To the Graduate Council:

I am submitting herewith a dissertation written by James J. Minesky entitled **"Development and Application of a Genetic Algorithm-Informational Modeling Approach to Exploratory Statistical Modeling of Lizard-Habitat Relationships"**. I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Ecology and Evolutionary Biology.

Arthur C. Echternacht, Major Professor

We have read this dissertation
and recommend its acceptance:

Accepted for the Council:

Associate Vice Chancellor
and Dean of The Graduate School

DEVELOPMENT AND APPLICATION OF A GENETIC ALGORITHM-INFORMATIONAL MODELING APPROACH TO EXPLORATORY STATISTICAL MODELING OF LIZARD-HABITAT RELATIONSHIPS

A Dissertation
Presented for the
Doctor of Philosophy Degree
The University of Tennessee, Knoxville

James J. Minesky
May 1999

Copyright © 1999 by James John Minesky

All rights reserved

DEDICATION

This dissertation is dedicated to all of the people who have played important roles, both in and out of the formal classroom, in helping me to develop and shape my thinking, professional skills, and values: my parents and grandparents, wife, siblings, extended family, friends, teachers and professors, scientists, authors, philosophers, and sages.

ACKNOWLEDGMENTS

First and foremost, I extend my utmost thanks and appreciation to my advisor, Sandy Echternacht, for the endless amount of time and energy that he devoted to my graduate experience. I am deeply indebted to him for his advice, support, and unending determination to see me finish my degree.

I thank my committee members, Susan Riechert, Hamparsum Bozdogan, and John Gittleman, for their time, efforts, and advice while serving on my committee. I have benefitted from their knowledge and many combined years of experience, as well as from their dedication to and compassion for scientific research.

A tremendous amount of thanks goes to two individuals who provided assistance with the field work reported in this dissertation: M. Paula Goetting-Minesky (for helping to establish and mark the study plots) and Dan MacDonald (for helping to establish study plots, survey plots for anoles, and record habitat data). Without their assistance the field work would have been much too difficult. Dan also provided invaluable companionship in the field by way of his unique perspective of the world, love of reptiles and the outdoors, and crazy stories. I cannot imagine what my field work would have been like without Dan's assistance and friendship.

Financial support from the Theodore Roosevelt Fund of the American Museum of Natural History (New York, NY) is greatly appreciated for funding, in part, the field work on green anole-habitat relationships conducted for Part 5 of this dissertation. In addition, special thanks goes to the Carlos C. Campbell Memorial Research Fellowship of the Great Smoky

Mountains Conservation Association (TN) for funding field work on green anoles that provided valuable insight into helping design and conduct the research on green anoles reported in Part 5. I also thank the Departments of Zoology (now extinct) and Ecology and Evolutionary Biology for teaching assistantships and the Graduate School for a Hilton Smith Graduate Fellowship awarded for two years at UTK.

I am grateful to the Tennessee Valley Authority for access to the study sites along the Little Tennessee River and Judith Bartlow at the TVA for her assistance. Kind appreciation is extended to Jerry Hughes, who lives near one of the green anole study sites, for his assistance when we encountered any minor difficulties at that site (such as getting stuck in the mud).

A great deal of thanks goes to David Sylwester for teaching statistics so well, for clearly pointing out early on that I was not really conducting statistical-hypothesis testing, but statistical modeling, and also suggesting that I take a course on informational statistics to be taught by Ham Bozdogan. I took David's excellent advice and became excited about applying informational statistics to my research. Much appreciation and thanks goes to Ham Bozdogan for teaching me about informational statistics and providing a keen perspective on many statistical issues.

Hang-Kwang (Hans) Luh deserves special thanks for writing the genetic algorithm (GA) code. He was always gracious enough to collaborate with me and to help with many statistical and computational difficulties. Without his knowledge and help the GA approach would not have been used in this dissertation.

I thank the following individuals for their valuable assistance with other field studies on green anoles and informative discussions about lizard biology and ecology which provided useful information for this dissertation: Dan MacDonald, Sandy Echternacht, M. Paula Goetting-Minesky, Julie McNamara, Chris Samblanet, Glen Gerber, Ed Michaud, Luke Hasty, and DeeDee Truett.

The UTK Computing and Administrative Systems (CAS) deserves thanks for providing computer access to and computer time on the VAX (now extinct at UTK) and UNIX computers. It is unfortunate that CAS had to undergo a large staff cut a few years ago. Thanks to all the people who work, and those who once worked, at CAS for their efforts at making the computing systems a tool for students, teachers, and researchers at UTK.

To my parents, Mary and Ron Minesky, I am forever grateful for teaching me the importance of a good education, providing many educational opportunities, and believing in me. My journey so far would have been much more difficult had it not been for their support and love. Tremendous gratitude also goes to my brother Dave and sister Monica (and their significant others), Smith and Judy Goetting (my in-laws), and my wife's siblings for their support.

I greatly thank two long-time Pennsylvania friends, George and Michelle Solomon (who have been in Tennessee for many years now), for their friendship and assistance during my years in Knoxville. Special thanks to Linda Phillips and Ken McFarland for their friendship, encouragement, and wonderful hospitality over the years.

My gratitude also goes to Stan MacNevin, Bob Scott, John Ardis, Tom Tavella, Bill Brimeley, Terry Ryan, Ernie DeCiccio, and others while in Knoxville for their words of wisdom and guidance. The Theatre groups at UTK are to be graciously thanked for many excellent entertaining and cultural productions, which, little do they know, provided wonderful distractions from the grind of graduate work.

Finally, my undying gratitude and deepest appreciation goes to my wife, Paula, for her love and support over the many years. I have learned many lessons from her about myself and about others. Most importantly, I thank her for being true to herself and for helping and supporting me to be true to myself.

ABSTRACT

Anolis carolinensis, an arboreal lizard common to the southeastern United States, has been studied often in lab settings, but infrequently in its natural habitats with respect to the ecology of this species. The current study conducted exploratory statistical modeling of associations between 18 habitat features and the occurrence of *A. carolinensis* in study plots at the northern distributional limits of this species in eastern Tennessee.

Statistical hypothesis-testing procedures and stepwise computer algorithms are commonly used by ecologists to analyze observational (non-experimental) multivariate data, such as the data analyzed in this study. However, such procedures and algorithms are frequently, but inappropriately, used to find the single supposedly "best" statistical model and/or support interpretations of the "importance" or causal nature of variables in the model. Thus, such analyses provide only a narrow scientific view of the multivariate data and the many potentially useful models.

The present study developed a genetic algorithm-informational modeling (GAIM) approach to a) reduce certain computational and statistical limitations imposed by stepwise algorithms and hypothesis-testing procedures, respectively, and b) conduct a wider exploration of any observational multivariate data set. The GAIM approach utilizes a genetic algorithm, which bases its searching power on biological and evolutionary concepts, and the informational approach to statistics, which bases its ability to rank and evaluate models on statistical likelihood and information theory. It is suggested that researchers use an approach that

provides a wider view of the data (e.g., finds many models that fit the data well instead of just one or a few models), such as the GAIM approach, to more fully explore observational multivariate data. The *set* of well-fitting models obtained from a GAIM analysis can then be used to propose combinations of variables or factors that could be investigated by experiments in order to test causal hypotheses and/or produce predictive models.

One hundred sixty-six plots were placed in four different habitats along the Little Tennessee River where *A. carolinensis* occurs. Plots were surveyed for the presence/absence of this species in summer and winter seasons and habitat variables, both in and adjacent to the plots, were measured. Logistic regression modeling using the GAIM approach was conducted separately on the summer and winter data sets. For the summer data, the most frequent variables in the final set of GA models were (including the intercept): distance to potential overwintering rock, summer canopy categorization, distance to habitat edge, herb/shrub/vine cover, summer sunlight index, ambient temperature, and standardized distance along the habitat edge from the west boundary of habitat.

For the winter data, the most frequent variables in the final set of models were (including the intercept): ambient temperature, presence of live overstory evergreen tree trunks, presence of overwintering rock, standardized distance along the habitat edge from the west boundary of habitat, distance to potential overwintering rock, and canopy cover categorization. In each data set, the variables which most frequently

occurred in the final model set were also the ones which most frequently possessed statistically significant parameter estimates.

The summer models suggest that further research on *A. carolinensis* might focus on a) sunlight and thermal factors and b) habitat features related to certain spatial scales beyond the summer home range scale. Future research might also examine responses of this species to winter habitat features such as a) shelter and potential basking sites, b) sunlight availability and temperature, and c) spatial features beyond the typical winter home range size. Methods using experimental control, or at least partial control, over field variables are needed to determine the specific responses of this species to key habitat features and the causal mechanisms underlying those responses. In addition, more studies are needed which take approaches based on biophysical and physiological ecology, especially if they can be linked to reproductive output, population ecology, and habitat use on local and regional scales.

TABLE OF CONTENTS

PAGE

PART 1 : INTRODUCTION

OVERVIEW 2

RATIONALE AND OBJECTIVES 5

LITERATURE CITED 13

APPENDIX TO PART 1 19

HABITAT SCALES 20

PART 2 : BACKGROUND ON *ANOLIS CAROLINENSIS*

BACKGROUND ON POLYCHROTIDAE AND *ANOLIS* . . . 29
Family: Polychrotidae 29
Background on Anolis 29

ANCESTRY, DISTRIBUTION AND HABITATS
OF *ANOLIS CAROLINENSIS* 31
Possible ancestry and colonization history 31
Distribution, climates, and physiographic provinces . . . 33
Vegetation types, ecoregions, and habitats 35

NATURAL HISTORY OF *ANOLIS CAROLINENSIS* 40
Territoriality, movements, and home range size 40
Reproduction 44
Growth, maturity, and longevity 46
Daily activity 49
Winter activity 50
Winter thermal physiology and metabolism 54

LITERATURE CITED 57

APPENDIX TO PART 2 66

PAGE

PART 3 : THE INFORMATIONAL APPROACH TO DATA ANALYSIS : AN ALTERNATIVE TO STATISTICAL HYPOTHESIS-TESTING PROCEDURES

INTRODUCTION 82
Statistics and biology 82
Why another statistical approach? 85

OVERVIEW OF BOTH THE CLASSICAL AND INFORMATIONAL APPROACHES TO STATISTICAL ANALYSIS 88
The classical approach 88
The informational approach 91

COMPARISONS OF THE CLASSICAL AND INFORMATIONAL APPROACHES TO STATISTICAL ANALYSIS 99

SOME PERTINENT LITERATURE ON THE FUNDAMENTALS AND APPLICATIONS OF THE INFORMATIONAL APPROACH 112
General and technical literature 112
Applications of the informational approach in biology . . 113

COMMENTS ON THE USE OF THE INFORMATIONAL APPROACH 115

CONCLUDING REMARKS 121

LITERATURE CITED 126

APPENDIX TO PART 3 135

PART 4 : THE GENETIC ALGORITHM WITH AN INFORMATIONAL CRITERION : AN ALTERNATIVE METHOD FOR STATISTICAL MODELING OF OBSERVATIONAL DATA

INTRODUCTION 142

PAGE

THE LOGISTIC REGRESSION MODEL 146
Overview 146
The multiple logistic regression model 148

THE PROBLEM OF MODEL SELECTION 155
The classical approach vs. the informational approach . . 155
The objective of observational studies 158
Searching a vast model space and the limitations of current procedures 161

THE GENETIC ALGORITHM AND ITS APPLICATION TO STATISTICAL MODELING OF OBSERVATIONAL DATA . . 168
Overview of a simple genetic algorithm 168
A GA-informational modeling approach for logistic regression 175
Practical matters 192

SUMMARY AND CONCLUDING REMARKS 195

LITERATURE CITED 210

APPENDIX TO PART 4 218

PART 5 : ASSOCIATIONS BETWEEN HABITAT FEATURES AND THE PRESENCE OF *ANOLIS CAROLINENSIS* AMONG FOUR HABITATS IN EASTERN TENNESSEE: AN ANALYSIS USING THE GAIM APPROACH

INTRODUCTION 221

STUDY SITES 226

METHODS 230
Habitat scales 230
Plots 232
Surveying of plots 233
Habitat variables and their measurement 236
Data analysis 238

PAGE

RESULTS 249
Summer models 249
Winter models 256

DISCUSSION 265
Statistical approach used in this study 265
Summer models 272
Winter models 283
Non-habitat factors 292
Limitations of this study 294
The role of observational studies of animal-habitat relationships in forming conservation and management plans 296
Final comments 301

LITERATURE CITED 302

APPENDIX TO PART 5 312

PART 6 : SUMMARY

SUMMARY OF PREVIOUS PARTS 371

VITA 378

LIST OF TABLES

TABLE PAGE

PART 1: INTRODUCTION

1-1 Definitions used in this dissertation for the different areas or scales of habitat used by animals 27

PART 2: BACKGROUND ON *ANOLIS CAROLINENSIS*

2-1 The occurrence of *Anolis carolinensis* in the physiographic provinces of the 11 states in which this species is presently found 67

2-2 The occurrence of *Anolis carolinensis* in the potential natural vegetation types defined by Küchler (1964) in the 11 states over this lizard's range 69

2-3 Ecoregion classification of Bailey (1976, 1980) and the occurrence of *Anolis carolinensis* 74

2-4 Specific habitats in which *Anolis carolinensis* has been reported to occur 77

PART 3 : THE INFORMATIONAL APPROACH TO DATA ANALYSIS : AN ALTERNATIVE TO STATISTICAL HYPOTHESIS-TESTING PROCEDURES

3-1 Selected examples of publications which have expressed concerns about statistical hypothesis-testing procedures and the use and/or abuse of such procedures in scientific research 136

3-2 Various statistical techniques commonly used by biologists and references which provide some explanation of how to use the informational approach in conjunction with these techniques 139

TABLE PAGE

PART 4 : THE GENETIC ALGORITHM WITH AN INFORMATIONAL CRITERION : AN ALTERNATIVE METHOD FOR STATISTICAL MODELING OF OBSERVATIONAL DATA

4-1 Example of a data matrix with sample data for three continuous independent variables (X_1, X_2, and X_3), one categorical independent variable with two design variables (X_4), and a column of ones representing the intercept term (X_0) 219

PART 5 : ASSOCIATIONS BETWEEN HABITAT FEATURES AND THE PRESENCE OF *ANOLIS CAROLINENSIS* AMONG FOUR HABITATS IN EASTERN TENNESSEE: AN ANALYSIS USING THE GAIM APPROACH

5-1 The original names, forms, and descriptions of the measurement of the original variables for the study of the relationship between the presence of *Anolis carolinensis* and habitat features in four habitats along the Little Tennessee River in Tennessee 313

5-2 The names, final forms, and descriptions of the final form of the habitat variables used in the summer data analysis for the study of the relationship between the presence of *Anolis carolinensis* and habitat features in four habitats along the Little Tennessee River in Tennessee 317

5-3 The names, final forms, and descriptions of the final form of the variables used in the winter data analysis for the study of the relationship between the presence of *Anolis carolinensis* and habitat features in four habitats along the Little Tennessee River in Tennessee 319

5-4 Univariate logistic regression summary information for the summer habitat variables 321

TABLE PAGE

5-5 Summer models: the best logistic regression models with 15 or fewer parameters from the genetic algorithm (GA) output for modeling the relationship between habitat variables and the presence of *Anolis carolinensis* in summer study plots . 332

5-6 Summer models: the logistic regression parameter values for some of the top GA models for the *Anolis carolinensis* - habitat models 335

5-7 Univariate logistic regression summary information for the winter habitat variables 338

5-8 Winter models: the 20 logistic regression models having the lowest criterion values from the genetic algorithm (GA) output for modeling the relationship between habitat variables and the presence of *Anolis carolinensis* in winter plots 347

5-9 Winter models: the final best logistic regression models with 15 or fewer parameters (k) and model variance < 3.00 from both the genetic algorithm (GA) output and subsequent subset analyses for modeling the relationship between habitat variables and the presence of *Anolis carolinensis* in winter study plots 355

5-10 The 2x2 contingency table for the summer survey data used to examine the possible association between *Anolis carolinensis* and other lizard species 369

LIST OF FIGURES

FIGURE PAGE

PART 5 : ASSOCIATIONS BETWEEN HABITAT FEATURES AND THE PRESENCE OF *ANOLIS CAROLINENSIS* AMONG FOUR HABITATS IN EASTERN TENNESSEE: AN ANALYSIS USING THE GAIM APPROACH

5-1 Summer GA models: box plots showing trends in the (a) lack-of-fit term (-2LogL), (b) complexity term, and (c) model selection criterion, ICOMP-IFIM, across the different model sizes, represented by *k* (number of estimated regression parameters), for the best 115 summer logistic regression models found by the genetic algorithm (GA) analysis . . . 324

5-2 Summer GA models: the frequency of independent variables in the best 115 logistic regression models from the genetic algorithm (GA) output modeling the relationship between habitat features and the presence of *Anolis carolinensis* in summer plots 328

5-3 Summer GA models: trends in the frequency of independent variables in the different model sizes (*k* levels) in the best 115 logistic regression models from the genetic algorithm (GA) output modeling the relationship between habitat features and the presence of *Anolis carolinensis* in summer plots 330

5-4 Winter GA models: the frequency of independent variables in the best 184 logistic regression models from the genetic algorithm (GA) output modeling the relationship between habitat features and the presence of *Anolis carolinensis* in winter plots 341

5-5 Winter GA models: box plots showing trends in the (a) lack-of-fit term (-2LogL), (b) complexity term, and (c) model selection criterion, ICOMP-IFIM, across the different model sizes, represented by *k* (number of estimated regression parameters), for the best 184 winter logistic regression models found by the genetic algorithm (GA) analysis 343

FIGURE PAGE

5-6 Final best winter models: box plots showing lack of any clear trend in the model selection criterion, ICOMP-IFIM, across the different model sizes, represented by k (number of estimated regression parameters), for the final best winter logistic regression models (n = 154) found by the genetic algorithm (GA) analysis and subsequent subset analysis . . 349

5-7 Final best winter models: the frequency of independent variables in the final best logistic regression models (n = 154) from the combined results of the genetic algorithm (GA) output and subsequent subset analysis modeling the relationship between habitat features and the presence of *Anolis carolinensis* in winter plots 351

5-8 Final best winter models: trends in the frequency of independent variables in the different model sizes (k levels) in the final best logistic regression models (n = 154) from the combined results of the genetic algorithm (GA) output and subsequent subset analysis modeling the relationship between habitat features and the presence of *Anolis carolinensis* in winter plots 353

5-9 Winter Model 1 (k = 15, ICOMP-IFIM = 143.36, model variance = 1.91): graphical presentation of logistic regression diagnostic measures 359

5-10 Winter Model 144 (k = 10, ICOMP-IFIM = 147.12, model variance = 1.38): graphical presentation of logistic regression diagnostic measures 364

PART 1 : INTRODUCTION

OVERVIEW

Considerable theoretical and empirical research has been conducted in an effort to understand the complex relationships between an organism and/or a species and its habitat. Habitat can potentially influence heat balance and physiology (Gates 1980, Porter 1989), growth (Porter 1989), reproduction and life history traits (Stearns 1976), individual fitness (Fretwell 1972), and abundance and distribution of populations and species (Hutchinson 1957, MacArthur 1972). An understanding of the interactions between an organism and its habitat is important for gaining insight into individual behavior, physiological performance of individuals, life history traits, population dynamics and the viability of populations, community structure and organization, and evolution.

Habitat is an important concept in ecology, but its definition, like that of niche, has varied among ecologists over time (see Udvardy 1959, Davis 1960, Whittaker et al. 1973, Kulesza 1975). In this dissertation, habitat is defined as the area or place that contains the physical, chemical, and biotic resources required by individuals or populations of a given species (see Davis 1960), or even a species itself. Such resources can include water, humidity, sunlight, heat, shade, nesting or egg-laying sites, food, structural vegetation, and refugia from both predators and potentially threatening weather conditions. The importance of habitat is that it contains or consists of the crucial resources which are required by organisms for survival, growth, and reproduction and which promote the continued existence of populations and species.

Concerns over the future of many animal populations and species have been expressed by both scientists and the public with increasing frequency during the 20th century as more humans and human developments (e.g., housing, roads, railways, and industrial developments) appeared to negatively affect many animal populations and their habitats. These concerns helped contribute to federal laws being passed in the United States which required that wildlife and their habitats, as well as other natural resources, be given consideration whenever human activities were planned and conducted on public lands (Morrison et al. 1992:7-9). Some of these laws included the National Environmental Policy Act (NEPA) of 1969, the Endangered Species Act of 1973, the Federal Land Policy and Management Act of 1976, and the National Forest Management Act (NFMA) of 1976 (see Morrison et al. 1992:8-9).

In order for wildlife to be given consideration in the process of planning human activities on public lands, knowledge of the associations and interactions between wildlife and their habitats is needed. Such knowledge existed for certain game species, but little insight was available about the habitat requirements of many animal species. This has led biologists to study many game and nongame animals in order to formulate models of relationships between animals and their habitats and to potentially predict how changes made to habitats might affect animal populations (Morrison et al. 1992:9). A number of such quantitative models can be found, for example, in Verner et al. (1986).

Increased study of animals and their habitats may have been fostered by federal legislation and public concern, but the quantitative nature of these

studies was driven by the increasingly quantitative approach being used in ecology. The actual emergence of ecology as a science, around the beginning of the 20th century, is said to have started when biologists began applying mathematical and experimental methods to analyzing community structure and succession, population dynamics, and organism-environment relations (Kingsland 1991). Not all "ecological" studies in the early part of the 20th century were either quantitative or experimental, but such approaches became more frequent. The use of mathematical models and statistics in ecology increased during the second half of the 20th century.

Researchers studying animal habitats also began to adopt more quantitative methods. In particular, the ability to analyze multiple variables at the same time and to handle large data sets efficiently and accurately was realized with both the development of multivariate statistical techniques and the availability of high-speed digital computers to ecologists. Those two developments, combined with certain ecological developments, provided a synthesis in the 1970s that produced multivariate statistical analysis of habitat requirements of animals (Shugart 1981).

The application of multivariate statistics to ecological data analysis has been an important aspect in the development of ecology as a quantitative science, as well as in the development of habitat studies. However, misuses and misapplications of statistics have occurred in these developments and continue to occur today. The rationale for the research presented in this dissertation, as discussed in the next section, is based on a)

certain concerns over the misuses of multivariate statistics in ecology and in studies of animal-habitat relationships and b) the desire to examine possible associations between green anoles (*Anolis carolinensis*) and habitat features by using multivariate analyses.

RATIONALE AND OBJECTIVES

Concerns about misuses of multivariate techniques (inclusive of multiple regression) and misinterpretations of subsequent results have been discussed with respect to ecological data in general (e.g., see James and McCulloch 1985, 1990), and animal-habitat data in particular (e.g., see Johnson 1981a, b, Karr and Martin 1981). Three general statistical concerns addressed in this dissertation are:

1. The heavy reliance in ecology on statistical hypothesis-testing procedures (of the "frequentist" or "classical" approach) and the very limited use of other statistical approaches, particularly for purposes of statistical model selection.
2. The belief among many ecologists that a single "best" model exists for any multivariate data set and the subsequent use of stepwise procedures to find the supposedly "best" model.
3. The formulation of causal inferences, rather than correlative descriptions, based on multivariate analysis of observational (non-experimental) data.

Statistical hypothesis-testing procedures (henceforth referred to simply as hypothesis-testing procedures) include such tests of significance as a *t*-test, *F*-test, Chi-square test, likelihood ratio test (or *G*-test), Mann-Whitney U-test, Kruskal-Wallis test, and Wilcoxon's signed-ranks test. Not unlike many researchers in the natural and social sciences, ecologists tend to use hypothesis-testing procedures far more frequently than other approaches,

such as the informational approach. Evidence to support this statement is not difficult to find; one only has to scan the 'Methods' sections of papers in ecological journals to find that the vast majority of researchers exclusively use hypothesis-testing procedures in their statistical analyses. A particular point of debate is the overreliance of ecologists on hypothesis-testing procedures for selection of an appropriate statistical model (or models) when other, often more advantageous, methods exist to help balance problems of overfitting and underfitting. For example, most ecologists use test procedures as the basis for model building (adding and removing variables) and model selection in linear regression, even though Mallows' C_p criterion has been available for such analysis for over 20 years.

Each statistical approach or methodology can be viewed as a statistical "tool" that may be more useful in some situations than other approaches or tools. Although hypothesis-testing procedures may perform well for a variety of research designs, data sets, and questions, other statistical approaches may be of equal or greater utility especially under certain conditions or with certain data. The informational approach, based on the ground-breaking work of Akaike (1973, 1974), is a viable alternative to hypothesis-testing procedures which has certain statistical advantages over such test procedures.

The informational approach uses statistical likelihood as part of a numerical criterion to help analysts select statistical models which best fit the data and, in turn, isolate meaningful variables that can be investigated via additional studies for possible causal relations. Many ecologists who work on capture-recapture data (see, e.g. Szymczak and Rexstad 1991,

Burnham and Anderson 1992, Lebreton et al. 1992) are rapidly adopting the informational approach for the analysis and estimation of survival and recapture rates. The informational approach is the approach used for statistical modeling in this dissertation.

The second important statistical concern is that many ecologists believe that an observational multivariate data set (i.e., multivariate data collected without an experimental design in which variables can be controlled by the investigator) will have a single "best" statistical model which has a superior fit to the data over alternative models. However, with any multivariate data set it is fairly probable that no single model will be better than all other models (Gorman and Toman 1966, Hocking 1983, McCullagh and Nelder 1989:8); such data can often be described equally well, statistically and biologically, by several or more models. In practice, many ecologists overlook this point and commonly use stepwise procedures to find a single, supposedly "best" model when analyzing multivariate data. However, stepwise procedures have certain limitations (see Beale 1970, Mantel 1970, Hocking 1976, 1983, Moses 1986, Myers 1986, James and McCulloch 1990) which make it unlikely that a single best model could be found if many independent variables are used. Analysts often forget these points and operate under the notion that a stepwise procedure does find the single best statistical model to fit the data.

The third statistical problem of interest regards the interpretation of observational data. Cautionary notes have been sounded about the dangers of making predictions (see Hocking 1983, Snee 1983) and causal interpretations (see Johnson 1981b, James and McCulloch 1985, 1990) based

on observational data that involves many variables. Such data help produce models or hypotheses about possible causation, but only controlled experiments provide rigorous tests of causal hypotheses/models (James and McCulloch 1990, Lubchenco and Real 1991). However, examples of risky interpretations and inferences about causation based on observational data can be found in the research published on animal-habitat relationships in such journals as *Ecology, Conservation Biology,* and the *Journal of Wildlife Management,* as well as in meetings/workshop proceedings (see, e.g., various papers in Capen 1981, Verner et al. 1986).

Researchers often collect observational multivariate data on animal-habitat relationships, use a stepwise procedure to find the supposedly "best" model, and then make either strong inferences about the causation between dependent and independent variables or specific predictions of outcomes, or both. In addition, specific management or conservation recommendations are often based on observational data and subsequent analysis, though such recommendations would be better based on a designed experimental approach (see Marzluff 1986).

The objectives of this study are two-fold. The first is to outline the framework of a new methodology, the genetic algorithm-informational modeling (GAIM) approach, for the analysis of observational multivariate data. The GAIM approach is proposed as an alternative to stepwise procedures for modeling multivariate data and as an attempt to address the three statistical concerns outlined previously in this section. The second objective is to apply the GAIM approach to observational *A. carolinensis*-habitat data in order to a) demonstrate how to use the GAIM approach and

b) obtain insight into possible associations between the presence of *A. carolinensis* and various habitat features. Such insight, along with what is already known about this species, can help formulate hypotheses about *A. carolinensis*-habitat relationships which could be tested by future research.

Anolis carolinensis is a small, mainly arboreal lizard found in the southern United States. Over the past 25 years, more than 1,000 papers have been published on *A. carolinensis*, the vast majority of which have been laboratory-based studies. The limited aspects of the ecology of this species which have been studied have been done so on only a few populations. Part 2 of this dissertation summarizes the rather limited number of studies which provide insight into the natural history of *A. carolinensis*. Because this species can potentially populate a diversity of habitats, Part 2 also provides a survey of the physiographic provinces, potential vegetation types, ecoregions, and descriptions of habitats in which *A. carolinensis* has been reported to occur.

The GAIM approach proposed in this dissertation utilizes the informational approach to statistics. Much of the valuable insight into the informational approach is scattered among many books, journal articles, and contributed volumes, with only a small number of articles being published in ecological journals. Thus, Part 3 provides an overview of the informational approach which includes many references to relevant literature published in the fields of statistics and ecology. Part 3 also discusses the heavy reliance on hypothesis-testing procedures by ecologists and some possible reasons for this reliance. The informational approach is

a viable alternative to hypothesis-testing procedures because it has some advantages over hypothesis-testing procedures, as described in Part 3.

An important problem in statistical modeling of multivariate data is how to find models that have a very good fit to an observational data set when a large number of independent variables exist and the researcher cannot easily evaluate all possible models. Part 4 of this dissertation addresses three aspects of this problem. First, although ecologists commonly use various stepwise procedures to find only one or a few models to fit a data set which has many independent variables, a solid case for abandoning the use of such procedures is made in Part 4.

Second, the concepts behind genetic algorithms (GAs), the way in which a simple GA works, and the use of a GA as an alternative searching algorithm to stepwise procedures for multiple regression are discussed in Part 4. A genetic algorithm (GA) is any computer algorithm which is based (although loosely) on the concepts of genetic recombination, natural selection, mutation, and biological evolution and can search for and find useful solutions to a problem when the problem has too many potential solutions to individually evaluate or rank (see Holland 1992a, b, Forrest 1993, Goldberg 1994). Luh et al. (submitted) proposed using a genetic algorithm in conjunction with an informational model-selection criterion for variable selection in multiple regression. Some researchers have apparently been using GAs for estimating statistical parameters including estimates of standard errors (see Chatterjee and Laudato 1995), but the use of a GA in conjunction with an informational criterion for the problem of statistical model selection was developed and used by H.-K. Luh, J. J.

Minesky, and H. Bozdogan in 1995 independently of the work by Chatterjee and Laudato (1995).

Third, Part 4 proposes a general methodology, the GAIM approach, as an alternative to the commonly used combination of stepwise algorithms and hypothesis-testing procedures for purposes of model selection when a large number of independent variables exist. Although Luh et al. (submitted) showed how to apply a GA to multiple regression modeling, they did not address the inadequacy of trying to find the single "best" model when using either stepwise procedures or GAs for modeling observational multivariate data. However, the GAIM approach proposed in Part 4 most importantly emphasizes the need to examine and report a *set* of very good models rather than one supposedly best model and provides the computer/GA and statistical methods to do so. The GAIM approach provides an analyst with a "wider" view of the models and data than could be obtained with stepwise procedures. Specific recommendations are provided in particular for multiple logistic regression modeling.

Application of the GAIM approach to the analysis of two different observational data sets on *A. carolinensis*-habitat associations is then presented in Part 5. The green anole, *Anolis carolinensis*, is an especially good candidate for many ecological studies because of its excellent colonizing abilities (Williams 1969), fairly wide distribution across the southern United States (Conant and Collins 1991), occurrence in a variety of habitats (see Part 2), small home range sizes (Gordon 1956, King 1966) compared to many other vertebrates, arboreal habits making it rather

visible, and sometimes high local abundance (Gordon 1956, personal observation).

In the present study, home range sized plots were surveyed for the presence of *A. carolinensis* in four different habitats along the Little Tennessee River during both the summer and winter activity seasons. A suite of habitat variables were also measured during those seasons. The GAIM approach was then used to find a *set* of statistical models that fit the data very well for the summer and winter seasons. The analyses can provide insight into *A. carolinensis*-habitat relationships to aid researchers in designing and conducting future observational and experimental studies on the habitat ecology of this species.

Habitat variables were measured at different spatial and temporal scales in the current study. The author's concepts and definitions of different habitat scales are provided in the Appendix (including Table 1-1) at the end of Part 1. Although only a few of the spatial scales were measured in this study, definitions of all the different scales provide readers with a more complete sense of the author's concepts and use of habitat scales for this research on *A. carolinensis*.

LITERATURE CITED

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267-281 *in* B. N. Petrov and F. Csáki, editors. Second international symposium on information theory. 1971. Akadémiai Kiadó, Budapest, Hungary. 451 pp.

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

Bailey, R. G. 1976. Ecoregions of the United States (map). U. S. Department of Agriculture, Forest Service Intermountain Region, Ogden, Utah, USA. Scale 1:7,500,000.

Bailey, R. G. 1983. Delineation of ecosystem regions. Environmental Management 7:365-373.

Bailey, R. G. 1995. Description of the ecoregions of the United States. 2nd ed. rev. and expanded. U. S. Department of Agriculture, Forest Service, Miscellaneous Publication No. 1391 (rev.), Washington D.C., USA. 108 pp. With separate map at 1:7,500,000.

Beale, E. M. L. 1970. A note on procedures for variable selection in multiple regression. Technometrics 12:909-914.

Burnham, K. P. and D. R. Anderson. 1992. Data-based selection of an appropriate biological model: the key to modern data analysis. Pages 16-30 *in* D. R. McCullough and R. H. Barrett, editors. Wildlife 2001: Populations. Elsevier Science Publishers, London, United Kingdom. 1163 pp.

Capen, D. E. (editor). 1981. The use of multivariate statistics in studies of wildlife habitat. USDA Forest Service General Technical Report RM-87, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado, USA. 249 pp.

Chatterjee, S. and M. Laudato. 1995. Genetic algorithms and their statistical applications. 1995 Proceedings of the Statistical Computing Section, American Statistical Association. American Statistical Association, Alexandria, VA, USA. 253 pp.

Conant, R. and J. T. Collins. 1991. A field guide to reptiles and amphibians. 3rd ed. Houghton Mifflin Company, Boston, Massachusetts, USA. 450 pp.

Davis, J. H. 1960. Proposals concerning the concept of habitat and a classification of types. Ecology 41:537-541.

Forrest, S. 1993. Genetic algorithms: principles of natural selection applied to computation. Science 261:872-878.

Fretwell, S. D. 1972. Populations in seasonal environment. Princeton University Press, Princeton, New Jersey, USA. 217 pp.

Gates, D. M. 1980. Biophysical ecology. Springer-Verlag, New York, New York, USA. 611 pp.

Goldberg, D. E. 1994. Genetic and evolutionary algorithms come of age. Communications of the ACM 37:113-119.

Gordon, R. E. 1956. The biology and demography of *Anolis carolinensis carolinensis* Voigt. Unpublished Ph. D. dissertation. Tulane University, New Orleans, Louisiana, USA. 263 pp.

Gorman, J. W. and R. J. Toman. 1966. Selection of variables for fitting equations to data. Technometrics 8:27-51.

Hocking, R. R. 1976. The analysis and selection of variables in linear regression. Biometrics 32:1-49.

Hocking, R. R. 1983. Developments in linear regression methodology: 1959-1982. Technometrics 25:219-230.

Holland, J. H. 1992a. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. 1st MIT Press edition. The MIT Press, Cambridge, Massachusetts, USA. 211 pp.

Holland, J. H. 1992b. Genetic algorithms. Scientific American July 1992: 66-72.

Hutchinson, G. E. 1957. Concluding remarks. Cold Spring Harbor Symposium on Quantitative Biology 22:415-427.

James, F. C. and C. E. McCulloch. 1985. Data analysis and the design of experiments in ornithology. Pages 1-63 *in* R. F. Johnston, editor. Current ornithology, vol. 2, Plenum Press, New York, New York, USA. 378 pp.

James, F. C. and C. E. McCulloch. 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? Annual Review of Ecology and Systematics 21:129-166.

Johnson, D. H. 1981a. The use and misuse of statistics in wildlife habitat studies. Pages 11-19 *in* D. E. Capen, editor. The use of multivariate statistics in studies of wildlife habitat. Rocky Mountain Forest and Range Experiment Station, U.S. Forest Service, General and Technical Report RM-87, U.S. Department of Agriculture, Fort Collins, Colorado, USA. 249 pp.

Johnson, D. H. 1981b. How to measure habitat - a statistical perspective. Pages 53-57 *in* D. E. Capen, editor. The use of multivariate statistics in studies of wildlife habitat. Rocky Mountain Forest and Range Experiment Station, U.S. Forest Service, General and Technical Report RM-87, U.S. Department of Agriculture, Fort Collins, Colorado, USA. 249 pp.

Karr, J. R. and T. E. Martin. 1981. Random numbers and principal components: further searches for the unicorn? Pages 20-24 *in* D. E. Capen, editor. The use of multivariate statistics in studies of wildlife habitat. Rocky Mountain Forest and Range Experiment Station, U.S. Forest Service, General and Technical Report RM-87, U.S. Department of Agriculture, Fort Collins, Colorado, USA. 249 pp.

King, F. W. 1966. Competition between two south Florida lizards of the genus *Anolis*. Unpublished Ph. D. dissertation. University of Miami, Coral Gables, Florida, USA. 97 pp.

Kingsland, S. 1991. Defining ecology as a science. Pages 1-13 *in* L. A. Real and J. H. Brown, editors. Foundations of ecology: classic papers with commentaries. The University of Chicago Press, Chicago, Illinois, USA. 905 pp.

Köppen, W. 1931. Grundriss der Klimakunde. Walter de Gruyter Company, Berlin, Germany. 388 pp.

Küchler, A. W. 1964. Potential natural vegetation of the conterminous United States. American Geographical Society Special Publication No. 36. American Geographical Society, New York, New York, USA. 116 pp. With separate map at 1:3,168,000.

Kulesza, G. 1975. Comment on "Niche, Habitat, and Ecotope". The American Naturalist 109:476-479.

Lebreton, J. D., K. P. Burnham, J. Clobert, and D. R. Anderson. 1992. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecological Monographs 62:67-118.

Lubchenco, J. and L. A. Real. 1991. Manipulative experiments as tests of ecological theory. Pages 715-7733 *in* L. A. Real and J. H. Brown, editors. Foundations of ecology: classic papers with commentaries. The University of Chicago Press, Chicago, Illinois, USA. 905 pp.

Luh, H.-K., J. J. Minesky, and H. Bozdogan. Submitted manuscript. Choosing the best predictors in regression analysis via the genetic algorithm with informational complexity as the fitness function.

MacArthur, R. H. 1972. Geographical ecology. Princeton University Press, Princeton, New Jersey, USA. 269 pp.

Mantel, N. 1970. Why stepdown procedures in variable selection. Technometrics 12:591-612.

Marzluff, J. 1986. Assumptions and design of regression experiments: the importance of lack-of-fit-testing. Pages 165-170 *in* J. Verner, M. L. Morrison, and C. J. Ralph, editors. Wildlife 2000: modeling habitat relationships of terrestrial vertebrates. The University of Wisconsin Press, Madison, Wisconsin, USA. 470 pp.

McCullagh, P. and A. J. Nelder. 1989. Generalized linear models. 2nd edition. Chapman and Hall, London, United Kingdom. 511 pp.

Morris, D. W. 1987. Ecological scale and habitat use. Ecology 68:362-369.

Morrison, M. L., B. G. Marcot, and R. W. Mannan. 1992. Wildlife-habitat relationships: concepts and applications. University of Wisconsin Press, Madison, Wisconsin, USA. 343 pp.

Moses, L. E. 1986. Think and explain with statistics. Addison-Wesley Publishing Company, Reading, Massachusetts, USA. 483 pp.

Myers, R. H. 1986. Classical and modern regression with applications. Duxbury Press, Boston, Massachusetts, USA. 359 pp.

Porter, W. P. 1989. New animal models and experiments for calculating growth potential at different elevations. Physiological Zoology 62:286-313.

Quarterman, E., M. P. Burbanck, and D. J. Shure. 1993. Rock outcrop communities: limestone, sandstone, and granite. Pages 35-86 *in* W. H. Martin, S. G. Boyce, and A. C. Echternacht, editors. Biodiversity of the southeastern United States: upland terrestrial communities. John Wiley and Sons, New York, New York, USA. 373 pp.

Shaffer, M. L. 1981. Minimum population sizes for species conservation. BioScience 31:131-134.

Shugart, H. H., Jr. 1981. An overview of multivariate methods and their application to studies of wildlife habitat. Pages 4-10 *in* D. E. Capen, editor. The use of multivariate statistics in studies of wildlife habitat. Rocky Mountain Forest and Range Experiment Station, U.S. Forest Service, General and Technical Report RM-87, U.S. Department of Agriculture, Fort Collins, Colorado, USA. 249 pp.

Skeen, J. N., P. D. Doerr, and D. H. Van Lear. 1993. Oak-hickory-pine forests. Pages 1-34 *in* W. H. Martin, S. G. Boyce, and A.C. Echternacht (editors), Biodiversity of the southeastern United States: upland terrestrial communities. John Wiley and Sons, New York, New York, USA. 373 pp.

Snee, R. D. 1983. Discussion. Technometrics 25:230-237.

Stearns, S. C. 1976. Life history tactics: a review of the ideas. Quarterly Review of Biology 51:3-47.

Stephenson, S. L., A. N. Ash, and D. F. Stauffer. 1993. Appalachian oak forests. Pages 255-304 *in* W. H. Martin, S. G. Boyce, and A.C. Echternacht (editors), Biodiversity of the southeastern United States: upland terrestrial communities. John Wiley and Sons, New York, New York, USA. 373 pp.

Szymczak, M. R. and E. A. Rexstad. 1991. Harvest distribution and survival of a gadwall population. Journal of Wildlife Management 55: 592-600.

Thiollay, J. M. 1989. Area requirements for the conservation of rainforest raptors and game birds in French Guiana. Conservation Biology 3:128-137.

Trewartha, G.T. 1968. An introduction to climate. 4th edition. McGraw-Hill, New York, New York, USA. 408 pp.

Udvardy, M. F. D. 1959. Notes on the ecological concepts of habitat, biotope, and niche. Ecology 40:725-728.

Verner, J., M. L. Morrison, and C. J. Ralph. (editors). 1986. Wildlife 2000: modeling habitat relationships of terrestrial vertebrates. University of Wisconsin Press, Madison, Wisconsin, USA. 470 pp.

Whittaker, R. H., S. A. Levin, and R. B. Root. 1973. Niche, habitat, and ecotope. The American Naturalist 107:321-338.

Williams, E. E. 1969. The ecology of colonization as seen in the zoogeography of anoline lizards on small islands. Quarterly Review of Biology 44(4): 345-399.

APPENDIX TO PART 1

HABITAT SCALES

Habitat has a multitude of scales or levels which, taken collectively, form a continuum across space or time. However, a complete set of definitions is difficult to find in any one published paper or book for the spectrum of habitat scales encountered by animals. This might be due to the fact that researchers either adopt or formally define only the particular habitat terms and scales that are most relevant to their specific topic or subdiscipline (such as animal ecology, plant ecology, biophysical ecology, conservation ecology, and landscape ecology). For example, habitat selection theory emerged, at least in part, from foraging theory and both areas often treat the term "microhabitat" as a foraging patch within the home range of an animal. Biophysical ecologists typically consider microhabitat to be a small spatial scale related to the organism's current interchange of heat, water, and gases with the immediate environment. In addition, some researchers have defined spatial scales of habitat, but only over the range of scales for which their research was conducted (see, e.g., Morris 1987).

Herein are provided some working definitions of a wide range of habitat spatial scales (see Table 1-1 for a summary of these definitions). General definitions of the smaller scales should be based on the basic biological activities of the organism, but the actual size of the habitat unit or the duration of the time frame will be determined by the particular species or population being studied and either the specific or collective activities being performed by the organisms. On the other hand, definitions of spatial scales beyond the area that is used during a typical

individual's lifetime do not need to be based primarily on the biological activities and resource acquisition of an individual. Instead, definitions of larger spatial scales should be based on 1) relevant structural, physical, and chemical factors in an area that includes a subpopulation, a population, or several populations, and 2) processes associated with expansion and contraction of populations and/or the range of the species. Larger scales should be defined in terms of such features as vegetational types, topography, aspect, soil types, and local or regional climatic features or categories, in conjunction with, when possible, any biological processes that occur on a scale equal to or greater than the area of a subpopulation or population and/or the duration of one generation.

Considering these concepts of habitat scales, "microhabitat", at the smaller end of the spatial continuum, is defined in this dissertation as the habitat used by an individual in a population, on average, in conjunction with a specific biological activity during a specific time or segment of a daily activity pattern. The activity can be performed for either directly obtaining one or more resources or conducting one or more functions not directly related to resource acquisition such as sleeping, molting, egg-laying, nesting, or displaying to or communicating with other individuals. Thus, a researcher could measure microhabitat variables related to an organism's site for either displaying to conspecifics, basking, nesting, sleeping, or so on. Note that any individual will likely encounter and use several or more different microhabitat patches over the course of a day because 1) several or more microhabitat patches might be used for acquiring a resource or conducting a certain activity, 2) each specific activity might require a

different microhabitat patch, and 3) each resource might be located in a different microhabitat.

At a spatial level above microhabitat, "seasonal-use" (SU) habitat is the habitat area used by a given individual in a population during a "season", where "season" is determined by both important biological activities and environmental factors. This definition considers the fact that many species or populations exhibit differences in the size and/or characteristics of a typical home range depending on the climatic season (e.g., summer vs. winter) and/or a particular biological activity "season" (e.g., reproduction vs. non-reproduction). For many territorial species, territories are defended with greater intensity at certain times of the year or life-cycle than others. In other species, territories are formed and defended during a particular season coinciding with specific activities. Thus, it is useful to define a habitat scale, such as the SU habitat, based on certain biological activities conducted during certain climatic and or biological seasons.

At a level above the SU habitat, "overall home range" (OHR) habitat can be defined as the habitat scale that approximately equals the area used by an individual in the population over either one complete cycle of defined alternating biological activity "seasons" or one complete cycle of climatic seasons (such as one year in non-tropical climates). The SU and OHR scales will be equal in those species or populations in which individuals do not seasonally change their spatial use of habitat.

For those species which change their use of habitat both across seasons and over an entire life-time, "life-time" (LT) habitat scale is the typical habitat area used over the life span of a given individual. This definition

is more easily used for non-migratory species, but could be applied to certain seasonal migratory species provided that biologically meaningful insight could be obtained through the use of the definition.

The term "macrohabitat" could possibly be used synonymously with OHR and/or LT habitat. If so, this use of macrohabitat is similar to the definition used by other biologists. For example, Morris (1987:363) defined macrohabitat as "... distinguishable units whose minimum area corresponds to that within which an average individual performs all of its biological functions (home range) during a typical activity cycle". He does not define the terms "minimum area" or "activity cycle" any further in a biological context. However, if 1) minimum area is taken as the area needed for resource acquisition sufficient for growth and reproduction and 2) activity cycle is taken as a cycle over the reproductive and non-reproductive seasons or over the active ("summer") and non-active ("winter") seasons, then OHR habitat and macrohabitat are similar.

Beyond the OHR and LT habitat scales, the "population level" (PL) habitat scale is that habitat area occupied by a particular population or subpopulation of a given species at a given time. Ecologists, conservationists, and resource managers are often interested in determining the size of the habitat needed for a minimum viable population (MVP). Shaffer (1981) first defined an MVP as "... the smallest isolated population having a 99% chance of remaining extant for 1000 years despite the foreseeable effects of demographic, environmental, and genetic stochasticity, and natural catastrophes.", but also suggested that other probabilities of survival or time frames could be used as well. The

definition of PL habitat does not address this point. In conservation biology, the "minimum dynamic area" (MDA) is defined as the size of habitat needed to maintain an MVP for a given species and estimates of MDA can be obtained from knowledge of home range sizes of individuals and groups (Thiollay 1989).

Although some animal ecologists do describe their study sites and/or study habitats in terms of Küchler's (1964) vegetation types or some accepted vegetational classification scheme, it appears that most animal ecologists do not provide such descriptions in their publications. Reporting such descriptions of study sites and habitats provides the readers with a greater sense of the habitat in which the species or community of species lives. Unless quantitative measurements of the habitat are reported, other researchers should be provided with at least a description of the study sites in terms of some accepted vegetational scheme in order to make mental comparisons of the reported study sites to their own sites.

For many terrestrial animals, descriptions of PL and MVP habitats could be given, at least partly, in terms of Küchler's (1964) potential natural vegetation types or other vegetational community classifications on a similar scale. For example, PL and/or MVP habitats for the Carolina chickadee (*Parus carolinensis*) could be described in terms of vegetational communities such as oak-hickory-pine forest (see Skeen et al. 1993), cedar glade (see Quarterman et al. 1993), and Appalachian oak forest (early successional stages, see Stephenson et al. 1993) instead of simply as "edge" or "forest". Part 2 of this dissertation provides a list of the potential natural

vegetation types (based on Küchler 1964) in which *A. carolinensis* is likely to inhabit throughout its range.

Beyond the PL and MDA habitat scales, ecologists often define or describe geographical areas in which either at least one population is found (such as a map of county occurrences) or all of the known viable populations of a species occur (the species "range"). Use of such geographical descriptions or maps often lack references to relevant climatic and vegetational information for terrestrial animals. However, the development of a classification scheme of ecoregions ("ecosystems of regional extent") by Bailey (1976, 1983, 1995) gives ecologists the opportunity to describe fairly large geographical areas occupied by populations or a species in terms of climatic, land form, and vegetational features.

Bailey's (1995) ecoregions scheme is a hierarchical classification of ecosystems as the map units. The largest unit is the domain, within which exist successively smaller ecosystem units, such as divisions and provinces. The map boundaries of domains and divisions are defined mainly on large scale, ecological climate zones based on the works of Köppen (1931) and Trewartha (1968). Large scale vegetational features are the basis for then subdividing each division into provinces (Bailey 1995). Further subdivision of provinces and even smaller units could be used by ecologists to describe the regional and local distribution of terrestrial and semiaquatic animal species at levels above that of MDA habitat. Part 2 of this dissertation lists the ecoregions (down to province) in which *A.*

carolinensis is known to occur as reconstructed by the author from many distributional reports for this species.

Table 1-1. Definitions used in this dissertation for the different areas or scales of habitat used by animals. All terms, except those provided with a cited reference, have been defined by the author.

Term	Definition
Microhabitat	The habitat area used by an individual in a population, on average, in conjunction with a specific biological activity (e.g., egg-laying, nesting, hibernating, communicating to conspecifics, or obtaining any type of resource) during a specific time or segment of a daily activity pattern.
Seasonal-Use (SU) Habitat	The habitat area used by a given individual in a population during a "season", where "season" is determined by both important biological activities conducted by organisms in the population and environmental factors in the habitat.
Overall Home Range (OHR) Habitat	The habitat area that approximately equals the area used by an individual in a population over one complete cycle of either defined alternating biological activity "seasons" or climatic seasons.
Life-Time (LT) Habitat	The typical habitat area used over the life span of an individual.
Population Level (PL) Habitat	The habitat area occupied by a particular population or subpopulation of a given species at a given time.
Minimum Dynamic Area (MDA)	The size of habitat needed to maintain a minimum viable population (MVP) for a given species (Thiollay 1989).
Range	The geographical area which encompasses all of the known viable populations of a species (based on usage of this term in the discipline of biogeography).

PART 2 : BACKGROUND ON *ANOLIS CAROLINENSIS*

BACKGROUND ON POLYCHROTIDAE AND *ANOLIS*

Family: Polychrotidae

The phylogenetics and taxonomy of the lizard family Iguanidae has been under scrutiny in recent years. Etheridge and de Queiroz (1988) showed that eight major suprageneric groups exist within the Iguanidae, but they could not ascertain clear relationships of the major groups to one another. Frost and Etheridge (1989) performed a phylogenetic analysis of the Iguania (Iguanidae, Agamidae, and Chamaeleonidae) and found evidence for partitioning Iguanidae into eight families. One of these families, Polychridae, includes the genus *Anolis* and several other related genera.

The Polychridae (= Polychrotidae) of Frost and Etheridge (1989) corresponds to the group which Etheridge and Williams (1985) and Etheridge and de Queiroz (1988) called the "anoloids" (anoles and their relatives). Members of Polychrotidae range from small to medium sized lizards and most species are arboreal. This family is neotropical in origin and its present distribution covers the southeastern United States, Mexico, Central America, the West Indies, and much of South America (Frost and Etheridge 1989).

Background on Anolis

The genus *Anolis* is a widespread and an ecologically diverse group within Polychrotidae. Although no exact total count has been made recently, it is believed that around 300 species of *Anolis* exist. Schwartz and Henderson (1991) list and describe 128 species found in the West Indies alone. Members of this genus occur in a wide variety of habitats including

grassy areas, mesic forests, rock outcrops, xeric woodlands, scrub, gardens, plantations, and urban areas (see Schwartz and Henderson 1991). Most *Anolis* are arboreal, but a few are saxicolous, semi-aquatic, or terrestrial (Etheridge and de Queiroz 1988). Anoles are primarily insectivorous, although some species may also eat earthworms, millipedes, spiders, ticks, snails, slugs, crustaceans, frogs, lizards, small birds, and/or fruit (see Schwartz and Henderson 1991).

Reproduction has not been studied in all species of *Anolis*, but for those in which it has females are known to usually produce one egg at a time with ovulation occurring alternately between the two ovaries (Fitch 1970, Smith et al. 1972, Fitch 1982). Fitch (1982) reviewed reproductive cycles in tropical reptiles and indicated that *Anolis* breeding seasons vary in length from a short season available each year in areas with distinct dry seasons to a year-round event in areas with sufficient temperature and rainfall. Also, the length and timing of the breeding season within a given species of *Anolis* appears to be related to climatic variation (Fitch 1982).

Many studies on the ecology, evolution, and behavior of various *Anolis* species have been conducted. However, since about 300 species occur in this genus it becomes apparent after examining much literature that a thorough knowledge of the biology of probably only a small number of species exists. For most species a rather limited and scattered understanding of the ecology, evolution, and behavior has been gained.

ANCESTRY, DISTRIBUTION, AND HABITATS OF *ANOLIS CAROLINENSIS*

Possible ancestry and colonization history

The green anole, *Anolis carolinensis,* is the only species within Polychrotidae which is native to the continental United States. Anoles are divided into two sections, the alpha and beta anoles, based on vertebral morphology and these sections show somewhat different geographical affinities (Etheridge 1960 cited in Williams 1969). Mexico and Jamaica have only beta anoles, Central America has both alphas and betas, South America has predominantly alphas, and Hispaniola, Puerto Rico and the Lesser Antilles all have only alphas. Cuba has both alpha and beta anoles and the United States has the alpha anole, *A. carolinensis.* Probably 12 or so species in the Caribbean belong to the *Anolis carolinensis* complex; a group of species which occur on the crowns and upper trunks of trees and which have green bodies (usually), certain scale features, and a moderate body size (less than 91 mm snout-vent length) relative to other anoles (Williams 1969). Since Mexico has only beta anoles the relationships of *A. carolinensis* are West Indian (Williams 1969).

Cuba has two species, *A. allisoni* and *A. porcatus,* which are very similar to *A. carolinensis.* It is suspected that *A. porcatus* is the ancestor to *A. carolinensis* (Williams 1969). Using allozyme information and estimates of genetic distance, Buth et al. (1980) indicated that the genetic distance between *A. porcatus* from Havana, Cuba and three populations of *A. carolinensis* from the U.S. was slightly higher than interpopulation, intraspecific distance estimates for other lizards. However, only two to three individuals were used from each population and no other species

were compared to these. The data of Buth et al. (1980) indicate a possible close relationship between these two species, but do not provide conclusive evidence that *A. porcatus* is the ancestor of *A. carolinensis*.

Shochat and Dessauer (1981) used antiserum to serum albumin from *A. carolinensis* to examine possible relationships between this species and other anoles, but did not find any close relationships between any West Indian anoles and *A. carolinensis*. Hass et al. (1993), using the same antiserum to serum albumin from *A. carolinensis* as Shochat and Dessauer (1981), obtained data on immunological distance units between this species and *A. allisoni* and *A. porcatus* from Cuba. They found that these two Cuban species were equally close to *A. carolinensis* in immunological distance. Though *A. porcatus*, rather than other West Indian anoles, is often thought to be the ancestor to *A. carolinensis*, more conclusive evidence is still needed to confirm this relationship.

Williams (1969) indicated that *A. carolinensis* and members of the *carolinensis* complex are rather good colonizers based on biogeographical evidence. It is quite clear that anoles of the Caribbean have been able to travel across open water by rafting to colonize many islands of the region (Williams 1969). It is likely that the ancestor to *A. carolinensis* colonized the mainland U.S. from Cuba since ocean currents are in that direction and the distance between Cuba and Florida is 100 miles or less (Williams 1969), regardless of the proof of the exact ancestor to *A. carolinensis*. Though the origins of *A. carolinensis* are neotropical, it has been able to colonize the mainland U. S. and push its distributional limits to areas of cold winters in

eastern Tennessee and North Carolina and to locations of low rainfall in central Texas.

The summaries in the next two sections on the distribution, climates, physiographic provinces, ecoregions, potential natural vegetation types, and habitats of *A. carolinensis* further illustrate the ability of this lizard of neotropical origin to colonize and inhabit fairly diverse areas.

Distribution, climates, and physiographic provinces

The distribution of *A. carolinensis* is from North Carolina and eastern Tennessee south through Florida and to Key West, throughout the Gulf Coast region, into southern Arkansas and southeastern Oklahoma, to eastern and central Texas, and is established in the lower Rio Grande Valley (Conant and Collins 1991). The northern limits of the distribution of *A. carolinensis* correspond approximately to the 50 °F (10 °C) "isocryme" as indicated by Gordon (1956), which appears to be the isotherm of the average annual low temperature. This northern limit is also generally south of or near the 4.4 °C average January isotherm (Wilson and Echternacht 1987:758, Fig. 1). Thus, temperature may play a role in determining the northern limits of this species. The western distributional limit in central Texas may be related to rainfall as it is approximated by areas where the average annual rainfall is less than 25 inches or 63.5 cm (Gordon 1956). Biogeographic analysis of 24 central Texas counties by Gehlbach (1991) suggests that the tallgrass prairie, which emerged with the start of warmer-drier climates in post-glacial time, is the major barrier to dispersal in an east-west direction for many terrestrial vertebrates (including *A. carolinensis*) in central Texas. Since central Texas is a

transition zone for many terrestrial vertebrates between an eastern deciduous forest region and a central-western evergreen woodland region (Gehlbach 1991), it is possible that the western distributional limit of *A. carolinensis* is defined by the interplay between climate and the structural habitat rather than some single aspect of climate alone.

The climate over the distribution of *A. carolinensis* falls into three groups/types according to the climate classification scheme of Köppen as modified by Trewartha (1968). In extreme southern Florida the climate is that of the Tropical Humid Climates group and the Tropical Wet-and Dry type (Trewartha 1968). This climate has hot summers and very mild winters with a rather small range in annual temperatures. Precipitation is seasonally distributed and a distinct dry season exists (Trewartha 1968).

In parts of south-central and southern Texas the climate is that of the Dry Climates group and the Steppe or Semiarid type. This climate has an annual loss of water from evapotransporation exceeding the annual gain from precipitation. The summers are hot and at least eight months have an average temperature over 50 ^{o}F (10 ^{o}C). Daily and annual temperature ranges can be rather large (Trewartha 1968).

The climate group and type found over most of the distribution of *A. carolinensis* is that of the Subtropical Climates group and the Subtropical Humid type. This climate occurs at middle latitudes in humid areas where the influence of both tropical and polar air masses is experienced. Summers are hot and winters are relatively mild. Most areas of the Subtropical Humid type experience freezing temperatures and frost and parts of the southern United States do experience severe cold spells.

Temperatures in Montgomery, Alabama and New Orleans, Louisiana have been recorded as low as -5 ° and 7 °F (or -20.56 and -13.89 °C), respectively (Trewartha 1968:299). Along the Gulf Coast temperatures as low as 10 °F (-12.22 °C) have been recorded and Trewartha (1968:299) states that "No other part of the world near sea level in these latitudes has such low winter minima.".

Anolis carolinensis occurs in many physiographic regions throughout the southern United States. Using both Fenneman's definitions and mapping of physiographic provinces (Fenneman 1931, 1938, United States Geological Survey 1970) and various distributional accounts from individual states, it is easily seen that *A. carolinensis* occurs in nearly all the major physiographic provinces in the 11 states in its range (Table 2-1[1]). In some physiographic provinces, such as the Coastal Plain, this species is fairly widespread. However, in other provinces, such as the Blue Ridge and the Appalachian Plateau, *A. carolinensis* occurs at elevations of 2250 ft or 686 m (Cochran 1938) or lower (Jones and Ressler 1927, King 1939, Johnson 1958) and often along stream or river edges.

Vegetation types, ecoregions, and habitats

The potential natural vegetation of the United States was described and mapped by Küchler (1964). Although some distributional accounts of *A. carolinensis* actually mention the Küchler vegetation type in which this lizard was found, most do not. However, a particular vegetation type can be inferred based on either a description of the vegetation of the area or a geographic overlap of the *A. carolinensis* location and a Küchler vegetation

type. Table 2-2 summarizes the results of a literature survey which documents either the occurence or non-occurence of *A. carolinensis* in Küchler vegetation types, as well as infers the possible presence of this lizard in such vegetation, in the southern United States. In general, this lizard is more common to many of the eastern forest types than to the central and eastern grasslands types and the western shrub and grassland types. However, some of the grassland and forest combinations (types 80, 90, 91, and 92) and some of the grassland types (type 79), all found in Florida, can provide conditions suitable for the occurrence of this species. Other grassland types, such as those in Louisiana and eastern and coastal Texas, possibly provide suitable habitat for this lizard.

The drier grasslands (types 65, 69, 74, 76), grassland and forest combinations (types 84-87), and western shrub and grasslands of Texas and Oklahoma (Table 2-2) are habitats within the Dry Climates group/Steppe type of Köppen. *A. carolinensis* does not occur in most of these drier habitats (see Table 2-2). However, in some of these dry regions in Texas it does occur, but probably does so in trees or shrubs along streams, rivers, or lakes where water is available rather than in the grassland vegetation. For example, Greg Sievert (personal communication) has observed adult and juvenile *A. carolinensis* in Big Bend National Park in Texas in shrubs and small willow trees along a stream, but not in the more characteristic grassland habitats of the region. The occurrences of *A. carolinensis* in counties with some of the drier habitats in Texas are thought to possibly

[1]All tables and figures can be found in the appendix at the end of the specific "Part" of this dissertation in which they are first cited.

represent introductions (Brown 1950:86, Dixon 1987:90), but this cannot be determined from distributional records alone.

Bailey (1976, 1980) described ecoregions of the United States as geographical ecosystems classified according to regional variations in landform, climate, and vegetation. An ecoregion is a continuous area and can "... be thought of as a geographical area over which the environmental complex, produced by climate, topography, and soil, is sufficiently uniform to permit development of characteristic types of ecological associations." (Bailey 1976). This concept differs from the "biome" concept of Shelford (1963) since a biome is based primarily on climax vegetation, whereas an ecoregion is defined by a number of ecological and environmental characteristics (Bailey 1976). Indeed, Bailey (1976) states that the proper classification of ecoregions should be based ecological associations of both plants and animals, but the data for such associations is lacking. Thus, the classification is based largely on climate and vegetation. This classification system can assist in management of land and resources, organization of resource inventory data, and interpretation of inventory data (including flora and fauna) (Bailey 1980).

The ecoregion concept consists of a hierarchy of levels and Bailey (1976) provided for nine such levels. Only the first four levels will be related to the distribution of *A. carolinensis*. A domain is the top level and is a subcontinental area having similar or related climates. A division is an area within a given domain which includes a single regional climate as described by Köppen's types and modified by Trewartha (1968). Next, a province is a broad vegetation region having the same or similar types of

zonal soils and a fairly uniform regional climate. The fourth level is the section which is an area within a province defined by the climatic climax as indicated by Küchler's (1964) potential natural vegetation types (Bailey 1976). Thus, a section level in Bailey's classification scheme often corresponds with either a single Küchler vegetation type or a combination of two or more types.

Table 2-3 indicates the ecoregions in which *A. carolinensis* occurs over its 11 state range. None of the references found on *A. carolinensis* actually mention Bailey's ecoregion classification so the association of this lizard with any section level has been inferred in Table 2-3 based on the distributional accounts of this lizard. In some cases information on occurrence of *A. carolinensis* in certain Küchler vegetation types (as in Table 2-2) indicates that this species is associated with the predominant vegetation in a given ecoregion section. Most of the distribution of *A. carolinensis* falls within the Subtropical and the Hot Continental divisions of the Humid Temperate domain of Bailey's (1976, 1980) ecoregion classification. The forests of the Subtropical division are mainly coniferous and mixed deciduous-coniferous forest, while those of the Hot Continental division are deciduous forests. Within the southern United States, *A. carolinensis* apparently occurs in all of the ecoregion sections of these two divisions, as well as the Everglades province of the Savanna division of the Humid Tropical domain in southern Florida (Table 2-3).

Green anoles are not likely to be common in the Prairie division and Desert division of Bailey's ecoregions in Texas since these regions receive less annual precipitation than other regions throughout this lizard's range.

However, natural populations and some possible introductions of *A. carolinensis* do occur in both central and southern Texas (Dixon 1987) so it is a faunal element of the Juniper-Oak-Mesquite (2522), Mesquite-Acacia (2523), and Tarbush-Creosote Bush (3212) sections (Table 2-3). It should be remembered that the vegetation within a given section is not entirely of one plant community or association, but that the vegetation may vary according to more local environmental conditions. Thus, as mentioned with the Küchler vegetation types of Table 2-2, *A. carolinensis* probably does not occur frequently or at all in the drier grasslands of Texas, but along riparian habitats in such areas.

The occurrence of *A. carolinensis* in many section levels of Bailey's (1976, 1980) ecoregion classifications and in many of Küchler's (1964) vegetation types is, in part, probably due to this lizard's ability to colonize and inhabit a wide variety of habitat conditions. Typically this species has been described as one which inhabits ecotonal or "edge" habitats (Gordon 1956) and somewhat open areas with dense ground cover (Dundee and Rossman 1989). However, it can be found deep in forests in Louisiana (Dundee and Rossman 1989) and areas with abundant vegetation and shade, including shady residential areas, in Alabama (Mount 1975).

Many of the early faunal surveys and distributional accounts of *A. carolinensis* in the literature gave brief accounts of the vegetation or habitats within which this species was found and Table 2-4 provides a summary of these early accounts as well as later ones. Not all of the literature accounts of the habitat for *A. carolinensis* are summarized in this table, but rather a representative cross-section is provided. Without

repeating much of the information already presented in the previous tables, Table 2-4 shows that *A. carolinensis* occurs in many different types of habitats throughout its range, including many human-dominated habitats and urban areas, evergreen woodlands, deciduous forests, swamp habitats, and the Everglades of Florida.

NATURAL HISTORY OF *ANOLIS CAROLINENSIS*

Territoriality, movements, and home range size

Many laboratory studies have examined the behavioral interactions between individuals of *A. carolinensis* regarding territoriality (see, e.g., Greenberg and Noble 1944, Cooper 1977, Crews 1980). However, the focus of the discussion here is on studies which have been conducted in the field.

Gordon (1956) found that aggressive encounters between adult males began during late February and early March in study sites in Louisiana. Males then establish and defend territories during April through August which coincides with the breeding season at his study areas. A male territory holder would begin each day looking around its territory usually from an observation post. The individual would challenge any intruders and court females within his territory. Only a small number of the interactions between males in the field actually resulted in physical contact or combat (such as biting) between the individuals as most encounters involved challenge and bluffing by one male. Encounters in which physical contact did occur between males took place during early spring (Gordon 1956). At a study site in South Carolina, Jenssen et al. (1995a) found that male *A. carolinensis* patrolled and defended a territory from

other males during the breeding season (through July), but not so immediately after the breeding season (August-September).

