach is that juvenile misclassification will be reflected in the following adult classifications. However, this model may be overfit to the training dataset and may not perform as well as anticipated with adult classification. In other words, using this minimum error model may not be useful in generalizing to the test data and determining group membership probabilities from unknown mixtures can be inaccurate (Hastie et al. 2009). Furthermore, as the proportions of adults from each area are unknown, this bias will go unnoticed.

Another approach to building the discriminatory function has been to use only the variables whose means are significantly different between groups as determined by analysis of variance (ANOVA). More ad hoc methods such as choosing all variables that can be successfully collected or, using variables that provide high classification accuracy in previous studies have also been used. Although some of these methods such as the minimum error model and ANOVA approach are an improvement to the ad hoc approaches mentioned, they may not maximize classification accuracy of test datasets as they do not directly address model complexity. Using a model selection technique to determine the combination of variables that closely characterizes each group, will be more powerful in assigning group membership of unknown entries (McLachlan 2004).

Selection procedures such as Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) are widely used in other fields of biology and is becoming
more prevalent in ecological studies (Johnson and Omland 2004). Both criteria measure
goodness of fit while penalizing for the number of model parameters entered. This
penalty ensures that models are not over parameterized and those with the smallest AIC
and BIC values are considered to be both powerful and parsimonious (Burnham and
Anderson 2002). These selection criteria have not been applied to otolith chemistry to
identify natal source but may offer a significant advantage over more commonly used
methods of building discriminant functions.

In this study I highlight the importance of determining the best model from
known individuals to optimize correct classification of unknown fish. Using observed
chemical markers from juvenile spotted seatrout otoliths (Cynoscion nebulosus), I
evaluated, for the first time, the power of AIC and BIC model selection techniques for
optimizing classification of simulated adults from each natal source. Specifically, I
employed both AIC and BIC selection criteria to determine the most parsimonious model
from data that exhibit heterogeneous variance-covariance matrices, a common occurrence
in otolith chemical datasets (Mercier et al. 2011). Using AIC, BIC, and the minimum
error models from training data, I simulated “unknown” fish drawn from the juvenile
distribution and determined classification accuracy using jackknifed quadratic
discriminant function analysis (QDFA). I emphasize the potential problems that can arise
when classification models are constructed without evaluating model complexity.

Material and Methods

Fish collection

I collected juvenile spotted seatrout from the Eastern Shore (ES) and Western
Shore (WS) seagrass habitats in the lower Chesapeake Bay with sampling locations nested within each habitat (Fig. 3). Fish were sampled at each location using an 8m otter trawl every two weeks from July to October during 2008. Six tows, each two minutes in length, were conducted at all locations. All fish collected were kept on ice until returned to the laboratory and frozen. Total and standard lengths were measured and fish were weighed before the otoliths were removed (See Chapter II for further details).

*Trace element chemistry*

Seatrat otoliths were extracted in a positive flow class 100 clean room following standard clean room protocol (see Dorval et al. 2005b). For trace element analysis, otolith solutions were analyzed using a Finnigan MAT Element 2 Inductively Coupled Plasma Mass Spectrometer (ICPMS). Samples were randomized within trays to account for machine drift and $^7$Li, $^{25}$Mg, $^{48}$Ca, $^{55}$Mn, $^{86}$Rb, $^{89}$Y, $^{138}$Ba, and $^{139}$La were measured under low resolution while $^{48}$Ca, and $^{88}$Sr measured under medium resolution mode (Chen and Jones 2006). The concentration of each element was calibrated using the NRC otolith standard (Sturgeon et al. 2005) and multi-element standards and instrument blank solutions of 2% HNO$_3$ were run every six samples and linearly interpolated between sample runs (Dorval et al. 2005b, Schaffler and Winkelman 2008). The limits of detection (LOD) were calculated as the mean blank values plus three standard deviations (Thorrold et al. 1997b, Wells et al. 2003). For elements that contained more than 25% of the samples below LOD were excluded from analyses (See Chapter II for further details). As otoliths are primarily composed of calcium, all elements were normalized to Ca and expressed as element-to-calcium ratios.
FIG. 3. Locations of sampling stations in the Chesapeake Bay: Western Shore (●) and Eastern Shore (●).
Stable isotopes

The remaining otolith was homogenized using an acid rinsed ceramic mortar and pestle and the resulting powder assayed for carbon (δ^{13}C) and oxygen (δ^{18}O) stable isotope concentrations using an automated Isoprime Micromass carbonate analyzer. Otolith samples and limestone standards were spiked with 102% phosphoric acid at 90°C with five standards for every ten otolith samples (Dorval et al. 2007). Isotope concentrations were measured based on the dual inlet method and coldfinger mode and data were corrected and stated in PDB (Peedee Belemnite) (See Chapter II for further details).

Model Selection

I determined whether data conformed to multivariate normality using Mardia’s test based on multivariate skewness and kurtosis (Khattree and Naik 1999). As elemental signatures did not conform to normality, Box-Cox transformations were applied to all variables in the dataset. To detect differences in the elemental signature between ES and WS, I used a multivariate analysis of variance (MANOVA). If significant differences in the multivariate signatures were found, I determined which variables were responsible for the differences using univariate analysis of variance (ANOVA). Additionally, to evaluate the power of each variable, the Mahalanobis distance ($D^2$) was adjusted to incorporate the heterogeneous variance-covariance matrices among groups using the following:

$$D^2 = (\mu_{ES} - \mu_{WS})'(\Sigma_{ES} + \Sigma_{WS})^{-1}(\mu_{ES} - \mu_{WS})$$

Several tests of information conveyed by variables rely on this distance calculation (see Rao 1965 for example) and this modification will in turn allow for the separation
distance among groups resulting from each individual variable to be determined.

Jackknifed quadratic discriminant function analysis (leave-one-out approach) was used for all combinations of all variables from the training dataset to determine the minimum error model. From this list of models, I also used AIC and BIC selection criteria to evaluate each model and determined how well these models performed against the minimum error model for the juvenile dataset using the following:

\[ AIC(J_k) = \sum_{g=1}^{q} N_g \log [\Lambda_g(J_k)] + (q - 1)k(k + 3) \]

\[ BIC(J_k) = \sum_{g=1}^{q} N_g \log [\Lambda_g(J_k)] + 0.5 (q - 1)k(k + 3) \log N \]

where \( g \) = number of groups, \( k \) = subset of variables used in the model out of \( K \), and \( \Lambda_g(J_k) = \sum_{g=1}^{q} \frac{1}{N_g W_k^{(g)}} \), where \( W_k^{(g)} \) is the maximum likelihood estimate of the partitioned sum-of-squares and cross-products using a \( k \) subset of variables for each group \( (g = ES \ or \ WS) \), and \( T_k \) is the maximum likelihood estimate of variance-covariance matrix corresponding to all variables in the model (Pynnönen 1987).

