YID3202C: Data Analysis in Environmental Studies
Semester One (August – November), AY 2016-17

Instructor: Jeffrey Park
Office: RC2-M-02
jeffrey.park@yale-nus.edu.sg
Office Hours: Tues 2-4pm, Thurs 4-6pm, or by appointment

If theory were perfect, observational data in environmental studies would be explained by simple mathematical models. In real life, observational data is more complicated. Measurements can be uncertain, which leads to variations in the data that make a perfect fit to theory impossible to achieve. More importantly, environmental processes have an intrinsic variability that makes exact prediction impossible. Surface temperature varies from day to day, from season to season, and from year to year, yet the overall climate appears to be warming. Earthquakes occur in a range of magnitudes with extreme variation in timing between events. Different environments in the ocean are characterized by groups of plankton species that may define an ecosystem, but the groupings of species may overlap. A time-sequence of monthly temperatures that appears random may correlate strongly with a time-sequence of temperatures halfway across the globe, suggesting a causal relation related to El Nino or another climate process. Understanding the statistics of environmental data is often key to understanding how mankind must adapt to the Earth system, and how mankind influences the environment.

This lecture course covers a number of data analysis techniques, and demonstrates their use in environmental and earth-science research. Students should have a working familiarity with computers, because problem sets will ask them to use a standard software package to apply some of the concepts and algorithms in lecture to datasets. The instructor will hold office hours to help students master the computational assignments, especially early in the semester. There is a midterm exam and a final exam, based more on problem-solving concepts than mathematical derivations or calculations. Students who practice derivations and calculations in the problem sets will perform better in the exams.

Pre-requisite: elementary calculus and some experience with computers. Concepts of multivariable calculus, matrices, vectors, and statistics are introduced in the class to analyze problems. Previous courses in these topics would be helpful, but are not required.

Learning Goals: Students will master elementary statistical concepts, gain proficiency in a common open-source statistical package for personal computers, and become familiar to a handful of salient research problems in environmental and earth science.

Assignments: problem sets assigned weekly (with some gaps for weeks with exams), two "midterm exams" that will test students' mastery of concepts and problem-solving skills, final exam. One oral presentation on a journal article that illustrates a technique or concept in the course, performed with a partner.

Exam dates: Week 5 Monday, Week 10 Monday

Plagiarism .... Is BAD! However, learning with other students can be GOOD.
It is OK to work together on problem sets, but every student should write his or her answer on his or her assignment separately. You will not learn if you cut and paste.

We use geological, geophysical, and climate-change data sets to illustrate and master a variety of data analysis techniques important to the study of environmental processes. The course draws on earth science topics but incorporates software and algorithms that applicable to a variety of fields. Our focus includes statistical distributions, linear-regression modeling, principal-component analysis, and time-series analysis. Students read and present papers on different datasets, and also perform a variety of exercises with these same or similar data sets. The datasets are mostly taken from earth and environmental sciences, including monthly temperature anomalies used to study anthropogenic global warming, flood-frequency data for river basins, seismic data that reveals the "tones" of Earth's natural vibrations after a large earthquake, and microfossil abundances in marine sediments.

Textbook (Uses public-domain R-software package)
Author: Michael J. Crawley.
Title: Statistics: An Introduction Using R
Published: Wiley, 2014
ISBN-13: 978-1118941096

The library has access to an online version of this book. Another book that you may find useful, maybe useful enough to purchase for your own use during the course and afterwards, when you have your own dataset to analyze:

R Graphics Cookbook 1st Edition
By Winston Chang (Author)
Paperback: 416 pages
Publisher: O'Reilly Media; 1 edition (January 6, 2013)
Language: English
ISBN-10: 1449316956
ISBN-13: 978-1449316952

R is open-source software that runs on most operating systems, including Apple, Windows and Linux. You can download it and get it running on your computer from here:

https://www.r-project.org

Download the R version 3.3.1, the latest one. AFTER you have downloaded and installed R, a graphical dashboard interface for the R software can be downloaded from https://www.rstudio.com. The baseline R software must be installed on your computer before rstudio can find it. You do not need the rstudio interface, but it keeps the software output somwhat better organized.

There are many packages that researchers have written for R that augment the basic software. For instance, the R Graphics Cookbook above works with packages called "ggplot2" and "gcookbook". These packages can be retrieved from the internet by commands within R. You don't need to find their webpages or private servers. At the ">" prompt within R, you just run these commands:

```
install.packages("ggplot2")
install.packages("gcookbook")
```

and later load these libraries from your own local R-installation with these commands

```
library(ggplot2)
library(gcookbook)
```

Useful libraries will be identified and recommended to the class as we progress through data-analysis topics.

