GEOL 425L
Data Analysis in the Earth & Environmental Sciences

Fall 2023

General Information

Where/When Class meets Tues-Thurs 14:00-15:20 in ZHS 200.
Lab meets Wed 15:30-17:70 in ZHS 130.

Instructors

Professor:	Julien Emile-Geay	ZHS 275	julieneg@usc.edu
Teaching Assistant:	Binhao Wang	ZHS 154	binhaowa@usc.edu

Office Hours Julien: Thurs 10-12pm or by appointment (ZHS 275).

Preparation MATH 125-126, Matrix Algebra

Overview

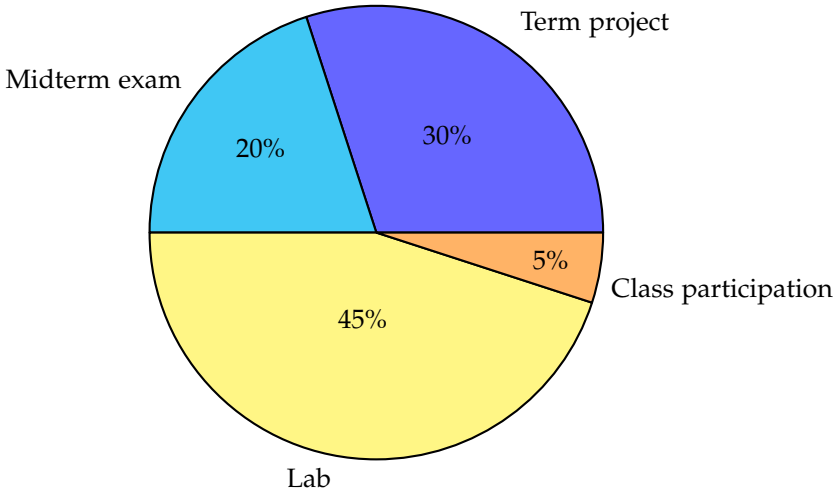
Objectives Scientific reasoning always involves data to some extent. This is the class where you learn how to reason about data, thinking critically about what they really allow you to say. Essential skills we teach are:

- Performing elementary calculations with real-world data.
- Visualizing data with error estimates and perform basic error propagation analyses.
- Computing and correctly interpreting a correlation coefficient.
- Computing, representing and correctly interpreting spectra.
- Performing basic linear regressions and least-squares fits.
- Being conversant with classic parametric and non-parametric statistical tests.
- Mastering basic data reduction techniques like principal component analysis
- Applying a number of these tools to your own research.

Philosophy The class is articulated around three main themes:

1. Living in an uncertain world
2. Living in the temporal world
3. Living in multiple dimensions

We begin each section of the class with an appropriate refresher in the underlying mathematical foundation (calculus, complex numbers, linear algebra and probability theory). We then describe the theory behind quantitative tools and then have students apply them to real-world problems from the solid and fluid Earth. in the form of weekly laboratory practicums and a final paper. By the end of the class, the goal is for you to realize that every scientific statement is probabilistic in nature. You will learn to reason quantitatively about a dataset from your field of study, and to write about it in a knowledgeable way.



Grade

The class will earn you 4 units, which means that it requires substantial work, every week. Please get in touch if you find yourself falling behind assignments for any reason. Proactive communication goes a long way.

Rules There aren't many rules for the course, but they're all important. First, read the assigned readings before you come to class. Second, turn everything in on time. Third, ask questions when you don't understand things; chances are you're not alone. Fourth, don't miss class or lab.

Computing We will be using Python as a computing/visualization package. Prior exposure to python, while not strictly necessary if you know other object-oriented languages, would be desirable. Many online tutorials exist for that, like this bootcamp.

If you have never programmed before, and still doubt that this would be a useful skill, look at the current job market. It simply is an indispensable skill in today's world. Some people like to learn by "sink or swim", but I recommend online tutorials prior to the class start for a smoother experience.

If you are already conversant with another programming language (e.g. Matlab, R, Julia), you may program in that for your final project, but still need to turn in a reproducible computational narrative ("notebook") of some kind. All major languages support that. If you wish to use a more esoteric language, you are on your own.

Term Paper Other than the laboratory practicums, the main assignment for this class is for you to write a paper that implements one or several techniques used in this class for your own work. This is worth about 1/3 of the grade and is usually underappreciated by students, who prefer to freak out over the midterm exam. So let it be known: the midterm will be easy, and mostly a measure of much you've come to class. The real work is in the weekly labs and term paper.

