

INTRODUCTION TO ENVIRONMENTAL DATA SCIENCE

William W. Hsieh

Statistical and machine learning methods have many applications in the environmental sciences, including prediction and data analysis in meteorology, hydrology and oceanography; pattern recognition for satellite images from remote sensing; management of agriculture and forests; assessment of climate change; and much more. With rapid advances in machine learning in the last decade, this book provides an urgently needed, comprehensive guide to machine learning and statistics for students and researchers interested in environmental data science. It includes intuitive explanations covering the relevant background mathematics, with examples drawn from the environmental sciences. A broad range of topics is covered, including correlation, regression, classification, clustering, neural networks, random forests, boosting, kernel methods, evolutionary algorithms and deep learning, as well as the recent merging of machine learning and physics. End-of-chapter exercises allow readers to develop their problem-solving skills, and online datasets allow readers to practise analysis of real data.

WILLIAM W. HSIEH is a professor emeritus in the Department of Earth, Ocean and Atmospheric Sciences at the University of British Columbia. Known as a pioneer in introducing machine learning to environmental science, he has written more than 100 peer-reviewed journal papers on climate variability, machine learning, atmospheric science, oceanography, hydrology and agricultural science. He is the author of the book *Machine Learning Methods in the Environmental Sciences* (Cambridge University Press, 2009), the first single-authored textbook on machine learning for environmental scientists. Currently retired in Victoria, British Columbia, he enjoys growing organic vegetables.

‘As a new wave of machine learning becomes part of our toolbox for environmental science, this book is both a guide to the latest developments and a comprehensive textbook on statistics and data science. Almost everything is covered, from hypothesis testing to convolutional neural networks. The book is enjoyable to read, well explained and economically written, so it will probably become the first place I’ll go to read up on any of these topics.’

– **Alan Geer**, *European Centre for Medium-Range Weather Forecasts (ECMWF)*

‘There is a need for a forward-looking text on environmental data science and William Hsieh’s text succeeds in filling the gap. This comprehensive text covers basic to advanced material ranging from timeless statistical techniques to some of the latest machine learning approaches. His refreshingly engaging style is written to be understood and is complemented by a plethora of expressive visuals. Hsieh’s treatment of nonlinearity is cutting-edge and the final chapter examines ways to combine machine learning with physics. This text is destined to become a modern classic.’

– **Sue Ellen Haupt**, *National Center for Atmospheric Research*

‘William Hsieh has been one of the “founding fathers” of an exciting new field of using machine learning (ML) in the environmental sciences. His new book provides readers with a solid introduction to the statistical foundation of ML and various ML techniques, as well as with the fundamentals of data science. The unique combination of solid mathematical and statistical backgrounds with modern applications of ML tools in the environmental sciences ... is an important distinguishing feature of this book. The broad range of topics covered in this book makes it an invaluable reference and guide for researchers and graduate students working in this and related fields.’

– **Vladimir Krasnopolsky**, *Center for Weather and Climate Prediction, NOAA*

‘Dr. Hsieh is one of the pioneers of the development of machine learning for the environmental sciences including the development of methods such as nonlinear principal component analysis to provide insights into the ENSO dynamic. Dr. Hsieh has a deep understanding of the foundations of statistics, machine learning, and environmental processes that he is sharing in this timely and comprehensive work with many recent references. It will no doubt become an indispensable reference for our field. I plan to use the book for my graduate environmental forecasting class and recommend the book for a self-guided progression or as a comprehensive reference.’

– **Philippe Tissot**, *Texas A & M University, Corpus Christi*

INTRODUCTION TO
ENVIRONMENTAL DATA SCIENCE

William W. Hsieh

University of British Columbia





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/9781107065550

© William W. Hsieh 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 TJ Books Limited, Padstow Cornwall
A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data

Names: Hsieh, William Wei, 1955– author.

Title: Introduction to environmental data science / William W. Hsieh.

Description: New York : Cambridge University Press, 2023. |

Includes bibliographical references and index.

Identifiers: LCCN 2022054278 | ISBN 9781107065550 (hardback)

Subjects: LCSH: Environmental sciences–Data processing. | Environmental
protection–Data processing. | Environmental management–Data processing. |
Machine learning.

Classification: LCC GE45.D37 H74 2023 | DDC 363.700285–dc23/eng20221219
LC record available at <https://lccn.loc.gov/2022054278>

ISBN 978-1-107-06555-0 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.

