Copyrighted Material

Statistics and Data Visualization in Climate Science with R and Python

SAMUEL S. P. SHEN
San Diego State University

GERALD R. NORTH
Texas A&M University

CAMBRIDGE
UNIVERSITY PRESS

Copyrighted Material

Copyrighted Material

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India
103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108842570

DOI: 10.1017/9781108903578

© Cambridge University Press & Assessment 2023

This publication is in copyright. Subject to statutory exception and to the provisions
of relevant collective licensing agreements, no reproduction of any part may take
place without the written permission of Cambridge University Press & Assessment.

First published 2023

Printed in the United Kingdom by CPI Group Ltd, Croydon CR0 4YY

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Shen, Samuel S., author. | North, Gerald R., author.
Title: Statistics and data visualization in climate science with R and Python / Samuel S. P. Shen, Gerald R. North.
Description: Cambridge ; New York : Cambridge University Press, 2023. | Includes bibliographical references and index.
Identifiers: LCCN 2023029571 | ISBN 9781108842570 (hardback) | ISBN 9781108829465 (paperback) | ISBN 9781108903578 (ebook)
Subjects: LCSH: Climatology – Statistical methods. | Climatology – Data processing. | Information visualization. | R (Computer program language) | Python (Computer program language)
Classification: LCC QC874.5 .S48 2023 | DDC 551.63/3–dc23/eng/20230816
LC record available at https://lccn.loc.gov/2023029571

ISBN 978-1-108-84257-0 Hardback

Cambridge University Press & Assessment has no responsibility for the persistence
or accuracy of URLs for external or third-party internet websites referred to in this
publication and does not guarantee that any content on such websites is, or will remain,
accurate or appropriate.

Copyrighted Material

Copyrighted Material

Contents

Preface page xiii
Acknowledgments xix
How to Use This Book xx

1 **Basics of Climate Data Arrays, Statistics, and Visualization** 1
1.1 Global Temperature Anomalies from 1880 to 2018 1
1.1.1 The NOAAGlobalTemp Dataset 2
1.1.2 Visualize the Data of Global Average Annual Mean Temperature 3
1.1.3 Statistical Indices 7
1.2 Commonly Used Climate Statistical Plots 10
1.2.1 Histogram of a Set of Data 10
1.2.2 Box Plot 11
1.2.3 Q-Q Plot 12
1.2.4 Plot a Linear Trend Line 15
1.3 Read netCDF Data File and Plot Spatial Data Maps 16
1.3.1 Read netCDF Data 16
1.3.2 Plot a Spatial Map of Temperature 18
1.3.3 Panoply Plot of a Spatial Map of Temperature 19
1.4 1D-Space-1D-Time Data and Hovmöller Diagram 20
1.5 4D netCDF File and Its Map Plotting 22
1.6 Paraview, 4DVD, and Other Tools 25
1.6.1 Paraview 25
1.6.2 4DVD 25
1.6.3 Other Online Climate Data Visualization Tools 27
1.6.4 Use ChatGPT as a Study Assistant 27
1.7 Chapter Summary 29
References and Further Reading 30
Exercises 31

2 **Elementary Probability and Statistics** 35
2.1 Random Variables 35
2.1.1 Definition 35
2.1.2 Probabilities of a Random Variable 39
2.1.3 Conditional Probability and Bayes' Theorem 39

Copyrighted Material

Copyrighted Material

2.1.4 Probability of a Dry Spell 40
2.1.5 Generate Random Numbers 41
2.2 PDF and CDF 42
2.2.1 The Dry Spell Example 42
2.2.2 Binomial Distribution 45
2.2.3 Normal Distribution 47
2.2.4 PDF and Histogram 50
2.3 Expected Values, Variances and Higher Moments of an RV 51
2.3.1 Definitions 52
2.3.2 Properties of Expected Values 53
2.4 Joint Distributions of X and Y 53
2.4.1 Joint Distributions and Marginal Distributions 53
2.4.2 Covariance and Correlation 56
2.5 Additional Commonly Used Probabilistic Distributions in Climate Science 56
2.5.1 Poisson Distribution 56
2.5.2 Exponential Distribution 58
2.5.3 Mathematical Expression of Normal Distributions, Mean, and the Central Limit Theorem 59
2.5.4 Chi-Square χ^2 Distribution 62
2.5.5 Lognormal Distribution 64
2.5.6 Gamma Distribution 66
2.5.7 Student's t-Distribution 67
2.6 Chapter Summary 68
References and Further Reading 69
Exercises 70