To determine the performance of each model in assigning fish of unknown origin to nursery habitat I simulated data sets each containing fish of "unknown" origin using the sample parameters from ES and WS juveniles. These simulated "unknown" fish were drawn from a multivariate normal distribution structured by the same mean and variance-covariance matrices as the juvenile fish. These simulated fish represent the subsequent adults from each habitat. Although I collected approximately equal proportions of juvenile fish between both ES and WS seagrass beds, this ratio may not be maintained in subsequent "unknown" collections. Therefore, I simulated seven datasets each with
different group sample size from each habitat while maintaining a total sample size of 100 fish. The datasets each contained “unknown” fish in the following proportions of ES to WS: 100:0, 80:20, 60:40, 50:50, 40:60, 20:80, 0:100. All simulated datasets were classified to natal habitats using each of the three models (AIC, BIC and minimum error) as determined from the juvenile training data. Overall performance of these jackknifed QDFA classification models were compared using misclassification error.

Results

I assayed a total of 106 juvenile spotted seatrout for this study (56 from ES, 50 from WS). All trace elements except lanthanum were above detection limits, which was subsequently removed from statistical analysis. Of the remaining nine discriminatory variables (seven trace elements and two stable isotope ratios), all were found to be univariate normal using the Shapiro-Wilks test for normality. Moreover, after Box-Cox transformation these data were also found to be multivariate normal using Mardia’s test. From the MANOVA, there was a significant difference in overall variable concentrations between ES and WS seagrass habitats (Pillai’s Trace = 0.77; $F_{9,96} = 35.4; P < 0.0001$). However, all variables were not significantly different between zones. From the univariate ANOVA, only manganese (Mn), rubidium (Rb), yttrium (Y), strontium (Sr), barium (Ba), and $\delta^{13}$C showed significant differences in mean concentration between ES and WS seagrass habitats.

From the modification of the Mahalanobis distance, I found that all variables were not equal in their power to classify fish. Distances ($D^2$) indicated that the most important variables in classifying fish to habitat were Ba and $\delta^{13}$C. I found that Ba was the most
powerful discriminatory variable with a distance of 137.5 (Table 4). In a QDFA classification model using barium as the only discriminatory variable for juvenile fish, I obtained an overall error rate of only 15.1%. The next most important variable δ\(^{13}\)C had a \(D^2\) value of 58.1. This large separation distance of carbon also reflected its classification power, as a model using only carbon resulted in an overall error rate of 21.7%. Less powerful variables were the minor elements rubidium (Rb) and manganese (Mn).

The best model selected by both AIC and BIC using the juvenile dataset differed by a few variables and in their classification success (Table 5). The model that performed the best based on AIC included 7 variables: Ba, δ\(^{13}\)C, Mn, Sr, Mg, δ\(^{18}\)O, and Li (ordered by \(D^2\)). Strong discriminators Ba and δ\(^{13}\)C were chosen in this model alongside less powerful variables Mn and Sr. The overall error rate for this model was 7.5% with an

<table>
<thead>
<tr>
<th>Variable</th>
<th>(D^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ba</td>
<td>137.5</td>
</tr>
<tr>
<td>δ(^{13})C</td>
<td>58.1</td>
</tr>
<tr>
<td>Rb</td>
<td>31.4</td>
</tr>
<tr>
<td>Mn</td>
<td>19.3</td>
</tr>
<tr>
<td>Y</td>
<td>6.8</td>
</tr>
<tr>
<td>Sr</td>
<td>6.1</td>
</tr>
<tr>
<td>Mg</td>
<td>3.9</td>
</tr>
<tr>
<td>δ(^{18})O</td>
<td>3.0</td>
</tr>
<tr>
<td>Li</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 4. Mahalanobis distances \((D^2)\) for all variables as a measure of variable importance.
accuracy of 91.0% on the ES and 94.0% on the WS for juvenile fish.

The model that performed the best according to the BIC method included 5 variables: Ba, δ^{13}C, Mn, Sr, and Mg. This model had an overall error rate of 8.4% with an accuracy of 89.2% on the ES and 94.0% on the WS for juvenile fish. The model that produced the lowest overall error contained six variables: Ba, Mn, Sr, Mg, δ^{18}O, and Li ordered by variable importance. Using this combination of variables for the juvenile dataset, I obtained an overall error rate of 4.7% with a success rate of 94.6% on the ES and 96% on the WS seagrass habitats.

Using each of these three models to classify the simulated “unknown” data back to natal origin I found the best model to determine group membership of fish was the AIC selection (Fig. 4). Even though this model did not have the lowest jackknifed error rate for the juvenile dataset, it performed the best overall in almost all of the simulated data sets. The only mixture where the AIC did not perform the best was the mixture that contained 20 fish from the ES and 80 from WS. However, even in this case it was the second best performing model. The BIC selection performed the next best while the

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Value</th>
<th>Juvenile Classification Error%</th>
<th>Variable Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best QDFA</td>
<td>n/a</td>
<td>4.7</td>
<td>Ba, Mn, Sr, Mg, δ^{18}O, Li</td>
</tr>
<tr>
<td>AIC</td>
<td>-172.99</td>
<td>7.5</td>
<td>Ba, δ^{13}C, Mn, Sr, Mg, δ^{18}O, Li</td>
</tr>
<tr>
<td>BIC</td>
<td>-103.30</td>
<td>8.4</td>
<td>Ba, δ^{13}C, Mn, Sr, Mg</td>
</tr>
</tbody>
</table>
Fig. 4. Overall error rate of simulated “unknown” fish to natal source using AIC (♦), BIC (□), and minimum error (×) models for each combination of simulated “unknowns”. Axis labels indicate the number of fish simulated from ES with the remaining fish out of 100 simulated from the WS.

Discussion

These simulations demonstrate the importance of selecting the subset of variables that are the most informative in classifying unknown individuals. In constructing a discriminant function, I found that the most parsimonious model to assign “unknown”
fish to natal origin was from the AIC section procedure. This model performed the best across a variety of “unknown” mixtures from the same distributions as the juvenile fish (error of misclassification < 6% in all simulations), even though this model did not maximize classification of the juveniles. Similarly, the BIC selection also performed well when classifying “unknown” fish to natal sources. However, the most commonly applied model selection procedure in the literature, the minimum error model from the juvenile training set, performed the worst.

