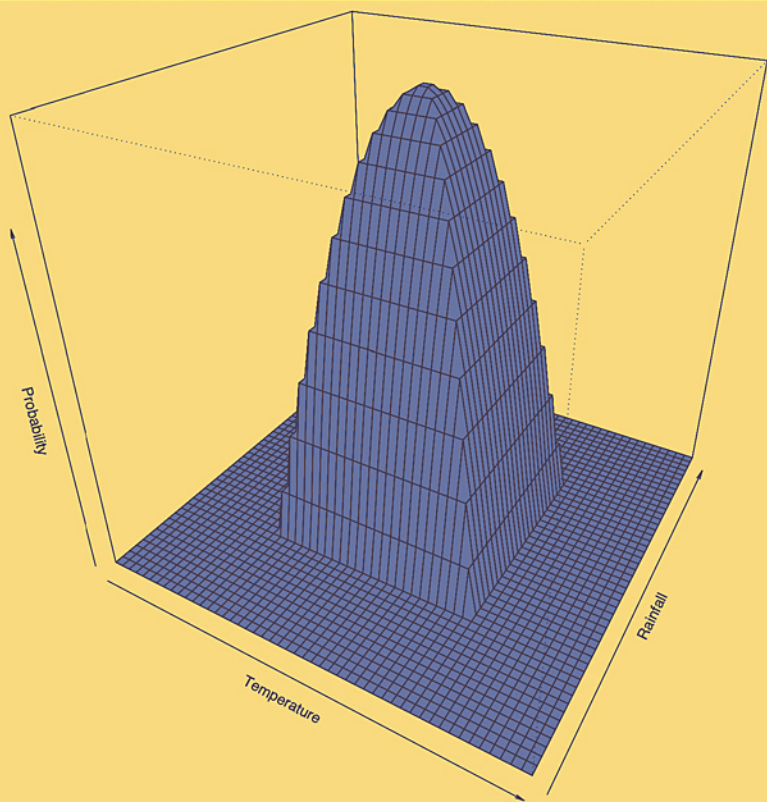


Chapman & Hall/CRC
Mathematical and Computational Biology Series

Niche Modeling

Predictions from Statistical Distributions



David Stockwell

 Chapman & Hall/CRC
Taylor & Francis Group

Chapman & Hall/CRC Mathematical and Computational Biology Series

Niche Modeling

Predictions from Statistical
Distributions

CHAPMAN & HALL/CRC

Mathematical and Computational Biology Series

Aims and scope:

This series aims to capture new developments and summarize what is known over the whole spectrum of mathematical and computational biology and medicine. It seeks to encourage the integration of mathematical, statistical and computational methods into biology by publishing a broad range of textbooks, reference works and handbooks. The titles included in the series are meant to appeal to students, researchers and professionals in the mathematical, statistical and computational sciences, fundamental biology and bioengineering, as well as interdisciplinary researchers involved in the field. The inclusion of concrete examples and applications, and programming techniques and examples, is highly encouraged.

Series Editors

Alison M. Etheridge
Department of Statistics
University of Oxford

Louis J. Gross
Department of Ecology and Evolutionary Biology
University of Tennessee

Suzanne Lenhart
Department of Mathematics
University of Tennessee

Philip K. Maini
Mathematical Institute
University of Oxford

Shoba Ranganathan
Research Institute of Biotechnology
Macquarie University

Hershel M. Safer
Weizmann Institute of Science
Bioinformatics & Bio Computing

Eberhard O. Voit
The Wallace H. Couter Department of Biomedical Engineering
Georgia Tech and Emory University

Proposals for the series should be submitted to one of the series editors above or directly to:

CRC Press, Taylor & Francis Group

24-25 Blades Court
Deodar Road
London SW15 2NU
UK

Published Titles

Cancer Modelling and Simulation

Luigi Preziosi

Computational Biology: A Statistical Mechanics Perspective

Ralf Blossey

Computational Neuroscience: A Comprehensive Approach

Jianfeng Feng

Data Analysis Tools for DNA Microarrays

Sorin Draghici

Differential Equations and Mathematical Biology

D.S. Jones and B.D. Sleeman

Exactly Solvable Models of Biological Invasion

Sergei V. Petrovskii and Bai-Lian Li

Introduction to Bioinformatics

Anna Tramontano

An Introduction to Systems Biology: Design Principles of Biological Circuits

Uri Alon

Knowledge Discovery in Proteomics

Igor Jurisica and Dennis Wigle

Modeling and Simulation of Capsules and Biological Cells

C. Pozrikidis

Niche Modeling: Predictions from Statistical Distributions

David Stockwell

Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems

Qiang Cui and Ivet Bahar

Stochastic Modelling for Systems Biology

Darren J. Wilkinson

The Ten Most Wanted Solutions in Protein Bioinformatics

Anna Tramontano

Chapman & Hall/CRC Mathematical and Computational Biology Series

Niche Modeling

Predictions from Statistical
Distributions

David Stockwell



Chapman & Hall/CRC

Taylor & Francis Group

Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an informa business

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2007 by Taylor & Francis Group, LLC
Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-58488-494-0 (Hardcover)
International Standard Book Number-13: 978-1-58488-494-1 (Hardcover)

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC) 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Stockwell, David R. B. (David Russell Bancroft)
Ecological niche modeling : ecoinformatics in application to biodiversity /
David R.B. Stockwell.
p. cm. -- (Mathematical and computational biology series)
Includes bibliographical references.
ISBN-13: 978-1-58488-494-1 (alk. paper)
ISBN-10: 1-58488-494-0 (alk. paper)
1. Niche (Ecology)--Mathematical models. 2. Niche (Ecology)--Computer
simulation. I. Title. II. Series.

QH546.3.S76 2006
577.8'2--dc22

2006027353

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

Contents

0.1	Preface	xix
0.1.1	Summary of chapters	xix
1	Functions	1
1.1	Elements	1
1.1.1	Factor	1
1.1.2	Complex	2
1.1.3	Raw	2
1.1.4	Vectors	2
1.1.5	Lists	3
1.1.6	Data frames	3
1.1.7	Time series	3
1.1.8	Matrix	4
1.2	Operations	4
1.3	Functions	6
1.4	Ecological models	9
1.4.1	Preferences	11
1.4.2	Stochastic functions	11
1.4.3	Random fields	18
1.5	Summary	21
2	Data	23
2.1	Creating	24
2.2	Entering data	25
2.3	Queries	26
2.4	Joins	28
2.5	Loading and saving a database	29
2.6	Summary	29
3	Spatial	31
3.1	Data types	31
3.2	Operations	34
3.2.1	Rasterizing	37
3.2.2	Overlay	37
3.2.3	Proximity	39
3.2.4	Cropping	40
3.2.5	Palette swapping	40
3.3	Summary	44

4	Topology	45
4.1	Formalism	45
4.2	Topology	47
4.3	Hutchinsonian niche	47
4.3.1	Species space	48
4.3.2	Environmental space	48
4.3.3	Topological generalizations	49
4.3.4	Geographic space	49
4.3.5	Relationships	50
4.4	Environmental envelope	51
4.4.1	Relevant variables	51
4.4.2	Tails of the distribution	51
4.4.3	Independence	52
4.5	Probability distribution	52
4.5.1	Dynamics	53
4.5.2	Generalized linear models	54
4.6	Machine learning methods	57
4.7	Data mining	58
4.7.1	Decision trees	59
4.7.2	Clustering	59
4.7.3	Comparison	59
4.8	Post-Hutchinsonian niche	60
4.8.1	Product space	61
4.9	Summary	63
5	Environmental data collections	65
5.1	Datasets	66
5.1.1	Global ecosystems database	88
5.1.2	Worldclim	89
5.1.3	World ocean atlas	90
5.1.4	Continuous fields	90
5.1.5	Hydro1km	91
5.1.6	WhyWhere	91
5.2	Archives	91
5.2.1	Traffic	92
5.2.2	Management	92
5.2.3	Interaction	92
5.2.4	Updating	92
5.2.5	Legacy	92
5.2.6	Example: WhyWhere archive	93
5.2.7	Browsing	93
5.2.8	Format	94
5.2.9	Meta data	94
5.2.10	Operations	95
5.3	Summary	95

