EAEE E4257: Environmental Data Analysis and Modeling – Spring 2023

Learning Outcomes:
This course will provide students with an understanding of fundamental statistical concepts for understanding and modeling environmental data. In this class we will focus on analyzing and modeling such data sets with a focus on space and time variations, and potential inter-relations between variables. The aim is to help students develop a conceptual understanding of statistics, rather than memorizing a set of routines and methodologies.

- How to think about statistical data and their analysis in the context of typical environmental problems
- Key techniques in statistical modeling and machine learning, their underlying ideas and applicability, and how to use them
- How to build, select, and check statistical models using modern computational techniques
- How to address core challenges in environmental data analysis, including spatial correlation, structured time series behavior, changepoints, and trends

Above all, students who take this course will learn to think critically about environmental data and models and will be able to understand and discuss conceptually the limitations of models they use or encounter "in the wild".

Time: Mon, Wed 11:40 pm – 12:55 pm; **Classroom:** Mudd 633

Instructor: Bolun Xu
Office: Mudd 918G
Email: bx2177@columbia.edu
Office hours: Mudd 918G (or class zoom link; please email me first)
Wed 2:30 pm – 3:30 pm

Teaching Assistant: Umar Salman
Email: uts2000@columbia.edu
Office hours: Mudd 918 Project Room
Mon 10:00 am – 11:00 am; Thu 2:00 pm – 3:00 pm

Graders:
Isabela Maria Yepes imy2103@columbia.edu
Yuelan Zhu yz4160@columbia.edu

Grading:
Homework: 70%
Final exam (take home): 30%

Pre-Requisites

These topics are pre-requisites for the course and will *not* be covered. We will post a **Homework 0** prior to the first day of class so that you can assess your knowledge.

- Matrix and vector algebra
- Discrete probability distributions (particularly Bernoulli, binomial, and Poisson)
- Continuous probability distributions (particularly normal, exponential, and uniform)
- Sample Estimates
- Correlation and Covariance

Reference textbooks:

[**JWHT**] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Available Online at https://www.statlearning.com/

[**HH**] Helsel, D.R. and R. M. Hirsch, 2002. Statistical Methods in Water Resources Techniques of Water Resources Investigations, Book 4, chapter A3.U.S. Geological Survey. 522 pages. Freely available at https://pubs.usgs.gov/tm/04/a03/tm4a3.pdf

[**Hengl**] Tomislav Hengl (2009). A Practical Guide to Geostatistical Mapping. Available Online at https://publications.jrc.ec.europa.eu/repository/handle/JRC38153

Homework, Exam, and Grading

Computing/Programming

All computation in this course will be done in the **R** language using **Jupyter Notebook**, and in particular we will use the tidyverse package ecosystem wherever possible. This is not to say that other languages or frameworks are not helpful -- we regularly other languages and other sets of packages. However, we feel that the wide availability of statistical methods already implemented in **R**, the interactive `RStudio` environment, and easy-to-learn tidyverse packages give us the most "bang for our buck".

We do not expect that students enrolling in this class have used **R** before, but basic programming knowledge is required. If you are not sure about your programming background, make sure you are comfortable with HW1, and talk to TA.

Homework

We will have eight homework: HW0 to HW7. HW0 will be about assessing your background in linear algebra and probability, and **will not be graded**; HW1 is an online tutorial for R to prepare you for coding in R. HW2-HW7 are computation homework in R.

Homework 1-7 each worth 10% of total course credit and the total homework credit is 70%.

Late policy: Homework is due 11:59PM on **Fridays** and must be submitted via Coursework. Submission portal will remain open for 24 hours after the due date and late submission receives 10% penalty of the full homework grade. No submission allowed after submission portal is closed.

You can discuss homework with your classmates, but homework writings/codes must be your own work. To receive full credit, students must thoroughly explain how they arrived at their solutions. If the homework is **computational** (as most will be), it should be turned in as a **Jupyter Notebook (.ipynb)** file, along with the **.html** file which is the result of running it. If the homework is not computational, it should be turned in as a **.pdf** file.

Final Exam

A 24-hour take-home final exam will be assigned during the final exam period. The final exam is in a R computation exam of similar format to Homework in .ipynb, and you will turn in your solution also in .ipynb and .html format.

Tentative Schedule

Lecture Calendar

Date	Topics	References
1/18	Introduction	
1/23	Introduction: Model estimation	HH Ch2
1/25	Introduction: Multi-variant distribution	HH Ch8
1/30	Linear regression	JWHT 3.1
2/1	Multiple linear regression	JWHT 3.2-3.3
2/6	Trend analysis: Auto-regressive models	HH 12.3-12.6
2/8	Trend analysis: Auto-regressive models	HH 12.3-12.6
2/13	Trend change analysis: Rank-sum tests	HH 4, HH 5.1-5.2
2/15	Trend change analysis: t-test and bootstrap	HH 5.3-5.4
2/20	Smooth trend: Mann-Kendal test	HH 12.2
2/22	Smooth trend: local regression	HH 10.3
2/27	Weighted least square	JWHT 7.6
3/1	KNN and KDE	JWHT 2.2-KNN
3/6	Generalized linear model	JWHT 4.6
3/8	Generalized linear model	
3/14	Spring break	
3/16	Spring break	
3/20	LASSO & PCA	JWHT 6.2
3/22	Clustering: K-means	JWHT 12.4.1
3/27	Clustering: Hierarchical clustering	JWHT 12.4.2
3/29	Support Vector Machines	JWHT 9.1-9.3
4/3	Spatial statistics: Intro	Hengl 1.1
4/5	Spatial statistics: Trend surface estimation	Hengl 1.2
4/10	Spatial statistics: Trend surface estimation	Hengl 1.2
4/12	Kriging: Intro	Hengl 1.3-1.4
4/17	Kriging: Application	Hengl 1.3-1.4
4/19	Bayesian analysis	
4/24	Markov Chain Monte Carlo	
4/26	Hidden Markov models	
5/1	Review	

Homework and Exam Calendar

Homework	Topics	Due date
HW0	Background in linear algebra and probabilities	01/27
HW1	R tutorial and set-up	01/27
HW2	Data visualization, distribution, and density	02/10
HW3	Linear regression and trends	02/24
HW4	Local regression	03/09
HW5	Generalized linear models	03/31
HW6	Clustering	04/14
HW7	Geostatistics / Kriging	04/28
Final exam	24-hour take home exam	05/08-05/09

Grading:

Homework: 70%
Final exam: 30%

Grade distributions (curve up if possible depends on the grade distribution):

A+	97%
A	94%
A-	90%
B+	87%
B	84%
B-	80%
C+	77%
C	74%
C-	70%
D	< 70%