Model selection is an important aspect to determining the most powerful and parsimonious model (McLachlan 2004). The minimum error model from the training dataset did not maximize classification success of the “unknowns” even though this combination of variables was the most accurate at assigning juvenile fish. These results confirm that training error is often not a good estimate of the test error. A minimum test error is produced at a specific model complexity as there is a tradeoff between bias (the distance between estimates and truth) and variance (the spread of the estimate around truth) in determining the most powerful model (Hastie et al. 2009). In this study I show that the model which produced a low classification error for the juvenile training set was over fit to these data and could not generalize adequately to the adult test data. In other words, the minimum error model from the juvenile dataset is highly specific to that sample, but not to the overall population.

The methods presented here offer an important advantage over some of the more commonly used methods to variable selection. Ad hoc approaches such as using all variables collected can introduce redundancy in the model. The performance of the discriminant rule does not improve with the addition of more discriminatory variables
(Williams 1983). In fact classification increases to a threshold and subsequently declines (McLachlan 2004). This peaking phenomenon is often disregarded by researchers but is an important issue in building a discriminant function. Additionally, using the minimum error rate as a selection criterion can be worthwhile practice, particularly when model selection techniques are not employed. However, as I demonstrate information regarding the underlying structure of the system may not be conveyed and the model can be over-parameterized.

The use of model selection techniques is particularly important in marine ecology. As technological advances allow for large amounts of information to be the collected, high dimensional datasets are routinely analyzed. For example, Higgins et al. (2010) collected 7 metrics each with a suite of variables to delineate the harvest origin of cod (*Gadus morhua*). In these large datasets model selection techniques can identify useful subsets of variables. In turn, this can reduce the burden of collecting costly metrics in the next sampling cycle that may not increase the separation distance between groups (Wood et al. 2005). Additionally, selection criteria can allow researchers to evaluate several competing models that may represent unique hypotheses (Burnham and Anderson 2001, Johnson and Omland 2004).

In this study I show that the correct assignment of unknown individuals to their respective groups relies heavily on the combination of variables used in the discriminatory function. From the Mahalanobis distance, I found a wide range of discriminatory power among variables. Barium was the most important variable in this system as noted by the large $D^2$ and it was selected by both AIC and BIC selection procedures. The importance of Ba to this system is not surprising as both shores exhibit
unique hydrology and the incorporation of Ba into the otolith is strongly linked to environmental properties (Dorval et al. 2005a). In the Chesapeake Bay, the dissolution of Ba is linked to the interaction of marine water and sediments. As oceanic waters enter the estuary and move across seagrass habitats, particularly on the ES, Ba is released from sediments (Coffey 1997). In the variable selection study using a linear dataset, I also found that Ba was the most important variable in separating seagrass habitats within the Chesapeake Bay (see Chapter II). Moreover, other studies that have used otolith microchemistry to identify movement for other species in the Bay also relied on Ba as a discriminatory variable (Thorrold et al. 1998b, Thorrold et al. 2001).

An important secondary discriminatory variable was δ^{13}C. This marker has been used in a variety of studies that determine the natal sources of fish (Thorrold et al. 2001, Rooker et al. 2008b). Research by Hanson et al. (2004) using otolith chemistry to classify fish to their nursery area found that δ^{13}C was an important discriminant marker. As the incorporation of δ^{13}C in fish otolith is regulated by a variety of factors such as water composition, temperature and diet it is unclear as to what are factors driving the difference in chemistry between ES and WS nursery areas. However, in applying the otolith tagging method it is not necessary to fully understand the drivers that make variable concentration unique among locations; the only requirement is that there is a difference in concentrations in order for that variable to be successfully used as habitat marker (Gillanders 2005b).

The ability to identify group membership of unknown fish is an important factor in understanding population dynamics and maintaining fishable stocks. Having methods that can accurately determine natal sources, migratory paths, or population structure are
relevant issues in marine ecology. I show the importance of model selection techniques to maximize the classification of unknown individuals and further demonstrate the importance of selecting informative variables to build the most parsimonious and powerful from the training dataset. This methodology has significant potential for accurately identifying the contributions of recruits from multiple sources in a given system.
CHAPTER IV

DEVELOPING A MULTI-YEAR HABITAT TAG FOR DETERMINING THE NATAL SOURCES OF FISH

Introduction

Nursery habitats have long been recognized for their central role in maintaining fishable populations and preserving ecosystem dynamics. Juvenile fish that settle on nursery areas are often conferred a survival advantage, as structurally complex areas provide shelter, and support a diverse prey field for developing fish (Orth and Moore 1984, Dahlgren et al. 2006). Globally, there has been widespread degradation of many coastal ecosystems where nursery habitats are located (Hinrichsen 1998). Loss of these nursery areas can have significant impacts on dependent fish populations, as many marine and estuarine populations are supplied by recruits from several nursery areas, with some locations contributing disproportionately more fish to the adult stock (Beck et al. 2001). For these populations, identifying the natal source of adults, and evaluating the long term contribution of different nursery areas to the adult population, is an important step for effective management (Gillanders et al. 2003). Consequently, much attention has focused on methods that determine natal sources of adult fish, with natural markers derived from the otolith chemistry emerging as powerful technique in identifying provenance.

Otolith chemistry can offer significant insight into the past environmental history experienced by the fish (Campana 2005). As the fish grows, trace elements and other
compounds from the surrounding water are incorporated into the calcium carbonate matrix of the otolith. Several experiments have demonstrated the correlation between trace elemental concentration in the otolith and ambient conditions for certain elements (Fowler et al. 1995, Secor et al. 1995, Bath et al. 2000, Walther and Thorrold 2006, Gibson-Reinemer et al. 2009). The elemental composition of the otoliths is also structured by physical parameters such as temperature and salinity that are often unique to a habitat (Collingsworth et al. 2010, Phillis et al. 2011). For example, Bath et al. (2000) showed that although otolith strontium (Sr) and barium (Ba) concentrations in otoliths of spot (*Leiostomus xanthurus*) reflected the relative proportion of these elements in ambient waters, temperature significantly affected incorporation of Sr. In the case of δ¹⁸O, ratios often reflect ambient conditions while the majority of δ¹³C deposited in otoliths is derived from dissolved inorganic carbon (DIC) with a small proportion coming from metabolic sources (Kalish 1991b, Thorrold et al. 1997a, Solomon et al. 2006). As otolith growth is continuous, and the deposited material is not reabsorbed, the chemical composition is considered a suitable tag for revealing patterns of origin and for retrospectively tracking movement in fish (Campana 1999, Gillanders 2005b, Dorval et al. 2007).