Late Work With assignments due virtually every week of the term, it's easy to fall behind. While it may seem desirable to take extra time to deepen your understanding of a subject, this will have a domino effect on subsequent assignments. As a result, lab assignments are due every Wednesday, one week after each lab session. A 5 points penalty for every late day will be assessed.

Reading

Class notes The notes are available as an e-book, last updated in July 2023. Despite multiple rounds of corrections, some typos remain, so it will highly benefit from your careful reading. Submitting comments, pointing out typos, asking questions (whether in class or via electronic interaction) will all count for class participation. If you miss class, it is *highly* recommended that you catch up with reading from the previous week before a lab, as it will save you (and your TA) a considerable amount of time.

Books The notes being necessarily partial, many of you will want to explore some subjects more deeply, so here is a short (non-exhaustive) list of useful books.

Undergraduate books

- Taylor, J.R., An Introduction to Error Analysis, University Science Books, 1997. URL.
  A very approachable perspective on error analysis, written by a physicist for readers equipped with minimal mathematical literacy. Very entertaining and quite effective.

Graduate books

- Gubbins, D. Time Series Analysis and Inverse Theory for Geophysicists, Cambridge University Press, 2004. URL. *A very succinct introduction to timeseries analysis, especially useful to geophysicists*.
- Wilks, D., Statistical methods in the atmospheric sciences, (3rd ed.), Academic Press, 2011. URL. *A bible for data analysis in the atmospheric and oceanic sciences*.
- Venegas, S. Statistical Methods for Signal Detection in Climate URL. *A great (and free!) set of notes describing just about every analysis method you will ever encounter in climate science*.

Advanced Books

- Menke, W.H.. Geophysical Data Analysis: Discrete Inverse Theory (Third Edition), URL.
  A modern classic in inverse theory, written for geophysicists.
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, R. Bayesian Data Analysis, URL. *The ultimate *practical* reference in Bayesian data analysis*.
- Hastie, T., Tibshirani, R., Friedman, J.:The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition). URL. *A very lucid exposition of all aspects of statistical learning, written by statisticians for non-statisticians who just want to use statistics, not philosophize about them. Highly recommended.*
- Wavelets in the Geosciences, URL *Very thorough mathematical introduction to wavelets, which are now a mainstay of data analysis in the Earth Sciences.*

Schedule

I Living in an uncertain world: probability and statistics

The first section of the class focuses on the fundamental problem of data analysis: uncertainties. This is the domain of probability theory and statistical inference.

Week 1 — August 21— Math Review

Tuesday: Calculus review: differentiation; integration; Taylor expansions and approximations.

Wednesday: Lab 0: Introduction to Python. Elementary Computations and Graphics

Thursday: Linear Algebra Review: Basis. Orthogonality. Matrix algebra. Invertibility.

Read: Notes, Appendix A & B. Chapter 1.

Week 2 — August 28— Probability Theory I.

Tuesday: Probability theory as extended logic. Probability calculus. Law of total probability.

Wednesday: Lab 1: Integration. Orthonormality. Spherical Harmonics.

Thursday: Bayes' theorem. Bayesian vs frequentist interpretation. Inference.

Read: Notes, Appendix B. Chapter 2.

Week 3 — September 4—Probability Theory II

Tuesday: Random Variables. Probability Laws. Distribution functions. Moments. Quantiles.

Wednesday: Lab 2: Matrix Inversion as applied to Earthquake Deformation

Thursday: Exploratory Data Analysis

Read: Notes, chapter 2, 3.

Week 4 — September 11—Probability Theory III

Tuesday: Classic distributions (discrete and continuous)

Wednesday: Lab 3: Exploratory Analysis of Climate Data

Thursday: Normal distribution. Central Limit Theorem. Error analysis.

Read: Notes, chapter 3.

Week 5 — September 18—Univariate Statistics I

Tuesday: Statistical estimation I: maximum likelihood principle. quality of estimators.

Wednesday: Lab 4: The normal distribution as an error analysis tool.

Thursday: Statistical estimation II: Bayesian Data Analysis.

Read: Notes, chapter 4, 5.

Week 6 — September 25— Univariate Statistics II

Tuesday: Confirmatory Data Analysis. Confidence Intervals

Wednesday: Lab 5: Unmixing Ice Ages. Testing for Drought. Fitting ocean currents.