3 Estimation and Decision-Making 73
3.1 From Data to Estimate 73
3.1.1 Sample Mean and Its Standard Error 73
3.1.2 Confidence Interval for the True Mean 77
3.2 Decision-Making by Statistical Inference 85
3.2.1 Contingency Table for Decision-Making with Uncertainty 86
3.2.2 Steps of Hypothesis Testing 86
3.2.3 Interpretations of Significance Level, Power, and p-Value 90
3.2.4 Hypothesis Testing for the 1997–2016 Mean of the Edmonton January Temperature Anomalies 91
3.3 Effective Sample Size 93
3.4 Test Goodness of Fit 103
3.4.1 The Number of Clear, Partly Cloudy, and Cloudy Days 103
3.4.2 Fit the Monthly Rainfall Data to a Gamma Distribution 104
3.5 Kolmogorov–Smirnov Test Using Cumulative Distributions 107

Copyrighted Material

Copyrighted Material

3.6 Determine the Existence of a Significant Relationship 112

3.6.1 Correlation and t-Test 112

3.6.2 Kendall Tau Test for the Existence of a Relationship 114

3.6.3 Mann–Kendall Test for Trend 115

3.7 Chapter Summary 116

References and Further Reading 118

Exercises 119

4 Regression Models and Methods 121

4.1 Simple Linear Regression 121

4.1.1 Temperature Lapse Rate and an Approximately Linear Model 121

4.1.2 Assumptions and Formula Derivations of the Single Variate Linear Regression 124

4.1.3 Statistics of Slope and Intercept: Distributions, Confidence Intervals, and Inference 136

4.2 Multiple Linear Regression 148

4.2.1 Calculating the Colorado TLR When Taking Location Coordinates into Account 148

4.2.2 Formulas for Estimating Parameters in the Multiple Linear Regression 150

4.3 Nonlinear Fittings Using the Multiple Linear Regression 152

4.3.1 Diagnostics of Linear Regression: An Example of Global Temperature 152

4.3.2 Fit a Third-Order Polynomial 157

4.4 Chapter Summary 160

References and Further Reading 161

Exercises 162

5 Matrices for Climate Data 165

5.1 Matrix Definitions 165

5.2 Fundamental Properties and Basic Operations of Matrices 168

5.3 Some Basic Concepts and Theories of Linear Algebra 173

5.3.1 Linear Equations 173

5.3.2 Linear Transformations 174

5.3.3 Linear Independence 175

5.3.4 Determinants 176

5.3.5 Rank of a Matrix 176

5.4 Eigenvectors and Eigenvalues 179

5.4.1 Definition of Eigenvectors and Eigenvalues 179

5.4.2 Properties of Eigenvectors and Eigenvalues for a Symmetric Matrix 182

5.5 Singular Value Decomposition 185

5.5.1 SVD Formula and a Simple SVD Example 185

Copyrighted Material

Copyrighted Material

5.6 SVD for the Standardized Sea Level Pressure Data of Tahiti and Darwin 191
5.7 Chapter Summary 193
References and Further Reading 195
Exercises 196