Because of the environmental and physiological information conveyed by the chemical signature of the otolith, the trace element and stable isotope composition is now commonly used in studies that evaluate provenance (Gillanders and Kingsford 1996, Yamashita et al. 2000, Thorrold et al. 2001, Rooker et al. 2008a, Wright et al. 2010). However, to successfully use a chemical tag to identify natal origin in populations with multiple nursery sources, it is first important to determine the chemical variables that best
assign fish to nursery areas and the temporal scale over which this tag can be applied. As the chemical composition of otoliths is influenced by a variety of factors, it is not surprising that chemical variables often convey disproportionate amounts of information regarding the habitat (Mercier et al. 2011). Therefore, selecting informative variables will result in a classification model that ideally captures the contrasts among areas. Additionally, carefully choosing variables to develop a parsimonious yet powerful classification model will allow for improved ability to classify unknown adults correctly and provide useful information regarding temporal variation of elemental concentrations across a variety of time scales.

Numerous studies have addressed temporal stability for otolith signatures in both yearly (Rooker et al. 2001, Dorval et al. 2005b, Mateo et al. 2010) and monthly time scales (Hamer et al. 2003). In many of these applications, there was significant variation in chemical signatures across years (Rooker et al. 2001, Gillanders 2002). This variation in chemical signatures limits the application of the habitat tag, as juveniles must be sampled each year to construct a habitat marker for subsequent adults. This temporal variation in habitat chemistry can be attributed to the multiple processes that structure the chemical environment on each nursery area. However even in very dynamic systems such as estuaries, where processes such as river discharge, tidal flow, lithology etc. contribute unique components to the overall chemical signature, the relationship among elements can remain stable over time (Smith 1977, Shumilin et al. 1993). In these instances, using model selection criterion to elucidate important variables from the system can lead to the development of a stable tag that can delineate habitats across longer time scales.

In this study, I determined whether a multi-year chemical tag for nursery habitats
in Chesapeake Bay can be retrospectively applied to historically short-term datasets. Specifically, I identify 1) a single best model for characterizing seagrass beds across the lower Bay using an AIC model selection procedure for each individual year, and 2) a multi-year habitat tag from a time series of otolith chemistry data. Temporal consistency in a habitat tag can greatly influence the way in which productive nursery areas are identified and managed. If a long-term tag exits, contributions of each natal source can be determined over a broader time scale. This information will allow for better management of critical nursery areas. Furthermore, as chemical analyses of otoliths are both time-consuming and costly, establishing a multi-year tag can make otolith chemical markers a practical solution to identifying nursery sources of adult fish.

Material and Methods

Fish collection

I collected juvenile spotted seatrout (*Cynoscion nebulosus*) from two major habitat zones in the lower Chesapeake Bay: the Eastern Shore (ES) and Western Shore (WS). These habitat zones ranged across the entire lower Bay with sampling locations nested within each zone (Fig. 5). Seatrout were sampled at each location from July to October in 2006 – 2008 using an 8 m otter trawl with 0.64 cm stretched mesh. Each location was sampled every two weeks with six 2 minute tows on each seagrass bed. Tow distances and locations were measured using a GPS navigation system. All fish were stored in Whirl-Pak bags and kept on ice until returned to the laboratory and subsequently frozen. Total and standard lengths (mm) were measured and fish were weighed (g) before the otoliths were extracted (see Chapter II for further details). Historical juvenile seatrout
collections from 2001–2002 were also used for this analysis and fish were collected using the same sampling design as the recent 2006–2008 collections.

**Trace element chemistry**

Otoliths were prepared for trace element assay following protocols stated in Dorval et al. (2005b, 2007). I analyzed otolith solutions using a Finnigan MAT Element 2 Inductively Coupled Plasma Mass Spectrometer (ICPMS). I measured $^7$Li, $^{25}$Mg, $^{48}$Ca, $^{55}$Mn, $^{86}$Rb, $^{89}$Y, $^{138}$Ba, and $^{139}$La in low resolution and $^{48}$Ca and $^{88}$Sr in medium resolution mode (Chen and Jones 2006). Element concentrations were calibrated using the NRC otolith standard (Sturgeon et al. 2005) and multi-element standards prepared from ultrapure stock solutions (Dorval et al. 2005b, Schaffler and Winkelman 2008). Instrument blank solutions of 2% HNO$_3$ were run every six samples and the limits of detection (LOD) were calculated as the mean blank values plus three standard deviations (Thorrold et al. 1997b, Wells et al. 2003). Elements that contained more than 25% of the samples below the LOD were removed from the analyses (See Chapter II for further details). As otoliths are primarily composed of calcium, all elements were normalized to Ca and expressed as element-to-calcium ratios.

**Stable isotope chemistry**

The remaining sagittal otolith was analyzed for carbon ($\delta^{13}$C) and oxygen ($\delta^{18}$O) concentrations. Otolith halves were homogenized and approximately 60-100 µg of otolith powder was used for analysis. Otolith samples and limestone standards were spiked with 102% phosphoric acid at 90°C with five standards for every ten otolith samples.
FIG. 5. Locations of sampling stations in the Chesapeake Bay: Western Shore (●) and Eastern Shore (◆).
Isotope concentrations were measured based on the dual inlet method and coldfinger mode (See Chapter II for further details). Data were corrected and stated in PDB (Peedee Belemnite).

**Statistical analysis**

Variables in each year class were first tested for multivariate normality using Mardia's test for multivariate skewness and kurtosis (Khattree and Naik 1999). Data did not meet multivariate normality and all variables were transformed using a Box-Cox transformation until residuals conformed to normality. To investigate elemental differences between habitats for individual year classes, I transformed the data for each year separately using Box-Cox transformations to achieve normality. Using these transformed data, I compared the elemental composition between ES and WS seagrass habitats in individual year classes using a univariate analysis of variance (ANOVA). To examine overall differences between habitat (ES and WS), all year classes were combined and simultaneously transformed using Box-Cox. These data were then tested for habitat effects using multivariate analysis of variance (MANOVA).