Thursday: Classic parametric tests: Z, T, F and χ^2 tests. Significance of correlations.

Read: Notes, chapter 6.

Week 7 — October 2— Midterm

Tuesday: Non-parametric tests.

Wednesday: Midterm Review

Thursday: MIDTERM EXAM

Read: Notes, chapter 6. Appendix B.

Week 8 — October 9— Circular Functions

Tuesday: Trigonometry & Complex Numbers Review

Wednesday: no lab

Thursday: Fourier series & transform.

Read: Appendix C.

FALL BREAK : Oct 12 – 15

II Living in the temporal world: timeseries analysis

Up to now we have considered data and their uncertainties; never their order. Timeseries analysis is all about finding patterns in sequential observations, and assessing their significance.

Week 9 — October 16— Timeseries Analysis I

Tuesday: Fourier transform: Important theorems.

Wednesday: Lab 6: Fourier Analysis & Synthesis.

Thursday: Linear Algebra redux: Functional Spaces. Projection. Least Squares

Read: Notes, chapter 7.

Week 10 — October 23— Timeseries Analysis II

Tuesday: Discrete Fourier Transform. Fourier Sampling Theory.

Wednesday: Class project problematization

Thursday: Spectral Analysis. Timeseries Modeling.

Read: Notes, chapter 7, 8, 9.

III Living in multiple dimensions: multivariate analysis

In the brief time that is allotted to us, we now tackle multivariate problems: problems involving space, time, or other dimensions, and the mathematical challenges they pose. A central theme is how to estimate parameters from uncertain data, or predict one variable given another.

Week 11 — October 30— The Multivariate Normal

Tuesday: Advanced Spectral Analysis.

Wednesday: Lab 7: Correlations

Thursday: The Multivariate Normal Distribution

Read: Notes, chapter 9 & 11.

Week 12 — November 6— Data Reduction

Tuesday: Diagonalization

Wednesday: Lab 8: Advanced Spectral Analysis

Thursday: Principal Component Analysis

Read: Notes, Appendix D; chapter 12.

Week 13 — November 13— Least Squares

Tuesday: Least Squares

Wednesday: Lab 9: SVD and Empirical Orthogonal Functions

Thursday: Univariate Linear regression

Read: Notes, chapter 13, 15.

Week 14 — November 20— Advanced Topics (optional)

Tuesday: Choosing from: Wavelet and Multiresolution Analysis. Singular Spectrum Analysis. Changepoint detection. Data Assimilation.

Thanksgiving Break Nov 22–26

Week 15 — November 27— Linear Regression

Tuesday: Multivariate Linear Regression

Wednesday: Lab 10: Linear regression

Thursday: Working with Geoscientific Data. Visualization & Sonification.

Read: Notes, chapter 14.

Dec 9—Final Project Due

IV Term Project

The high point of this course is an individual research project where you apply the methods learned over the semester to a dataset of your choosing, demonstrating working knowledge of the material. The ideal project will take data that you or your lab generated, and use it to make advance your own research. If you are not currently research-active, or are too lazy to Google a dataset, I can supply you with one, but you'll have much more fun investigating a topic of your choosing. Here are a few recommendations to make it a pleasant experience for everyone involved:

Overview

- State the problem and purpose (what you want to accomplish with the data)
- Describe the approach and techniques to be used to accomplish the stated goals
- Pick $p \geq 1$ datasets of at least $n = 128$ points (higher n and p are desirable, but not mandatory).
- Analyze the data, computing uncertainties whenever possible and investigating the sensitivity to key parameters.
- Interpret the results of each technique used.
- Discuss the successes/failures of the approaches used.
- Provide an overall conclusion.

Methods
Acceptable methods include:

- Exploratory data analysis: density estimation, low-order moments, autocorrelation, range, etc.
- Some form of curve fitting (e.g. interpolation)
- using the data to form and evaluate one or more hypotheses
- If timeseries: some form of spectral analysis and/or filtering
- If multivariate: principal component analysis, correlations and/or regression
- (Grad only) changepoint analysis, analysis of unevenly-spaced or time-uncertain data, wavelet analysis, cross-spectral analysis.

If you do not plan on using any of these, get the green light from me first.

Timeline
Please pick a dataset as early as possible in the semester. The data generators among you can start with a preliminary dataset, since it will be trivial to extend your analysis to the whole dataset once you have more data. The papers are **due by 23:59 PST on Dec 9**. Please do yourself a favor and do not wait until the last possible minute to get started. As a safeguard, the lab session of Week 10 will be devoted to a preliminary analysis of your dataset. You should aim to have data on hand **at least two weeks before that**.