6 Covariance Matrices, EOFs, and PCs 200
6.1 From a Space-Time Data Matrix to a Covariance Matrix 200
6.2 Definition of EOFs and PCs 207
6.2.1 Defining EOFs and PCs from the Sample Covariance Matrix 207
6.2.2 Percentage Variance Explained 209
6.2.3 Temporal Covariance Matrix 211
6.3 Climate Field and Its EOFs 213
6.3.1 SVD for a Climate Field 213
6.3.2 Stochastic Climate Field and Covariance Function 214
6.4 Generating Random Fields 215
6.5 Sampling Errors for EOFs 220
6.5.1 Sampling Error of Mean and Variance of a Random Variable 221
6.5.2 Errors of the Sample Eigenvalues 222
6.5.3 North's Rule-of-Thumb: Errors of the Sample Eigenvectors 223
6.5.4 EOF Errors and Mode Mixing: A 1D Example 224
6.5.5 EOF Errors and Mode Mixing: A 2D Example 232
6.5.6 The Original Paper of North's Rule-of-Thumb 236
6.5.7 When There Is Serial Correlation 236
6.6 Chapter Summary 237
References and Further Reading 238
Exercises 239

7 Introduction to Time Series 243
7.1 Examples of Time Series Data 243
7.1.1 The Keeling Curve: Carbon Dioxide Data of Mauna Loa 244
7.1.2 ETS Decomposition of the CO_2 Time Series Data 246
7.1.3 Forecasting the CO_2 Data Time Series 250
7.1.4 Ten Years of Daily Minimum Temperature Data of St. Paul, Minnesota, USA 251
7.2 White Noise 254
7.3 Random Walk 257
7.4 Stochastic Processes and Stationarity 262
7.4.1 Stochastic Processes 262
7.4.2 Stationarity 262
7.4.3 Test for Stationarity 263

Copyrighted Material

Copyrighted Material

7.5 Moving Average Time Series 264
7.6 Autoregressive Process 267
7.6.1 Brownian Motion and Autoregressive Model AR(1) 267
7.6.2 Simulations of AR(1) Time Series 268
7.6.3 Autocovariance of AR(1) Time Series when $X_0 = 0$ 270
7.7 Fit Time Series Models to Data 273
7.7.1 Estimate the MA(1) Model Parameters 274
7.7.2 Estimate the AR(1) Model Parameters 275
7.7.3 Difference Time Series 275
7.7.4 AR(p) Model Estimation Using Yule-Walker Equations 276
7.7.5 ARIMA(p, d, q) Model and Data Fitting by R 276
7.8 Chapter Summary 281
References and Further Reading 283
Exercises 284

8 Spectral Analysis of Time Series 287
8.1 The Sine Oscillation 287
8.2 Discrete Fourier Series and Periodograms 292
8.2.1 Discrete Sine Transform 292
8.2.2 Discrete Fourier Transform 294
8.2.3 Energy Identity 312
8.2.4 Periodogram of White Noise 313
8.3 Fourier Transform in $(-\infty, \infty)$ 314
8.4 Fourier Series for a Continuous Time Series on a Finite Time Interval $[-T/2, T/2]$ 315
8.5 Chapter Summary 318
References and Further Reading 321
Exercises 322

9 Introduction to Machine Learning 329
9.1 K-Means Clustering 329
9.1.1 K-Means Setup and Trivial Examples 330
9.1.2 A K-Means Algorithm 335
9.1.3 K-Means Clustering for the Daily Miami Weather Data 337
9.2 Support Vector Machine 346
9.2.1 Maximize the Difference between Two Labeled Points 347
9.2.2 SVM for a System of Three Points Labeled in Two Categories 350
9.2.3 SVM Mathematical Formulation for a System of Many Points in Two Categories 353
9.3 Random Forest Method for Classification and Regression 359
9.3.1 RF Flower Classification for a Benchmark Iris Dataset 360

Copyrighted Material

Copyrighted Material

9.3.2 RF Regression for the Daily Ozone Data of New York City 366

9.3.3 What Does a Decision Tree Look Like? 370

9.4 Neural Network and Deep Learning 372

9.4.1 An NN Model for an Automized Decision System 373

9.4.2 An NN Prediction of Iris Species 379

9.5 Chapter Summary 381

References and Further Reading 383

Exercises 384

Index 387

Copyrighted Material