Having a multi-year chemical tag would allow for nursery habitats of fish to be correctly identified regardless of the time period in which it was caught, or the cohort to which it belonged. In order to correctly classify fish in dynamic systems, the discriminatory function should be constructed using informative variables. To best determine whether a single discriminatory function can be built for all years, I first investigated the elemental concentrations that best characterize each year using AIC model selection procedures as first described by Pynnönen (1987) (See Chapter III).
\[ AIC(k) = \sum_{g=1}^{q} N_g \log[A_g(J_k)] + (q - 1)k(k + 3) \]

where \( g = \) number of groups, \( k = \) subset of variables used in the model out of \( K \), and

\[ A_g(J_k) = \sum_{g=1}^{q} \frac{1}{N_g} \frac{W_k^{(g)}}{|T_k|}; \text{where } W_k^{(g)} \text{ is the maximum likelihood estimate of the} \]

partitioned sum-of-squares and cross-products using a \( k \) subset of variables for each group \( (g = \text{ES or WS}) \), and \( T_k \) is the maximum likelihood estimate of variance-covariance matrix corresponding to all variables in the model. Using this criterion, a single best discriminant model was constructed for each year. Using each year-specific model, juveniles were classified to ES or WS habitats using jackknifed (leave-one-out) quadratic discriminant function analysis (QDFA). The classification accuracy for each year class was then determined by comparing the predicted group membership to the actual group membership. From these models, elements that repeatedly show significant differences are a good indication of a stable habitat tag over time. Additionally, to evaluate the power of each variable, I modified the Mahalanobis distance \( (D^2) \) equation to determine whether specific variables were better at separating fish across multiple years (see Chapter III).

To determine whether a multi-year model can be constructed for the Chesapeake Bay seatrout population, data from 2006, 2007, and 2008 were pooled to form the training data for the multi-year tag. Using this dataset, AIC selection procedures were employed to obtain a single best discriminatory model. To determine if this multiyear tag could be successful applied to other year classes, I used a jackknifed QDFA to calculate the classification probability for juvenile fish collected in 2001 and 2002. Classification
success was determined by the percentage of fish that correctly classified to their natal source. All analyses were completed using SAS software.

**Results**

*Signatures for individual years*

A total of 444 juvenile spotted seatrout were used in this study. There was variability in the numbers of fish collected in each year, however greater than 30 samples (considered to be a large sample size) were collected for each habitat zone in all years (Table 6). In many years, lanthanum was below detection limits and was subsequently removed from all analyses. A total of seven trace element ratios (Li, Mg, Mn, Rb, Sr, Y, and Ba) and two stable isotope ratios ($\delta^{13}C$, $\delta^{18}O$) were used in all years for statistical analyses. In all year classes, data did not meet multivariate normality and data for each year were transformed using Box-Cox transformations. For each year, I found a significant habitat effect using MANOVA (Table 7). Investigation of the elemental concentrations that led to this overall difference between ES and WS using ANOVA showed that many elements were significantly different between these two habitat zones (Table 6). However, several elements were found to have no difference with respect to their mean concentrations between habitats in multiple years. Rubidium and Sr showed no concentration differences in three of the five year classes. Similarly, Mn, Li, and $\delta^{18}O$ were found to have no concentration differences in two years, with Mg showing no difference in only one year class (Table 6).

The suite of discriminatory variables selected to correctly classify fish differed slightly between years (Table 8). There was considerable overlap in variables in each
TABLE 6. Total number of fish collected for each habitat (ES and WS) and the ANOVA results of variables that show significant differences or no difference in mean concentrations between habitats for all years ($\alpha = 0.05$).

<table>
<thead>
<tr>
<th>Year</th>
<th>ES</th>
<th>WS</th>
<th>Total</th>
<th>Number of fish</th>
<th>ANOVA results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Significant elements</td>
</tr>
<tr>
<td>2001</td>
<td>43</td>
<td>42</td>
<td>85</td>
<td>Mg, Mn, Ba, Y, $\delta^{18}$O</td>
<td>Li, Rb, Sr, $\delta^{13}$C</td>
</tr>
<tr>
<td>2002</td>
<td>44</td>
<td>44</td>
<td>88</td>
<td>Li, Mg, Rb, Y, Ba, $\delta^{13}$C, $\delta^{18}$O</td>
<td>Mn, Sr</td>
</tr>
<tr>
<td>2006</td>
<td>40</td>
<td>23</td>
<td>63</td>
<td>Li, Y, Ba, $\delta^{13}$C, $\delta^{18}$O</td>
<td>Mg, Mn, Rb, Sr</td>
</tr>
<tr>
<td>2007</td>
<td>68</td>
<td>34</td>
<td>102</td>
<td>Li, Mg, Mn, Sr, Y, Ba, $\delta^{13}$C</td>
<td>Rb, $\delta^{18}$O</td>
</tr>
<tr>
<td>2008</td>
<td>56</td>
<td>50</td>
<td>106</td>
<td>Mg, Mn, Rb, Sr, Y, Ba, $\delta^{13}$C</td>
<td>Li, $\delta^{18}$O</td>
</tr>
</tbody>
</table>

TABLE 7. Overall habitat effect for each year using MANOVA ($\alpha = 0.05$).

<table>
<thead>
<tr>
<th>Year</th>
<th>Statistic</th>
<th>Value</th>
<th>$F$</th>
<th>Num Df</th>
<th>Den Df</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>Wilks' Lambda</td>
<td>0.21</td>
<td>30.56</td>
<td>9</td>
<td>75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Pillai's Trace</td>
<td>0.79</td>
<td>30.56</td>
<td>9</td>
<td>75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Hotelling-Lawley Trace</td>
<td>3.67</td>
<td>30.56</td>
<td>9</td>
<td>75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>2002</td>
<td>Wilks' Lambda</td>
<td>0.27</td>
<td>23.49</td>
<td>9</td>
<td>78</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Pillai's Trace</td>
<td>0.73</td>
<td>23.49</td>
<td>9</td>
<td>78</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Hotelling-Lawley Trace</td>
<td>2.71</td>
<td>23.49</td>
<td>9</td>
<td>78</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>2006</td>
<td>Wilks' Lambda</td>
<td>0.23</td>
<td>19.67</td>
<td>9</td>
<td>53</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Pillai's Trace</td>
<td>0.77</td>
<td>19.67</td>
<td>9</td>
<td>53</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Hotelling-Lawley Trace</td>
<td>3.34</td>
<td>19.67</td>
<td>9</td>
<td>53</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>2007</td>
<td>Wilks' Lambda</td>
<td>0.45</td>
<td>12.31</td>
<td>9</td>
<td>92</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Pillai's Trace</td>
<td>0.55</td>
<td>12.31</td>
<td>9</td>
<td>92</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Hotelling-Lawley Trace</td>
<td>1.20</td>
<td>12.31</td>
<td>9</td>
<td>92</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>2008</td>
<td>Wilks' Lambda</td>
<td>0.24</td>
<td>33.86</td>
<td>9</td>
<td>96</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Pillai's Trace</td>
<td>0.76</td>
<td>33.86</td>
<td>9</td>
<td>96</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Hotelling-Lawley Trace</td>
<td>3.17</td>
<td>33.86</td>
<td>9</td>
<td>96</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
AIC model: Mg, Li, Ba, and $\delta^{18}$O were used in the classification function in all five years, while Mn, Rb, Sr, Y, and $\delta^{13}$C were chosen in some. Using the AIC models to classify juvenile fish to natal source, each individual year’s AIC model accurately classified fish with an overall error of less than 10%. Moreover, I found that in all cases successful classification of juvenile fish to the ES seagrass habitats was slightly higher than WS seagrass beds in all years, with the exception of 2008.