Contents

Preface	<i>Page xv</i>
Notation Used	xviii
List of Abbreviations	xix
1 Introduction	1
1.1 Statistics and Machine Learning $\boxed{\text{A}}^{\odot}$	1
1.2 Environmental Data Science $\boxed{\text{A}}^{\odot}$	6
1.3 A Simple Example of Curve Fitting $\boxed{\text{A}}^{\odot}$	9
1.4 Main Types of Data Problems $\boxed{\text{A}}^{\odot}$	12
1.4.1 Supervised Learning $\boxed{\text{A}}^{\odot}$	15
1.4.2 Unsupervised Learning $\boxed{\text{A}}^{\odot}$	16
1.4.3 Reinforced Learning $\boxed{\text{A}}^{\odot}$	17
1.5 Curse of Dimensionality $\boxed{\text{A}}^{\odot}$	17
2 Basics	19
2.1 Data Types $\boxed{\text{A}}^{\odot}$	19
2.2 Probability $\boxed{\text{A}}^{\odot}$	19
2.3 Probability Density $\boxed{\text{A}}^{\odot}$	22
2.4 Expectation and Mean $\boxed{\text{A}}^{\odot}$	24
2.5 Variance and Standard Deviation $\boxed{\text{A}}^{\odot}$	25
2.6 Covariance $\boxed{\text{A}}^{\odot}$	26
2.7 Online Algorithms for Mean, Variance and Covariance $\boxed{\text{C}}^{\odot}$	27
2.8 Median and Median Absolute Deviation $\boxed{\text{A}}^{\odot}$	29
2.9 Quantiles $\boxed{\text{A}}^{\odot}$	30
2.10 Skewness and Kurtosis $\boxed{\text{B}}^{\odot}$	32
2.11 Correlation $\boxed{\text{A}}^{\odot}$	33
2.11.1 Pearson Correlation $\boxed{\text{A}}^{\odot}$	33
2.11.2 Serial Correlation $\boxed{\text{A}}^{\odot}$	36
2.11.3 Spearman Rank Correlation $\boxed{\text{A}}^{\odot}$	38
2.11.4 Kendall Rank Correlation $\boxed{\text{A}}^{\odot}$	39
2.11.5 Biweight Midcorrelation $\boxed{\text{B}}^{\odot}$	40
2.12 Exploratory Data Analysis $\boxed{\text{A}}^{\odot}$	41
2.12.1 Histograms $\boxed{\text{A}}^{\odot}$	41
2.12.2 Quantile–Quantile (Q–Q) Plots $\boxed{\text{B}}^{\odot}$	42

2.12.3	Boxplots [A]⊙	45
2.13	Mahalanobis Distance [A]⊙	47
2.13.1	Mahalanobis Distance and Principal Component Analysis [B]⊙	47
2.14	Bayes Theorem [A]⊙	49
2.15	Classification [A]⊙	52
2.16	Clustering [A]⊙	54
2.17	Information Theory [B]⊙	56
2.17.1	Entropy [B]⊙	58
2.17.2	Joint Entropy and Conditional Entropy [B]⊙	59
2.17.3	Relative Entropy [B]⊙	60
2.17.4	Mutual Information [B]⊙	61
	Exercises	62
3	Probability Distributions	65
3.1	Binomial Distribution [A]⊙	66
3.2	Poisson Distribution [B]⊙	68
3.3	Multinomial Distribution [B]⊙	68
3.4	Gaussian Distribution [A]⊙	69
3.5	Maximum Likelihood Estimation [A]⊙	73
3.6	Multivariate Gaussian Distribution [B]⊙	75
3.7	Conditional and Marginal Gaussian Distributions [C]⊙	77
3.8	Gamma Distribution [B]⊙	78
3.9	Beta Distribution [B]⊙	80
3.10	Von Mises Distribution [C]⊙	82
3.11	Extreme Value Distributions [C]⊙	83
3.12	Gaussian Mixture Model [B]⊙	86
3.12.1	Expectation–Maximization (EM) Algorithm [B]⊙	90
3.13	Kernel Density Estimation [B]⊙	91
3.14	Re-expressing Data [A]⊙	93
3.15	Student <i>t</i> -distribution [B]⊙	95
3.16	Chi-squared Distribution [B]⊙	97
	Exercises	99
4	Statistical Inference	101
4.1	Hypothesis Testing [A]⊙	101
4.2	Student <i>t</i> -test [A]⊙	104
4.2.1	One-Sample <i>t</i> -test [A]⊙	105
4.2.2	Independent Two-Sample <i>t</i> -test [A]⊙	105
4.2.3	Dependent <i>t</i> -test for Paired Samples [A]⊙	107
4.2.4	Serial Correlation [A]⊙	107
4.2.5	Significance Test for Correlation [A]⊙	109
4.3	Non-parametric Alternatives to <i>t</i> -test [B]⊙	111
4.3.1	Wilcoxon–Mann–Whitney Test [C]⊙	111
4.3.2	Wilcoxon Signed-Rank Test [C]⊙	114
4.4	Confidence Interval [A]⊙	115