6	Examples	97
6.0.1	Model skill	97
6.0.2	Calculating accuracy	99
6.1	Predicting house prices	99
6.1.1	Analysis	100
6.1.2	P data and no mask	104
6.1.3	Presence and absence (PA) data	105
6.1.4	Interpretation	106
6.2	Brown Treesnake	107
6.2.1	Predictive model	107
6.3	Invasion of Zebra Mussel	109
6.4	Observations	113
7	Bias	115
7.1	Range shift	116
7.1.1	Example: climate change	116
7.2	Range-shift Model	117
7.3	Forms of bias	120
7.3.1	Width r and width error	120
7.3.2	Shift s and shift error	123
7.3.3	Proportional p_e	123
7.4	Quantifying bias	123
7.5	Summary	125
8	Autocorrelation	127
8.1	Types	128
8.1.1	Independent identically distributed (IID)	128
8.1.2	Moving average models (MA)	128
8.1.3	Autoregressive models (AR)	129
8.1.4	Self-similar series (SSS)	129
8.2	Characteristics	130
8.2.1	Autocorrelation Function (ACF)	130
8.2.2	The problems of autocorrelation	136
8.3	Example: Testing statistical skill	137
8.4	Within range	139
8.4.1	Beyond range	139
8.5	Generalization to 2D	140
8.6	Summary	141
9	Non-linearity	143
9.1	Growth niches	144
9.1.1	Linear	145
9.1.2	Sigmoidal	145
9.1.3	Quadratic	147
9.1.4	Cubic	154

9.2	Summary	155
10	Long term persistence	157
10.1	Detecting LTP	159
10.1.1	Hurst Exponent	162
10.1.2	Partial ACF	163
10.2	Implications of LTP	166
10.3	Discussion	171
11	Circularity	173
11.1	Climate prediction	173
11.1.1	Experiments	174
11.2	Lessons for niche modeling	177
12	Fraud	179
12.1	Methods	181
12.1.1	Random numbers	181
12.1.2	CRU	184
12.1.3	Tree rings	186
12.1.4	Tidal Gauge	186
12.1.5	Tidal gauge - hand recorded	188
12.2	Summary	190
	References	191
	Index	199

List of Tables

1.1	R contains a spreadsheet-like data editor called with the <i>edit</i> command.	4
1.2	Some basic types in the R language.	5
1.3	Some basic operations in the R language.	5
1.4	Some useful built-in functions.	9
2.1	Example data consisting of field observations with locations. .	24
4.1	Links between geographic, environmental and species spaces.	46
6.1	Example of a contingency table with a perfect score.	98
6.2	Accuracies for each single variable of Legates Willmott Annual temperatures and precipitation data.	102
8.1	Both r^2 and RE statistics erroneously indicate skill of the random model on in-range data.	139
8.2	While r^2 indicates skill on the smoothed model on out-of-range data, RE indicates little skill for the random model.	140
9.1	Global temperatures and temperature reconstructions.	145
9.2	Slope and correlation coefficient of temperature reconstructions with temperature.	147
10.1	Estimates of Hurst exponent for all series.	162
11.1	Correlations of random model resulting from out of range validation of different experiments.	176
12.1	Statistics from digit frequency of random and fabricated data: df - management up or down, z score, chi-square value, and distance of distribution from expected.	184
12.2	Indices of first and second digit frequency in CRU global temperatures.	186
12.3	First and second digit indices for tree ring data.	186
12.4	Digit frequency of tidal data by hand and by instrument. . .	188

List of Figures

1.1	The bitwise OR combination of two images, A representing longitude and B a mask to give C representing longitude in a masked area.	7
1.2	Basic functions used in modeling: linear, exponential or power relationships.	10
1.3	Basic functions used to represent niche model preference relationships: a step function, a truncated quadratic, exponential and a ramp.	12
1.4	Cyclical functions are common responses to environmental cycles, both singly and added together to produce more complex patterns.	13
1.5	A series with IID errors. Below, ACF plot showing autocorrelation of the IID series at a range of lags.	15
1.6	A moving average of an IID series. Below, the ACF shows oscillation of the autocorrelation of the MA at increasing lags.	16
1.7	A random walk from the cumulative sum of an IID series. Below, the ACF plot shows high autocorrelation at long lags.	17
1.8	Lag plots of periodic, random, moving average and random walk series.	18
1.9	An IID random variable in two dimensions.	19
1.10	An example of a Gaussian field, a two dimensional stochastic variable with autocorrelation.	20
1.11	The ACF of 2D Gaussian field random variable, treated as a 1D vector.	20
3.1	Example of a simple raster to use for testing algorithms.	32
3.2	Example of a raster from an image file representing the average annual temperature in the continental USA.	33
3.3	Examples of vector data, a circle and points of various sizes.	35
3.4	A contour plot generated from the annual temperature raster map.	36
3.5	Simulated image with distribution of values shown in a histogram.	37
3.6	Application of an overlay by multiplication of vectors. The resulting distribution of values is shown in a histogram.	38