Aggression and territoriality in the field are exhibited by adult females, but can be more difficult to observe since females tended to be less conspicuous than males (Gordon 1956). Territories of females overlapped with a territory or home range of a male. In some cases more than one adult female resided within at least part of an adult male's territory. In such cases a dominant female existed and encounters would occur between the dominant female and the others, as well as among subordinate females. Encounters usually took place when one female entered the basking site of another or intruded into the immediate vicinity of a feeding female (Gordon 1956).

Movements of marked individuals have been examined for *A. carolinensis* in a few cases. Gordon (1956) estimated average horizontal distances moved for an "unrestricted" data set and a "restricted" data set. The unrestricted data on any marked individual included any and all movements between recaptures during the year, including movements to and from "hibernacula" (overwintering sites). The restricted data, however, included the more typical movements made during the year, such as movements within a territory or home range, and excluded the more infrequent movements which were not made on a routine basis. Examples of these excluded infrequent movements, which often covered somewhat larger distances, were movements to enter overwintering sites, reestablish a territory or home range, and feed opportunistically when

termites appeared in large numbers in certain spots during nuptial flights (Gordon 1956).

For the unrestricted data on all size classes combined, females moved an average distance of around 13 ft (3.96 m) at the Bridge City site and 22 ft (6.71 m) at the Plauche site, whereas males moved an average distance of 14 ft (4.27 m) at the Bridge City site and 21 ft (6.40 m) at the Plauche site. Overwintering areas were in closer proximity to summer habitat at Bridge City than Plauche, therefore accounting for these smaller distances for Bridge City anoles (Gordon 1956). For the restricted data covering the more typical movements over the year, the average distance moved for females was 8 ft (2.44 m) at the Bridge City site and 12 ft (3.66 m) at the Plauche site, whereas the average for males was 10 ft (3.05 m) at the Bridge City site and 12 ft (3.66 m) at the Plauche site (Gordon 1956). The Bridge City site was dominated by woody perennials, whereas the Plauche site was dominated by herbaceous annuals (Gordon 1956), so perhaps some differences here could be related to habitat structure.

King (1966) estimated horizontal movement distances for *A. carolinensis* at a southern Florida study site. He found that the average distance moved by adult males was 2.89 $\pm$ 1.92 m and by juveniles was 2.13 $\pm$ 0.86 m over a 12 week period. King could not obtain an average for adult females since only one such individual was recaptured.

By observing and videotaping focal males, Jenssen et al. (1995a) were able to estimate distances moved in association with several behaviors during part of the breeding season (May-July) and post-breeding season (August-September). Average distances moved by these males in

conjunction with any given event were fairly short during the breeding season and even shorter during the post-breeding season. For example, the greatest average distance moved per bout was 2.3 m/bout during travel events (i.e., moving from one perch to another without pausing for > 60 seconds and without conducting other behaviors) during the breeding season. During the breeding season however, adult males moved frequently, spent 25% of their day moving through their territories, and showed high rates of display and locomotion compared with other lizards (Jenssen et al. 1995a).

Gordon (1956) also provided some rough estimates of home range sizes for two populations in Louisiana using the "restricted" data from known individuals captured three or more times. He assumed that a home range was somewhat circular and the average distance moved between recaptures by any individual approximated half the diameter of a home range. To get the diameter of the home range size Gordon simply doubled the average distance moved. Whether or not this provides an accurate measure of home range diameter is not known.

Some reports of territory sizes of adult males were given by Gordon (1956), but no comprehensive or systematic estimates were made. One adult male living near the sow-pen area of one site, for example, had a territory of approximately 15 by 16 ft (4.57 by 4.88 m) and a maximum height of about 13 ft (3.96 m). Gordon (1956) indicated that territory shape and size were often determined by the vegetation structure and boundaries were often defined by the natural breaks in the vegetation. Jenssen et al. (1995a) reported a mean home range volume of 173.6 m^3 (standard error

37.9) for males during May-September in South Carolina, but did report the non-vertical distances used to calculate this volume.

Reproduction

Numerous laboratory studies have been conducted on the behavioral, hormonal, and neuroendocrine aspects of courting, reproduction, and the reproductive cycle of *A. carolinensis* (see Crews 1980). However, only a few studies have examined aspects of reproduction in wild populations and most of the available information comes from Louisiana populations. The information presented here is mainly from studies of green anoles either collected in the field or studied in the field, rather than from laboratory-maintained anoles.

Very little aggression occurs between male *A. carolinensis* in Louisiana from October to February, but by 1 April males have established individual territories (Gordon 1956). Females establish their home ranges overlapping one or more male territories. Mating in *A. carolinensis* begins in late March in Louisiana (Hamlett 1952) and has been observed as early as 1 April in an eastern Tennessee population (Minesky, personal observation). The peak period for mating in eastern Tennessee is from mid-May to mid-July (Wade 1981). Copulation continues through August in Louisiana (Hamlett 1952, Gordon 1956), but ends by August (only one copulation observed in early August) in South Carolina (Jenssen et al. 1995a).

The gravid condition of females occurs from early April through August in Louisiana (Hamlett 1952, Dessauer 1955, Gordon 1956) and from early May through mid-August in Tennessee (Wade 1981). Occasionally a few gravid females occur in March (Gordon 1956) and on into September

(Hamlett 1952) or October (Cagle 1948). During the early part of the reproductive season in Louisiana less than half of all females are gravid, but after 1 May nearly all females are gravid (Hamlett 1952). Cagle (1948) found that in a May sample in Louisiana the percentage of gravid females increased with the size class of the females.

Females ovulate one egg at a time and ovulation alternates between the two ovaries (Hamlett 1952). Some females do contain two oviducal eggs, one in each oviduct, at the same time (Cagle 1948, Hamlett 1952, Gordon 1956, King 1966, Wade 1981), with larger females apparently being more likely to have two oviducal eggs than smaller females for both Louisiana (Cagle 1948, Gordon 1956) and Tennessee (Wade 1981) populations. Ovulation occurs about every 13-14 days, on average, in the lab (Hamlett 1952). Eggs are usually laid one per clutch based on observations of gravid females brought into the lab (Hamlett 1952, Gordon 1956, Michaud 1990) and females held in outdoor enclosures (Michaud 1990). Oliver (1955:244) reported that females in Florida usually lay two eggs per clutch, but this may be incorrectly based on the fact that females can have two oviducal eggs, at different developmental stages, present at the same time. Eggs are laid from late April through August in Louisiana (Hamlett 1952) and southern Florida (King 1966). Predation on eggs does occur as Tinkle (1959) reported that eggs of *A. carolinensis* were found in the stomach of one *Lampropeltis getulus* in Louisiana.

Hatchlings first appear during June in Louisiana (Gordon 1956) and during July in Texas (Michael 1972), Florida (King 1966), and eastern Tennessee (Minesky, unpublished data). The body size of hatchlings is

about 22-25 mm from eggs laid in the lab (Hamlett 1952) and 19.4-24.4 mm for eggs collected in the field and then brought into the lab (Gordon 1956) for Louisiana individuals. In southern Florida, field caught hatchlings (with open umbilicus) were 19-22 mm SVL (King 1966). Individuals 22-26 mm SVL and weighing 0.2-0.4 g, many of which still have a visible yolk scar, have been observed in a population in eastern Tennessee (Minesky, unpublished data). In a non-captive Texas population, young with umbilical scars generally ranged from 22-27 mm SVL (Michael 1972).

Growth, maturity, and longevity

Actual growth rates for non-captive *A. carolinensis* have been published by only a few researchers. Gordon (1956) reported average daily increments in body length for various size classes of two different populations in Louisiana and Michael (1972) published growth rates for a Texas population. Their results indicated that hatchlings grow rather rapidly (usually 0.2 mm/day or more for Louisiana and 2.5 mm/month for Texas), males grow more rapidly than females, and larger females and males grow the slowest of any size class prior to winter. Growth during winter occurs in many individuals in Louisiana although it is slower than growth in other seasons (Gordon 1956).

Michael (1972) found that in a Texas population the growth rates of August and September hatchlings were not significantly different, but both were significantly less than July hatchlings prior to winter. By November, the difference between mean SVLs for the August hatchlings and the September ones was 3.4 mm, but by the following April this same difference was only 1.1 mm. This suggests that the September hatchlings,

although entering winter at a smaller size than August hatchlings, may be growing faster in the early spring and closing the size gap. Similar comparisons with July hatchlings could not be made due to low survival rates of July individuals (Michael 1972). Growth rates on captive animals have also been reported by Fox and Dessauer (1958) and Michaud (1990).

Sexual maturity is defined as having at least one oviducal egg for females and as having mature spermatazoa for males. Females in Louisiana become mature after reaching a minimum body size of 45 mm (Gordon 1956), 46 mm (Cagle 1948), or 45-48 mm (Hamlett 1952). The smallest female with an oviducal egg from a sample of Tennessee individuals was 48 mm SVL (see Wade 1981:58), but eggs have been palpated in some Tennessee females as small 44 mm (Michaud 1990) and 45 mm (Minesky, unpublished data) SVL. In central Florida near Orlando females typically mature at 44-45 mm SVL, but some sexually mature females as small as 42 mm SVL have been found (Michaud 1990), whereas in the Miami area the smallest female with an oviducal egg was 41 mm SVL (King 1966). Females in a Texas population reach sexual maturity at 45 mm SVL by their second summer (Michael 1972). From his field study of *A. carolinensis* in Louisiana Gordon (1956) indicated that adult size was attained by all individuals which were marked as hatchlings and then recovered during the mating season after their first winter season. Tennessee females appear to reach sexual maturity during the spring or summer after their first winter (personal observation).

Male *A. carolinensis* in Louisiana mature at a body size of approximately 45 mm SVL and sexual maturity is reached at least during

the mating season after the first winter (Gordon 1956). Fox and Dessauer (1958), studying captive green anoles, found that males under 55 mm could produce sperm, but lacked completely developed accessory sex organs. Thus, sexual maturity in terms of being capable of mating is not attainable by males less than 55 mm. Males in a non-captive Texas population can reach lengths of 55 mm by 18 months and 60 mm by 36 months after hatching, but most of the matings seem to be done by those males over 60 mm SVL (Michael 1972).

Survival of hatchlings and juveniles is considered to be rather low (Gordon 1956, Michael 1972) although precise estimates are not readily available. Accurate estimates of longevity of *A. carolinensis* are also scarce. Gordon (1956:252) stated that "No accurate measure of longevity was obtained." and reported that the longest period between recoveries of marked anoles was 522 and 587 days. He also indicated that the turnover rate in the population each year was very high. On a two acre study area in northern Florida, Oliver (1955) reported that 98% of the 200 marked green anoles were not found only 12 months after first being captured and no individuals were found after 16 months. In southern Florida King (1966) found a low frequency of recapture occurred over about half a year study time, possibly indicating a high turnover rate in the population and short life span.

Michael (1969) considered one marked male initially captured at 58 mm SVL to be at least five years old at the time it was hit and killed by a car. However, subsequent data on growth rates on this urban population in Texas (Michael 1972) show that males average about 55 mm SVL after

12-15 months of life rather than the 24 months or more initially reported by Michael (1969). This would suggest that this individual was probably first caught in its second October of life and therefore four years old at the time of its death. Regardless, his data do report one of the longest lived non-captive green anoles for any study. Based on mark-recapture data in eastern Tennessee, *A. carolinensis* can reach an estimated minimum age of two years and some individuals have attained at least 4 years of age (Minesky, unpublished data).

Daily activity

Activity by *A. carolinensis* occurs primarily during daylight hours, although some movement and feeding in warmer months has been observed at night under bright moonlight conditions in Louisiana (Gordon 1956). This species sleeps at night in trees and other vegetation in warm weather and in or under suitable cover during winter (Gordon 1956). Male territory holders would become active shortly after dawn in Louisiana. At overwintering sites during winter, Gordon (1956) observed activity and even feeding as early as 0907-0931 hrs when air temperatures were 67-69 °F (19.44-20.56 °C).

During April in northern Florida, this species is most active from 0800 to about 1000 or 1100 hrs and then again from 1600 to about 1800 or 1900 hrs (Oliver 1955). King (1966) counted the number of active green anoles at various times of day in March and August in southern Florida. He found the peak numbers of active individuals occurred around 1130 hrs in March, but very little activity occurred before 0930 or after 1530 hrs. Active

numbers peaked at 0930 and 1530 hrs in August, while mid-day showed the fewest active individuals (King 1966).

Time spent by adult male *A. carolinensis* on various activities between 0900 and 1900 hrs during May-July and a post-breeding period (August-September) was reported by Jenssen et al. (1995a) in South Carolina. During May-July, 49% of the daily time budget (on average), was spent in various activities vs. 51% of the time spent being stationary. Most of this activity time was spent "traveling", moving from one perch to another with only short pauses (< 60 seconds) and no other activities taking place at the time (such as foraging, displaying, avoiding predators, etc.). During the August-September period however, adult males were active only 21% of the time during the day in which most of this time was spent slowly creeping from one perch site to another (Jenssen et al. 1995a).

Winter activity

In the face of cold temperatures in winter, lizards (and ectotherms in general) can either become dormant (by means of hibernation or cold-induced torpor) or remain active and pay the associated energy costs (Ragland et al. 1981). Winter activity by ectotherms can potentially take place for extended periods on a daily basis, for short periods on a daily basis, for extended periods only during favorable weather, or for short periods only during favorable weather. Activity levels of ectotherms during winter can range from low to moderate levels, such as emergence from winter refugia and facultative basking during favorable weather conditions, to high levels, such as active thermoregulation, foraging, and digestion.

The term hibernation has been used by Hamilton (1948), Neill (1948b), Dessauer (1953), Gordon (1956), Michael (1972), Michael and Bailey (1972), and Crews (1980) in association with the reduction in or apparent lack of activity by *Anolis carolinensis* during winter (see also Gregory 1982). Whether or not one considers the behavior of this species during winter as either that of "activity" or "hibernation" (or even some combination of activity and cold-induced inactivity) really depends on how these terms are defined. Neill (1948b:107) rather simplistically defined hibernation as "... to include any sort of retreat from winter conditions, whether or not actual dormancy is involved.". Gregory (1982), in a review of reptilian hibernation, indicated that any period of winter dormancy has usually been termed hibernation and stated "In this chapter, winter dormancy in reptiles is called hibernation." (Gregory 1982:56). This rather loose definition of hibernation apparently stems from both the fact that specific features of reptilian hibernation have not been determined and the idea that behavioral definitions of hibernation may be just as useful as physiological ones (see Gregory 1982). Thus, for *A. carolinensis* and other reptiles, hibernation has been defined predominantly by behavior rather than physiology (or even a combination of both).

Even though Neill (1948b) and Gordon (1956) used the term hibernation regarding winter behavior in *A. carolinensis*, both also reported this species to be active on warm winter days. Activity occurred when temperatures were 60°F (15.56°C) or greater and inactivity occurred at temperatures of 55°F (12.78°C) or less (Gordon 1956). A wide variety of other reptiles also show both activity and inactivity during the winter (see

Gregory 1982). The term "discontinuous hibernator" has been used to describe species which are at times active in winter, but this classification does not easily distinguish such animals from winter-active ones (Gregory 1982).

During rather cold periods and on cloudy days in the winter, *A. carolinensis* will take cover. This species has been found in or under rotted logs (Hamilton 1948, Gordon 1956) and in rotting stumps, piles of fallen logs, holes in fence posts, pits filled with concrete gravel, and in clusters of peach baskets (Gordon 1956) in Louisiana, under rotted logs, bark scraps and under the bark of old stumps in Georgia (Neill 1948b), and in rock crevices in Tennessee (Minesky, personal observation and this study) during winter.

Activity has been reported for *A. carolinensis* during winter such as emergence from cover and basking when weather conditions and temperatures were mild (Gordon 1956, Ragland et al. 1981, Gatten et al. 1988, Gibbons and Semlitsch 1991, Jenssen et al. 1996). In Tennessee near its northern distributional limits, *A. carolinensis* has been observed basking on sunny winter days even when ambient air temperatures are between -1.0 and 1.0 ^{o}C. Body temperatures on such days may be as low as 10.4 to 15 ^{o}C when anoles begin basking or have been active for probably a short time (Echternacht and Minesky, unpublished data). During cold, cloudy weather, rainy days, and at night in the winter this lizard seeks shelter in rock crevices or other suitable cover (personal observation).

At one study site along the Little Tennessee River at least four lizard species (*Sceloporus undulatus, Cnemidophorus sexlineatus, Eumeces*

fasciatus, and *Scincella lateralis*) are sympatric with *A. carolinensis* and have been seen during spring or summer for several years. It is an extremely rare occurrence for any individuals of these other lizard species to be active during December, January, or February at this site and yet it is common to see *A. carolinensis* during this time (personal observation).

Another reason why "hibernation" may not be an appropriate term to describe the winter behavior of *A. carolinensis* is because some growth can apparently occur during winter. Growth in body length occurs in juvenile and adult *A. carolinensis* in Louisiana during winter, although rates of growth are slower than those observed in other seasons (Gordon 1956). Small amounts of growth, as well as shedding of skin, have also been observed in a few winter individuals in Tennessee (Minesky, unpublished data). It would seem rather unusual for an animal which "hibernates" to undergo a measurable amount of growth.

Jenssen et al. (1996) found that *A. carolinensis* in South Carolina emerged from cover on sunny days from November through mid-April. During this time, observations made between 1000-1500 h on sunny days with temperatures between 12-32 ^{o}C revealed that these lizards spent 66.8% of their time in the sun. Of their total emergence time, only 6.4% was spent foraging, while 92.2% was spent being stationary (i.e., at rest), on average (Jenssen et al. 1996). So although *A. carolinensis* shows basking activity and some foraging, the average time spent being stationary during winter was higher than that for May-July (50.7%) and August-September (78.8%) periods in South Carolina (Jenssen et al. 1995a). Overall average distance moved during emergence time in winter was considerably less by

adults (females and males combined: 2.66 cm per h from November-February and 1.68 cm per h from March-April, Jenssen et al. 1996) than that in summer months (26 m per h from May-July and 8 m per h from August-September, Jenssen et al. 1995a).

All of these behavioral observations suggest that *A. carolinensis* shows some activity (although sometimes reduced compared with other seasons) and none of the behavior associated with "hibernation", particularly in comparison with any sympatric lizard species. Therefore, the term hibernation will not be used here to describe the winter behavior of *A. carolinensis.* Some studies, as discussed in the next subsection, suggest that the physiology and metabolism of *A. carolinensis* during winter are reduced compared to that during other seasons.

Winter thermal physiology and metabolism

Geographic variation within *A. carolinensis* occurs with respect to Critical Thermal Minimum (CTMin) values. Wilson and Echternacht (1987), comparing adult male *A. carolinensis,* found that those from eastern Tennessee had significantly lower CTMin values than adult males from southern Georgia and central Florida after all lizards had been acclimated to cold in the lab.

Gatten et al. (1988) found that aerobic metabolism during exercise did not change seasonally in *A. carolinensis* from Tennessee. However, January and March animals had lower post-exercise lactate levels than animals from other seasons, thus indicating less glycolytic metabolism was occurring during exercise. Gatten et al. (1988) concluded that, because glycolysis provides the majority of ATP during intense muscular activity,

the ability of *A. carolinensis* to fuel locomotion by glycolytic metabolism was reduced in anoles from winter and early spring compared to those from other seasons. Thus, *A. carolinensis* is somewhat active in winter, but may have a reduced capability to fuel muscular activity via glycolysis.

Ragland et al. (1981) compared sympatric winter field-caught *A. carolinensis* and *Cnemidophorus sexlineatus* from east-central Alabama in terms of oxygen consumption. *C. sexlineatus* is a species known to be inactive from late August to at least April. The winter lizards were brought into the laboratory and oxygen consumption, measured at 10, 20, and 30 ^{o}C, was higher for *A. carolinensis* than *C. sexlineatus* (Ragland et al. 1981). This indicates that *A. carolinensis* is not nearly as "dormant" in winter as *C. sexlineatus* (Ragland et al. 1981) from a metabolic standpoint.

Other studies have also measured resting metabolic rates (via oxygen consumption) in *A. carolinensis* at various temperatures (see Gatten et al. 1988 for review and Jenssen et al. 1996:207). Results from the different studies are very similar, despite some differences in protocol, within the temperature at which lizards were tested. In general, the various studies show that resting metabolism at 10 ^{o}C can be between seven to nine times lower than that at 30 ^{o}C.

Jenssen et al. (1996) concluded that *A. carolinensis* in South Carolina during the winter often raise their body temperatures above the ambient temperature, but do not precisely regulate body temperatures. They characterized winter *A. carolinensis* as "passive thermal generalists" because these lizards moved infrequently and did not shuttle between sunny and shaded patches. If the anoles were to actively thermoregulate to

maintain optimal body temperatures observed in spring and summer, they suggest, then the metabolic cost would be 2.5 times that observed for the mean winter body temperature of 23 ^{o}C. Such a scenario would cause anoles in South Carolina to utilize much of their limited lipid reserves because foraging activities were infrequent and night-time temperatures were probably too low for sufficient digestion. This thermal passivity during winter was then speculated to be an "adaptive compromise" between an inability to remain dormant and the need to conserve energy during any winter activity (Jenssen et al. 1996). However, Jenssen et al. (1996:207) admit that the regressions of body temperature and air temperature on the hour of day and the fairly constant average body temperatures during the day "... provides evidence of some thermoregulation ..." by *A. carolinensis* during winter.

Obviously, the winter behavior of *A. carolinensis* is different from other reptiles which are dormant for the entire winter (such as other lizard species which are sympatric with this species). Whether *A. carolinensis* actually has a physiological inability to remain in a prolonged dormancy and/or adaptations for conserving energy by being thermally passive during winter is still open for debate and in need of further research.

LITERATURE CITED

Ashton, R. E., Jr. and P. S. Ashton. 1985. Handbook of reptiles and amphibians of Florida: part 2-lizards, turtles, and crocodilians. Windward Publishing, Inc., Miami, Florida, USA. 191 pp.

Bailey, R. G. 1976. Ecoregions of the United States (map). U. S. Department of Agriculture, Forest Service Intermountain Region, Ogden Utah, USA. Scale 1:7,500,000.

Bailey, R. G. 1980. Descriptions of the ecoregions of the United States. U. S. Department of Agriculture, Miscellaneous Publication No. 1391, Washington D.C., USA. 77 pp.

Brown, B. C. 1950. An annotated check list of the reptiles and amphibians of Texas. Baylor University Press, Waco, Texas, USA. 257 pp.

Bryant, W. S., W. C. McComb, and J. S. Fralish. 1993. Oak-hickory forests (western mesophytic/oak-hickory forests). Pages 143-201 *in* W. H. Martin, S. J. Boyce, and A. C. Echternacht, editors. Biodiversity of the southeastern United States: upland terrestrial communities. John Wiley and Sons, Inc., New York, New York, USA. 373 pp.

Buth, D. G., G. C. Gorman, and C. S. Lieb. 1980. Genetic divergence between *Anolis carolinensis*, and it Cuban progenitor, *Anolis porcatus*. Journal of Herpetology 14(3):279-284.

Cagle, F. R. 1948. A population of the Carolina anole. Natural History Miscellanea 15:1-5.

Cochran, D. M. 1938. An addition to the lizard fauna of Tennessee. Copeia 1938(2):90.

Cooper, W. C. 1977. Information analysis of agonistic behavioral sequences in the male iguanid lizard *Anolis carolinensis*. Copeia 1977: 721-735.

Conant, R. and J. T. Collins. 1991. A field guide to reptiles and amphibians. 3rd ed. Houghton Mifflin Company, Boston, Massachusetts, USA. 450 pp.

Corrington, J. D. 1927. Field notes on some amphibians and reptiles at Biloxi, Mississippi. Copeia 1927(165):98-102.

Crews, D. 1980. Interrelationships among ecological, behavioral, and neuroendocrine processes in the reproductive cycle of *Anolis carolinensis* and other reptiles. Pages 1-74 *in* J. S. Rosenblatt, R. A. Hinde, C. Beer, and M. C. Busnel, editors. Advances in the study of behavior. Academic Press, New York, New York, USA. 377 pp.

Dalrymple, G. H. 1988. The herpetofauna of Long Pine Key, Everglades National Park, in relation to vegetation and hydrology. Pages 72-86 *in* R. C. Szaro, K. E. Severson, and D. R. Patton, technical coordinators. Management of amphibians, reptiles, and small mammals in North America. USDA Forest Service General Technical Report RM-166, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado, USA. 458 pp.

Dessauer, H. C. 1953. Hibernation of the lizard, *Anolis carolinensis.* Proceedings of the Society of Experimental Biology and Medicine 82: 351-353.

Dessauer, H. C. 1955. Seasonal changes in the gross organ composition of the lizard, *Anolis carolinensis*. Journal of Experimental Zoology 128:1-12.

Dixon, J. R. 1987. Amphibians and reptiles of Texas. Texas A&M University Press, College Station, Texas, USA. 434 pp.

Dowling, H. G. 1957. A review of the amphibians and reptiles of Arkansas. Occasional Papers of the University of Arkansas Museum, No. 3, University of Arkansas, Fayetteville, Arkansas, USA. 51 pp.

Duellman, W. E. and A. Schwartz. 1958. Amphibians and reptiles of southern Florida. Bulletin of the Florida State Museum, Biological Sciences 3(5):181-324. University of Florida, Gainesville, Florida, USA.

Dundee, H. A. and D. A. Rossman. 1989. The amphibians and reptiles of Louisiana. Louisiana State University Press, Baton Rouge, Louisiana, USA. 300pp.

Engels, W. L. 1952. Vertebrate fauna of North Carolina coastal islands II. Shackleford Banks. American Midland Naturalist 47(3):702-742.

Etheridge, R. 1960. The relationships of the anoles (Reptilia: Sauria: Iguanidae): an interpretation based on skeletal morphology. Ph.D. dissertation. University Microfilms, Inc., Ann Arbor, Michigan, USA. 236 pp.

Etheridge, R. and K. de Queiroz. 1988. A phylogeny of Iguanidae. Pages 283-368 *in* R. Estes and G. Pregill, editors. Phylogenetic relationships of lizard families: essays commemorating Charles L. Camp. Stanford University Press, Stanford, California, USA. 631 pp.

Etheridge, R. and E. E. Williams. 1985. Notes on *Pristidactylus* (Squamata: Iguanidae). Breviora 483:1-18.

Fenneman, N. M. 1931. Physiography of western United States. McGraw-Hill Book Company, New York, New York, USA. 534 pp.

Fenneman, N. M. 1938. Physiography of eastern United States. McGraw-Hill Book Company, New York, New York, USA. 714 pp.

Fitch, H. S. 1970. Reproductive cycles of lizards and snakes. University of Kansas Museum of Natural History Miscellaneous Publication No. 52: 1-247.

Fitch, H. S. 1982. Reproductive cycles in tropical reptiles. Occasional Papers of the Museum of Natural History, University of Kansas No. 96:1-53.

Fox, W. and H. C. Dessauer. 1958. Growth rates in captive male green anoles. Herpetologica 14:196-197.

Frost, D. R. and R. Etheridge. 1989. A phylogenetic analysis and taxonomy of Iguanian lizards (Reptilia: Sauria). The University of Kansas Museum of Natural History Miscellaneous Publication No. 81:1-65.

Gatten, R. E. Jr., A. C. Echternacht, and M. A. Wilson. 1988. Acclimatization versus acclimation of activity metabolism in a lizard. Physiological Zoology 61:322-329.

Gehlbach, F. R. 1991. The east-west transition zone of terrestrial vertebrates in central Texas: a biogeographic analysis. Texas Journal of Science 43(4):415-427.

Gibbons, J. W. and R. D. Semlitsch. 1991. Guide to the reptiles and amphibians of the Savannah River site. The University of Georgia Press, Athens, Georgia, USA. 131 pp.

Gilmore, R. G., Jr., and S. C. Snedaker. 1993. Mangrove forests. Pages 165-198 *in* W. H. Martin, S. J. Boyce, and A. C. Echternacht, editors. Biodiversity of the southeastern United States: lowland terrestrial communities. John Wiley and Sons, New York, New York, USA. 502 pp.

Gordon, R. E. 1956. The biology and demography of *Anolis carolinensis carolinensis* Voigt. Unpublished Ph. D. dissertation. Tulane University, New Orleans, Louisiana, USA. 263 pp.

Greenberg, B. and G. Noble. 1944. Social behavior of the American chameleon (*Anolis carolinensis* Voigt). Physiological Zoology 17:392-439.

Greenberg, C. H., D. G. Neary, and L. D. Harris. 1994. Effect of high-intensity wildfire and silvicultural treatments on reptile communities in sand-pine scrub. Conservation Biology 8:1047-1057.

Gregory, P. T. 1982. Reptilian hibernation. Pages 53-154 in C. Gans and F. H. Pough, editors. Biology of the reptilia. Vol. 13. Academic Press, London, Great Britain. 345 pp.

Gunderson, L. H. and W. F. Loftus. 1993. The everglades. Pages 199-255 *in* W. H. Martin, S. J. Boyce, and A. C. Echternacht, editors. Biodiversity of the southeastern United States: lowland terrestrial communities. John Wiley and Sons, New York, New York, USA. 502 pp.

Hamilton, W. J., Jr. 1948. Hibernation site of the lizards *Eumeces* and *Anolis* in Louisiana. Copeia 1948:211.

Hamlett, G. W. D. 1952. Notes on breeding and reproduction in the lizard *Anolis carolinensis*. Copeia 1952:183-185.

Hass, C. A., S. B. Hedges, and L. R. Maxson. 1993. Molecular insights into the relationships and biogeography of West Indian anoline lizards. Biochemical Systematics and Ecology 21:97-114.

Huheey, J. E. and A. Stupka. 1972. Amphibians and reptiles of Great Smoky Mountains National Park. The University of Tennessee Press, Knoxville, Tennessee, USA. 98 pp.

Jenssen, T. A., N. Greenberg, and K. A. Hovde. 1995a. Behavioral profile of free-ranging male lizards, *Anolis carolinensis*, across breeding and post-breeding seasons. Herpetological Monographs 9:41-62.

Jenssen, T. A., J. D. Congdon, R. U. Fischer, R. Estes, D. Kling, and S. Edmands. 1995b. Morphological characteristics of the lizard *Anolis carolinensis* from South Carolina. Herpetologica 51:401-411.

Jenssen, T. A., J. D. Congdon, R. U. Fischer, R. Estes, D. Kling, S. Edmands, and H. Berna. 1996. Behavioral, thermal, and metabolic characteristics of a wintering lizard (*Anolis carolinensis*) from South Carolina. Functional Ecology 10:201-209.

Johnson, R. M. 1958. A biogeographic study of the herpetofauna of eastern Tennessee. Unpublished Ph.D. dissertation. University of Florida, Gainesville, Florida, USA. 221 pp.

Jones, J. P. and B. C. V. Ressler. 1927. The occurrence of *Anolis carolinensis* Voigt in eastern Tennessee. Copeia 1927:86-87.

King, W. 1939. A survey of the herpetology of Great Smoky Mountains National Park. American Midland Naturalist 21:531-582.

King, F. W. 1966. Competition between two south Florida lizards of the genus *Anolis*. Unpublished Ph.D. dissertation. University of Miami, Coral Gables, Florida, USA. 97 pp.

Köppen, W. 1931. Grundriss der Klimakunde. Walter de Gruyter Company, Berlin, Germany. 388 pp.

Küchler, A. W. 1964. Potential natural vegetation of the conterminous United States. American Geographical Society Special Publication No. 36. American Geographical Society, New York, New York, USA. 116 pp. With separate map at 1:3,168,000.

Lee, D. S. 1969. Floridian herpetofauna associated with cabbage palms. Herpetologica 25:70-71.

Lohoefener, R. and R. Altig. 1983. Mississippi herpetology. Mississippi State Research Center Bulletin 1, Mississippi State, Mississippi, USA. 66 pp.

Martoff, B. S., W. M. Parker, J. R. Bailey, and J. R. Harrison III. 1980. Amphibians and reptiles of the Carolinas and Virginia. The University of North Carolina Press, Chapel Hill, North Carolina, USA. 264 pp.

Michael, E. D. 1969. A longevity record for a non-captive *Anolis carolinensis*. Herpetologica 25:318.

Michael, E. D. 1972. Growth rates in *Anolis carolinensis*. Copeia 1972:575-577.

Michael, E. D. and T. F. Bailey. 1972. Hibernation sites of *Anolis carolinensis* and *Sceloporus undulatus*. Texas Journal of Science 24: 351-353.

Michaud, E. J. 1990. Geographic variation of life history traits in the lizard, *Anolis carolinensis*. Ph.D. dissertation. University of Tennessee, Knoxville, Tennessee, USA. 174 pp.

Mount, R. H. 1975. The reptiles and amphibians of Alabama. Auburn University Agricultural Experiment Station, Auburn, Alabama, USA. 347 pp.

Neill, W. T. 1948a. The lizards of Georgia. Herpetologica 4:153-158.

Neill, W. T. 1948b. Hibernation of amphibians and reptiles in Richmond County, Georgia. Herpetologica 4:107-114.

Neill, W. T. 1950. Reptiles and amphibians in urban areas of Georgia. Herpetologica 6:113-116.

Oliver, J. 1955. The natural history of North American amphibians and reptiles. D. Van Nostrand, Princeton, New Jersey, USA. 359 pp.

Ragland, I. M., L. C. Wit, and J. C. Sellers. 1981. Temperature acclimation in the lizards *Cnemidophorus sexlineatus* and *Anolis carolinensis*. Comparative Biochemistry and Physiology 70A: 33-36.

Raun, G. G. 1959. Terrestrial and aquatic vertebrates of a moist, relict area in central Texas. Texas Journal of Science 11:158-171.

Richardson, C. J. and J. W. Gibbons. 1993. Pocosins, Carolina bays, and mountain bogs. Pages 257-310 *in* W. H. Martin, S. J. Boyce, and A. C. Echternacht, editors. Biodiversity of the southeastern United States: lowland terrestrial communities. John Wiley and Sons, New York, New York, USA. 502 pp.

Schwartz, A. and R. W. Henderson. 1991. Amphibians and reptiles of the West Indies: descriptions, distributions and natural history. University of Florida Press, Gainesville, Florida, USA. 720 pp.

Secor, S. M. and C. C. Carpenter. 1984. Distribution maps of Oklahoma reptiles. Oklahoma Herpetological Society Special Publication No.3: 1-57.

Sharitz, R. R. and W. J. Mitsch. 1993. Southern floodplain forests. Pages 311-372 *in* W. H. Martin, S. J. Boyce, and A. C. Echternacht, editors. Biodiversity of the southeastern United States: lowland terrestrial communities. John Wiley and Sons, New York, New York, USA. 502 pp.

Shelford, V. E. 1963. The ecology of North America. University of Illinois Press, Urbana, Illinois, USA. 610 pp.

Shochat, D. and H. C. Dessauer. 1981. Comparative immunological study of albumins of *Anolis* lizards of the Caribbean Islands. Comparative Biochemistry and Physiology 68A:67-73.

Sievert, G. and L. Sievert. 1993. A field guide to reptiles of Oklahoma. Oklahoma Department of Wildlife Conservation, Oklahoma City, Oklahoma, USA. 104 pp.

Sinclair, R., W. Hon, and B. Ferguson. 1970. Amphibians and reptiles of Tennessee. Tennessee Game and Fish Commission, Nashville, Tennessee, USA. 29 pp.

Smith, H. M., G. Sinelnik, J. D. Fawcett, and R. E. Jones. 1972. A survey of the chronology of ovulation in anoline lizard genera. Transactions of the Kansas Academy of Science 75:107-120.

Stalter, R. and W. E. Odum. 1993. Maritime communities. Pages 117-163 *in* W. H. Martin, S. J. Boyce, and A. C. Echternacht, editors. Biodiversity of the southeastern United States: lowland terrestrial communities. John Wiley and Sons, New York, New York, USA. 502 pp.

Stout, I. J. and W. R. Marion. 1993. Pine flatwoods and xeric pine forests of the southern (lower) coastal plain. Pages 373-446 *in* W. H. Martin, S. J. Boyce, and A. C. Echternacht, editors. Biodiversity of the southeastern United States: lowland terrestrial communities. John Wiley and Sons, New York, New York, USA. 502 pp.

Taylor, R. J. and H. Laughlin. 1964. Additions to the herpetofauna of Bryan county, Oklahoma. Southwestern Naturalist 9:41-43.

Thompson, E. F., Jr. 1982. A guide to the amphibians, reptiles, and mammals of South Carolina. Published by the author. Printed by The State Printing Company, Columbia, South Carolina, USA. 150 pp.

Tinkle, D. W. 1959. Observations of reptiles and amphibians in a Louisiana swamp. American Midland Naturalist 62:189-205.

Trewartha, G. T. 1968. An introduction to climate. 4th ed. McGraw-Hill Book Company, New York, New York, USA. 408 pp.

Truett, D. 1993. Gene flow among populations of the green anole, *Anolis carolinensis* (Sauria: Polychridae), in the northern extreme of its range: biogeographical implications. M.S. thesis. University of Tennessee, Knoxville, Tennessee, USA. 55 pp.

United States Geological Survey. 1970. The national atlas of the United States of America. Washington, D. C., USA. 417 pp.

Wade, J. K. 1981. A comparative study of reproduction in two populations of the lizard, *Anolis carolinensis.* Ph.D. dissertation. University of Tennessee, Knoxville, Tennessee, USA. 90 pp.

Webb, R. G. 1970. Reptiles of Oklahoma. University of Oklahoma Press, Norman, Oklahoma, USA. 370 pp.

Williams, E. E. 1969. The ecology of colonization as seen in the zoogeography of anoline lizards on small islands. Quarterly Review of Biology 44:345-399.

Wilson, M. A. and A. C. Echternacht. 1987. Geographic variation in the critical thermal minimum of the green anole, *Anolis carolinensis* (Sauria: Iguanidae), along a latitudinal gradient. Comparative Biochemistry and Physiology 87A:757-760.

APPENDIX TO PART 2

Table 2-1. The occurrence of *Anolis carolinensis* in the physiographic provinces of the 11 states in which this species is presently found. Physiographic provinces are primarily those of Fenneman (1931, 1938). Note that this species is present in nearly all of the physiographic provinces in these states. In the more mountainous provinces, *A. carolinensis* is generally found along river or stream edges at lower elevations and not at the higher elevations. An asterisk indicates that it is uncertain as to whether these locations represent introduced or natural populations of *A. carolinensis*.

PHYSIOGRAPHIC PROVINCES	OCCURRENCE OF *ANOLIS CAROLINENSIS*	REFERENCES
Coastal Plain	Yes	Neill (1948a), Brown (1950), Oliver (1955), Gordon (1956), Duellman and Schwartz (1958), Tinkle (1959), Webb (1970), Mount (1975), Martoff et al. (1980), Thompson (1982), Lohoefener and Altig (1983), Secor and Carpenter (1984), Ashton and Ashton (1985), Dixon (1987), Dundee and Rossman (1989), Gibbons and Semlitsch (1991), Sievert and Sievert (1993)
Piedmont	Yes	Neill (1948a), Mount (1975), Martoff et al. (1980), Thompson (1982), Truett (1993)
Blue Ridge	Yes	Jones and Ressler (1927), Cochran (1938), King (1939), Neill (1948a), Johnson (1958), Huheey and Stupka (1972), Truett (1993)
Ridge and Valley	Yes	Jones and Ressler (1927), Neill (1948a), Johnson (1958), Sinclair et al. (1970), Mount (1975), Truett (1993)

Table 2-1 (continued)

PHYSIOGRAPHIC PROVINCES	OCCURRENCE OF *ANOLIS CAROLINENSIS*	REFERENCES
Appalachian Plateau	Yes	Johnson (1958), Sinclair et al. (1970), Mount (1975), Truett (1993), D. MacDonald (personal communication)
Interior Low Plateaus	Yes	Johnson (1958), Sinclair et al. (1970), Truett (1993)
Ouchita	Yes	Dowling (1957), Webb (1970), Secor and Carpenter (1984), Sievert and Sievert (1993)
Ozark Plateaus	No	
Central Lowland	No	
Great Plains	Yes*	Dixon (1987)
Basin and Range	Yes*	G. Sievert (personal communication)

Table 2-2. The occurrence of *Anolis carolinensis* in the potential natural vegetation types defined by Küchler (1964) in the 11 states over this lizard's range. Only a few of the references actually indicated that *A. carolinensis* occurs in a given Küchler vegetation type. Most references simply give a geographic location, while a few others also provided a description of the vegetation from which one can infer a possible Küchler vegetation type. These situations are reflected in the letter codes for occurrence of this species: Y=yes, reported to occur in the Küchler vegetation type, L=likely to occur in the particular Küchler vegetation type since this species is reported in the geographic region associated with the particular Küchler vegetation type and either some description of vegetation is provided or this lizard commonly occurs throughout the region, U=uncertain that this lizard is associated with the predominant natural vegetation type of the region since the reference reports it in the geographic area, but does not describe the vegetation along with the specific distributional account, O=occurs in vegetation other than the predominant Küchler vegetation type of the region, I=not clear as to whether the occurrence possibly represents a recent introduction into the area or a natural population, and N=no known report in the vegetation type was found in the literature which was surveyed.

VEGETATION TYPES	OCCURRENCE OF *ANOLIS CAROLINENSIS*	REFERENCES
EASTERN FORESTS		
Needleleaf Forests		
Southeastern spruce-fir forest (97)	N	
Broadleaf Forests		
Oak-hickory forest (100)	Y	Bryant et al. (1993)
Mixed mesophytic forest (103)	U	Sinclair et al. (1970), Mount (1975)
Appalachian oak forest (104)	Y	Jones and Ressler (1927)
Mangrove (105)	Y	Gilmore and Snedaker (1993)

Table 2-2 (continued)

VEGETATION TYPES	OCCURRENCE OF *ANOLIS CAROLINENSIS*	REFERENCES
EASTERN FORESTS (continued)		
Broadleaf and Needleleaf Forests		
Northern hardwoods (106)	N	
Oak-hickory-pine forest (111)	L	Brown (1950), Mount (1975), Martoff et al. (1980), Dixon (1987), Dundee and Rossman (1989)
Southern mixed forest (112) (includes pine forests such as pine flatwoods)	Y	Stout and Marion (1993),
Southern floodplain forest (113)	Y	Tinkle (1959), Sharitz and Mitsch (1993)
Pocosin (114)	Y	Richardson and Gibbons (1993)
Sand pine scrub (115)	Y	Ashton and Ashton (1985), Greenberg et al. (1994), Stout and Marion (1993)
Subtropical pine forest (116)	Y	Stout and Marion (1993)
CENTRAL AND EASTERN GRASSLANDS		
Grasslands		
Grama-buffalo (65)	N	
Bluestem-grama prairie (69)	N	
Sandsage-bluestem prairie (70)	N	

Table 2-2 (continued)

VEGETATION TYPES	OCCURRENCE OF *ANOLIS CAROLINENSIS*	REFERENCES
CENTRAL AND EASTERN GRASSLANDS (continued)		
Grasslands (continued)		
Shinnery (71)	N	
Sea oats prairie (72)	U	Brown (1950), Dixon (1987)
Northern cordgrass prairie (73)	U	Martoff et al. (1980)
Bluestem prairie (74)	U	Brown (1950), Dixon (1987), Dundee and Rossman (1989)
Blackland prairie (76)	U	Brown (1950), Dixon (1987)
Bluestem-sacahuista prairie (77)	U	Brown (1950), Dixon (1987)
Southern cordgrass prairie (78)	U[a]	Dixon (1987)
	U	Dundee and Rossman (1989)
Palmetto prairie (79)	L	Ashton and Ashton (1985)
CENTRAL AND EASTERN GRASSLANDS		
Grassland and Forest Combinations		
Marl-Everglades (80)	Y	Gunderson and Loftus (1993)
Mosaic of Bluestem prairie/Oak-hickory forest (82)	N	

Table 2-2 (continued)

VEGETATION TYPES	OCCURRENCE OF *ANOLIS CAROLINENSIS*	REFERENCES
CENTRAL AND EASTERN GRASSLANDS (continued)		
Grassland and Forest Combinations (continued)		
Cedar glades (83)	N	
Cross timbers (84)	U	Brown (1950), Dixon (1987)
Mesquite-buffalo grass (85)	N[a]	
Juniper-oak savanna (86)	U[a]	Brown (1950: 20 mi. n. San Antonio) Dixon (1987)
Mesquite-oak savanna (87)	U[a]	Dixon (1987)
Fayette prairie (88)	U	Brown (1950), Dixon (1987)
Blackbelt (89)	U	Mount (1975), Lohoefener and Altig (1983)
Live oak-sea oats (90)	Y	Stalter and Odum (1993)
Cypress savanna (91)	L	Duellman and Schwartz (1958), Dalrymple (1988)
Everglades (92)	Y	Gunderson and Loftus (1993)

Table 2-2 (continued)

VEGETATION TYPES	OCCURRENCE OF *ANOLIS CAROLINENSIS*	REFERENCES
WESTERN SHRUB AND GRASSLAND[b]	U/I[a]	Brown (1950), Dixon (1987)
	O/I	G. Sievert (personal communication)

[a] All of the counties in Texas which are suggested by Dixon (1987) as representing possible introductions have two or more Küchler vegetation types. Therefore, it is difficult to determine the specific vegetation type without the precise location of the distributional account.

[b] The Küchler vegetation types are not given for these categories since few reports of *A. carolinensis* exist from the region and it is unclear as to whether or not the populations are natural or introductions.

Table 2-3. Ecoregion classification of Bailey (1976, 1980) and the occurrence of *Anolis carolinensis*. None of the references specifically mentions the presence of this lizard in any ecoregion section, but simply provide a geographic area and sometimes vegetation descriptions. Thus, the occurrence of *A. carolinensis* in a particular ecoregion section is inferred from the distributional accounts and any descriptions of vegetation. The vegetation within a section was classified by Bailey (1976, 1980) according to one or more of the natural potential vegetation types of Küchler (1964), but variation within a section does occur. Populations of *A. carolinensis* occur in a variety of vegetational communities in most ecoregion sections, but probably occur in vegetation other than the predominant Küchler type in central and Texas and parts of the Rio Grande Valley. Note that this table only lists those ecoregion sections in which *A. carolinensis* is reported to occur.

DOMAIN DIVISION PROVINCE SECTION	REFERENCES
2000 Humid Temperate	
2200 Hot Continental	
2210 Eastern Deciduous Forest	
2211 Mixed Mesophytic Forest	Sinclair (1970), Mount (1975)
2214 Appalachian Oak Forest	Jones and Ressler (1927), Cochran (1938), King (1939), Johnson (1958), Sinclair et al. (1970), Huheey and Stupka (1972), Martoff et al. (1980)
2215 Oak-Hickory Forest	Sinclair et al. (1970), Lohoefener and Altig (1983)

Table 2-3. (continued)

DOMAIN DIVISION PROVINCE SECTION	REFERENCES
2300 Subtropical	
2310 Outer Coastal Plain Forest	
2311 Beech-Sweetgum-Magnolia-Pine-Oak Forest	Brown (1950), Mount (1975), Lohoefener and Altig (1983), Ashton and Ashton (1985), Gibbons and Semlitsch (1991)
2312 Southern Floodplain Forest	Tinkle (1959), Lohoefener and Altig (1983), Dundee and Rossman (1989), Sharitz and Mitsch (1993)
2320 Southeastern Mixed Forest	Brown (1950), Dowling (1957), Webb (1970), Secor and Carpenter (1984), Dixon (1987), Dundee and Rossman (1989)
2500 Prairie	
2510 Prairie Parkland	
2512 Oak + Bluestem Parkland	Brown (1950), Secor and Carpenter (1984), Dixon (1987)
2520 Prairie Brushland	
2522 Juniper-Oak-Mesquite	Brown (1950), Dixon (1987)
2523 Mesquite-Acacia	Brown (1950), Dixon (1987)

Table 2-3. (continued)

DOMAIN DIVISION PROVINCE SECTION	REFERENCES
3000 Dry	
3200 Desert	
3210 ChihuahuanDesert	
3212 Tarbush-Creosote Bush	G. Sievert (personal communication)
4000 Humid Tropical	
4100 Savanna	
4110 Everglades	Duellman and Schwartz (1958), Ashton and Ashton (1985)

Table 2-4. Specific habitats in which *Anolis carolinensis* has been reported to occur. The entries in this table are by no means an absolutely complete list of habitats, but rather represent a broad survey of reports of habitats for this species.

HABITAT	LOCATION	REFERENCE
Sabal palmetto	Florida	Lee (1969)
Mesophytic hammocks	southern Florida	Duellman and Schwartz (1958)
	Florida	Ashton and Ashton (1985)
Hydric hammocks	Florida	Ashton and Ashton (1985)
Tropical hammocks	Florida	Ashton and Ashton (1985)
	Everglades National Park, Florida	Dalrymple (1988)
Xeric oak hammocks	Florida	Ashton and Ashton (1985)
Postoak-blackjack community	central Texas	Raun (1959)
Oak-pine-heath community	eastern Tennessee	King (1939), Johnson (1958)
Open oak-pine community	eastern Tennessee	Johnson (1958)
Mature oak-pine community	South Carolina	Jenssen et al. (1995a), Jenssen et al. (1996)

Table 2-4. (continued)

HABITAT	LOCATION	REFERENCE
Pine flatwoods	Florida	Ashton and Ashton (1985)
Pine forest	southern Florida	Duellman and Schwartz (1958)
	Everglades National Park, Florida	Dalrymple (1988)
Sand pine-rosemary scrub	Florida	Ashton and Ashton (1985)
Sand-pine scrub: mature forests and various managed stands	Florida	Greenberg et al. (1994)
Longleaf pine-turkey oak	Florida	Ashton and Ashton (1985)
Second-growth pine forests	western boundary of Great Smoky Mountains National Park, Tennessee	Huheey and Stupka (1972)
Mangrove swamps	Florida	Ashton and Ashton (1985)
Reeds and grasses	southern Florida	Duellman and Schwartz (1958)

Table 2-4. (continued)

HABITAT	LOCATION	REFERENCE
Cypress swamps and domes	Florida	Ashton and Ashton (1985)
Gum swamps and river swamps	Florida	Ashton and Ashton (1985)
Cypress-gum swamps: along dry ridges with willow, maple, and groundsel trees	southern Louisiana	Tinkle (1959)
Carolina bays	South Carolina	Gibbons and Semlitsch (1991)
Peat bogs	central Texas	Raun (1959)
Trees and shrubs in floodplains	eastern Tennessee	Johnson (1958)
	central Texas	Raun (1959)
	Oklahoma	Taylor and Laughlin (1964), Webb (1970)
	South Carolina	Jenssen et al. (1995b)
Sea grape	southern Florida	King (1966)
Barrier islands	Shackleford Banks, North Carolina	Engels (1952)
	coastal Mississippi	Lohoefener and Altig (1983)

Table 2-4. (continued)

HABITAT	LOCATION	REFERENCE
Dredge spoil islands	near Cape Canaveral, Florida	T. Campbell (personal communication)
Gardens	southern Florida Florida	Duellman and Schwartz (1958) Ashton and Ashton (1985)
Farmlands, fields, and disturbed areas	Florida Louisiana	Ashton and Ashton (1985) Gordon (1956)
Cemeteries with numerous vines	Louisiana	D. MacDonald (personal communication)
Around human homes	southern Florida Georgia Alabama Texas	Duellman and Schwartz (1958) Neill (1950) Mount (1975) Michael (1972)
Human habitations, golf courses, and trash piles	Florida	Ashton and Ashton (1985)
Walls and fences along beaches	Mississippi	Corrington (1927)

PART 3 : THE INFORMATIONAL APPROACH TO DATA ANALYSIS : AN ALTERNATIVE TO STATISTICAL HYPOTHESIS-TESTING PROCEDURES

"May I repeat: statistics is a tool, for us practitioners, and we should use whatever tool is most appropriate for getting at the question we want to answer."

C. Toft (1990:359)

INTRODUCTION

Statistics and biology

Statistics can be viewed as a discipline which provides principles and methods used for 1) designing the collection of data, 2) summarizing, analyzing, and interpreting sample data, and then 3) answering questions and/or drawing generalities about the phenomenon being studied (Johnson and Bhattacharyya 1996:3-14). Statistics can be an important tool for researchers in biology because some uncertainty always exists in any attempt to detect and describe patterns in the data, answer specific questions, and/or draw conclusions or inferences.

Barnett (1973) described and reviewed several approaches to statistics, including a considerable discussion of the classical (also called the frequentist or hypothesis-testing) approach and a brief mention of the informational (or information) approach. The informational approach was still in its infancy at the time that Barnett's (1973) comparative work was published. Today, many publications can be found on the theory, methodology, and applications of the informational approach (e.g., see Bozdogan 1994a, b, c and the references therein).

The classical approach practiced today emphasizes the estimation of parameters and the testing of a hypothesis regarding the estimated parameters for understanding data and making inferences. The word

"hypothesis" is used throughout this chapter to mean a *statistical* hypothesis, which is not necessarily the equivalent of a biological hypothesis. The informational approach views data analysis as methods for selecting the best model(s) to fit the data at hand. Models are ranked and selected by using a numerical criterion, based on mathematical likelihood, that is calculated for each model under consideration.

In this chapter, comparisons are made between only the informational and classical approaches because 1) the informational approach is used in this dissertation as an alternative to the classical approach and 2) the classical approach is most familiar to and most often used by biologists.

Although applications of the informational approach have been increasing over the past 20 years in many scientific fields, most biologists rely almost exclusively on the classical approach for data analysis. At least four possible interconnected reasons can possibly explain this reliance on the classical approach. First, the classical approach is largely (and sometimes solely) the methodology in which most biologists have been formally educated. Most statistics courses taught in statistics departments and/or biology degree programs in the United States focus primarily on the classical approach, despite the fact that other approaches are useful for data analysis.

Second, the classical approach has a longer history and association with biology than other approaches. The classical approach is largely rooted in the work of R. A. Fisher, J. Neyman, and E. S. Pearson during the first half of 20th century. Although these statisticians disagreed on a number of issues, the concepts and methods used in the classical approach represent a

combination of the principles they established (Barnett 1973). The long association between the classical approach and biology is evident from the early applications of the methods of Fisher to agriculture, genetics, and evolution, and of Neyman and Pearson to various biological problems. As a result of the classical approach's long history, this approach is the one adopted by the majority of college statistics textbooks and most computer software packages for data analysis. Also, the long association between biology and the classical approach probably accounts, in part, for this approach being utilized more frequently than other approaches in most biology lab manuals and most statistically-oriented publications in biology.

Third, most publications in biological journals which either discuss the merits of certain statistical methods or inform biologists about particular statistical techniques have done so mainly within the context of the classical approach. For example, many such publications appearing in ecological journals during the present decade have been in the context of hypothesis-testing procedures (see, e.g., James and McCulloch 1990, Petranka 1990, Seaman and Jaeger 1990, Simberloff 1990, Toft 1990, Dutilleul 1993, Legendre 1993, Potvin and Roff 1993, Scheiner and Gurevitch 1993, Shaw and Mitchell-Olds 1993, Trexler and Travis 1993, Johnson 1995, Smith 1995). Such works have performed greatly needed services for ecologists: that of interpreting much of the statistical literature and/or providing statistical guidelines. However, some researchers in ecology have demonstrated the usefulness of the informational approach over the classical approach for analyzing certain data (see Burnham and Anderson 1992, Lebreton et al. 1992, Burnham et al. 1995a, b).

Lastly, the classical approach has predominated over other approaches, in statistics as a whole, partly because of the viewpoints of influential statisticians, most notably R. A. Fisher, and the path that these views directed the field of statistics (Akaike 1994). This idea of Fisher's influence on the discipline of statistics centers around his view and use of mathematical likelihood. Basically, Fisher's restricted view of likelihood limited the use of likelihood to many of the procedures and tests practiced in the classical approach (Akaike 1994). However, the view of likelihood put forth by Akaike (1973) and others expands statistical analysis into the realm of the informational approach. These ideas and views of likelihood will be discussed further in a later section.

Why another statistical approach?

The use of statistics in the biological sciences has increased rapidly over the past 50 - 60 years (Sokal and Rohlf 1995:5-6). With such an increase, most biologists have probably felt overwhelmed at times by the many statistical methodologies and controversies that have appeared in the literature. So why do biologists need yet another approach, such as the informational approach, for their statistical "toolboxes"?

First, because the nature of science *requires* that scientists look for better ways to investigate phenomena and to answer questions. Scientists should not become complacent by thinking that the traditional methods of inquiry are the only ones that are of any use, particularly when potentially more useful methods are developed.

Second, criticisms and concerns about the classical approach have appeared in the literature of many disciplines (e.g., statistics, psychology,

sociology, education, and biology), mainly over aspects of hypothesis-testing procedures. Concerns have been expressed about certain philosophical underpinnings and statistical methods of hypothesis-testing procedures, as well as the overemphasis of and/or overreliance on these procedures for scientific inquiry (see Table 3-1 for examples of concerns). Some of these concerns will be discussed later when comparisons between the classical and informational approaches are made. Alternative methods which have been suggested for either complete replacement of hypothesis-testing procedures or accompanied use include : graphical examination (Deming 1975), estimation of parameters and standard errors (Salsburg 1985, Jones and Matloff 1986, Yoccoz 1991), estimation of the standard error of each mean for use in multiple comparisons of means (Perry 1986), estimation of confidence intervals (Jones and Matloff 1986, Matloff 1991, Yoccoz 1991), use of model selection and model checking techniques (Stewart-Oaten 1995), and use of the informational approach for model selection (Akaike 1973, Sakamoto et al. 1986, Bozdogan 1987, Burnham and Anderson 1992).

Last, the informational approach has certain advantages over hypothesis-testing procedures, both statistical and philosophical. For example, the informational approach's modeling viewpoint and use of model-selection criteria fits nicely with the goals of analyzing many types of biological data where determining patterns and/or finding a good set of descriptor or predictor variables are important. Biologists might find data analysis to be more straightforward using the informational approach than using hypothesis-testing procedures because 1) models are ranked and

selected based on their numerical criterion values and 2) alpha values, P-values, and statistical probability tables are not needed. In addition, new practitioners of statistics might be able to better grasp concepts and methods of data analysis by using the informational approach (by itself or combined with another approach) than by relying heavily on hypothesis-testing procedures. A more thorough discussion of advantages of the informational approach will be discussed later.

The main purpose of this chapter is to provide both a basic overview of the informational approach and general comparisons of it with the classical approach in order to show that the informational approach is a viable statistical alternative to hypothesis-testing procedures. Researchers can evaluate the comparisons being made, read the literature cited herein, and then make free choices about which methods of data analysis to use rather than become restricted by tradition or dogma. Some researchers, as mentioned previously, have broadened the statistical analysis of ecological data by using the informational approach (see Burnham and Anderson 1992, Lebreton et al. 1992, Burnham et al. 1995a, b). It is in this spirit of broadening the discussion about methods of data analysis in biology and giving biologists more choices that this chapter (and much of this dissertation) is presented.

OVERVIEW OF BOTH THE CLASSICAL AND INFORMATIONAL APPROACHES TO STATISTICAL ANALYSIS

The classical approach

The classical approach has two main aims : 1) the estimation of parameter values and 2) the testing of a statistical hypothesis, both usually performed for a given model (Barnett 1973). Parameter estimation involves using sample data in the point estimation of a particular feature (parameter) of the population and/or the estimation of a region or confidence interval within which the parameter is expected to reside (Barnett 1973). The objective of hypothesis testing is to decide whether a supposition about some feature or parameter of a population is well supported by the sample data (Johnson and Bhattacharyya 1996:327).

The bulk of most textbooks on classical statistics is composed of hypothesis-testing methods and the most common 'tools' in many ecologists' statistical tool boxes consist of hypothesis-testing procedures. Thus, discussions in this chapter of the criticisms of the classical approach and comparisons of this approach to the informational approach will often center around such hypothesis-testing procedures. The basic steps of hypothesis testing within the classical approach can be summarized as follows (see Johnson and Bhattacharyya 1996:327-335, Sokal and Rohlf 1995:157-169):

1. Define the null hypothesis and the alternative hypothesis.
2. Choose a test statistic to evaluate the null hypothesis based on its appropriateness for the data and on the expected distribution of the data.

3. Specify an alpha level (probability of making a type I error) in order to define the rejection region. Examine the power of the test to protect against a type II error given the particular alpha level and sample size.

4. Calculate the test statistic for the given data.

5. Make a decision based on the alpha level and the defined rejection region : either reject the null hypothesis if the calculated test statistic is greater than the critical value of the test statistic or do not reject the null hypothesis if the test statistic value is less than the critical value. Calculate the significance level (*P*-value) for the test.

Null and alternative hypotheses are usually stated in terms of the parameters being estimated (such as means, variances, regression coefficients, etc.). A researcher states his or her idea about the true value of the parameter in the form of the alternative hypothesis (H_1). The negation of this claim about the parameter is the null hypothesis (H_0). The choice of a specific test procedure is then based on the type of data, the expected distribution underlying the data, and the hypothesis to be tested. Many different test procedures exist and each one has specific assumptions about the data and its distribution.

Specifying an alpha level is up to the discretion of the researcher, although in many scientific fields 0.05 is the standard, albeit arbitrary, value accepted. The alpha value defines the region of the distribution of parameter values for which values of the estimated parameter would lead to the rejection of the null hypothesis in favor of the alternative one. The alpha value represents the probability of rejecting H_0 when H_0 is indeed true (i.e., probability of making a type I error).

The result of a test procedure leads to a decision to either reject H_0 or to accept H_0. Acceptance of the null hypothesis is usually taken to mean that the evidence was not strong enough to discredit H_0, rather than to suggest that H_0 is actually true (although researchers often incorrectly suggest that H_0 is "true"). The decision regarding H_0 is often used as a basis for inference and is then interpreted in the context of the research question.

The significance level (*P*-value) of the test indicates how likely (or unlikely) is the particular decision regarding H_0 given both the assumed probability distribution and the idea that many samples could be taken. Thus, if the *P*-value is 0.04, then 96 of 100 random samples would likely produce estimated parameter values leading to the rejection of the null hypothesis. A *P*-value of 0.04 also means that a Type I error likely occurs in 1 out of 25 samples. Another way to view a *P*-value is as a measure of the strength of the rejection of H_0. The smaller the *P*-value the stronger the evidence supposedly is against H_0. However, the *P*-value is <u>not</u> a measure of the probability of the truth of H_0.

The overview given above on the classical approach is generally that which can be found in most introductory statistics books. This overview is indeed very brief because it is assumed that the reader already has a working knowledge of the classical approach and hypothesis-testing procedures. Although the basic steps of hypothesis testing are presented here in a somewhat factual manner, some criticisms of hypothesis-testing procedures will be discussed at various times in this chapter.

The informational approach

The informational approach utilizes a statistical modeling framework and has a different view of statistical likelihood than that found in the classical approach. The informational approach uses the likelihood term as a component of a criterion used in the model selection process. Thus, it is first necessary to discuss statistical models, model selection, and the different views of likelihood.

A "model", in ordinary English-usage, is a replica of an object or a description of some object or occurrence. The model can describe something using simply words, a mathematical formulation, or a statistical representation. In statistics, a model is "... something whose structure, and hence behaviour, corresponds in some sense to that of a particular reality or phenomenon." and the 'structure' of a model contains components of chance (Gilchrist 1984:14). A statistical model can be seen as a probability distribution (Sakamoto et al. 1986) or a description or an expression of the important features of the data in terms probabilities (Bozdogan, personal communication). The term "model" is henceforth used to mean a statistical model. A regression model, where independent variables are used to describe the variation in the dependent variable, is a well known example of a statistical model. Because data have some error associated with them, any model based on the data has some uncertainty also.

Given these definitions of a model and the fact that some amount of uncertainty exists is any model, the aim of statistical modeling is to build models to a data set and determine which model (or models) best describes or explains the phenomenon underlying the data. Even in the biological

literature, the importance of selecting the best model is echoed by Burnham and Anderson (1992:16):

> "Future data analysis through model building and selection should begin with an array of models that seem biologically reasonable. Then, the central problem of data analysis is selection of an appropriate model as the basis of inference."

Selection of the most appropriate model to describe the data is not a trivial problem in statistics; much time and effort over the years has been applied to this problem. The need for proper model evaluation and selection has lead to the development of model selection criteria, including those used in the informational approach.

One important distinction between the classical and informational approaches is the way in which likelihood is viewed and used. According to Akaike (1994), Fisher's great accomplishment was the development of the concept of mathematical likelihood, but that his main view of and *use* of likelihood was *restricted* to the estimation of parameters. Although Fisher may have been aware of other possible uses of likelihood, he restricted his use of likelihood in this manner and thus restricted the potential use of the log likelihood term as a general criterion of a model's fit to the data. This latter use of likelihood as a criterion is the view taken by those using the informational approach. Given this restricted view and use of likelihood by Fisher, Akaike (1994:29) suggests the following scenario in the development of modern statistics :

> "A framework was then established to view the test of significance as the basic procedure for the solution of the problems of specification and restrict the estimation to the parameter of a given model. Thus the test and estimation formed a paradigm to make statistics into what was called a normal science by Kuhn (1970).

> However, it seems that the use of test procedures advocated by this paradigm eventually produced a very restricted image of statistics in applications which was conditioned by the availability of proper test procedures."

In other words, Fisher's restricted view of likelihood led the field of statistics down a road which limited the use of likelihood to estimating parameters values of a *given model* and testing the statistical significance of those parameter values. However, for over 25 years Akaike and others have expanded Fisher's concept of likelihood and linked it with the probabalistic concept of entropy (or information) to develop the informational approach to statistics. This approach uses numerical, information-based criteria in order to compare alternative models and to select the model(s) which best fits (fit) the data at hand, rather than limiting the analysis to estimating parameters for only one or a few models.

The informational approach began to develop with the work of Kullback and others (see Barnett 1973:263-66 and Sakamoto et al. 1986:37-55). This approach reached new heights when Akaike (1973, 1974) proposed an information-based criterion (now called Akaike's Information Criterion or AIC) as a numerical criterion for evaluating two or more statistical models in a model selection problem.

The objective in any problem of model selection is to evaluate the closeness or goodness-of-fit of each model to the data. If the "true" model is known, then the Kullback-Leibler (K-L) information quantity (negentropy) can be used as a measure of the closeness of a proposed model to the true model (Akaike 1973, Sakamoto et al. 1986, Bozdogan 1987). In reality, however, the true model is not known, only the sample data are at hand, and an estimate of the K-L information quantity is needed. Akaike

(1973) linked together likelihood theory and information theory by showing that the mean log likelihood is an estimator of the K-L information quantity. He thus demonstrated that the log likelihood could be used as : 1) a measure of the fit between a model and the data and 2) part of a numerical criterion for model selection. The log likelihood could be calculated using maximum likelihood estimation (MLE) procedures at the estimated parameter values for a given model.

However, selection of the best model(s) cannot be based solely on the log likelihood term. The Principle of Parsimony states that the best explanation or description is the simplest one which is capable of capturing the essential aspects of the phenomenon being studied. Statistically, adherence to parsimony would help reduce both the risk of overfitting a model (Sakamoto et al. 1986, Bozdogan 1987, 1988a, b, 1990) and the problems associated with overfitting (Burnham and Anderson 1992, Bozdogan, forthcoming book). Thus, the analyst should select the simplest statistical model that sufficiently explains or describes the phenomenon.