4.4.1	Confidence Interval for Population Mean	B ☺	116
4.4.2	Confidence Interval for Correlation	B ☺	118
4.5	Goodness-of-Fit Tests	B ☺	119
4.5.1	One-Sample Goodness-of-Fit Tests	B ☺	119
4.5.2	Two-Sample Goodness-of-Fit Tests	B ☺	121
4.6	Test of Variances	B ☺	124
4.7	Mann–Kendall Trend Test	B ☺	125
4.8	Bootstrapping	A ☺	126
4.9	Field Significance	B ☺	131
	Exercises		134
5	Linear Regression		137
5.1	Simple Linear Regression	A ☺	137
5.1.1	Partition of Sums of Squares	A ☺	139
5.1.2	Confidence Interval for Regression Parameters	B ☺	141
5.1.3	Confidence Interval and Prediction Interval for the Response Variable	B ☺	143
5.1.4	Serial Correlation	B ☺	144
5.2	Multiple Linear Regression	A ☺	145
5.2.1	Gauss–Markov Theorem	B ☺	147
5.2.2	Partition of Sums of Squares	B ☺	148
5.2.3	Standardized Predictors	A ☺	148
5.2.4	Analysis of Variance (ANOVA)	B ☺	149
5.2.5	Confidence and Prediction Intervals	B ☺	151
5.3	Multivariate Linear Regression	B ☺	153
5.4	Online Learning with Linear Regression	C ☺	154
5.5	Circular and Categorical Data	A ☺	156
5.6	Predictor Selection	C ☺	158
5.7	Ridge Regression	B ☺	161
5.8	Lasso	C ☺	165
5.9	Quantile Regression	C ☺	167
5.10	Generalized Least Squares	C ☺	168
5.10.1	Optimal Fingerprinting in Climate Change	C ☺	170
	Exercises		170
6	Neural Networks		173
6.1	McCulloch and Pitts Model	B ☺	174
6.2	Perceptrons	A ☺	175
6.2.1	Limitation of Perceptrons	B ☺	178
6.3	Multi-layer Perceptrons	A ☺	180
6.3.1	Comparison with Polynomials	B ☺	185
6.3.2	Hidden Neurons	A ☺	186
6.3.3	Monotonic Multi-layer Perceptron Model	C ☺	187
6.4	Extreme Learning Machines	A ☺	189
6.4.1	Online Learning	B ☺	193
6.4.2	Random Vector Functional Link	C ☺	194

6.5	Radial Basis Functions	\boxed{B}^{\odot}	195
6.6	Modelling Conditional Distributions	\boxed{B}^{\odot}	199
6.6.1	Mixture Density Network	\boxed{B}^{\odot}	201
6.7	Quantile Regression	\boxed{C}^{\odot}	204
6.8	Historical Development of NN in Environmental Science	\boxed{B}^{\odot}	207
6.8.1	Remote Sensing	\boxed{B}^{\odot}	208
6.8.2	Hydrology	\boxed{B}^{\odot}	212
6.8.3	Atmospheric Science	\boxed{B}^{\odot}	213
6.8.4	Oceanography	\boxed{B}^{\odot}	213
	Exercises		214
7	Non-linear Optimization		216
7.1	Extrema and Saddle Points	\boxed{A}^{\odot}	216
7.2	Gradient Vector in Optimization	\boxed{A}^{\odot}	219
7.3	Back-Propagation	\boxed{B}^{\odot}	221
7.4	Training Protocol	\boxed{A}^{\odot}	224
7.5	Gradient Descent Method	\boxed{A}^{\odot}	225
7.6	Stochastic Gradient Descent	\boxed{B}^{\odot}	227
7.7	Conjugate Gradient Method	\boxed{C}^{\odot}	229
7.8	Quasi-Newton Methods	\boxed{C}^{\odot}	232
7.9	Non-linear Least Squares Methods	\boxed{C}^{\odot}	234
7.10	Evolutionary Algorithms	\boxed{B}^{\odot}	236
7.11	Hill Climbing	\boxed{C}^{\odot}	238
7.12	Genetic Algorithm	\boxed{C}^{\odot}	239
7.13	Differential Evolution	\boxed{C}^{\odot}	241
	Exercises		244
8	Learning and Generalization		245
8.1	Mean Squared Error and Maximum Likelihood	\boxed{A}^{\odot}	245
8.2	Objective Functions and Robustness	\boxed{A}^{\odot}	247
8.3	Variance and Bias Errors	\boxed{A}^{\odot}	250
8.4	Regularization	\boxed{A}^{\odot}	251
8.4.1	Weight Penalty	\boxed{A}^{\odot}	251
8.4.2	Early Stopping	\boxed{A}^{\odot}	254
8.5	Cross-Validation	\boxed{A}^{\odot}	255
8.6	Hyperparameter Tuning	\boxed{A}^{\odot}	258
8.7	Ensemble Methods	\boxed{A}^{\odot}	261
8.7.1	Bagging	\boxed{A}^{\odot}	263
8.7.2	Error of Ensemble	\boxed{B}^{\odot}	263
8.7.3	Unequal Weighting of Ensemble Members	\boxed{C}^{\odot}	266
8.7.4	Stacking	\boxed{C}^{\odot}	268
8.8	Dropout	\boxed{B}^{\odot}	269
8.9	Maximum Norm Constraint	\boxed{B}^{\odot}	271
8.10	Bayesian Model Selection	\boxed{B}^{\odot}	272
8.11	Information Criterion	\boxed{B}^{\odot}	273
8.11.1	Bayesian Information Criterion	\boxed{C}^{\odot}	273