Assessing the elemental distribution on each habitat zone, box plots showed that some of the elements were considerably different between ES and WS habitats (Fig. 6). For example, Ba concentration on the ES was considerable higher than on the WS and this contrast was maintained across all five year classes. In examining discriminatory power using the modified $D^2$, I found that Ba was the most informative variable. This element was consistently the most powerful discriminator with the exception of 2002

<table>
<thead>
<tr>
<th>AIC Value</th>
<th>Best model</th>
<th>Classification Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 -219.5 Li, Mg, Mn, Rb, Sr, Y, Ba, $\delta^{18}$O</td>
<td>93.0 90.5 8.3</td>
<td></td>
</tr>
<tr>
<td>2002 -170.4 Li, Mg, Rb, Sr, Y, Ba, $\delta^{13}$C, $\delta^{18}$O</td>
<td>90.9 90.9 9.1</td>
<td></td>
</tr>
<tr>
<td>2006 -151.9 Li, Mg, Mn, Ba, Y, Sr, $\delta^{13}$C, $\delta^{18}$O</td>
<td>100.0 91.3 4.4</td>
<td></td>
</tr>
<tr>
<td>2007 -133.6 Li, Mg, Mn, Ba, Rb, $\delta^{13}$C, $\delta^{18}$O</td>
<td>95.6 85.3 9.6</td>
<td></td>
</tr>
<tr>
<td>2008 -173.0 Li, Mg, Mn, Ba, Sr, $\delta^{13}$C, $\delta^{18}$O</td>
<td>91.1 94.0 7.5</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 8. The best model for each year determined by the AIC selection method and the correct classification probability for each shore.
Fig. 6. Box plots of untransformed otolith elements for ES (dark bars) and WS (open bars) across a five year time period for the Chesapeake Bay. Box represent range of 1st and 3rd quartiles while whiskers indicate the 90% and 10% of the data.
and 2006, where Ba was the third and second most important variable respectively (Table 9).

Similarly, fish collected on ES habitats maintained higher concentrations of Y in their otoliths than fish collected on WS grass beds. For other elemental variables, the difference between ES and WS fish were not as consistent. For example, $\delta^{13}C$ was higher on the WS seagrass beds in the majority of the years with the exception of 2001 when concentrations were very similar and 2007 where the ES fish maintained a higher $\delta^{13}C$ concentration than WS fish. Conversely, for some variables such as Mg, the differences in concentrations changed over time, but the direction of the contrast was consistent.

**Developing a multiyear habitat tag**

The pooled 2006-2008 data achieved multivariate normality after Box-Cox transformations. Results of MANOVA for these pooled data indicated that chemical

<table>
<thead>
<tr>
<th>Year</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>Ba</td>
<td>Y</td>
<td>$\delta^{18}O$</td>
</tr>
<tr>
<td>2002</td>
<td>$\delta^{18}O$</td>
<td>$\delta^{13}C$</td>
<td>Ba</td>
</tr>
<tr>
<td>2006</td>
<td>Y</td>
<td>Ba</td>
<td>$\delta^{18}O$</td>
</tr>
<tr>
<td>2007</td>
<td>Ba</td>
<td>Sr</td>
<td>$\delta^{13}C$</td>
</tr>
<tr>
<td>2008</td>
<td>Ba</td>
<td>$\delta^{13}C$</td>
<td>Rb</td>
</tr>
</tbody>
</table>

**Table 9.** Three most important variables for separating groups in each year with corresponding Mahalanobis distance ($D^2$) for ES and WS habitats.
signatures differed significantly between habitats (Pillai’s trace = 0.58; $F_{9,257} = 39.0$; $P < 0.0001$) and there were significant year (Pillai’s trace = 0.93; $F_{18,516} = 25.1$; $P < 0.0001$), and year × habitat effects (Pillai’s trace = 0.48; $F_{18,516} = 9.14$; $P < 0.0001$). Additionally, the ANOVA for these pooled data indicated that all otolith elemental concentrations differed significantly between ES and WS fish. The AIC procedure selected the best multi-element model as: Mn, Rb, Sr, Y, Ba, $\delta^{13}$C, $\delta^{18}$O. The $D^2$ values for these pooled data also reflected elements that were important in separating fish in the individual years. Barium ($D^2=153.2$) was the most important variable followed by Y ($D^2=22.4$) and third most important was Mn ($D^2=17.4$).

The AIC multi-year tag performed well in separating fish from 2006-2008. In 2006, fish were correctly classified to ES habitat with 100% accuracy while the WS habitat had 78.3% accuracy; in 2007, fish classified correctly to ES with 82.4% and to

<table>
<thead>
<tr>
<th>Year</th>
<th>Classification success</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES</td>
</tr>
<tr>
<td>2001</td>
<td>88.4</td>
</tr>
<tr>
<td>2002</td>
<td>70.5</td>
</tr>
<tr>
<td>2006</td>
<td>100.0</td>
</tr>
<tr>
<td>2007</td>
<td>82.4</td>
</tr>
<tr>
<td>2008</td>
<td>91.1</td>
</tr>
</tbody>
</table>

TABLE 10. Classification success using a multi-year, multi-element discriminatory tag derived from fish spawned in 2006 – 2008 to delineate habitats across all years, including years where fish were not used to develop the tag (depicted in bold).
WS with 91.2%; in 2008, there was 91.1% correct assignment to ES and 90.0% to WS (Table 10). In applying this tag to classify juvenile fish spawned in 2001 and 2002, classification success was 88.4% to ES seagrass habitats and 92.7% for WS in 2001, while fish were correctly classified to natal habitat with 70.5% accuracy to ES and 88.6% to WS in 2002 (Table 10).