AIC and all other informational criteria for model selection take into consideration *both* the fit of the model and the Principle of Parsimony by utilizing the basic form : criterion = lack-of-fit term + penalty term. The lack-of-fit term is a measure of the discrepancy between the model and the data and is estimated using MLE procedures. The smaller the lack-of-fit the better the given model is for describing the data; the larger this term becomes then the poorer the fit. If more and more parameters are added to the model the lack-of-fit decreases. However, selection of the best model cannot be accomplished by the lack-of-fit term alone because the problem of

overfitting would not be addressed, hence a penalty term is incorporated into the criterion.

AIC is specifically defined as:

$$AIC = -2\ln L(\hat{\theta}_k) + 2k\,, \tag{3.1}$$

where 'ln $L(\hat{\theta}_k)$' is the maximum loglikelihood value when MLE methods are used to estimate the parameter values for the model, ln is the natural logarithm, and k is the total number of estimated parameters in the model (Akaike 1973, Sakamoto et al. 1986). The first term in AIC is a measure of the lack-of-fit of a given model to the data and is often part of the standard output of many statistical software packages. The $2k$ term of AIC is the penalty (or complexity) term of the model; the more parameters in the model, the larger the penalty term becomes. The penalty term helps address the Principle of Parsimony. This second term also accounts for the bias associated with using MLE procedures to estimate the fit between the model and data (Sakamoto et al. 1986, Bozdogan 1987).

In model selection problems AIC is calculated for various alternative models which are considered to be alternative representations of the "true" underlying structure of the data or the population from which the data were sampled (see, e.g., analyses in Sakamoto et al. 1986). The model with the smallest criterion is considered the "best" model to describe the data at hand. The practical importance of AIC is that the evaluation of alternative models takes into account both the goodness of fit and the number of estimated parameters in each model, thereby directly addressing concerns

about under- and overfitting the data (Sakamoto et al. 1986, Bozdogan 1987, Burnham and Anderson 1992). These points will be discussed later in more detail.

The use of numerical criteria in problems of model selection were used in statistics prior to AIC. For example, in multiple regression problems the adjusted R^2 and Mallows' C_p have been used to measure the quality of fit of a model and to compare competing models. Akaike's development of AIC simply extended the idea of a numerical criterion into a new realm by combining aspects of information theory and statistics and by viewing likelihood in a different way from that of Fisher and the classical approach.

After Akaike's ground-breaking work other informational criteria have been developed including: consistent AIC (CAIC) and CAIC with Fisher information (CAICF) (Bozdogan 1987), the information-theoretic measure of complexity (ICOMP) (Bozdogan 1988a, b), and Bayesian modifications of AIC and other Bayesian criteria such as Kashyap's Criterion (see Bozdogan 1990). These criteria generally follow the basic ideas put forth by Akaike's development of AIC, but have different second terms than AIC.

Bozdogan (1988a, b) developed the model selection criterion called ICOMP with a penalty term defined in terms of the interdependencies among model components such as among parameter estimates and among residuals (Bozdogan 1988a, 1990) rather than simply as a multiple of the number of estimated parameters. Two approaches can be taken with ICOMP to calculate the penalty term. The first approach uses an information-based measure of complexity of the estimated covariance

matrix of the parameter estimates as the penalty term (Bozdogan 1990). The second approach uses a measure of the complexity of the inverse-Fisher information matrix (IFIM) (Bozdogan 1990). In both cases complexity is based on the complexity measure of van Emden (1971, cited in Bozdogan 1988a, 1990). For any symmetric matrix, M, with b number of rows and b number of columns complexity is :

$$C_1[M] = (1/2)\{b\ln[\mathrm{tr}(M)/b] - \ln[\det(M)]\}, \tag{3.2}$$

where: ln = the natural logarithm,

tr = the trace of the matrix,

det = the determinant of the matrix.

In the first approach ICOMP is defined as (following Bozdogan 1990, but using 2 times the complexity as indicated by Bozdogan and Haughton 1998) :

$$\mathrm{ICOMP} = -2\ln L(\hat{\theta}_k) + 2\{C_1[\widehat{\mathrm{Cov}}(\theta)] + C_1[\widehat{\mathrm{Cov}}(\varepsilon)]\}. \tag{3.3}$$

The first term is the same as the first term in AIC. C_1 denotes the measure of complexity of a covariance matrix defined by van Embden (1971, cited in Bozdogan 1988a, 1990). $\widehat{\mathrm{Cov}}(\theta)$ is the estimated covariance matrix of the estimated parameter values and $\widehat{\mathrm{Cov}}(\varepsilon)$ is the estimated covariance matrix of the estimated residual terms, ε. These matrices are obtained using maximum likelihood estimation procedures. The covariance matrices contain information about variances and covariances which can be used to

measure interdependencies among terms (recall, for example, that a correlation matrix can be obtained from a covariance matrix). A more detailed formula for ICOMP in equation (3.3) is:

$$\begin{aligned} \text{ICOMP} = -2\ln L(\hat{\theta}_k) + 2\{(k/2)\ln[\text{tr}(\widehat{\text{Cov}(\theta)})/k] - \\ (1/2)\ln[\det(\widehat{\text{Cov}(\theta)})] + (n/2)\ln[\text{tr}(\widehat{\text{Cov}(\varepsilon)})/n] - \\ (1/2)\ln[\det(\widehat{\text{Cov}(\varepsilon)})]\} . \end{aligned} \tag{3.4}$$

Sample size is denoted by n and all other notation follows that given in equations (3.1) through (3.3).

In the second approach to ICOMP, the penalty term is calculated as the complexity of the estimated inverse-Fisher information matrix over the entire parameter space. IFIM measures the accuracy of the model, as well as the complexity of the parameter estimates and provides a way to see how different covariance structures in different models might influence the accuracy of the parameter estimates (Bozdogan 1990). This approach also models the random error terms, ε, as independent and/or dependent. The formula for ICOMP-IFIM (following Bozdogan 1990, but again using 2 times the complexity according to Bozdogan and Haughton 1998) is:

$$\begin{aligned} \text{ICOMP-IFIM} = -2\ln L(\hat{\theta}_k) + 2\{(r/2)\ln[\text{tr}(\hat{F}^{-1})/r] - \\ (1/2)\ln[\det(\hat{F}^{-1})]\} , \end{aligned} \tag{3.5}$$

where: $\hat{F}^{-1}$ = the estimated inverse-Fisher information matrix of the parameter estimates and

r = rank or dimension of $\hat{F}^{-1}$.

The second term here represents two times the complexity of the inverse Fisher information matrix (Bozdogan 1990).

Like AIC, both approaches to ICOMP are derived from information theory, adhere to the Principle of Parsimony, and choose the best overall model based on the minimum criterion value. However, unlike AIC, ICOMP incorporates the interdependencies of parameter estimates and residual terms into the criterion (Bozdogan 1990). This consideration of the interdependency of model components is of practical importance in model selection in multivariate data sets where either parameter estimates or error terms may be correlated to some degree. ICOMP essentially considers the "better" models to be those having a small lack of fit to the data and having lesser amounts of interdependency in their structure (Bozdogan 1990).

Interested readers can find more thorough statistical coverages of AIC in Akaike (1973, 1974), Sakamoto et al. (1986), and Bozdogan (1987) and of the approaches to ICOMP in Bozdogan (1988a, b, 1990). Many examples of applications of the informational approach to statistical data analysis can be found in Sakamoto et al. (1986) and Bozdogan (1994a, b, c).

COMPARISONS OF THE CLASSICAL AND INFORMATIONAL APPROACHES TO STATISTICAL ANALYSIS

The classical and informational approaches differ with respect to certain philosophical and statistical points. One difference, as mentioned previously, was Fisher's view of likelihood which led to the restricted use of likelihood for testing statistical hypotheses about parameters of a given

model (Akaike 1994). Thus, hypothesis-testing has come to play a major role in data analysis and inference in the classical approach. The informational approach, however, takes a modeling viewpoint and uses the log likelihood as the basis for developing numerical criteria to distinguish how well various competing models fit a data set. Thus, this approach considers data analysis, in part, to consist of optimizing a criterion for selecting the best model(s) to fit the data at hand.

An example will illustrate this difference in viewpoint between these two approaches. Consider the analysis of variance (ANOVA) commonly used by biologists. Given two treatment groups, A and B, and one control group, C, a researcher wants to determine whether differences exist between the control and treatments and between the two treatments themselves. Using the classical approach, the null and alternative hypotheses could be stated as: H_0: $\mu A = \mu B = \mu C$ and H_1: $\mu A \neq \mu B \neq \mu C$. A test statistic (F-test) would be calculated to test the null hypothesis. Regardless of the specific alternative hypothesis, the analysis here is always a test between only two competing hypotheses. If the null is rejected, then further hypothesis tests (i.e., post-hoc test procedures, multiple comparisons) are needed to determine which means are different.

The informational approach handles the ANOVA problem as one of selecting the best model supported by the data. If all possible outcomes were biologically reasonable then the researcher would have the following models to evaluate:

Model Number	Model	Number of Estimated Means
1	$\mu A = \mu B = \mu C$	1
2	$(\mu A = \mu B) \neq \mu C$	2
3	$(\mu A = \mu C) \neq \mu B$	2
4	$(\mu B = \mu C) \neq \mu A$	2
5	$\mu A \neq \mu B \neq \mu C$	3

Note that Model 1 corresponds to the null hypothesis of the classical approach and is the simplest model because all means are equal and only one mean has to be estimated from the data. The model with the lowest information criterion value, such as AIC, would be the model that best fits the data. Note that the estimation of parameters and the evaluation and selection of the best-fitting model are all done as one process in this analysis. This is unlike the classical approach where more than one stage of hypothesis-testing procedures must be performed.

Many statisticians and researchers realize that analysis of complex data not only can be viewed as a problem of model selection rather than of strict hypothesis-testing, but actually *requires* a modeling approach. One reason for this particular viewpoint regarding complex data is because statistical hypotheses are either difficult to state or irrelevant to the biological questions being asked. Even if one could state a statistical hypothesis to be tested, the nature of the complex data and the goals of the analysis do not provide simple direct links between statistical hypotheses and biological hypotheses.

Closely linked to the reason stated above is that the analyst's goals in analyzing complex data are to uncover patterns or relationships and to understand and simplify the complexity in the data rather than to test

statistical and/or biological hypotheses. Such a goal has been discussed at length with respect to analysis of ecological and evolutionary data (e.g., Quinn and Dunham 1983), but Toft (1990:359) makes the point clearly:

> "In fact, many practitioners, including those in fields other than ecology, are using statistical methods primarily to understand "information" in the data, rather than primarily as strict hypothesis-testing. Many fields examining complex phenomena, like ecological processes, are turning to multivariate procedures simply to detect patterns in the data; no hypotheses are tested (i.e., these have been called "hypothesis-free" procedures)."

Toft (1990) did not mention the informational approach, but it is easy to see that the informational approach fits nicely into what she called "hypothesis-free" procedures (regardless of whether or not this refers to statistical or biological hypotheses).

As an example, suppose a researcher is interested in examining patterns of microhabitat use for nesting sites by three species of passerines over several seasons in a given deciduous forest in the Great Smoky Mountains. How can this be stated in terms of a single, statistical null hypothesis when several or more variables are to be measured? It cannot. If only one microhabitat variable, say nest height, was examined, then the biological null hypothesis and the statistical null hypothesis could be identically stated as: H_0: mean height for A = mean height for B = mean height for C. For the multivariate case, however, the data might be analyzed by discriminant analysis to see how the species separate out in multivariate space. Separation of the species in multivariate space would help the analyst understand and quantify any patterns in the use of nesting microhabitats among the species.

Techniques often used in the modeling of complex biological data include multiple regression, multi-way contingency tables (log linear models), discriminant analysis, principal component analysis, and factor analysis. The main goal of using these techniques is to *find the best models* which uncover patterns or relationships, simplify complex data, and/or serve as the foundation for inference. Over the later part of this century *selection* of the best model(s) to fit the data when using such techniques has been conducted with hypothesis-testing procedures, but this usage is being questioned more frequently (see, e.g., Burnham and Anderson 1992, Lebreton et al. 1992). Use of the informational approach for model selection, however, has certain advantages over hypothesis-testing procedures, which should be of interest to biologists. These advantages center around the following issues: 1) problems of overfitting and underfitting the data, 2) problems concerning alpha levels and *P*-values, and 3) comparisons of nested versus non-nested models.

Any approach to model selection must consider the potential problems of overfitting and underfitting the data. Overfitting is when more parameters are included in the model than are needed to adequately describe or explain the essential attributes of the data. These extra parameters do not improve the fit of the model to the data. When overfitting occurs the model possesses high variances associated with parameter estimates (see, e.g., Myers 1986, Burnham and Anderson 1992). Perfect fit could be obtained to the data by having as many parameters as observations, but this would only result in extremely complex models with excessive variances. It is undesirable to have either a perfectly fit model or

an overfit model because such models become too specific and lose predictive power (ability to be used for prediction with other data sets).

Underfitting occurs when too few parameters are included in the model to capture the essential information in the data. This causes variances to become too small and squared bias to become large (see, e.g., Myers 1986, Burnham and Anderson 1992:Fig. 1). Bias is the expected value of a parameter estimate minus the true value of the parameter. Squared bias is simply the square of this difference. Model bias can be thought of as the distance between the fitted model and the true model. Underfitting produces a larger distance between the fitted model and the true model and unreasonably small variances and parameter estimates. As more parameters are fitted to the data the model bias decreases, but variance becomes larger (see Burnham and Anderson 1992:Fig. 1). Ideally, models should be chosen which have small variances and small bias.

The informational approach directly addresses this need to balance between overfitting and underfitting because the actual number of parameters in the model (model size) or an estimate of model complexity are incorporated into the model-selection criterion. However, hypothesis-testing procedures only directly compare the goodness-of-fit of one model to that of another model. An analyst has to perform several or more hypothesis tests and make multiple decisions based on those tests in order to compare several competing models of different sizes and begin to address overfitting and underfitting. Some researchers simply rely on automatic software programs, such as stepwise algorithms, to perform model selection, but such procedures do not guarantee that the best model

will be selected or that the problem of over- and underfitting will be adequately addressed.

This requirement of having to perform multiple test procedures when conducting multiple tests with the classical approach leads to the issue of lack of control of an overall error rate (alpha level). In the case of ANOVA the researcher can control the overall alpha level of the post-hoc procedures. However, before one performs any post-hoc tests the actual ANOVA is conducted to determine if any differences in means exist. If the ANOVA has an alpha level of 0.05 and the post-hoc alpha level is controlled to 0.05, then what is the overall error rate for drawing inferences from the entire analysis? Is it 0.05 plus 0.05 or is it somehow simply 0.05? With the informational approach, only one analysis need be performed with an ANOVA in order to determine which group means differ, if any (refer to the earlier ANOVA example).

Given the way in which many classical methods to multiple decisions and modeling are conducted (see, e.g., Bishop et al. 1975:155-168, Fienberg 1980:56-80, McCullagh and Nelder 1989:1-5), one can see that this problem of the unknown, overall alpha level is not just restricted to ANOVA, but occurs with many applications of hypothesis testing. Use of the informational approach (or any methodology which uses a criterion to rank models) appears to address this problem because one overall analysis can be performed in order to select the best model for the data (e.g., see Sakamoto et al. 1986). Lebreton et al. (1992) used AIC to model survival in marked populations and addressed the overall alpha issue when they stated (p.111):

> "We recommend use of Akaike's Information Criterion here as a way to assist in selecting a basic model from the global model. Then specific biological questions can be addressed by using only a few formal tests between this AIC-selected model and neighboring ones, thus limiting the increase in the overall risk of rejection of at least one null hypothesis otherwise caused by multiple tests."

The viewpoint held by others is that the informational approach can *entirely* replace hypothesis-testing procedures for performing model selection (see Akaike 1973, Sakamoto et al. 1986). One overall analysis can be conducted in place of multiple significance tests, thus avoiding inherent problems of unknown overall alpha levels associated with the classical approach (see, e.g., Bozdogan 1988*b*). Models with similar criterion values could be further compared by examining diagnostic measures and considering the analyst's biological knowledge and insights, rather than using hypothesis tests.

Other problems concerning alpha levels that are inherent in the classical approach are the arbitrariness of selecting an alpha level and the overemphasis on Type I error. Why do most researchers choose 0.05 as this value? It seems this number has become a 'Magic Number' against which a null hypothesis is either accepted or rejected (Toft 1990) without any consideration of any other information. In the ecological literature, for example, this strong adherence to a critical value of 0.05 for hypothesis testing has been criticized (see Toft and Shea 1983, Petranka 1990, Toft 1990). Toft (1990:360) even stated that "... most of us forget that it's a convention and treat it as omniscience.".

Type II error is just as critical as Type I error in many biological studies, particularly with exploratory analyses and multivariate analyses. However, choosing an alpha value of 0.05 mistakenly places more emphasis on Type I than Type II error. With specific reference to statistical model selection, Burnham and Anderson (1992:20) stated:

> "The 0.05 α-level is not considered appropriate because then too much of the emphasis is on type I error when type II error is equally important. Yet, by realizing this, authors were admitting that the problem of data-based model selection was not one of classical null hypothesis testing."

Various suggestions to remedy the abuse and misuse of alpha levels have been made. In the ecological literature, for example, it has been suggested that the exact critical value of each test be reported in a journal article to allow readers to independently assess the significance (see Petranka 1990) and that the power of a test should be determined and attention be paid to Type II error as well as to Type I error (Toft and Shea 1983, Toft 1990). These suggestions are good ways to deal with problems concerning alpha levels, but researchers can avoid such problems inherent in hypothesis-testing procedures by using the informational approach. Competing models can be ranked according to their informational criterion values. Data analysis thus focuses on selection of the best model(s) given the data at hand rather than testing hypotheses and having various problems related to imprecise and/or unknown alpha values.

Another advantage of the informational approach centers around both the real meaning of *P*-values and conclusions or inferences based on unknown samples vs. the data at hand. Analysts often forget that the actual *P*-value does not solely reflect the data at hand, but reflects what

would be expected to occur if the researcher would repeatedly sample the study population. Hence, the *P*-value is based on data that the researcher *never* obtained (see, e.g., Carver 1978:385). On the other hand, the value of a criterion such as AIC or ICOMP is based on the data at hand and the given model for which the criterion is calculated, not necessarily on unknown repeated samples. Conclusions and/or inferences drawn from tests and associated *P*-values are partly based on non-existing data, whereas conclusions from analyses using the informational approach are based more on the data at hand.

The informational approach also has the advantage, in more complex modeling situations, of allowing comparisons of non-nested models, whereas classical test statistics limit comparisons to that of only nested models (see Akaike 1985, Bozdogan 1988b, Burnham and Anderson 1992). For example, say that survival estimates for three age classes, S1, S2, and S3, in an animal population are to be compared. The possible outcomes (models) are:

Model Number	Model
1	$S1 \neq S2 \neq S3$
2	$(S1 = S2) \neq S3$
3	$(S1 = S3) \neq S2$
4	$S1 \neq (S2 = S3)$
5	$S1 = S2 = S3$

Model 1 is essentially the full model; it has the most parameters to estimate (three) of any model. Models 2-5 are all nested within Model 1 (i.e., they are all subsets or special cases of Model 1). Any model could be compared

to Model 1 using hypothesis-testing procedures, such as the likelihood ratio test statistic. However, Models 2 and 3, for example, are not special cases of one another (i.e., are not nested) and classical model selection cannot compare these two models (or any other non-nested models). Thus, hypothesis-testing procedures have limitations on the actual number of models which can be compared. Indeed, researchers modeling survival rates from telemetry and capture-recapture records have been able to compare non-nested models using AIC (Szymczak and Rexstad 1991, Burnham and Anderson 1992, Lebreton et al. 1992), but this could not have been accomplished using the classical approach.

In addition to these statistical advantages, the informational approach can provide an easy, straight-forward method to analyze data and present the analysis to readers, particularly for complex data, for several reasons. First, the analyst does not have to rely on numerous different test statistics and associated statistical tables. Instead, one can use MLE procedures with a particular technique (ANOVA, regression, discriminant analysis, etc.) in order to estimate the loglikelihood term for a given model, which is then used to calculate the criterion value. Second, the analyst does not have to rely on multiple tests and make multiple decisions based on those tests. Instead, the competing models can be ranked according to their AIC (or other criterion) values and the initial process of selecting the best models simply involves choosing those models with the lowest values.

Third, for a given analysis the whole process of estimating parameters, calculating criterion values, and ranking the models can be performed with one computer program (or several linked routines). Fourth, it is not

uncommon for readers of journal articles or attendees at meetings to be somewhat confused by the presentation of numerous statistical tests, *P*-values, and a bewildering array of significance stars or asterisks when a researcher summarizes the analysis of complex or multivariate data. However, an analyst using the informational approach can easily summarize and present the results by using tables which show the best models and their sizes (number of parameters), criterion values, penalty term, and the parameters or variables present. Readers can then easily see the best model for each given level of model size, along with which parameters are present. Graphs can also be used to show model diagnostics for the various models which have the lowest criterion values.

The informational approach, when used as a *tool* for model selection, is more likely than hypothesis-testing procedures to require that the analyst use biological information when selecting the best models to fit complex data. This idea is discussed in Part 4 of this dissertation, but can be summarized as follows. Many complex or multivariate data sets are unlikely to have one clearly "best" model based on just the ranking of criterion values. In such cases the analyst is forced to use an informational criterion as an initial measure for selecting the best models. Then, additional statistical information (e.g., diagnostic measures) and the analyst's biological knowledge and expertise should guide the selection of the final model or models (Burnham and Anderson 1992, see Part 4 of this dissertation for specific details of these different stages of model selection).

Unfortunately, users of hypothesis-testing procedures often base their analyses or model selection either entirely or largely on the results of

significance tests, thereby allowing the statistical analysis to completely dictate their findings. Perhaps this is not an inherent short-coming of the classical approach, but a fault of our collective misuse of this approach. Many critics of hypothesis testing have said that significance tests are relied upon too heavily by researchers in a variety of fields. Thus, substantive knowledge is often under-utilized for analysis and inference. In part, this problem is possibly caused by use of the word "significant" in the hypothesis-testing framework. A "significant" test result is often incorrectly interpreted to mean a "biologically important" result and the analyst does not then use his/her full biological knowledge in the interpretation of the data.

No such language regarding "significance" exists within the framework of the informational approach. Some guidelines do exist which suggest that models with AICs differing by only one or two can be considered as being equivalent (see Sakamoto et al. 1986:84-85). However, these are just guidelines and not "dogma" (at least until misguided users turn the guidelines into dogma). Every analyst must use biological information to decide which model is best suited for describing the data (and answering the relevant questions) when two or more models have similar AIC values.

SOME PERTINENT LITERATURE ON THE FUNDAMENTALS AND APPLICATIONS OF THE INFORMATIONAL APPROACH

General and technical literature

Unfortunately, no one or two books are presently available to serve as a complete resource on the informational approach that provide both a basic and comprehensive discussion (including many univariate and multivariate techniques) and also covers readily available computer software that would easily permit most biologists to quickly begin utilizing the informational approach. However, biologists who have had at least a few statistics courses could learn the basics of the informational approach by reading Sakamoto et al. (1986). Biologists who have taken some graduate-level statistics will probably find useful both Sakamoto et al. (1986) and the Proceedings of the first US/Japan Conference on the Frontiers of Statistical Modeling (Bozdogan 1994a, b, c).

Some of the more common statistical techniques often used by biologists are listed in Table 3.2 along with references which provide methodologies and/or applications of the informational approach with such techniques. For example, Sakamoto et al. (1986) give background information on the application of AIC with ANOVA. In addition, interested readers could read Rosenblum (1994) for comparisons of both the various informational criteria and these criteria with hypothesis-testing methods for one factor ANOVA.

Multiple regression (linear or logistic) and multivariate techniques such as log linear models for analysis of multi-way contingency tables, discriminant analysis, and principle components analysis are also listed in

Table 3.2. Hypothesis-testing procedures are often used in conjunction with many multivariate techniques, but the main objective with such techniques is really selection of the most appropriate model(s) rather than hypothesis testing, per se. For example, many statistics textbooks take the hypothesis-testing approach to modeling, but discuss the use of Mallows' C_p as one way to address the balance between overfitting and underfitting in regression analyses (e.g., see Myers 1986). The use of Mallows' C_p in multiple regression analyses *suggests* that analysts indeed know that hypothesis-testing procedures are often inadequate when it comes to wrestling with overfitting and underfitting. Like Mallows' C_p, model selection criteria used in the informational approach can provide a way to balance between overfitting and underfitting. However, informational criteria, unlike Mallows' C_p, can be used for more than just regression analyses.

Although Table 3.2 does not provide a comprehensive review of the literature, it does give readers a place to start to see how both statisticians and researchers are using the informational approach for selection of appropriate models based on the data at hand. A forthcoming book on statistical modeling and the informational approach is being completed by Hamparsum Bozdogan which should be of interest to many biologists.

Applications of the informational approach in biology

The use of AIC is a practical alternative to the classical approach for problems of model selection with biological data sets (Burnham and Anderson 1992). Researchers in other disciplines, such as engineering and psychometrics, were apparently exposed to the applications of AIC in their

respective literature earlier than biologists. For example, not long after Akaike's 1973 paper the use of AIC with multidimensional scaling was published in the psychometric literature by Takane (1978). In 1986, the Psychometric Society held a symposium on AIC at its annual meeting and then published four feature papers on AIC in its journal *Psychometrika* (see Akaike 1987, Bozdogan 1987, Sclove 1987, Takane et al. 1987). Comparatively, the advantageous aspects of the informational approach have been only "recently" utilized by researcher in the biological sciences. Nevertheless, biologists are beginning to use informational criteria with the modeling framework as an alternative to hypothesis-testing procedures.

One of the earlier, and now readily accepted, applications of the informational approach in ecology has been in the estimation and modeling of survival rates from capture-recapture data (see Huggins 1991, Szymczak and Rexstad 1991, Burnham and Anderson 1992, Lebreton et al. 1992, Anderson et al. 1994, Burnham et al. 1995a, b, Spendelow et al. 1995). It is not uncommon to see researchers in ecology and wildlife biology using AIC to select the best model(s) which provide estimates of survival rates of different age, size, and/or sex classes within an animal population. Other applications of the informational approach in research on animal populations include selecting the best model for describing fish growth (Tsangridis and Filippousis 1994), assessing the factors associated with mortality of rainbow trout caught by sportfishing (Schisler and Bergersen 1996), and modeling the relationships between habitat features and the

presence of black bears (van Manen 1994, van Manen and Pelton 1993) and green anoles (Minesky, Part 5 of this dissertation).

In toxicology, AIC has been used recently in modeling the uptake of methyl mercury by red blood cells of rats (Wu 1995) and in determining which characteristics of metal ions could best be used to predict the relative toxicity of those metals in freshwater conditions (McCloskey et al. 1996). Examples of the use of the informational approach in epidemiological research include modeling and estimating the rate of spread of HIV in a cohort of men (Byers et al. 1988), selection of statistical models to determine the genetic risk factors associated with lung cancer (Sellers et al. 1994), and determining both the genetic and environmental factors associated with physiological lung functioning (Chen et al. 1996). In physiology and medicine, AIC has been used in multivariate autoregressive modeling of feedback systems and homeostasis in humans (see Wada et al. 1994).

The informational approach using AIC has also been applied to research in biochemistry, molecular biology, and genetics. AIC, along with classical hypothesis testing, has been used in computer software to help researchers fit curves to enzyme kinetic data (Perella 1988). AIC has also been used to examine linkage relationships among genetic loci (Shiraishi 1988, Na'iem et al. 1993). Thus, the informational approach to statistical analysis is being successfully applied to a wide variety of biological research.

COMMENTS ON THE USE OF THE INFORMATIONAL APPROACH

Any statistical methodology has limitations. Obviously, small sample sizes are always a concern in any analysis. With the informational

approach the lack-of-fit term is a function of sample size and the penalty term (when expressed as a multiple of k) is a function of the number of parameters. It has been recommended that when using AIC the number of estimated parameters be less than $2\sqrt{n}$ ($n/2$ at most, where n = sample size) (Sakamoto et al. 1986:83). With any statistical technique, a larger sample size is preferable to a smaller one.

Use of AIC assumes that the true model is included within the global model or set of models being considered (Burnham and Anderson 1992). The global model here refers to the most general model and the one with the correct 'structure' (i.e., constraints placed on parameters) for the given data (Burnham and Anderson 1992). This assumption applies to all statistical modeling approaches and all model selection criteria, not just those used in the informational approach. In addition, if the researcher fails to measure a biologically important variable, then any criterion or statistical test procedure would be unable to evaluate and select the true model. One has to be practical and realize that such situations are related more to one's imperfect prior knowledge than to a limitation of the criterion being used. Researchers should always keep in mind that 1) one or more truly meaningful variables might not have been measured, 2) the "true" model might not actually be in the set of models being evaluated, and 3) no single study defines the "truth" in science. This last point serves to emphasize that statistical modeling can help to uncover and approximate the truth about a phenomenon, but one or a few models obtained from a study must be further verified or validated by additional independent studies. These studies should be in the form of modeling of

the same phenomenon by other researchers and testing the model's predictions or conclusions (hopefully by means of experiments). Sound conclusions or inferences cannot result from one study alone. Additional research must bear out the inferences and the scientific community must then reach a concensus.

AIC is not a formal test of significance of a model (Gilchrist 1984:161), nor are any informational criteria. AIC does not show that a model has a statistically significant fit to the data or that two models are significantly different in the same sense that 'significant' is used in the context of the classical approach. The informational and classical approaches are philosophically and statistically different. To say then that AIC is less useful because it is not a formal test would be to reject the advantages of the informational approach over the classical approach and to consider the classical paradigm as the only legitimate approach to statistical analysis. Again, how one views likelihood is important here. The Fisherian view leads the analyst to use likelihood as the basis for hypothesis-testing of a *given* model. The informational view formalized by Akaike leads the analyst to use likelihood as the basis for model selection criteria without a strict need of alpha values, statistical tables, and *P*-values. Failure by an analyst to see how analyses can take place without hypothesis-testing procedures and alpha values might be related to either being unaware of the informational view of likelihood or being unable to break out beyond the classical Fisherian view of likelihood.

If two or more statistical models have very similar values of the criterion no single best model exists to describe the data. Sakamoto et al.

(1986) consider any difference greater than 1 to 2 between the AIC values of two models to be of importance, otherwise the models can be considered statistically equivalent in their fit to the data. Burnham and Anderson (1992) point out that in situations where models have nearly equal AIC values, biological factors must be considered if a single best model is to be chosen.

Too often, however, it is assumed that a single best model or a single best explanation exists for complex data sets. Look, for example, at how stepwise selection procedures are often used in multiple regression and multivariate analyses to supposedly find a single 'best' model (combination of variables) by evaluating only a small number of the total possible models that exist for the given data. Researchers often report the model produced by a stepwise analysis as 'the best' model, but statisticians have warned that stepwise procedures will likely produce only a good model and not the 'best' model (see James and McCulloch 1990 for an overview of problems associated with using stepwise procedures). In some cases, no single best model exists based on either statistical or biological grounds. Some variables may be essentially equivalent substitutes for other variables. As McCullagh and Nelder (1989:23) stated:

> "Note that even if we could define exactly what is meant by an optimum model in a given context, it is most unlikely that the data would indicate a clear winner among the potentially large number of competing models. We must anticipate that, clustered around the 'best' model will be a set of alternatives almost as good and not statistically distinguishable."

Therefore, assumptions that every large, complex data set will have a uniquely best model may likely be incorrect. The informational approach,

through the use of criteria to rank and compare models, can be used to identify a single best model when such a model exists. However, this approach also forces analysts to admit that, for certain data sets, several alternative models may be equally as good and that no single best model exists. In addition, reporting the criterion values of a number of competing models allows readers of the published research to compare such models for themselves, rather than to rely on the analyst's limited results and interpretation coming from stepwise analysis.

An important practical consideration regarding the use of the informational approach by biologists at the present time is that many computer routines in statistical software do not provide the direct output of AIC. However, some routines do calculate AIC for certain analyses. For example, SAS (SAS Institute Inc. 1989a) outputs AIC values for linear regression models (in PROC REG) and for logistic regression models (in PROC LOGISTIC). In time series analysis, one can obtain AIC values for multivariate autoregressive models from the PROC STATESPACE procedure in SAS (Brocklebank and Dickey 1986). Analysts should check the manuals of their favorite statistical software to see which routines calculate AIC directly. If one can obtain the likelihood term for each competing model, then AIC values can be calculated by hand or in a separate routine by the user.

ICOMP and ICOMP-IFIM are not calculated directly by most statistical software packages. However, these criteria can be calculated using MATLAB (1989), SAS IML (SAS Institute 1989b), or other software which can perform matrix algebra. Using such software the analyst can write the

necessary code to calculate both the -2log likelihood term and the complexity term for each model. An important first step is finding the correct formulas for these terms for the particular analysis being used. Procedures to calculate likelihood terms for many techniques can be found in various statistics publications and in some statistical software manuals. The general formulas for complexity terms in ICOMP and ICOMP-IFIM were provided earlier in this chapter and can also be found in many of the publications of H. Bozdogan. For ICOMP, the analyst specifically needs to calculate the estimated covariance matrix of the estimated parameter values ($\widehat{Cov}(\theta)$) and the estimated covariance matrix of the estimated residual terms ($\widehat{Cov}(\varepsilon)$) for each model based on the technique being used (i.e., ANOVA, linear regression, logistic regression, log-linear models, etc.). The formulas for these covariance matrices can be obtained from many statistics books and statistical software manuals. For ICOMP-IFIM, the analyst specifically needs to write the computer code to calculate the estimated inverse Fisher information matrix and the necessary formulas can be found in certain statistical textbooks and statistical software packages. In some cases these matrices needed for the complexity terms of ICOMP and ICOMP-IFIM can be obtained from the output in some software packages and then incorporated into another routine to calculate the numerical values of the complexity terms and the criteria.

For those researchers who are not confident in writing their own computer code to calculate AIC, ICOMP, or ICOMP-IFIM, all is not lost. Certain computer routines can be found in some books (e.g., several FORTRAN routines appear in Sakamoto et al. 1986). Also, as more

researchers use the informational approach and write computer programs for their data analysis, more programs will become available to all researchers. Eventually, the increased use of the informational approach will also cause statistical software producers to incorporate routines to provide users with AIC, ICOMP, and ICOMP-IFIM as standard or optional output.

CONCLUDING REMARKS

To many biologists who use the classical approach, data analysis is usually equivalent to testing statistical hypotheses. Testing the statistical significance of parameter values is a central objective, but such significance is not necessarily equivalent to biological importance. Statistical modeling is performed in many analyses of biological data, but this is usually conducted using hypothesis-testing procedures. However, some biologists are using the informational approach which is known to be a viable alternative to such procedures.

Data analysis using the informational approach involves statistical modeling and model selection by using numerical criteria which serve for the evaluation and comparison of alternative models. Thus, modeling is conducted without the use of hypothesis-testing procedures. The criteria are derived from combining statistical likelihood theory and information theory as first shown by Akaike (1973). Estimation of parameter values and evaluation of the various models are done in one overall analysis. Practitioners of the informational approach focus on the selection of the most appropriate model(s) to describe the data at hand, realizing that the

data support only a certain amount of inference. The informational approach has certain statistical advantages over hypothesis-testing procedures which were discussed in this chapter.

Toft's (1990) statement, quoted at the beginning of this chapter, was made specifically about a debate over the uses and advantages of non-parametric versus parametric tests. However, her comment could be applied to any discussion about choosing an appropriate statistical method data analysis. It is in this spirit of "using the most appropriate tool" that the informational approach to statistical analysis is used in this dissertation and should be considered by biologists for addition to their statistical tool boxes. Biologists should be both aware of alternatives to the classical approach and open to the possibility of expanding their ability to better analyze their data. Use of the informational approach would certainly expand this ability.

Some researchers and statisticians would say that one only needs to use the informational approach and not statistical hypothesis-testing procedures. At the present time, whether or not a researcher uses the informational approach exclusively or in conjunction with hypothesis-testing procedures will largely depend on practical considerations, the researcher's viewpoints, and the viewpoints of journal editors and reviewers. Healthy discussions of the utility of the informational approach for analysis of biological data should continue. Biologists must ask themselves "Does the classical approach have *primacy* over all other statistical approaches?". Many biologists are beginning to find the answer is "No" and that the informational approach has great utility in the

biological sciences. Perhaps Baxter's (1991:356-357) words (appearing as a book review) should be kept in our thoughts about data analysis :

> "... Professor Akaike is quietly assembling his own theory of statistical estimation based on entropy, information and likelihood, centred around the Akaike information criterion (AIC), ... and that this theory is more likely to survive than most, being based on data and common sense rather than dogma."

Are statistics and data analysis evolving? Changes certainly *are* occurring in how data analysis is conducted and in how researchers and statisticians view the discipline of statistics. Will a new synthesis emerge? Whether a new synthesis will emerge or the various approaches will continue to be separate in practice (with a researcher using just one approach on a given data set) remains to be seen. Certainly the rapid growth of computers in research and recent advances in statistical and graphics software are shaping the way biologists conduct data analysis. What might emerge as a new synthesis of data analysis and modeling? One possibility might be the combination of graphical methods for visual display and summary of data and models, informational criteria (or Mallows' C_p in some cases) for model selection, mathematical diagnostics and graphical diagnostics (along with some confidence intervals and/or certain hypothesis-testing procedures) for model checking and diagnostics assessment. Indeed, many statisticians have been preaching such a path to statistical enlightenment (although their model selection process is based on hypothesis-testing procedures rather than the informational approach), but few biologists firmly practice the complete faith.

Multivariate analysis is one area that appears to be changing, though perhaps slowly, along the lines described above. Statistical modeling of complex or multivariate data should, and often does, occur in stages. First, model selection criteria, such as AIC or ICOMP, can be used to rank the many competing models. Second, the initial 'best' models can be examined further using diagnostic measures and graphical techniques. Then, the final stage of selection of the best models include biological interpretation of models and parameter values along with common sense. This whole process is somewhat counter to the way model selection is currently practiced by many biologists who rely strictly on hypothesis tests, P-values, and stepwise algorithms. However, Part 4 of this dissertation presents a new method for analysis of observational ecological data whereby the informational approach is combined with a genetic algorithm for use with multiple logistic regression models.

Should a new synthesis of data analysis or statistical methods be nurtured, taught, and used by biologists and statisticians alike? In this researcher's opinion, yes. The tools now available to biologists for data analysis go far beyond the mere realm of hypothesis-testing procedures. It would be a mistake for undergraduate and graduate programs in the biological sciences to teach students only the classical approach. This would *narrowly* train biological researchers about statistics and data analysis. Students in the biological sciences should be exposed to a variety of statistical methods and approaches in order to provide the students with the all of powerful tools they will use in their professional careers. Teaching the informational approach, along with other approaches and

methods, would be a step toward diversifying the statistical skills of new biologists.

LITERATURE CITED

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267-281 *in* B. N. Petrov and F. Csáki, editors. Second international symposium on information theory. 1971. Akadémiai Kiadó, Budapest, Hungary. 451 pp.

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

Akaike, H. 1985. Prediction and entropy. Pages 1-24 *in* A. C. Atkinson and S. E. Fienberg, editors. A celebration of statistics: the ISI centenary volume. Springer-Verlag, New York, New York, USA. 606 pp.

Akaike, H. 1987. Factor analysis and AIC. Psychometrika 52:317-332.

Akaike, H. 1994. Implications of informational point of view on the development of statistical science. Pages 27-38 *in* H. Bozdogan, editor. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. 1992. Vol. 3, engineering and scientific applications. Kluwer Academic Publishers, Dordrecht, The Netherlands. 346 pp.

Anderson, D. R., K. P. Burnham, and G. C. White. 1994. AIC model selection in overdispersed capture-recapture data. Ecology 75:1780-1793.

Barnett, V. 1973. Comparative statistical inference. John Wiley and Sons, London, United Kingdom. 287 pp.

Baxter, L. A. 1991. Book review of A celebration of statistics: the ISI centenary volume. Journal of the Royal Statistical Society A154:356-357.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. Discrete multivariate analysis: theory and practice. The MIT Press, Cambridge, Massachusetts, USA. 557 pp.

Bozdogan, H. 1987. Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. Psychometrika 52:345-370.

Bozdogan, H. 1988a. ICOMP: a new model-selection criterion. Pages 599-608 *in* H. H. Bock, editor. Classification and related methods of data analysis: proceedings of the first conference of the international classification societies. North-Holland, Amsterdam, The Netherlands. 750 pp.

Bozdogan, H. 1988b. Selecting loglinear models and subset selection of variables in multiway contingency tables using Akaike's Information Criterion (AIC). Pages 609-616 *in* H. H. Bock, editor. Classification and related methods of data analysis. Proceedings of the First Conference of the International Classification Societies. North-Holland, Amsterdam, The Netherlands. 750 pp.

Bozdogan, H. 1990. On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. Communications in Statistics: Theory and Methods 19:221-278.

Bozdogan, H. (editor). 1994a. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. 1992. Vol. 1, theory and methodology of time series analysis. Kluwer Academic Publishers, Dordrecht, The Netherlands. 277 pp.

Bozdogan, H. (editor). 1994b. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. 1992. Vol. 2, multivariate statistical modeling. Kluwer Academic Publishers, Dordrecht, The Netherlands. 413 pp.

Bozdogan, H. (editor). 1994c. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. 1992. Vol. 3, engineering and scientific applications. Kluwer Academic Publishers, Dordrecht, The Netherlands. 346 pp.

Bozdogan, H. and D. M. A. Haughton. 1998. Informational complexity criteria for regression models. Computational Statistics & Data Analysis 28:51-76.

Bozdogan, H. and S. L. Sclove. 1984. Multi-sample cluster analysis with varying parameters using Akaike's Information Criterion. Annals of the Institute of Statistical Mathematics 36:163-180.

Bozdogan, H., S. L. Sclove, and A. K. Gupta. 1994. AIC-replacements for some multivariate tests of homogeneity with applications in multisample clustering and variable selection. Pages 199-232 *in* H. Bozdogan, editor. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. 1992. Vol. 2, multivariate statistical modeling. Kluwer Academic Publishers, Dordrecht, The Netherlands. 413 pp.

Brocklebank, J. C. and D. A. Dickey. 1986. SAS® system for forecasting time series. SAS Institute Inc., Cary North Carolina, USA. 240 pp.

Burnham, K. P. and D. R. Anderson. 1992. Data-based selection of an appropriate biological model: the key to modern data analysis. Pages 16-30 *in* D. R. McCullough and R. H. Barrett, editors. Wildlife 2001: Populations. Elsevier Science Publishers, London, United Kingdom. 1163 pp.

Burnham, K. P., D. R. Anderson, and G. C. White. 1995a. Selection among open population capture-recapture models when capture probabilities are heterogeneous. Journal of Applied Statistics 22:611-624.

Burnham, K. P., G. C. White, and D. R. Anderson. 1995b. Model selection strategy in the analysis of capture-recapture data. Biometrics 51:888-898.

Byers, R. H., Jr., W. M. Morgan, W. W. Darrow, L. Doll, H. W. Jaffe, G. Rutherford, N. Hessol, and P. M. O'Malley. 1988. Estimating AIDS infection rates in the San Francisco cohort. AIDS 2:207-210.

Carver, R. P. 1978. The case against statistical testing. Harvard Educational Review 48:378-399.

Chen, Y., S. L. Horne, D. C. Rennie, and J. A. Dosman. 1996. Segregation analysis of two lung function indices in a random sample of young families: the Humboldt family study. Genetic Epidemiology 13:35-47.

Deming, W. E. 1975. On probability as a basis for action. American Statistician 29:146-152.

Dutilleul, P. 1993. Spatial heterogeneity and the design of ecological field experiments. Ecology 74:1646-1658.

Fienberg, S. E. 1980. The analysis of cross-classified categorical data. The MIT Press, Cambridge, Massachusetts, USA. 198 pp.

Flury, B. D. and B. E. Neuenschwander. 1994. Modelling principal components with structure. Pages 183-198 *in* H. Bozdogan, editor. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. 1992. Vol. 2, multivariate statistical modeling. Kluwer Academic Publishers, Dordrecht, The Netherlands. 413 pp.

Gilchrist, W. 1984. Statistical modelling. John Wiley and Sons, Chichester, United Kingdom. 339 pp.

Guttman, L. 1985. The illogic of statistical inference for cumulative science. Applied Stochastic Models and Data Analysis 1:3-9.

Huggins, R. M. 1991. Some practical aspects of a conditional likelihood approach to capture experiments. Biometrics 47:725-732.

James, F. C. and C. E. McCulloch. 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? Annual Review of Ecology and Systematics 21:129-166.

Johnson, D. H. 1995. Statistical sirens: the allure of nonparametrics. Ecology 76:1998-2000.

Johnson, R. A. and G. K. Bhattacharyya. 1996. Statistics : principles and methods. 3rd edition. John Wiley and Sons, Inc., New York, New York, USA. 720 pp.

Jones, D. and N. Matloff. 1986. Statistical hypothesis testing in biology: a contradiction in terms. Journal of Economic Entomology 79:1156-1160.

Lebreton, J. D., K. P. Burnham, J. Clobert, and D. R. Anderson. 1992. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecological Monographs 62:67-118.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74:1659-1673.

MATLAB. 1989. Pro-MATLAB for VAX/VMS Computers. The Math Works, Inc., South Natick, Massachusetts, USA. 356 pp.

Matloff, N. S. 1991. Statistical hypothesis testing: problems and alternatives. Environmental Entomology 20:1246-1250.

Meehl, P. E. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology 46:806-834.

McCloskey, J. T., M. C. Newman, and S. B. Clark. 1996. Predicting the relative toxicity of metal ions using ion characteristics: Microtox registered bioluminescence assay. Environmental Toxicology and Chemistry 15:1730-1737.

McCullagh, P. and J. A. Nelder. 1989. Generalized linear models. 2nd edition. Chapman and Hall, London, United Kingdom. 511 pp.

Morrison, D. E. and R. E. Henkel (editors). 1970. The significance test controversy: a reader. Aldine Publishing Company, Chicago, Illinois, USA. 333 pp.

Myers, R. H. 1986. Classical and modern regression with applications. Duxbury Press, Boston, Massachusetts, USA. 359 pp.

Na'iem, M., Y. Tsumura, K. Uchida, T. Nakamura, and K. Ohba. 1993. Linkage of allozyme loci in Japanese red pine (*Pinus densiflora*). Canadian Journal of Forest Research 23: 680-687.

Perrella, F. W. 1988. EZ-FIT: a practical curve-fitting microcomputer program for the analysis of enzyme kinetic data on IBM-PC compatible computers. Analytical Biochemistry 174:437-447.

Perry, J. N. 1986. Multiple-comparison procedures: a dissenting view. Journal of Economic Entomology 79:1149-1155.

Petranka, J. W. 1990. Caught between a rock and a hard place. Herpetologica 46:346-350.

Potvin, C. and D. A. Roff. 1993. Distribution-free and robust statistical methods: viable alternatives to parametric statistics? Ecology 74:1617-1628.

Pratt, J. W. 1976. A discussion of the question: for what use are tests of hypotheses and tests of significance. Communications in Statistics, Theory and Methods A5:779-787.

Quinn, J. F. and A. E. Dunham. 1983. On hypothesis testing in ecology and evolution. American Naturalist 122:602-617.

Roberts, H. V. 1976. For what use are tests of hypotheses and tests of significance. Communications in Statistics, Theory and Methods A5:753-761.

Rosenblum, E. P. 1994. A simulation study of information theoretic techniques and classical hypothesis tests in one factor ANOVA. Pages 319-346 *in* H. Bozdogan, editor. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. 1992. Vol. 2, multivariate statistical modeling. Kluwer Academic Publishers, Dordrecht, The Netherlands. 413 pp.

Sakamoto, Y. 1982. Efficient use of Akaike's Information Criterion for model selection in high dimensional contingency table analysis. Metron 40:257-275.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. Akaike information criterion statistics. KTK Scientific Publishers, Tokyo, Japan. (copublished with D. Reidel Publishing Company, Dordrecht, Holland). 290 pp.

Salsburg, D. S. 1985. The religion of statistics as practiced in medical journals. The American Statistician 39:220-223.

SAS Institute Inc. 1989a. SAS/STAT® user's guide. Version 6, Fourth Edition, Volume 2. Cary, North Carolina, USA. 846 pp.

SAS Institute Inc. 1989b. SAS/IML® software: usage and reference. Version 6, First Edition. Cary, North Carolina, USA. 501 pp.

Scheiner, S. M. and J. Gurevitch, editors. 1993. Design and analysis of ecological experiments. Chapman and Hall, New York, New York, USA. 445 pp.

Schisler, G. J. and E. P. Bergersen. 1996. Postrelease hooking mortality of rainbow trout caught on scented artificial baits. North American Journal of Fisheries Management 16:570-578.

Sclove, S. L. 1987. Application of model-selection criteria to some problems in multivariate analysis. Psychometrika 52:333-343.

Seaman, J. W., Jr., and R. G. Jaeger. 1990. Statisticae dogmaticae: a critical essay on statistical practice in ecology. Herpetological 46:337-346.

Sellers, T. A., P.-L. Chen, J. D. Potter, J. E. Bailey-Wilson, H. Rothschild, and R. C. Elston. 1994. Segregation analysis of smoking-associated malignancies: evidence for Mendelian inheritance. American Journal of Medical Genetics 52:308-314.

Shaw, R. G. and T. Mitchell-Olds. 1993. ANOVA for unbalanced data: an overview. Ecology 74:1638-1645.

Shiraishi, S. 1988. Linkage relationships among allozyme loci in Japanese black pine, *Pinus thunbergii* Parl. Silvae Genetica 37:60-66.

Simberloff, D. 1990. Hypotheses, errors, and statistical assumptions. Herpetologica 46:351-357.

Smith, S. M. 1995. Distribution-free and robust statistical methods: viable alternatives to parametric statistics. Ecology 76:1997-1998.

Sokal, R. R. and F. J. Rohlf. 1995. Biometry : the principles and practice of statistics in biological research. 3rd edition. W. H. Freeman and Company, New York, New York, USA. 887 pp.

Spendelow, J. A., J. D. Nichols, I. C. T. Nisbet, H. Hays, G. D. Cormons, J. Burger, C. Safina, J. E. Hines, and M. Gochfeld. 1995. Estimating annual survival and movement rates of adults within a metapopulation of roseate terns. Ecology 76:2415-2428.

Stewart-Oaten, A. 1995. Rules and judgments in statistics: three examples. Ecology 76:2001-2009.

Szymczak, M. R. and E. A. Rexstad. 1991. Harvest distribution and survival of a gadwall population. Journal of Wildlife Management 55: 592-600.

Takane, Y. 1978. A maximum likelihood method for nonmetric multidimensional scaling: 1. the case in which all empirical pairwise orderings are independent - theory and applications. Japanese Psychological Research 20:7-17, 105-114.

Takane, Y., H. Bozdogan, and T. Shibayama. 1987. Ideal point discriminant analysis. Psychometrika 52:371-392.

Toft, C. A. 1990. Reply to Seaman and Jaeger: an appeal to common sense. Herpetologica 46:357-361.

Toft, C. A. and P. J. Shea. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. American Naturalist 122:618-625.

Trexler, J. C. and J. Travis. 1993. Nontraditional regression analyses. Ecology 74:1629-1637.

Tsangridis, A. and N. Filippousis. 1994. Analysis of two models for picarel (*Spicara smaris* L.) growth using Schnute's micro-simplex nonlinear estimation procedure. Fisheries Research 20:181-189.

Tsuruta, S. and Y. Nogami. 1986. AIC in log linear model for contingency tables with Poisson and multinomial designs. Journal of the Japanese Statistical Society 16:165-172.

van Embden, M. H. 1971. An analysis of complexity. Mathematical Centre Tracts 35. Mathematisch Centrum, Amsterdam, The Netherlands. Cited in Bozdogan 1988a, 1990.

van Manen, F. T. 1994. Black bear habitat use in Great Smoky Mountains National Park. Ph.D. dissertation. University of Tennessee, Knoxville, Tennessee, USA. 212 pp.

van Manen, F. T. and M. R. Pelton. 1993. Data-based modelling of black bear habitat using GIS. Pages 323-329 *in* I. D. Thompson, editor. Proceedings of the International Union of Game Biologists XXI Congress: forests and wildlife .... towards the 21st century, Volume 1. Canadian Forest Service, Chalk River, Ontario, Canada. 379 pp.

Wada, T., T. Koyama, and M. Shigemori. 1994. Multivariate autoregressive modeling for analysis of biomedical systems with feedback. Pages 293-317 *in* H. Bozdogan, editor. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. 1992. Vol. 2, multivariate statistical modeling. Kluwer Academic Publishers, Dordrecht, The Netherlands. 413 pp.

Wu, G. 1995. Prediction of uptake of methyl mercury by rat erythrocytes using a two-compartment model. Archives of Toxicology 70:34-42.

Yoccoz, N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bulletin of the Ecological Society of America 72:106-111.

APPENDIX TO PART 3

Table 3-1. Selected examples of publications which have expressed concerns about statistical hypothesis-testing procedures and the use and/or abuse of such procedures in scientific research. Publications listed here either broadly discuss concerns with hypothesis-testing procedures or express concerns in the context of specific types of analyses or problems. Names of authors and date of publication are given under "Publication". "Discipline" describes the type of journal in which the publication appeared and/or the scientific audience mainly addressed by the publication[a]. "Concerns/Criticisms" provides a brief summary of the expressed concerns about the use of and/or overreliance on hypothesis-testing procedures[b].

Publication	Discipline[a]	Concerns/Criticisms[b]
Morrison and Henkel (1970)*	Phil., Soc. sci., Stat.	1, 2, 3, 4, 7, 8, 9, 10, 11, 12
Deming (1975)	Stat.	1, 2, 4, 11
Pratt (1976)	Stat.	9
Roberts (1976)	Stat.	9
Carver (1978)	Ed.	1, 2, 12
Meehl (1978)	Soc. sci., Stat.	1, 2, 3, 4, 12
Guttman (1985)	Stat.	
Salsburg (1985)	Med., Stat.	7, 8, 10, 11, 12
Jones and Matloff (1986)	Biol.	1, 2, 4, 10, 11
Perry (1986)	Biol., Ent.	1, 2, 3, 4, 9
Bozdogan (1988*b*)	Stat.	6, 7
Matloff (1991)	Biol.	1, 2, 4, 11
Yoccoz (1991)	Biol., Ecol.	1, 2, 3, 10
Burnham and Anderson (1992)	Biol., Ecol.	5, 6, 7
Johnson (1995)	Ecol.	1, 2, 4, 10, 11
Stewart-Oaten (1995)	Ecol.	4, 7, 11

[a]Biol. = biology, in general, Ecol. = ecology, Ed. = education, Med. = medicine, Phil. = philosophy of science, Soc. sci. = social sciences (sociology, psychology), and Stat. = statistics.

[b]1 = statistical significance of a hypothesis-testing procedure does not automatically indicate substantive importance

2 = hypothesis-testing procedures are a function of sample size (e.g., the larger the sample the more likely the test result will show statistical significance)

3 = rejection of a null hypothesis could as easily be due to violation of one or more assumptions underlying the test procedure as due to real differences among parameters being tested

4 = the null hypothesis is nearly always false in practice (i.e., researchers often already know that differences exist - why else would they collect the data?); some authors suggest researchers should focus instead on the magnitude and substantive meaning of the differences

5 = hypothesis-testing procedures usually start with an assumed model without checking the model's validity and/or usually only two models can be compared at a time with a single, given test

6 = hypothesis-testing procedures have certain statistical weaknesses when the task is to select an appropriate model to fit the data (e.g., too much emphasis placed on type I error and/or comparison of non-nested models is impossible)

7 = choice of alpha level is arbitrary

8 = hypothesis-testing procedures most often set up a decision to be made between only two competing hypotheses, but relevant research questions are not usually simple 'yes or no' questions

9 = hypothesis-testing procedures have a limited role to play in data analysis and inference; these procedures should be used for only certain types of questions and analyses

10 = researchers rely too heavily on the use of statistical hypothesis-testing procedures for purposes of inference and/or inquiry when other methods are equally or more useful for such purposes

11 = hypothesis tests and statistical significance do not reveal all aspects of the data that are needed to make informed decisions or to take action

12 = for various reasons based on logic and philosophy, hypothesis-testing procedures either a) are not superior to other methods of reasoning and inquiry or b) represent a distortion of the scientific method

*Morrison and Henkel (1970) provide a collection of author's concerns about hypothesis-testing procedures that go beyond the list above and summarize such concerns (see their p.305-311) based on both statistical and philosophical issues.

Table 3-2. Various statistical techniques commonly used by biologists and references which provide some explanation of how to use the informational approach in conjunction with these techniques.

Technique	Reference(s)
ANOVA	Sakamoto et al. (1986:202-222)
Simple contingency tables	Sakamoto et al. (1986:121-137)
Log linear models (multi-way contingency tables)	Sakamoto (1982), Tsuruta and Nogami (1986), Bozdogan (1988*b*)
Multiple linear regression	Sakamoto et al. (1986:180-184)
Polynomial regression	Sakamoto et al. (1986:165-179)
Multiple logistic regression	van Manen and Pelton (1993), van Manen (1994), Minesky (Parts 4 and 5 of this dissertation)
Factor analysis	Akaike (1987)
Principle components analysis (with some structure in data)	Flury and Neuenschwander (1994)
Multivariate tests of the homogeneity of covariance matrices	Bozdogan and Sclove (1984), Bozdogan et al. (1994)

PART 4 : THE GENETIC ALGORITHM WITH AN INFORMATIONAL CRITERION: AN ALTERNATIVE METHOD FOR STATISTICAL MODELING OF OBSERVATIONAL DATA

INTRODUCTION

The nature and complexity of the physical and biological worlds often require ecologists to take a multi-stage research approach in order to propose causal models and eventually uncover causal factors. The first stage occurs when well-defined questions have not yet been formulated and observational (non-experimental) data are collected on a large set of biologically relevant variables regarding a general question or phenomenon. In this early exploratory stage of inquiry, ecologists frequently analyze such data using multiple regression methods and multivariate statistical techniques. The results from this first stage are next used to construct possible causal models. Then, the actual testing of causal models and the possible uncovering of causal factors *requires* an experimental approach (James and McCulloch 1990).

A key problem in the analysis of the large set of independent variables in any observational study is to find the best statistical models which capture the essence of the data with smaller, more parsimonious sets of variables than the full set. This process of selecting the best statistical models is non-trivial when many variables are used and the number of possible models is in the tens of thousands or more. In such cases ecologists typically select one (or a few) models by using stepwise computer algorithms which build models by adding or removing one variable at a time to an initial model. However, stepwise algorithms have been considered inadequate for use with multiple regression modeling (Hocking 1976, 1983; Moses 1986). James and McCulloch (1990) stated that stepwise procedures are misused with multiple regression and multivariate analysis

in ecology and systematics and that researchers should avoid using such procedures.

The purpose of this part of the dissertation is to introduce ecologists to the use of a genetic algorithm (GA) in combination with informational model-selection criteria for conducting statistical model selection (based on the initial work of Luh et al., submitted manuscript) for multiple regression and multivariate techniques. This method is proposed as an alternative to the commonly used, but rather problematic stepwise procedures. A GA is a searching algorithm which utilizes certain principles of genetics, evolution, and natural selection to find the best solutions to a given problem when thousands or even millions of potential solutions exist (Holland 1992a, Forrest 1993, Goldberg 1994). An informational criterion provides a numerical value, based on a combination of statistical and informational theory, which includes a measure of both the fit of the model to the given data and the number of estimated parameters or complexity of each model (see Akaike 1973, Sakamoto et al. 1986, Bozdogan 1987, 1988a, b). The importance of such a criterion is that it is calculated for each statistical model and it provides a valuable means to rank and compare competing models and then to select the most parsimonious model (or set of models) for a given data set.

In the proposed methodology, a random population of statistical models is generated for the data at hand and the value of an informational criterion is calculated for each model in this population. Then, the GA allows models to recombine and produce a new generation of models. The value of the informational criterion is calculated for each new model in

this second generation. Whether or not a given model survives into a subsequent generation and/or recombines with other models depends on the model's "fitness" value, represented by its numerical value of the informational criterion, and some elements of chance. Thus, the GA searches the vast set of models to find the best models by means of a sampling process based on concepts of evolution and natural selection.

In general, it is proposed that researchers who perform statistical modeling on observational data use this GA-informational methodology to obtain a *set* of 'best' models, rather than continue to rely on stepwise procedures which provide a limited view of the data by finding only a few good models out of the vast set of possible models. The set of best models obtained from a GA can potentially provide *more* insight into the data than just one or two models obtained from stepwise procedures, thus enabling researchers to better use observational studies for refining questions and designing experiments. The purpose of any observational study and its use of multivariate analysis is not to support strong inferences, but to provide insight and information which then contributes to the next phase of research (see James and McCulloch 1990). The use of a GA and informational criteria can assist ecologists in this aim when observational data are analyzed. In particular, this alternative approach is explained for the case of multiple logistic regression and some suggestions are made to possibly improve logistic regression modeling of observational data.

Much of the work presented here has been the product of a collaboration between the author and Drs. Hang-Kwang Luh and

Hamparsum Bozdogan. Dr. Luh wrote the computer code for the GA and contributed to discussions on the applications of the GA to ecological studies. Dr. Bozdogan provided both expertise in the informational approach to statistical modeling and a computer program for estimating both the maximum loglikelihood term and the regression parameters for the case of logistic regression. The author expanded this existing program, with the assistance of Dr. Bozdogan, to calculate both the estimated variance and one of the informational model selection criteria (called ICOMP-IFIM) for the logistic regression model. The author also developed the guidelines and methodology for both the handling of the GA output and the subsequent selection of the best variables and models. This development of methodology was critiqued by Drs. Luh and Bozdogan.

Presentation of the proposed methodology starts with an introduction of the multiple logistic regression model and comments on some important assumptions about this model that researchers must consider. Then, a review is presented on the problem of selecting the best statistical model (combination of independent variables) when many independent variables exits. Finally, an overview of GAs is provided along with the potential application of a GA to statistical modeling and guidelines for such application. Part 5 of this dissertation then applies the proposed methodology to statistical modeling of actual field data in order to describe possible green anole-habitat relationships in four habitats in eastern Tennessee.

THE LOGISTIC REGRESSION MODEL

Overview

Logistic regression has been used for data analysis in a wide variety of research in the biological and health sciences where the dependent variable usually has just two possible values. This technique can be useful in the analysis of certain ecological data (Trexler and Travis 1993). A thorough treatment of the logistic regression model and its applications can be found in Hosmer and Lemeshow (1989). A brief overview is provided here followed by a more mathematical and statistical review.

Recall that linear regression is a technique for modeling the linear relationship between a dependent variable (also known as the outcome or response) and one or more independent variables (also called the predictor or explanatory variables or simply, the covariates). The dependent and independent variables are measured as continuous variables, but the model can include categorical (discrete) independent variables.

Logistic regression and linear regression both belong to the family of generalized linear models (McCullagh and Nelder 1989). However, logistic regression differs from linear regression in several ways. First, the outcome variable (Y) in logistic regression is binary rather than continuous. The outcome can be any response variable which exhibits a dichotomy and is usually recorded as 0 ("failure") or 1 ("success"), although the model can be expanded to handle outcomes with three or more discrete values. Independent variables (xs) can be continuous, categorical or a combination of both.

A second difference is that the mean of the outcome, given some values of the independent variables, exists as a non-linear function, unlike the linear relationship in linear regression. This mean is often called the conditional mean or the expected value of Y given x. A graph of the conditional mean values versus the values of x produces the familiar S-shaped logistic curve. Thus, the logistic distribution, rather than the normal distribution as for linear regression, is relevant to logistic regression (Hosmer and Lemeshow 1989:5-6). Other distributions can be used in the regression-type analyses of binary data, but the logistic model is often preferred because it is easily used, very flexible, and biologically interpretable (Hosmer and Lemeshow 1989:6).

A third difference between logistic and linear regression regards the error terms. In linear regression the errors are assumed to have a normal distribution with a mean of zero and a constant variance across the range of x values. In logistic regression the distribution of errors is binomial with a mean of zero and a variance equal to the expected value of Y given x (noted as $E(Y|x)$) times 1 minus $E(Y|x)$ (Hosmer and Lemeshow 1989:7). If the observed value of the outcome is 1, then the error equals 1 - $E(Y|x)$; otherwise if the observed value of the outcome is 0, then the error equals [- $E(Y|x)$]. Thus, the distribution of the errors is binomial in logistic regression, rather than normal and the value of the error depends on the value(s) of the independent variables.

Lastly in logistic regression, iterative maximum likelihood estimation (MLE) methods must be used for proper estimation of parameters, rather than the least squares method, because of the binomial nature of the data.

The interested reader is referred to McCullagh and Nelder (1989) for a review of MLE methods used to calculate these parameter values. Despite these differences some of the same principles and methods used in linear regression can be applied or adapted to logistic regression analysis (see Hosmer and Lemeshow 1989).

The multiple logistic regression model

The multiple logistic regression model for the case where all independent variables are continuous variables (at least interval in scale) is defined as (see Hosmer and Lemeshow 1989:25-26):

$$E(Y \mid x) = \pi(x) = \frac{e^{(ß_0 + ß_1 x_1 + ß_2 x_2 + \ldots + ß_p x_p)}}{1 + e^{(ß_0 + ß_1 x_1 + ß_2 x_2 + \ldots + ß_p x_p)}}, \qquad (4.1)$$

where: $E(Y \mid x) = \pi(x)$ = the conditional mean of Y given x,

x = the vector of independent variables $(x_1, x_2, \ldots, x_p)$,

p = the number of independent (continuous) variables,

e = the base of the natural logarithm (value = 2.71828),

$ß_0$ = the intercept parameter, and

$ß_1, ß_2, \ldots, ß_p$ = the regression parameters for the respective x.

Here $\pi(x)$ also represents the probability that the outcome is a "success" (Y=1) given some value of x.

If the data set has one or more independent variables which are categorical, such as sex or habitat type, then each of those variables can be coded as a design (dummy) variable. A categorical variable with three

levels (categories) requires only two design variables. In general, the number of design variables needed for a given categorical variable is simply 1 - c, where c is the total number of levels for that independent variable. Thus, 1 - c regression parameters would be estimated for each categorical variable in a model. If d represents the total number of categorical independent variables in the model, then the total number of regression parameters to be estimated would be:

$$k = 1 + p + \sum_{1}^{d} (1 + c) . \tag{4.2}$$

Since p independent variables measured on a continuous scale exist in the model, the number of parameters needed to be estimated for those variables is also p. The '1' in the above formula is included because one parameter estimate for the intercept is needed in the model (when the outcome is binary). The design variables and their parameters can be incorporated into the multiple logistic regression model of equation (4.1) (see Hosmer and Lemeshow 1989:26-27 for details). For simplicity, equation (4.1) will be used to denote the logistic regression model whether the model contains only continuous variables or both categorical and continuous variables (the specific equation for the model with both continuous and categorical variables can be found in Hosmer and Lemeshow 1989: 26-27). Note, however, if all the independent variables in the logistic regression model are categorical, then the model is equivalent

to a loglinear model (multiway contingency table) when associations among the independent variables are not included (Freeman 1987:258-261).