8.11.2 Akaike Information Criterion C^{\ominus}	275
8.12 Bayesian Model Averaging C^{\ominus}	278
8.13 Interpretable ML C^{\ominus}	280
Exercises	281
9 Principal Components and Canonical Correlation	283
9.1 Principal Component Analysis (PCA) A^{\ominus}	283
9.1.1 Geometric Approach to PCA A^{\ominus}	284
9.1.2 Eigenvector Approach to PCA A^{\ominus}	284
9.1.3 Real and Complex Data C^{\ominus}	288
9.1.4 Orthogonality Relations A^{\ominus}	289
9.1.5 PCA of the Tropical Pacific Climate Variability A^{\ominus}	290
9.1.6 Scaling the PCs and Eigenvectors A^{\ominus}	298
9.1.7 Degeneracy of Eigenvalues A^{\ominus}	299
9.1.8 A Smaller Covariance Matrix A^{\ominus}	299
9.1.9 How Many Modes to Retain B^{\ominus}	300
9.1.10 Temporal and Spatial Mean Removal B^{\ominus}	302
9.1.11 Singular Value Decomposition B^{\ominus}	302
9.1.12 Missing Data C^{\ominus}	304
9.2 Rotated PCA B^{\ominus}	305
9.2.1 E-frame Rotation B^{\ominus}	308
9.2.2 A-frame Rotation B^{\ominus}	311
9.2.3 Advantages and Disadvantages of Rotation A^{\ominus}	315
9.3 PCA for Two-Dimensional Vectors C^{\ominus}	317
9.4 Canonical Correlation Analysis (CCA) B^{\ominus}	320
9.4.1 CCA Theory B^{\ominus}	320
9.4.2 Pre-filter with PCA B^{\ominus}	324
9.5 Maximum Covariance Analysis B^{\ominus}	326
Exercises	327
10 Unsupervised Learning	330
10.1 Clustering B^{\ominus}	330
10.1.1 Distance Measure B^{\ominus}	331
10.1.2 Model Evaluation B^{\ominus}	332
10.2 Non-hierarchical Clustering B^{\ominus}	337
10.2.1 K -means Clustering B^{\ominus}	337
10.2.2 Nucleated Agglomerative Clustering C^{\ominus}	338
10.2.3 Gaussian Mixture Model C^{\ominus}	339
10.3 Hierarchical Clustering C^{\ominus}	339
10.4 Self-Organizing Map C^{\ominus}	343
10.4.1 Applications of SOM C^{\ominus}	345
10.5 Autoencoder A^{\ominus}	347
10.6 Non-linear Principal Component Analysis B^{\ominus}	349
10.6.1 Overfitting C^{\ominus}	358
10.6.2 Closed Curves C^{\ominus}	360
10.7 Other Non-linear Dimensionality Reduction Methods C^{\ominus}	363

