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

CAMBRIDGE
UNIVERSITY PRESS

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/9781107065550

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

Contents

978-1-107-06555-0 — Introduction to Environmental Data Science
William W. Hsieh
Frontmatter
More Information

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 $\boxed{\text{A}}$ for core material, $\boxed{\text{B}}$ for generally useful material and $\boxed{\text{C}}$ 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 $\boxed{\text{A}}$ 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 $\mathbf{x}^{\mathrm{T}}$ is a row vector, for example:

$$\mathbf{x}^{\mathrm{T}} = [x_1, x_2, \ldots, x_m] \quad \text{and} \quad \mathbf{x} = [x_1, x_2, \ldots, x_m]^{\mathrm{T}} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}, \tag{1}$$

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

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 $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, ..., $\mathbf{x}^{(n)}$. The total dataset containing m variables measured n times can be arranged in either of the matrix forms

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

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

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

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 $\mathrm{E}[\ldots]$ or $\langle \ldots \rangle$. The natural logarithm is denoted by ln or log.

Abbreviations

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

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