Different methods can be used for coding categorical variables as design variables for computational purposes, but not all statistical software uses the same method. The method which uses a referent group (reference cell) is the most commonly used. Interested readers can find details of this method and others for coding categorical variables in Hosmer and Lemeshow (1989:47-56).

Recall that for continuous independent variables the logit transformation provides a way to linearize the relationship between each of those variables and the outcome (Hosmer and Lemeshow 1989:6-7). The logit transformation is simply:

$$\text{logit} = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right]. \tag{4.3}$$

The logit transformation on $\pi(x)$ produces values which can potentially range from $-\infty$ to $+\infty$ (depending on the values of x), rather than values bounded by 0 and 1. The logit can be thought of in terms of statistical odds. Remember that Y is binary (0 or 1) in the case of logistic regression. Thus, the odds is the probability that $Y = 1$ given the values of the independent variables (x) divided by one minus the probability that $Y = 1$ given the values of the independent variables (x). The natural logarithm of the odds is simply the logit. In addition, the logit function allows equation (4.1) to be expressed in linear form (Hosmer and Lemeshow 1989:25):

$$\text{logit} = ß_0 + ß_1 x_1 + ß_2 x_2 + \ldots + ß_p x_p \,. \tag{4.4}$$

It is this linear form of the logit that is used in logistic regression analysis.

In order to fit a model to the data the regression parameters, $ß_s$, must be estimated. Estimation of logistic regression parameters uses the principle of maximum likelihood whereby the estimated value for any given parameter is the one which maximizes the likelihood function (Hosmer and Lemeshow 1989:8-9, 25-27). The log of the likelihood function is simpler to work with and is defined for a logistic regression model as:

$$\ln L(\theta_k) = \sum_{i=1}^{n} [(y_i)\ln(\pi(x_i)) + (1 - y_i)\ln(1 - \pi(x_i))] \,. \tag{4.5}$$

The symbols are the same as those used in equation (4.1). The y_i are the observed values of the outcome for each observation from $i = 1, 2, \ldots, n$, the $\pi(x_i)$ are the fitted values of equation (4.1) using the observed values of all of the independent variables for each observation, and n is the total number of observations (see Hosmer and Lemeshow 1989:25-28 for further details).

Differentiation of the loglikelihood function produces likelihood equations used for obtaining maximum likelihood estimates of the ßs (see Hosmer and Lemeshow 1989:27-28) for the equations). In practice, the actual estimation of the parameters requires using iterative methods (see

McCullagh and Nelder 1989). Many packaged statistical programs calculate the estimated parameter values and the loglikelihood value for a given logistic regression model for the user.

Two important assumptions deserve discussion with respect to the logistic regression and linear regression models. The first important difference between these regression models is that linear regression assumes a linear relationship between the values of the outcome and independent variables. In logistic regression, however, each continuous independent variable is assumed to have a linear relationship, not with the outcome values or the expected mean, but with the logit values (Hosmer and Lemeshow 1989:84-85, 88-91, McCullagh and Nelder 1989:107-109). Whether researchers in ecology are checking their data for linearity in the logit is not known because no mention of this assumption has been made in many papers reporting the use of logistic regression (see, e.g., Buehler et al. 1991, Burger et al. 1994, Diller and Wallace 1994, Larsen et al. 1994, Chandler et al. 1995, Coker and Capen 1995, DeLong et al. 1995, Drewien et al. 1995, Gorenzel and Salmon 1995, Nadeau et al. 1995, Hinsley et al. 1996, Kindvall 1996).

Checking the logit assumption of linearity could be done either before model selection begins or after a final model has been selected and further refinement of the variables is an issue. A variable with a distinct non-linear logit pattern could be excluded from models during the model building process based on *only* its non-linear logit form and *not* on its true association with the dependent variable. Thus, it seems that the best stage to check this assumption would be before the model selection process

begins. This way each variable is given the fullest opportunity to enter models in the selection process. Researchers should check the logit assumption at the univariate stage or at least decide *a priori* whether or not non-linearity in the logit is important from both a statistical and biological viewpoint; otherwise biologically relevant variables could be incorrectly excluded from models.

Hosmer and Lemeshow (1989) provide some suggestions on transforming or categorizing a continuous variable in order to overcome violations of the logit assumption. Several methods for examining the logit linearity assumption can be found in Hosmer and Lemeshow (1989:84-86, 89-91). One simple and graphical way involves looking at a plot of the logit versus the grouped values of the independent variable in question. Several steps are required to produce such a plot, including grouping the values of the independent variable into perhaps ten groups of approximately equal sizes (details are given in Hosmer and Lemeshow 1989:84-85, 90). A linear regression r^2 of approximately 0.7 or greater could be considered as sufficient evidence of linearity in the logit. Failure to meet such a criterion would then lead to consideration of possible transformations or categorizations of the variable. If simple and interpretable transformations do not produce linearity in the logit, then categorization of the variable in question can be conducted.

Hosmer and Lemeshow (1989) give some advice on the possible methods of categorization of independent variables, but some common sense can work well. The specific categorization of a continuous variable can be based on natural break-points in the data which might be found by

examining the logit plot. A variable not showing natural break-points could be split into three to five groups of roughly equal size, where each group is really a cell (or category level) in a contingency table. Examination of the proportions and odds-ratios across cells is then useful in determining the proper categorization of the variable (see Hosmer and Lemeshow 1989:97-98 for an example). Those cells with similar proportions of the outcome 'presence' and similar odds-ratios can be combined. Also, any cells with zero counts should be changed such as by combining them with non-zero cells in a meaningful way (Hosmer and Lemeshow 1989:84).

The second major difference in the assumptions between linear and logistic regression is that the model variance is assumed to equal one in logistic regression because of the binomial nature of the outcome (McCullagh and Nelder 1989:124-126). However, ecological binomial data may often show overdispersion (variance > 1) and variance should therefore be estimated. Researchers in ecology are most likely assuming model variance = 1 when conducting logistic regression analysis because no estimates of model variance are being reported (see, e.g., Brennan et al. 1986, Capen et al. 1986, Johnson and Temple 1986, Smith and Connors 1986, Diefenbach and Owen 1989, Buehler et al. 1991, van Manen and Pelton 1993, Burger et al. 1994, Diller and Wallace 1994, Larsen et al. 1994, Bartlett 1995, Chandler et al. 1995, Coker and Capen 1995, DeLong et al. 1995, Drewien et al. 1995, Gorenzel and Salmon 1995, McNay and Voller 1995, Nadeau et al. 1995, Hinsley et al. 1996, Kindvall 1996). A discussion of the merits of logistic regression in ecological research (Trexler and Travis 1993)

also fails to mention this assumption or what to do if model variance is greater than one. In addition, statistical software packages often used by ecologists apparently assume variance = 1 when conducting calculations and performing model selection (personal observations).

The variance assumption can possibly have major consequences because hypothesis-testing procedures often used in the model selection process do not incorporate information about overdispersion into the selection of an appropriate model. Researchers analyzing capture-recapture data also face this problem and some analysts have recommended that estimates of variance be incorporated into the process of model selection (see Lebreton et al. 1992). One way to incorporate an estimate of variance into the model selection process is to use informational model-selection criteria such as ICOMP or ICOMP-IFIM. A more detailed discussion of how to incorporate an estimate of model variance into the modeling process in logistic regression will be provided in a later section.

THE PROBLEM OF MODEL SELECTION

The classical approach vs. the informational approach

The evaluation of competing statistical models and the selection of a suitable model for description and/or inference can be viewed as a central part of any data analysis, not only for regression (see Sakamoto et al. 1986, Burnham and Anderson 1992). An important question is "What method or approach should be used for statistical model selection?". Ecologists typically answer this question by adopting statistical hypothesis-testing procedures.

With hypothesis-testing procedures an alpha value (the probability of a Type I error) is used to judge whether a statistical parameter based on the sample data is different from a hypothesized value of the parameter, given an underlying probability distribution for a sample. The null hypothesis states the value of the parameter against which the estimated sample parameter is to be compared. The alpha value provides some acceptable level of the probability (most often at 0.05) of incorrectly concluding that the value of the sample parameter is statistically different from the value stated in the null hypothesis when both are really the same value. Thus, the emphasis in the classical approach is on the testing of statistical null hypotheses.

Ecologists rely mainly on hypothesis-testing procedures and tests of significance, not only for univariate analyses, but also for statistical modeling of multiple regression and multivariate data. Some statisticians and researchers, however, have been drawing attention to the informational approach as a viable alternative to hypothesis-testing procedures (see Akaike 1973, Sakamoto et al. 1986, Bozdogan 1987, 1988a, b, 1990, 1994a, b, c). Burnham and Anderson (1992) and Lebreton et al. (1992) have recommended the use of such model-selection criteria, based on the work of Akaike (1973) and others, for statistical modeling of ecological data. Recall that the main points about the informational approach, as discussed in Part 3, are:

1. Statistical analysis is viewed as a process of evaluating various statistical models being fit to a given data set and selecting the best models according to the values of an informational criterion (e.g., see Sakamoto et al. 1986, Bozdogan 1987, 1988a, b).

2. Information-based criteria are used to evaluate each model's fit to the data and to provide a relative method of ranking and comparing models.

3. The models with the lowest numerical values of the criterion are the models which best fit the data at hand.

4. An informational criterion has two components, a lack-of-fit term and a penalty term.

5. The lack-of-fit term measures how poorly the given model fits the data; the smaller the value of this term the better the model fits the data. This term is calculated as -2(loglikelihood), where maximum likelihood estimation (MLE) procedures are used to estimate the parameter values under the given model.

6. The penalty term can be a multiple of the number of parameters estimated in the model (such as in Akaike's Information Criterion (AIC)), or a measure of the complexity of the model based on the model's covariance or correlational structure among the parameters.

7. The penalty term provides a way to balance problems of over- or underfitting the data and to adhere to the Principle of Parsimony.

A number of criteria, other than information-theoretic criteria, for the selection of the best model (or subset of independent variables) have been used in linear regression (see Hocking 1976:14-21). The *Cp* criterion first described by Mallows in the 1960s (see Gorman and Toman 1966, Mallows 1973, Hocking 1976) for linear regression has also been considered for model selection in logistic regression analysis (see Hosmer and Lemeshow 1989:121-125). In a way, statisticians and researchers using such criteria have demonstrated (either knowingly or unknowingly) that regression analyses are not statistical hypothesis-testing problems, but rather are problems in optimizing some criterion which estimates the fit of each

model to the data. This is exactly what informational statisticians have been advocating for many statistical analyses, not just for regression: that statistical analyses can often be viewed as a process of optimization rather than a process of testing statistical null hypotheses.

The informational approach, with its use of criteria and an optimization process, does have some advantages over hypothesis-testing procedures, as described previously in Part 3. These important advantages of the informational approach are that :

1. It performs well in a wide-variety of applications (e.g., see Bozdogan 1994a, b, c), without the need to use alpha levels, *P*-values, and statistical tables.

2. It allows for the comparison of non-nested models, unlike hypothesis-testing procedures.

3. It provides a straightforward way to help address the problem of overfitting of models to the data by directly incorporating a penalty term into model-selection criteria.

4. The results from the criterion values and the initial model selection process are fairly easy to interpret.

More detailed discussions of the informational approach can be found in Akaike (1973), Sakamoto et al. (1986), Bozdogan (1987, 1988a, b, 1990), Burnham and Anderson (1992), and Lebreton et al. (1992).

The objective of observational studies

An observational study is one in which the data were collected without the design and controls of a scientific experiment. James and McCulloch (1990:130-132, see especially their Figure 1) outlined the stages of a general research procedure to caution and remind ecologists of the fact that the

objective of any observational study is to produce *descriptive* models, not to provide strong inferences about causation as is often done. The first stage is to use observational data to produce descriptive models. Multiple regression and multivariate analyses are tools of exploratory data analysis which help to produce such models. Next, insight about causation (from various sources) in conjunction with descriptive models from observational studies should be used to develop causal models. Finally, the actual testing of causal models is then conducted with controlled experiments (or perhaps quasiexperiments, James and McCulloch 1990). The nature of observational data does *not* lend itself to making strong inferences about causation or sound predictions. Only data obtained from well-designed, controlled experiments can provide sound predictions or inferences about causation (James and McCulloch 1990).

Keeping this research procedure in mind, how should researchers analyze observational data? The answer depends partly on the total number of possible statistical models, or "model space". The model space is an exponential function of the number of independent variables and can be vast in many exploratory observational studies. For example, if a researcher measured 16 independent variables for analysis with logistic regression, then 17 total covariates exist (16 plus the intercept) and the model space is comprised of $2^{17} - 1 = 131{,}071$ possible logistic regression models. Twenty covariates places the model space over one million possible models. Regardless of whether one uses the informational approach or any other statistical approach, the only guaranteed way to find the best model is to evaluate all possible models ("exhaustive searching").

However, this can be time consuming and costly when large numbers of variables are analyzed.

A second point of concern is whether a single best model actually exists for cases of a vast model space. It is quite likely that no single model will be better than all other models (Gorman and Toman 1966, Hocking 1983, McCullagh and Nelder 1989:8). This point is further emphasized by some principles that the analyst should keep in mind according to McCullagh and Nelder (1989). First, all models are incorrect, but some models will be more useful than others (see Box 1979). The analyst should try to find the most useful models. Second, the analyst should not "... fall in love with one model to the exclusion of alternatives." (McCullagh and Nelder 1989:8). This is particularly true for vast model spaces when observational data are involved.

Observational data should be analyzed so that sufficient information and insight are obtained in order to construct useful descriptive models (remember the outline of the stages in a research procedure discussed by James and McCulloch 1990:130-132). The best way to produce good descriptive models in this research process is to examine a large number of models in order to gain more insight about variables and the data than by simply searching for a single best model. In addition, this wider view of the data could better assist researchers in developing causal models and designing experiments to help uncover causal mechanisms and relationships.

Unfortunately, the current practice of analyzing observational data does not often follow such guidelines and objectives as those outlined by

McCullagh and Nelder (1989) and James and McCulloch (1990). Too often analysts search the model space for a single best model, but the techniques commonly used cannot necessarily find a best model in a vast model space. A discussion of some searching and modeling procedures and their limitations are presented next.

Searching a vast model space and the limitations of current procedures

During the 1960s and 70s various computer algorithms were developed, such as stepwise procedures and branch-and-bound algorithms, to address the problem of searching a vast model space (see Hocking 1976, 1983 for a bit of the history). "Stepwise procedures or algorithms" here means any of the automated computer programs which use a stepwise selection (based on Efroymson 1960), forward selection (FS), or backward elimination (BE) process of adding or deleting one variable at a time to a model (see Hocking 1976, 1983 for specific definitions). In actual practice, the most common way that ecologists conduct model selection for either multiple regression or multivariate analysis is to use some type of stepwise algorithm in conjunction with hypothesis-testing procedures.

Stepwise algorithms and hypothesis-testing procedures are used to decide whether a given variable should enter or leave the specific model being examined. For example, stepwise procedures for linear regression often use *F*-tests. For each step in the analysis the *F*-ratio is calculated based on values of the residual sum of squares and the residual mean squares (or on incremental increases in R^2). A variable is added to the initial model at a given step ("forward selection" process) if that variable maximizes the *F*-ratio over those *F*-ratios of the other candidate variables. Some minor

differences in methods exist depending on whether the algorithm is a process of FS, BE, or stepwise regression. Interested readers can find further details in Hocking (1976:8-9), Myers (1986:117-122), and Sokal and Rohlf (1995:654-659) for linear regression.

Stepwise procedures for logistic regression use the likelihood ratio chi-square test rather than *F*-tests because error terms follow a binomial distribution (Hosmer and Lemeshow 1989:106). Regardless of whether a stepwise procedure is used for logistic regression, linear regression, or a multivariate analysis, the analyst must pre-select a critical value to set a "decision rule" for allowing a variable to be either included or excluded from the model.

Although stepwise procedures are commonly used they have serious limitations and are often misused (ecologists have been notably warned by James and McCulloch 1990:137-138 of some problems). One limitation of any stepwise algorithm is that it may find good models, but can easily miss finding excellent models or even the best *set* of models (Mantel 1970, Hocking 1976, Moses 1986). Indeed, the purpose of stepwise procedures for multiple regression "... is simply to find the smallest set of predictor variables that does an adequate job of prediction." (Sokal and Rohlf 1995:661). This is a potential limitation of any searching-type algorithm because only an exhaustive search can guarantee that the best models will be found. However, this limitation is a greater one for stepwise algorithms than some other searching algorithms because stepwise procedures search a "local" space rather than a "global" model space. Stepwise procedures add or delete one variable at a time and therefore search only along one or a

few of the many potentially useful searching paths. The total number of possible models (2^k - 1) that is actually evaluated with FS algorithms is only $k(k + 1)/2$, at most (Beale 1970, Mantel 1970). Stepwise algorithms that use both FS and BE procedures potentially evaluate more models, but still relatively few. Any stepwise procedure evaluates only a relatively small set of models out of the total possible number.

Another limitation is that the combined use of hypothesis-testing procedures and stepwise algorithms restricts the available searching space because such procedures can compare only nested models. For example, model "A" containing variables X_1, X_2, and X_4 and model "B" containing X_1, X_3, and X_4 are non-nested. Both models could be tested individually against the model possessing X_1, X_2, X_3, and X_4 because "A" and "B" are nested in this larger model. However, models "A" and "B" could not be compared to each other by using any hypothesis-testing procedure because neither model is a nested subset of the other.

The final results of a stepwise search can vary from one run to another depending on various conditions specified for the search which points out that a single best model may be an illusion. The final model obtained greatly depends on the pre-selected *P*-values and the type of stepwise procedure used (Myers 1986:122). Also, results of stepwise procedures are not consistent among analyses of the same data (Moses 1986:356-357).

Another method for variable selection and model building is one employing an interactive method, whereby subject-matter information and certain statistical techniques can be incorporated in the analysis (Henderson and Velleman 1981). For example, interactive methods for multiple linear

regression suggested by Henderson and Velleman (1981) can involve several or more techniques and steps in which the analyst plays an active role. First, exploratory methods (based on Tukey 1977, for example) are used to find potentially skewed univariate distributions and unusual patterns or data points. Second, the analyst can uncover possible nonlinear relationships by plotting y against each independent variable (remember, linear regression assumes a linear relationship between y and each x). Third, the information obtained from the first two steps form the basis for re-expressing (i.e., transforming or combining) variables "... in order to improve symmetry and linearity where this seems advisable." (Henderson and Velleman 1981:395). Fourth, correlations and partial correlations are used to find the better candidate variables for beginning the model building (maximum partial residuals and maximum absolute residuals can be used to aid this step). Fifth, partial regression plots of the better candidate variables are examined to help select which independent variable(s) to actually place in the model. Finally, after the final regression model is built, residuals and diagnostic measures can be examined. It should be noted that Henderson and Velleman (1981) outlined such steps as one possible approach to interactive regression, not as a formal, rigid methodology.

The advantage of the interactive approach is that decisions about model selection are made much more so by the analyst than the computer. However, some potential disadvantages of an interactive approach do exist, especially when used on observational data with many variables. In practice, the emphasis of interactive approaches is on selecting a single

final model. Again, this may be sufficient for experimental data, but such a goal should not be pursued with observational studies where the analyst really should explore many potentially good models. Though an interactive approach allows the analyst greater control of selecting variables and provides an in-depth look at a small group of models, it does not really evaluate many models for large data sets and does not permit an in-depth look at the *overall data or model space*. A second disadvantage is that interactive approaches often proceed in a "stepwise" manner in practice, even though automated stepwise procedures are not used, because the analyst tends to add variables to the model one at a time rather than in pairs or groups (personal observation). A large number of combinations of different variables or groups of variables are not examined. Although in-depth examinations are made of univariate regressions and relationships between pairs of independent variables, these are not always sufficient to determine how groups of three or more variables will work in concert together. The simple addition of one or two variables at a time might occur because of the time and effort it takes for the interactive analyst to examine the data and make decisions. Later, it will be shown how the GA-informational approach proposed here (and by Luh et al., submitted manuscript) can incorporate some of the initial steps of an interactive approach and also overcome these disadvantages by adding or deleting groups of variables to candidate models and evaluating a large number of candidate models.

Alternatives to the commonly used stepwise procedures should be sought after by ecologists. For example, when a relatively small set of

variables (12 or fewer) exists for an observational data set, analysts should take advantage of available computer technology and model-selection criteria (such as Mallows' C_p, AIC, ICOMP, etc.) to initially evaluate all possible models. Then, additional statistical examination and biological knowledge can help obtain a set of "best" models or a list of good variables to provide insight for constructing preliminary causal models to be tested by experiments.

When a large number of variables exist, analysts could use algorithms based on the work of Furnival and Wilson (1974) to obtain sets of models rather than just a few models. Such algorithms, often called branch-and-bound (or leaps-and-bound or best-subsets) algorithms, search for the "best" models of a given size (number of variables or estimated parameters). For example, if 20 independent variables exist for the regression problem, then the analyst could obtain the five best models for each subset having 19 variables, 18 variables, 17 variables, and so on. One problem with the branch-and-bound algorithms is that hypothesis-testing procedures and simple criteria are often employed. Some analysts use R^2 measures as the criterion to select the best regression models with these algorithms, but one can use Mallows' C_p instead (as is preferred for logistic regression by Hosmer and Lemeshow 1989). The use of informational criteria with branch-and-bound algorithms has been fairly unexplored. Another difficulty is that branch-and-bound algorithms are not always available for use with every technique in many software packages. For example, some software supports a branch-and-bound (best subsets) routine for linear regression, but not for logistic regression. Analysts must take additional

time and steps to apply the best subsets linear regression software to logistic regression by using the application suggested by Hosmer et al. (1989; see also Nordberg 1981, Nordberg 1982, Hosmer and Lemeshow 1989:118-126).

In summary, the methods commonly used to find a single best model to describe a multivariate observational data set are inappropriate for such purpose for several reasons. First, stepwise algorithms, because of their searching methods and combination with hypothesis-testing procedures as described in this section, search only a limited part of the overall model space. Second, analysts typically conduct only one or two stepwise runs on a data set even though the results of stepwise searching can vary from one run to another. Third, it is *extremely unlikely* that a single statistical model is clearly superior to all other models for any complex observational data set (Gorman and Toman 1966, Hocking 1983, McCullagh and Nelder 1989). Finally, finding and relying upon only a single model for a multivariate observational data set is not in keeping with scientific research goals. That is, the analysis of observational data should obtain the best insight into such data by obtaining descriptive models. Then, the observational models can be used, along with other information, to propose possible causal models that can be tested by experimental methods in order to provide possible causal explanations (James and McCulloch 1985, 1990).

Analysts should strive to obtain a wider view of any multivariate observational data than is provided by the current use of stepwise algorithms. Also, valuable methods and searching algorithms should not be restricted to being used only with hypothesis-testing procedures or to

being available in software packages for only certain techniques. Recently Luh et al. (submitted manuscript) proposed combining the informational approach with a genetic algorithm (GA) as a reliable method to search a vast model space. Their application was to linear regression, but their method can be extended to any multivariate modeling problem. In the next section, an introduction to a GA and its application to logistic regression modeling are provided, along with an explanation of how the GA can produce a large *set* of the best models for examination and how to interpret various results.

THE GENETIC ALGORITHM AND ITS APPLICATION TO STATISTICAL MODELING OF OBSERVATIONAL DATA

Overview of a simple genetic algorithm

The work of John Holland and his students during the 1960s and early '70s laid the groundwork for what are known as genetic algorithms (GAs; see Goldberg 1989, Ch. 4 for a general history of the development and the early application of GAs). Much of this early work was then formally published by Holland (see the new edition, Holland 1992a, of his original work) showing that GAs are optimization algorithms useful for solving complex problems. Many works on GAs have been published since 1975 and GAs have been applied to problem solving and optimization in such diverse fields as engineering, game-theory, political science, artificial intelligence, image processing and pattern recognition, biology, business, economics, and the social sciences (Goldberg 1989:125-142).

A GA uses concepts of genetics, natural selection, and evolution to search for the best solutions to a specific problem for which a vast number

of possible solutions exist (Holland 1992a, b). Many different types of GAs exist, but a simple GA is the type used in this dissertation. A simple GA has the following features (see Goldberg 1994:113):

1. 'strings' composed of binary codes and of a fixed length,
2. a population of strings which is finite in size, and
3. three basic operators which are selection, crossover (recombination), and mutation.

In a simple GA, a solution or answer to a specific problem is represented by a 'string'. Each string is rather analogous to a chromosome because both contain information coded along its length as a series of units. The chromosome's information units are genes, whereas the strings units are called 'bits'. The location of a bit can be referred to as a 'locus', analogous to a genetic locus on a chromosome. Each bit has a binary code, in any simple GA, whereby information is simply coded in an either-or manner (e.g., "0" or "1"). The *combination* of the binary information of all the bits on a given string forms a specific solution to the problem that is being solved. Simple GAs have strings of constant and equal lengths (fixed-length property).

In any complex problem, the quality of the solutions will greatly differ. Some strings will have excellent solutions, others will be of average quality, and still others will be of poor quality. The quality of a string's solution is analogous to the fitness value of an individual organism in a population. Indeed, the entire set of all possible solutions to a problem can be viewed as a "fitness landscape". The distribution of fitness values of strings determines the topology of the fitness landscape. Some simple

problems will have a single best solution and thus, a single fitness peak in the landscape. Many complex problems, however, will have different solutions of equal (or approximately equal) quality and therefore, will have a fitness landscape of many peaks and valleys.

A GA provides a probabilistic strategy of searching for the strings with the highest quality solution or fitness in any fitness landscape. A simple GA starts with a randomly chosen, initial group of strings that compose the starting 'population' of potentially interbreeding strings. Each subsequent population has the same number of strings as the initial population. Then, the GA searches for the best solutions in the solution space (or fitness landscape) by using concepts of natural selection, recombination of information (crossing-over), and mutation.

This searching process can be viewed as the evolution of a population of solutions such that the composition of the population changes over time and that some solutions of high fitness values appear during the history of the evolving population. High fitness solutions will be found by the GA because variations exist in the fitness values among solutions and the natural selection operator acts on this variation. Variation continues to exist from one generation of solutions to the next because of the recombination (mating) and mutation of solutions that takes place in the GA. Thus, the population of solutions "evolves" and *solutions with high fitness values will appear during the history of the solutions without all of the possible solutions having to be evaluated by the algorithm* (Holland 1992a, b, Goldberg 1994).

How exactly is a simple GA conducted? First, the algorithm randomly generates each binary-coded bit for a string (a solution). The user of a simple GA sets the number of bits based on the amount of information needed by a string to provide a real solution to the given problem. Next, the GA continues to randomly generate one string at a time until a preselected number of strings is obtained, thus forming the initial population. The quality of the solution of each string in the population must then be evaluated based on a defined fitness function. For example, say that the problem at hand is to find the shortest total distance traveled (or time spent travelling) by a traveling salesperson who must routinely visit clients who are spread out over a wide area (i.e, the "traveling salesman" problem). Each string would represent the sequence of clients (locations) in order of their visitation by the salesperson. The fitness function would be the total distance traveled over a given sequence of clients (route). In this example, higher fitnesses would actually be the lower distances traveled. The next several steps in a simple GA involve the three important operators: selection, recombination (crossover), and mutation.

The selection process chooses the better solutions, based on their fitness values, and then the recombination or crossover operator "mates" those strings (i.e, recombines their information) in order to produce a new generation with potentially better solutions. Different methods exist for conducting the selection process and forming the mating pool. One method is to calculate the mean fitness value of the population which then represents a cutoff point. Then, only those strings whose fitness values are

at or above the cutoff point are selected for the mating pool. Another method is to subtract the fitness value of each model from the highest value in the population. Next, the average of these differences and the ratio of a model's difference value to the mean difference is calculated for each model in that population. Then, the GA selects those models whose ratio is greater than one to form the mating pool. Regardless of the selection method used, the objective is to choose a group of strings with the higher quality solutions (fitness values) in the present population to form the mating pool.

The next step in a GA is to conduct the mating of the selected strings via a recombination operator. Sexual reproduction is a mechanism that can produce considerable genetic variety among the offspring of any two parents. Such variety is due to genetic recombination that takes place during meiosis (to form gametes) and zygote formation when one egg and and one sperm unite out of all the numerous possible combinations of genetically different gametes. This concept provides a model for genetic recombination, but in practical terms the actual mating or recombination process in a simple GA is performed like the crossover process between homologous chromosomes during meiosis.

The recombination process starts by choosing a pair of parents from the mating pool. For any parent pair, a point between two adjacent loci on the strings is randomly selected as the crossover point. Often in GAs, each point between such loci has an equal probability of being chosen, unlike the crossover probability found among real chromosomes. The strings of the two parents, are broken into two pieces at the same crossover point. Then,

each piece downstream of the crossover point is reconnected to the upstream piece on the other parent. For example, say that parent string *A* has the binary sequence 1 1 1 0 0 and parent *B* has 1 0 1 0 1. The GA might randomly choose the crossover point as the location between locus 3 and 4 and the parent strings each break apart at that point. The recombination process would then combine the 1 1 1 piece of parent *A* with 0 1 of parent *B* to form the offspring *A1* string as 1 1 1 0 1. Likewise, offspring *B1* would be 1 0 1 0 0 because 1 0 1 of Parent *B* combined with the 0 0 piece from parent *A*. Because simple GAs have strings of fixed length, the parent strings must break at the same point here and the offspring are always the same length (number of bits or loci) as the parents. The process of recombination continues to occur as new pairs of parents are chosen one pair at a time, until the number of strings in the second population (new generation) is equal to that of the initial population. GAs could have population sizes varying or increasing, but for simplicity this discussion will consider population size to be constant from one generation to the next.

Some flexibility exists with the recombination process because the analyst can program a specific crossover rate (i.e., the percentage of times that crossing-over actually takes place in a population) into the GA. A crossover rate of 1 means that crossing-over (mating) occurs in 100% of the pairings between parents and the subsequent generation is composed of only offspring strings. A rate of zero means that no crossing-over occurs between any parents, thus, the next generation is formed solely of members of the mating pool. The higher the crossover rate the more likely that the population in each generation consists of new offspring strings and

therefore, new solutions are introduced into the searching process. Obviously, tradeoffs exist because a high rate of crossover runs the risk that good strings from the current population do not become part of the next generation. However, low crossover rates run the risk of slowing down the search because too few new strings are produced.

Recombination among strings in the mating pool of the initial population produced the second population (generation). This second population can consist entirely of offspring strings or of a mixture of offspring and parent strings, depending on the specific crossover rate set by the analyst. Each subsequent population of new strings is produced from the previous population in the same manner that produced the second population from the initial population. It should be noted that the analyst chooses and sets the number of generations that the GA performs.

Selection and crossover are not the only means by which new strings are produced. Mutation of strings in a simple, binary-coded GA can produce new strings by simply changing the code of a single bit. For example, a string with the binary code 1 0 1 0 1 could have a random mutation occur at locus five so that this string now becomes 1 0 1 0 0. In simple GAs, a site for mutation is chosen randomly along a string. Any string is subject to mutation, but the analyst specifies the probability or rate (usually low) of mutation. Mutation is useful in a GA because it can produce diversity among strings and help the searching/sampling process 'jump' out of a particular area in the fitness landscape and perhaps into a yet unexplored area.

Overall, the operators in a simple GA proceed similarly to processes occurring in biological populations. Strings with the higher fitness values produce more offspring than strings with lower fitnesses by means of selection and recombination. A consequence of the selection and recombination processes, especially when conducted over many generations, is that good information from the better parent strings is combined to form potentially better offspring solutions to the problem.

A GA-informational modeling approach for logistic regression

The selection of appropriate statistical models out of a vast model space is an optimization problem and the combination of a GA and an informational model-selection criterion can be a potentially effective method of solving this problem. The combined use of a GA and an informational criterion for statistical model selection was described by Luh et al. (submitted manuscript) for linear regression. Recently, Bearse et al. (1997) and Bearse and Bozdogan (1998) have applied the basic GA methodology of Luh et al. (submitted manuscript) to other regression techniques.

The basic approach of Luh et al. (submitted manuscript) is presented here, but also provided are both specific suggestions for modeling complex observational data and recommendations for applications to logistic regression which Luh et al. (submitted manuscript) did not address. The general aspects of the genetic algorithm-informational modeling, or GAIM, approach outlined here can be used for any multivariate analysis, not just for logistic regression.

Two important problems must be addressed by the analysis of observational data sets where a vast model space exists. First, the analyst must find the best models out of the numerous possible models. Second, the analyst should examine a wide "field" of potentially useful models, rather than search for a single best model. The GAIM approach suggested here can address those two problems.

In the GAIM approach, a statistical model is represented as a string of bits (i.e, a combination of independent variables). Each string represents a unique combination of variables. The GA then searches for the best models, mimicking aspects of genetics and biological evolution as previously described, while using an informational criterion as the fitness function.

Before proceeding with the actual GA for logistic regression, the analyst should preview univariate logistic regression models for each independent variable. In addition, the assumption of linearity in the logit should be checked by examining plots as described earlier (also see Hosmer and Lemeshow 1989:84-86, 89-91). Any continuous variable that does not meet this linearity assumption should be transformed (and thus remain continuous in scale) and re-checked for linearity. If transformations do not satisfy the linearity assumption, then the continuous variable can be converted into a categorical variable following as described earlier (see Hosmer and Lemeshow 1989:97-98). All categorical variables should be checked for any cells with zero or low counts. Any necessary changing of categorical variables or combining of categories can be made in statistically appropriate and biologically substantive ways.

One difficulty with a categorical variable is the proper reading of the design variables in the GA code. A dichotomous categorical variable is handled easily by the GA because only one design column in needed in the data (X) matrix to code for such a variable (as is the case for any continuous variable as well). However, categorical variables with three or more levels require some care in the GA. For example, suppose a data set has three independent variables which are continuous in scale (X_1, X_2, and X_3), but the fourth variable (X_4) is categorical with three levels (see Table 4-1). The three levels might represent age classes of subjects (e.g., juveniles, subadults, and adults) or habitats (meadow, deciduous forest, and evergreen forest). Three levels requires the coding of two design (dummy) variables in the data matrix. Observations would be coded as 1 0 in the first category level, as 0 1 in the second level, and 0 0 in the third level. Whenever X_4 is used in the analysis the GA code would have to be written so that both columns representing the design variables would be read and entered into the model.

The analyst should also examine correlations between independent variables or examine logistic regression models consisting of pairs of variables. This would be done to determine whether two independent variables which were highly correlated would pose problems for matrix operations. It could also provide some insight into which variables might have similar and redundant information. The GA code can be written to prevent two or more highly correlated variables from entering the same models together. The code would allow one such variable to enter a model

if randomly chosen, but then prevent the correlated variable from subsequently entering that model.

With any multivariate technique analysts should likewise check basic statistical assumptions and examine plots of the data before actually running the GA. The form of the variables should be checked and any categorization or transformation of variables should be performed before starting the GA runs. All checks of this nature can be performed using various exploratory techniques and plots. Some hypothesis-testing procedures can be used also for inspection of simple models with one or two variables, but I prefer using informational criteria for these analyses.

The first step in the actual GA for statistical modeling is to encode the various possible combinations of independent variables in order to correctly represent the possible models in the model space. This means that each logistic regression model is encoded as a string of zeros and ones. Each locus or bit on a string represents either the presence (1) or absence (0) of a particular independent variable. The left-most locus (the 'starting' locus or locus 0) represents the intercept term. Locus 1 (the locus immediately to the right of locus 0) represents the independent variable X_1, locus 2 represents X_2, locus 3 represents X_3, and so on through the pth independent variable (where p = the total number of independent variables). In this encoding scheme each string has the same length (number of loci = $p + 1$), but has a different sequence of binary codes to represent a unique model. For example, consider the data in Table 4-1. The string of 1 1 1 1 1 represents the model consisting of (in sequence from left to right) the intercept term, X_1, X_2, and X_3 (the first three continuous

variables), and X_4 (the categorical variable). The string of 1 0 1 0 1 represents the model consisting of the intercept term, X_2, and X_4.

The next step is to randomly generate the bits for each model one at a time. Each bit in a model is randomly assigned either a '0' or a '1'. This process of randomly generating models continues until the prespecified number of models (population size) is reached to form the initial 'population'. An optimal population size is not known for the use of a GA for statistical model searching. Such a size likely depends in part on the complexity and size of the data set. It is suggested that analysts choose a moderate population size of 30-75 models until further research sheds insight on this issue.

The fitness of each model is then determined by calculating the model's criterion value. Informational criteria, such as AIC, ICOMP, and ICOMP-IFIM, are useful measures of a model's fit to the data, can be used with any multivariate technique, and are highly suitable to model selection in conjunction with a GA. On the other hand, hypothesis-testing procedures are neither easily used nor desirable for use with a GA. For logistic regression, any informational criterion can be used as the fitness function. However, recall that the analyst should include an estimate of the model's variance or adjust the criterion based on model variance because logistic regression models assume that variance = 1, but ecological data may often exceed this assumed variance. One method to include an estimate of variance in the criterion value for logistic regression is discussed later.

Another important technical point regards calculation of the fitness value whenever a categorical variable with two or more design variables is in the model. Suppose, for example, that the model 1 0 1 0 1 occurs in the initial population for the data example in Table 4-1. Variable X_4 at locus 5 represents a categorical variable with two columns of design variables in the data matrix. The GA code must read both of the columns of those design variables in the data matrix, along with the column for the intercept and the column for X_2, in order to properly calculate the criterion value for model 1 0 1 0 1.

The next step is to select the mating pool based on the fitness (informational criterion) values of all models in the initial population. A GA used by the author (and written by Hang-Kwang Luh) selects the mating pool based on the following procedure :

1. for each model in the current population, subtract the criterion value from the highest criterion value to obtain a 'difference' value,
2. calculate the average of these difference values,
3. divide the difference value of each model by the average difference to obtain the 'difference ratio', and
4. select only those models with difference ratios > 1 to enter the mating pool.

The analyst does not know the actual criterion values that will appear in different populations. Thus, standardized rules and numerical values across all generations in the GA, such as those described above, are needed for efficient selection of the mating pool. The rule of selecting only those models with difference ratios > 1 for the mating pool work well for

informational criteria because the lowest criterion values represent the best models. This rule would have to be changed if the model-selection criterion being used had large numerical values associated with the best models. The number of models selected to be in the mating pool equals half of the population size. This helps guarantee that the subsequent population will be the same size as the current population because any mating between two models produces only two offspring models.

The chance that a given model in the mating pool actually 'mates' (undergoes the crossing-over or recombination process) is directly proportional to its difference ratio. For example, a model with a ratio = 8 is four times more likely to mate with another model than a model with a ratio = 2.

The rules and methods of the overall selection process provide that the better the fitness (i.e., the lower the informational criterion value) then the greater the probability a model is selected to be a parent model (i.e., selected for both the mating pool and the actual mating process). In this way selection of parent models mimicks the natural selection process in biological populations.

Two models are chosen to mate with each other based on these rules and methods of the selection process. Mating is performed by means of the crossover (recombination) operator based on the process of crossing-over between two homologous chromosomes during meiosis, as described previously for a simple GA. The method used in the GAIM approach is a simple one-point breaking and crossing scheme in which a single point

between adjacent loci is chosen at random as the break point. Other schemes could be used whereby two or more crossover points are chosen.

After the first two models mate two more models are selected out of the mating pool for mating. Note that either of the first two parents are still eligible to be chosen again as a parent. The process of selecting parent models from the mating pool is a lottery-type process in which subsequent drawings are performed with replacement and the chance of being chosen is always proportional to a model's difference ratio. Recall that any two parent models do not always mate unless the crossover rate = 1. A crossover rate of 0.7, for example, means that the chance of mating between any two chosen parents is 70%. If the two parents do not mate, then they themselves are placed in the next generation since no offspring were produced. The process of drawing parent models and producing offspring models continues until the new population is equal in size to the initial population.

The GA then calculates the criterion values for each model in the new population. The selection and mating process is then conducted just as it was for the initial population and another new offspring population is produced. These processes continue one generation after the other until the GA has produced the total number of populations set by the analyst at the beginning of the run. Thus, the GAIM approach selects the better models in a given population, recombines their information (variable combinations) to form new (offspring) models, and then continues these processes over a number of generations. Thus, the GA, in conjunction

with an informational model-selection criterion, is able to sample or search for the best models in the model space.

The GAIM approach does not simply end here with the selection of the model with the lowest criterion value as "*the best*" model for the data. For any large observational data set it is quite likely that a single best model (i.e., one that is far superior to all others) does not exist (e.g., see McCullagh and Nelder 1989:23). Thus, one objective of the GAIM approach is to use the GA as a sampling mechanism so that *many* models can be initially examined from the vast model space. The informational criterion serves as an *initial* model selection criterion.

The analyst should first sort the output file from the GA according to the values of the criterion (and secondarily by the number of estimated parameters, k, in the model). Next, the criterion values should be examined to see just how close models are to the actual lowest value found and how close each model is to the models ranked immediately above and below it. One is reminded that the *differences* between AIC values of the candidate models are important, not the actual criterion values, and that differences larger than 1-2 between two models are statistically relevant (Sakamoto et al. 1986). Thus, a difference in AIC of <2 suggests that the two models are statistically equivalent. Analysts should adopt similar guidelines, along with common sense, with other informational criteria.

One scenario that could occur for a vast model space is that one model truly has a much lower criterion value than all other GA models. Before accepting this model which has the minimum criterion value (Model 1) as the "best model" from the GA, the analyst should perform a few checks.

The first check would be to see whether Model 1 is a nested subset of some of the good GA models which had the next lower criterion values. If not and these other good models have more estimated parameters than Model 1, then the analyst should delete one or two variables from the good models to see whether the new smaller models have criterion values equal to or lower than Model 1. For example, suppose Model 1 from the logistic regression analysis using a GA has the variables X_0, X_1, X_2, X_3, X_5, and X_6 (k=6) and Model 2 has the variables X_0, X_2, X_3, X_5, X_6, X_7, and X_8 (k=7). Model 1 is not nested in Model 2 so the analyst should calculate criterion values of subsets of Model 2 which lack X_7 or X_8 (as well as both of these variables) and compare the values to that of Model 1.

A second check to make, when one model appears to have a much lower criterion value than all other GA models, is to use logistic regression diagnostics on Model 1 and perhaps some of the other models. To balance and complete the analysis, the analyst can use the diagnostics to further examine the fit of Model 1 (and perhaps a few other good models) and gain insight into how well the model actually fits the data. The use of logistic regression diagnostics will be discussed later in this section.

Other scenarios of results from observational data will occur in which many candidate models may often have similar criterion values. Therefore, the analyst should take a different strategy from the usual viewpoint of simply selecting "one best model". The GAIM approach includes taking additional steps such as examining the relationship between the criterion values and k, as well as reporting the frequency distribution among a set of "best models".

The first step to take is to define a *set* of best models obtained from the combined runs of the GA. One way is to define this set of the "best GA models" is to choose those models which have criterion values within 2-3 of Model 1 (the minimum criterion model from all GA runs combined). Another way is to take the criterion value of each Model 1 from each GA run, calculate the median criterion value of the Model 1s, and then obtain the best GA models as those models within 2-3 units from this median value. The range of 2-3 as the cutoff point in these cases allows for a more conservative and wider view of the data (but still includes good models) than perhaps a tighter cutoff defined as models within 1-2 units of the minimum criterion model.

The next step after defining the best GA models is to examine these models for possible trends. Some trends to be examined are the relationships between k (the number of estimated parameters in a model) and the lack-of-fit terms, penalty terms, and informational criterion values of the best GA models. This can be accomplished using box plots or other graphical representations.

For example, suppose the best GA models for a particular analysis (defined by one of the methods mentioned previously) contained 100 models ranging from $k = 5$ to 9. One possible outcome would be that no distinct trend existed between k and the criterion values in this set of best GA models. This situation might suggest that deletion of some variables from the larger best models does not alter the fit of the smaller best models to the data. If the analyst found that the models with $k = 5$ or 6 were indeed subsets of the models with $k = 8$ or 9, then the smaller models might be the

better group of models. Any number of trends between the criterion values and k can occur. The point is that the analyst should examine the best GA models for any trends to gain better insight into the data.

In all cases, the analyst should also examine the estimated variance of the logistic regression models. The model-selection criterion can be used as a primary criterion, but the estimated variances could be incorporated into the penalty term of the criterion and/or used as a secondary criterion. For logistic regression models, the estimated variances of the best GA models could be examined for potentially large variances and used as a secondary criterion. Some models even among the best GA models could possibly have unreasonable variances (such as variances > 3; see Lebreton et al. 1992 for comments about large model variances in multinomial models). The analyst could further exclude any models which have larger variances from the best GA models.

Another step to take in order to gain insight into the data is to construct a frequency distribution of the variables which occur in the best GA models. The percentage or proportion of models in which a variable occurs is plotted by the list of variables in the analysis. The analyst can use this wide view of the data to identify a group of potentially useful variables around which causal models could be proposed. This group of most frequent variables in the best GA models, the analyst's knowledge of the biological meaning of these variables, and information from other studies can then be incorporated into the building of causal models and the design of experiments to test the causal models. Thus, the frequency distribution of variables among a set of best GA models provides some of the most

useful information because it gives the analyst a wider view of the data than could be obtained from the conventional use of stepwise procedures.

The criterion values tell the analyst which models best fit the data at hand, but do not confirm exactly how well a given model fits the data across all of the observations. The use of logistic regression diagnostics such as ΔX^2, ΔD, Δß, and h (see Pregibon 1981, Hosmer and Lemeshow 1989:149-157) can certainly play an important role in model checking, verify how well a model fits the data, and be used to compare specific aspects of the fit of competing models. If the set of best GA models has more than 10-20 models, then it may be cumbersome to obtain and scrutinize the diagnotic measures for all the best models. However, a few of the best GA models could be examined to verify that the models do fit the data well.

For logistic regression, Hosmer and Lemeshow (1989) recommend obtaining the diagnostic measures for each covariate pattern rather than for each observation when the number of covariate patterns (J) is much less than the sample size (n). Their recommendation makes good sense and analysts should check different software to see whether the diagnostics are calculated for each observation or for each covariate pattern. The LR (stepwise logistic regression) routine in BMDP software (Engleman 1988), for instance, calculates the building blocks for the diagnostics for each covariate pattern. Users can take the output from this routine and put it into a spread sheet, calculate other diagnostic measures from that output, and graphically display and interpret the results. Hosmer and Lemeshow 1989:158-168) provide some guidelines on the interpretation of logistic regression diagnostics.

However, if the analyst finds that the model(s) is(are) not correct, based on the diagnostics, other measures which assess the goodness-of-fit of a model, or biological knowledge, then other models could be carefully considered. Hosmer and Lemeshow (1989:168-170) discuss conditions which the analyst should consider in such cases for logistic regression models. It is possible that one or more biologically important variables were not measured in the study. The analyst can do little in this situation to correct this problem because retroactive collection of data is often unrealistic.

Another possibility for explaining the poor fit of models is that the identification of the scale of certain variables is insufficient. Recall that during the early stages of analysis the assumption of linearity in the logit should have been checked and variables violating this assumption were then transformed or made into categorical variables. Also, categorical variables were checked for zero cells and non-significant cells and possible rescaled. Imprecise scale selection of these variables could potentially contribute to a lack of fit in the models (Hosmer and Lemeshow 1989:170). More precise methods for improve scale selection than those discussed in this text have been published, but the methods can be computationally difficult (see Hosmer and Lemeshow 1989:170 for references to these methods). Hopefully, the analyst's biological knowledge will help to identify biologically relevant variables during the design phase, to collect the data in an accurate way, and to produce reasonable scale changes in troublesome variables so that additional procedures of scale selection are not needed.

Another condition to consider is that extra variation might be contributing to the lack of fit. Inclusion of the estimated variance into the penalty term of the criterion might help reduce the chance that the best models exhibit problems of this kind, but it is certainly no guarantee. Variances of the best GA models should be inspected as suggested previously. Some methods to handle the extra variation incorporate additional parameters into a model and interested readers should see Hosmer and Lemeshow (1989:170) for references.

In order to incorporate an estimate of variance into a model-selection criterion for logistic regression, the criterion employed is ICOMP-IFIM defined by Bozdogan (1990, 1994d; and using "2" times the penalty term as suggested by Bozdogan and Haughton 1998) as:

$$\text{ICOMP-IFIM} = -2\ln L(\hat{\theta}_k) + 2[C_1(\hat{F}^{-1})], \tag{4.6}$$

where: $\ln L(\theta_k)$ = the maximum loglikelihood value when maximum likelihood estimation (MLE) methods are used to estimate the parameter values of the model

ln = the natural logarithm,

$C_1(\hat{F}^{-1})$ = the maximal informational complexity of the estimated inverse-Fisher information matrix (F^{-1}).

The first term in equation (4.6) is a measure of the lack of fit of a given model to the data at hand. This term is the same one found in AIC for logistic regression and was defined in equation (4.5). The second term of

ICOMP-IFIM is the penalty term in the form of a scalar measure of the informational complexity of the inverse-Fisher information matrix. The complexity term in ICOMP-IFIM is defined as (Bozdogan 1990, 1994d):

$$C_1(\hat{F}^{-1}) = (r/2)\ln[\mathrm{tr}(\hat{F}^{-1})/r] - (1/2)\ln[\det(\hat{F}^{-1})] \tag{4.7}$$

where: r = rank or dimension of $\hat{F}^{-1}$,

$\hat{F}^{-1}$ = the estimated inverse-Fisher information matrix,

tr = the trace of a matrix, and

det = the determinant of a matrix.

It is in the calculation of $\hat{F}^{-1}$, and therefore in the penalty term, that the estimate of model variance can be incorporated.

The estimated inverse-Fisher information matrix is a block diagonal matrix defined for logistic regression (Bozdogan, personal communication) as:

$$\hat{F}^{-1} = \begin{bmatrix} (\mathrm{var})(\widehat{\mathrm{Cov}(\beta)}) & 0' \\ 0 & 2(\widehat{\mathrm{var}})^2/n \end{bmatrix} \tag{4.8}$$

where: $\widehat{\mathrm{Cov}(\beta)}$ = the estimated covariance matrix of the maximum-likelihood estimated logistic regression parameters,

0 = a k by 1 vector of zeros (and k = the number of estimated parameters in the model),

0' = a 1 by k vector of zeros,

$\widehat{\mathrm{var}}$ = the estimated model variance, and

n = the total number of observations.

The estimated covariance matrix of the parameters is calculated as the inverse of the matrix **X'VX** (using formal notation, $\widehat{Cov}(\hat{\beta}) = [\mathbf{X'VX}]^{-1}$). The matrix **X'VX** is called the information matrix (Hosmer and Lemeshow 1989:28), but is not the same thing as the Fisher information matrix. **X** is just the data matrix itself containing the observed data for all n subjects or observations. The **X** matrix also contains, as its first column, a column of ones to represent the intercept term for each observation. **X'** is simply the transpose of **X**. The **V** matrix is a diagonal matrix made of n columns and n rows, in which the diagonal elements are $\pi_i(1 - \pi_i)$ for $i = 1, 2, 3, ..., n$ (see Hosmer and Lemeshow 1989:29).

This approach with ICOMP-IFIM allows the analyst to incorporate model variance directly into the penalty structure of the model via the F^{-1} term. For those models with the same covariance structures, models with a larger variance will generally have a larger penalty term. In addition, those models with lower correlations or associations among model parameters generally have less complex covariance structure and therefore lower penalty terms (Bozdogan 1990, 1994d). Thus, ICOMP-IFIM gives consideration to both the model variance and the degree of multicollinearity among model parameters in the evaluation of each candidate model. If the analyst used hypothesis-testing procedures, then such considerations would have to be done as an analysis separated from, rather than intimately part of, the comparison of candidate models.

Variance can be estimated (as its biased form) for each candidate model and incorporated into the calculation of ICOMP-IFIM as:

$$\widehat{\text{var}} = \text{Pearson's chi-square goodness-of-fit statistic}/n$$

$$= \frac{\sum_{i=1}^{n} \left([y_i - \pi(x_i)]^2 / \pi(x_i)[1 - \pi(x_i)] \right)}{n} \tag{4.9}$$

The notation follows that of equations (4.1) and (4.5). The division by n to give a biased estimate is in keeping with the idea that some bias exists in the estimated parameters (Bozdogan, personal communication). Other methods of calculating model variance could be used (e.g., dividing Pearson's statistic given above by either $(n - 1)$ or the number of covariate patterns in the model). The analyst should decide which method best serves the data and modeling goals.

Incorporating the estimated variance into ICOMP-IFIM is by no means the only way to consider extra variation or overdispersion in logistic regression models. Estimated variance could potentially be incorporated into other informational criteria (such as a possible variance inflation factor into AIC). However, model variance would not be so easily incorporated into standard hypothesis-testing procedures used for model selection. Researchers using logistic regression should evaluate model variance and report the method they used to incorporate such variance into the model selection process when models have variances exceeding one.

Practical matters

A few issues of practical importance should be mentioned regarding the use of a GA and informational criteria for model selection. First,

because one objective of the GAIM approach is to obtain insight from a potentially large set of models, the GA's output file can be rather large depending on the pieces of information from each model that the analyst saves. For example, storing the estimated parameter values of each model obtained from the GA could potentially make an output file large and cumbersome for subsequent analysis. This of course depends on the computing resources available to the analyst. If large output files are a potential problem, then the analyst could save a simple output file consisting of the following for each logistic regression model obtained in the GA: the binary coding, the -2(loglikelihood) term, the number of estimated parameters, the criterion value, and the estimated variance. An alternative would be to save only the models (along with their ßs and other output) which have criterion values below some cutoff. This cutoff value could be based on the criterion of a good model that was obtained from either the analyst's expectations of the data, a stepwise analysis, or a quick-and-dirty GA.

Another practical point to consider regards being able to easily identify the models and their variables in the output. The analyst can convert the binary coded bits of each model in the output file to easily recognized labels for the variables, such as either X_1, X_2, X_3-type representations or abbreviated names. The models and their summary information could then be read into a spreadsheet or graphics package in order to produce desired tables and/or graphs.

Third, Wald statistics and the associated *P* values for parameter values (regression coefficients or estimated ßs) of candidate logistic regression

models do not need to be calculated by the GA. However, such values and statistics can be obtained for the best GA models by using standard statistical software. Statistical hypothesis-testing procedures could be used to supplement the GAIM approach, perhaps in the context of diagnostic analysis and model checking, particulary when presenting the results to audiences unfamiliar with the informational approach to statistical modeling.

Lastly, an important practical matter regards obtaining the actual code for a GA since statistical software commonly used by ecologists does not have such codes. The code for a simple GA and the calculation of the informational criterion can be written in any one of the many computer languages or even by using some software packages which have many defined functions to make programming easier. One key is to be sure the language or software can perform the matrix operations (such as obtaining the determinant, trace, and inverse of a large matrix) needed for calculating model-selection criteria and possibly manipulating data matrices.

Public-domain codes do exist for various GAs (see Goldberg 1994:115 for sources of such codes). These public-domain codes would no doubt have to be modified and linked with programs to calculate model-selection criteria. Analysts should also see Davis (1991) for discussions and guidelines on how to apply GAs to problems of optimization. Ecologists could collaborate with statisticians and computer scientists in order to develop GAs and build upon the ideas presented here. GAs are becoming extremely popular and useful in many disciplines. Many computer scientists are familiar with GAs and could easily write a GA for statistical model selection

problems. Collaborations among scientists could produce novel GAs for statistical analyses of vast model spaces and improve upon the simple GA presented in this dissertation.

SUMMARY AND CONCLUDING REMARKS

Observational (non-experimental) studies have been and continue to be an important part of the research procedure in ecology. This research procedure consists of several stages, of which observational studies are typically the first (see James and McCulloch 1985:4, 1990:132). Observational studies help lead to producing descriptive models. Information about possible causation, from a variety of sources, and a descriptive model (or models) are used to form a causal model. Experiments then test the causal model(s) to provide for strong inferences, valid predictions, and/or insight about causal mechanisms. It is this overall process which leads to the goal of uncovering causation (James and McCulloch 1985, 1990). Other discussions about ecological research procedures seem to overlap with that described by James and McCulloch (1990) (see, e.g., Hom and Cochran 1991:464-466, Scheiner 1993:3-7), and most ecologists would probably agree with this basic procedure.

Despite the importance of observational studies in the overall research procedure, serious problems often occur with both the application and the subsequent interpretation of multivariate analysis of observational data (James and McCulloch 1990). This part of the dissertation has focused on three interrelated problems regarding multivariate analysis of observational data:

1) the reliance of ecologists on stepwise procedures,
2) the tendency of analysts to select only one or a few statistical models from a vast model space, and
3) the mistake of jumping directly from exploratory analyses to conclusions and inferences about causal mechanisms.

The GA-informational modeling (GAIM) approach described in this part of the dissertation can potentially help analysts address these problems.

Stepwise algorithms are commonly used by ecologists, particularly in conjunction with hypothesis-testing procedures, to select variables and build statistical models. However, this approach has a number of limitations and problems, namely that stepwise procedures:

1) do not necessarily find the best model or even the best set of models (Mantel 1970, Hocking 1976, Moses 1986),
2) search only a limited part of the model space (Mantel 1970, Beale 1970), and
3) provide the analyst with only a few good models, and therefore a limited view of the data and models.

James and McCulloch (1990) have emphasized, based on these points and additional reasons, that ecologists should stop using stepwise procedures.

Another problem with the current approach to multivariate analysis of observational data is the tendency of analysts to search for and select a single model. For most data sets with many variables a single best model probably does not exist (Gorman and Toman 1966, Hocking 1983, McCullagh and Nelder 1989:8). This is particularly pertinent to observational studies because models obtained from such studies should not be viewed as an end-point, but rather as a means to formulate descriptive models that then help construct initial causal models to be

tested by experiments. Yet, analysts often search their observational data using stepwise procedures in order to find "the best" model. Even interactive methods of variable selection (e.g., Henderson and Velleman 1981) are sometimes used to find a single model. Such approaches to analyzing observational data produce an extremely narrow, restricted view of the data because the intention is to find a single best model rather than to find a large set of best models or the best subsets of models.

Analysis of observational data often goes from selecting a single best model to then making strong inferences about causal mechanisms and/or predictions based on this "best" model. Such a process is a problem of *overinterpretation* of the data because the analyst jumps directly from the exploratory analysis of observational data to conclusions of cause-and-effect (James and McCulloch 1990). Observational studies are best used in the early stages of inquiry in order to further refine future research questions and to help develop experiments. Observational studies can produce descriptive models, but only controlled experiments (or quasiexperimental designs) can truly allow researchers to make firm conclusions about causation or to make useful predictions (James and McCulloch 1990).

Serious problems occur in the research procedure (the procedure outlined by James and McCulloch 1990) because of the combined effects of relying too heavily on stepwise procedures, believing that a single best model must be found in all cases, and jumping from the analysis of an observational study to conclusions of cause-and-effect. The GAIM approach to analyzing multivariate observational data can potentially help ecologists address these problems. The GAIM approach emphasizes that

analysts can obtain a wide view of their multivariate observational data by using a genetic algorithm (GA) to search a vast model space and by examining both a *set* of best models and the frequency of variables in that best set.

The GAIM approach is conducted in stages, some of which have been advocated by statisticians in other contexts for a many years. In Stage 1, the analyst inspects the multivariate data and obtains a preliminary view. This stage involves such activities as checking assumptions about the data with respect to the statistical techniques being used, conducting univariate analyses, graphing the data one or two variables at a time, transforming or rescaling variables, and possibly combining variables.

In Stage 2 of the GAIM approach, which is rather new to statistical analysis, the analyst uses a genetic algorithm (GA), in conjunction with an informational model-selection criterion, to search or sample the vast model space and find the best combinations of variables to fit the data. GAs can find solutions to complex problems by using concepts of genetic recombination, natural selection, mutation, and evolution. GAs have a proven history of success with problem-solving in many areas (Goldberg 1994, Forrest 1993, Holland 1992a, b) and they can be successfully applied to statistical analysis of large observational data sets. Likewise, the informational approach to statistical modeling also has a history of success in engineering and the sciences (see Bozdogan 1994a, b, c). This approach views modeling as a problem in optimization of a criterion rather than as a problem in using statistical hypothesis-testing procedures. Specifically, the

informational approach, unlike hypothesis-testing procedures, can rank and compare both nested and non-nested models.

The analyst should perform at least two or three independent runs of the GA in Stage 2 on the data. The output produced by each GA consists of many more models than that produced by stepwise procedures. The analyst then ranks the GA models based on their values of a criterion such as AIC or ICOMP.

In Stage 3 of the GAIM approach the analyst: 1) selects a "window" of values of the criterion in order to define the set of best GA models, 2) plots the criterion values of the best GA models with respect to k (or other measures of interest), 3) decides whether other criteria (such as model variance and/or biological considerations) should be used to further redefine the set of best GA models, 4) examines the frequencies of independent variables among the best GA models to determine whether some variables are more common than others, and 5) when possible, uses diagnostic measures to obtain further insight.

The use of a GA in the GAIM approach has certain advantages. First, biologists can easily understand what the GA is doing as it searches the model space because a GA is based on basic biological concepts. Second, a great deal of flexibility can be programmed into a GA. The user can be given options for choosing the number of generations, population size, mutation rate, recombination (crossover) rate, and the number of models to be produced in the final output.

The utility of the GAIM approach is that its results are much better suited to the actual purpose of analyzing observational data than the

results obtained from stepwise procedures; that is, to produce descriptive models which can be used to help formulate causal models. This utility stems from the fact that the GAIM approach focuses on a *set* of best GA models and the variables most frequently found in that set rather than on the parameter values of the variables found in a supposedly single "best" model.

Just because an equation can be fitted to observational data does not mean that the parameter values can be interpreted with great confidence or that the equations permit strong inference. In regression analysis, researchers must "... recognize that the fitting of equations to observational data (as opposed to data from carefully designed experiments) is, at best, a risky business." (Hocking 1983:226). This is risky because errors occur in both the dependent variable and each of the independent variables, sampling is not adequate in the experimental region, and correlations are often high among many independent variables (Snee 1983:230).

Likewise, it is risky with any multivariate technique to fit equations to observational data and then attempt to make solid inferences about the numerical values of the parameters or about causation. Multivariate analysis of observational data is a descriptive or correlative analysis and an exploratory type of investigation rather than a confirmatory one (James and McCulloch 1990). Some multivariate methods can be confirmatory (i.e., statistical inferences or conclusions can be made and extended beyond the sample to a larger population). However, confirmatory conclusions can be valid only when a truly random sample has been obtained from a

large population (Tukey 1980, James and McCulloch 1990) and conditions mentioned above (see Snee 1983) are not problematic.

It cannot be emphasized enough that any multivariate observational data set does not lend itself easily to the selection of one best model for the purposes of making solid inferences and predictions. Observational data often leads to the formation of descriptive models which, when combined with additional observations and research findings, help researchers construct initial causal models (James and McCulloch 1990). Such causal models can also be thought of as biological hypotheses. Testing competing hypotheses by means of experiments is a method of inductive inference. Scientific knowledge and understanding often advances most rapidly when the experiments can cleanly eliminate all but one of the alternatives in this inductive process (Platt 1964).

An important purpose of collecting and analyzing observational data is to help researchers develop alternative hypotheses (or models) to be tested by well-designed, controlled experiments. Much of what this purpose entails, as well as what the GAIM approach advocates, is the process of scientific inquiry called abduction. The philosopher Charles S. Peirce is credited with formally describing abduction. Sometimes abduction is defined as the initial process of producing alternative explanatory hypotheses (see, e.g. Akaike 1994). Some philosophers and scientists, however, define abduction as a process of making observations, forming hypotheses to explain the observations, and the subsequent process of evaluating the alternative hypotheses and then deciding which one is the best explanatory hypothesis (see Josephson 1994:8-9). Peirce seems to have

considered that creativity and originality in thinking comes mainly from proposing hypotheses or suggestions, rather than from induction alone whereby hypotheses are merely tested. For simplicity, consider abduction here to mean the process of generating hypotheses.

Akaike (1994) convincingly argues that: 1) R. A. Fisher's view of likelihood was limited to the idea of estimating parameters while assuming a given model (and model structure), 2) the Fisherian view of statistics put abductive inference out of the realm of consideration by theoretical statisticians, and 3) the introduction of AIC and the informational approach changed the way in which likelihood was viewed such that the loglikelihood can be seen as a general criterion for comparing models which may have distributional structures that are different. The informational approach uses numerical criteria, which are justified in theory and proven by successful application, to evaluate and compare competing models or hypotheses.

Akaike (1994:33) emphasizes that the concept of a true model depends on the modeling objectives (such as the construction and application of the model) and "... the basic choice of a model is realized only through the mental activity of the researcher.". The informational viewpoint allows for some creativity and subjective activity in the process of generating hypotheses ("abduction"), whereas the hypothesis-testing framework is limited to the inductive testing of a given hypothesis versus the null hypothesis (Akaike 1994). Akaike (1994) further states: "... the generation of an innovative hypothesis is always highly dependent on personal activity, it is obvious that blind adherence to the concept of objectivity must be

eliminated to regain the creative power of statistical methods in scientific activities.".

One could extend Akaike's statements to the current application of stepwise algorithms, in conjunction with hypothesis-testing procedures, which is based on this blind adherence to statistical objectivity and a desire to find a single best model. However, such an unthinking approach does not provide researchers with the full potential of both the statistical muscle and creative power needed in the abductive process when analyzing observational multivariate data.

For an observational data set which has a small number of variables, chances are fairly good that one or a few clearly best models exist. However, as the number of variables increases linearly the number of models increases exponentially to produce a vast model space. In such cases it is unlikely that a minimum AIC model exists and the analyst would not have a clear choice of a single best model (Akaike 1985:16, 1994:34).

The lack of a clearly best model might be seen as a deterrence. Akaike's point is well taken, but the GAIM approach shows that insight can still be obtained from a vast model space for several reasons. First, this approach deals with observational multivariate data in which the analyst does not expect a single best model to even exist. Second, the nature of the analyses with such data is exploratory, not confirmatory. Third, the GAIM approach emphasizes the use of an informational criterion as an initial screening criterion, not as the definitive criterion. Other statistical measures, as well as biological knowledge, can be used to further screen and select the best

models produced by the GA. Fourth, the objective is to obtain a *set* of the best GA models and then examine the frequency of variables in that set. Individual variables or combinations of variables that occur most frequently in the best GA models provide insight into possible relationships to investigate in future experiments. Thus, the GAIM approach, though using an informational criterion on a vast model space, can provide a method for obtaining insight from observational data and directing the formation of biological hypotheses.

The GAIM approach emphasizes that:

1. Analysts should *not* expect a single best model to exist for observational multivariate data which have a vast model space.
2. The use of a GA, in conjunction with an informational criterion, can help the analyst find a *set* of very good models to fit the data.
3. An informational criterion serves as an *initial* statistical criterion for the selection of models, but additional statistical and/or biological criteria can be used to select the best models from the GA output.
4. The frequency distribution of the variables which appear in the final "best" set of models and the analyst's insight and expertise can then be used together to formulate descriptive models.
5. The descriptive models and insight from other observations, studies, and published papers can be used to construct initial causal models.

Points 2-5 above fit into the process of abduction, rather than into induction. The initial causal models or hypotheses are then tested by performing well-designed, controlled experiments in a process of inductive inference to determine the best hypothesis.