10.8 Non-linear Canonical Correlation Analysis [C]⊙	365
Exercises	369
11 Time Series	372
11.1 Fourier Analysis [A]⊙	372
11.1.1 Fourier Series [A]⊙	373
11.1.2 Discrete Fourier Transform [A]⊙	374
11.1.3 Continuous-Time Fourier Transform [B]⊙	375
11.1.4 Discrete-Time Fourier Transform [B]⊙	376
11.2 Windows [A]⊙	376
11.3 Spectrum [A]⊙	379
11.3.1 Effects of Window Functions [A]⊙	380
11.3.2 Trend Removal [A]⊙	382
11.3.3 Nyquist Frequency and Aliasing [A]⊙	382
11.3.4 Smoothing the Spectrum [A]⊙	384
11.3.5 Confidence Interval [B]⊙	385
11.3.6 Examples [B]⊙	385
11.3.7 Fast Fourier Transform [B]⊙	386
11.3.8 Relation with Auto-covariance [B]⊙	387
11.3.9 Rotary Spectrum for 2-D Vectors [C]⊙	387
11.3.10 Wavelets [C]⊙	390
11.4 Cross-Spectrum [B]⊙	393
11.5 Filtering [A]⊙	395
11.5.1 Periodic Signals [A]⊙	396
11.5.2 Ideal Filters [A]⊙	397
11.5.3 Finite Impulse Response Filters [B]⊙	398
11.6 Averaging [A]⊙	401
11.6.1 Moving Average Filters [A]⊙	402
11.6.2 Grid-Scale Noise [B]⊙	402
11.6.3 Linearization from Time-Averaging [B]⊙	404
11.7 Singular Spectrum Analysis [B]⊙	405
11.7.1 Multichannel Singular Spectrum Analysis [B]⊙	407
11.8 Auto-regressive Process [B]⊙	410
11.8.1 AR(p) Process [B]⊙	411
11.8.2 AR(1) Process [B]⊙	412
11.8.3 AR(2) Process [C]⊙	413
11.9 Box–Jenkins Models [C]⊙	414
11.9.1 Moving Average (MA) Process [C]⊙	414
11.9.2 Auto-regressive Moving Average (ARMA) Model [C]⊙	414
11.9.3 Auto-regressive Integrated Moving Average (ARIMA) Model [C]⊙	415
Exercises	416

12 Classification	418
12.1 Linear Discriminant Analysis $\boxed{\text{A}}^{\ominus}$	419
12.1.1 Fisher Linear Discriminant $\boxed{\text{B}}^{\ominus}$	421
12.2 Logistic Regression $\boxed{\text{A}}^{\ominus}$	424
12.2.1 Multiclass Logistic Regression $\boxed{\text{B}}^{\ominus}$	425
12.3 Naive Bayes Classifier $\boxed{\text{B}}^{\ominus}$	427
12.4 K -nearest Neighbours $\boxed{\text{B}}^{\ominus}$	428
12.5 Extreme Learning Machine Classifier $\boxed{\text{A}}^{\ominus}$	430
12.6 Cross-Entropy $\boxed{\text{A}}^{\ominus}$	432
12.7 Multi-layer Perceptron Classifier $\boxed{\text{A}}^{\ominus}$	434
12.8 Class Imbalance $\boxed{\text{A}}^{\ominus}$	436
Exercises	438
13 Kernel Methods	440
13.1 From Neural Networks to Kernel Methods $\boxed{\text{B}}^{\ominus}$	441
13.2 Primal and Dual Solutions for Linear Regression $\boxed{\text{B}}^{\ominus}$	442
13.3 Kernels $\boxed{\text{B}}^{\ominus}$	444
13.4 Kernel Ridge Regression $\boxed{\text{B}}^{\ominus}$	448
13.5 Advantages and Disadvantages $\boxed{\text{B}}^{\ominus}$	449
13.6 Pre-image Problem $\boxed{\text{C}}^{\ominus}$	451
13.7 Support Vector Machines (SVM) $\boxed{\text{B}}^{\ominus}$	453
13.7.1 Linearly Separable Case $\boxed{\text{B}}^{\ominus}$	454
13.7.2 Linearly Non-separable Case $\boxed{\text{B}}^{\ominus}$	458
13.7.3 Non-linear Classification by SVM $\boxed{\text{B}}^{\ominus}$	460
13.7.4 Multi-class Classification by SVM $\boxed{\text{C}}^{\ominus}$	461
13.7.5 Support Vector Regression $\boxed{\text{C}}^{\ominus}$	462
13.8 Gaussian Processes $\boxed{\text{B}}^{\ominus}$	463
13.8.1 Learning the Hyperparameters $\boxed{\text{B}}^{\ominus}$	466
13.8.2 Other Common Kernels $\boxed{\text{C}}^{\ominus}$	467
13.9 Kernel Principal Component Analysis $\boxed{\text{C}}^{\ominus}$	469
Exercises	471
14 Decision Trees, Random Forests and Boosting	473
14.1 Classification and Regression Trees (CART) $\boxed{\text{A}}^{\ominus}$	474
14.1.1 Relative Importance of Predictors $\boxed{\text{B}}^{\ominus}$	479
14.1.2 Surrogate Splits $\boxed{\text{B}}^{\ominus}$	480
14.2 Random Forests $\boxed{\text{A}}^{\ominus}$	482
14.2.1 Extremely Randomized Trees (Extra Trees) $\boxed{\text{B}}^{\ominus}$	487
14.3 Boosting $\boxed{\text{A}}^{\ominus}$	487
14.3.1 Gradient Boosting $\boxed{\text{B}}^{\ominus}$	488
Exercises	492