Discussion

Important variables

The differences in otolith chemistry between ES and WS fish were primarily driven by Ba levels in otoliths. Dorval et al. (2007) showed that Ba concentrations in the Chesapeake Bay is largely regulated by salinity. Additionally, the spatial differences in Ba between ES and WS seagrass beds can also be attributed to the conservative behavior of this element in estuaries and the unique hydrology of the region (Coffey et al. 1997). The Western shore is predominantly influenced by riverine discharge, while the Eastern shore seagrass beds are regulated by marine waters (Boicourt et al. 1999). As Ba release occurs through particle interaction with oceanic waters, riverine and other sediments accumulated on seagrass beds will slowly release Ba as seawater moves across the habitat (Coffey et al. 1997). Therefore, with the stratified, unidirectional circulation of oceanic waters moving into the Bay and the movement of freshwater seaward, the stability of the Ba difference across shores is not surprising.

Although a stable natural tag can be derived from Ba values only, there were several secondary variables that were also important in separating fish. Yttrium (Y) was the second most important variable. With advances in technology, rare earth element
concentrations in otoliths are now routinely obtained and recent studies have found Y to have value as a discriminatory marker (Leakey et al. 2009, Mercier et al. 2011) However, compared to the other trace elements little is known about the factors regulating the concentration of Y in otoliths and the environmental or physiological features which make this an important marker for spotted seatrout in the Bay remains unclear.

The third most important marker was Mn, this element has been used in a variety of system to delineate fish (Thorrold et al. 1998a, Hanson et al. 2004, Clarke et al. 2009). The Mn concentration in otoliths has been shown to be independent of temperature and salinity (Martin and Thorrold 2005, Hamer and Jenkins 2007) but maintains a positive correlated with ambient concentration (Dorval et al. 2007). In the Chesapeake Bay, the Mn concentration in waters is non-conservative and is regulated by redox reactions (Eaton 1979). Reducing conditions have been shown to develop in the lower Eastern shore sediments and Mn is remobilized by reducing sediments (Hannigan et al. 2010). Therefore, this element can be useful in separating ES fish from those on the WS.

The stable isotope ratios of $\delta^{13}$C and $\delta^{18}$O were used in the year-specific AIC models, and not surprisingly these markers were also used in the multi-year AIC model to separate ES and WS habitats. Unlike $\delta^{13}$C, the $\delta^{18}$O otolith signature has been shown to reflect ambient conditions (See Campana 1999 for review). Although $\delta^{18}$O is strongly linked to temperature gradients in the environment, the accuracy of assigning spotted seatrout to natal habitats in the Bay has been attributed to the differences in water chemistry due to the many river drainages that influence the Western Shore nursery beds (Dorval et al. 2005b). Conversely, the stable isotope of carbon is regulated by a variety of factors such as temperature, water composition and diet (Campana 1999). Therefore, a
clear reason for the separation caused by $\delta^{13}\text{C}$ cannot be inferred.

The combination of stable isotope ratios and trace elements to delineate nursery habitats is becoming more prevalent in the use of otolith chemistry as a natural tag. The information conveyed by stable isotope ratios is often unique to that expressed by trace element values. These markers are regulated by environmental processes that can be specific to estuaries and within-estuary sites such as ambient concentration and temperature (Kalish 1991a, Høie et al. 2004). In many cases the introduction of stable isotopes to the classification model increases the ability to separate groups. For example, Thorrold et al. (1998b) found that classification of juvenile weakfish (*Cynoscion regalis*), a congener to spotted seatrout, increased with the addition of $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ isotope values. In other instances the stable isotope ratios can be used without trace elements to accurately separate fish (Gao et al. 2001, Rooker et al. 2008b, Pruell et al. 2011).

**Multi-year tag**

The application of otolith chemical tags have been limited by the temporal variability often associated with changes in environmental processes through time. However in this study, I demonstrate that even in very dynamic systems such as the Chesapeake Bay, a multi-year tag can be developed that can be retrospectively applied to historical datasets on a short term basis. This high degree of correct classification stemmed from the consistent spatial variation in otolith chemistry between juveniles residing on seagrass beds on the Eastern and Western Shores of the Bay. Otolith elements that contributed to this difference were often the same, indicating that the unique variation in biogeochemical signatures and the physical parameters structuring these
habitats can be relatively stable.

It is frequently noted in the literature that multi-year tags cannot be developed for many systems, as there is considerable interannual variation in elemental concentrations among sites (Gillanders 2002, Swearer et al. 2003, Gillanders 2005b, Elsdon et al. 2008). Hamer et al. (2003) found significant variation in otolith chemistry between adjacent year classes and suggested that chemical tags developed for one year class could not be subsequently applied to juveniles from other years classes. Similarly, Gillanders and Kingsford (2000) found that chemical tags could not be applied to adjacent years. However, there are instances when multi-year tags can be developed for a system. Hanson et al. (2004) found that temporal differences in fish otolith compositions were not significant when fish collections ranged over large spatial scales. Similarly, Brown (2006a) was able to classify fish to coastal or estuarine habitats with almost 80% accuracy by pooling data over several years to develop a classification model. The results of this study corroborates the findings by Brown (2006a), but on a much finer spatial scale, as fish were also classified within an estuary with a high probability of success using a tag derived from multi-year data.

Even though the annual signatures had different variable combinations, when data were pooled and analyzed across years a strong tag was discerned. This tag was driven by a selection of variables that showed consistent relative behaviors in each individual year. In this study, Ba was a significant driver for the differences observed between Eastern and Western shore seagrass habitats. Although the absolute concentrations in otolith concentrations of Ba changed between the habitats over the years, the relative difference remained constant. Barium was always higher in the ES fish otoliths than those in the WS
fish. Therefore, a main requirement for using an element is that the relative difference remains consistent through time. The stability and strong differences between ES and WS makes Ba a very powerful discriminatory variable and using Ba alone can provide good baseline estimates of contributions for each area.

In applying a multi-year tag, it should be noted that the discriminatory model must be valid for a given time period. In this study the multi-year tag was applied to historical collections that were relatively recent. However, caution should be exercised in using this tag to discriminate nursery areas for fish from much older cohorts, as its long-term stability is unknown. When the multi-year tag was applied to the year-specific chemical signature, classification success was less than that obtained from the individual year tag. Not unexpectedly, there is a loss of information as a result of pooling year classes. However, a multi-year tag can be used to provide a baseline estimate of historical production over time when only adult cores are available and the system remains stable.