Besides helping researchers formulate both descriptive and initial causal models, the GAIM approach can provide better opportunities than stepwise procedures to compare models from other studies against a *richer*

set of original models for purposes of external validation of observational studies. The validation process is much more limited in scope with stepwise procedures because each researcher only reports a single "best" model and its estimated parameter values, thus providing a very narrow view of the data and models to the scientific community. Ideally, external validation would involve a comparison of *sets* of models and variable frequencies from a number of studies. Model validation and scientific inference might better be served if a set of best models and the frequency distribution of variables were reported, as suggested in the GAIM approach.

Weak inferences about causation could possibly be obtained when independently conducted observational studies produce the same results, provided that the same study methods and statistical procedures were used. However, the strongest inferences come from well-designed, controlled experiments that can eliminate competing hypotheses in favor of one hypothesis.

Conservation decisions and wildlife management plans are being based on inferences and predictions obtained from observational data published in ecology, conservation biology, and wildlife biology journals. Such decision making may be too risky because multivariate observational data and the estimated parameter values fit to such data do not necessarily permit strong inferences and reliable predictions, particularly for data that has a vast model space. Habitat modeling is one area of analysis being used to provide explicit conservation and/or management plans for many species even though the habitat studies: 1) use stepwise procedures (or other limited searching methods) on multivariate observational data,

2) report only one or a few of the actual models fitted to the data, and 3) lack external validation by other studies on the same species and habitats.

No doubt decisions must be made, but the risks of making strong inferences and predictions based on observational data from a single study must be addressed by researchers, policy makers, and the public. Conservation and management would be better served in the long term if researchers followed the research procedure reiterated by James and McCulloch (1990) to ultimately test causal mechanisms and policy makers based their decisions on this entire research procedure, not simply on an observational study. If experiments were not possible in certain field situations, then weak inferences could possibly be obtained from several independently conducted, observational studies (on the same species or habitats) which did not rely upon stepwise procedures and the statistical hypothesis-testing framework for a very limited view of the data. The GAIM approach is potentially useful for providing a wider view of the data than stepwise procedures, helping to formulate descriptive and initial causal models, and acting as a tool for richer external validation.

No doubt abuses of the GAIM approach can occur. Some abuses of the stepwise methods have occurred because of the improper use of such methods and the subsequent interpretations, rather than because of claims made by developers of such methods (see Hocking 1976:8-9, Hosmer and Lemeshow 1989:87). With the GAIM approach, for example, one could ignore nearly all of the output and simply pick the model with the lowest criterion value to be the single "best" model. By no means is such a method recommended here. Instead, it is recommended that analysts use

the methods discussed in Stage 3 of the GAIM approach and interpret the results in light of the comments of Hocking (1983), James and McCulloch (1985, 1990), Moses (1986), and others regarding the purposes and limitations of observational studies. As for overinterpretation or misinterpretation of results, the GAIM approach itself cannot guard against these problems. Analysts themselves must guard against problems of interpretation, but perhaps the GAIM approach will help analysts adopt a different philosophy toward interpreting the analysis and objectives of observational data.

This chapter also discussed specific changes in the way logistic regression analysis of observational data can be conducted from that which is presently being used by ecologists. First, analysts should carefully examine the assumption of linearity in the logit for each independent variable and make any necessary transformations or scale changes in those variables in violation of this assumption.

Second, the approach suggested here utilizes the informational approach instead of hypothesis-testing procedures. This is in keeping with the growing philosophy that statistical modeling is not a hypothesis testing problem per se, but rather a problem in optimizing some criterion for model selection (see Sakamoto et al. 1986, Bozdogan 1987, 1988a, b, Burnham and Anderson 1992, Lebreton et al. 1992).

Third, it is suggested that a GA, in conjunction with an informational model-selection criterion, be used to search a vast model space for the best models. This is extremely different, both methodologically and

philosophically, from the use of stepwise procedures for logistic regression modeling.

Fourth, the assumption that model variance is equal to one for logistic regression models should be checked because standard statistical packages typically do not check and ecological data may not satisfy this assumption. Analysts should incorporate estimated variance or a variance-inflation factor into the calculation of the model-selection criterion. One method suggested here was the incorporation of a model's estimated variance into the calculation of its ICOMP-IFIM value.

The GAIM approach is potentially applicable to multivariate analyses, not just multiple regression analysis. Ecologists should end their heavy reliance on both stepwise procedures and the idea that a single best model exists for any large observational data set. Alternative methods for searching and evaluating statistical models exists for observational data. Data sets containing >12 variables could be analyzed using branch-and-bound algorithms or the GAIM approach discussed here. Data sets with $\leq$12 variables can be analyzed more effectively by using already existing algorithms to enumerate all possible models and to calculate the values of a model-selection criterion, such as Mallows' C_p for regression cases or an informational criterion, rather than by using a stepwise procedure. Criterion values can be used to make an initial selection of the best models. Then the analyst can follow the procedures used in Stage 3 of the GAIM approach to obtain a wide view of the data and models which can ultimately be used to formulate initial causal models or hypotheses about causation.

The GAIM approach discussed in this chapter is in its infancy. Further research needs to examine the performance of the GAIM approach in the statistical modeling of observational data and in helping researchers produce descriptive models, build initial causal models, and design experiments to test causal models. A wide variety of data sets should be explored with the proposed GAIM approach. It could be particularly interesting for researchers to model old data sets using the GAIM approach and then compare the new GAIM results to their previous results obtained from stepwise procedures. The two methods may produce similar results in some cases, but for many complex, multivariate data sets it is expected that the GAIM approach will provide different insights.

LITERATURE CITED

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267-281 *in* B. N. Petrov and F. Csáki, editors. Second international symposium on information theory. Akadémiai Kiadó, Budapest, Hungary. 451 pp.

Akaike, H. 1985. Prediction and entropy. Pages 1-24 *in* A. C. Atkinson and S. E. Fienberg, editors. A celebration of statistics: the ISI centenary volume. Springer-Verlag, New York, New York, USA. 606 pp.

Akaike, H. 1994. Implications of informational point of view on the development of statistical science. Pages 27-38 *in* H. Bozdogan, editor. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. 1992. Vol. 3, engineering and scientific applications. Kluwer Academic Publishers, Dordrecht, The Netherlands. 346 pp.

Bartlett, J. G. 1995. Relative abundance of breeding birds and habitat associations of select neotropical migrant songbirds on the Cherokee National Forest, Tennessee. M.S. thesis, University of Tennessee, Knoxville, Tennessee, USA. 143 pp.

Beale, E. M. L. 1970. A note on procedures for variable selection in multiple regression. Technometrics 12:909-914.

Bearse, P. M. and H. Bozdogan. 1998. Subset selection in vector autoregressive models using the genetic algorithm with informational complexity as the fitness function. Systems Analysis Modeling Simulation (SAMS) 31:61-91.

Bearse, P. M., H. Bozdogan, and A. M. Schlottmann. 1997. Empirical econometric modelling of food consumption data using a new informational complexity approach. Journal of Applied Econometrics 12:563-592.

Box, G. E. P. 1979. Robustness in the strategy of scientific model building. Pages 201-236 *in* R. L. Launer and G. N. Wilkinson, editors. Robustness in statistics. Academic Press, New York, New York, USA. 296 pp.

Bozdogan, H. 1987. Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. Psychometrika 52:345-370.

Bozdogan, H. 1988a. ICOMP: a new model-selection criterion. Pages 599-608 *in* H. H. Bock, editor. Classification and related methods of data analysis: proceedings of the first conference of the international classification societies. North-Holland, Amsterdam, The Netherlands. 750 pp.

Bozdogan, H. 1988b. Selecting loglinear models and subset selection of variables in multiway contingency tables using Akaike's Information Criterion (AIC). Pages 609-616 *in* H. H. Bock, editor. Classification and related methods of data analysis: proceedings of the first conference of the international classification societies. North-Holland, Amsterdam, The Netherlands. 750 pp.

Bozdogan, H. 1990. On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. Communications in Statistics: Theory and Methods 19:221-278.

Bozdogan, H. (editor). 1994a. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. Vol. 1, theory and methodology of time series analysis. Kluwer Academic Publishers, Dordrecht, The Netherlands. 277 pp.

Bozdogan, H. (editor). 1994b. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. Vol. 2, multivariate statistical modeling. Kluwer Academic Publishers, Dordrecht, The Netherlands. 413 pp.

Bozdogan, H. (editor). 1994c. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. Vol. 3, engineering and scientific applications. Kluwer Academic Publishers, Dordrecht, The Netherlands. 346 pp.

Bozdogan, H. 1994d. Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. Pages 69-113 *in* H. Bozdogan, editor. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. Vol. 2, multivariate statistical modeling. Kluwer Academic Publishers, Dordrecht, The Netherlands. 413 pp.

Bozdogan, H. and D. M. A. Haughton. 1998. Informational complexity criteria for regression models. Computational Statistics & Data Analysis 28:51-76.

Brennan, L. A., W. M. Block, R. J. Gutiérrez. 1986. The use of multivariate statistics for developing habitat suitability index models. Pages 177-182 *in* J. Verner, M. L. Morrison, and C. J. Ralph, editors. Wildlife 2000: modeling habitat relationships of terrestrial vertebrates. The University of Wisconsin Press, Madison, Wisconsin, USA. 470 pp.

Buehler, D. A., T. J. Mersmann, J. D. Fraser, and J. K. D. Seegar. 1991. Effects of human activity on bald eagle distribution on the northern Chesapeake Bay. Journal of Wildlife Management 55:282-290.

Burger, L. D., L. W. Burger, and J. Faaborg. 1994. Effects of prairie fragmentation on predation on artificial nests. Journal of Wildlife Management 58:249-254.

Burnham, K. P. and D. R. Anderson. 1992. Data-based selection of an appropriate biological model: the key to modern data analysis. Pages 16-30 *in* D. R. McCullough and R. H. Barrett, editors. Wildlife 2001: Populations. Elsevier Science Publishers, London, United Kingdom. 1163 pp.

Capen, D. E., J. W. Fenwick, D. B. Inkley, and A. C. Boynton. 1986. Multivariate models of songbird habitat in New England forests. Pages 171-175 *in* J. Verner, M. L. Morrison, and C. J. Ralph, editors. Wildlife 2000: modeling habitat relationships of terrestrial vertebrates. The University of Wisconsin Press, Madison, Wisconsin, USA. 470 pp.

Chandler, S. K., J. D. Fraser, D. A. Buehler, and J. K. D. Seegar. 1995. Perch trees and shoreline development as predictors of bald eagle distribution on Chesapeake Bay. Journal of Wildlife Management 59:325-332.

Coker, D. R. and D. E. Capen. 1995. Landscape-level habitat use by brown-headed cowbirds in Vermont. Journal of Wildlife Management 59:631-637.

Davis, L. editor. 1991. Handbook of genetic algorithms. Van Nostrand Reinhold, New York, New York, USA. 385 pp.

DeLong, A. K., J. A. Crawford, and D. C. DeLong, Jr. 1995. Relationships between vegetational structure and predation of artificial sage grouse nests. Journal of Wildlife Management 59:88-92.

Diefenbach, D. R. and R. B. Owen, Jr. 1989. A model of habitat use by breeding American black ducks. Journal of Wildlife Management 53: 383-389.

Diller, L. V. and R. L. Wallace. 1994. Distribution and habitat of *Plethodon elongatus* on managed, young growth forests in north coastal California. Journal of Herpetology 28:310-318.

Drewien, R. C., W. M. Brown, and W. L. Kendall. 1995. Recruitment in Rocky Mountain greater sandhill cranes and comparison with other crane populations. Journal of Wildlife Management 59:339-356.

Efroymson, M. A. 1960. Multiple regression analysis. Pages 191-203 *in* A. Ralston and H. S. Wilf, editors. Mathematical methods for digital computers. Volume 1. John Wiley and Sons, Inc., New York, New York, USA. 293 pp.

Engleman, L. 1988. Stepwise logistic regression. Pages 1013-1046 *in* W. J. Dixon, editor. BMDP Statistical Software. University of California Press, Berkeley, CA, USA. 1234 pp.

Forrest, S. 1993. Genetic algorithms: principles of natural selection applied to computation. Science 261:872-878.

Freeman, D. H. 1987. Applied categorical data analysis. Marcel Dekker, Inc., New York, New York, USA. 318 pp.

Furnival, G. M. and R. W. Wilson. 1974. Regression by leaps and bounds. Technometrics 16:499-511.

Goldberg, D. E. 1989. Genetic algorithms in search, optimization, and machine learning. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, USA. 412 pp.

Goldberg, D. E. 1994. Genetic and evolutionary algorithms come of age. Communications of the ACM 37:113-119.

Gorenzel, W. P. and T. P. Salmon. 1995. Characteristics of American crow urban roosts in California. Journal of Wildlife Management 59:638-645.

Gorman, J. W. and R. J. Toman. 1966. Selection of variables for fitting equations to data. Technometrics 8:27-51.

Henderson, H. V. and P. F. Velleman. 1981. Building multiple regression models interactively. Biometrics 37:391-411.

Hinsley, S. A., P. E. Bellamy, I. Newton, and T. H. Sparks. 1996. Influences of population size and woodland area on bird species distributions in small woods. Oecologia 105:100-106.

Hocking, R. R. 1976. The analysis and selection of variables in linear regression. Biometrics 32:1-49.

Hocking, R. R. 1983. Developments in linear regression methodology: 1959-1982. Technometrics 25:219-230.

Holland, J. H. 1992a. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT Press, Cambridge, Massachusetts, USA. 211 pp.

Holland, J. H. 1992b. Genetic algorithms. Scientific American July 1992:66-72.

Hom, C. L. and M. E. Cochran. 1991. Ecological experiments: assumptions, approaches, and prospects. Herpetologica 47:460-473.

Hosmer, D. W., Jr. and S. Lemeshow. 1989. Applied logistic regression. John Wiley and Sons, Inc., New York, New York, USA. 307 pp.

James, F. C. and C. E. McCulloch. 1985. Data analysis and the design of experiments in ornithology. Pages 1-63 *in* R. F. Johnston, editor. Current ornithology, vol. 2, Plenum Press, New York, New York, USA. 378 pp.

James, F. C. and C. E. McCulloch. 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? Annual Review of Ecology and Systematics 21:129-166.

Johnson, R. G. and S. A. Temple. 1986. Assessing habitat quality for birds nesting in fragmented tallgrass prairies. Pages 245-249 *in* J. Verner, M. L. Morrison, and C. J. Ralph, editors. Wildlife 2000: modeling habitat relationships of terrestrial vertebrates. The University of Wisconsin Press, Madison, Wisconsin, USA. 470 pp.

Josephson, J. R. 1994. Conceptual analysis of abduction. Pages 5-30 *in* J. R. Josephson and S. G. Josephson, editors. Abductive inference: computation, philosophy, technology. Cambridge University Press, Cambridge, United Kingdom. 306 pp.

Kindvall, O. 1996. Habitat heterogeneity and survival in a bush cricket metapopulation. Ecology 77:207-214.

Larsen, D. T., P. L. Crookston, and L. D. Flake. 1994. Factors associated with ring-necked pheasant use of winter food plots. Wildlife Society Bulletin 22:620-626.

Lebreton, J. D., K. P. Burnham, J. Clobert, and D. R. Anderson. 1992. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecological Monographs 62:67-118.

Luh, H.-K., J. J. Minesky, and H. Bozdogan. Submitted manuscript. Choosing the best predictors in regression analysis via the genetic algorithm with informational complexity as the fitness function.

Mallows, C. L. 1973. Some comments on C_p. Technometrics 15:661-675.

Mantel, N. 1970. Why stepdown procedures in variable selection. Technometrics 12:591-612.

McCullagh, P. and A. J. Nelder. 1989. Generalized linear models. 2nd edition. Chapman and Hall, London, United Kingdom. 511 pp.

McNay, R. S. and J. M. Voller. 1995. Mortality causes and survival estimates for adult female Columbian black-tailed deer. Journal of Wildlife Management 59:138-146.

Moses, L. E. 1986. Think and explain with statistics. Addison-Wesley Publishing Company, Reading, Massachusetts, USA. 483 pp.

Myers, R. H. 1986. Classical and modern regression with applications. Duxbury Press, Boston, Massachusetts, USA. 359 pp.

Nadeau, S., R. Décarie, D. Lambert, and M. St-Georges. 1995. Nonlinear modeling of muskrat use of habitat. Journal of Wildlife Management 59:110-117.

Nordberg, L. 1981. Stepwise selection of explanatory variables in the binary logit models. Scandinavian Journal of Statistics 8:17-26.

Nordberg, L. 1982. On variable selection in generalized linear and related regression models. Communications in Statistics, Series A 11:2427-2449.

Platt, J. R. 1964. Strong inference. Science 146:347-353.

Pregibon, D. 1981. Logistic regression diagnostics. The Annals of Statistics 9:705-724.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. Akaike information criterion statistics. KTK Scientific Publishers, Tokyo, Japan. 290 pp.

Scheiner, S. M. 1993. Introduction: theories, hypotheses, and statistics. Pages 1-13 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Chapman and Hall, New York, New York, USA. 445 pp.

Smith, K. G. and P. G. Connors. 1986. Building predictive models of species occurrence from total-count transect data and habitat measurements. Pages 45-50 *in* J. Verner, M. L. Morrison, and C. J. Ralph, editors. Wildlife 2000: modeling habitat relationships of terrestrial vertebrates. The University of Wisconsin Press, Madison, Wisconsin, USA. 470 pp.

Snee, R. D. 1983. Discussion. Technometrics 25:230-237.

Sokal, R. R. and F. J. Rohlf. 1995. Biometry: the principles and practice of statistics in biological research. 3rd edition. W. H. Freeman and Company, New York, New York, USA. 887 pp.

Trexler, J. C. and J. Travis. 1993. Nontraditional regression analyses. Ecology 74:1629-1637.

Tukey, J. W. 1977. Exploratory data analysis. Addison-Wesley, Reading, Massachusetts, USA. 688 pp.

Tukey, J. W. 1980. We need both exploratory and confirmatory. The American Statistician 34:23-25.

van Manen, F. T. and M. R. Pelton. 1993. Data-based modelling of black bear habitat using GIS. Pages 323-329 *in* I. D. Thompson, editor. Proceedings of the International Union of Game Biologists XXI Congress: forests and wildlife .... towards the 21st century, Volume 1. Canadian Forest Service, Chalk River, Ontario, Canada. 379 pp.

APPENDIX TO PART 4

Table 4-1. Example of a data matrix with sample data for three continuous independent variables (X_1, X_2, and X_3), one categorical independent variable with two design variables (X_4), and a column of ones representing the intercept term (X_0).

X_0	X_1	X_2	X_3	X_4
1	10	2	9	1 0
1	12	8	5	1 0
1	10	4	7	1 0
1	16	8	5	1 0
1	8	2	11	0 1
1	10	4	13	0 1
1	14	6	11	0 1
1	10	4	9	0 1
1	18	2	13	0 0
1	16	8	15	0 0
1	20	10	13	0 0
1	20	8	19	0 0

PART 5 : ASSOCIATIONS BETWEEN HABITAT FEATURES AND THE PRESENCE OF *ANOLIS CAROLINENSIS* AMONG FOUR HABITATS IN EASTERN TENNESSEE: AN ANALYSIS USING THE GAIM APPROACH

INTRODUCTION

Considerable theoretical and empirical research has been conducted in an effort to understand the complex relationships between an individual organism and/or a species and its habitat. Habitat can potentially influence heat balance and physiology (Gates 1980, Porter 1989), growth (Porter 1989), reproduction and life history traits (Stearns 1976), individual fitness (Fretwell 1972), abundance and distribution of populations and species (Hutchinson 1957, MacArthur 1972), and species diversity (MacArthur 1964). Thus, understanding interactions and relationships between an organism or a population and its habitat(s) is critical to understanding individual behavior, individual physiological performance, life history traits, population dynamics and the viability of populations, community structure and organization, and evolution.

Concerns over the future of many animal populations and species contributed to passing federal laws in the United States which required that wildlife and their habitats, as well as other natural resources, be given consideration whenever human activities are planned and conducted on public lands (Morrison et al. 1992:7-9). Protection and conservation of species and their habitats also became a growing concern among scientists and the public. Scientific information and insight are needed in order to meet the requirements of federal laws and to conserve and manage species and their habitats. Thus, many studies of animal-habitat relationships have been and continue to be conducted to meet these needs (Morrison et al. 1992).

Researchers studying animal-habitat relationships often use multivariate statistical techniques (inclusive of multiple regression) to analyze observational (non-experimental) data. The application of such techniques has been an important aspect in the development of a quantitative approach to animal-habitat studies. However, misuses and misapplications of multivariate statistics have been reported to occur in animal-habitat studies, in particular (e.g., see Johnson 1981a, b), and in ecological research, in general (e.g., see James and McCulloch 1985, 1990). Two problems critical to the analysis and interpretation of observational multivariate data exist: 1) using stepwise algorithms to find a supposedly single "best" model and 2) inappropriately making specific predictions and/or causal inferences.

Many observational data sets may have 15 or more biologically relevant variables. Statistical model selection with such data can be a difficult process because multicollinearity can occur and the total number of possible models, or "model space", may be vast (i.e., in the hundreds of thousands of models or more). Researchers typically use stepwise algorithms, such as "forward selection", "backward elimination", and "stepwise selection" procedures (see Hocking 1976, 1983 for reviews), in conjunction with hypothesis-testing procedures in order to find a single "best" model to fit the data.

Analysts often believe that a stepwise procedure actually finds the single "best" model or at least report the use of such a procedure as having found the "best" model. A few points are thus warranted regarding model selection via stepwise procedures. First, stepwise procedures search and

evaluate a very small proportion of the model space (Beale 1970, Mantel 1970), providing researchers with only a narrow view of their data and model space. Second, stepwise algorithms cannot compare non-nested models when using hypothesis-testing procedures. Considering such facts it is no surprise that statisticians have stated that stepwise procedures cannot find the single best model (if it exists) or even the best set of models (see e.g., Mantel 1970, Hocking 1976, Moses 1986). Third, it is fairly probable that no single model will be better than all other models for any multivariate data set (Gorman and Toman 1966, Hocking 1983, McCullagh and Nelder 1989:8); such data can often be described equally well, statistically and biologically, by several or more models. James and McCulloch (1990) have suggested that ecologists stop using stepwise procedures on multivariate data.

The second problem critical to the analysis and interpretation of observational multivariate data is when interpretation of the results incorrectly leaps from a correlative description to a causal explanation (James and McCulloch 1990). For example, analysts of observational data have often taken the supposedly "best" model obtained from a stepwise procedure and made specific inferences about the causation between dependent and independent variables and/or precise predictions. Cautionary notes have been sounded about the dangers of making such causal interpretations (see Johnson 1981b, James and McCulloch 1985, 1990) and predictions (see Hocking 1983, Snee 1983) based on observational data. Such data help produce models or hypotheses about possible causation, but controlled experiments provide the actual tests of causal hypotheses or

models upon which causal inferences can be based (James and McCulloch 1990, Lubchenco and Real 1991). The different steps or phases of a research procedure should be 1) collect multivariate observational data and conduct exploratory analysis of the data, 2) formulate descriptive models based on that exploratory analysis, 3) use the descriptive models and information from other studies about possible causation to propose causal hypotheses or models, 4) test the causal model(s) by preferably using a controlled field or lab experiment (see James and McCulloch 1985, 1990).

Based on the issues and concerns mentioned above, the genetic algorithm-informational modeling (GAIM) approach has been suggested (Part 4 of this dissertation) as an alternative to stepwise procedures for analyzing observational multivariate data. The GAIM approach emphasizes the need to find and report a *set* of very good models and provides the computer and statistical tools to do so. The GAIM approach uses a genetic algorithm (GA) in conjunction with an informational model-selection criterion to help the analyst select a *set* of very good models (see Part 4 of this dissertation) and obtain a wider view of the data and models.

GAs are computer algorithms which can find very good solutions to complex problems in which hundreds of thousands or more possible solutions exist. GAs, which are based on the concepts of biological evolution, natural selection, and genetic recombination, have been used successfully in many problem-solving applications (Holland 1992a, b, Forrest 1993, Goldberg 1994). The ability to find a set of well-fitting statistical models in a vast model space can be achieved by using an

informational criterion, such as Akaike's Information Criterion (AIC; Akaike 1973) or Bozdogan's Informational Complexity criterion (ICOMP; Bozdogan 1988a, 1990, Bozdogan and Haughton 1998), as the fitness function in a GA (see Part 4 of this dissertation).

The present study examines potential associations between habitat features and the presence of *Anolis carolinensis* (Sauria: Polychrotidae) based on observational data from field studies. *Anolis carolinensis* is a small, mainly arboreal lizard found in the southern United States, is a member of a genus of tropical origin, and is the only native anoline species in the continental United States. Populations of *A. carolinensis* in eastern Tennessee represent some of the most northern populations of this species. Despite being common in many habitats across its distribution, most research on this species is laboratory-based and quantitative field research is considerably lacking on the relationship between habitat and *A. carolinensis.*

The objectives of this study are two-fold. The first is to demonstrate the application of the GAIM approach to the analysis of observational data on *A. carolinensis*-habitat relationships. The second objective is to obtain insight, in the form of descriptive models, into possible associations between the presence of *A. carolinensis* in home range sized plots, during a summer and a winter season, and various habitat features across four habitat types in eastern Tennessee. Such insight, along with what is already known about this species, could be used to formulate causal models or hypotheses about *A. carolinensis*-habitat relationships which could be tested by future experiments. In this way, the ultimate goal of testing

causal models, via an experimental approach (as emphasized by James and McCulloch 1990) would be better served than by simply making causal inferences based on analysis of observational data alone.

STUDY SITES

The study was conducted on wooded slopes adjacent to the Little Tennessee River (Monroe and Blount Counties, Tennessee) in an area straddling the Blue Ridge and Ridge and Valley Physiographic Provinces. The natural potential vegetation is classified as Appalachian oak forest (Küchler 1964). The lower Little Tennessee River valley has been inhabited by humans for at least 12,000 yr B.P. and dating back to the Paleo-Indian cultural period (Delcourt et al. 1986). The last Native American group to inhabit this valley was the Overhill Cherokee during the 1700s and early 1800s before they were supplanted by European-Americans (Chapman 1985). Botanical evidence from archaeological deposits (Chapman and Shea 1981, Chapman et al. 1982) along with the independent paleoecological record from natural ponds (Delcourt et al. 1986) suggests that humans have influenced the vegetation of the valley for about the past 10,000 yr.

Early alterations of the local vegetation by humans occurred during the late Archaic cultural period (5000 to 2800 yr B.P.) when cultivation and associated land-clearing was first taking place (Delcourt et al. 1986). More extensive changes in the valley's vegetation occurred over the past 300 yr through land clearing and cultivation, which extended into the uplands, by the Overhill Cherokee and European-Americans (Delcourt et al. 1986). In

the 20th century the valley has been altered by the building of dams and associated land-use changes. The last dam built on the Little Tennessee River was Tellico Dam which impounded the lowest reaches of the river after its flood gates closed in 1979.

Selection of habitat sites for this study was based on five criteria. First, a site had to have a southerly aspect since *A. carolinensis* appears to be limited to such slopes in eastern Tennessee (personal observation). Second, a distinct edge had to be present on the southern part of a site, where edge is defined as a line of distinct change in the structural vegetation and landscape such as occurs between a wooded slope and either a pasture, road, river, or power line right-of-way. Third, a site had to be in close proximity (<100 m) to other areas where *A. carolinensis* occurs. Fourth, a site had to be connected with other wooded habitat in either an east or west direction (parallel to habitat edge). The third and fourth criteria both ensured that sites were not isolated patches of habitat and not inaccessible to colonization by *A. carolinensis*. Lastly, a site had to be readily accessible and capable of being surveyed and sampled without an extremely large cost in time. Sites with steep slopes or vertical rock faces without accessible ledges were too difficult to sample.

Four habitat sites were chosen which fulfilled these criteria. Habitat A is river bluff site with a narrow strip of vegetation between the river and bluff face and with short steep slopes and ledges leading to the bluff face. An abandoned railroad bed (now a rail-less dirt path) lies between the river and bluff. The width of vegetated habitat along the river is approximately 0.5-4 m and from 6-20 m between the dirt path and vertical bluff face. The

habitat is dominated by deciduous trees with common species being black locust (*Robinia pseudoacacia*), hackberry (*Celtis occidentalis*), redbud (*Cercis canadensis*), and ash spp. (*Fraxinus* spp.). Other trees include various oaks (*Quercus* spp.), honey locust (*Gleditsia triacanthos*), black walnut (*Juglans nigra*), winged elm (*Ulmus alata*), slippery elm (*Ulmus rubra*), sycamore (*Platanus occidentalis*), eastern red cedar (*Juniperus virginiana*), and an occasional tulip poplar (*Liriodendron tulipifera*), mimosa (*Albizia julibrissin*), and box elder (*Acer negundo*). Various woody shrubs and vines and the herbaceous flatseed sunflowers (*Verbesina occidentalis* - common; and *Verbesina virginica* - occasional) occur throughout much of the site. Numerous cracks and fissures in the south-facing bluff provide potential refuge sites for *A. carolinensis* during cold weather.

Habitat B is a wooded, south-facing slope with two rock seams running in an east-west direction; one near the crest of the slope and another at about mid-slope. This habitat is adjacent to Habitat A and the seams of rock are actually exposed rock extending laterally from the bluff. Deciduous trees dominate the slope with common trees being various oak (spp.) and hickory (spp.), black locust, and hackberry. Other trees include sycamore, redbud, walnut, and a few mimosa and eastern red cedar. Various woody shrubs and vines and the herbaceous flatseed sunflowers are also present. The south edge at Habitat B is created by the river for a short distance and a pasture for the larger part. The pasture extends up along the east edge of the habitat where a small hollow runs between Habitat B and the slopes to the immediate east. A small herd of cattle graze

in the pasture and occasionally along the slope of Habitat B itself and within flat areas of Habitat A. Both Habitat A and B are under federal land ownership as part of the land acquisition and impoundment created by the Tellico Dam project administered by the Tennessee Valley Authority (TVA).

Habitats C and D are located about 6.4 and 8.2 river km (5.5 and 7.3 km straight-line distance), respectively, upriver from Habitats A and B. A state highway runs between the Little Tennessee River and Habitats C and D. Habitat D is approximately 2.2 river km (2.0 km straight-line distance) upstream from Habitat C. Habitat C is a wooded hillside dominated by pine (*Pinus* spp.) with some oaks present. No rock seams or outcrops occur within the sampled area, but a small bluff and rock outcrop (where *A. carolinensis* is present) occurs just to the west of Habitat C with continuous forest present between the two areas. The habitat edge is created by a treeless patch maintained for local electrical lines running between the site and the highway.

Habitat D is a wooded slope bordered to the east by a small body of water backed-up by the impoundment of Chilhowee Dam and to the south by a state highway. The vegetation of the sampled area is a mixture of both deciduous and evergreen trees consisting of pines and oaks. Farther up the slope and outside the sampled area pines dominate. Ground cover at this site includes some woody shrubs and sapling trees, but very little herbaceous cover. The southern edge at this habitat is created by a narrow treeless area between the site and the highway. A small rock outcrop occurs between Habitat D and the highway, but no bluff or rock outcrops

occur within the sampled habitat. Both Habitat C and D have a history of past fires (at least ground fires) as indicated by fire scars on fallen logs and the lower portions of tree trunks.

METHODS

Habitat scales

Habitat is defined in this study as the area or place that contains the physical, chemical, and biotic resources required by individuals or populations of a given species (see Davis 1960), or even a species itself. Such resources can include water, humidity, sunlight, heat, shade, nesting or egg-laying sites, food, structural vegetation, and refugia from both predators and potentially threatening weather conditions.

In the present study, surveys for the presence of lizards were conducted in small plots (radius = 2.5 m, surface area approximately = 19.6 m^2) which approximated the area used by *A. carolinensis* as summer territories and winter home ranges in eastern Tennessee (J. J. Minesky, unpublished data). This size is similar to that for this species in Louisiana (Gordon 1956) and South Carolina (Jenssen et al. 1995). Habitat variables were measured at various scales ranging from within the plots to the categories of four habitat types (sites) themselves.

The plots represent the "seasonal-use" (SU) habitat scale defined here as the habitat area used by a typical individual in a population during a certain climatic season and/or a biological "season" (e.g., reproductive vs. non-reproductive seasons). The summer "season" and winter "season" are defined in the subsection "*Surveying of plots*". Because *A. carolinensis*

uses somewhat different habitat patches in the course of a given year it is important to define the SU habitat scale based on such seasons. Within the SU habitat an individual will encounter several or more "microhabitats", each being a habitat patch used by an individual in conjunction with a specific activity during a specific time or segment of a daily activity pattern. The activity can be performed for either directly obtaining one or more resources or conducting one or more functions not directly related to resource acquisition such as sleeping, molting, egg-laying, nesting, or displaying to or communicating with other individuals.

At a level above the SU habitat, "overall home range" (OHR) habitat is the habitat scale that approximately equals the area used by an individual in the population over either one complete cycle of defined alternating biological activity "seasons" or one complete cycle of climatic seasons (such as one calendar year). *Anolis carolinensis* often uses different home ranges during summer vs. winter seasons in eastern Tennessee, thus it is important to make the distinction between SU and OHR habitat scales. For those individuals which change their use of SU or OHR habitat and/or location of home ranges over an entire life-time, "life-time" (LT) habitat is the habitat area used over the typical life span of a non-migratory individual. The term "macrohabitat", depending on how it has been defined by other researchers, may be similar in scale to that of OHR or LT habitats.

The "population level" (PL) habitat scale is that habitat area occupied by a particular population or subpopulation of a species during a given unit of time, such as a season or year. This scale is does not necessarily address the

concept of minimum viable population (MVP, Shaffer 1981) or "minimum dynamic area" (MDA, Thiollay 1989) because PL habitat is defined for time scales of interest shorter than those typically used with MVPs.

Each habitat or site in the present study represents a different PL habitat because a) the total area available in each habitat could support *A. carolinensis* territories/home ranges and a potential population for at least two consecutive seasons and b) basic vegetational and physical features were distinct from adjacent habitats, but fairly similar within a given habitat. No distinction was made between potential "source" and "sink" populations or habitats, but hatchlings were observed in all four habitats.

Plots

The actual area used for possible sampling ("sampled area") within each habitat excluded the extreme east and west areas of a habitat because those areas either made a transition into or formed a distinct boundary with the adjacent habitat. The actual sampled area within each habitat varied in dimensions due to differences in size and physical features of the four habitats. The approximate length along the southern edge and the maximum distance up the slope, respectively, for each of the sampled areas were: 236 and 22 m (Habitat A), 60 and 45 m (Habitat B), 67 and 45 m (Habitat C), and 26 and 45 m (Habitat D). The number of circular plots, used as sampling units, within each habitat was proportional to the approximate total area of each habitat to be sampled such that 30-35% of each habitat was sampled. Thus, the number of plots used was 51 for A, 43 for B, 51 for C, and 21 for D for a total of 166 plots.

Nearly all plots (160) were randomly located in the habitats using randomly generated, whole integer coordinates for a) the distance (m) along the southern edge from the western corner of the sampled area and b) the distance (m) up the slope from the southern edge. The remaining six plots were placed non-randomly in Habitats A and C to sample some locations which might have been missed by random coordinates due to local variations in the topography of the sites (such as on bluff ledges or in small depressions). Plot centers were marked with either rebar steel rods, wooden stakes, or paint marks.

Surveying of plots

Each plot was surveyed twice for the presence of *A. carolinensis*, once in summer and once in winter. The summer survey period was 22 May through 22 July 1991 and only the presence of adult *A. carolinensis* (females $\geq$ 45 mm and males $\geq$ 50 mm SVL) was recorded. This "season" or period was chosen because it was a time when adults had firmly established summer territories, mating was taking place, and adult activity and visibility were still high (personal observations). Monthly counts of *A. carolinensis* along transects showed that the number of adult individuals visually observed declines from April through August and September, but numbers are most consistent between June and July (unpublished data). These differences between April and August/September are probably due to a combination of changes in lizard activity, mortality, and in leaf and vegetation cover. Hatchlings first appear in early July and continue to be produced through late September or early October so that peak densities of young-of-the-year occur in September and October in eastern Tennessee.

Thus, hatchlings and juveniles were not used in the summer analysis because their numbers changed too dramatically over the summer and early-fall, which would certainly influence the probability of finding a lizard in a plot.

The presence of juveniles, along with adults, was recorded in the winter surveys because no further recruitment of individuals occurs after mid-autumn. The winter survey period ("season") was defined as the time from to late December through early March for the following reasons. First, *A. carolinensis* in eastern Tennessee begin seeking shelter from low overnight temperatures as early as late September, but this process of shifting from summer to winter microhabitat is not completed until late October for adults and mid- to late-November for juveniles (personal observations). In March anoles begin to move further away from over-wintering shelters, at least during the day, in response to warmer temperatures. Second, visibility in the habitat is not constant until after leaf fall has been completed sometime after early November. The actual survey period for the winter study started on 31 December 1991, but because of frequent cloudy conditions it extended through 21 March 1992.

Plots within each habitat were randomly selected and the same random sequence was used for both summer and winter studies. If, however, two plots were adjacent (plot edges within 1.5 m of each other), then both plots were surveyed consecutively on the same day in order to minimize the chances of counting the same lizard in two different plots on different days. True random selection of the order of habitats themselves was not feasible since Habitats A and B were separated by a considerable distance from both

C and D and travel time wasted potential time for surveying. Time available for surveying plots was more effectively utilized by visiting Habitats A and B in conjunction since they were adjacent and by visiting Habitats C and D in conjunction since they were closer to each other than to the other habitats. Thus, the following standard, alternating sequence was adopted for visiting the habitats during the survey: ABCD, DCBA, BADC, CDAB, ABDC, DCAB, BACD, and CDBA, where the sequence within a block of four letters represents the order of visitation of these habitats on a given survey day and the sequence of eight blocks of letters represents the different orders of habitats to visit on different survey days. After the sequence CDBA was completed, the surveying sequence was repeated beginning with ABCD. Plots in all four habitats were surveyed on any given day as long as weather conditions permitted.

Surveying was only conducted during mostly sunny to sunny conditions starting about 3-4 hr after sunrise and ending about 7-8 hr after sunrise. Actual survey times were between 1000 and 1400 hr (Eastern Daylight Savings Time) for summer and between 1100 and 1430 hr (Eastern Standard Time) for winter. These times are ones of considerable activity by *A. carolinensis* in eastern Tennessee (personal observation). During the winter season no plots were surveyed if ambient air temperature was below 5.0°C, regardless of time of day and sky conditions.

Each plot was surveyed by either one observer (J. Minesky) or two observers (J. Minesky and D. MacDonald). Many plots could be identified before entering so the survey often began with the observer(s) scanning the vegetation before going into the plot. One observer always had binoculars

to assist in scanning tree canopy. The observer(s) visually scanned the plot from the ground up to the top of the trees in summer and from the ground up to about 2 m in winter. Often toward the end of the survey time the lower vegetation was searched more closely by moving branches, stems, and vines in order to see within the vegetation matrix. The presence, number, size class and, when possible, the sex, of any *A. carolinensis* was recorded for each plot. In addition, the presence and number of lizards of any other species both within and adjacent to the plot were noted.

Initially, all plots were to be surveyed for a total of 20 observer-minutes. However, early in the summer study it was realized that all 166 plots might not be surveyed before the cutoff time of late July because of the narrow window of time available each day for surveying, the sometimes rapidly changing weather conditions, and the occurrence of cloudy weather for several days at a time. Thus, the first 82 plots were surveyed for the full 20 observer-minutes, but the remaining plots were surveyed only until at least one *A. carolinensis* was seen or the 20 observer-minute mark occurred, whichever came first. This ensured that all plots were surveyed before the specified end of the survey period.

Habitat variables and their measurement

Habitat variables, regardless of the spatial or temporal scales, were measured with reference to each plot (remember that a plot represents a potential home-range or SU habitat). Based on previous qualitative field observations, the directly-measured and subsequently derived variables (Table 5-1) were thought to have some biological association with the presence of *A. carolinensis*. Some variables contain information about

structural features within the plot (e.g., those relating to tree trunk sizes and numbers, herb/shrub/vine cover, presence of dead fallen woody material, and presence of rock), whereas other variables contain information about habitat features both within and down the slope of a plot (e.g., sunlight and canopy cover). Variables measuring distance to habitat edge and distance to potential over-wintering rock span across different spatial scales (from SU to OHR scales or beyond) to which *A. carolinensis* might respond. Because *A. carolinensis* is ectothermic, the ambient air temperature at 1-1.25 m above ground level was measured after surveying each plot to consider the possible influence of temperature (which could not be controlled in the study).

Some habitat variables measured in the summer season were used in the winter analysis and vice versa. This permitted consideration of possible temporal influences of habitat variables which might be relevant from a green anole's perspective (e.g., perspective between summer and winter SU habitat scales). For instance, although a green anole presumably does not require the protective shelter and thermal properties of rocks during summer, the probability of an anole being present in a plot in summer might be higher if the plot were closer to rocks. This scenario might be due to the shorter distance the individual would need to move to find shelter at the onset of cooler weather.

All measurements of tree trunk number and size, herb/shrub/vine cover, and presence of dead fallen woody material were made during the summer season. No tree falls were noted between summer and winter which would have required re-measurement of the tree variables.

Herb/shrub/vine cover was not remeasured in the winter because very little if any evergreen ground cover existed in each plot.

Data analysis

Data were analyzed using logistic regression, a statistical technique which is applicable both to analysis of certain ecological data (Trexler and Travis 1993) and extensively used in the modeling of animal-habitat relationships (e.g., Capen et al. 1986, Brennan et al. 1986, Johnson and Temple 1986, Smith and Conners 1986, van Manen and Pelton 1993, Diller and Wallace 1994). Logistic regression is useful for modeling the relationship between a binary dependent variable and a set of continuous and/or discrete independent variables (Hosmer and Lemeshow 1989, Trexler and Travis 1993). The value of the dependent variable in the present study was either "presence" or "absence" of *A. carolinensis* in a plot as determined by surveying the plot within a given season. The habitat variables were the independent variables.

Three major differences exist in both the philosophy and methodology of statistical modeling between this study and that of other studies using logistic regression. First, the objective here was to find a set of models which fit the data well rather than to search for a supposedly single "best" model. The "best" model is the one which has a distinctly superior statistical fit to the data over all other models. Second, the GAIM approach was used rather than stepwise procedures (in conjunction with hypothesis-testing procedures) to analyze the data. Third, an estimate of model variance for each logistic regression model was used in the analysis rather

than assuming that model variance = 1, as researchers commonly assume with logistic regression. These points are explained further as the methods of data analysis are outlined below.

Stepwise procedures typically search and evaluate a very limited part of any vast model space. A forward selection (FS) procedure, for example, would evaluate $k(k + 1)/2$ models *at most*, where k is the total number of independent variables (Beale 1970, Mantel 1970). In this study with 19 variables (including the intercept), only 190 out of 524,287 ($= 2^{19} - 1$) models would be evaluated at best by a FS procedure. Researchers can evaluate many more models with a GA than with a stepwise procedure because of a) the differences between the searching abilities of GAs and stepwise procedures and b) the ease with which the programming code in a GA can be modified by the researcher (see Part 4). Thus, a GA always has the ability to evaluate more models than a stepwise procedure and give a researcher a wider view of the data.

As Forrest (1993:875) indicated, GAs can find very good solutions to a problem, but are not appropriate for problems ".... in which it is important to find the exact global optimum.". In other words, a GA cannot necessarily find 'the single best' model in the problem of statistical model selection when the model space is vast and many good models exist. The real utility of a GA is that it can find *many* very good models to fit the data when the model space is vast. Use of a GA is therefore appropriate for multivariate analysis of observational data where the analyst needs to conduct exploratory analysis and obtain more insight into the data (such as

in obtaining a set of well-fitting models out of the vast model space) than can be provided by stepwise procedures.

The three basic components or stages of the GAIM approach (see Part 4) were used to find a set of very good models for both the summer and winter data. The first stage involves checking both the form of the variables and assumptions about the data, transforming or rescaling variables, and conducting univariate analyses. Each categorical variable was checked for both zero and low cell (category) counts for both the "present" and "absent" responses and appropriate recombinations of such cells (categories) with other cells were made. For example, Habitats C and D had low cell counts for the "present" response (only two and one, respectively) for the summer data (but not for winter). Thus, Habitats C and D were combined into one habitat category ("Pine/Mixed").

For continuous variables the assumption of linearity in the logit was examined where the logit is defined as the natural logarithm of the quantity [Pr(Y="present")/Pr(Y="absent")]. This assumption was examined at the univariate stage rather than after the multiple regression model was obtained because a) biologically relevant variables can be associated with the outcome variable in ways that are not linear in the logit and b) elimination of such relevant variables based only on non-linearity in the logit is not a desired event during the model selection process. Examination of the linearity assumption followed a method based on that of Hosmer and Lemeshow (1989:2-6, 85) which involved a) dividing a continuous variable into groups with approximately similar numbers of observations, b) calculating for each group: the mid-point value, the

proportion of observations with the outcome as "present", and the logit, and then c) plotting the logit versus the group mid-point for each group. Usually ten groups of approximately equal sizes were used for this analysis but, in a few cases, fewer groups had to be used due to either the ordinal nature or the limited number of observed values of some continuous variables. A linear regression r^2 of approximately 0.7 or greater was considered sufficient evidence of linearity in the logit. Failure to meet this criterion led to consideration of possible transformations or categorizations of the variable. Most of the continuous variables for both summer and winter data exhibited non-linearity in the logit. Categorization of all variables violating the logit linearity assumption was then conducted because suitable, simple transformations were not found.

The specific categorization of any continuous variables which were non-linear in the logit was based either on natural break-points in the data or on examination of the proportions and odds-ratios across categories. Initially, those variables not showing natural break-points were split into three to five categories of roughly equal size. Those categories with similar proportions of the outcome "presence" and similar odds-ratios were combined. Also, any zero cells were combined with non-zero cells. Most categorizations of habitat variables were accomplished by making simple dichotomies, but some variables required three or four categories. The category with the lowest proportion of plots having green anoles present for each variable was designated as the reference cell (i.e., the category to which others are compared as in the calculation of odds ratios; see Hosmer and Lemeshow 1989:45-50). The final form and description of each

explanatory variable is given in Tables 5-2 (summer data) and 5-3 (winter data).

Finally in stage one, univariate logistic regression models were fit to the data and the univariate likelihood ratio test statistics (*G*), ICOMP-IFIM values, regression parameters (ßs), and Wald statistics were examined for both the original and rescaled variables. In all cases where rescaling was needed, the fit of the rescaled variable was better, based on ICOMP-IFIM values, than the fit of the original variable.

The second stage of the GAIM approach involves the use of a GA to find a set of models that fit the data very well. A description of the basic workings of a GA can be found in "The Genetic Algorithm And Its Application To Statistical Modeling of Observational Data" in Part 4 of this dissertation, as well as in Goldberg (1989) and Holland (1992a, b). The GA used in this study was written by Dr. Hang-Kwang Luh in MATLAB (The Math Works, Inc. 1989). Each categorical variable (regardless of the number of categories) and continuous variable was represented as a bit on the model "string". Any categorical variable with more than two categories simply had all of the design variable columns in the data matrix either enter or exit a model whenever the bit representing the variable itself entered or exited a model (see "A GA-Informational Modeling Approach for Logistic Regression" in Part 4).

An informational model-selection criterion was used in the fitness function of the GA. Akaike (1973) first proposed the use of an information criterion for model selection and since then information criteria have been utilized in a variety of scientific fields for the purpose of statistical model

selection (e.g., see Bozdogan 1994a, b, c). The use of informational criteria for model selection in ecological research has been increasing over the past 10 years, particularly in the analysis of capture-recapture data (see Szymczak and Rexstad 1991, Burnham and Anderson 1992, and Lebreton et al. 1992, Anderson et al. 1994, Burnham et al. 1995a, b). Aspects of the informational approach, including some advantages over hypothesis-testing procedures, can be found in Sakamoto et al. (1986), Bozdogan (1987, 1988a, b, 1990), Burnham and Anderson (1992), and Lebreton et al. (1992), as well as in Part 3 of this dissertation.

The specific model selection criterion used in the GA's fitness function was ICOMP-IFIM. This criterion is defined by Bozdogan (1990, 1994d) as:

$$\text{ICOMP-IFIM} = -2(\text{Loglikelihood}) + 2[C_1(\text{IFIM})]. \tag{5.1}$$

Models with the lowest ICOMP-IFIM values are considered to have the better fit to the given data. The first term is the maximum likelihood estimate of the lack-of-fit of the model to the data (the same as that in AIC): lower values indicate a better fit than do higher values. The second term represents the measure of complexity of the estimated inverse-Fisher information matrix (IFIM) and acts as a penalty. Two times C_1[IFIM] is used here based on the formulation by Bozdogan and Haughton (1998). This complexity or penalty term provides information on the degree of association among the model parameters: those models with lower correlations or associations among model parameters generally have less complex covariance structure and therefore lower penalty terms (Bozdogan

1990, 1994d). Complex covariance structures and high multicollinearity are often undesirable qualities in multivariate models. Thus, $2C_1$[IFIM] provides a way to penalize those models with high multicollinearity and the complex covariance structure and to incorporate this information directly into the model-selection criterion. Consideration of covariance complexity and multicollinearity when using hypothesis-testing methods for model selection often must be done as a separate process from the direct comparison of competing models (e.g., see methods for evaluation of multicollinearity in logistic regression models by Marx and Smith 1990). Equations for calculating C_1[IFIM] for logistic regression were given in Part 4 (equations 4.7 and 4.8).

The usual assumption with logistic regression models is that model variance = 1. Researchers who use logistic regression on biological data typically do not check this assumption (or at least do not report estimated variances if different from one) and most software provides logistic regression output based on variance = 1. However, estimating the variance would be appropriate because many binomial and multinomial data structures can exhibit the undesirable property of over-dispersion (variance > 1). Thus, variance was estimated for each candidate model as the Pearson X^2 value divided by n and incorporated into the calculation of ICOMP-IFIM in the GA (see Part 4 of this dissertation and equation 4.9 for details). All other components of ICOMP-IFIM being equal among competing models, those models with larger variances would have larger ICOMP-IFIM values, thus reflecting a poorer fit, than models with smaller variances.

All 18 variables and the intercept were entered into the GA for each analysis. For the summer analysis, four separate runs of the GA were performed (for a total of 11,300 models) with the following number of runs, number of generations, and models per generation, respectively: 2:50:50, 1:80:60, and 1:30:50. For the winter analysis, three separate runs of the GA were performed, each with 50:50 (for a total of 7500 models). For both summer and winter analyses, the point mutation rate (the frequency of switching of a single, randomly selected bit from either one to zero or vice versa) was 0.01 per generation and the crossover rate (the probability of mating or crossover between two chosen strings) was 0.7. Two highly associated variables in the summer analysis, LDS and SMOS, were not allowed to enter the same summer models together. However, either LDS or SMOS was allowed to enter a model, when randomly selected in a string, if that model lacked the variable's highly associated counterpart.

GA programs, including the calculation of ICOMP-IFIM, were written in MATLAB (The Math Works, Inc. 1989) and executed on either a VAX mainframe computer (University of Tennessee Computing and Administrative Systems) or a Power Macintosh 8100/100 (Department of Ecology and Evolutionary Biology, University of Tennessee). All analyses of the GA output were conducted by the author.

Statistical hypothesis-testing procedures were used to a) provide a familiar point of reference to readers who are unfamiliar with informational criteria and b) supplement the informational modeling approach for examination of certain candidate models. When such procedures were conducted, estimated parameter values (regression

coefficients), Wald statistics, and the associated *P* values were obtained using the LOGISTIC procedure in SAS (SAS Institute Inc. 1989) and an alpha level of 0.10 was used as a "guidepost" rather than as a "Magic Number" (see Toft 1990).

The third stage of the GAIM approach involves a) selecting a "window" of criterion values in order to define the set of best models found by the GA, 2) deciding whether other criteria (such as model variance and/or biological considerations) should be used to further redefine the set of best GA models, 3) plotting the criterion values of the best GA models with respect to *k* (the number of estimated parameters in the model) or other measures of interest, 4) examining the frequencies of independent variables among the best GA models to determine whether some variables are more common than others, and 5) when possible, using diagnostic measures to obtain further insight.

Models having the lowest ICOMP-IFIM values were obtained from the GA runs to form the initial best GA models. Model variance was used as a secondary criterion to further refine this set of models. Models having variances considered to be too large (≥ 3.00) were excluded from further examination. Graphs of ICOMP-IFIM vs. *k* and the frequency of each independent variable were examined (using JMP) in the final set of best GA models. The phrase "best GA models" means that the reported models were the best models that the *GA runs actually found*, not that *all* of the possible "best" models were found. These "best GA" models are simply very good models based on the criteria (ICOMP-IFIM and, secondarily, model variance).

For certain models, the logistic regression diagnostics of ΔX^2_j, ΔD_j, $\Delta \beta_j$, and h_j were used to further examine how well a model fit the data across all observations and to compare specific alternative models. Calculations of ΔX^2_j, ΔD_j, and $\Delta \beta_j$ first involved obtaining residual terms and h_j for each covariate pattern (i.e., each unique combination of observed values of the independent variables in a model) by using certain commands in the LR procedure in BMDP (see Engleman 1988). All residuals and the subsequent diagnostics were thus formulated for each covariate pattern x_j (where **x** represents the observed values of the independent variables and the total J covariate patterns for a model are indexed by $j = 1, 2, ..., J$) rather than for each of the n observations as suggested by Hosmer and Lemeshow (1989:152).

Residual measures obtained from the LR procedure in BMDP were the standardized Pearson residuals, $(r_{Sj})^2$, and the deviance residuals, $(d_j)^2$. These residuals, along with values for h_j and the predicted probabilities for each covariate pattern were put into a spreadsheet in JMP and used to calculate ΔX^2_j, ΔD_j, and $\Delta \beta_j$ based on Pregibon (1981) and Hosmer and Lemeshow (1989:149-156).

The h_j values are diagonal elements of the hat matrix (**H**), a matrix which contains information from both the data (design) matrix and the estimated probabilities. A measure of leverage is provided by h_j. When the predicted probability of a covariate pattern is between 0.1 and 0.9, then large h_j values (upper bound of which is 1) can be interpreted as a large distance from the mean and a large leverage on the estimated parameter values (Hosmer and Lemeshow 1989:154).

ΔX^2_j is based on the Pearson residual and represents the change in the Pearson X^2 statistic when all observations with a given covariate pattern are deleted from the model. ΔD_j is based on the deviance residual and represents the change in the deviance statistic when all subjects with a given covariate pattern are deleted from the model. A large value in either ΔX^2_j or ΔD_j indicates that the particular covariate pattern is poorly fit by the model. $\Delta ß_j$ is a generalized estimate of the standardized change in the logistic regression parameters between the model with all the covariate patterns included and the model with a particular covariate pattern excluded from the model under examination. Thus, large $\Delta ß_j$ values indicate which covariate patterns have the greatest influence on the estimated parameters of a given model (Hosmer and Lemeshow 1989).

Because the distribution of diagnostic measures is not known for most logistic regression cases under the hypothesis that the model fits the data, a graphical assessment of these measures was used following methods outlined by Hosmer and Lemeshow (1989:157-166). Graphs of the diagnostics $\Delta ß_j$, ΔX^2_j and ΔD_j versus the predicted probability were plotted using JMP software. For ΔX^2_j and ΔD_j, focus was on covariate patterns with large values relative to other patterns and where "large" was considered to be > 2.71 (the critical value for X^2 with alpha = 0.10 and df = 1). This definition of "large" for ΔX^2_j and ΔD_j, in order to identify poorly fit covariate patterns, is more rigorous than the definition of > 4.0 (alpha = 0.05 and df = 1) used by Hosmer and Lemeshow (1989:163). Both numerical value and visual inspection should guide the analyst in determining what is "large". Graphically, the analyst should pay attention to any points "...

that fall some distance from the balance of the data plotted." (Hosmer and Lemeshow 1989:162) in plots involving diagnostics.

Both graphical and numerical information was then viewed in combination to find covariate patterns which were not fit well by the model. The greatest concern should be placed on covariate patterns which exhibit a) moderate to high leverage (large h_j values), b) large influence on the values of the estimated parameters (large $\Delta\beta_j$), and c) poor fit (large ΔX^2_j or ΔD_j), all within the range of estimated probabilities between 0.1 and 0.9 where leverage values are expected to be relatively large. Such patterns can have the greatest potential impact on interpretations and conclusions about a model, whereas those patterns which have poor fit and high leverage can be "biologically plausible" depending on their configurations of the covariates (e.g., see Hosmer and Lemeshow 1989:164-167).

RESULTS

Summer models

Anolis carolinensis were present in 45 of 166 plots (27.1%) during the summer survey. Anoles were present in 36 of 51 plots (70.6%) in Habitat A, 6 of 43 plots (14.0%) in B, and 3 of 72 plots (4.2%) in the Pine/Mixed habitat (Habitats C and D combined). A summary of each variable in relation to the presence/absence of *A. carolinensis* is given in Table 5-4 for both categorical and continuous variables. Univariate logistic regression models (with both the intercept and each variable in its final form) had criterion values ranging from 126.60 (for HABS) to 194.60 (for LOSD; Table 5-4). Except for LOSD, each explanatory variable had a univariate ICOMP-IFIM

value less than that of the intercept-only model (194.21), although criterion values for STMD, HSSD, SSSD, and WSUN were close to (within 2-3 units) that of the intercept-only model. This suggested that each of the variables alone, except for LOSD and perhaps STMD, HSSD, SSSD, and WSUN, had a strong to moderate association with the presence of *A. carolinensis* in summer plots. In general, the best univariate variables, based simply on ICOMP-IFIM values, were HABS, DPOR, LDS, WCD, DES, and NLU.

Because of a singularity problem when variables LDS and SMOS occurred together in any model, two different "full" models had to be fit; one without LDS and one without SMOS. Both of the "full" models fit the data better than the intercept-only model by both the classical statistical method (without SMOS: $G = 114.92$, df = 23, $P \leq 0.0001$; without LDS: $G = 114.79$, df = 22, $P \leq 0.0001$) and by the informational method (without SMOS: ICOMP-IFIM = 143.81; without LDS: ICOMP-IFIM = 143.57). However, most regression parameters in both of the "full" models had associated P values > 0.10. For the model without SMOS, 19 of 24 parameters (including that of the intercept) showed $P > 0.10$ and only four parameters showed $P < 0.05$ (range: 0.0010 to 0.0312). The results for the model without LDS were nearly identical. This situation (rather small P value in the model test statistic, but larger values associated with the regression parameters themselves) sometimes occurs in linear regression where it suggests the possible existence of multicollinearity (Neter et al. 1985:278-282, Moses 1986:353-355). The parameter estimates of EVG and ESSD, for the full model without SMOS, were both highly correlated with the parameter estimate of the intercept (r = -0.90 and -0.87, respectively), as

well as with each other (r = 0.95). In addition, the WSUN parameter was moderately correlated with DPOR1 and DPOR2 (r = -0.62 and -0.53, respectively). A similar situation occurred among parameter estimates of the full model without LDS. Despite the presence of some collinearity all 18 variables and the intercept were included in the analysis using the GA (with the one restriction involving LDS and SMOS) in order to examine the utility of using ICOMP-IFIM with the GA in such cases.

The four GA runs produced models with a wide range of ICOMP-IFIM values (103.56-211.58). When criterion values were sorted from lowest to highest, many models differed only slightly (< 0.50) from the models that ranked immediately above or below. Thus, initial examination of the GA models was confined to those models within a specific cutoff value which was defined as 3.00 plus the median criterion value of the single best models from each of the GA runs. The median of the best four criterion values was 103.92; therefore, models with criterion values $\leq$ 106.92 were examined. This produced 115 models (henceforth called the "best summer GA models") used for the examination of the frequency of each variable and trends in the fit of models. Model variance was not used as a secondary criterion for these 115 models because variances were near 1.0 (range: 0.60-0.86).

Possible trends in the lack-of-fit, complexity, and ICOMP-IFIM values across the number of parameters (k) were examined graphically by grouping models according to k, such that each group had at least 5% of the 115 best GA models. Box plots showing the 10th, 25th, 75th, and 90th quantiles and median values for each group suggested that the lack-of-fit

values decreased with increasing k (Fig. 5-1a), whereas the complexity values increased with increasing k (Fig. 5-1b). Obviously, a trade-off exists between these two components of ICOMP-IFIM; those models with $k = 16$ or 17 generally fit the data better than those with ≤ 13 parameters with respect to just the lack-of-fit term, but the larger models tended to have greater complexity values than the smaller models. Considerable overlap in ICOMP-IFIM values occurred, with no clear increase or decrease in this criterion across k for these 115 models (Fig. 5-1c). In general for the best summer GA models, the smaller models (k = 11-12) were at least equivalent to the larger models (k = 16 or 17) in their overall fit to the data because ICOMP-IFIM values of the smaller models tended to be equivalent or slightly smaller than those values for larger models.

Overall, five variables (INT, DPOR, SSSD, STMD, and DAES) occurred in 100%, three variables (SCAN, DES, and HSSD) were in 75.0-99.9%, and two variables (HABS and ROCK) were in 50.0-74.9% of the best summer GA models (Fig. 5-2). Of the nine variables which occurred in less than half of these models, six were related to either the number or size of standing tree trunks (LDS, SMOS, LOSD, NLU, EVG, and ESSD) and two were related directly to winter canopy/sunlight conditions (WCD and WSUN).

Because the smaller models seemed to fit the data as well as the larger models, the frequency of variables among different model size classes was examined. Four size classes of the best summer GA models, k = 11-13, 14, 15 , and 16-17, were designated with at least 25 models per size class (32, 29, 25, and 29 models, respectively). DPOR, SSSD, STMD, DAES, and the

intercept were present in 100% of the models within each of the model size classes (Fig. 5-3). DES occurred in all models regardless of *k*, except for one model in the k = 11-13 group. SCAN appeared in 100% of the models with 15 or more parameters, but in 79.3% and 59.4% of the models with 14 and 11-13 parameters, respectively. The percent occurrence for HSSD was highest in the largest model class, but stayed fairly constant among the other model classes (65.5%-76.0%). HABS was more infrequent (53.1%) in the smallest model class compared to the other size classes (range: 58.6-72.0%). ROCK showed a large decline in occurrence from the k = 16-17 models (69.0%) to the k = 11-13 models (37.5%). The remaining variables all occurred in less than 50% of the models across all size classes with LDS, SMOS, LOSD, and WCD each showing greater than two-fold declines from k = 16-17 to k = 11-13 models (Fig 5-3).

Hypothesis-testing procedures on the estimated parameter values (Wald X^2 tests, SAS PROC LOGISTIC) were used to supplement the previous graphical analyses. Because the data are observational and not based on an experimental design, the estimated logistic regression parameter values and their associated *P* values should be interpreted as approximations rather than as exact or highly reliable values. Models with 16 or 17 parameters typically had four or more parameters each with $P >$ 0.10. This finding, along with the graphical results of Figures 5.1c, 5.2, and 5.3, suggested that simpler models ($k < 16$) should be examined more closely. Eighty-six of the 115 best summer GA models had fewer than 16 parameters.

The variables which were most frequent among the best summer GA models had statistically significant parameter estimates more often than the least frequent variables. To illustrate this point, Table 5-5 simply shows a maximum of ten models for each model size under $k = 16$ with the lowest ICOMP-IFIM values. The variables which occurred in 100% of the best 115 GA models (INT, DPOR, SSSD, STMD, and DAES) were also the variables which always had significant parameters in the best models shown in Table 5-5. The least frequent variables (frequency <60%) among the best 115 GA models always had non-significant parameter estimates (LDS, SMOS, LOSD, EVG, ESSD, WCD, DFW, and ROCK) or were completely absent (NLU) from the best models in Table 5-5. This suggests that these least frequent variables had little to contribute statistically to models which already contained INT, DPOR, SSSD, STMD, and DAES.

Variables such as HABS, SCAN, and HSSD, which occurred with moderate frequency overall (60-85%; Fig. 5-2), had significant parameter estimates in 29.7% (11/37), 65.6% (21/32), and 45.2% (14/31), respectively, of the models in Table 5-5 in which they each occurred. Among these three variables, the one that occurred most frequently among the best GA models, SCAN, had the highest ratio of significant to non-significant parameter estimates among models in Table 5-5, whereas the least frequent variable, HABS, had the lowest ratio. The statistical significance of HABS, SCAN, or HSSD in any model appeared to be related to whether or not certain other variables were present. The estimated parameter for DES2, but not DES1, was significant in all but one of the models in Table 5-5. This suggests that DES either may not really contribute to any given model or

may simply need to be categorized or put in a different form or scale in future modeling efforts. The latter interpretation is the one currently preferred.

INT, DPOR, DES, SSSD, STMD, and DAES more consistently a) occurred in the best summer GA models and b) possessed significant parameter estimates among those GA models than any other variables. Moderate consistency was exhibited by SCAN and HSSD. Together, these eight variables were also the ones which formed Model 1 of the summer GA results. Model 1 has both the lowest ICOMP-IFIM value and the smallest number of parameters ($k = 11$) of all the summer GA models. Interestingly, Model 1 is a subset of 75/114 (65.8%) of the best GA models (deletion of one, two, or three variables from those models leads to Model 1). For example, deletion of either variable 13 (ROCK), 1 (HABS), 3 (LDS) and 13, or 7 (LOSD) and 10 (ESSD), from Models 2, 4, 5, and 7 respectively, all lead to Model 1 (see Table 5-5).

For comparison with Model 1, parameter estimates for the best summer GA models which had 11 parameters, as well as Models 2-19, are shown in Table 5-6. All parameter estimates in Model 1 had associated P values < 0.10 except for DES1. By comparison, at least two parameters had associated P values > 0.10 in the the two other GA models with 11 parameters and among GA Models 2-19 (Table 5-6). These non-significant parameters most often belonged to those variables which occurred less frequently among the best summer GA models, such as HABS, LDS, ESSD, DFW, and ROCK.

Deletion of any single habitat variable from Model 1 produced only models with higher ICOMP-IFIM values (range: 106.34-148.64) than this model. The subset model of GA Model 1 which consisted of INT, DPOR, SCAN, DES, SSSD, STMD, and DAES (HSSD deleted) came the closest to the ICOMP-IFIM value of Model 1. Based on ICOMP-IFIM, Model 1 seems to provide a better fit to the data than any of these subsets. However, because all possible models were not evaluated, some possibility remains that other models (either larger or smaller than $k = 11$) exist which are equivalent to or better than Model 1 for the summer data.

The emphasis on the summer results is not on any one model or on the specific estimates of any given parameters, but rather on the frequency of variables and the frequency of the approximate statistical significance of variables among the best GA models. Overall, the results suggest that those variables which occurred most frequently were also the ones which were most frequently significant based on classical testing procedures. The variables which occurred in 100% of the best summer GA models, INT, DPOR, SSSD, STMD, and DAES, also constituted all but two of the variables which occurred in Model 1 from the GA results.

Winter models

Anolis carolinensis were present in 62 of 166 plots (37.3%) during the winter survey. The frequency of occurrence in plots according to habitat type was: 34 of 51 (66.7%) in Habitat A, 16 of 43 (37.2%) in B, 5 of 51 (9.8%) in C, and 7 of 21 (33.3%) in D. Univariate logistic regression models (containing each variable in its final form and the intercept) had ICOMP-IFIM values ranging from 180.14 to 222.21 (Table 5-7). Each explanatory

variable had a univariate ICOMP-IFIM value less than that of the intercept-only model (219.52), except for NLOW. However, ICOMP-IFIM for the intercept-only model was within two to three units of those univariate models of EVG, SOTW, DFW, HSCW, and SSWD. Thus, each variable alone, except for perhaps NLOW, EVG, SOTW, DFW, HSCW, and SSWD, showed a strong to moderate association with the presence of *A. carolinensis* in winter plots. The best univariate variables appeared to be ROCK, DPOR, HAB, WSUN, WTM, and WCD.