3.7	Smoothing of simulated image, first in the x direction, then in the y direction.	39
3.8	A hypothetical niche model of preference for crime given environmental values.	42
3.9	The hypothetical prediction of probability of crime, after palette swapping.	43
4.1	The logistic function transforms values of y from $-\infty$ to ∞ to the range $[0, 1]$ and so can be used to represent linear response as a probability.	55
5.1	The components and operation of the WhyWhere SRB data archive for ecological niche modeling.	93
6.1	Predicted price increases >20% using altitude 2.5 minute variable selected by WhyWhere from the dataset of 528 All Terrestrial variables.	101
6.2	Predicted price increases greater than 20% using annual climate averages and presence only data	102
6.3	Frequency of P and B environmental values for precipitation. The histogram of the proportion of grid cells in the precipitation variable in the locations where metro areas with appreciation greater than 20% (solid line showing presence or P points) and the proportion of values of precipitation for the entire area (dashed line showing background B).	103
6.4	Predicted price increases of less than 10% with locations as black squares.	103
6.5	Frequency of environmental variables predicting house price increases <10%. Note in this case the response if the P point (solid lines) is unimodal.	104
6.6	The distribution of the Brown Treesnake predicted from March precipitation by WhyWhere. Black is zero or low suitability, dark grey is medium and light grey is highly suitable environment.	108
6.7	The histogram of the response of the Brown Treesnake (y axis) to classes of March precipitation (x axis). Dashed bars represent the frequency of the precipitation class in the environment, while solid bars represent the frequency of the BTS occurrences in that precipitation class.	109
6.8	An effective protocol for predicting the potential distribution of invasive species is to develop a model on the home range of a species then predict the distribution using the same environmental variables in the area of interest.	110

6.9 A simple approach to simulating the spread of an invasive species is to develop a series of predictions by moving a cut value from the peak of the probability distribution to the base. 111

6.10 The nested sequence of predicted ranges, based on movement of the cut value. 112

6.11 Evaluation of the accuracy of the prediction of invasion trajectory, with time before present on the x axis and value of cut probability on y axis. Observations above the diagonal are correct predictions, while observations below the diagonal are incorrect predictions. 113

7.1 Theoretical model of shift in species distribution from change in climate. Dashed circle marked O is old range, solid circle marked N is new range and I is intersection area. 118

7.2 The change in the areas of intersection of a square and circle for different shifts (s) and widths (r). 119

7.3 Combined effect of shift and width error. 121

7.4 Combined effect of shift and shift error 122

7.5 Combined effect of shift, shift error, width error and proportional error. 124

8.1 Plots of the global temperatures (CRU), the simulated series random, walk, ar(1), and sss. 131

8.2 Probability distributions for the differenced variables. 132

8.3 Autocorrelation function (ACF) of the simulated series, with decay in correlation plotted as lines. Degree of autocorrelation is readily seen from the rate of decay and compared with temperatures (CRU). 133

8.4 Highly autocorrelated series are more clearly shown when plotting on a log plot. The IID and simple Markov AR1.67 series decline most rapidly. Note also that the autocorrelation of the moving average of CRU temperatures tends to decline more rapidly than the raw CRU series. 134

8.5 Lag plot of the processes CRU, IID, CRU30, AR1.67, walk, and SSS. Autocorrelated series exhibit strong diagonals. 135

8.6 As reconstruction of past temperatures generated by averaging random series that correlate with CRU temperature during the period 1850 to 2000. 138

9.1 Reconstructed smoothed temperatures against proxy values for eight major reconstructions. 146

9.2 Fit of a logistic curve to each of the studies. 148

9.3 Idealized chronology showing tree-rings and the two possible solutions due to non-linear response of the principle (solid and dashed line) after calibration on the end region marked C. . 150