15 Deep Learning	494
15.1 Transfer Learning A ☺	498
15.2 Convolutional Neural Network B ☺	499
15.2.1 Convolution Operation B ☺	499
15.2.2 Pooling B ☺	502
15.2.3 AlexNet B ☺	503
15.2.4 Residual Neural Network (ResNet) B ☺	505
15.2.5 Data Augmentation C ☺	506
15.2.6 Applications in Environment Science B ☺	506
15.3 Encoder–Decoder Network B ☺	507
15.3.1 U-net B ☺	508
15.4 Time Series C ☺	510
15.4.1 Long Short-Term Memory (LSTM) Network C ☺	510
15.4.2 Temporal Convolutional Network C ☺	513
15.5 Generative Adversarial Network C ☺	514
Exercises	517
16 Forecast Verification and Post-processing	518
16.1 Binary Classes B ☺	519
16.1.1 Skill Scores for Binary Classes B ☺	524
16.2 Multiple Classes C ☺	527
16.3 Probabilistic Forecasts for Binary Classes B ☺	528
16.3.1 Reliability Diagram B ☺	529
16.4 Probabilistic Forecasts for Multiple Classes B ☺	531
16.5 Continuous Variables B ☺	532
16.5.1 Forecast Scores B ☺	532
16.5.2 Skill Scores B ☺	533
16.6 Probabilistic Forecasts for Continuous Variables B ☺	535
16.7 Minimizing Loss C ☺	535
16.8 Spurious Skills A ☺	536
16.9 Extrapolation A ☺	538
16.10 Post-processing C ☺	541
16.11 Downscaling C ☺	543
16.11.1 Reduced Variance C ☺	546
Exercises	547
17 Merging of Machine Learning and Physics	549
17.1 Physics Emulation and Hybrid Models C ☺	550
17.1.1 Radiation in Atmospheric Models C ☺	550
17.1.2 Clouds C ☺	551
17.1.3 Turbulent Fluxes C ☺	553
17.1.4 Hybrid Coupled Atmosphere–Ocean Modelling C ☺	554
17.1.5 Wind Wave Modelling C ☺	555
17.2 Physics-Informed Machine Learning A ☺	556
17.2.1 Soft Constraint C ☺	556

<i>Contents</i>	xiii
17.2.2 Hard Constraint \square^{\ominus}	558
17.3 Data Assimilation and ML \square^{\ominus}	560
17.3.1 3D-Var \square^{\ominus}	562
17.3.2 4D-Var \square^{\ominus}	563
17.3.3 Neural Networks in 4D-Var \square^{\ominus}	564
Exercises	568
Appendices	569
A Trends in Terminology	569
B Lagrange Multipliers	569
References	573
Index	613

Cambridge University Press & Assessment
978-1-107-06555-0 — Introduction to Environmental Data Science
William W. Hsieh
Frontmatter
[More Information](#)

Preface

Modern data science has two main branches – statistics and machine learning – analogous to physics containing classical mechanics and quantum mechanics. Statistics, the much older branch, grew out from mathematics, while the advent of the computer and computer science in the post–World War II era led to an interest in intelligent machines, henceforth artificial intelligence (AI), and machine learning (ML), the fastest growing branch of AI. As quantum mechanics arrived in the 1920s with a fuzzy, random view of nature, which made many physicists, including Einstein, uncomfortable, the ML models too have been disapprovingly called ‘black boxes’ from their use of a large number of parameters that are opaque in practical problems. Quantum mechanics was eventually accepted, and a modern physicist learns both classical mechanics and quantum mechanics, using the former on everyday problems and the latter on atomic-scale problems. Similarly, a modern data scientist learns both statistics and machine learning, choosing the appropriate statistical or ML tool based on the particular data problem.

Environmental data science is the intersection between environmental science and data science. Environmental science is composed of many parts – atmospheric science, oceanography, hydrology, cryospheric science, ecology, agricultural science, remote sensing, climate science, and so on. Environmental datasets have their unique characteristics, for example most non-environmental datasets used in ML contain discrete or categorical data (alphabets and numbers in texts, colour pixels in an image, etc.), whereas most environmental datasets contain continuous variables (temperature, air pressure, precipitation amount, pollutant concentration, sea level height, streamflow, crop yield, etc.). Hence, environmental scientists need to assess astutely whether data methods developed from non-environmental fields would work well for particular environmental datasets.