The full model for the winter data had a better fit than the intercept-only model (ICOMP-IFIM = 154.59 vs. 219.52; $G = 105.95$, df = 23, $P \leq 0.0001$). However, the full model had 10 of 24 parameters with associated P values > 0.10, but only ten parameters with P values between 0.05 and 0.0001 and one parameter with a P value ≤ 0.0001. As with the summer data, this situation suggested the possibility of some multicollinearity among the habitat variables. Correlations between the parameter estimates of the intercept and other variables in the full model were moderate ($0.50 \leq |r| \leq 0.63$) in five cases (NLUW, NLOW, SOTW, ESWD, and WSUN) and slightly high in one case ($r = -0.76$; EVG). The correlations between parameter estimates of variables other than the intercept were moderate in four cases (between SOTW and NLOW, SOTW and LDW, ESWD and EVG, and WSUN and WCD).

Despite possible multicollinearity, the full set of variables was used in each GA run for the same reason stated for the summer analysis. ICOMP-IFIM values from the three winter GA runs combined ranged from 143.61 to 229.93. These values differed by less than 0.50 between a given model

and the model ranked just above or below for many pairs of adjacent models. The initial cutoff value for the winter GA models, calculated in the same manner as that for the summer analysis, was 146.97 and a total of 184 winter GA models had criterion values below this cutoff. Within this set of the best winter GA models, four variables other than the intercept (WTM, EVG, ROCK, and DAEW) were present in 100%, five variables (NLUW, DPOR, NLOW, SCW, and DEW) occurred in 75.0-99.9%, and six variables (HAB, LDW, SOTW, ESWD, WCD, and HSCW) each occurred in 50.0-74.9% of those models (Fig. 5-4). Unlike the findings for the summer models, only three variables (DFW, SSWD, and WSUN) were found in fewer than half of the best GA models.

Both the lack-of-fit and complexity terms had a rather linear trend with k; lack-of-fit decreased and complexity increased with increasing k values (Fig. 5-5a,b). In general, the smaller models tended to be equivalent in ICOMP-IFIM values to larger models, but some of the largest models (k = 21-23) had the lowest criterion values (Fig. 5-5c). Overall, 77.7% (143/184) of the best GA models had 17 or more parameters.

Most of the top twenty GA models had 17 or more parameters and/or a variance $\geq$ 3.00 (Table 5-8). For example, GA Model 1 possessed 21 parameters and a variance = 5.80. Only GA Models 2 and 13 out of the top 20 had both $k < 17$ and variance < 3.00. In addition, each model with $k \geq 17$ had three or more parameter estimates which likely contributed statistically little to a model (i.e., associated P value > 0.10). These combined qualities (i.e., large model size, large variance, and possible non-contributing parameters) are undesirable in a model despite a low ICOMP-

IFIM value. However, among the best 184 GA models, one, one, three, five, 11, and 18 models had 11, 12, 13, 14, 15, and 16 parameters, respectively and a variance < 3.00. For example, Models 2 and 13, both had a variance < 3.00 and fewer parameters than the other top 20 GA models. Model 13 is a subset of both Models 1 and 2, but it is the only subset of these models (found with the GA) which had both $k \leq 15$ and ICOMP-IFIM rank less than 40th.

Deletion of two or more of the variables which possessed non-significant parameters in Model 1 produced smaller sized models which had ICOMP-IFIM values < 145.00 and variances < 3.00. It was decided, therefore, that further analysis of the winter models should be conducted by obtaining a limited number of subset models from a selected group of GA models. Another option would have been to conduct two or three more GA runs, possibly altering the rates of mutation and/or crossing over or increasing the size of the population and/or number of generations. Because most of the top GA models were subsets of either Model 1, 2, 3, or 6, these four models served as the parent models for the subset analysis. Only those subsets which included the intercept term were enumerated here because all of the best 184 GA models possessed the intercept. All possible subsets having 11-14 habitat variables were obtained from Models 1, 3, and 6 (which all had 15 habitat variables). For Model 2 (13 habitat variables), all possible subsets having 6-12 habitat variables were enumerated. Results from each of the "subset analyses" on the four parent models were combined and duplicate models were deleted so that a given model was represented only once.

The lowest ICOMP-IFIM value found for any of these subset models was 143.36 (k = 15 and variance = 1.91), which was slightly lower than that of GA Model 1. Because this model with the lowest ICOMP-IFIM value of any of the winter models had 15 parameters and a variance under 3.00, further examination of the winter models was then restricted to those models which satisfied all of the following criteria: a) $k < 16$, b) variance < 3.00, and c) ICOMP-IFIM value below 147.28 (= 3.00 plus the median of the lowest criterion values from the four subset analyses). This produced 154 winter models (referred to as the "final best" winter models) from the combined GA and subset analyses which were below the new cutoff value.

Model variances and the number of parameters among the final best models ranged from 1.25 to 2.83, and from 10 to 15, respectively. Considerable similarity in the lack-of-fit terms (-2 loglikelihood), complexity values, and variances existed among the final winter models. No clear increase or decrease occurred in ICOMP-IFIM values across k among the final best winter models (Fig. 5-6). This suggests that these models are probably similar in their fit to the data regardless of their size. Among the top ten models with the lowest ICOMP-IFIM values, at least one model from each size group was represented.

INT, WTM, EVG, ROCK, and DAEW each occurred in 100%, DPOR and SCW each occurred in over 90%, and DEW occurred in 77.9% of the final best models (Fig. 5-7). HSCW, SSWD, NLOW, and SOTW occurred in 64.3%, 57.1%, 67.5%, and 51.3%, respectively, of these models. Variables which occurred in < 50.0% of the final best models were (in decreasing order of frequency): LDW, WCD, ESWD, NLUW, HAB, and DFW. WSUN

was absent from all of the 154 final best models. Those variables which occurred in over 95% of the best 184 GA models also occurred in over 95% of the final best winter models. The greatest declines in frequencies of occurrence for the habitat variables from the best 184 GA models to the final 154 winter models was seen with HAB (from 67.4% to 6.5%), NLUW (from 82.1% to 27.3%), LDW (from 66.8% to 46.8%), ESWD (from 70.7% to 28.6%), and DFW (from 32.6% to 0.65%).

In order to examine changes in the frequency of variables among model classes of the final best models, four size classes were designated with at least 15% of the 154 total models in each class (k = 15, n = 62; k = 14, n = 40; k = 13, n = 26; k = 10-12, n = 26). Besides the intercept (INT), the variables WTM, EVG, ROCK, and DAEW occurred in 100% and DPOR and SCW were in at least 83.9% and 92.3%, respectively, of each the model classes (Fig. 5-8). WSUN was not found in any of these best models and DFW was found in only one model. Large declines in variable frequency from larger (k = 14 or 15) to smaller models ($k \leq 12$) were observed for NLUW (58.1% to 0%), SOTW (61.3% to 23.1%), ESWD (42.5% to 0%), and WCD (45.2% to 7.7%). More moderate to smaller declines were seen in NLOW (75.0% to 50.0%), LDW (55.0% to 34.6%), HSCW (73.1% to 42.3%), SSWD (60.0% to 46.2%), and DEW (82.5% to 69.2%).

The models with the lowest ICOMP-IFIM values among the final best models for a given model size class (k = 10 through 15; maximum of ten models per size class) are shown in Table 5-9. Model 1 possessed INT, WTM, DPOR, NLOW, LDW, EVG, SOTW, SCW, ROCK, HSCW, SSWD, DEW, and DAEW. The regression parameters of NLOW, LDW, SOTW,

HSCW, and SSWD in Model 1 had associated *P* values > 0.10, but the other parameters were significant (P < 0.10; DEW had a *P* value = 0.10).

Among the models shown in Table 5-9, INT, WTM, DPOR, EVG, ROCK, and DAEW always had significant parameter values (*P* < 0.10); SCW had significant parameter values in all but a few cases across those models. WTM2 was non-significant in all of these models even though WTM1 was significant. As with the summer models, this variable may simply need to be categorized or treated in a different form or scale in future modeling efforts. For the present models, WTM is considered to be relevant because of the potential influence that ambient air temperature has on anole activity in the winter. NLOW, ESWD, WCD, DFW, and SSWD always had non-significant parameter estimates, and LDW and SOTW were significant in only one of the models in Table 5-9. HSCW and DEW were significant in some models, but not others; their statistical significance was probably related to which other variables were present in a given model. Results from Figures 5-6 and 5-8 and the patterns of the statistical significance of parameters in Table 5-9 suggest that those variables which occurred most frequently were also the ones which were most often significant based on classical testing procedures.

Although Model 1 had the lowest ICOMP-IFIM value, other models had fairly similar criterion values, as well as equal or lower numbers of parameters, complexity values, and variances to Model 1 (Table 5-9). For example, Models 2-6 and 8-12 all have criterion values within 1.50 of Model 1, 14 or fewer parameters, complexity values less than Model 1, and variances similar or less than that of Model 1. In particular, Model 4 (*k* =

13, ICOMP-IFIM = 144.20) and Model 5 (k = 12, ICOMP-IFIM = 144.36) appear to be very good alternative models to Model 1. Thus, no single model appears to be vastly superior to all other models, based on ICOMP-IFIM, complexity, and model variance, within the set of final best models for the winter data.

Differences between any two of the top 12 models were mainly due to either inclusion or deletion of either NLOW, LDW, and SOTW (and occasionally WCD), or some combination thereof, from a model. Recall that these variables were ones which occurred at moderate (50-70%) or low (< 50%) frequencies among the final best 154 winter models. It appears that some of these less frequently occurring variables might be able to statistically substitute for one another in the winter models.

Logistic regression diagnostic measures were examined for each of the models which had the lowest ICOMP-IFIM value within size class k = 10, 12, 13, and 15 among the final best winter models. Diagnostic results were similar among these four models, but for simplicity results are presented for final winter Models 1 (k = 15, ICOMP-IFIM = 143.36) and 144 (k = 10, ICOMP-IFIM = 147.12). Graphical examination of diagnostic measures indicated that only a few covariate patterns for either Model 1 or 144 possessed moderate to high values for all three diagnostic measures: leverage (h_j), $\Delta\beta_j$, and poor fit (either ΔX^2_j or ΔD_j). Such covariate patterns are indicated by arrows in Figures 5-9 and 5-10.

Only a small proportion of the covariate patterns of Model 1 had h_j values that fell away from the balance of the data or even had moderately large leverage values; at least 90% of the covariate patterns had h_j values

under 0.20 (Fig. 5-9a). The plot of predicted probability and $\Delta\beta_j$ shows that only six covariate patterns had fairly large values of $\Delta\beta_j$ relative to the other data (Fig. 5-9b). $\Delta\beta_j$ is calculated using both h_j and ΔX^2_j (see Hosmer and Lemeshow 1989:155-156), so a large $\Delta\beta_j$ value could result when either h_j or ΔX^2_j (or both) are moderate to large. Only two of the covariate patterns that had moderate to large values of $\Delta\beta_j$ also possessed moderate to exceptionally large leverage values (points indicated by the arrows in Fig. 5-9b). The other data with large $\Delta\beta_j$ had relatively low leverage values so the major constituent of their large $\Delta\beta_j$ was from ΔX^2_j.

The plots ΔX^2_j (Fig. 5-9c) and ΔD_j (Fig. 5-9d) versus predicted probability for final winter Model 1 indicated that 10% or less of the covariate patterns had diagnostic values greater than the conservative cutoff of 2.71. Only a few such points in both of those plots also had both moderate to large h_j and $\Delta\beta_j$ values (as indicated by arrows in Figs 5-9c, d).

Similar findings regarding diagnostics were observed for final winter Model 144. Only a few covariate patterns had fairly large h_j values, whereas about 11 points possessed leverage values of relatively moderate size between 0.2 and 0.3 (Fig. 5-10a). The plot of $\Delta\beta_j$ versus predicted probability showed that only two covariate patterns fall away from the balance of the data (Fig. 5-10b) and it is only those two points which also have both moderate to large h_j and poor fit.

The plots of ΔX^2_j (Fig. 5-10c) and ΔD_j (Fig. 5-10d) versus predicted probability for Model 144 indicated that seven covariate patterns had diagnostic values greater that 2.71 and also fell away from the balance of the data. However, only one of these covariate patterns also possessed a large

leverage value (= 0.33) and large $\Delta\beta_j$, and only one possessed a relatively moderate leverage (= 0.21) and large $\Delta\beta_j$.

Recall that covariate patterns which have both moderate to large values of leverage (h_j) and poor fit (as measured by moderate to large values ΔX^2_j and ΔD_j), as well as moderate to large $\Delta\beta_j$, can have the greatest potential impact on interpretations and conclusions about a model. Those patterns which have poor fit, but low leverage can be "biologically plausible" (i.e., are unusual, but can have reasonable biological explanations for their values). By such guidelines, both winter Models 1 and 144 fit the data quite well because each model had only two covariate patterns which possessed poor fit, moderate to large $\Delta\beta_j$, and moderate to high leverage.

DISCUSSION

Statistical approach used in this study

Unlike the conventional methods used to analyze observational multivariate data, namely stepwise algorithms with hypothesis-testing procedures, the present study demonstrates the genetic algorithm-informational modeling (GAIM) approach which:

1. takes advantage of a GA's ability to evaluate many more models in one overall analysis of a vast model space than could be evaluated by stepwise algorithms,
2. uses an informational model-selection criterion as the primary criterion, rather than hypothesis-testing methods, to rank and choose models,
3. emphasizes the need to examine the frequency of independent variables among a set of "best" models found by the GA, and
4. de-emphasizes the often misdirected activity of choosing a "single-best" model for observational multivariate data.

GAs have a proven record of finding valuable solutions to complex problems that have extremely large numbers of potentially good and bad solutions (e.g., see Holland 1992a, b, Forrest 1993, Goldberg 1994), mainly because of the way in which the actual searching for solutions takes place. GAs can treat each solution to a specific problem as a unique "string" of information, similar to the genetic information on a chromosome. Then, a randomly selected population of strings or solutions is evaluated on its ability to address the problem by means of a defined "fitness" function or criterion. Fitness being used in a sense similar to that of Darwinian or evolutionary fitness. The strings are then "mated" in a manner mimicking chromosomal crossing-over during meiosis. The offspring are evaluated for their "fitness" or ability to solve the problem.

Whether or not an offspring string is selected to form a new mating pool is based on the fitness value of the string. Those selected strings are mated and another "generation" of solutions is produced. This process continues for many generations in a process mimicking biological evolution, such that very good solutions to a problem are uncovered. Readers are referred to Goldberg (1989, 1994) and Holland (1992a, b) for both overviews and more details about GAs.

The GAIM approach described in Part 4 and its application in the current study show how a GA can be used as an alternative to stepwise selection algorithms. Two factors are important for using a GA as a searching algorithm in problems of statistical modeling: the ability to represent a statistical model as a string of variables and the ability to evaluate and rank models by some fitness function or criterion. It is a

simple process to represent models as strings of information. It is also fairly easy to define the fitness function by using an informational model-selection criterion, such as AIC or ICOMP-IFIM.

GAs can do two important things over stepwise algorithms. First, GAs can form new models by adding or removing two or more variables at a time, whereas stepwise algorithms typically only handle changes in one variable at a time. Second, GAs used in conjunction with an informational criterion can compare any models, unlike stepwise algorithms which compare only nested models when used in conjunction with statistical hypothesis-testing procedures. Statistical modeling in the present study primarily utilized an informational approach instead of hypothesis-testing procedures. This is in keeping with the growing philosophy that statistical modeling is not a hypothesis testing problem per se, but rather a problem in optimizing some criterion for model selection (see Sakamoto et al. 1986, Bozdogan 1987, 1988a, b, Burnham and Anderson 1992, Lebreton et al. 1992). Model selection criteria are often formulated to include measures of both the lack-of-fit and model size or complexity into one numerical summary value.

Because of the two attributes mentioned in the paragraph above and the nature of its searching operations, a GA can provide a much wider and diverse search of the model space and produce many more models in one analysis that a stepwise algorithm. Granted, a GA which operates for more than just a few "generations" takes longer to run than a single stepwise search on a given data set. However, the analyst must consider speed versus the total amount of information gained.

Stepwise procedures have been used regularly in the multivariate analysis of ecological data, but their use and misuse have been criticized (see James and McCulloch 1990 for an overview, as well as Part 4 of this dissertation). One important potential misuse is when an analyst uses stepwise algorithms to search for a single "best" model (James and McCulloch 1990). When a data set has many variables it is unlikely that a single model will be clearly superior over all other models (Gorman and Toman 1966, Hocking 1983, McCullagh and Nelder 1989:8).

The GAIM approach emphasizes the need to find a *set* of very good models rather than to find a single "best" model when it comes to analyzing multivariate observational data. With such a data set it is unlikely that a single best model will exist to capture the essence of the data. By finding a set of good models and reporting them as such, the analyst shows the scientific community a wider variety of models than would be shown from results of stepwise algorithms. In addition, other scientists can consider and evaluate a wider variety of models found by the GAIM approach, instead of having to accept a supposedly single "best" model reported from the stepwise searching.

Likewise, by reporting a set of best models rather than a single best model researchers provide others with both a *wider view* of the researchers' findings and the opportunity to compare models from subsequent studies against a *richer set* of original models for purposes of both external validation and formulation of initial causal models. The validation process is much more limited in scope if each researcher of a

particular phenomenon only reports a single model which, if the study used stepwise selection, is likely to be just one of a number of good models.

Another problem with observational multivariate data is that analysts will often interpret the uncovered correlations or associations in terms of causation and/or confirmatory conclusions. Although some statistical methods can be confirmatory (i.e., statistical inferences can be made and extended beyond the sample to a larger population), confirmatory conclusions are valid only under certain conditions (see, e.g., Snee 1983:230, Tukey 1980, James and McCulloch 1990). Typically, such conditions do not exist with observational (non-experimental) data.

The GAIM approach adheres to the fact that observational multivariate data should be analyzed in an exploratory fashion and/or correlational manner, whereby the analyst uses the results of the analysis to propose possible hypotheses and questions for further research. This has been suggested as the appropriate role of observational data in ecology (see James and McCulloch 1985, 1990). Controlled experiments, not observational studies, are best suited for addressing and testing hypotheses regarding causal mechanisms (James and McCulloch 1985, 1990, Lubchenco and Real 1991). The GAIM approach provides a researcher and the scientific community that will scrutinize the research with a wider view of the data by providing more models and a frequency distribution of variables among the best GA-produced models. These outputs from the GAIM approach can potentially provide more insight into the data so that a richer process of generating hypotheses and suggesting experimental studies can take place.

Such an "abductive" process (i.e., that of proposing hypotheses) is an important role of observational studies.

In addition to using the GAIM approach, the present study verified and took into consideration specific assumptions to logistic regression analysis which seem to be either overlooked or unreported by many analysts. Specifically, analysts using logistic regression should examine whether:

1. continuous variables are linear in the logit, and
2. model variance differs considerably from 1.0.

Consideration of these two factors can be performed in association with or directly part of the GAIM approach.

This study checked the assumption of linearity in the logit for all of the continuous variables and changed the form of those variables which were non-linear in the logit prior to performing model selection. Hosmer and Lemeshow (1989:84-86, 89-91) mention this logit assumption and provide some suggestions, such as transforming or categorizing the variable, to overcome violations. Checking this assumption could be done either before model selection begins or after a final model (or set of models) has (have) been selected and further refinement of the variables is an issue.

A variable with a distinct non-linear association (in the logit) with the dependent variable could be excluded from models during the model building/selection process based on only its non-linear logit form and not on its true association with the dependent variable. The best stage to check this assumption might then be before starting the model selection process. This was the approach used in this study in order to give each variable the fullest opportunity to enter models in the selection process. The logit

response function, as determined by the animal's association with a given habitat variable, should not automatically be assumed to be linear over the range of the continuous habitat variable.

Many ecologists who use logistic regression are not reporting any examination of the logit assumption for continuous variables (see, e.g., Buehler et al. 1991, Burger et al. 1994, Diller and Wallace 1994, Larsen et al. 1994, Chandler et al. 1995, Coker and Capen 1995, DeLong et al. 1995, Drewien et al. 1995, Gorenzel and Salmon 1995, Nadeau et al. 1995, Hinsley et al. 1996, Kindvall 1996, Munger et al. 1998). Either the assumption is not being checked or it is not being reported, but one cannot tell by simply reading the publications previously mentioned. Researchers should be sure to check this assumption or at least decide *a priori* whether or not non-linearity in the logit is important from both a statistical and biological viewpoint. Otherwise, relevant habitat variables measured on a continuous scale might be incorrectly excluded from models simply because the variables were not linear in the logit.

Model variance is assumed to equal one for logistic regression models because of the binomial nature of the outcome variable (McCullagh and Nelder 1989:124-126). Because ecological binomial data may often exhibit overdispersion (variance > 1.0), model variance should be estimated. In this study, the variance of each model was estimated instead of assuming model variance = 1. Researchers likely assume a model variance of 1 when using logistic regression in animal-habitat studies as no mention of this variance is being reported (see, e.g., Brennan et al. 1986, Capen et al. 1986, Johnson and Temple 1986, Smith and Connors 1986, Diefenbach and Owen

1989, Buehler et al. 1991, van Manen and Pelton 1993, Burger et al. 1994, Diller and Wallace 1994, Chandler et al. 1995, Coker and Capen 1995, DeLong et al. 1995, Gorenzel and Salmon 1995, Nadeau et al. 1995, Kindvall 1996, Munger et al. 1998). A discussion by Trexler and Travis (1993) of the merits of logistic regression in ecological research also fails mention this assumption or what to do if model variance is greater than one.

The variance assumption may seem to be a minor point, but it can influence the selection of models. That is one reason why some researchers performing model selection on capture-recapture data, where others often assume model variance equals one because they are using multinomial models, recommend estimating model variance and incorporating these estimates into the model selection process (Lebreton et al. 1992). In the current study, an estimate of variance for each model was both incorporated into the model's criterion (ICOMP-IFIM) value and considered in its own right when selecting the best set of models after ICOMP-IFIM values were used to tentatively select the best models. It should be further emphasized that hypothesis-testing procedures do not incorporate information about extra variance into the selection of an appropriate model in logistic regression analyses.

Summer models

The most frequently occurring variables among the best set of GA models describing the associations between habitat features and the presence of *Anolis carolinensis* in summer plots were (including the intercept): distance to potential overwintering rock (DPOR), summer canopy categorization (SCAN), distance to habitat edge (DES),

herb/shrub/vine cover (HSSD), summer sunlight index (SSSD), ambient temperature (STMD), and standardized distance along the habitat edge (DAES). These were also the same variables which most frequently possessed significant parameter estimates (by classical hypothesis-testing methods) and which occurred in the model with the lowest ICOMP-IFIM value.

It should be noted that the following interpretations of habitat variables and their parameters are not being suggested as reflecting the biological "importance" of the independent variables as is incorrectly done in many multivariate analyses (see James and McCulloch 1990:136-138). Rather, the interpretations here only suggest *how* the probability of the presence of *A. carolinensis* in a summer plot *might* be related to these habitat variables in a biological manner. Methods using experimental control, or at least partial control, over field variables would be needed to possibly interpret the importance and/or causal aspects of the variables and their parameter estimates. Such rigorous interpretation is not attempted here because of the observational nature of the data.

Anolis carolinensis has been described as an "edge" species (e.g., see Gordon 1956) meaning that it is often associated with ecotones or habitat edges where vegetation is present, but any overstory canopy is open or thin enough to allow at least moderate amounts of sunlight to reach the lizards' habitat patches. Though all four sampled habitats did have a considerable "edge" component, plots were sampled as far as 40-45 m from the habitat edge in three of the habitats and in closed, partially open, and open canopy structures. The combination of variables that occurred most frequently

among the best summer GA models seems to qualitatively support the general characterization that *A. carolinensis* is an edge species.

Given the other variables in the model, SCAN, DES, SSSD, and to some extent DAES, reflect the possible associations between *A. carolinensis* and edge conditions and/or canopy gaps in woodland habitats. The specific categories used for SCAN and SSSD and the positive values of their respective estimated parameters suggest that this species might be associated more so with open and partially open canopies than closed canopies and more so with moderate to moderately high than extreme (low and high) levels of sunlight in the habitats sampled. In Louisiana, *A. carolinensis* was most often found in open areas, clearings, and along edges during the breeding season rather than in dense, shaded woodlands (Gordon 1956).

DES provides a categorization of the distance to the distinct habitat edge present in each of the habitats (see Table 5-2). *Anolis carolinensis* was most likely to be found at the intermediate distance of 8-14 m from the habitat edge (DES2 category). The large (non-significant) *P* value of the DES1 parameter when this variable occurs with other variables among many of the top models suggests a need to restructure DES, possibly so that it consists of just two categories in future models. It is possible that other variables in the model may influence the parameter value of DES1, especially if the DES1 category was associated with less suitable habitat features (such as farther distances from rock).

Lizard body temperatures can be affected by environmental factors such as ambient air temperature, substrate temperature, and solar radiation in

quantifiable ways (see, e.g., Bartlett and Gates 1967, Porter and Gates 1969, Porter 1989). The two sunlight/canopy-related variables, SCAN and SSSD, and DES are possibly appearing in the top summer models because sunlight can be a resource for the ectothermic *A. carolinensis*. Air and substrate temperatures and solar radiation can be influenced by canopy cover and the amount of sunlight reaching potential perch locations (e.g., shrubs, tree trunks below canopy level, and lower parts of the tree canopy) in wooded habitats.

Measurement of SSSD incorporates information regarding sunlight and both overstory and understory canopy cover, perhaps more so than SCAN, because it was measured as the sunlight striking a horizontal surface at approximately 1.5 m above the ground (see Tables 5-1 and 5-2). The positive parameter estimates of SSSD among the best models suggest that *A. carolinensis* might be less likely to occur in the highest or lowest sunlight patches than in patches with moderate sunlight. Such patches with moderate sunlight could potentially provide lizards with more opportunities to shuttle between sunlight and shade in order reduce the chances of overheating or being at suboptimal body temperatures for very long.

Sunlight and canopy might also play a role in such non-thermal aspects of the ecology of *A. carolinensis* as visual acuity, color perception, and communication. This species uses physical displays, including head bobbing and extending the colored throat-fan, as a means of intraspecific communication in its natural habitats (Gordon 1956, Jenssen et al. 1995). Being diurnal, *A. carolinensis* probably has greater visual acuity in well

lighted conditions than in shaded ones. Color and its perception in wooded habitats are influenced by the amount of light and canopy cover (Endler 1993). Thus, sunlight and canopy cover might play a role in the perception of color and visual communication in *A. carolinensis* during the spring and summer seasons when courtship and territorial defense are taking place.

Ambient air temperature (STMD) in the summer plots was one of the weaker variables among univariate models in terms of its association with the presence of *A. carolinensis*. However, the fact that this variable appears in all of the best GA models suggests that it may provide useful information when in conjunction with other habitat variables. The positive parameter estimates of STMD suggest that *A. carolinensis* might be more likely to occur in plots with moderate temperatures than cooler (< 25.5°C) or warmer (> 28.8°C) ones. The range of temperatures during the actual surveys was not as great as the full range of temperatures occurring over any given summer day. However, it is possible that STMD appears in so many models because it interplays with habitat features in ways not directly measured here and/or stands in for some habitat or microclimatic parameter that was not measured.

Herb/shrub/vine ground cover (HSSD) occurred frequently among the best summer models. The positive parameter values, given the other habitat variables also present in the best models, suggest that *A. carolinensis* might be more likely to be associated with lower to moderate ground cover than with either the lowest ($\leq$ 19.0%) or the highest ($\geq$ 42.0%) amounts of such cover. Whether these intermediate levels of ground

cover are related to aspects of feeding, detection of ground predators, or thermoregulation are not known. It is possible though that adults have home ranges possessing moderate ground cover, but which are near or adjacent to areas with more ground cover where hatchlings and juveniles could have close access to high amounts of such cover. The presence of adults, not juveniles, was the observed outcome variable in the summer plots and adults perch more often on trees than juveniles which often perch on herbs and shrubs (Part 4 and unpublished data).

Distance-related variables DPOR and DAES, and even SCAN and SSUN, given the other variables also in the best models, suggest that the presence of *A. carolinensis* might be associated with habitat features on a spatial scale larger than an individual's summer home range. SCAN and SSUN were measured in a way that shows that sunlight in a plot and canopy cover are related to canopy structure beyond the scale of a plot's own size (see Table 5-1). The amount of sunlight reaching an anole's location along a slope or bluff area is influenced by the degree of canopy openness or closure both adjacent to and directly above the anole's location. Thus, the scale beyond the home range itself can have some relationships to sunlight levels in the summer home range or territory.

Distance to potential over-wintering rock (DPOR) is probably not thermally critical to green anoles during the summer. However, summer home ranges might occur fairly close to potential overwintering rocks if survival advantages occur to minimize the distance needed to travel in order to reach such rocks once overnight temperatures begin to drop in autumn. Positive parameter estimates for DPOR suggest that the

probability of occurrence of *A. carolinensis* is highest when the distance to potential rock shelter is $\leq$ 10 m, followed by distance category 11-20 m (DPOR2). In Louisiana, median distances moved by marked individuals to over-wintering sites were 43 ft (~13.1 m) at Bridge City and 68 ft (~20.7 m) at Plauche (Gordon 1956:167). Slightly larger distances to winter cover for Louisiana anoles could reflect either behavioral differences or actual physiological differences from Tennessee anoles, perhaps in response to autumn climatic differences between these locations.

In many habitats and microhabitats in Tennessee, fewer anoles are seen in locations far from rocks (personal observations). Although a maximum distance which could be travelled by this small, ectothermic lizard to reach a suitable overwintering site is not known, it seems reasonable to suggest that survival might be enhanced by locating a summer home range as close to such sites as possible. The best strategy might be to remain rather close to the rock shelter during summer so that an anole would not get caught out in a fast moving cold front in autumn. This distance to rock shelter (or other suitable cover) could possibly be associated with some locomotor or behavioral limitations of *A. carolinensis* to move to such shelter when overnight temperatures begin to first decline significantly. Adult *A. carolinensis* typically begin moving to overwintering rocks around mid-September (personal observation).

The high frequency of DPOR in these summer models might suggest that these lizards have a limited distance that they venture away from potential overwintering rocks, even in the summer. An alternative interpretation is that the rocks themselves create open areas or gaps in the

canopy because of the difficulty trees might have in growing on or between closely clustered rocks. Thus, lizards might associate more closely to rocks because of the canopy gaps produced and a greater availability of sunlight during the reproductive and growing season (mainly spring and summer).

Another distance-related variable, distance along the habitat edge (DAES), appears in all of the top summer GA models. Green anoles were more likely to occur in plots away from the east and west ends of the sampled habitats than near those ends, as suggested by the positive parameter value and the specific categorization used for DAES (see Tables 5-6 and 5-2). All four of the habitats have either distinct habitat changes or transitions with other habitats at their east and west boundaries. DAES possibly suggests that *A. carolinensis* shows a gradient in its distribution in these riparian and edge habitats.

The central parts of these habitats might have the more suitable habitat features, whereas the east and west ends of each habitat transition or change into habitats less suitable for these lizards. For example, Habitat A (river bluff) shows both a gradual decline in the amount of rock present and an increase in canopy cover at its western end. Habitat B exhibits a decrease in available rock at its east end and Habitat D makes a fairly abrupt turn to the north along a small body of water not far from its east end. In designing this study the actual areas to be sampled within each habitat were chosen so as to avoid any overlapping with or distinct changes into adjacent habitats. However, *A. carolinensis* might respond to transitions or changes into adjacent habitats over a distance larger than that anticipated when the boundaries of the sampled areas were first delineated.

Observational data and interpretations of such data are best suited for providing preliminary insight into a phenomenon or pattern and suggesting possible causal questions or models to test by means of experiments (James and McCulloch 1985, 1990). The summer analysis presented here on *A. carolinensis*-habitat relationships attempts to adhere to these roles of observational research. The exploratory analysis of the summer data suggests that overall factors to examine in future research on *A. carolinensis* might be a) sunlight and thermal factors and b) habitat features related to certain spatial scales beyond the actual home range scale of this species.

Experimental manipulations of canopy cover in summer home ranges could be performed. Patches that supported an adult male and one or more adult females during one or more summers could be shaded in different amounts. The plots would then be surveyed for changes in anole occurrence over one or more summers following manipulation to examine the possible effect of shading on probability of occurrence of *A. carolinensis*.

Many studies of *Anolis* lizards encompassing a habitat or niche component have focused on a spatial scale equivalent to the local perch area or territory (e.g., Rand 1964, Schoener 1968, 1975, Jenssen 1973) or have been conducted within one study site or habitat. The present study suggests that future studies of *A. carolinensis* should perhaps consider the importance of various spatial scales when examining this lizard's ecology and patterns of habitat use. Field and/or lab experiments might be able to estimate the extent to which the presence and/or abundance of *A.*

carolinensis is associated or even influenced by habitat variables at different spatial scales.

One approach to this spatial aspect would be to alter the distances from overwintering rock to home ranges/territories that were known to be used each summer. This could be performed by hiding the rocks from view and accessibility from the anoles or by making the rocks of low quality winter shelter (e.g., put structures in to keep the rocks heavily shaded at all times). Only the habitat of the rocks and extremely close to rocks would be altered. Would anoles continue to utilize summer territories which themselves continue to be good summer patches, but which are now located 20, 30 or 40 m from suitable sunny winter rocks because the closer rocks are now shaded and unsuitable for winter shelter and basking sites?

Another experimental approach, highlighting both the spatial aspect and possible thermal and biological relevance of overwintering rocks would be to supplement habitats with such rock shelters. A sample of rather similar habitat patches could be selected, whereby the patches are fairly far from overwintering rocks and the probability of occurrence of *A. carolinensis* is fairly low. Half of the sample would have artificial overwintering rock shelter constructed within or very near each patch, whereas the other half of the sample remained as it was (control group). Presence of *A. carolinensis* before and after the habitat manipulations would be measured to see the effect of the manipulations on occurrence of this anole. In addition, thermal aspects of the habitat patches would be measured before and after the manipulations by means of thermocouples

and/or thermocouples inside copper lizard models. The thermal data would also be compared between the two groups of patches.

The variables DES and DAES may suggest some relationship of *A. carolinensis* occurrence to the larger physical features of the habitat itself. The locations directly on the habitat edge and near the east and west boundaries, where the habitats transition into other habitats, might be less suitable to anoles than core locations. To test this, one could alter habitat features at the east and west boundaries to make the locations more suitable and census plots in such locations before and after manipulations for changes in the presence of *A. carolinensis*. The alterations should be done in a manner to make habitat features in the boundary plots more like those in the more centrally located plots. In turn, centrally located plots (in terms of DAES) could be modified to resemble the initial state of the plots at the east and west habitat boundaries of the sites used in this study.

Habitat models alone will not likely provide useful tools to effectively predict the occurrence or abundance of a species, particularly when based on observational data. Experimental approaches are very much needed. Some experimental studies of *Anolis* habitat use have been performed (see Sexton and Heatwole 1968 for an outdoor caged-experiment). In addition to experimental studies using habitat manipulations, more field research is needed in the areas of thermal relations, biophysical ecology, energetics, food/prey availability, and community ecology in order to better understand any patterns in the distribution and abundance of *A. carolinensis* in the northern part of its range. Useful research on this anole would include studies taking a similar approach to that of Riechert and

Tracy (1975), Dunham et al. (1989), and Porter (1989) whereby researchers try to relate biophysical and physiological ecology to reproductive output and population ecology.

Winter models

The most frequently occurring variables among the set of best models describing the associations between habitat features and the presence of *Anolis carolinensis* in winter plots were (including the intercept): ambient temperature (WTM), presence of live overstory evergreen tree trunks (EVG), presence of overwintering rock (ROCK), distance along the habitat edge from west boundary of habitat (DAEW), distance to potential overwintering rock (DPOR), and canopy cover categorization (SCW). All of these variables occurred in 100% of the best final winter models, except for DPOR and SCW which occurred in over 90% of those models (all seven of these variables were also the ones which most often possessed significant parameter estimates via hypothesis-testing procedures). This suggests that at least three habitat aspects in combination seem to be related to the occurrence of *A. carolinensis* in winter plots: a) shelter and potential basking sites (ROCK, DPOR), b) sunlight availability and temperature (EVG, SCW, WTM), and c) spatial features (DAEW, and to some extent DPOR).

Several winter variables which occurred in approximately 50-80% of the final model set, DEW, HSCW, SSWD, NLOW, and SOTW, did not frequently have significant parameter estimates. However, they might account for some information not completely accounted by the more frequent variables. These less frequent variables do seem to lend some support to the three key habitat aspects mentioned above. DEW has a

relationship to spatial features and SSWD, NLOW, and SOTW has some relationship to canopy and sunlight conditions. These last three variables might be capturing some information related to summer sunlight/canopy, more than winter, because they measure summer sunlight in a plot (SSWD) and number (NLOW) and size totals of deciduous trees (see Tables 5-1 and 5-3). Such information might have relevance to winter conditions and/or suggest spatial and temporal relevance of such information to the lizards for their use of summer and winter home ranges that are likely to be in close proximity (< 20 m apart). As for HSCW (herb/shrub/vine foliage ground cover), it may be weakly associated with habitat use of juveniles and subadults, both in winter and summer, as such individuals tend to be associated with that type of habitat structure more than adults (unpublished data).

What can be stated about the possible relationship between *A. carolinensis* occurrence and the three aspects of habitat uncovered by the analyses? First, the probability of *A. carolinensis* occurring in a patch is associated with available overwintering rock, given the other habitat variable also in the models. Shelter from cold temperatures is needed by *A. carolinensis*, as it is for ectotherms living in any seasonal temperate zone. The presence of rocks which possess crevices can provide shelter from cold winter temperatures, both during the day and night.

Rocks along south-facing slopes can also provide potential basking sites for *A. carolinensis*. This anole emerges from refugia on sunny winter days at sites in this study as well as other sites across the species range (Ragland et al. 1981, Gatten et al. 1988, Jenssen et al. 1996). Green anoles in eastern

Tennessee will perch in sunny areas, basking on rocks, bare soil and sometimes fallen logs (unpublished data) and on various structures in other parts of its range (e.g., Louisiana - Gordon 1956). Perhaps *A. carolinensis* does not possess the physiological capabilities to become dormant, as do other lizards in temperate zones, because it is of tropical origin (i.e., phylogenetic constraint). Research is needed to address such physiological capabilities or limitations of this species relative to typical winter dormant lizards.

Second, sunlight and temperature appear to be associated with the occurrence of *A. carolinensis* in winter plots. This is directly related to the aspects discussed above regarding basking and lack of complete dormancy in winter. This species appears to have a need to bask and raise its body temperature. Lab experiments show that seasonal gonadal recrudescence is enhanced by daily fluctuations in temperature and high body temperature (T_b around 32 ^{o}C; Noeske and Meier 1977, Licht 1971, 1973). However, feeding in winter does not seem to occur very often for *A. carolinensis* in South Carolina (Jenssen et al. 1996) or eastern Tennessee (personal observations), although the number and diversity of arthropods active on and near the rock faces on sunny winter days suggests that anoles have the opportunity to do so. Raising T_b to higher levels than those typically observed in the field would impose considerable metabolic costs to these lizards in winter(Jenssen et al. 1996). Based on such considerations and on results from their work, Jenssen et al. (1996) hypothesized that northern *A. carolinensis* compromise between raising *T*b high enough for physiological

purposes (perhaps gonadal recrudescence) and too high to inflict detrimental metabolic costs.

Regardless, because of physiological needs and/or possible phylogenetic constraints to become dormant, *A. carolinensis* likely requires winter habitat (including microhabitat) features that provide *both* thermal and solar radiation resources to allow for elevated T_b during the day and shelter from the cold at night (and on cloudy days). Exposed rocks, with crevices, located along south-facing slopes provide such features. Also, the microclimate around the rocks is warmer than surrounding habitat; where many rocks or a bluff occur the thermal environment is usually warmer than adjacent areas without many rocks. The rocks can also act as a heat sink late in the day and into the evening hours in winter. Exposed tree roots or logs with cracks can also provide shelter and basking sites for *A. carolinensis*, but more anoles tend to utilize rocks and bluffs during winter in eastern Tennessee (unpublished data).

Third, spatial features such as DAEW, DPOR, and possibly DEW may be useful in describing associations between *A. carolinensis* and habitat features. DAEW might be of relevance to winter anoles for the same reasons as explained for the summer models. That is, the west and east ends of the study sites here typically showed transitions into habitats that were less suitable for or absent of anoles.

Even though ROCK occurred in all of the models in the final winter set, DPOR still appeared in most of those models. Perhaps the mere characterization of presence/absence of rock (ROCK variable) in a plot was not sufficient information. Some patches of habitat did not have

"potential overwintering rocks" (i.e., those possessing crevices or holes - see Table 5-1) within them, but did have bare soil, fallen logs, and small flat rocks upon which anoles could bask. Such plots were sometimes within a few meters or so of potential overwintering rock (such as plots in category DPOR1 - see Tables 5-3 and 5-7). Thus, DPOR appears to contribute to the winter models in a way that also suggests a spatial component to the ROCK variable.

Overall, these results and interpretations can suggest hypotheses and questions to be examined by future research. One question to examine, as suggested by this research and other field observations, regards the interplay between rocks, sunlight availability/canopy structure, and thermal features of habitat patches. Artificial rock shelters could be constructed in the field, some of which are unshaded (control) and others which are partially shaded treatment, and placed in replicate habitats. Alternatively, the researcher could partially shade existing overwintering rocks and leave others in sunlight. The treatments and control rocks could be surveyed for the presence and abundance of *A. carolinensis*.

Lizard body temperatures and air and substrate temperatures would be recorded. In addition, copper lizard models with thermocouples could be placed in all the treatments and controls to obtain data over many days under various climatic conditions during winter. The data would provide a means to compare the treatment and controls on the basis of lizard and lizard model body temperatures. Using these data, biophysical modeling could provide profiles of possible thermal advantages of unshaded versus partially shaded rocks.

Using a similar design, one could individually enclose each treatment and control and stock the enclosures with male-female pairs. If raising body temperature does influence gonadal recrudescence, then habitat manipulations related to shaded versus partially shaded rocks, and subsequent differences in thermal microhabitat conditions, could be used to test such effects on the level of recrudescence or onset of reproduction in spring and number of eggs (= clutches) produced in *A. carolinensis*.

The winter models frequently possessed the presence of overwintering rock (ROCK) and both summer and winter analyses showed that DPOR was one of the most frequently occurring variables among the best models found. In locations where rocks with crevices are absent or far away (or at distances not easily perceivable by these anoles), anoles probably use dead stumps or trees, fallen logs, or along roots of trees below the surface of the ground for overwintering sites. If the thermal quality of such alternative sites is lower than that of rocks, then the probability of *A. carolinensis* being in an area far from rock crevices might be lower than in areas close to such cover. Studies of the thermal differences among the various winter shelters and winter basking sites, the physiological costs and benefits of the use of these microhabitats, and the ability of these lizards to move to the shelters across a wide range of distances are needed. Such studies would require an approach based on biophysical and physiological ecology.

How do population estimates, reproductive output, and overwintering survival rates differ among habitats, such as the four different habitats used in this study? Such parameters should be compared among habitats which differ in the availability of rocks with crevices for winter shelter, canopy

cover, and solar radiation levels. Do different habitats also differ in important population and reproductive parameters? The fact that *A. carolinensis* occurs in a variety of microhabitats and habitats, but with apparent different probabilities of occurrence, in eastern Tennessee presents a great opportunity to try to link biophysical models with reproductive and population parameters in a manner following research by Dunham et al. (1989) and Porter (1989).

Although presence-absence data were analyzed in this study, densities of *A. carolinensis* per winter plot varied both within and among the four habitats. Given the relatively warm winters that occur in eastern Tennessee, could anoles survive which occurred in plots that lacked overwintering rock (or were greater than some minimum distance from rock) if the winter happened to be particularly severe? This point, again, suggests that the thermal quality of various overwintering structures should be examined, but also in context of the severity or mildness of a given winter season. It is also possible that the presence of *A. carolinensis* in some winter plots could be a function of the generally mild winters that are occurring. Thus, plots or habitats that lack overwintering rock might represent "marginal" patches or "sink" habitats, respectively, when winters are relatively mild. During severe winters, such plots or habitats might not support *A. carolinensis*. In addition, could the winter survival of *A. carolinensis* in plots that lack rock or are relatively far from rock be a key to northward range expansion of the species. Using an approach similar to that taken by Porter and Tracy (1983) on the distributional limits of the desert iguana (*Dipsosaurus dorsalis*), evaluation of habitat features,

measurement of climatic and microclimatic variables, and biophysical modeling could possibly help address these questions.

Relationships between habitat features and distribution of *A. carolinensis* should be examined on large spatial scales. Given the results of this study and what can be uncovered from additional studies like this one, it might be expected that the availability of rock crevices and/or bluffs, levels of solar radiation, and profiles of winter temperatures at various locations and spatial scales would be related to the geographic distribution across the northern part of this species range. What can such variables tell us about the northern limits to the range of *A. carolinensis*? What will happen to these northern distributional limits with changes in the global climate? Biophysical modeling, combined with habitat and environmental measurements, might help answer such questions.

The analyses conducted in the present study utilized data from four habitats in eastern Tennessee at the northern distributional limits of *A. carolinensis*. Such analyses should be conducted in similar habitats at different latitudes between, for example, Tennessee and southern Georgia. The lower latitudes would probably have rather different variables appear in the models and would likely lack variables such as presence of overwintering rock and distance to rock. A comparative approach such as this could provide insight into different associations between the occurrence of this anole and various habitat features. Physiological data and biophysical models to accompany the habitat analyses would provide quantitative ways to compare important physiological and thermal

parameters related to similar structural habitats along a latitudinal gradient.

Regarding the statistical analysis itself, the winter data analysis proved to be a more elaborate process than for the summer data. Perhaps this was because of more collinearity among winter variables and/or the possibility that more winter variables were able to substitute for one another than in summer data. Regardless, subset analyses on the winter GA results were conducted to find smaller models having equally good fit as the slightly larger models first found by the three GA runs. Perhaps additional GA runs were needed to uncover smaller models. Similarly, larger numbers of generations or sizes of each population may have produced better results given the somewhat complex nature of the winter data. More frequent crossovers and/or mutation rate may have been needed.

The point is that a single, standard GA is not going to produce perfectly tidy results for every data set. One alternative approach is to run one pass with the GA, examine the preliminary results, and check whether or not some smaller subset models fit the data better (in terms of their informational criterion values). If a good number of smaller, better models were missed, then parameters in the GA such as mutation rates, crossover rates, population size, and/or generation size, could be altered. In addition, a researcher could use a more sophisticated GA than the one used in the present study (computer scientists are now producing a wide variety of GAs). It is just a matter of more research being needed to determine the full applicability and range of potential benefits of using the more sophisticated GAs for statistical modeling.

Non-habitat factors

The current study examined possible associations between the presence of *A. carolinensis* and only habitat features. The presence of an animal in a patch or a larger area, however, can be associated with or even influenced by factors such as competitors, predators, food availability and/or quality, disease, and parasites.

For *A. carolinensis*, potential competitors would most likely be other lizards and insectivorous birds, but very little information exists to evaluate any positive or negative associations between green anoles and other vertebrates. The occurrence of other lizards, regardless of the species, was recorded during the survey of each plot as either "present" (non-anoline lizard either inside the plot or within 2 m of plot edge) or "absent" (non-anoline lizard not observed inside the plot or within 2 m of plot edge). Because some plots were surveyed only long enough to find *A. carolinensis*, only plots which were sampled for at least 19 observer minutes were used in this analysis. A two-way contingency table was used to analyze this limited data for the possible association (whether positive or exclusionary) between *A. carolinensis* and all other lizard species combined. In only one case was a non-anoline individual actually seen within a given plot when a non-anoline species was considered "present" according to the categorization.

Results from the summer data do not support the possible association between the presence/absence of *A. carolinensis* and that of other lizards ($X^2 = 0.930$, df = 1, $P = 0.335$, Table 5-10). For the winter data, there were only two cases of another lizard species within or adjacent to any of the

plots (one plot with and one without *A. carolinensis*), thus providing similar findings to that for the summer data. In general along the Little Tennessee River, *A. carolinensis* seems to occur in greater numbers than any other lizard species (personal observations and data from the present study). Data on insectivorous bird numbers at the time of this study are lacking, but qualitative observations in the four study habitats and several others suggest that *A. carolinensis* is the most common diurnal vertebrate in these habitats. Thus, the limited information available shows no support for possible associations between *A. carolinensis* and all other lizards combined, but it cannot be said that such associations or competitive interactions which might influence *A. carolinensis* distribution are indeed absent.

As for predation, not a single predatory attempt on *A. carolinensis* has been witnessed by the author during eight years of field work in Tennessee. This is not to say that predation on *A. carolinensis* does not occur because tail autotomy occurs (personal observations), but that the lack of data on predatory events prevents any information about predation from being thoroughly evaluated. Snakes (e.g., *Coluber constrictor* (Colubridae)), have been observed attacking primarily ground-dwelling lizards such as *Eumeces* spp. (Scincidae) and *Cnemidophorus sexlineatus* (Teiidae) in one of the habitats (personal observations). *Anolis carolinensis* is arboreal to semi-arboreal and therefore might be difficult for many snakes to capture. In addition, the habitat in which potential snake predators have been most frequently observed, Habitat A, is also the habitat in which *A. carolinensis* is most abundant (personal observations).

The survey methods in this study were used to determine the presence of lizards and not potential predators. Potential snake and bird predators on *A. carolinensis* are rather widely-foraging, habitat generalists. Therefore, evaluating possible associations between the occurrence of anoles and such predators would require different survey methods than those used in this study.

Data on diseases and parasites is also lacking for *A. carolinensis* in Tennessee. Disease and/or parasite factors might be related indirectly to the presence of *A. carolinensis* in a given plot via their potential relationships with the abundance of lizards in a given habitat or locality. Lizard or saurian malaria (*Plasmodium*) is one possible factor, but its existence has not yet been studied in Tennessee populations. Decreases in several hematological, physiological, reproductive and behavioral parameters in *Sceloporus occidentalis* have been correlated with lizard malarial infection (see Schall 1983), but whether lizard malaria influences the distribution of *A. carolinensis* either within or among habitats is not known. Information is also lacking on other parasites and their ecological implications for *A. carolinensis.*

Limitations of this study

Any scientific study has limitations on the extent to which knowledge is gained and/or inferences can be made due to the nature of the study design, sampling scheme, and data structure. The present study is no exception. The summer and winter studies reported here were observational in nature and not experimental, thus causal relationships between the presence of *A. carolinensis* and habitat features cannot be

made. Useful descriptions of associations, but not clear causal relationships, can be obtained from observational studies analyzed with regression analyses (linear: Moses 1986:357, James and McCulloch 1990:137-138; logistic: James and McCulloch 1990:144-145). It would also be inappropriate, for example, to say that the variables in the best model(s) are the most "important" variables either statistically or biologically to explain the presence of this species. Observational studies with many regressor variables simply cannot assess the biological importance of variables, particularly when stepwise selection procedures are used (James and McCulloch 1990:136-138).

This study was exploratory rather than confirmatory due to both its observational nature and the sampling design. Although plots were randomly placed and randomly surveyed within habitats, the four habitats themselves were not randomly selected from among all habitats along the Little Tennessee River. Various constraints (e.g., travel time between sites, available time frame for the study, access to sites, and limited man-power) forced sites to be chosen which could be adequately sampled given such constraints. Thus, inferences here can probably be extended only to the specific types of habitat which were sampled and not to all habitats in eastern Tennessee or even along the Little Tennessee River. Such a limitation is common among field studies where the overall sampling may not have been entirely random in design.

Just because certain variables occur most frequently does not automatically infer that they are the most "biologically important" variables. Likewise, variables which did not occur frequently among the

best GA models are not necessarily biologically unimportant. For example, winter sunlight estimates (WSUN) did not appear in the final set of winter models. This does not mean that sunlight, per se, is not important because anoles bask in sunlight during sunny winter days. Other sunlight and/or canopy variables in the final models may have been able to capture some of the information about sunlight better than WSUN given the other variables that also appeared in the models.

Biological importance of habitat variables must ultimately be verified through experimental approaches, if possible, not via statistical modeling of observational data. Thus, the most frequent variables among the best GA models could be the ones that receive the attention and interest of future experimental research as mentioned previously.

The role of observational studies of animal-habitat relationships in forming conservation and management plans

The search for models, based on observational multivariate data, as a tool for testing "hypotheses", predicting ecological relationships or outcomes, and/or forming conservation policy and management plans is fairly popular in ecology, conservation biology, and wildlife management. Many such studies can be found in the the edited volume *Wildlife 2000* (Verner et al. 1986), the *Journal of Wildlife Management, Ecology,* and *Conservation Biology.* Unfortunately, the use of observational data for purposes of prediction is often risky and unreliable (see discussion of this topic in Part 4 as related to Hocking 1983, Snee 1983).

Researchers have also incorrectly used statistical modeling of observational data to "test hypotheses" or determine the effects or

influences of one or many things on another. In a study of fish assemblages in streams and beaver ponds, for example, multiple regression on a set of observational data was used to test "... the hypothesis that change in species richness per pond with pond age was a result of physical habitat changes." (Snodgrass and Meffe 1998:931). After forcing two other variables into models, the researchers used stepwise selection indicating that significant predictive power of pond age beyond that of other variables would be demonstrated if any of the age categories were still found in the model. Such "tests" and assessments of "predictive power" based on non-experimental data can only be rather weak tests of hypotheses if we are to believe most statisticians and the cautions of James and McCulloch (1985, 1990).

Morris (1987) used multivariate techniques and linear regression analysis to test hypotheses about which spatial scales small rodents were "selecting" in his study of habitat selection processes based on observational data. The non-experimental approach used by Morris (1987) cannot make inferences or draw conclusions about which scale of habitat is being selected by the animals. His observational study could suggest hypotheses to test regarding habitat scales, but only an experimental approach could confirm or refute hypotheses about which scales are being selected.

Studies using multivariate statistical methods to analyze observational data can be viewed as an exploratory stage (James and McCulloch 1985, 1990) in the research process, whereby associations or correlations are found. Exploratory data analysis on observational data helps lead to the

formulation of causal models (working models or research hypotheses) which then can be tested using experimental or quasi-experimental designs. The uncovering of causal relationships most likely culminates from research which progresses in stages beginning with simple observations and/or natural history information, proceeding to the building of descriptive models based on observational studies and exploratory data analysis, followed by formulation of initial causal models and the proper design of experiments to test causal relationships, and then ending in the confirmation of causation using confirmatory statistical methods (see James and McCulloch 1990:130-132).

Habitat models based on observational data can assist in the formulation of initial causal models or hypotheses regarding the relationships between an animal and habitat features. Hypotheses should then be investigated by means of field experiments or quasi-experiments, but not via purely observational studies.

Field experiments may be difficult to do, especially in obtaining enough replicates to test all possible combinations of conditions or habitat variables. Researchers could look at controlling one or two habitat factors at a time and having enough replicates for such a study. The analysis can then concentrate on determining how much variation in the presence/absence data of the animal is accounted for by controlling for effects of those two variables.

Researchers wanting to ultimately gain insight on the causal mechanisms behind any ecological phenomenon should follow the research procedure described by James and McCulloch (1990). An example

of how the full research procedure, from observational field studies to laboratory experiments, has been used to study habitat use/selection by a spider can be found in the work of Dr. Susan E. Riechert. First, field studies were conducted on *Agelenopsis aperta* to examine associations between reproductive success both food availability and the thermal environment (Riechert and Tracy 1975). Observational data were then collected on microhabitat use to examine potential differences between sites occupied by spiders and the general habitat (Riechert 1976, 1979). Then, lab experiments were conducted to test field observations and two specific hypotheses regarding habitat selection (Riechert 1985). An approach using biophysical and physiological ecology that examines the influences of environmental features on the population ecology of a lizard can be found in the work of Dunham et al. (1989).

Statistical inferences based on a single observational study are extremely weak and can be invalid. Researchers should not make strong statements about causal relationships based on findings from an observational study. Likewise, conservation and management plans, if possible, should not be based solely on the results of habitat modeling of observational data. The validity and reliability of forming conservation and management plans on just observational field data has never been fully demonstrated to the satisfaction of many scientists.

Statistical inferences are the strongest when based on sound experimental studies and a variety of data, rather than on observational studies. When possible, conservation and management decisions should be based on a diversity of information coming from observational studies,

experimental studies, mathematical modeling and simulations, biophysical and physiological ecology, and expert knowledge on the population, species, or community in question.

The reality is that ecologists, conservationists, and managers are caught in a dilema. Management and conservation decisions must be made regardless of the amount or quality of the information that is available. Observational data and models resulting from the exploratory analysis of such data are all that exist in many cases. In addition, conducting experimental studies can be costly and time-consuming with respect to many species and their habitats. What can be done?

When experimental studies are too difficult to conduct, mathematical and/or biophysical models do not yet exist for the given species, and decisions must be based on observational data alone, then at least four things should occur. First, all stake holders should be made explicitly aware of the fact that observational data do not provide a necessarily "true" or complete picture of ecological organisms and systems, nor does such data provide for accurate quantitative predictions. Second, predictions and forecasts should be carried out in qualitative, rather than specific quantitative, terms. For example, predictions of the influence of potential habitat alterations to change either the probability of occurrence of an organism or the density of organisms in a population should be made in qualitative categories such as "likely", "highly likely", "unlikely", "highly unlikely" to produce a change. Third, conservation and management decisions should be made with caution and an understanding of the limitations of observational data and exploratory analyses. Lastly,

monitoring programs should always be established for species and habitats in need of management and/or conservation decisions. In this way the imperfect knowledge could be modified, if necessary, in accordance with new data and insight, in keeping with an adaptive management approach.

Final comments

Many studies of *A. carolinensis* have been conducted in laboratory settings, but it seems that findings from lab studies have shed little knowledge directly on the ecology of this species. If we are to better understand the ecology of *A. carolinensis*, then lab-based studies must be pertinent to the ecology of this species. In addition, more experimental or quasi-experimental approaches need to be conducted in the field in order to understand causal mechanisms.

The GAIM approach used in this study for statistical modeling of observational multivariate data provides a way to take a wider view of an observational multivariate data set and models to fit the data than conventional methods which search for a single "best" model by using stepwise procedures. Then, the GAIM results can provide a set of very good models and a set of very good variables *to act as a starting point for future experimental studies*. This approach follows the research program and roles of observational studies in ecological research described by James and McCulloch (1985, 1990).

No doubt, much additional work is needed to determine the full practicality and overall utility of the GAIM approach for statistical modeling problems. It is hoped that researchers will rigorously investigate such matters with respect to multivariate modeling.

LITERATURE CITED

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267-281 *in* B. N. Petrov and F. Csáki, editors. Second international symposium on information theory. Akadémiai Kiadó, Budapest, Hungary. 451 pp.

Anderson, D. R., K. P. Burnham, and G. C. White. 1994. AIC model selection in overdispersed capture-recapture data. Ecology 75:1780-1793.

Bartlett, P. N. and D. M. Gates. 1967. The energy budget of a lizard on a tree. Ecology 48:315-322.

Beale, E. M. L. 1970. A note on procedures for variable selection in multiple regression. Technometrics 12:909-914.

Bozdogan, H. 1987. Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. Psychometrika 52:345-370.

Bozdogan, H. 1988a. ICOMP: a new model-selection criterion. Pages 599-608 *in* H. H. Bock, editor. Classification and related methods of data analysis: proceedings of the first conference of the international classification societies. North-Holland, Amsterdam, The Netherlands. 750 pp.

Bozdogan, H. 1988b. Selecting loglinear models and subset selection of variables in multiway contingency tables using Akaike's Information Criterion (AIC). Pages 609-616 *in* H. H. Bock, editor. Classification and related methods of data analysis: proceedings of the first conference of the international classification societies. North-Holland, Amsterdam, The Netherlands. 750 pp.

Bozdogan, H. 1990. On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. Communications in Statistics: Theory and Methods 19:221-278.

Bozdogan, H. (editor). 1994a. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. Vol. 1, theory and methodology of time series analysis. Kluwer Academic Publishers, Dordrecht, The Netherlands. 277 pp.

Bozdogan, H. (editor). 1994b. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. Vol. 2, multivariate statistical modeling. Kluwer Academic Publishers, Dordrecht, The Netherlands. 413 pp.

Bozdogan, H. (editor). 1994c. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. Vol. 3, engineering and scientific applications. Kluwer Academic Publishers, Dordrecht, The Netherlands. 346 pp.

Bozdogan, H. 1994d. Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. Pages 69-113 *in* H. Bozdogan, editor. Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach. Vol. 2, multivariate statistical modeling. Kluwer Academic Publishers, Dordrecht, The Netherlands. 413 pp.

Bozdogan, H. and D. M. A. Haughton. 1998. Informational complexity criteria for regression models. Computational Statistics & Data Analysis 28:51-76.

Brennan, L. A., W. M. Block, R. J. Gutiérrez. 1986. The use of multivariate statistics for developing habitat suitability index models. Pages 177-182 *in* J. Verner, M. L. Morrison, and C. J. Ralph, editors. Wildlife 2000: modeling habitat relationships of terrestrial vertebrates. The University of Wisconsin Press, Madison, Wisconsin, USA. 470 pp.

Buehler, D. A., T. J. Mersmann, J. D. Fraser, and J. K. D. Seegar. 1991. Effects of human activity on bald eagle distribution on the northern Chesapeake Bay. Journal of Wildlife Management 55:282-290.

Burger, L. D., L. W. Burger, and J. Faaborg. 1994. Effects of prairie fragmentation on predation on artificial nests. Journal of Wildlife Management 58:249-254.

Burnham, K. P. and D. R. Anderson. 1992. Data-based selection of an appropriate biological model: the key to modern data analysis. Pages 16-30 *in* D. R. McCullough and R. H. Barrett, editors. Wildlife 2001: Populations. Elsevier Science Publishers, London, United Kingdom. 1163 pp.

Burnham, K. P., D. R. Anderson, and G. C. White. 1995a. Selection among open population capture-recapture models when capture probabilities are heterogeneous. Journal of Applied Statistics 22:611-624.

Burnham, K. P., G. C. White, and D. R. Anderson. 1995b. Model selection strategy in the analysis of capture-recapture data. Biometrics 51:888-898.

Capen, D. E., J. W. Fenwick, D. B. Inkley, and A. C. Boynton. 1986. Multivariate models of songbird habitat in New England forests. Pages 171-175 *in* J. Verner, M. L. Morrison, and C. J. Ralph, editors. Wildlife 2000: modeling habitat relationships of terrestrial vertebrates. The University of Wisconsin Press, Madison, Wisconsin, USA. 470 pp.

Chandler, S. K., J. D. Fraser, D. A. Buehler, and J. K. D. Seegar. 1995. Perch trees and shoreline development as predictors of bald eagle distribution on Chesapeake Bay. Journal of Wildlife Management 59:325-332.

Chapman, J. 1985. Tellico archaeology: 12,000 years of Native American history. Tennessee Valley Authority, Knoxville, Tennessee, USA. 136 pp.

Chapman, J. and A. B. Shea. 1981. The archaeobotanical record: early Archaic period to contact in the lower Little Tennessee River valley. Tennessee Anthropologist 6:61-84.

Chapman, J., P. A. Delcourt, P. A. Cridlebaugh, A. B. Shea, and H. R. Delcourt. 1982. Man-land interaction: 10,000 years of American Indian impact on native ecosystems in the lower Little Tennessee River valley. Southeastern Archaeology 1:115-121.

Coker, D. R. and D. E. Capen. 1995. Landscape-level habitat use by brown-headed cowbirds in Vermont. Journal of Wildlife Management 59:631-637.

Davis, J. H. 1960. Proposals concerning the concept of habitat and a classification of types. Ecology 41:537-541.

Diefenbach, D. R. and R. B. Owen, Jr. 1989. A model of habitat use by breeding American black ducks. Journal of Wildlife Management 53: 383-389.

Delcourt, P. A., H. R. Delcourt, P. A. Cridlebaugh, and J. Chapman. 1986. Holocene ethnobotanical and paleoecological record of human impact on vegetation in the Little Tennessee River valley, Tennessee. Quarternary Research 25:330-349.

DeLong, A. K., J. A. Crawford, and D. C. DeLong, Jr. 1995. Relationships between vegetational structure and predation of artificial sage grouse nests. Journal of Wildlife Management 59:88-92.

Diller, L. V. and R. L. Wallace. 1994. Distribution and habitat of *Plethodon elongatus* on managed, young growth forests in north coastal California. Journal of Herpetology 28:310-318.

Drewien, R. C., W. M. Brown, and W. L. Kendall. 1995. Recruitment in Rocky Mountain greater sandhill cranes and comparison with other crane populations. Journal of Wildlife Management 59:339-356.

Dunham, A. E., B. W. Grant, and K. L. Overall. 1989. Interfaces between biophysical and physiological ecology and the population ecology of terrestrial vertebrate ectotherms. Physiological Zoology 62:335-355.

Endler, J. A. 1993. The color of light in forests and its implications. Ecological Monographs 63:1-27.

Engleman, L. 1988. Stepwise logistic regression. Pages 1013-1046 *in* W. J. Dixon, editor. BMDP Statistical Software. University of California Press, Berkeley, CA, USA. 1234 pp.

Forrest, S. 1993. Genetic algorithms: principles of natural selection applied to computation. Science 261:872-878.

Fretwell, S. D. 1972. Populations in seasonal environment. Princeton University Press, Princeton, New Jersey, USA. 217 pp.

Gates, D. M. 1980. Biophysical ecology. Springer-Verlag, New York, New York, USA. 611 pp.

Gatten, R.E., Jr., A. C. Echternacht, and M. A. Wilson. 1988. Acclimatization versus acclimation of activity metabolism in a lizard. Physiological Zoology 6:322-329.

Goldberg, D. E. 1989. Genetic algorithms in search, optimization, and machine learning. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, USA. 412 pp.

Goldberg, D. E. 1994. Genetic and evolutionary algorithms come of age. Communications of the ACM 37:113-119.

Gordon, R. E. 1956. The biology and biodemography of *Anolis carolinensis carolinensis* Voigt. Unpublished Ph.D. dissertation, Tulane University, New Orleans, Louisiana, USA. 263 pp.

Gorenzel, W. P. and T. P. Salmon. 1995. Characteristics of American crow urban roosts in California. Journal of Wildlife Management 59: 638-645.

Gorman, J. W. and R. J. Toman. 1966. Selection of variables for fitting equations to data. Technometrics 8:27-51.

Hinsley, S. A., P. E. Bellamy, I. Newton, and T. H. Sparks. 1996. Influences of population size and woodland area on bird species distributions in small woods. Oecologia 105:100-106.

Hocking, R. R. 1976. The analysis and selection of variables in linear regression. Biometrics 32:1-49.

Hocking, R. R. 1983. Developments in linear regression methodology: 1959-1982. Technometrics 25:219-230.

Holland, J. H. 1992a. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. 1st MIT Press edition. The MIT Press, Cambridge, Massachusetts, USA. 211 pp.

Holland, J. H. 1992b. Genetic algorithms. Scientific American July 1992: 66-72.