9.4	Nonlinear growth response to a simple sinusoidal driver (e.g. temperature) at three optimal response points (dashed lines).	150
9.5	Nonlinear growth response to two out of phase simple sinusoidal drivers (e.g. temperature and rainfall) at three response points. Solid and dashed lines are climate principles; dotted lines the response of the proxies.	151
9.6	Example of fitting a quadratic model of response to a reconstruction. As response over the given range is fairly linear, reconstruction does not differ greatly.	152
9.7	Reconstruction from a linear model fit to the portion of the graph from 650 to 700.	152
9.8	A linear model fit to years 600 to 800 where the proxies show a significant downturn in growth.	153
9.9	Reconstruction from a quadratic model derived from data years 700 to 800, the period of ideal nonlinear response to the driving variable.	154
9.10	Reconstruction resulting from a quadratic model calibrated from 750 to 850 with two out of phase driving variables, as shown in Figure 9.5	154
10.1	One way of plotting autocorrelation in series: the ACF function at lags 1 to k.	160
10.2	A second way of plotting autocorrelation in series: the ACF at lag 1 of the aggregated processes at time scales 1 to k.	160
10.3	The log-log plot of the standard deviation of the aggregated simulated processes vs. scale k.	161
10.4	Lag 1 ACF of the proxy series at time scales from 1 to 40.	163
10.5	Lag 1 ACF of temperature and precipitation at time 1 to 40 with simulated series for comparison.	164
10.6	Log-log plot of the standard deviation of the aggregated temperature and precipitation processes at scales 1 to 40 with simulated series for comparison.	165
10.7	Plot of the partial correlation coefficient of the simple diagnostic series IID, MA, AR and SSS.	167
10.8	Plot of the partial correlation coefficient of natural series CRU, MBH99, precipitation and temperature.	168
10.9	A: Order of magnitude of the s.d. for FGN model exceeds s.d. for IID model at different H values.	169
10.10	Confidence intervals for the 30 year mean temperature anomaly under IID assumptions (dashed line) and FGN assumptions (dotted lines).	170
11.1	A reconstruction of temperatures generated by summing random series that correlate with temperature.	174

12.1	Expected frequency of digits 1 to 4 predicted by Benford's Law.	180
12.2	Digit frequency of random data.	182
12.3	Digit frequency of fabricated data.	183
12.4	Random data with section of fabricated data inserted in the middle.	183
12.5	The same data above differenced with lag one.	184
12.6	First and second digit frequency of CRU data.	185
12.7	Digit frequency of tree-ring data.	187
12.8	Digit significance of tree-ring series.	187
12.9	Digit frequency of tidal height data, instrument series.	188
12.10	Digit frequency of tidal height data - hand recorded.	189
12.11	Digit significance of hand recorded set along series.	189

0.1 Preface

Niche modeling is a relatively new field of research aimed at helping us to understand the response of species to their environment and predicting their distribution. The practice of niche modeling uses tools from mathematics and statistics, data management and geographic spatial analysis. The first six chapters are concerned with fundamentals, programming, theory and examples of niche modeling. When used in conjunction with more detailed and specific texts and manuals, students and researchers may successfully do niche modeling for the first time.

Successful niche modeling also requires an understanding of the limitations and potential pitfalls of prediction. Due to the importance of avoiding errors, the last six chapters are devoted to sources of errors. All are relatively novel topics in the field: autocorrelation, bias, long term persistence, non-linearity, circularity and fraud, and should be of interest to researchers.

While a statistical language like R or S-plus is not essential, it provides a way of describing these main concepts, showing someone how to use them, and hands on experience at solving problems through examples. It is assumed that readers have a basic knowledge of mathematics and programming.

Above all, successful niche modeling requires deep understanding of the process of creating and using probability distributions in multidimensional spatial and temporal application. Here simplified examples complement the rigor and completeness that can be found in the literature. The generality of the approach is illustrated by examples as diverse as invasive species dynamics, predicting house price increases, and detecting management of data or fraud.

I think there are many advantages in developing depth of intuition, such as capacity to develop novel approaches, and avoiding gross errors. Off-the-shelf statistical packages are tailored exactly to applications but can hide problematic complexity. Recipe book implementations fail to educate users in the details, assumptions and pitfalls of the analysis. As each situation is a little different, packages may not be able to adapt to the specific need of their study. Understanding of the basics, and the pitfalls, also creates confidence for communicating the results.

