EAEE E4257: Environmental Data Analysis and Modeling – Spring 2026

Learning Outcomes:
This course will provide students with fundamental statistical concepts for understanding and modeling environmental data. In this class, we will focus on analyzing and modeling such data sets, focusing 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 10:10 pm – 11:25 pm;   **Classroom:** 420 Pupin

Instructor: Bolun Xu

Office:	Mudd 918G
Email:	bx2177@columbia.edu
Office hours:	Mudd 918G Tue 3 pm – 4 pm

Teaching Assistant: Emily Logan

Email:	el3350@columbia.edu
Office hours:	Mudd 924 Project Room Thu 2:00 pm – 3:00 pm

Graders: Pina Pi

Email:	rp3255@columbia.edu

Grading:

Homework:	60%
Quiz:	30%
Attendance:	10%

Pre-Requisites

(STAT4001W OR IEOR3658E OR IEOR4150E) AND ENGI1006E

These topics are prerequisites for the course and will *not* be covered. We will post a **Homework 1** before 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 HW1 is an online tutorial for R to prepare you for coding in R. HW2-HW7 are computation homework in R.HW0 and HW1 each worth 5%, while Homework 2-7 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.

In case your computer broke down, you can use on campus computing resources here: https://www.cuit.columbia.edu/computer-lab-technologies/locations-hours. It is your responsibility to make sure you back up your code/process on the cloud such as Google drive or Github, and submit your homework on time. Missing homework submission due to computer issues will not be allowed.

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.

Plagiarism Review. Homework will be reviewed using Turnitin from Coursework for plagiarism review (over class submissions and previous year submissions). Students who are found plagiarizing will receive zero credit for the homework and receive a warning at the first offense, and will be reported to SEAS per school policy requirements.

Grading inquiries. Please contact TA (not Graders) directly for grading concerns.

Midterm and Final Exam

The midterm and final exams will be 30 minutes in class. The format is closed book, closed notes, and no calculator. Each exam will include 15 multiple-choice and true or false questions and test you on the conceptual understanding of the course topics and ability to understand sample codes. A sample midterm exam and solutions are available on Coursework.

Use of AI Tools for Homework

Students are permitted to use AI-based tools (e.g., large language models or coding assistants) to support homework, including help with coding syntax, debugging, and exploring alternative approaches. However, students are fully responsible for all submitted work. In particular:

- You must understand, verify, and validate any code, results, figures, or explanations generated with AI assistance.
- All analyses, interpretations, and conclusions must reflect your own reasoning and be consistent with course concepts.
- You must be able to explain and defend your solution, code logic, and results if asked.
- Blindly copying AI-generated outputs, submitting incorrect or unverified results, or using AI to replace your own thinking constitutes academic misconduct and will be treated as plagiarism.
- All homework submissions are subjected to Turnitin plagiarism review built in Coursework.

You can check out AI tools from Columbia: https://www.cuit.columbia.edu/content/ai-services

Honor Code:

Please revise the Columbia SEAS Honor Code: https://www.wikicu.com/SEAS_Honor_Code

Tentative Schedule

Lecture Calendar (W(eek) Number and Monday/Wedesday)

Date	Topics	References
W1W	Introduction	
W2M	Introduction: Model estimation	HH Ch2
W2W	Introduction: Multi-variant distribution	HH Ch8
W3M	Introduction: Multi-variant distribution	HH Ch8
W3W	Linear regression	JWHT 3.1
W4M	Multiple linear regression	JWHT 3.2-3.3
W4W	Multiple linear regression	JWHT 3.2-3.3
W5M	Trend analysis basics	HH 12.3-12.6
W5W	Trend analysis: Auto-regressive models	HH 12.3-12.6
W6M	Trend analysis: Auto-regressive models	HH 12.3-12.6
W6W	Quiz 1 / Rank-sum tests	HH 4, HH 5.1-5.2
W7M	Trend change analysis: Rank-sum tests	HH 4, HH 5.1-5.2
W7W	Trend change analysis: t-test and bootstrap	HH 5.3-5.4
W8M	Smooth trend: Mann-Kendal test	HH 12.2
W8W	Generalized linear model	JWHT 4.6
W9M	Spring break	
W9W	Spring break	
W10M	LASSO & PCA	JWHT 6.2
W10W	Classification	JWHT 9.1-9.3
W11M	Code recitation	
W11W	Clustering	JWHT 12.4
W12M	Quiz 2 / Code recitation	
W12W	Clustering	JWHT 12.4
W13M	Class Cancelled	
W13W	Class Cancelled	
W14M	Spatial statistics: Intro	Hengl 1.1
W14W	Spatial statistics: Trend surface estimation	Hengl 1.2
W15M	Spatial statistics: Trend surface estimation	Hengl 1.2
W15W	Kriging: Intro	Hengl 1.3-1.4
W16M	Kriging: Applications	Hengl 1.3-1.4

Homework and Exam Calendar

Homework	Topics	Due date
HW1	Data visualization, distribution, and density	2/6
HW2	Linear regression and trends	2/20
HW3	Local regression	3/13
HW4	Generalized linear models	4/3
HW5	Clustering	4/17
HW6	Geostatistics / Kriging	5/1

Grade distributions (curve up may be possible depending on the grade distribution):

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