This book is an introduction to environmental data science, attempting to balance the yin (ML) and the yang (statistics) when teaching data science to environmental science students. Written as a textbook for advanced undergraduates and beginning graduate students, it should also be useful for researchers and practitioners in environmental science. The reader is assumed to know multivariate calculus, linear algebra and basic probability.

Sections are marked by the flags **A** for core material, **B** for generally useful material and **C** for more specialized material, and emojis indicating the level of technical difficulty for students – ☺ (easy), ☺ (moderately easy), ☺ (moderate), ☺ (moderately difficult) and ☺ (difficult). For instance, an instructor giving a one-term course would select topics mainly from sections **A** and if the students have limited mathematical background, skip topics marked by ☺ and ☺.

The **book website** www.cambridge.org/hsieh-ieds contains downloadable datasets needed for some of the exercises provided in this book, and the solutions to most of the exercises. Readers of the printed book (with only greyscale figures) can also download a file containing coloured figures.

~~~~~

How this book came about: With an undergraduate degree in mathematics and physics and a PhD (1981) in physical oceanography, I had, prior to 1992, zero knowledge of machine learning and very little of statistics! Through serendipity, I met Dr Benyang Tang, who introduced me to neural network models from ML. After learning this exotic topic (often the hard way) and training some graduate students in this direction, I wrote my first book *Machine Learning Methods in the Environmental Sciences*, published by Cambridge University Press in 2009. This was actually a gruelling ordeal lasting over eight years, so I thought that would be my last book.

However, three things happened on my way to retirement: (i) For a long time, machine learning was a fringe topic in the environmental sciences, but over the last five or six years, it has broken into the mainstream and has been growing exponentially. With so many fascinating new advances, I felt like a young boy unable to leave a toy store. (ii) ML and statistics have been taught separately from different books, which seems unnatural as I gradually view the two as the yin and yang side of a larger data science. Of course, this view is idiosyncratic as every ML/statistics researcher would have his/her own unique view. (iii) At conferences, enthusiastic graduate students told me that they had got into this research area from having read my first book – such comments were heartwarming to an author and made all the hard work worthwhile. So I dropped my serene retirement plans for one more book!

Writing this book has been a humbling learning experience for me. For such a vast, diverse subject, it is impossible to cover all important areas, and contributions from many brilliant scientists have regrettably been omitted.

~~~~~

I have been fortunate in having supervised numerous talented graduate students, postdoctoral fellows and research associates, many of whom taught me more than I taught them. In particular, Dr Benyang Tang, Dr Aiming Wu and Dr Alex Cannon have, respectively, contributed the most to my research group during the early, mid and late phases of my research career, especially in helping my graduate students with their research projects.

The support from the editorial team led by Dr Matt Lloyd at Cambridge University Press was essential for bringing this book to fruition. The book has also benefitted from the comments provided by many colleagues who carefully read various draft chapters.

Although retired, I remain connected, as professor emeritus, to the Department of Earth, Ocean and Atmospheric Sciences at the University of British Columbia. Having moved from Vancouver to Victoria in 2016, I am grateful to the School of Earth and Ocean Sciences, University of Victoria, for giving me Visiting Scientist status.

Without the loving support from my family (my wife, Jean, and my daughters, Teresa and Serena) and the strong educational roots planted decades ago by my parents and my teachers, especially my PhD supervisor, Professor Lawrence Mysak, I could not have written this book.

Notation Used

In general, scalars are typeset in italics (e.g. x or J), vectors are denoted by lower case bold letters (e.g. \mathbf{x} or \mathbf{a}) and matrices by upper case bold letters (e.g. \mathbf{X} or \mathbf{A}). The elements of a vector \mathbf{a} are denoted by a_i , while the elements of a matrix \mathbf{A} are written as A_{ij} or $(\mathbf{A})_{ij}$. A column vector is denoted by \mathbf{x} , while its transpose \mathbf{x}^T is a row vector, for example:

$$\mathbf{x}^T = [x_1, x_2, \dots, x_m] \quad \text{and} \quad \mathbf{x} = [x_1, x_2, \dots, x_m]^T = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}, \quad (1)$$

and the inner or dot product of two vectors $\mathbf{a} \cdot \mathbf{x} = \mathbf{a}^T \mathbf{x} = \mathbf{x}^T \mathbf{a}$.