Writing
Just because this is a relatively mathematical class, does not mean that you can get away with poor writing. As emphasized above, communicating your results is at least as important as the analysis itself, so I'll want to see some clear reasoning about data. We shall assume familiarity with the principles of scientific writing, and I'll expect succinct, lucid analyses of what the data say. We're on the same side here: I don't want to read a long paper any more than you want to write one, so make every word count. Exact length is unimportant, but in general I expect about 5-10 pages of *double-spaced* text, not including figures: 1-2 pages for the introduction (motivation, presentation of dataset), 1-2 pages for the results, and 1-2 pages for the discussion/conclusion.

Graphics
Given how important graphics are to written and oral presentations, it's staggering how mediocre most published figures are. Early on in this course, you will be learn how to properly label and annotate your figures, design them to eliminate chart junk, and to export any figure in vector format. Failing to apply these principles will result in 5 points being deducted from your paper.

Reproducibility
Another key feature that you will hopefully learn in this class is that the ability to reproduce past analyses is central to the scientific process itself. Accordingly, you will be sharing code and data.

Format
The project itself should be submitted via Blackboard as a zip file containing:

1. A Jupyter Notebook, appropriately commented, containing all the code, figures and interpretation.
2. the data necessary to run the notebook (if less than 10Mb – otherwise talk to me).
3. a PDF rendition of your notebook, for easier commenting and grading.

I will not accept Microsoft, Apple, OpenOffice, or any other proprietary format. Work turned in using those formats will not be looked at.

V Academic Conduct

Collaboration

In this class, you are expected to submit work that demonstrates your individual mastery of the course concepts. All assignments are expected to be completed individually.

Integrity

The University of Southern California is foremost a learning community committed to fostering successful scholars and researchers dedicated to the pursuit of knowledge and the transmission of ideas. Academic misconduct is in contrast to the university's mission to educate students through a broad array of first-rank academic, professional, and extracurricular programs and includes any act of dishonesty in the submission of academic work (either in draft or final form).

This course will follow the expectations for academic integrity as stated in the USC Student Handbook. All students are expected to submit assignments that are original work and prepared specifically for the course/section in this academic term. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s). Students suspected of engaging in academic misconduct will be reported to the Office of Academic Integrity.

Other violations of academic misconduct include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

For more information about academic integrity see the USC Student Handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Plagiarism

Presenting someone else's ideas as your own, either verbatim or recast in your own words, is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Section 11, Behavior Violating University Standards. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct.

Discrimination

Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the Office of Equity and Diversity or to the Department of Public Safety . This is important for the safety whole USC community. Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person. The Relationship and Sexual Violence Prevention and Services provides 24/7 confidential support, and the sexual assault resource center webpage describes reporting options and other resources.

Support Systems

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the American Language Institute, which sponsors courses and workshops specifically for international graduate students. The Office of Disability Services and Programs for students with disabilities and helps arrange the relevant accommodations. If an officially declared emergency makes travel to campus infeasible, USC Emergency Information will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.

Artificial Intelligence

You are allowed to use generative AI (e.g., ChatGPT) in this class to help climb the coding learning curve. Learning to use AI is an emerging skill, so keep in mind the following:

- AI tools are permitted to help you generate coding strategies, but you are responsible for implementing them and checking that the results make sense. Be warned that you will rarely get a good answer on the first try.
- If you provide minimum-effort prompts, you will get low-quality results. You will need to refine your prompts to get good outcomes. This will take work.
- Proceed with caution when using AI tools and do not assume the information provided is accurate or trustworthy. If it gives you a number or fact, assume it is incorrect unless you either know the correct answer or can verify its accuracy with another source. You will be responsible for any errors or omissions provided by the tool. It works best for topics you understand.
- AI is a tool, but one that you need to acknowledge using. Please include a paragraph at the end of any assignment that uses AI explaining how (and why) you used AI and indicate/specify the prompts you used to obtain the results what prompts you used to get the results. Failure to do so is a violation of academic integrity policies.
- Be thoughtful about when AI is useful. Consider its appropriateness for each assignment or circumstance. The use of AI tools requires attribution. You are expected to clearly attribute any material generated by the tool used.

Feel free to ask me if you are unsure about what constitutes unauthorized assistance on an exam or assignment, or what information requires citation and/or attribution.