Hosmer, D. W., Jr. and S. Lemeshow. 1989. Applied logistic regression. John Wiley and Sons, New York, New York, USA. 307 pp.

Hutchinson, G. E. 1957. Concluding remarks. Cold Spring Harbor Symposium on Quantitative Biology 22:415-427.

James, F. C. and C. E. McCulloch. 1985. Data analysis and the design of experiments in ornithology. Pages 1-63 *in* R. F. Johnston, editor. Current ornithology, vol. 2, Plenum Press, New York, New York, USA. 378 pp.

James, F. C. and C. E. McCulloch. 1990. Multivariate analysis in ecology and ststematics: panacea or Pandora's box? Annual Review of Ecology and Systematics 21:129-166.

Jenssen, T. A. 1973. Shift in the structural habitat of *Anolis opalinus* due to congeneric competition. Ecology 54:863-869.

Jenssen, T. A., N. Greenberg, and K. A. Hovde. 1995. Behavioral profile of free-ranging male lizards, *Anolis carolinensis,* across breeding and post-breeding seasons. Herpetological Monographs 9:41-62.

Jenssen, T. A., J. D. Congdon, R. U. Fischer, R. Estes, D. Kling, S. Edmands, and H. Berna. 1996. Behavioural, thermal, and metabolic characteristics of a wintering lizard (*Anolis carolinensis*) from South Carolina. Functional Ecology 10:201-209.

Johnson, D. H. 1981a. The use and misuse of statistics in wildlife habitat studies. Pages 11-19 *in* D. E. Capen, editor. The use of multivariate statistics in studies of wildlife habitat. Rocky Mountain Forest and Range Experiment Station, U.S. Forest Service, General and Technical Report RM-87, U.S. Department of Agriculture, Fort Collins, Colorado, USA. 249 pp.

Johnson, D. H. 1981b. How to measure habitat - a statistical perspective. Pages 53-57 *in* D. E. Capen, editor. The use of multivariate statistics in studies of wildlife habitat. Rocky Mountain Forest and Range Experiment Station, U.S. Forest Service, General and Technical Report RM-87, U.S. Department of Agriculture, Fort Collins, Colorado, USA. 249 pp.

Johnson, R. G. and S. A. Temple. 1986. Assessing habitat quality for birds nesting in fragmented tallgrass prairies. Pages 245-249 *in* J. Verner, M. L. Morrison, and C. J. Ralph, editors. Wildlife 2000: modeling habitat relationships of terrestrial vertebrates. The University of Wisconsin Press, Madison, Wisconsin, USA. 470 pp.

Kindvall, O. 1996. Habitat heterogeneity and survival in a bush cricket metapopulation. Ecology 77:207-214.

Küchler, A. W. 1964. Potential natural vegetation of the conterminous United States. American Geographical Society Special Publication No. 36. American Geographical Society, New York, New York, USA. 116 pp. With separate map at 1:3,168,000.

Lebreton, J. D., K. P. Burnham, J. Clobert, and D. R. Anderson. 1992. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecological Monographs 62:67-118.

Licht, P. 1971. Regulation of the annual testis cycle by photoperiod and temperature in the lizard *Anolis carolinensis*. Ecology 52:240-252.

Licht, P. 1973. Influence of temperature and photoperiod on the annual ovarian cycle in the lizard *Anolis carolinensis*. Copeia 1973:465-472.

Lubchenco, J. and L. A. Real. 1991. Manipulative experiments as tests of ecological theory. Pages 715-7733 *in* L. A. Real and J. H. Brown, editors. Foundations of ecology: classic papers with commentaries. The University of Chicago Press, Chicago, Illinois, USA. 905 pp.

MacArthur, R. H. 1964. Environmental factors affecting bird species diversity. American Naturalist 98:387-398.

MacArthur, R. H. 1972. Geographical ecology. Princeton University Press, Princeton, New Jersey, USA. 269 pp.

Mantel, N. 1970. Why stepdown procedures in variable selection. Technometrics 12:591-612.

Marx, B. D. and E. P. Smith. 1990. Weighted multicollinearity in logistic regression: diagnostics and biased estimation techniques with an example from lake acidification. Canadian Journal of Fisheries and Aquatic Sciences 47:1128-1135.

MATLAB®. 1989. Pro-MATLAB for VAX/VMS computers. The Math Works, South Natick, Massachusetts, USA. 356 pp.

McCullagh, P. and A. J. Nelder. 1989. Generalized linear models. 2nd edition. Chapman and Hall, London, United Kingdom. 511 pp.

Morris, D. W. 1987. Ecological scale and habitat use. Ecology 68:362-369.

Morrison, M. L., B. G. Marcot, and R. W. Mannan. 1992. Wildlife-habitat relationships: concepts and applications. The University of Wisconsin Press, Madison, Wisconsin, USA. 343 pp.

Moses, L. E. 1986. Think and explain with statistics. Addison-Wesley Publishing Company, Reading, Massachusetts, USA. 483 pp.

Munger, J. C., M. Gerber, K. Madrid, M. Carroll, W. Petersen, and L. Heberger. 1998. U.S. National Wetland Inventory classifications as predictors of the occurrence of Columbia spotted frogs (*Rana luteiventris*) and Pacific treefrogs (*Hyla regilla*). Conservation Biology 320-330.

Nadeau, S., R. Décarie, D. Lambert, and M. St-Georges. 1995. Nonlinear modeling of muskrat use of habitat. Journal of Wildlife Management 59:110-117.

Neter, J., W. Wasserman, and M. H. Kutner. 1985. Applied linear statistical models: regression, analysis of variance, and experimental designs. 2nd ed. Richard D. Irwin, Homewood, Illinois, USA. 1127 pp.

Noeske, T. A. and A. H. Meier. 1977. Photoperiodic and thermoperiodic interaction affecting fat stores and reproductive indexes in the male green anole, *Anolis carolinensis*. Journal of Experimental Zoology 202:97-102.

Porter, W. P. 1989. New animal models and experiments for calculating growth potential at different elevations. Physiological Zoology 62:286-313.

Porter, W. P. and D. M. Gates. 1969. Thermodynamic equilibria of animals with environment. Ecological Monographs 39:227-244.

Porter, W. P. and C. R. Tracy. 1983. Biophysical analyses of energetics, time-space utilization, and distributional limits. Pages 55-83 *in* R. B. Huey, E. R. Pianka, and T. W. Schoener, editors. Lizard ecology: studies of a model organism. Harvard University Press, Cambridge, Massachusetts, USA. 501 pp.

Pregibon, D. 1981. Logistic regression diagnostics. The Annals of Statistics 9:705-724.

Ragland, I. M., L. C. Wit, and J. C. Sellers. 1981. Temperature acclimation in the lizards *Cnemidophorus sexlineatus* and *Anolis carolinensis*. Comparative Biochemistry and Physiology 70A:33-36.

Rand, A. S. 1964. Ecological distribution in anoline lizards of Puerto Rico. Ecology 45:745-752.

Riechert, S. E. 1976. Web site selection in a desert spider, *Agelenopsis aperta* (Gertsch.). Oikos 27:311-315.

Riechert, S. E. 1979. Games spiders play II: resource assessment strategies. Behavioral Ecology and Sociobiology 6:121-128.

Riechert, S. E. 1985. Decisions in multiple goal contexts: habitat selection of the spider, *Agelenopsis aperta* (Gertsch). Zool. Tierpsychol. 70:53-69.

Riechert, S. E. and C. R. Tracy. 1975. Thermal balance and prey availability: bases for a model relating web-site characteristics to spider reproductive success. Ecology 56:265-285.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. Akaike information criterion statistics. KTK Scientific Publishers, Tokyo, Japan. 290 pp.

SAS Institute Inc. 1989. SAS/STAT® user's guide. Version 6, Fourth Edition, Volume 2, Cary, North Carolina, USA. 846 pp.

Sexton, O. W. and H. Heatwole. 1968. An experimental investigation of habitat selection and water loss in some anoline lizards. Ecology 49:762-767.

Schall, J. J. 1983. Lizard malaria: parasite-host ecology. Pages 84-100 *in* R. B. Huey, E. R. Pianka, and T. W. Schoener, editors. Lizard ecology: studies of a model organism. Harvard University Press, Cambridge, Massachusetts, USA. 501 pp.

Schoener, T. W. 1968. The *Anolis* lizards of Bimini: resource partitioning in a complex fauna. Ecology 49:704-726.

Schoener, T. W. 1975. Presence and absence of habitat shift in some widespread lizard species. Ecological Monographs 45:233-258.

Shaffer, M. L. 1981. Minimum population sizes for species conservation. BioScience 31:131-134.

Smith, K. G. and P. G. Connors. 1986. Building predictive models of species occurrence from total-count transect data and habitat measurements. Pages 45-50 *in* J. Verner, M. L. Morrison, and C. J. Ralph, editors. Wildlife 2000: modeling habitat relationships of terrestrial vertebrates. The University of Wisconsin Press, Madison, Wisconsin, USA. 470 pp.

Snee, R. D. 1983. Discussion. Technometrics 25:230-237.

Snodgrass, J. W. and G. K. Meffe. 1998. Influence of beavers on stream fish assemblages: effects of pond age and watershed position. Ecology 79:928-942.

Stearns, S. C. 1976. Life history tactics: a review of the ideas. Quarterly Review of Biology 51:3-47.

Szymczak, M. R. and E. A. Rexstad. 1991. Harvest distribution and survival of a gadwall population. Journal of Wildlife Management 55: 592-600.

Thiollay, J. M. 1989. Area requirements for the conservation of rainforest raptors and game birds in French Guiana. Conservation Biology 3:128-137.

Toft, C. A. 1990. Reply to Seaman and Jaeger: an appeal to common sense. Herpetologica 46:357-361.

Trexler, J. C. and J. Travis. 1993. Nontraditional regression analyses. Ecology 74:1629-1637.

van Manen, F. T. and M. R. Pelton. 1993. Data-based modelling of black bear habitat using GIS. Pages 323-329 *in* I. D. Thompson, editor. Proceedings of the International Union of Game Biologists XXI Congress: forests and wildlife .... towards the 21st century, Volume 1. Canadian Forest Service, Chalk River, Ontario, Canada. 379 pp.

APPENDIX TO PART 5

Table 5-1. The original names, forms, and descriptions of the measurement of the original variables for the study of the relationship between the presence of *Anolis carolinensis* and habitat features in four habitats along the Little Tennessee River in Tennessee. Cat indicates a variable is categorical; Con indicates a variable is continuous in scale.

Original Variable	Original Form	Description of Original Measurement
HAB	Cat	The different habitats; Habitats A (HABA), B (HABB), C (HABC; reference cell), or D (HABD)
DPOWR	Cat	Distance from the plot center to the nearest potential overwintering rock (rock with crevices or holes into which lizards could crawl); four levels: $\leq$ 10 (DPOWR1), 11-20 (DPOWR2), 21-30 (DPOWR3), and $\geq$ 31 m (reference cell)
ROCK	Cat	Presence or absence (reference cell) of potential overwintering rock within a plot
NLOV	Con	Number of live overstory tree trunks (those $\geq$ 75 mm in diameter at breast height (DBH)) in a plot
NLU	Con	Number of live understory tree trunks (those $\leq$ 74 mm DBH) in a plot
EVG	Cat	Presence or absence (reference cell) of any live overstory evergreen tree trunks in a plot
SMEV	Con	Sum of DBHs (mm) of all live evergreen tree trunks in a plot
LDBH	Con	DBH (mm) of the largest live tree trunk in a plot
SMBG	Con	Sum of the DBHs (mm) of all the live overstory tree trunks in a plot

Table 5-1. (continued).

Original Variable	Original Form	Description of Original Measurement
DFW	Cat	Presence or absence (reference cell) of any dead fallen woody trunks, logs, branches, or limbs; all items had to have at least 0.5 m length within a plot and any branches or limbs also had to project at least 0.5 m above ground level; logs or trunks had to be $\geq$ 3 cm in diameter; these criteria eliminated any small material which was unlikely to be used by lizards for perch sites
HSFC	Con	Herb, shrub, and vine foliage cover during summer measured by looking down on four ground transects (each running N, S, E, and W from plot center to the edge of a plot) from a height of 1.5 m above the ground and counting the presence or absence of such cover at 20 cm intervals along each transect; expressed as the % of 96 total transect counts with cover present
SCAN	Cat	Categories of summer canopy cover based on inspecting the relative amounts of sunlight vs. shade on both the ground and vertical vegetation in a plot and by viewing the canopy within about a 120° area facing south from plot center during mid- to late summer; three levels: open canopy, gap, or along habitat edge (plot dominated largely by sunlight; SCAN1), partially open canopy (plot with relatively equal mixture of sun and shade; SCAN2), or closed canopy (plot dominated by shade; reference cell)
WCAN	Cat	Winter canopy visually categorized at the end of winter survey using the same methods and categorizations as for SCAN

Table 5-1. (continued).

Original Variable	Original Form	Description of Original Measurement
SSUN	Con	Summer sunlight as estimated from scoring patches of sun or shade flecks on a Sun/Shade Board (SSB); the SSB is a board approximately 2.5 m long, painted white, and divided into 40 rectangles (starting 30 cm from the proximal end of the board and continuing its entire length) with each rectangle approximating the area of three to four bodies of adult green anoles lying side-by-side (55 mm long and 35 mm wide) and representing a patch of sunlight suitable for basking by an adult; each rectangle was scored as sun (1), approximately equal sun and shade (0.5), or shade (0), by holding the SSB parallel to the ground at 1.5 m above the plot center and orienting the SSB along each of five different randomly selected compass directions selected from eight primary directions (N, NE, E, SE, S, SW, W, and NW); to be scored as either sun or shade a rectangle had to have more than half of its area in either full sun or full shade; any rectangle having filtered sunlight/partial shading over half or more of its area or having equal areas of full sun and full shade was scored as 0.5; a plot's overall estimate of sunlight was the sum of the 200 rectangle scores and expressed as a %
WSUN	Con	Winter sunlight estimated in same manner as SSUN within a plot, but during winter season
DE	Con	Distance from plot center to the habitat edge as measured to the nearest whole m
STEMP	Con	Summer air temperature (ambient) taken 1 m above ground within a plot after surveying a plot

Table 5-1. (continued).

Original Variable	Original Form	Description of Original Measurement
DAE	Con	Distance (to nearest whole m) along the habitat edge from the western end of a site to the plot center; because the four habitats differed greatly in their sizes this distance was standardized by dividing the distance value of a given plot by the total distance along the habitat edge for that particular habitat so that all values were between zero and one

Table 5-2. The names, final forms, and descriptions of the final form of the habitat variables used in the summer data analysis for the study of the relationship between the presence of *Anolis carolinensis* and habitat features in four habitats along the Little Tennessee River in Tennessee. Cat indicates a variable is categorical; Con indicates a variable is continuous in scale. The numbering scheme given here for the variables will be used for the summer variables in other tables.

Summer Variable	Final Form	Description of Final Summer Form
1. HABS	Cat	Refinement of HAB because HABD had similar odds ratio as reference cell HABC; three levels: Habitats A (HABA), B (HABB), and C and D combined (reference cell)
2. DPOR	Cat	Refinement of DPOR since original DPOWR3 had a zero cell; three levels: $\leq$ 10 (DPOR1), 11-20 (DPOR2), or $\geq$ 21 m (reference cell)
3. LDS	Cat	Categorization of LDBH due to non-linearity in logit and lack of trees in some plots; four levels: live trunks absent or largest diameter at breast height (DBH) $\leq$ 74 (LDS1), 75-149 (LDS2), 150-235 (LDS3), or $\geq$ 236 mm DBH (reference cell)
4. SMOS	Cat	Categorization of SMBG due to non-linearity in logit and lack of trees in some plots; three levels: live trees absent or only live trees $\leq$ 74 mm present (SMOS1), sum 75-189 mm (SMOS2), or sum $\geq$ 190 mm DBH (reference cell)
5. SCAN	Cat	Same as original
6. DES	Cat	Categorization of DE due to non-linearity in logit; three levels: $\leq$ 7 (DES1), 8-14 (DES2), or $\geq$ 15 m (reference cell)
7. LOSD	Cat	Categorization of NLOV due to non-linearity in logit; two levels: $\leq$ 1 or $\geq$ 2 (reference cell)

Table 5-2. (continued).

Summer Variable	Final Form	Description of Final Summer Form
8. NLU	Con	Same as original
9. EVG	Cat	Same as original
10. ESSD	Cat	Categorization of SMEV due to non-linearity in logit and lack of trees in some plots; two levels: live evergreens absent or sum $\leq$ 99, or otherwise sum $\geq$ 100 mm DBH (reference cell)
11. WCD	Cat	Refinement of WCAN since closed canopy level did not occur for any plot during the winter; two levels: open, gap, or edge vs. partially open canopy (reference cell)
12. DFW	Cat	Same as original
13. ROCK	Cat	Same as original
14. HSSD	Cat	Categorization of HSFC; two levels: 19.0-41.9% or otherwise (reference cell)
15. SSSD	Cat	Categorization of SSUN due to non-linearity in logit; two levels: 34.5-74.0% or otherwise (reference cell)
16. WSUN	Con	Same as original
17. STMD	Cat	Categorization of STEMP due to non-linearity in logit; two levels: 25.6-28.7°C or otherwise (reference cell)
18. DAES	Cat	Categorization of DAE; two levels: 0.269-0.599 or otherwise (reference cell)

Table 5-3. The names, final forms, and descriptions of the variables used in the winter data analysis for the study of the relationship between the presence of *Anolis carolinensis* and habitat features in four habitats along the Little Tennessee River in Tennessee. Cat indicates a variable is categorical; Con indicates a variable is continuous in scale. The numbering scheme given here for the variables will be used for the winter variables in other tables.

Winter Variable	Final Form	Description of Final Winter Form
1. HAB	Cat	Same as original
2. NLUW	Cat	Categorization of NLU due to non-linearity in logit; three levels: ≤ 1 (NLUW1), 2 (reference cell), or ≥ 3 (NLUW2)
3. WTM	Cat	Categorization of WTEMP due to non-linearity in logit; three levels: 10.5-14.9°C (WTM1), 15.0-20.0°C (WTM2), or ≤ 10.4 or ≥ 20.1°C (reference cell)
4. DPOR	Cat	Same as categorization for summer analysis
5. NLOW	Cat	Categorization of NLOV due to non-linearity in logit; two levels: ≤ 2 or ≥ 3 (reference cell)
6. LDW	Cat	Categorization of LDBH due to non-linearity in logit and some plots did not have trees; two levels: DBH of largest live trunk $\leq$ 185 mm, or no live trunks present or largest live DBH $\geq$ 186 mm (reference cell)
7. EVG	Cat	Same as original
8. SOTW	Cat	Categorization of SMBG due to non-linearity in logit and some plots did not have trees; two levels: only live understory trunks present or the sum of the DBHs of all live overstory trunks $\leq$ 189 mm, or no live trunks present or sum of the DBHs of all live overstory trunks $\geq$ 190 mm DBH (reference cell)

Table 5-3. (continued).

Winter Variable	Final Form	Description of Final Winter Form
9. ESWD	Cat	Categorization of SMEV due to non-linearity in logit; two levels: live trees absent or sum of live evergreen trunks $\geq$ 100 mm DBH, or only live deciduous trunks present or sum of live evergreen trunks $\leq$ 99 mm DBH (reference cell)
10. SCW	Cat	Refinement of SCAN; two levels: open/gap/edge, or otherwise (reference cell)
11. WCD	Cat	Same as categorization for summer analysis
12. DFW	Cat	Same as original
13. ROCK	Cat	Same as original
14. HSCW	Cat	Categorization of HSFC due to non-linearity in logit; two levels: $\leq$ 22.9 or $\geq$ 34.0 %, or otherwise (reference cell)
15. SSWD	Cat	Categorization of SSUN due to non-linearity in logit; two levels: SSUN $\leq$ 27.9 or $\geq$ 49.6, or otherwise (reference cell)
16. WSUN	Con	Same as original
17. DEW	Cat	Categorization of DE due to non-linearity in logit; two levels: $\leq$ 7 or $\geq$ 17 m, or otherwise (reference cell)
18. DAEW	Cat	Categorization of DAE; two levels: 0.300-0.599, or otherwise (reference cell)

Table 5-4. Univariate logistic regression summary information for the summer habitat variables. Variable No. and Name refer to the number and abbreviated name for each habitat variable, respectively, as given in Table 5-2 and Outcome is the outcome of the survey of each plot (P= presence and A= absence of *Anolis carolinensis*). ICOMP-IFIM is the model selection criterion value for the univariate logistic regression model which includes the habitat variable and the intercept term. a. For each categorical variable the following summary information is reported: the categories (Level) of a variable (the reference cell is given last and p= presence and a= absence of a particular habitat feature; see Table 5-2 for descriptions), the numbers of observed plots having *Anolis carolinensis* present or absent in each level (Outcome), and ICOMP-IFIM for the univariate model. b. For each continuous variable (see Table 5-2 for descriptions) the following summary information is reported: ICOMP-IFIM for the univariate model and, for each possible outcome of the plot survey, the 25% quantile (Q), median, 75% quantile, mean, and standard deviation of the mean (Stand. Dev.).

a.

Independent Variable	Level	Outcome P	Outcome A	ICOMP-IFIM
1. HABS	A	36	15	126.60
	B	6	37	
	C/D	3	69	
2. DPOR	1	39	39	151.64
	2	4	14	
	3	2	68	
3. LDS	1	13	34	174.24
	2	22	18	
	3	8	32	
	4	2	37	
4. SMOS	1	13	34	181.37
	2	19	18	
	3	13	69	
5. SCAN	1	26	29	180.30
	2	17	62	
	3	2	30	

Table 5-4. (continued).

a.

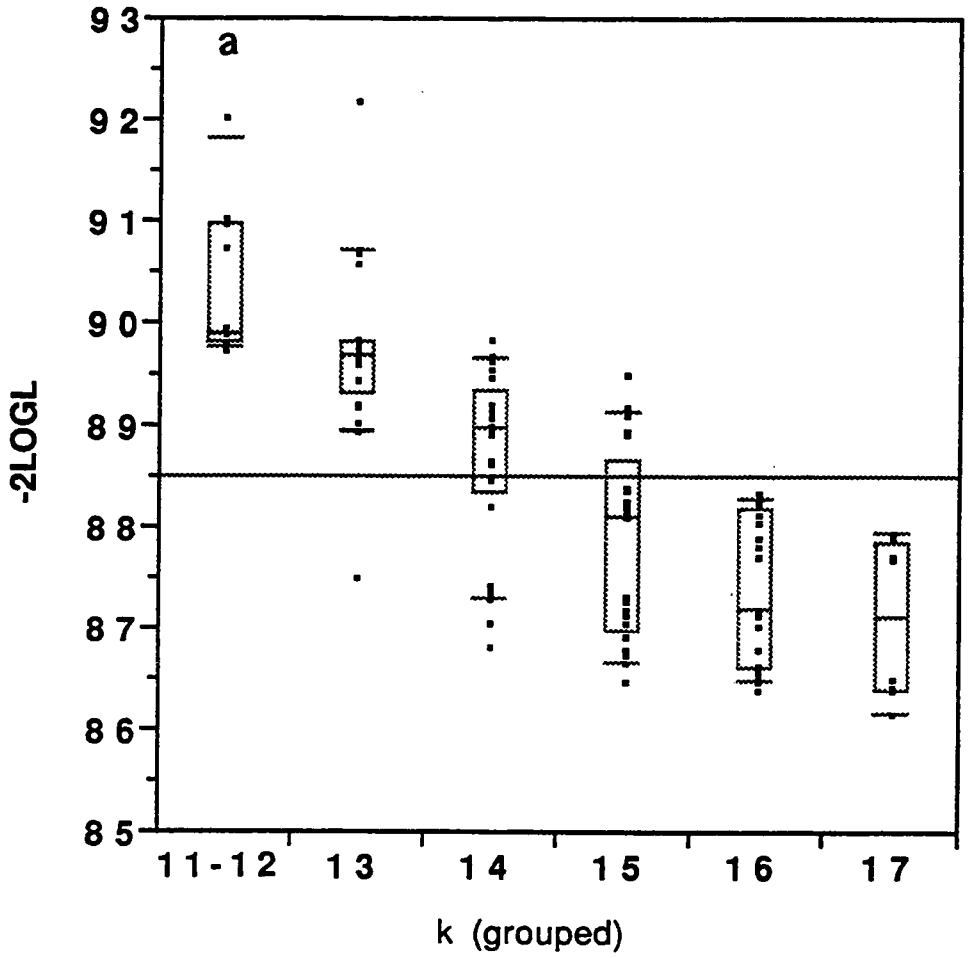
Independent Variable	Level	Outcome P	Outcome A	ICOMP-IFIM
6. DES	1	12	31	176.32
	2	23	22	
	3	10	68	
7. LOSD	1	31	70	194.60
	0	14	51	
9. EVG	p	2	33	184.81
	a	43	88	
10. ESSD	1	42	88	188.95
	0	3	33	
11. WCD	1	44	77	174.87
	0	1	44	
12. DFW	p	27	100	187.72
	a	18	21	
13. ROCK	p	18	20	186.23
	a	27	101	
14. HSSD	1	33	68	192.29
	0	12	53	
15. SSSD	1	20	34	191.83
	0	25	87	
17. STMD	1	27	56	193.49
	0	18	65	
18. DAES	1	20	29	189.38
	0	25	92	

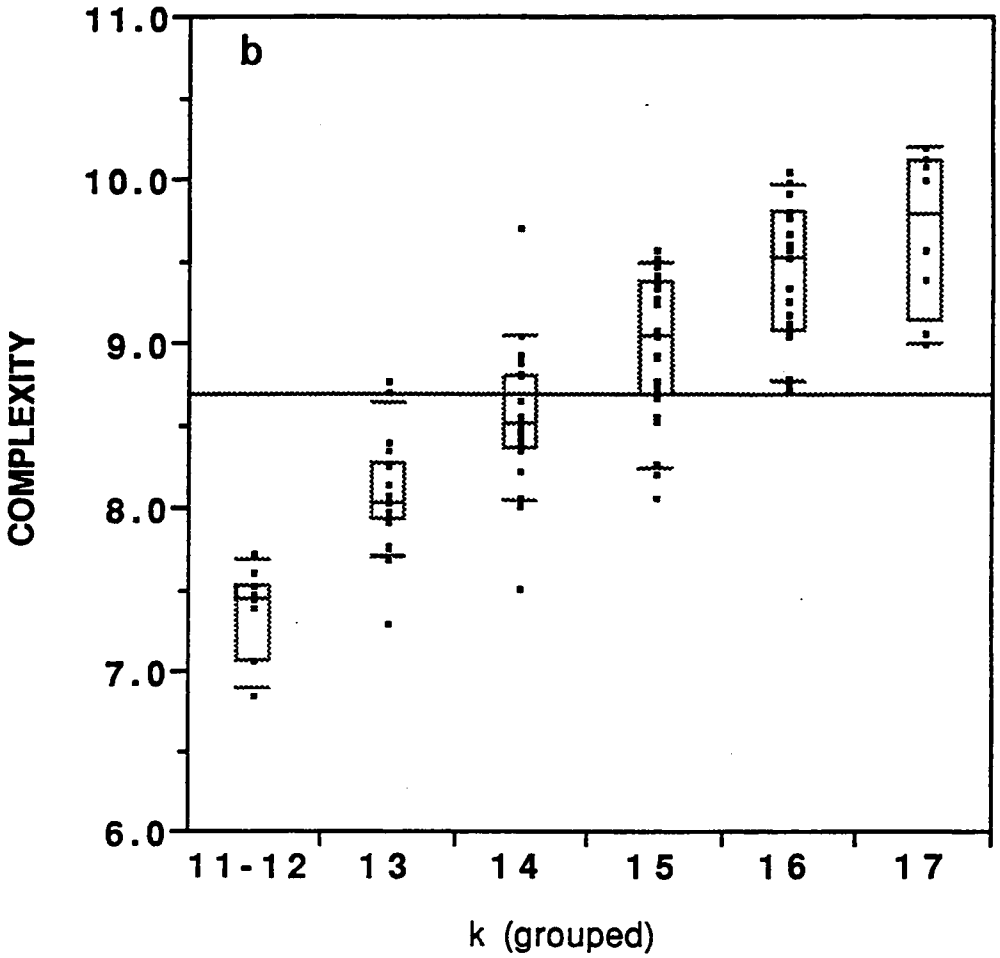
Table 5-4. (continued).

b.

Independent Variable	Outcome Variable	25% Q	Median	75% Q	Mean	Stand. Dev.	ICOMP-IFIM
8. NLU	P	0	1	3	1.87	2.46	178.06
	A	0	0	1	0.64	1.02	
16. WSUN	P	80.0	89.5	96.0	85.93	11.96	192.58
	A	62.0	80.5	89.0	75.07	17.50	

Fig. 5-1. Summer GA models: box plots showing trends in the (a) lack-of-fit term (-2LogL), (b) complexity term, and (c) model selection criterion, ICOMP-IFIM, across the different model sizes, represented by k (number of estimated regression parameters), for the best 115 summer logistic regression models found by the genetic algorithm (GA) analysis. Some models with different k values were grouped together so that no level of k had less than 5% of the 115 total models. The line across the graph parallel to the X-axis shows the mean value of the given term for the 115 models. The line within each box represents the median for the given level of k. The 25% and 75% quantiles are represented by the ends of a box, while the 10% and 90% quantiles are shown as the short lines outside the ends of a box.


a
-2LOGL
93
92
91
90
89
88
87
86
85
11-12
13
14
15
16
17
k (grouped)


b
11.0
10.0
9.0
8.0
7.0
6.0
COMPLEXITY
11-12
13
14
15
16
17
k (grouped)

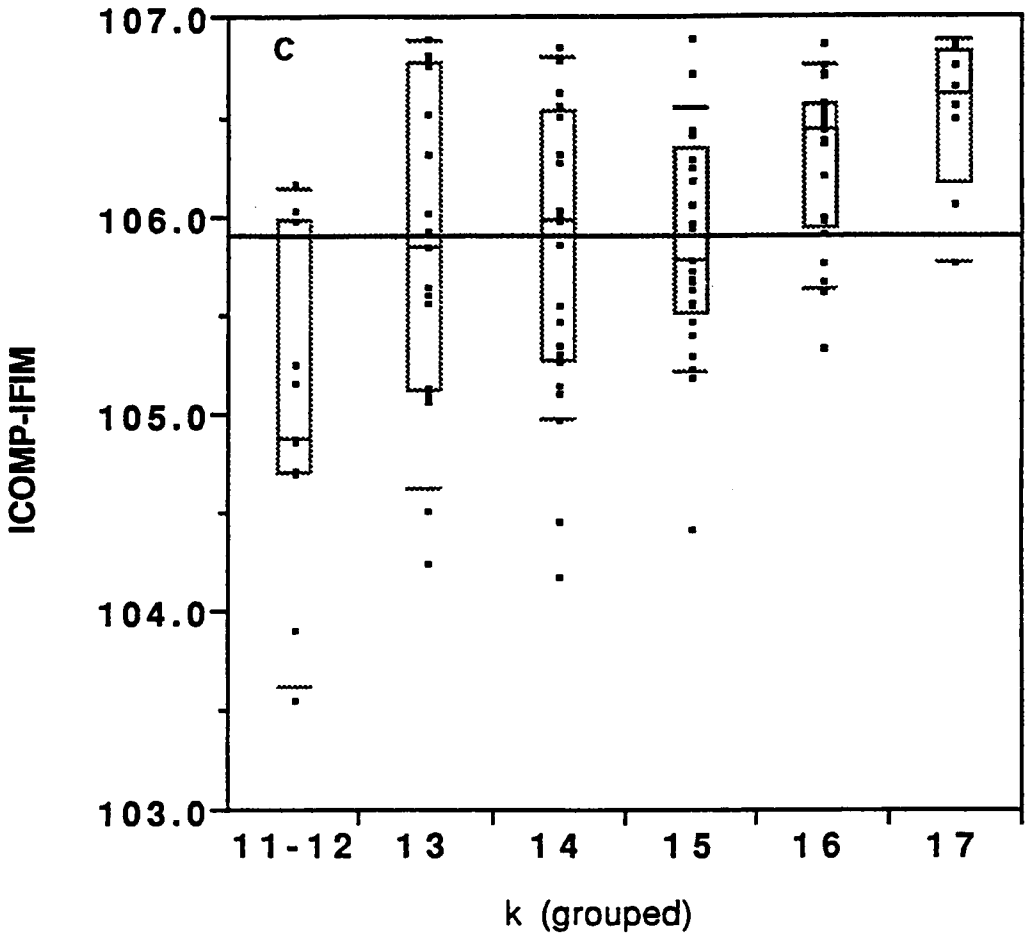

C
107.0
106.0
105.0
104.0
103.0
ICOMP-IFIM
11-12
13
14
15
16
17
k (grouped)

Fig. 5-2. Summer GA models: the frequency of independent variables in the best 115 logistic regression models from the genetic algorithm (GA) output modeling the relationship between habitat features and the presence of *Anolis carolinensis* in summer plots. Percent represents the percentage of best summer GA models in which a given variable occurred. Variable acronyms are defined in Table 5-2.

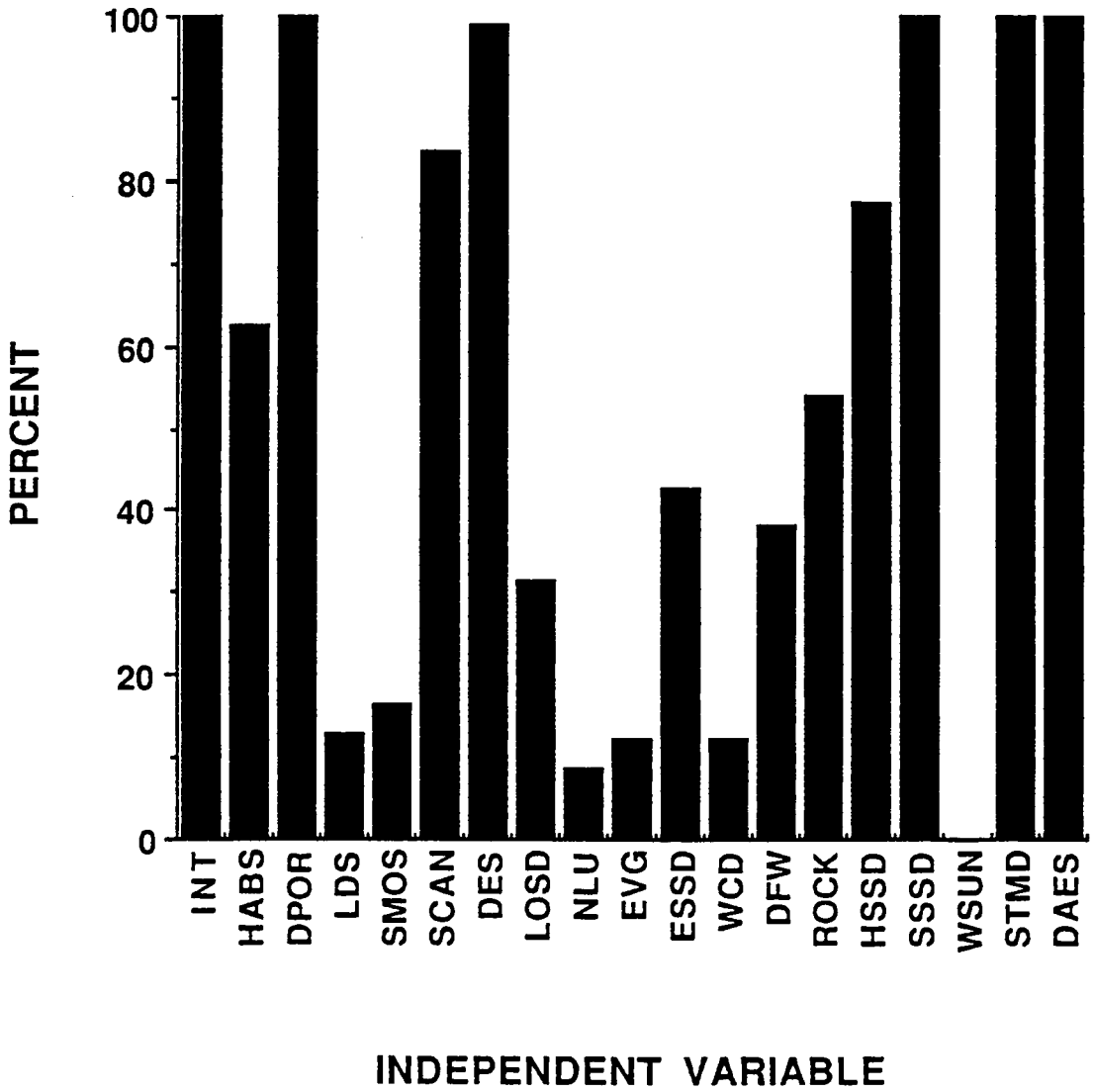

PERCENT
100
80
60
40
20
0
INT
HABS
DPOR
LDS
SMOS
SCAN
DES
LOSD
NLU
EVG
ESSD
WCD
DFW
ROCK
HSSD
SSSD
WSUN
STMD
DAES
INDEPENDENT VARIABLE

Fig. 5-3. Summer GA models: trends in the frequency of independent variables in the different model sizes (k levels) in the best 115 logistic regression models from the genetic algorithm (GA) output modeling the relationship between habitat features and the presence of *Anolis carolinensis* in summer plots. (a) k = 16 to 17 estimated parameters (n = 29). (b) k = 15 estimated parameters (n = 25). (c) k = 14 estimated parameters (n = 29). (d) k = 11 to 13 estimated parameters (n = 32). Percent represents the percentage of best summer GA models of a specific k size in which a given variable occurred. Variable acronyms are defined in Table 5-2.

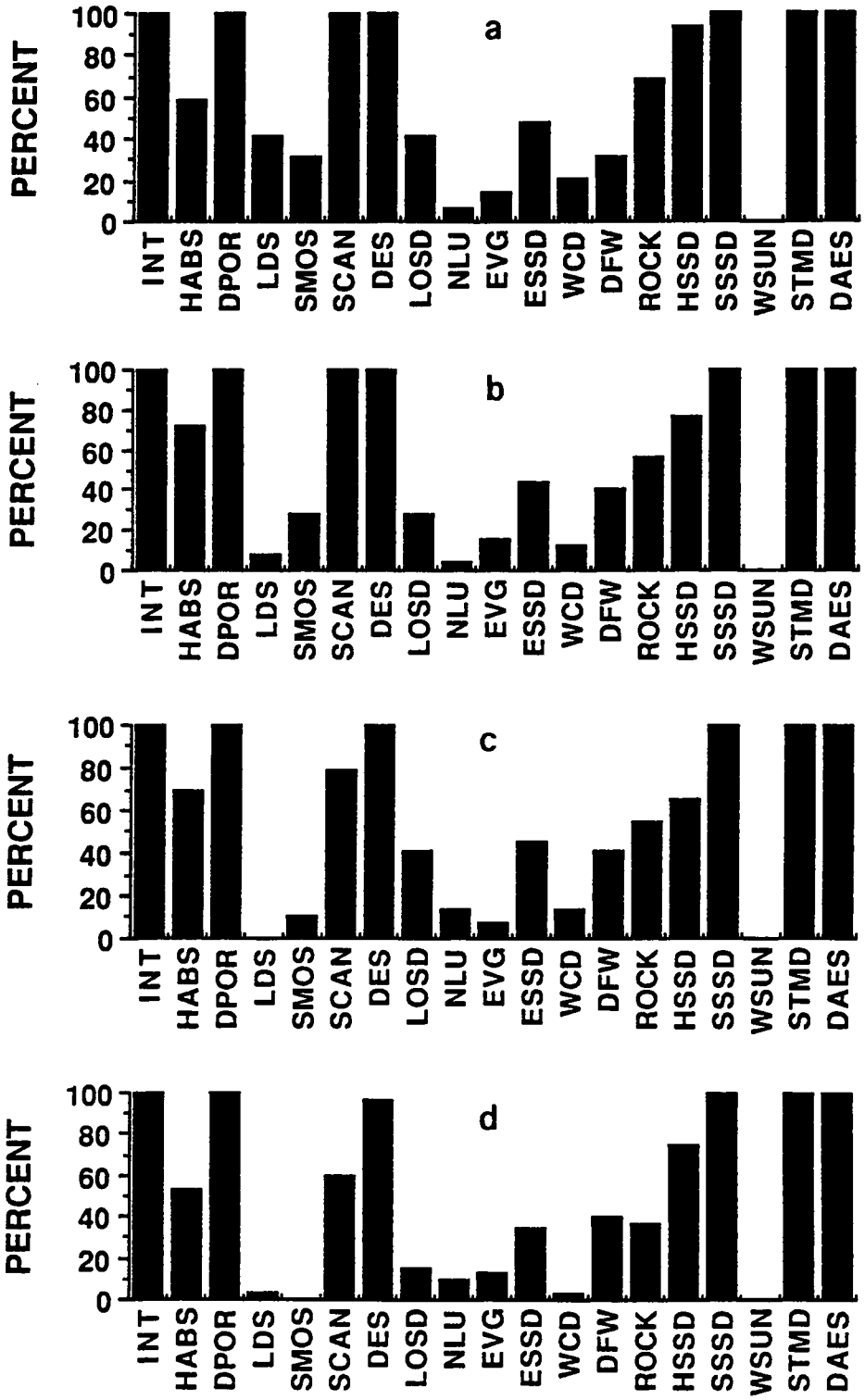

PERCENT
100
80
60
40
20
0
a
b
c
d
INT
HABS
DPOR
LDS
SMOS
SCAN
DES
LOSD
NLU
EVG
ESSD
WCD
DFW
ROCK
HSSD
SSSD
WSUN
STMD
DAES

Table 5-5. Summer models: the best logistic regression models with 15 or fewer parameters from the genetic algorithm (GA) output for modeling the relationship between habitat variables and the presence of *Anolis carolinensis* in summer study plots. The heading Model No. gives the rank order of each model relative to all other GA models based on the ICOMP-IFIM (Informational Complexity of the Inverse Fisher Information Matrix) values, where the lowest values represent the better models. *k*= the number of estimated regression parameters in the model, -2LogL= (-2) times the loglikelihood value (or lack-of-fit term) for the model, Comp= the complexity (penalty) term calculated as the of complexity of the inverse Fisher information matrix, Var= the estimated model variance, and the numbers[a] under the heading of Independent Variables identify the independent variables which are present in a model (a dot indicates the given variable is absent from the model). Variable numbers in bold type are those whose parameter estimates were statistically significant at $P < 0.10$ (categorical variables with two or more parameter estimates are shown in bold type only if half or more of the estimates for a given variable meet this significance level). For simplicity, only the top ten GA models, at the most, are shown for each *k* level.

Model No.	ICOMP -IFIM	*k*	-2LogL	Comp	Var	Independent Variables
1	103.56	11	89.83	6.87	0.72	**0** . **2** . . **5** **6** . . . . **14** **15** . **17** **18**
20	105.17	11	91.03	7.07	0.63	**0** **1** **2** . . . **6** . . . . 14 **15** . **17** **18**
67	106.18	11	92.04	7.07	0.68	**0** **1** **2** . . . **6** . . . . . 12 . . **15** . **17** **18**
2	103.92	12	89.76	7.08	0.70	**0** . **2** . . **5** **6** . . . . . . 13 **14** **15** . **17** **18**
8	104.70	12	89.92	7.39	0.86	**0** 1 **2** . . 5 **6** . . . . **15** . **17** **18**
9	104.73	12	89.80	7.46	0.72	**0** . **2** . . **5** **6** . . . . . 12 . **14** **15** . **17** **18**
10	104.88	12	89.92	7.48	0.64	**0** **1** **2** . . . **6** . . . 10 . . 13 . **15** . **17** **18**
11	104.88	12	89.97	7.46	0.64	**0** **1** **2** . . . **6** . . . . . 12 13 . **15** . **17** **18**
23	105.27	12	89.83	7.72	0.72	**0** . **2** . . **5** **6** . . 9 . . . . **14** **15** . **17** **18**

Table 5-5. (continued).

Model No.	ICOMP -IFIM	*k*	-2LogL	Comp	Var	Independent Variables																
59	105.99	12	90.76	7.62	0.63	**0**	**1**	**2**	.	.	.	**6**	.	.	.	10	.	.	.	14	**15**	.	**17**	**18**
63	106.04	12	90.99	7.53	0.63	**0**	**1**	**2**	.	.	.	**6**	.	.	.	.	.	12	.	14	**15**	.	**17**	**18**
4	104.25	13	87.52	8.36	0.73	**0**	1	**2**	.	.	**5**	**6**	.	.	.	.	.	.	.	14	**15**	.	**17**	**18**
7	104.52	13	88.96	7.78	0.64	**0**	.	**2**	.	.	**5**	**6**	7	.	.	10	.	.	.	**14**	**15**	.	**17**	**18**
14	105.07	13	89.19	7.94	0.60	**0**	**1**	**2**	.	.	.	**6**	.	.	.	10	.	.	13	14	**15**	.	**17**	**18**
15	105.10	13	89.03	8.03	0.64	**0**	.	**2**	.	.	**5**	**6**	7	.	9	.	.	.	.	**14**	**15**	.	**17**	**18**
16	105.11	13	89.72	7.70	0.70	**0**	.	**2**	.	.	**5**	**6**	.	.	.	.	.	12	13	**14**	**15**	.	**17**	**18**
18	105.15	13	89.61	7.77	0.70	**0**	.	**2**	.	.	**5**	**6**	.	.	.	10	.	.	13	**14**	**15**	.	**17**	**18**
37	105.58	13	89.68	7.95	0.85	**0**	1	**2**	.	.	5	**6**	.	.	.	10	.	.	.	.	**15**	.	**17**	**18**
38	105.62	13	89.71	7.95	0.86	**0**	**1**	**2**	.	.	5	**6**	.	.	.	.	.	12	.	.	**15**	.	**17**	**18**
39	105.62	13	89.75	7.93	0.63	**0**	**1**	**2**	.	.	.	**6**	7	.	.	10	.	.	13	.	**15**	.	**17**	**18**
42	105.66	13	89.76	7.95	0.70	**0**	.	**2**	.	.	**5**	**6**	.	.	9	.	.	.	13	**14**	**15**	.	**17**	**18**
3	104.18	14	89.16	7.51	0.69	**0**	.	**2**	.	4	**5**	**6**	.	.	.	.	.	.	13	**14**	**15**	.	**17**	**18**
6	104.46	14	86.82	8.82	0.67	**0**	1	**2**	.	.	5	**6**	.	.	.	.	.	.	13	14	**15**	.	**17**	**18**
12	104.98	14	87.06	8.96	0.67	**0**	1	**2**	.	.	**5**	**6**	7	.	.	.	.	.	.	14	**15**	.	**17**	**18**
13	104.98	14	87.30	8.84	0.72	**0**	1	**2**	.	.	**5**	**6**	.	.	.	10	.	.	.	14	**15**	.	**17**	**18**
17	105.12	14	89.02	8.05	0.63	**0**	.	**2**	.	.	**5**	**6**	7	.	.	.	.	12	13	**14**	**15**	.	**17**	**18**
19	105.16	14	87.35	8.90	0.72	**0**	1	**2**	.	.	**5**	**6**	.	.	.	.	.	12	.	14	**15**	.	**17**	**18**

Table 5-5. (continued).

Model No.	ICOMP -IFIM	*k*	-2LogL	Comp	Var	Independent Variables
24	105.27	14	88.52	8.37	0.75	**0** **1** **2** . . 5 **6** . . . 10 . . 13 . **15** . **17** **18**
25	105.28	14	89.23	8.02	0.61	**0** **1** **2** . 4 . **6** . . . . . . 13 14 **15** . **17** **18**
27	105.31	14	88.22	8.54	0.70	**0** 1 **2** . . 5 **6** 7 . . . . . 13 . **15** . **17** **18**
28	105.31	14	87.44	8.94	0.73	**0** 1 **2** . . **5** **6** . . . . 11 . . 14 **15** . **17** **18**
5	104.43	15	88.29	8.07	0.76	**0** . **2** 3 . **5** **6** . . . . . . 13 **14** **15** . **17** **18**
21	105.19	15	88.11	8.54	0.81	**0** . **2** 3 . **5** **6** . . . 10 . . . **14** **15** . **17** **18**
22	105.23	15	86.68	9.27	0.67	**0** 1 **2** . . 5 **6** . . . 10 . . 13 14 **15** . **17** **18**
26	105.30	15	86.48	9.41	0.64	**0** 1 **2** . . 5 **6** 7 . . . . . 13 14 **15** . **17** **18**
31	105.41	15	86.67	9.37	0.67	**0** 1 **2** . . 5 **6** . . . . . 12 13 14 **15** . **17** **18**
33	105.48	15	87.33	9.07	0.73	**0** 1 **2** . 4 5 6 . . . . 14 **15** . **17** **18**
35	105.57	15	89.13	8.22	0.69	**0** . **2** . 4 5 **6** . . . . . 12 13 **14** **15** . **17** **18**
36	105.58	15	88.41	8.59	0.77	**0** 1 **2** . 4 5 **6** . . . . . . 13 . **15** . **17** **18**
41	105.64	15	86.81	9.42	0.67	**0** 1 **2** . . 5 **6** . . 9 . . . 13 14 **15** . **17** **18**
44	105.68	15	86.93	9.38	0.67	**0** 1 **2** . . 5 **6** 7 . . 10 . . . 14 **15** . **17** **18**

[a]Summer variables are: 0= Intercept, 1= HABS, 2= DPOR, 3= LDS, 4= SMOS, 5= SCAN, 6= DES, 7= LOSD, 8= NLU, 9= EVG, 10= ESSD, 11= WCD, 12= DFW, 13= ROCK, 14= HSSD, 15= SSSD, 16= WSUN, 17= STMD, 18= DAES (see Tables 5-1 and 5-2 for defintions of variables).

Table 5-6. Summer models: the logistic regression parameter values for some of the top GA models for the *Anolis carolinensis* - habitat models. Estimated parameter values (Param.) and their associated standard errors (SE), Wald X^2 statistics (Wald), and *P* values are given for those GA models with 11 parameters (Models 1, 20, and 67). The degrees of freedom associated with each Wald statistic equals one for any given parameter estimate. For GA Models 2-19 the minimum and maximum values of the estimated parameters, standard errors, Wald statistics, and associated *P* values among those models are provided (if a parameter appeared in only one of these models then that model's number is given in parentheses and superscript in the column 'Param.'). See Table 5.1 and 5.2 for definitions of variables.

	Model 1		Model 20		Model 67		Models 2 - 19	
Variable	Param. (SE)	Wald (*P*)	Param. (SE)	Wald (*P*)	Param. (SE)	Wald (*P*)	Param. (SE)	Wald (*P*)
INT	-10.41	31.73	-7.59	25.63	-7.12	23.83	-10.39, -6.55	19.24, 31.36
	(1.85)	(<0.0001)	(1.50)	(<0.0001)	(1.46)	(<0.0001)	(1.43, 2.07)	(<0.0001)
HABS A	--	--	1.95	2.99	2.15	3.19	1.44, 2.55	1.33, 4.18
	--	--	(1.13)	(0.084)	(1.21)	(0.074)	(1.17, 1.28)	(0.041, 0.248)
HABS B	--	--	0.37	0.09	0.25	0.04	0.25, 0.74	0.04, 0.33
	--	--	(1.20)	(0.760)	(1.22)	(0.838)	(1.23, 1.32)	(0.566, 0.842)
DPOR1	5.18	23.93	3.53	6.64	3.34	5.96	3.55, 5.39	6.10, 23.14
	(1.06)	(<0.0001)	(1.37)	(0.010)	(1.37)	(0.015)	(1.07, 1.52)	(0.0001, 0.014)
DPOR2	3.66	8.99	2.24	2.45	2.06	2.00	2.06, 3.94	1.87, 9.15
	(1.22)	(0.003)	(1.43)	(0.118)	(1.46)	(0.158)	(1.22, 1.63)	(0.003, 0.155)
LDS1	--	--	--	--	--	--	-0.04 (5)	0.001
	--	--	--	--	--	--	(1.03)	(0.973)

Table 5-6. (continued).

Variable	Model 1 Param. (SE)	Model 1 Wald (*P*)	Model 20 Param. (SE)	Model 20 Wald (*P*)	Model 67 Param. (SE)	Model 67 Wald (*P*)	Models 2 - 19 Param. (SE)	Models 2 - 19 Wald (*P*)
LDS2	--	--	--	--	--	--	0.77	0.57
	--	--	--	--	--	--	(1.01)	(0.449)
LDS3	--	--	--	--	--	--	0.12	0.01
	--	--	--	--	--	--	(1.07)	(0.910)
SMOS1	--	--	--	--	--	--	-0.43[3]	0.36
	--	--	--	--	--	--	(0.72)	(0.550)
SMOS2	--	--	--	--	--	--	0.13	0.03
	--	--	--	--	--	--	(0.76)	(0.861)
SCAN1	2.21	4.45	--	--	--	--	0.996, 2.54	0.83, 5.03
	(1.05)	(0.035)	--	--	--	--	(1.06, 1.19)	(0.025, 0.362)
SCAN2	2.26	5.08	--	--	--	--	1.38, 2.34	1.97, 5.36
	(1.00)	(0.024)	--	--	--	--	(0.98, 1.09)	(0.021, 0.161)
DES1	0.86	1.07	0.60	0.49	0.47	0.31	-0.002, 0.82	0.000, 0.90
	(0.83)	(0.302)	(0.86)	(0.484)	(0.85)	(0.579)	(0.83, 1.00)	(0.342, 0.999)
DES2	1.68	5.84	1.47	3.73	1.54	4.14	1.35, 1.85	3.09, 6.30
	(0.70)	(0.016)	(0.76)	(0.053)	(0.76)	(0.042)	(0.70, 0.79)	(0.012, 0.079)

Table 5-6. (continued).

Variable	Model 1 Param. (SE)	Model 1 Wald (*P*)	Model 20 Param. (SE)	Model 20 Wald (*P*)	Model 67 Param. (SE)	Model 67 Wald (*P*)	Models 2 - 19 Param. (SE)	Models 2 - 19 Wald (*P*)
LOSD	--	--	--	--	--	--	-0.56, -0.44	0.45, 0.78
	--	--	--	--	--	--	(0.63, 0.65)	(0.378, 0.501)
EVG	--	--	--	--	--	--	-0.22 (15)	0.02
	--	--	--	--	--	--	(1.46)	(0.883)
ESSD	--	--	--	--	--	--	-0.61, -0.37	0.09, 0.22
	--	--	--	--	--	--	(1.21, 1.30)	(0.641, 0.762)
DFW	--	--	--	--	0.19	0.07	-0.15, 0.30	0.003, 0.17
	--	--	--	--	(0.71)	(0.793)	(0.68, 0.73)	(0.680, 0.957)
HSSD	1.42	4.53	0.68	1.06	--	--	0.57, 1.41	0.72, 4.47
	(0.67)	(0.033)	(0.66)	(0.303)	--	--	(0.66, 0.71)	(0.035, 0.396)
SSSD	1.33	4.09	1.73	7.00	1.72	7.04	1.28, 1.68	3.62, 6.74
	(0.66)	(0.043)	(0.66)	(0.008)	(0.65)	(0.008)	(0.64, 0.68)	(0.009, 0.057)
STMD	1.49	5.88	1.22	4.09	1.19	3.90	1.05, 1.50	2.90, 5.87
	(0.61)	(0.015)	(0.60)	(0.043)	(0.61)	(0.048)	(0.60, 0.64)	(0.015, 0.089)
DAES	2.27	11.22	2.26	10.86	2.15	10.15	2.07, 2.29	9.20, 11.27
	(0.68)	(0.001)	(0.69)	(0.001)	(0.67)	(0.001)	(0.67, 71)	(0.001, 0.002)

Table 5-7. Univariate logistic regression summary information for the winter habitat variables. Variable No. and Name refer to the number and abbreviated name for each habitat variable, respectively, as given in Table 5-3 and Outcome is the outcome of the survey of each plot (P= presence and A= absence of *Anolis carolinensis*). ICOMP-IFIM is the model selection criterion value for the univariate logistic regression model which includes the habitat variable and the intercept term. a. For each categorical variable the following summary information is reported: the categories (Level) of a variable (the reference cell is given last and p= presence and a= absence of a particular habitat feature; see Table 5-3 for descriptions), the numbers of observed plots having *Anolis carolinensis* present or absent in each level (Outcome), and ICOMP-IFIM for the univariate model. b. For the continuous variable WSUN (see Table 5-3 for description) the following summary information is reported: ICOMP-IFIM for the univariate model and, for each possible outcome of the plot survey, the 25% quantile (Q), median, 75% quantile, mean, and standard deviation of the mean (Stand. Dev.).

a.

Independent Variable	Level	Outcome P	Outcome A	ICOMP-IFIM
1. HAB	A	34	17	186.61
	B	16	27	
	C	5	46	
	D	7	14	
2. NLUW	1	48	86	216.32
	2	2	12	
	3	12	6	
3. WTM	1	19	36	202.48
	2	33	23	
	3	10	45	
4. DPOR	1	47	31	182.55
	2	7	11	
	3	8	62	
5. NLOW	1	51	82	222.21
	0	11	22	

Table 5-7. (continued).

a.

Independent Variable	Level	Outcome P	Outcome A	ICOMP-IFIM
6. LDW	1	40	45	214.16
	0	22	59	
7. EVG	p	8	27	217.27
	a	54	77	
8. SOTW	1	29	32	216.72
	0	33	72	
9. ESWD	1	48	59	214.31
	0	14	45	
10. SCW	1	29	26	212.77
	0	33	78	
11. WCD	1	57	64	202.64
	0	5	40	
12. DFW	p	42	85	217.86
	a	20	19	
13. ROCK	p	31	7	180.14
	a	31	97	
14. HSCW	1	50	70	218.58
	0	12	34	
15. SSWD	1	50	67	217.00
	0	12	37	
17. DEW	1	30	24	209.81
	0	32	80	
18. DAEW	1	22	19	215.00
	0	40	85	

Table 5-7. (continued).

b.

Independent Variable	Outcome Variable	25% Q	Median	75% Q	Mean	Stand. Dev.	ICOMP-IFIM
16. WSUN	P	82.0	89.0	95.5	87.0	10.5	200.66
	A	58.5	76.25	88.25	72.6	17.7	

Fig. 5-4. Winter GA models: the frequency of independent variables in the best 184 logistic regression models from the genetic algorithm (GA) output modeling the relationship between habitat features and the presence of *Anolis carolinensis* in winter plots. Percent represents the percentage of best winter GA models in which a given variable occurred. Variable acronyms are defined in Table 5-3.

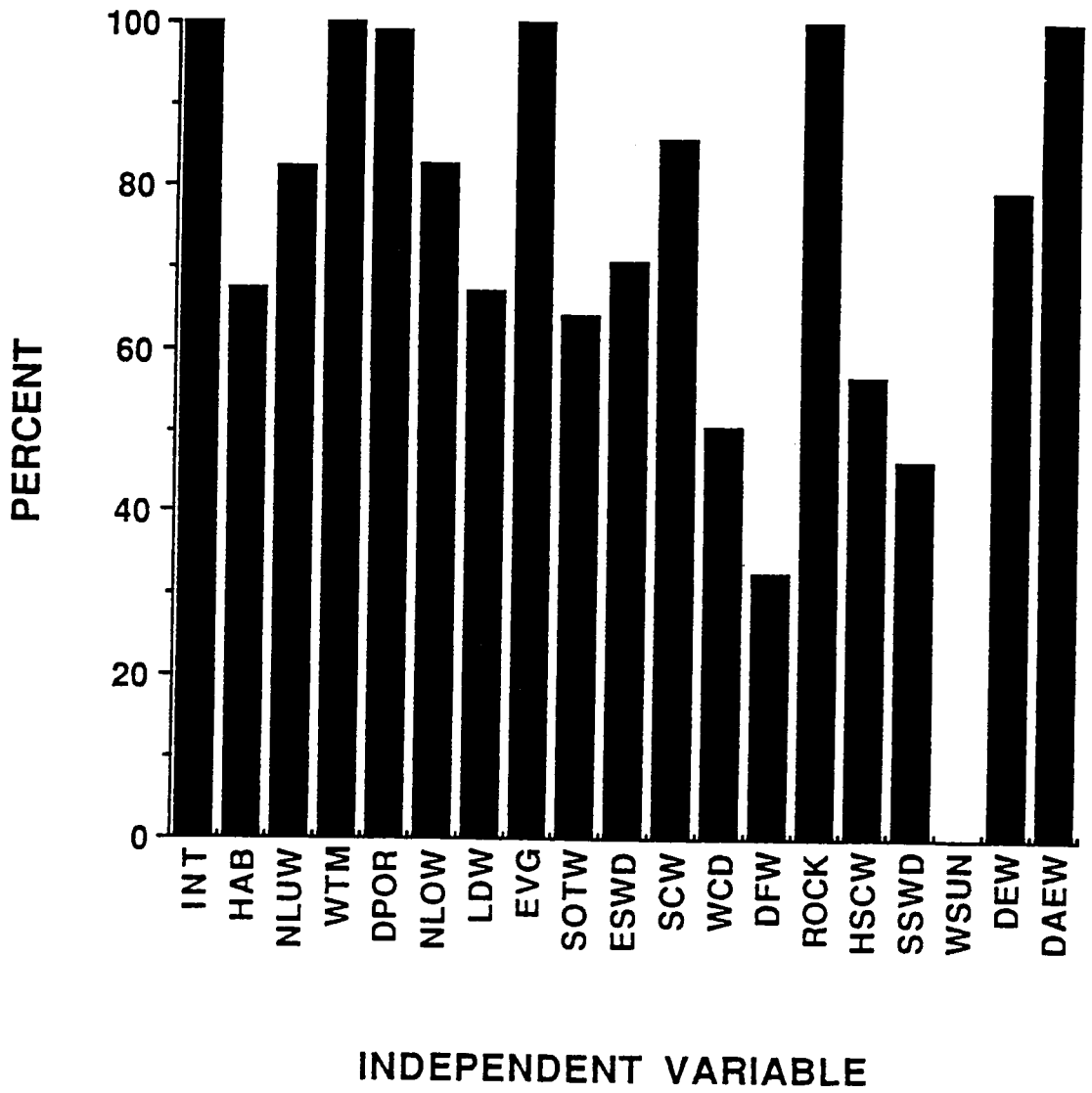
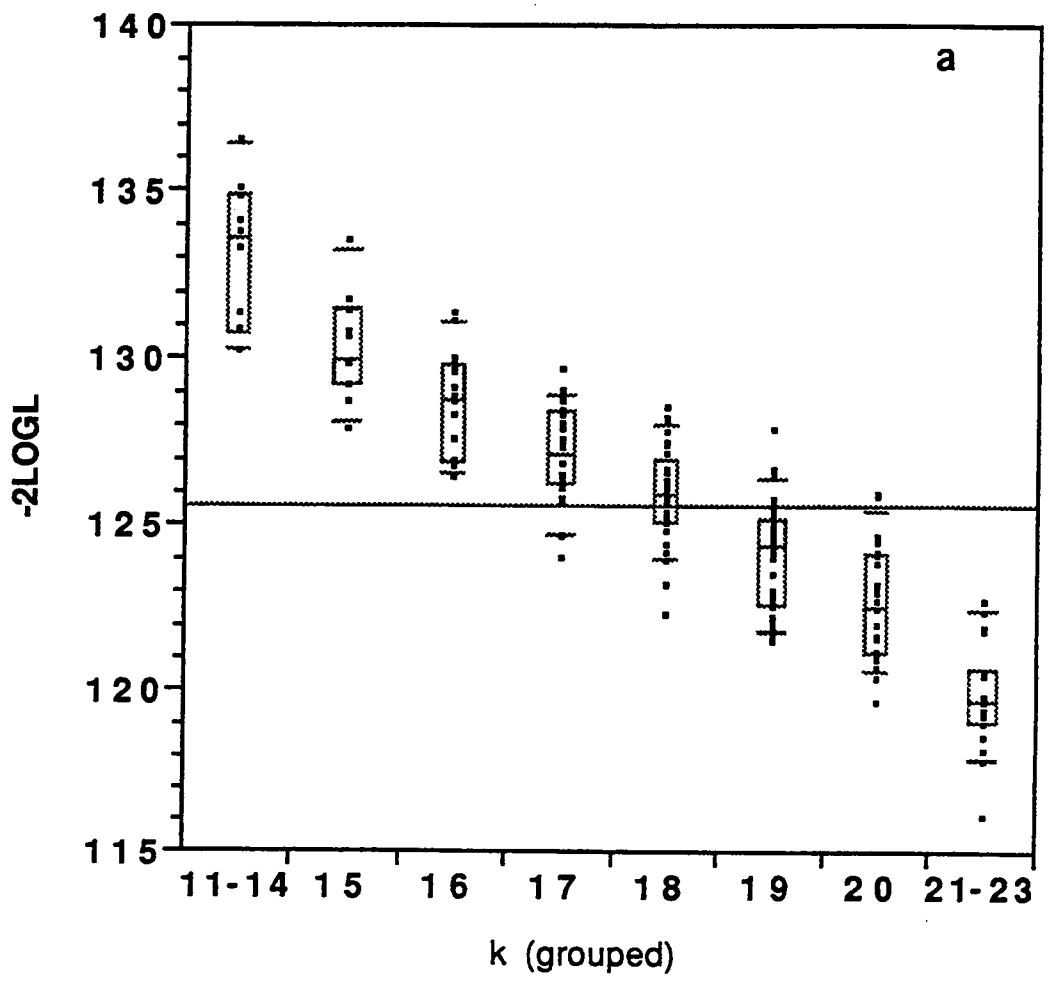
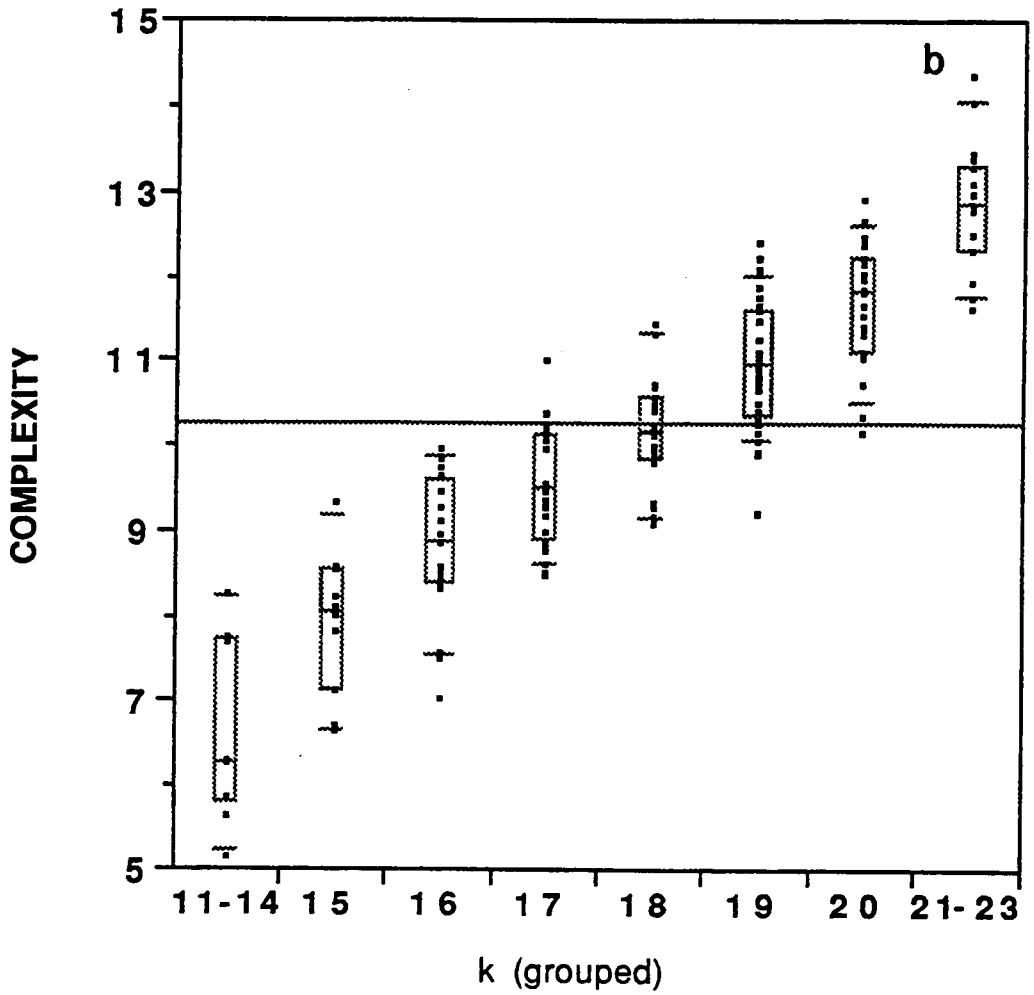

PERCENT
100
80
60
40
20
0
INT
HAB
NLUW
WTM
DPOR
NLOW
LDW
EVG
SOTW
ESWD
SCW
WCD
DFW
ROCK
HSCW
SSWD
WSUN
DEW
DAEW
INDEPENDENT VARIABLE

Fig. 5-5. Winter GA models: box plots showing trends in the (a) lack-of-fit term (-2LogL), (b) complexity term, and (c) model selection criterion, ICOMP-IFIM, across the different model sizes, represented by k (number of estimated regression parameters), for the best 184 winter logistic regression models found by the genetic algorithm (GA) analysis. Some models with different k values were grouped together so that no level of k had less than 5% of the 184 total models. The line across the graph parallel to the X-axis shows the mean value of the given term for the 184 models. The line within each box represents the median for the given level of k. The 25% and 75% quantiles are represented by the ends of a box, while the 10% and 90% quantiles are shown as the short lines outside the ends of a box.


a
140
135
130
125
120
115
-2LOGL
11-14
15
16
17
18
19
20
21-23
k (grouped)


15
13
11
9
7
5
COMPLEXITY
b
11-14
15
16
17
18
19
20
21-23
k (grouped)

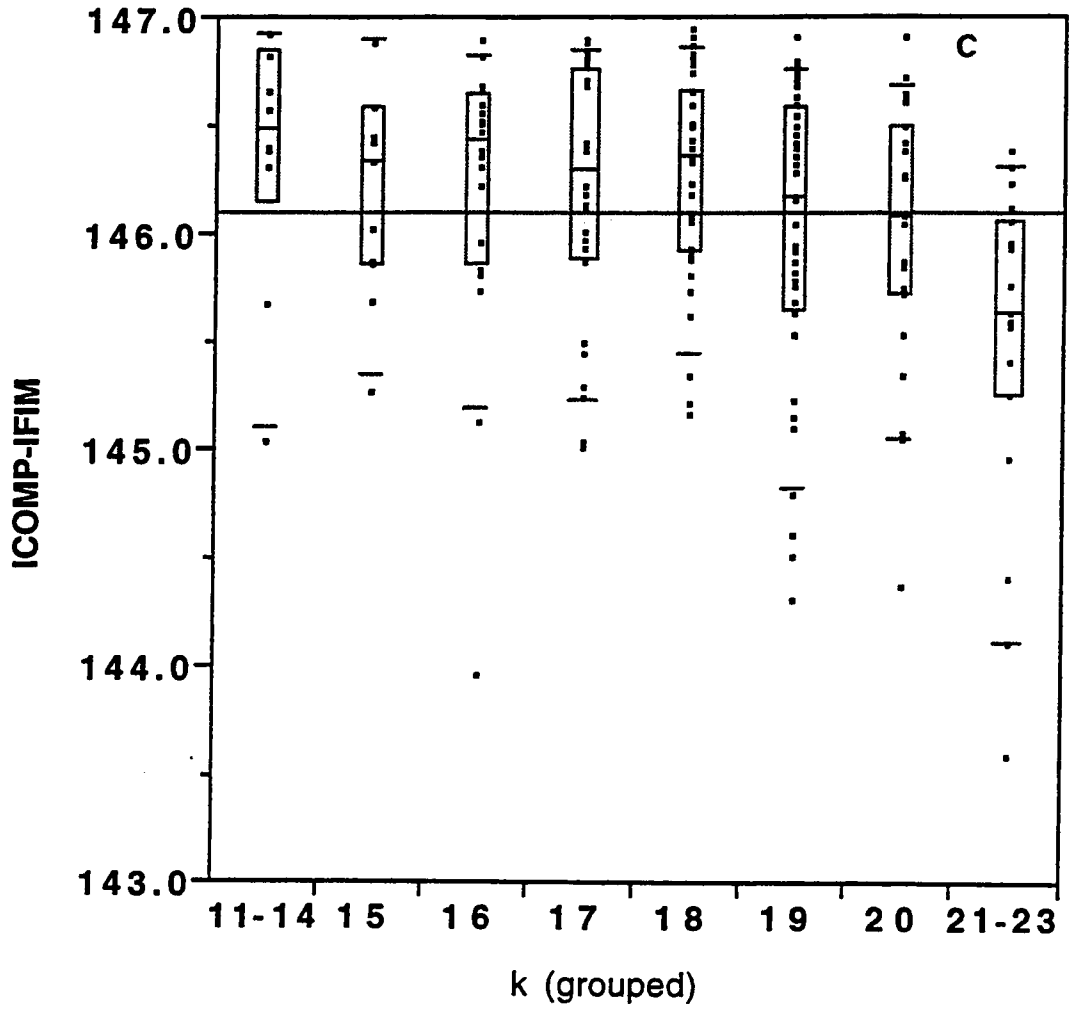

147.0
146.0
145.0
144.0
143.0
ICOMP-IFIM
C
11-14
15
16
17
18
19
20
21-23
k (grouped)

Table 5-8. Winter models: the 20 logistic regression models having the lowest criterion values from the genetic algorithm (GA) output for modeling the relationship between habitat variables and the presence of *Anolis carolinensis* in winter plots. The heading Model No. gives the rank order of each model relative to all other GA models based on the ICOMP-IFIM (Informational Complexity of the Inverse Fisher Information Matrix) values, where the lowest values represent the better models. *k*= the number of estimated regression parameters in the model, -2LogL= (-2) times the loglikelihood value (or lack-of-fit term) for the model, Comp= the complexity (penalty) term calculated as the complexity of the inverse Fisher information matrix, Var= the estimated model variance, and the numbers[a] under the heading of Independent Variables identify the independent variables which are present in a model (a dot indicates the given variable is absent from the model).