The results of this study demonstrate the importance of model selection in discerning multi-year, multi-element tags for a habitat. The development of overall markers for a system can greatly influence the way in which fisheries are managed. As historical fish production can be identified, critical nursery areas can be conserved in order to ensure population persistence.
CHAPTER V

CONCLUSION

Classification of fish to natal sources is a central issue in understanding how nursery habitats influence fisheries population dynamics. Recruitment of fish to the adult stock depends heavily on nursery habitat quality, as anthropogenic stressors intensify on many nursery systems such as seagrass habitats, having accurate methods to identify sources of recruits is central to effective habitat-based fisheries management. In this dissertation I investigated statistical approaches to handle otolith chemistry data, a commonly used natural tag for delineating origin. Using variable and model selection techniques for spotted seatrout (*Cynoscion nebulosus*) on their nursery seagrass habitats, I developed classification models from otolith elemental compositions that minimize the probability of misclassification for subsequent survivors. Additionally, I examined the temporal stability of otolith variables and by selecting those variables that best characterize the system. By applying these techniques to a time series of otolith chemical data I was able to construct a discriminatory model that shows considerable promise as a short-term multi-year habitat tag for spotted seatrout in the Chesapeake Bay.

The chemical composition of the otolith has been shown to be a valuable tool in elucidating the natal sources of fish as it reflects the environmental contrast among habitats in its chemistry. However, the information conveyed by each chemical variable can be divided into three distinct groups: informative, redundant, or uninformative. By using Rao’s test of additional information for data that exhibit equality in variance-
covariance matrices, I found that many of the commonly used chemical markers were not informative or conveyed similar information in discriminating fish from multiple sources. In the lower Chesapeake Bay system only two variables, Ba and δ¹³C, were necessary in constructing a parsimonious model from the 2007 dataset that contained all three metrics (trace elements, stable isotopes, and growth). Although Rao’s test is well known in the statistical literature, this is the first application in an otolith chemistry study.

This study provides the first application of model selection techniques for choosing a best DFA classification model. I examined AIC and BIC approaches for data that did not meet the assumption of equality among variance-covariance matrices, thus a situation where Rao’s test cannot be used. Otolith chemistry data often fail to meet this assumption; therefore the methods used here can be applied to a variety of applications using otolith chemical data. To determine whether an AIC or BIC selection performed well with fish that subsequently survived from each natal area, I simulated “unknown” fish from each source to represent adult fish drawn from the juvenile distribution. It was necessary to use simulated data for this study as the source location of observed adults would have been unknown and the probability of misclassification could not be correctly evaluated.

I found that the AIC model selection was the most powerful model in identifying the source of recruits. The commonly used minimum error model did not result in the lowest prediction error of unknown fish from the same cohort. It is widely hypothesized that the accuracy of assignment of adult fish to natal sources is a reflection of the classification success from the juvenile DFA. However, I demonstrate that the minimum error model from the training data is not reflected in the accuracy of assignment in the
test data as models can be over-parameterized. These results emphasize the importance of choosing the best number of informative variables to construct a classification model and it is the first study to directly examine the issue of balancing bias and variance in building a habitat marker.

Model selection techniques to identify a multi-year model can significantly improve our ability to discern variables that remain stable over time. In this study a multi-year habitat model created from fish collected in 2006 – 08 correctly identified the natal source of fish collected in 2001 and 2002 with 90.6% and 79.5% overall classification success respectively. Additionally, having a multi-year marker reduces the number of juvenile collections that are needed to obtain the chemical fingerprint of each nursery area in the system making this methodology a cost effective and practical approach to determining critically nursery areas.

In all experiments presented in this study, the most informative variable for correctly classifying fish to natal sources was Ba. Because of the unique circulation patterns observed in the Bay, the Ba concentrations found in otoliths from fish residing on Eastern shore seagrass nurseries were higher than those residing on Western shore nurseries. This difference in Ba was stable across time and indicates that even in very dynamic systems the difference in concentrations of specific variables across nursery areas can be consistent across time.

The statistical approaches presented in this study offer significant improvement to some of the more commonly used methods for constructing DFA classification models. Rao’s test is a direct measure of variable inclusion for LDFA as this method is a parametric approach to identifying variable importance when groups have a shared
variance-covariance matrix. The AIC and BIC selection criteria presented, although only used in the statistical literature, provides a baseline approach to choosing QDFA models from data that do not meet assumptions of equality among variance-covariance matrices.

In conclusion, as coastal and near shore nursery areas become altered through anthropogenic forcing, following the provenance of adults back to their nursery habitat becomes increasing important. By using variable and model selection techniques to determine the best DFA models to separate groups, important information on the system drivers can be better elucidated. As long term habitat tags can be established through this methodology, vital information such as adult production from each nursery area and habitat-specific survivorship can also be determined.
LITERATURE CITED


Chen, Z., and C. M. Jones. 2006. Simultaneous determination of 33 major, minor, and trace elements in juvenile and larval fish otoliths by high resolution double focusing sector field inductively coupled plasma mass spectrometry. Pages 1-18 in Winter Conference on Plasma Spectrochemistry, Tucson, Arizona, USA.


trout (*Oncorhynchus mykiss*): distinctiveness and utility for detecting origins and movement. Canadian Journal of Fisheries and Aquatic Sciences 66:513-524.


Goshorn, D. M. 2006. Large-scale restoration of eelgrass (*Zostera marina*) in the Patuxent River, Maryland. Maryland Dept. of Natural Resources, Maryland, USA.


Hanson, P. J., C. C. Koenig, and V. S. Zdanowicz. 2004. Elemental composition of otoliths used to trace estuarine habitats of juvenile gag Mycteroperca microlepis along the west coast of Florida. Marine Ecology Progress Series 267:253-265.


Jones, J. B., and S. E. Campana. 2009. Stable oxygen isotope reconstruction of ambient temperature during the collapse of a cod (Gadus morhua) fishery. Ecological Applications 19:1500-1514.


VITA

Stacy Kavita Beharry was born on June 18, 1982 in Trinidad and Tobago. She was awarded a full academic scholarship to Morgan State University of Baltimore, Maryland to pursue a Bachelors of Science Degree in Biology in 2001. She obtained her Biology degree alongside a concentration in Mathematics at Morgan State University and graduated with the highest grade point average of the graduating class in May of 2005. Stacy enrolled in Old Dominion University’s Doctor of Philosophy in Oceanography degree program (4600 Elkhorn Ave., Norfolk, VA 23529) in August 2005. Her major professor was Dr. Cynthia M. Jones at the Center for Quantitative Fisheries Ecology (800 West 46th St., Norfolk, VA 23508) Prior to completing her Doctoral degree in December 2011, Stacy was awarded a Knauss Marine Science Policy Fellowship to work as a policy fellow in the executive branch of government for one year. She will complete her fellowship at the National Science Foundation headquarters, Directorate for Geosciences in Arlington, VA.