GEOL 425L
Data Analysis in the Earth & Environmental Sciences

Fall 2019

General Information

Where/When Class meets Wed-Fri 10:30-11:50am in ZHS 130.
Lab meets Fri 2:00-3:50 in ZHS 130.

Instructors

Professor:	Julien Emile-Geay	ZHS 275	julieneg@usc.edu
Teaching Assistant:	Alan Juarez	ZHS 266	alanjuar@usc.edu

Office Hours Julien: Wed 2-5pm or by appointment (ZHS 275).

Preparation MATH 125-126, Matrix Algebra

Overview

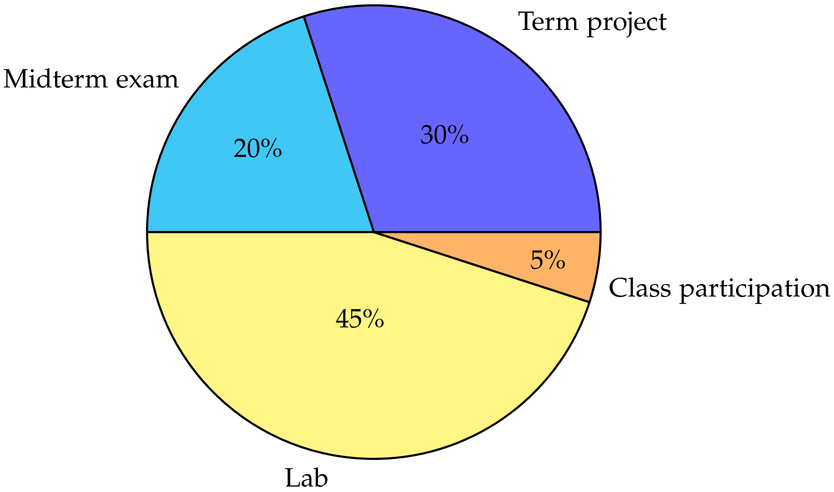
Objectives Scientific reasoning always involves data to some extent. This is the class where you learn how to reason about data, thinking critically about what they really allow you to say. Essential skills we teach are:

- 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 very substantial work, every week. I do not believe in curving grades; if everybody gets an A, I'll pop some bubbly.

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 mainly using Python as a computing/visualization package. If you have never programmed before, this will be your trial by fire. Be reassured: in this modern world, learning how to program is as fundamental as reading and writing, so this is a skill you have to learn at some point, and there isn't really any other way to learn programming than just doing it. We will provide large amounts of code, and if you're smart, reusing it will save you months of work on your thesis or make you a data analysis hotshot that any company will want to hire.

If you are already conversant with another programming language (e.g. Matlab, R), you may program in that for your final project. (Excel/Visual Basic does not count as a programming language).

Term Papers 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 Friday, 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, first published in May 2014. Despite multiple rounds of corrections, some typos are still lurking, so it will highly benefit from your careful reading. Submitting comments, pointing out typos, asking questions about them (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 notes 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 26— Introduction & Math Review

Wednesday: Introduction to Data Analysis; Calculus review

Friday: Linear Algebra Review

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

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

Week 2 — September 2— Linear Algebra. Probability Theory I.

Wednesday: Linear Algebra Review: Basis. Projection. Orthogonality. Functional Spaces

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

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

Read: Notes, Appendix B. Chapter 2.

Week 3 — September 9—Probability Theory II

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

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

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

Read: Notes, chapter 2, 3.

Week 4 — September 16—Probability Theory III

Wednesday: Exploratory Data Analysis

Friday: Classic distributions (discrete and continuous)

Friday: Lab 3: Exploratory Analysis of Rainfall Data

Read: Notes, chapter 3.

Week 5 — September 23—Univariate Statistics I

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

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

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

Read: Notes, chapter 4, 5.

Week 6 — September 30— Univariate Statistics II

Wednesday: Statistical estimation II: Bayesian Data Analysis.

Friday: Confirmatory Data Analysis. Classic Parametric Tests: Z, T.

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

Read: Notes, chapter 6.

Week 7 — October 7— Univariate Statistics III

Wednesday: F and χ^2 tests. Non-parametric tests. Significance of correlations.

Friday: Trigonometry & Complex Numbers Review

Friday: NO LAB: midterm review

Read: Notes, chapter 6. Appendix C.

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 8 — October 14— Midterm

Wednesday: MIDTERM EXAM

FALL BREAK : Oct 17 – 18

Week 9 — October 21— Timeseries Analysis I

Wednesday: Fourier series & transform. Important theorems.

Friday: Discrete Fourier Transform. Fourier Sampling Theory.

Friday: Lab 6: Fourier Analysis & Synthesis.

Read: Notes, chapter 7.

Week 10 — October 28— Timeseries Analysis II

Wednesday: FFT. Practical Spectral Analysis.

Friday: Timeseries Modeling.

Friday: Class project problematization

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 — November 4— The Multivariate Normal

Wednesday: Advanced Spectral Analysis.

Friday: The Multivariate Normal Distribution

Friday: Lab 7: Correlations

Read: Notes, chapter 11, Appendix D.

Week 12 — November 11— PCA

Wednesday: Diagonalization: Singular Value Decomposition and Eigensystems

Friday: Principal Component Analysis

Friday: Lab 8: Advanced Spectral Analysis

Read: Notes, chapter 12.

Week 13 — November 18— Least Squares Fitting

Wednesday: Least Squares

Friday: Univariate Linear regression

Friday: Lab 9: SVD and Empirical Orthogonal Functions

Read: Notes, chapter 13, 15.

Thanksgiving Break Nov 27–Dec 1

Week 14 — December 2— Linear Regression

Wednesday: Multivariate Linear Regression

Friday: Working with Geoscientific Data. Visualization & Sonification.

Friday: Lab 10: Linear regression

Read: Notes, chapter 14.

Dec 8—Final Project Due

IV Term Project

The meat 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 fundamental advances in 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 linear 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 on Dec 8**. 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. At a minimum, I expect access to your code, if not your data.

Format

All final written work should be turned in double-spaced as pdf files. I will not accept Microsoft, Apple, OpenOffice, or any other proprietary format. Work turned in using those formats will not be looked at and subsequent pdf files will be considered late work.

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

1. the paper
2. the code, appropriately commented (so I can understand it, and so you can remember what you did when you look at it a few months from now).
3. the data, in a machine-readable format (with the script to read them in)

A Jupyter Notebook is an acceptable submission format as well.