There are two class sessions per week. After the first few weeks, a pair of students in the course will be assigned to present a paper that covers either the techniques covered in class, the scientific data assigned in the problem sets, or both. The papers to be presented are listed in the syllabus. The choice of journal articles is flexible, and may be altered during the course by the prof, or by a class request, assuming sufficient time for students to adjust to a change. There will be problem sets that apply different algorithms to relevant data sets, assigned roughly weekly.

The general sequence of topics. With supplemental reading in the form of scientific journal articles

Week 1

Organizational. Probability, Statistics, Random Variables, Computer Software Exercises

Perception of climate change, James Hansen, Makiko Sato, and Reto Ruedy, Proceedings of the National Academy of Sciences of the USA, v109, p. E2415-E2423, 2012, DOI: 10.1073/pnas.1205276109.

READING: Crawley, Chapter 1

Week 2

Types of Statistical Distributions for Environmental Data I

Earthquake Hazard After a Mainshock in California, PAUL A. REASENBERG, LUCILE M. JONES, Science, v243, pp. 1173-1176, 1989, DOI: 10.1126/science.243.4895.1173

Week 3

Types of Statistical Distributions for Environmental Data II

Statistics of extremes in hydrology, R. W. Katz, M. B. Parlange, P. Naveau, Advances in Water Resources, v25, p1287–1304, 2002, doi:10.1016/S0309-1708(02)00056-8.

Week 4

Least Squares Model-Fitting. Goodness-of-Fit

Salinity change in the subtropical Atlantic: Secular increase and teleconnections to the North Atlantic Oscillation, Brad E. Rosenheim, Peter K. Swart, Simon R. Thorrold, Anton Eisenhauer, and Philippe Willenz, GEOPHYSICAL RESEARCH

LETTERS, VOL. 32, L02603, doi:10.1029/2004GL021499, 2005

Week 5 (First Midterm Exam)
Least Squares Model-Fitting. Jackknife & Bootstrap Uncertainty Estimates
Assessing the Quality of Earthquake Catalogues: Estimating the Magnitude of Completeness and Its Uncertainty, Jochen Woessner and Stefan Wiemer, Bulletin of the Seismological Society of America, Vol. 95, pp. 684–698, 2005, doi: 10.1785/0120040007

Week 6
Multiparameter Least Squares Model-Fitting.
Aspherical earth structure from fundamental spheroidal-mode data, Masters, G., T. H. Jordan, P. G. Silver, and F. Gilbert, Nature, v298, p609-613, 1982.

Week 7
Principal Component Analysis
Using principal components analysis (PCA) with cluster analysis to study the organic geochemistry of sinking particles in the ocean, Jianhong Xue, Cindy Lee, Stuart G. Wakeham, Robert A. Armstrong, Organic Geochemistry, v.42, p. 356–367, 2011, doi:10.1016/j.orggeochem.2011.01.012.

Week 8
Spectrum estimation. Discrete Fourier Transform
Long-term trends and cycles in the hydrometeorology of the Amazon basin since the late 1920s, J. A. Marengo, Hydrol. Process. 23, 3236–3244 (2009)

Week 9
Spectrum estimation II
*Earth's Free Oscillations Excited by the 26 December 2004 Sumatra-Andaman Earthquake, Park, J., T. R. Song, J. Tromp, E. Okal, S. Stein, G. Roult, E. Clevede, G. Laske, H. Kanamori, P. Davis, J. Berger, C. Braitenberg, M. Van Camp, X. Lei, H. Sun, H. Xu, and S. Rosat, **Science, 308**, 1139-1144, 2005.*

Week 10 (Second Midterm Exam)
Autoregressive Processes
Seismic Excitation of the Polar Motion, 1977-1993, B.F. CHAO, R. S. GROSS, and Y.-B. HAN, Pure and Applied Geophys., Vol. 146, 407-419 (1996)

Week 11
Data Tapers and Spectral Leakage: Slepian Tapers
On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform, F. J. HARRIS, PROCEEDINGS OF THE IEEE, VOL. 66, pp51-83, 1978.

Week 12
Coherence and Correlation
SPATIAL CORRELATIONS OF INTERDECADAL VARIATION IN GLOBAL SURFACE TEMPERATURES, Michael E. Mann and Jeffrey Park, Geophys. Res. Letts., V20, NO. 11, PAGES 1055-1058, (1993)

Week 13
Multivariate Spectra: Climate Data

Joint Spatiotemporal Variability of Global Sea Surface Temperatures and Global Palmer Drought Severity Index Values, Apipattanavis S, McCabe GJ, Rajagopalan B, et al., JOURNAL OF CLIMATE Vol22 6251-6267 (2009)