In many environmental problems, \mathbf{x} can denote m different variables or measurements of a variable (e.g. temperature) at m different stations. The measurements are often taken repeatedly at different times up to n times, yielding $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}$. The total dataset containing m variables measured n times can be arranged in either of the matrix forms

$$\begin{bmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} x_{11} & \dots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nm} \end{bmatrix}, \quad (2)$$

with each matrix being simply the transpose of the other. In my first book (Hsieh, 2009), the first matrix form was used, but the second form has become increasingly widely used, probably due to the way data are typically arranged in spreadsheets. Hence, in this book, the data matrix \mathbf{X} is written as

$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nm} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)T} \\ \vdots \\ \mathbf{x}^{(n)T} \end{bmatrix}. \quad (3)$$

The probability for discrete variables is denoted by upper case P , whereas the probability density for continuous variables is denoted by lower case p . The expectation is denoted by $E[\dots]$ or $\langle \dots \rangle$. The natural logarithm is denoted by \ln or \log .

Abbreviations

AAO	Antarctic Oscillation	ECMWF	European Centre for Medium-Range Weather Forecasts
AIC	Akaike information criterion	EDA	exploratory data analysis
ANOVA	analysis of variance	EEOF	extended empirical orthogonal function
ANN	artificial neural network	ELM	extreme learning machine
AO	Arctic Oscillation	ENSO	El Niño-Southern Oscillation
AR	auto-regressive	EOF	empirical orthogonal function
ARIMA	auto-regressive integrated moving average	ES	environmental science
ARMA	auto-regressive moving average	ET	extra trees (extremely randomized trees)
BIC	Bayesian information criterion	ETS	equitable threat score
BLUE	best linear unbiased estimator	FFNN	feed-forward neural network
BMA	Bayesian model averaging	FFT	fast Fourier transform
BS	Brier score	GA	genetic algorithm
CART	classification and regression tree	GAN	generative adversarial network
CCA	canonical correlation analysis	GBM	gradient boosting machine
CCDF	complementary cumulative distribution function	GCM	general circulation model or global climate model
CDF	cumulative distribution function	GEV	generalized extreme value distribution
CDN	conditional density network	GP	Gaussian process model
CI	confidence interval	GSS	Gilbert skill score
CNN	convolutional neural network	HSS	Heidke skill score
ConvLSTM	convolutional long short-term memory model	IC	information criterion
CRPS	continuous ranked probability score	i.i.d.	independent and identically distributed
CSI	critical success index	IPCC	Intergovernmental Panel on Climate Change
CTFT	continuous-time Fourier transform	IQR	interquartile range
DE	differential evolution	IR	infrared
DFT	discrete Fourier transform	KDA	kernel density estimation
DL	deep learning	KNN	K -nearest neighbours
DNN	deep neural network	LDA	linear discriminant analysis
DNS	direct numerical simulation (in computational fluid dynamics)	LSTM	long short-term memory model
DTFT	discrete-time Fourier transform	MA	moving average
EA	evolutionary algorithm	MAD	median absolute deviation
		MAE	mean absolute error
		MCA	maximum covariance analysis
		MDN	mixture density network
		ME	mean error

MJO	Madden–Julian Oscillation	PI	prediction interval
ML	machine learning	PNA	Pacific–North American pattern
MLP	multi-layer perceptron neural network	POD	probability of detection
MLR	multiple linear regression	POFD	probability of false detection
MOS	model output statistics	PSS	Peirce skill score
MSE	mean squared error	QBO	Quasi-Biennial Oscillation
MSSA	multichannel singular spectrum analysis	QRNN	quantile regression neural network
NAO	North Atlantic Oscillation	RBF	radial basis function
NASA	National Aeronautics and Space Administration (USA)	RCM	regional climate model
NCAR	National Center for Atmospheric Research (USA)	ReLU	rectified linear unit
NCEP	National Centers for Environmental Prediction (USA)	RF	random forest
NLCCA	nonlinear canonical correlation analysis	RMSE	root mean squared error
NLCPCA	nonlinear complex PCA	RNN	recurrent neural network
NLPC	nonlinear principal component	ROC	relative operating characteristic
NLPCA	nonlinear principal component analysis	RPCA	rotated principal component analysis
NLSSA	nonlinear singular spectrum analysis	RPS	ranked probability score
NN	neural network	SGD	stochastic gradient descent
NOAA	National Oceanic and Atmospheric Administration (USA)	SLP	sea level pressure
NWP	numerical weather prediction	SOI	Southern Oscillation Index
OSELM	online sequential extreme learning machine	SOM	self-organizing map
PC	principal component	SS	skill score
PCA	principal component analysis	SSA	singular spectrum analysis
PDF	probability density function or probability distribution function	SSE	sum of squared errors
		SSR	sum of squares due to regression
		SST	sea surface temperature; sum of squares (total)
		SVD	singular value decomposition
		SVM	support vector machine
		SVR	support vector regression
		SWE	snow water equivalent
		TS	threat score
		UAS	unmanned aerial systems
		XGBoost	extreme gradient boosting