0.1.1 Summary of chapters

1. Functions This chapter summarizes major mathematical types, operations and relationships encountered both in the book and in niche modeling. This and the following two chapters could be treated as a tutorial in the R language. For example, the main functions for representing the

inverted U shape characteristic of a niche – step, Gaussian, quadratic and ramp functions – are illustrated both graphically and in R code. The chapter concludes with the ACF and lag plots, in one or two dimensions.

2. **Data** This chapter shows a simple biodiversity database using R. By using data frames as tables, it is possible to replicate the basic spreadsheet and relational database operations with R's powerful indexing functions, eliminating conversion problems as data is moved between systems while learning more about R.
3. **Spatial** R and image processing operations can perform many of the elementary spatial operations necessary for niche modeling. While these do not replace a GIS, it demonstrates generalization of arithmetic concepts to images and efficient implementation of simple spatial operations.
4. **Topology** Set theory helps to identify the basic assumptions underlying niche modeling, and the relationships and constraints between these assumptions. The chapter shows the standard definition of the niche as environmental envelopes around all ecologically relevant variables is equivalent to a box topology. A proof is offered that the Hutchinsonian environmental envelope definition of a niche when extended to large or infinite dimensions of environmental variables loses desirable topological properties. This argues for the necessity of careful selection of a small set of environmental variables.
5. **Environmental data collections** Management of data for niche modeling is poorly served by user-developed files stored in a local directory. A wide variety of data sets are currently available, and better quality niche modeling will result from using data in true archives – shared by many studies and trusted with the highest level of quality. A number of sources of data are described and access issues discussed.
6. **Examples** The three examples of niche models here were selected to contradict three main misconceptions of niche modeling. The house price increase example shows a niche that is bimodal and not an inverted U. The second example of the Brown Treesnake shows an asymptotic response with respect to precipitation. The third example of the zebra mussel shows how dynamic models of the spread of invasive species can be developed from the niche model, contrary to the view that niche models are restricted to equilibrium approaches.
7. **Bias** Here a simple theoretical model of range-shift is used to estimate the magnitude of potential bias in estimates of changes in range area due to climate change.
8. **Autocorrelation** This chapter shows the problem of validating models on autocorrelated data using internal or external validation. Holding

back data at random is shown to be inadequate to determine the skill of a model when the data are autocorrelated, particularly when using smoothed data.

- 9. Nonlinearity** Procedures with linear assumptions are not reliable when the responses are non-linear. Here using simulations and a linear model for reconstructing past temperatures, niche model-like tree responses create artifacts including signal degradation, loss of variance, temporal shifts in peaks, and period doubling.
- 10. Long Term Persistence** The natural world is more uncertain and more indeterministic than modeled using classical statistics. Here we show evidence that temporal and spatial natural series display LTP, or scale invariant distributions. These results provide no justification for models with preferred spatial or temporal scale, which greatly underestimate confidence limits.
- 11. Circularity** A major source of error is due to conclusions encoded into the assumptions of the methodology, so allowing no other conclusion than the one obtained. Here we show a potential approach to the problem of quantifying circular reasoning. By feeding random data with the same noise and autocorrelation properties into a methodology, one obtains a null model with benchmarks for rejection regions, and expectations incorporating hidden model assumptions.
- 12. Fraud** The accidental or fraudulent management of results can be detected using the distributional modeling methods of niche modeling. The second digit distribution postulated by Benford's Law allows detection of fabricated data in natural time series drawn from a single distribution. The approach is applied to a range of natural data.

I would like to express my thanks to providers of data used to illustrate issues in niche modeling. The Brown Treesnake point data were from a listing of the Australian Museum holdings provided by Gordon Rodda. Zebra Mussel occurrence data were provided by Amy J. Benson. Temperature reconstruction data were provided by Steve McIntyre. Thank you also to the San Diego Supercomputer Center, University of California San Diego, and to the National Center for Ecological Analysis and Synthesis, University of California Santa Barbara, for providing financial support and office space, funded under a sabbatical research program by the United States National Science Foundation. The development and refinement of some of the sections of the book were assisted by exchanges via a weblog. Steve McIntyre, Demetris Koutsoyiannis, Martin Ringo, and anonymous correspondent TCO were particularly helpful. I would also like to express my deep appreciation for my wife Siriluck and two children, Lena and Victoria.