Model No.	ICOMP -IFIM	*k*	-2LogL	Comp	Var	Independent Variables																		
1	143.61	21	117.88	12.87	5.80	0	1	2	3	4	5	.	7	8	9	10	11	.	13	14	15	.	17	18
2	143.97	16	129.91	7.03	1.63	0	.	.	3	4	5	6	7	8	.	10	11	.	13	14	15	.	17	18
3	144.12	21	119.44	12.34	3.32	0	1	2	3	4	5	6	7	8	9	10	11	.	13	14	.	.	17	18
4	144.33	19	124.53	9.90	3.79	0	1	.	3	4	5	.	7	8	9	10	11	.	13	14	15	.	17	18
5	144.38	20	120.39	12.00	7.18	0	1	2	3	4	5	6	7	8	9	10	.	.	13	.	15	.	17	18
6	144.43	21	119.39	12.52	6.72	0	1	2	3	4	5	6	7	8	9	10	11	.	13	.	15	.	17	18
7	144.53	19	121.56	11.48	4.36	0	1	2	3	4	5	6	7	8	9	10	.	.	13	.	.	.	17	18
8	144.63	19	122.79	10.92	2.15	0	1	2	3	4	5	6	7	8	.	10	.	.	13	14	.	.	17	18
9	144.81	19	121.55	11.63	3.60	0	1	2	3	4	5	6	7	8	9	10	11	.	13	.	.	.	.	18
10	144.98	23	116.18	14.40	6.74	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	.	17	18
11	145.03	17	124.65	10.19	2.40	0	.	2	3	4	5	6	7	8	9	10	.	.	13	14	.	.	17	18

Table 5-8. (continued).

Model No.	ICOMP -IFIM	*k*	-2LogL	Comp	Var	Independent Variables
12	145.05	17	126.44	9.30	1.37	0 . 2 3 4 5 6 7 8 . 10 11 . 13 14 . . 17 18
13	145.06	13	133.29	5.88	1.62	0 . . 3 4 . . 7 . . 10 11 . 13 14 15 . 17 18
14	145.06	20	120.67	12.19	5.83	0 1 2 3 4 5 . 7 8 9 10 11 . 13 . 15 . 17 18
15	145.09	20	120.96	12.06	2.71	0 1 2 3 4 5 . 7 8 9 10 11 . 13 14 . . 17 18
16	145.12	19	123.06	11.03	3.23	0 1 2 3 4 5 6 7 8 . 10 . . 13 . 15 . 17 18
17	145.14	16	126.91	9.12	3.04	0 . 2 3 4 5 . 7 . 9 10 . . 13 14 15 . 17 18
18	145.16	19	121.83	11.66	2.63	0 1 2 3 4 5 6 7 8 9 . . . 13 14 . . 17 18
19	145.18	18	125.27	9.95	3.61	0 1 2 3 4 . 6 7 8 9 10 . . 13 . . . 17 18
20	145.23	18	122.32	11.46	5.24	0 1 2 3 4 5 6 7 8 9 . . . 13 . 15 . . 18

[a]Winter variables are: 0= Intercept, 1= HAB, 2= NLUW, 3= WTM, 4= DPOR, 5= NLOW, 6= LDW, 7= EVG, 8= SOTW, 9= ESWD, 10= SCW, 11= WCD, 12= DFW, 13= ROCK, 14= HSCW, 15= SSWD, 16= WSUN, 17= DEW, 18= DAEW (see Tables 5-1 and 5-3 for definitions of variables).

Fig. 5-6. Final best winter models: box plots showing lack of any clear trend in the model selection criterion, ICOMP-IFIM, across the different model sizes, represented by k (number of estimated regression parameters), for the final best winter logistic regression models (n = 154) found by the genetic algorithm (GA) analysis and subsequent subset analysis. Some models with different k values were grouped together so that no level of k had less than 5% of the 1 total models. The line across the graph parallel to the X-axis shows the mean value of ICOMP-IFIM for the final winter models. The line within each box represents the median ICOMP-IFIM for the given level of k. The 25% and 75% quantiles are represented by the ends of a box, while the 10% and 90% quantiles are shown as the short lines outside the ends of a box.

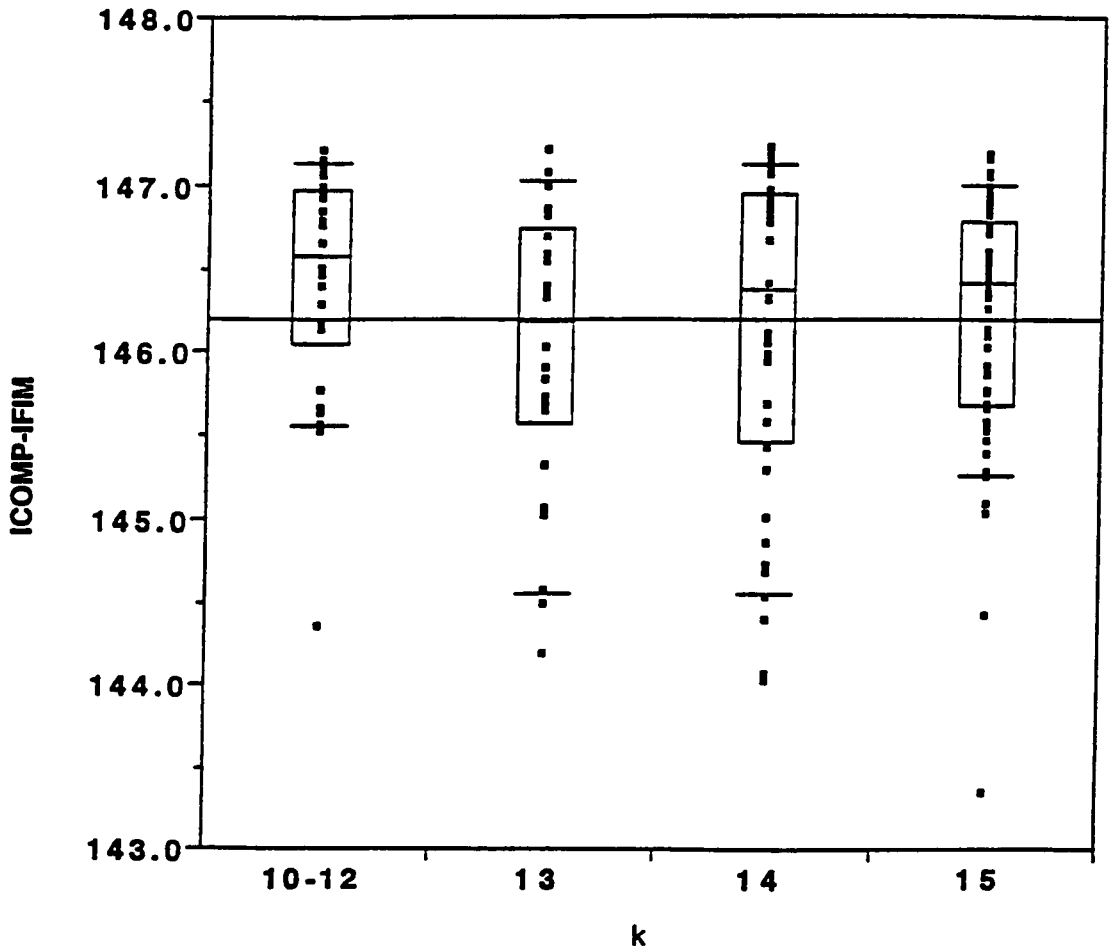

148.0
147.0
146.0
145.0
144.0
143.0
ICOMP-IFIM
10-12
13
14
15
k

Fig. 5-7. Final best winter models: the frequency of independent variables in the final best logistic regression models (n = 154) from the combined results of the genetic algorithm (GA) output and subsequent subset analysis modeling the relationship between habitat features and the presence of *Anolis carolinensis* in winter plots. Percent represents the percentage of final best winter models in which a given variable occurred. Variable acronyms are defined in Table 5-3.

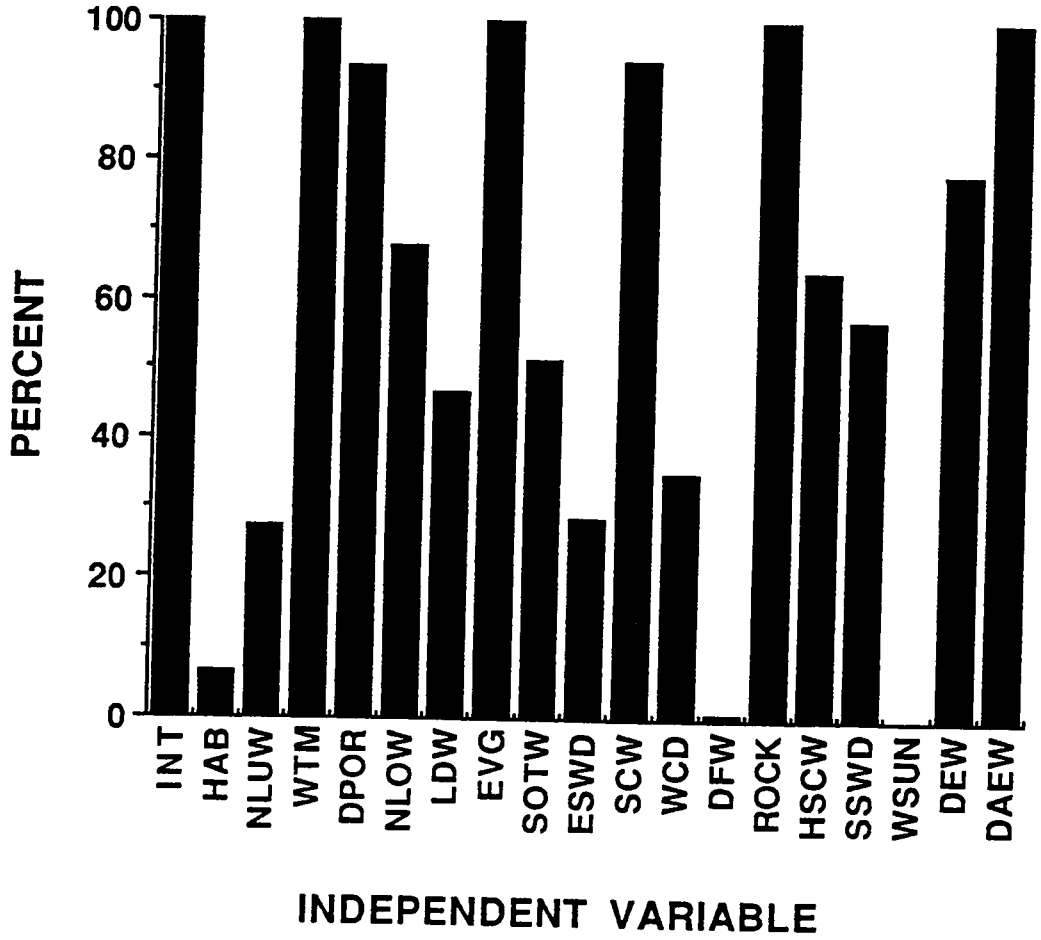

PERCENT
100
80
60
40
20
0
INT
HAB
NLUW
WTM
DPOR
NLOW
LDW
EVG
SOTW
ESWD
SCW
WCD
DFW
ROCK
HSCW
SSWD
WSUN
DEW
DAEW
INDEPENDENT VARIABLE

Fig. 5-8. Final best winter models: trends in the frequency of independent variables in the different model sizes (k levels) in the final best logistic regression models (n = 154) from the combined results of the genetic algorithm (GA) output and subsequent subset analysis modeling the relationship between habitat features and the presence of *Anolis carolinensis* in winter plots. (a) k = 15 estimated parameters (n = 62). (b) k = 14 estimated parameters (n = 40). (c) k = 13 estimated parameters (n = 26). (d) k = 10 to 12 estimated parameters (n = 26). Percent represents the percentage of final best winter models of a specific k size in which a given variable occurred. Variable acronyms are defined in Table 5-3.

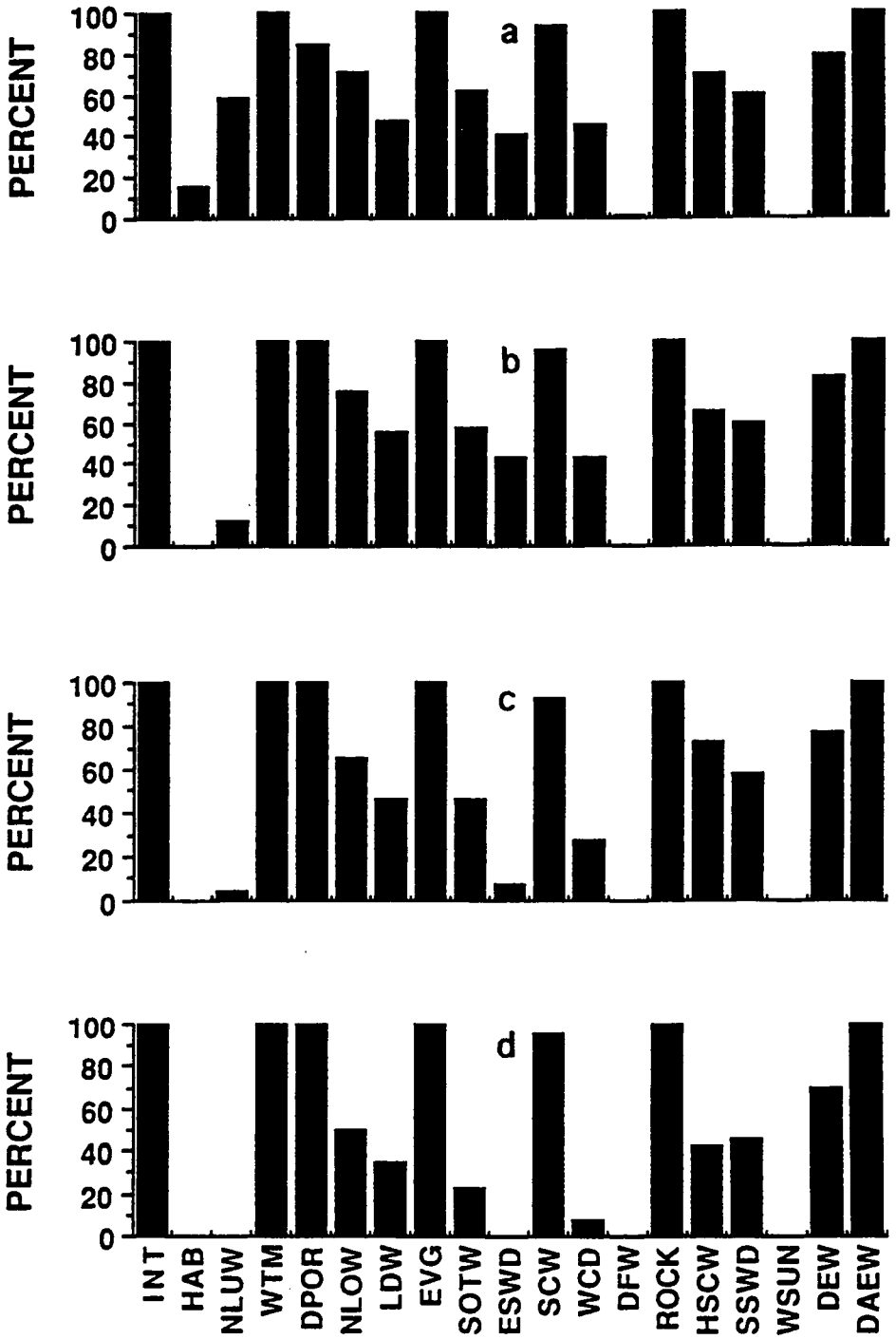

PERCENT
100
80
60
40
20
0
a
b
c
d
INT
HAB
NLUW
WTM
DPOR
NLOW
LDW
EVG
SOTW
ESWD
SCW
WCD
DFW
ROCK
HSCW
SSWD
WSUN
DEW
DAEW

Table 5-9. Winter models: the final best logistic regression models with 15 or fewer parameters (k) and model variance < 3.00 from both the genetic algorithm (GA) output and subsequent subset analyses for modeling the relationship between habitat variables and the presence of *Anolis carolinensis* in winter study plots. The table shows only a maximum of the ten best models within each k group. The headings for each column are the same as those in Table 5-5 except that Model No. represents the rankings after models with variance ≥ 3.00 and $k \geq 16$ were removed (Model 1, however, had the lowest ICOMP-IFIM value of any models found by either the GA or subset analyses). The numbers[a] representing the independent variables identify those variables in the winter models (note that these numbers do not necessarily correspond to the same variables as in the summer models). Variable numbers in bold type follow the same convention as in Table 5-5, but an underlined variable number represents a parameter with an associated P value = 0.100.

Model No.	ICOMP -IFIM	k	-2LogL	Comp	Var	Independent Variables
144	147.12	10	137.55	4.78	1.38	**0** . . **3** **4** . . **7** . . **10** . . **13** . . . 17 **18**
28	145.53	11	135.21	5.16	1.35	**0** . . **3** **4** . . **7** . . **10** . . **13** 14 . . **17** **18**
65	146.18	11	136.30	4.94	1.71	**0** . . **3** **4** . . **7** . . **10** . . **13** . 15 . 17 **18**
113	146.85	11	136.28	5.29	1.50	**0** . . **3** **4** 5 . **7** . . **10** . . **13** . 15 . . **18**
124	146.95	11	136.58	5.19	1.33	**0** . . **3** **4** 5 . **7** . . **10** . . **13** . . . 17 **18**
136	147.07	11	137.02	5.02	1.50	**0** . . **3** **4** . 6 **7** . . **10** . . **13** . . . 17 **18**
152	147.22	11	136.45	5.38	1.23	**0** . . **3** **4** 5 . **7** . . 10 . . **13** 14 . . . **18**
5	144.36	12	133.52	5.42	1.78	**0** . . **3** **4** . . **7** . . **10** . . **13** 14 15 . **17** **18**
31	145.57	12	134.50	5.53	1.30	**0** . . **3** **4** 5 . **7** . . **10** . . **13** 14 . . 17 **18**

Table 5-9. (continued).

Model No.	ICOMP -IFIM	*k*	-2LogL	Comp	Var	Independent Variables
36	145.64	12	134.89	5.38	1.43	**0** . . **3** **4** . 6 **7** . . **10** . . **13** 14 . . 17 **18**
38	145.67	12	135.00	5.34	1.33	**0** . . **3** **4** . . **7** 8 . **10** . . **13** 14 . . **17** **18**
49	145.78	12	134.92	5.43	1.69	**0** . . **3** **4** 5 . **7** . . **10** . . **13** . 15 . 17 **18**
63	146.14	12	135.80	5.17	1.89	**0** . . **3** **4** . 6 **7** . . **10** . . **13** . 15 . 17 **18**
67	146.30	12	134.90	5.70	1.48	**0** . . **3** **4** 5 . **7** . . 10 . . **13** 14 15 . . **18**
77	146.41	12	135.08	5.66	1.29	**0** . . **3** **4** . . **7** . . **10** 11 . **13** 14 . . **17** **18**
81	146.47	12	136.21	5.13	1.67	**0** . . **3** **4** . . **7** 8 . **10** . . **13** . 15 . 17 **18**
83	146.48	12	135.99	5.24	1.53	**0** . . **3** **4** . 6 **7** 8 . **10** . . **13** . . . 17 **18**
4	144.20	13	132.40	5.90	1.74	**0** . . **3** **4** 5 . **7** . . 10 . . **13** 14 15 . 17 **18**
8	144.49	13	133.25	5.62	1.90	**0** . . **3** **4** . 6 **7** . . **10** . . **13** 14 15 . **17** **18**
10	144.58	13	133.36	5.61	1.74	**0** . . **3** **4** . . **7** 8 . **10** . . **13** **14** 15 . **17** **18**
15	145.02	13	133.83	5.60	1.46	**0** . . **3** **4** . 6 **7** 8 . **10** . . **13** 14 . . **17** **18**
17	145.06	13	133.29	5.88	1.62	**0** . . **3** **4** . . **7** . . **10** 11 . **13** **14** 15 . **17** **18**
24	145.33	13	133.87	5.73	1.25	**0** . . **3** **4** 5 . **7** 8 . **10** . . **13** 14 . . 17 **18**
37	145.65	13	134.36	5.65	1.60	**0** . . **3** **4** 5 . **7** 8 . **10** . . **13** . 15 . 17 **18**
42	145.70	13	134.15	5.77	1.44	**0** . . **3** **4** 5 6 **7** 8 . **10** . . **13** . . . 17 **18**
43	145.72	13	134.28	5.72	1.36	**0** . . **3** **4** 5 6 **7** . . **10** . . **13** 14 . . 17 **18**

Table 5-9. (continued).

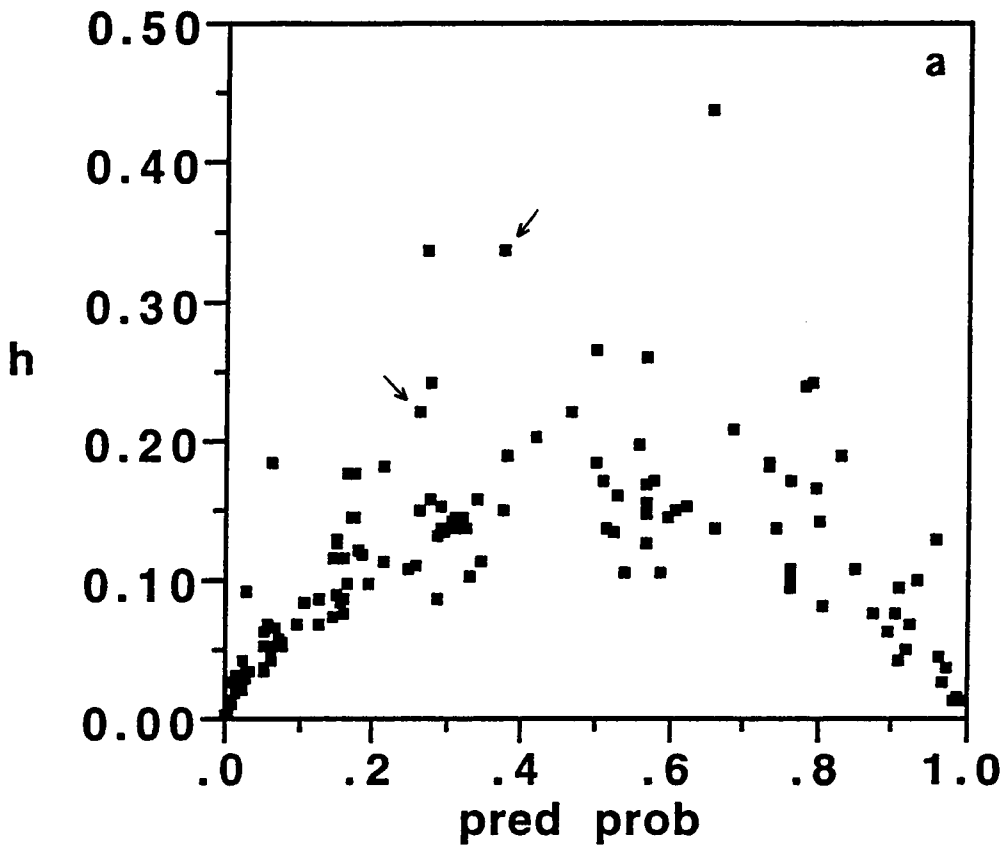
Model No.	ICOMP -IFIM	*k*	-2LogL	Comp	Var	Independent Variables
44	145.74	13	134.91	5.41	1.93	**0** . . **3** **4** . 6 **7** 8 . **10** . . **13** . 15 . 17 **18**
2	144.03	14	131.74	6.15	1.66	**0** . . **3** **4** 5 . **7** 8 . **10** . . **13** 14 15 . **17** **18**
3	144.07	14	132.36	5.85	1.94	**0** . . **3** **4** . 6 **7** 8 . **10** . . **13** 14 15 . **17** **18**
6	144.40	14	132.24	6.08	1.83	**0** . . **3** **4** 5 6 **7** . . 10 . . **13** 14 15 . 17 **18**
9	144.53	14	132.26	6.13	1.38	**0** . . **3** **4** 5 6 **7** 8 . **10** . . **13** 14 . . 17 **18**
11	144.68	14	131.99	6.34	1.54	**0** . . **3** **4** 5 . **7** . . **10** 11 . **13** 14 15 . 17 **18**
12	144.73	14	132.62	6.05	1.90	**0** . . **3** **4** 5 6 **7** 8 . **10** . . **13** . 15 . 17 **18**
13	144.86	14	131.15	6.86	2.69	**0** . . **3** **4** 5 . **7** . 9 10 . . **13** 14 15 . **17** **18**
14	145.01	14	132.14	6.44	2.51	**0** . . **3** **4** . . **7** 8 9 **10** . . **13** **14** 15 . **17** **18**
23	145.30	14	133.09	6.10	1.58	**0** . . **3** **4** . . **7** 8 . **10** 11 . **13** **14** 15 . **17** **18**
26	145.43	14	133.10	6.17	1.73	**0** . . **3** **4** . 6 **7** . . **10** 11 . **13** 14 15 . **17** **18**
1	143.36	15	130.27	6.55	1.91	**0** . . **3** **4** 5 6 **7** 8 . **10** . . **13** 14 15 . 17 **18**
7	144.42	15	131.14	6.64	1.45	**0** . . **3** **4** 5 . **7** 8 . **10** 11 . **13** **14** 15 . **17** **18**
16	145.04	15	132.25	6.39	1.78	**0** . . **3** **4** . 6 **7** 8 . **10** 11 . **13** 14 15 . **17** **18**
18	145.10	15	129.03	8.04	1.70	**0** . **2** **3** **4** 5 . **7** . . **10** . . **13** 14 15 . 17 **18**
19	145.10	15	131.91	6.59	1.60	**0** . . **3** **4** 5 6 **7** . . **10** 11 . **13** 14 15 . 17 **18**

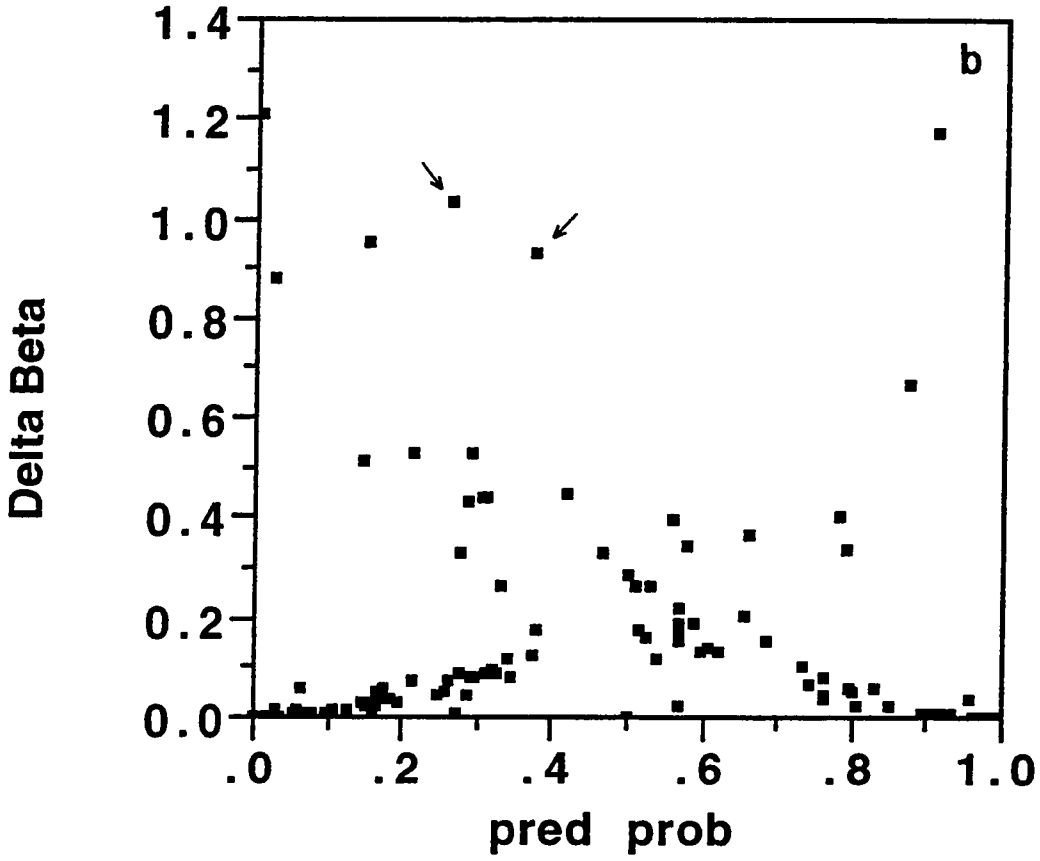
Table 5-9. (continued).

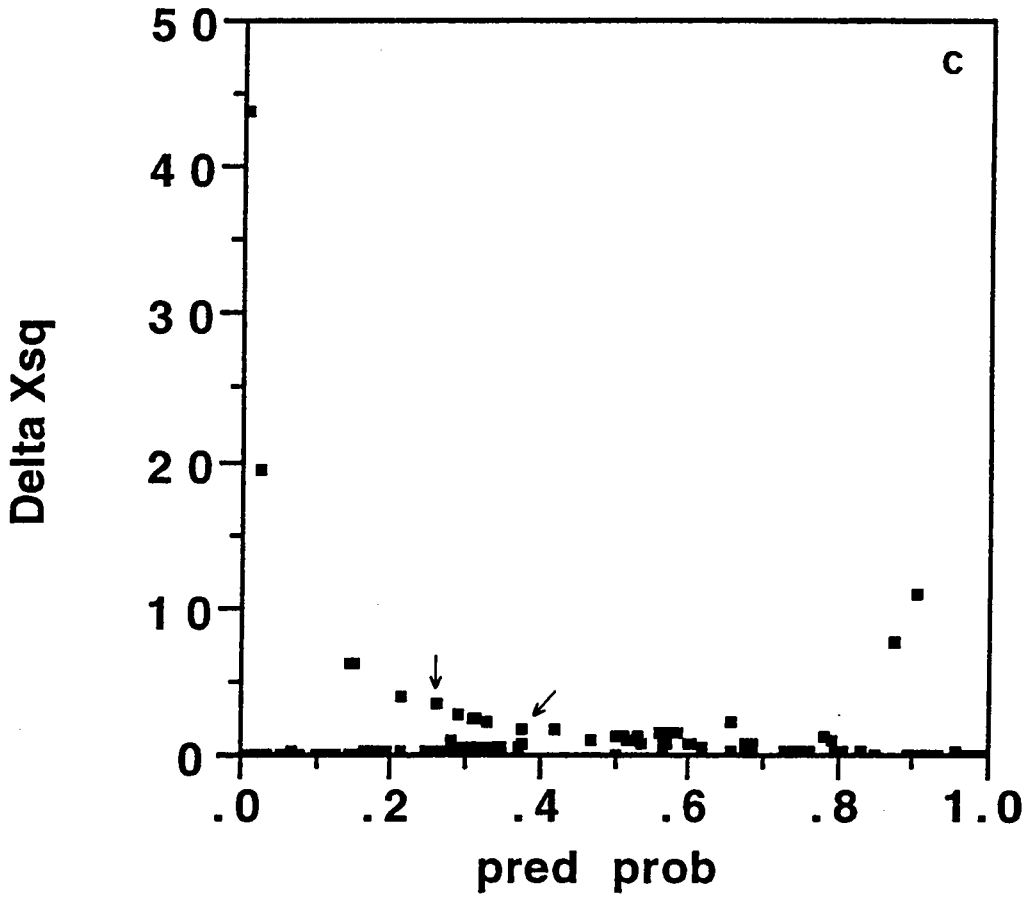
Model No.	ICOMP -IFIM	*k*	-2LogL	Comp	Var	Independent Variables																	
20	145.25	15	130.72	7.27	2.27	**0**	.	.	**3**	**4**	5	.	**7**	.	9	**10**	11	.	**13**	**14**	15	.	**17**	**18**
21	145.28	15	131.84	6.72	1.38	**0**	.	.	**3**	**4**	5	6	**7**	8	.	**10**	.	12	**13**	14	.	.	**17**	**18**
22	145.29	15	129.09	8.10	1.66	**0**	.	**2**	**3**	**4**	.	6	**7**	8	.	**10**	.	.	**13**	14	.	.	**17**	**18**
25	145.39	15	132.09	6.65	1.27	**0**	.	.	**3**	**4**	5	6	**7**	8	.	**10**	11	.	**13**	14	.	.	**17**	**18**
27	145.47	15	128.13	8.67	1.62	**0**	.	**2**	**3**	**4**	5	**6**	**7**	**8**	.	**10**	.	.	**13**	.	.	.	17	**18**

[a]Winter variables are: 0= Intercept, 1= HAB, 2= NLUW, 3= WTM, 4= DPOR, 5= NLOW, 6= LDW, 7= EVG, 8= SOTW, 9= ESWD, 10= SCW, 11= WCD, 12= DFW, 13= ROCK, 14= HSCW, 15= SSWD, 16= WSUN, 17= DEW, 18= DAEW (see Tables 5-1 and 5-3 for definitions of variables).

Fig. 5-9. Winter Model 1 (k = 15, ICOMP-IFIM = 143.36, model variance = 1.91): graphical presentation of logistic regression diagnostic measures. (a) Plot of leverage (h_j) versus predicted probability. In general, h_j is expected to be: small when the predicted (estimated) probability is between 0-0.1 and 0.9-1.0, moderate to small when predicted probability is between 0.3-0.7, and large when predicted probability is between 0.1-0.3 and 0.7-0.9 (Hosmer and Lemeshow 1989:157). (b) Plot of $\Delta\beta_j$ versus predicted probability. $\Delta\beta_j$ measures the change in the estimated parameter values of the logistic regression model, in general, when a particular covariate pattern is deleted from the model. Note that only a few points stand out or fall away from the others (in this case those ≥ 0.6). (c) Plot of ΔX^2_j versus predicted probability. ΔX^2_j measures the change in the Pearson chi-square statistic, a summary measure of the goodness-of-fit of a model, when a particular covariate pattern is removed from the model. (d) Plot of ΔD_j versus predicted probability. ΔD_j measures the change in the deviance, a summary measure of the goodness-of-fit of a model, when a particular covariate pattern is removed from the model. Model 1 fits the winter data fairly well across the covariate patterns because only about 10% of the covariates had either ΔX^2_j or ΔD_j greater than the conservative cutoff of 2.71. Overall, the model fits very well as suggested by the fact that only two points, those indicated by the arrows, had moderate to high values for leverage and poor fit (large values of h_j, $\Delta\beta_j$, and either ΔX^2_j or ΔD_j).


0.50
0.40
0.30
0.20
0.10
0.00
h
a
.0
.2
.4
.6
.8
1.0
pred prob


b
Delta Beta
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
.0
.2
.4
.6
.8
1.0
pred prob


5 0
4 0
3 0
2 0
1 0
0
Delta Xsq
c
.0
.2
.4
.6
.8
1.0
pred prob

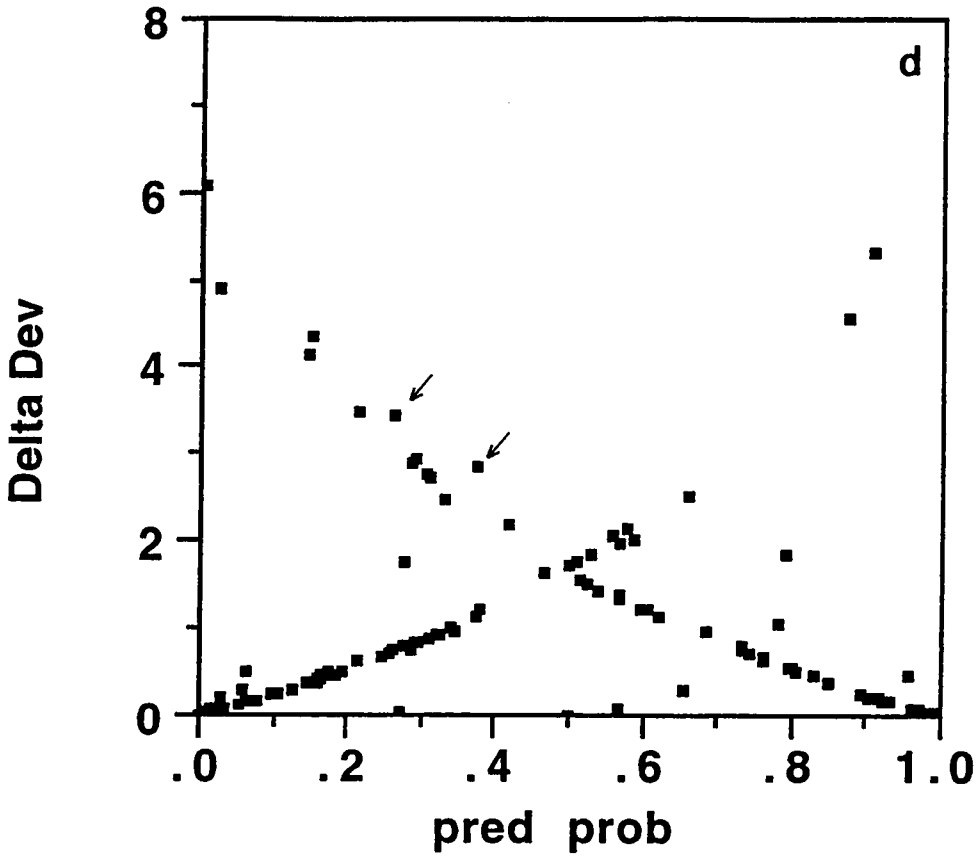
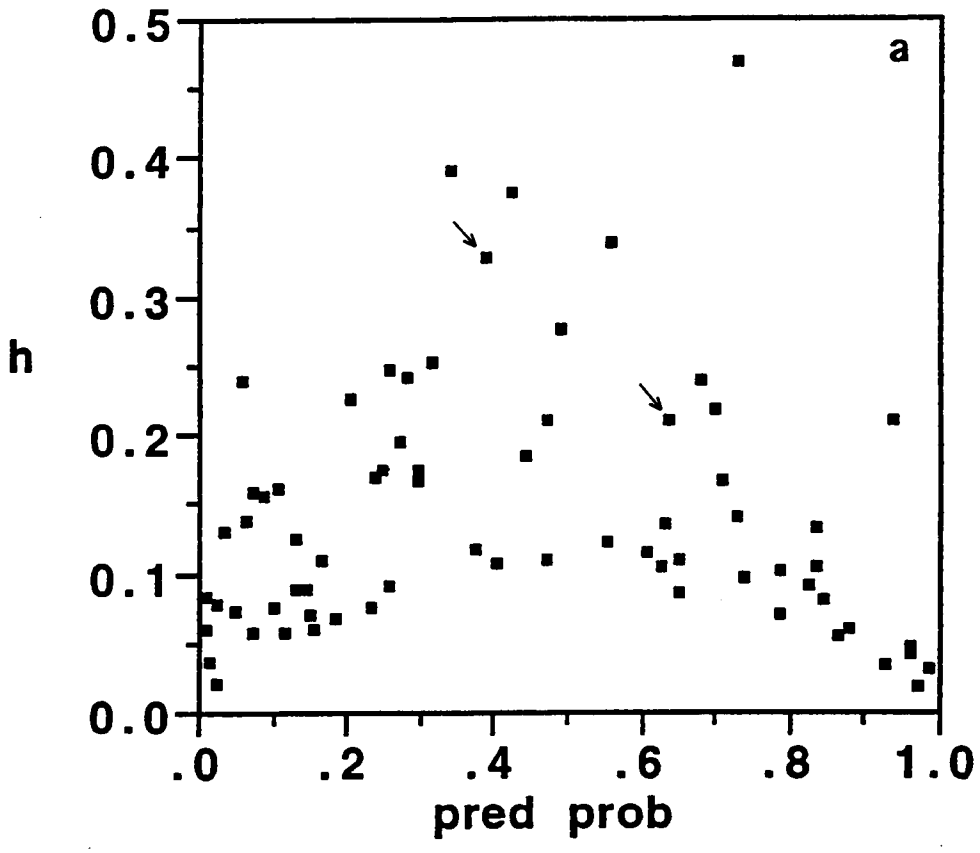
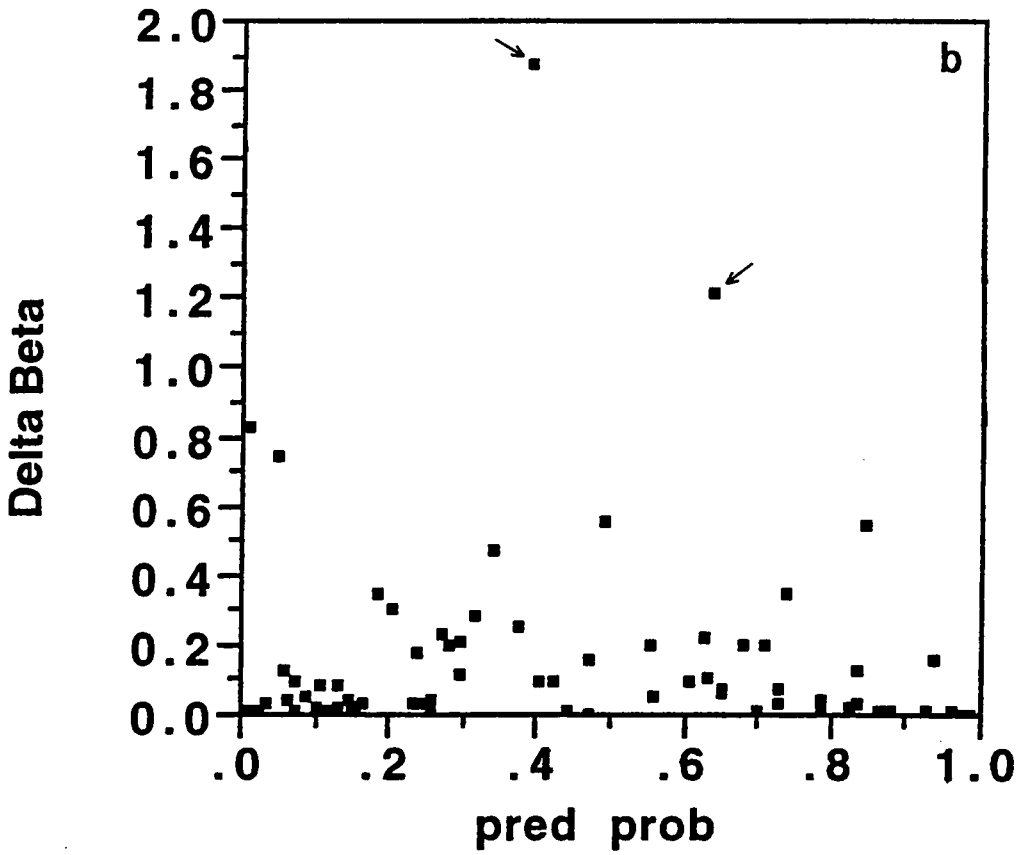
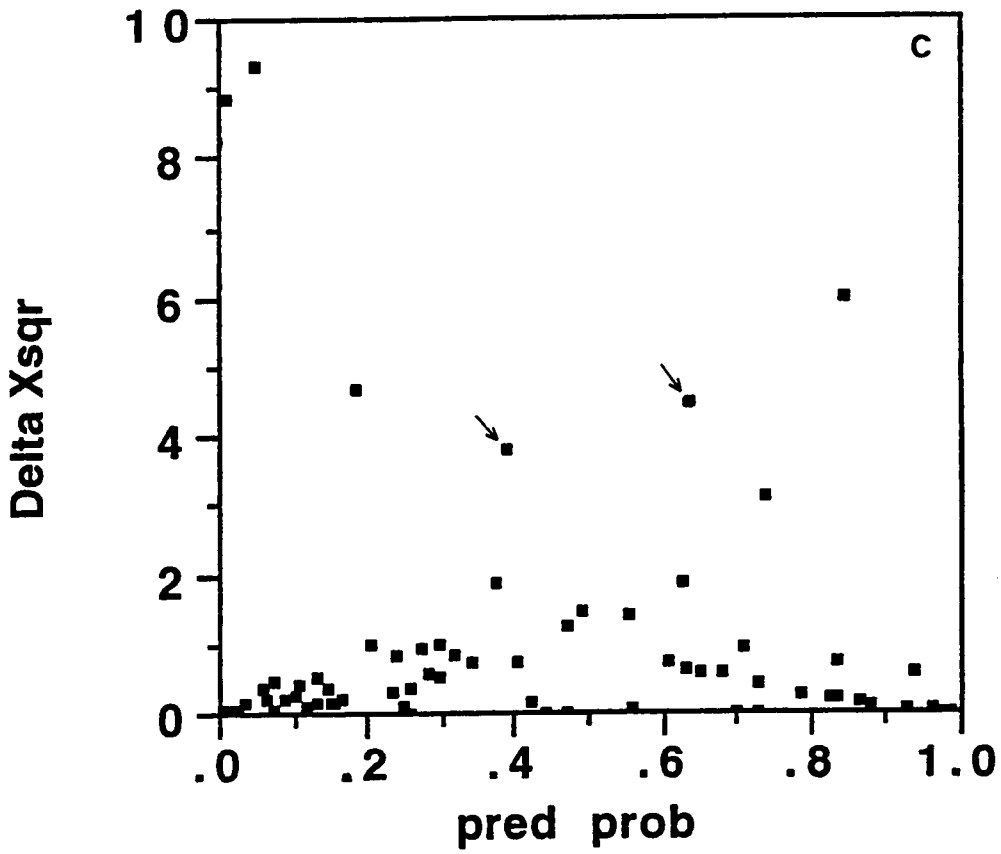

d
Delta Dev
8
6
4
2
0
.0
.2
.4
.6
.8
1.0
pred prob

Fig. 5-10. Winter Model 144 (k = 10, ICOMP-IFIM = 147.12, model variance = 1.38): graphical presentation of logistic regression diagnostic measures. (a) Plot of leverage (h_j) versus predicted probability. In general, h_j is expected to be: small when the predicted (estimated) probability is between 0-0.1 and 0.9-1.0, moderate to small when predicted probability is between 0.3-0.7, and large when predicted probability is between 0.1-0.3 and 0.7-0.9 (Hosmer and Lemeshow 1989:157). (b) Plot of $\Delta\beta_j$ versus predicted probability. $\Delta\beta_j$ measures the change in the estimated parameter values of the logistic regression model, in general, when a particular covariate pattern is deleted from the model. Note that only a few points stand out or fall away from the others (in this case those ≥ 0.6). (c) Plot of ΔX^2_j versus predicted probability. ΔX^2_j measures the change in the Pearson chi-square statistic, a summary measure of the goodness-of-fit of a model, when a particular covariate pattern is removed from the model. (d) Plot of ΔD_j versus predicted probability. ΔD_j measures the change in the deviance, a summary measure of the goodness-of-fit of a model, when a particular covariate pattern is removed from the model. Model 144 fits the winter data fairly well across the covariate patterns because only about 10% of the covariates had either ΔX^2_j or ΔD_j greater than the conservative cutoff of 2.71. Overall, the model fits very well as suggested by the fact that only two points, those indicated by the arrows, had moderate to high values for leverage and poor fit (large values of h_j, $\Delta\beta_j$, and either ΔX^2_j or ΔD_j).


0.5
0.4
0.3
0.2
0.1
0.0
h
a
.0
.2
.4
.6
.8
1.0
pred prob


b
Delta Beta
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
.0
.2
.4
.6
.8
1.0
pred prob


10
8
6
4
2
0
Delta Xsqr
c
.0
.2
.4
.6
.8
1.0
pred prob

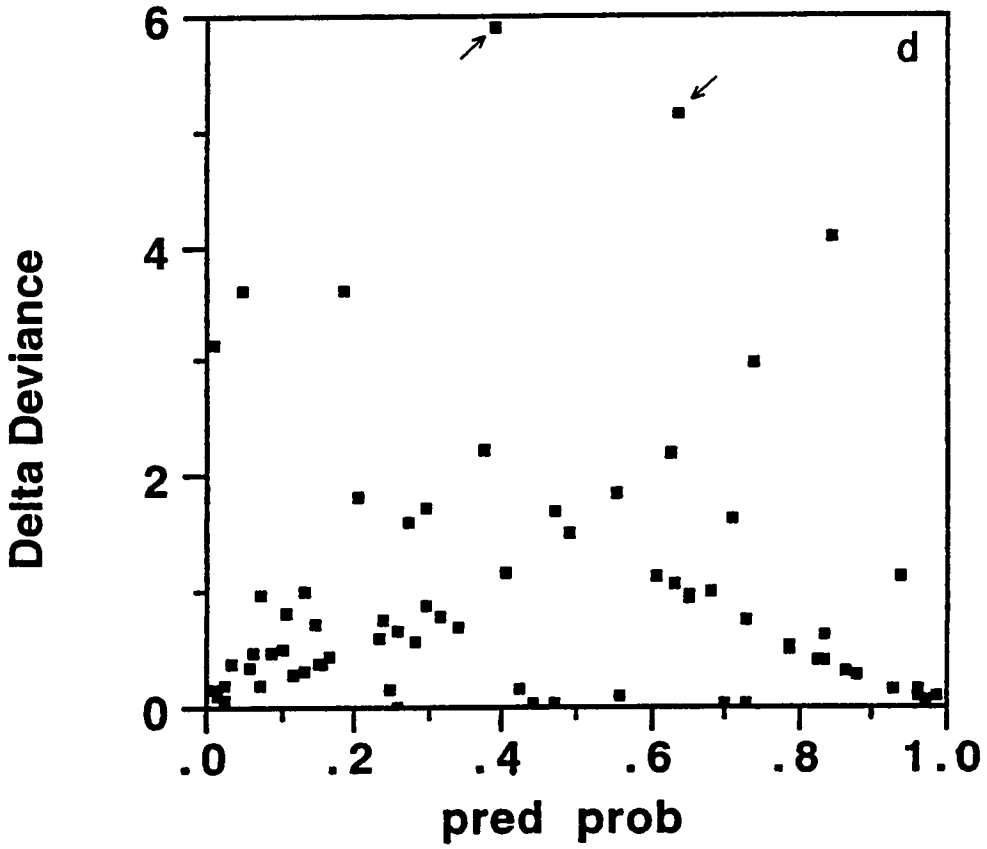

6
4
2
0
d
Delta Deviance
.0
.2
.4
.6
.8
1.0
pred prob

Table 5-10. The 2x2 contingency table for the summer survey data used to examine the possible association between *Anolis carolinensis* and other lizard species. Only data from the field plots which were surveyed for at least 19 observer minutes were used in the analysis. Presence (P) for *A. carolinensis* (AC) was determined as the observation of at least one adult individual inside the plot, whereas Absence (A) was defined as the lack of observation of any adult inside the plot. Presence (P) and Absence (A) of lizard species other than *A. carolinensis* (OTHER) were determined as the observation of at least one individual (adult or juvenile) of any non-anoline lizard and the lack of observation of any non-anoline lizard either inside the plot or within 2 m of the plot edge, repectively. Numbers in parentheses represent the "expected" number for each cell given the observed counts in the table.

		OTHER	
		P	A
AC	P	3 (1.84)	17 (18.16)
	A	10 (11.16)	111 (109.84)

PART 6 : SUMMARY

SUMMARY OF PREVIOUS PARTS

The research reported in this dissertation examined possible associations between habitat features and the occurrence of *Anolis carolinensis* in four habitats along the Little Tennessee River. Because observational (non-experimental) data were used to examine such possible associations, a new approach to exploratory statistical modeling, the genetic algorithm-informational modeling (GAIM) approach, was developed and applied to these data.

Habitat defines the biotic and abiotic parameters and sets the limits within which an organism must live, grow, and reproduce. Thus, understanding the relationships and interactions between an organism and its habitat is fundamental to gaining insight into the behaviors and physiological performances of individuals and populations, the life history traits, population ecology, and evolution of populations and species, and community structure and dynamics. Because of the scientific importance of habitat to an organism's biology, as well as public and legislative concerns over loss of habitats and species, much research has focused on relationships and interactions between an organism and its habitat.

Many ecological studies of animal-habitat relationships are based on observational multivariate data. Ecologists often use stepwise algorithms, hypothesis-testing procedures and multivariate techniques (which also include both multiple linear and multiple logistic regression) to analyze such data sets. Typically, ecologists run the data through a stepwise algorithm, using hypothesis-testing procedures to evaluate the merits of adding or removing a variable to a given model, in order to find a single

"best" model that supposedly fits the data better than any other models. In addition, analysts often use this single "best" model to draw inferences about causation and/or make quantitative predictions. Some of the problems inherent to using stepwise algorithms and hypothesis-testing procedures in such a manner to analyze observational multivariate data are summarized below.

1. Stepwise algorithms are not guaranteed to find the single best model because they search and evaluate only a very small number of the total possible models.

2. For most observational multivariate data sets, it is unlikely that a single model fits the data exceedingly better than all other models. Thus, it is often inappropriate for analysts to interpret the results as if the best model was found.

3. The problem of model selection is really not one of testing hypotheses, statistical or otherwise, but one of evaluating and comparing competing models.

4. Selection of a single supposedly "best" model provides a limited and narrow scientific view of the data.

5. Observational data, because of their non-experimental nature, are not well suited for being the basis on which to make predictions and draw causal inferences. Statisticians and certain ecologists have cautioned against the dangers of such "over-interpretation" of observational data.

The GAIM approach was developed in this dissertation research as a means to potentially avoid such problems associated with stepwise algorithms, hypothesis-testing procedures, and over-interpretation of observational data.

The GAIM approach combines the utility of an informational model-selection criterion with the searching power of a genetic algorithm (GA).

The informational approach to statistical analysis, which uses informational criteria for model selection, is an alternative to hypothesis-testing procedures and has been recently used in some ecological research. The following are important points about this statistical approach, particularly with respect to the use of informational criteria in the GAIM methodology:

1. The informational approach views statistical analysis not as statistical hypothesis-testing, but as a process of model evaluation and selection whereby selection is based on the numerical values of an informational criterion.

2. An informational criterion is used to evaluate each model's fit to the data and to provide a method of ranking and comparing models relative to one another.

3. The models with the lowest numerical values of the criterion are the models which best fit the data.

4. An informational criterion has two components, a lack-of-fit term and a penalty term.

5. The lack-of-fit term, calculated as -2(loglikelihood) using maximum likelihood estimation procedures, measures how poorly the given model fits the data.

6. The penalty term can be a multiple of the number of parameters estimated in the model or a measure of the complexity of the model's covariance or correlational structure among the parameters.

7. The penalty term provides a way to balance problems of over- or underfitting the data and to adhere to the Principle of Parsimony.

An informational criterion can statistically rank and compare any models for a given data set, but how can an analyst handle the logistics of comparing many models when thousands or more potential models (i.e., different combinations of variables) exist and stepwise algorithms evaluate

only a very limited number of models? A GA can be used to search such a vast model space for models that fit well to the data. A GA is a computer algorithm based loosely on genetic and evolutionary concepts. GAs have been applied in many disciplines and with great success to a wide variety of searching and problem solving cases.

In the GAIM approach, a GA can search for models that fit well to a given multivariate data set when thousands or more statistical models exist. An informational model-selection criterion (such as AIC or ICOMP-IFIM) is used by the GA to statistically rank and compare the various models generated by the GA's searching methods. Hypothesis-testing procedures would be of little value in a GA because they cannot, unlike informational criteria, rank and compare any two models unless one model is a subset of the other model.

The GAIM approach allows the analyst to examine the frequency of variables that appear in a *set* of well-fitting models obtained from the GA instead of searching for a single "best" model. How well the models fit the data is determined initially by the values of the informational criterion for the models found by the GA. Analysts can give subsequent consideration to other appropriate statistical and biological information about the models in order to refine the set of models found by the GA. Thus, use of the GAIM approach can allow the analyst to take a wider and richer view of the data than could be obtained from a single well-fitting model (or even a few models) found by stepwise algorithms and hypothesis-testing procedures. The results from the GAIM analysis could be used to propose sets of variables or features that analysts should consider in future observational

studies and/or experiments. This seems to be a more appropriate way to view the analysis of observational multivariate data, given the statistical and interpretational limitations of such data.

The GAIM approach was applied to the exploratory analysis of possible associations between 18 habitat features and the presence of *Anolis carolinensis* among four habitats in eastern Tennessee. Very little ecological information, based on field studies, exists on *A. carolinensis*, especially at the northern limits of this anole's range. Thus, this study represents fundamental exploratory research that can potentially provide some direction for future ecological research on this species.

The application of the GAIM approach to the summer data showed that the most frequent variables in the final set of models were (including the intercept): distance to potential overwintering rock, summer canopy categorization, distance to habitat edge, herb/shrub/vine cover, summer sunlight index, ambient temperature, and standardized distance along the habitat edge from the west boundary of habitat. These variables were also the ones which most frequently possessed statistically significant parameter estimates. The summer results suggest that further research on *A. carolinensis* might focus on a) sunlight/canopy and thermal factors and b) habitat features (such as habitat edges, canopy gaps, and overwintering rocks) related to certain spatial scales beyond the summer home range scale.

For the winter data, the most frequent variables in the final set of models were (including the intercept): ambient temperature, presence of live overstory evergreen tree trunks, presence of overwintering rock,

standardized distance along the habitat edge from the west boundary of habitat, distance to potential overwintering rock, and canopy cover categorization. These variables were also the ones which most frequently possessed statistically significant parameter estimates. Future research might examine ecological, physiological, and biophysical responses of this species to winter habitat features such as a) shelter and potential basking sites, b) sunlight availability and temperature, and c) spatial features beyond the typical winter home range size.

Anolis carolinensis occurs in a diversity of potential natural vegetation types, ecoregions, and habitats across its geographic range in the southeastern and southcentral United States. The research presented in this dissertation examined *A. carolinensis*-habitat associations in only one ecoregion and only a few habitats out of all of those in which this species occurs. Detailed studies of the habitat ecology of *A. carolinensis* are lacking for many habitats and ecoregions. One ecological study alone does not constitute the "definitive work" on either a population or a species. The nature of science and of statistical data analysis and interpretation requires that patterns uncovered or conclusions drawn from a data set must be verified by additional studies of the same kind.

The research presented in this dissertation is an exploratory study, not a confirmatory one. Thus, it should be a starting point for further habitat studies of *Anolis carolinensis* both in Tennessee and other parts of its range. Investigations are much needed into the many aspects of the ecology of *A. carolinensis* that this dissertation research either does not or could not address. Methods using experimental control, or at least partial

control, over field variables are needed to determine the specific responses of this species to key habitat features and the causal mechanisms underlying those responses. In addition, much could be gained from studies that take approaches based on biophysical and physiological ecology, especially if they can be linked to reproductive output, population ecology, and habitat use on local and regional scales. If biologists are to understand the habitat ecology, population biology, distribution and biogeography, and evolution of this species, then such ecological studies, including basic habitat modeling, will be required.

VITA

James J. Minesky was born in Pennsylvania into a family whose three previous generations were primarily factory and steel workers. He attended first through eighth grades at Saint Gertrude's Grade School (Vandergrift, Pennsylvania). He then attended the public school system in Vandergrift and graduated from Kiski Area Senior High School.

Having spent much of his childhood and teenage years enjoying the outdoors with family, finding and watching animals, and living on a hill above a river polluted from mine acid drainage and sewage and industrial effluents, Jim developed an interest in environmental issues and biology. He entered The Pennsylvania State University (New Kensington Campus) and, with little surprise, enjoyed studying biology and ecology. Jim graduated from Penn State (University Park, Pennsylvania) with a Bachelor of Science degree in biology.

After graduating from Penn State, he attended graduate school at the University of Maryland (College Park, Maryland) for one academic year. He then transferred to the Biology Department at Saint Louis University (St. Louis, Missouri), where he later received a Master of Science degree in biology. During graduate studies, he developed an interest in statistics.

While-and after completing his master's degree, Jim worked as a medical research technician, assisting Dr. H. Brent Clark with studies of central nervous system development, at both the Department of Pathology, Washington University School of Medicine (St. Louis, Missouri) and Memorial Hospital (Springfield, Illinois).

After working those three-and-one half years in medical research, he entered the doctoral program in the Department of Zoology at the University of Tennessee, Knoxville (UTK), specifically to study lizard habitat ecology with Dr. Arthur (Sandy) Echternacht. During the latter part of his doctoral program, he worked as full-time instructor at UTK, first in the General Biology Program (Division of Biology) and then in the Department of Botany. The doctoral degree in Ecology and Evolutionary Biology was awarded in May 1999.

Jim is presently working as a program coordinator for a Pennsylvania-based, non-profit environmental organization.