CEE 5984 Data Analysis for Environmental Science and Engineering

The Charles E. Via, Jr. Department of Civil and Environmental Engineering
Virginia Polytechnic Institute and State University
Fall Semester, 2020

Instructor:	Dr. Megan A. Rippy	E-Mail: mrippy@vt.edu
Office:	OWML 207	Office Phone: (703) 361-5606 x112
Office Hours:	by appointment (*please email me with questions or to schedule a Zoom meeting)*	

Course Description

Catalog Description:	This course covers a range of data curation and analysis techniques including 1) graphical interpretation of data (biplots, path diagrams etc.), 2) resampling-based statistics for hypothesis testing, regression, and quantifying statistical confidence, and 3) multivariate statistics, including principal component analysis, correspondence analysis, cluster analysis, and structural equation modeling. This course will be computer intensive and require using Matlab and R to complete assignments.
Prerequisites:	Undergraduate statistics
Textbook/ Other reading:	There is no official textbook for this course. Reading material will be made available through Canvas for specific lectures.
Course notes:	Lecture notes and data sets will be made available through Canvas.
Data packages:	Students should download R and Matlab to complete their assignments

Course Objectives

Students that complete this course should be able to:

1) Select appropriate techniques for visualizing data of different forms and create high resolution graphics in Matlab or R.
2) Assess statistical confidence about means and slopes using nonparametric bootstrap approaches.
3) Use multivariate statistics like PCA to identify and extract dominant patterns from data or generate composite variables for further evaluation in regression or structural equation models.
4) Identify common assumptions and limitations of multivariate statistical analyses including PCA and CA
5) Understand the difference between cluster analysis, classification/regression trees, and PCA and when to use each approach

Topics and Reading Assignments (*subject to modification*)

MODULE 1 - Lectures 1-4: Introduction to Matlab and R (data curation, visualization, and graphics)

- *Lecture 1:* Introduction to the Course and software packages (description of packages, how to download Matlab and R, package interface, loading data, basic plotting)
- *Lecture 2:* In-class coding exercise: loading data, defining variables, making/exporting basic plots
- *Lecture 3:* Data curation: What is it, why is it necessary and common coding tools (for loops, if/else statements, working with NaN's and different data formats)
- *Lecture 4:* In-class data curation exercise
- *Lecture 5*: using graphs to effectively convey results (different plot forms, error and outliers, appropriate axis scaling)
- *Lecture 6*: In-class graphing exercise

MODULE 2 - Lectures 7-10: Overview of standard distributions, traditional methods for estimating confidence intervals, and parametric and nonparametric methods for estimating confidence intervals

- *Lecture 7:* Standard distributions and their use to represent data (mean, median, mode, and confidence intervals
- *Lecture 8:* In-class coding exercise – representing data using parametric distributions
- *Lecture 9:* Bootstrapping statistical confidence
- *Lecture 10:* In-class coding exercise – application of bootstrapping to assess statistical confidence

MODULE 3 - Lectures 11-16: Linear regression, multiple linear regression, & generalized linear models

- *Lecture 11*: Linear regression (mean squared error, maximum likelihood, and bootstrapped CI about the slope of a best-fit line)
- *Lecture 12*: In-class coding exercise – linear regression using MSE and ML
- *Lecture 13*: MLR and AICc
- *Lecture 14*: In-class coding exercise – performing multiple linear regression and identifying the most parsimonious model
- *Lecture 15*: GLM
- *Lecture 16*: In-class coding exercise – evaluating GLMs using R

MODULE 4 – Lectures 17-20: Multivariate Statistics Part 1: Cluster Analysis and Classification/Regression Trees

- *Lecture 17:* Hierarchical agglomerative clustering
- *Lecture 18:* In-class coding exercise – clustering variables using average Euclidian distance
- *Lecture 19:* Classification/Regression trees and cross validation
- *Lecture 20:* In-class coding exercise – implementing a classification tree

MODULE 5 - Lectures 21-27: Multivariate Statistics Part 2: Ordination methods

- *Lecture 21*: Principal component analysis (PCA), Correspondence analysis (CA), and biplots
- *Lecture 22*: In-class coding exercise – practice performing PCA
- *Lecture 23*: Stopping Rules and Bootstrapped Confidence Bounds for PCA
- *Lecture 24*: In-class coding exercise – coding a resampling-based stopping rule
- *Lecture 25*: Identifying Drivers part 1: Regressing Variables in PC Space, and Constrained Ordinal Space (CCA)
- *Lecture 26*: In-class coding exercise – identifying drivers through regressing variables in PC space
- *Lecture 27*: Identifying Drivers part 2: Evaluating composite variables using MLR and Structural Equation Models

Lecture 28 and 29: Final Presentations

GRADING POLICIES

Course Grade:

Problem Sets (5)..... 75%

Final Project..... ..... 25%

Research question (2%)

Preliminary statistical analyses (5%)

Final presentation (15%)

Graded rubrics for 2 other presenters (3%)

Classroom Participation.....up to 5% extra credit

Problem Sets

Five problem sets that require applying the data analysis techniques covered in class will be assigned for the course. Problem sets (and due dates) will be available on Canvas. Completed assignments (and any associated R or Matlab codes) should be turned in online through Canvas. Assignments are due before midnight on their due date. Late assignments will be penalized 1/2 letter grade for each day they are late.

You may work with other students on your problem sets. However, each student must prepare their own codes, graphs and write-ups. You may not copy another student's work. Duplicate assignments will result in an F for that assignment for both parties.

Final Project

The final project will be a **6-minute** presentation of a research project requiring data analysis that is of interest to you. You may use data from your personal research or one of the datasets used in class. You will need to 1) articulate your research question and why it is of interest (4-5 references from the literature are required to support background information), 2) describe your dataset and the statistical techniques you used to answer your question (you should use at least 2 data analysis techniques presented in class) 3) describe your results, referencing appropriate figures (minimum of 2), and 4) briefly discuss the implications of your results (why are they interesting to your field? What do they tell us that we didn't know before?). There will be 2 minutes for questions at the end of your presentation. You will be stopped at the 6 minute mark so please keep track of your time. Presentations that go over will lose points

You will be required to articulate your research question by lecture 16 (Oct 15) and submit preliminary statistical analyses for review by lecture 24 (Nov 12). These submissions count for 7% of your grade.

Final presentations will occur on the last two days of class (Dec 3th and 8th). You will need to turn a copy of your talk in to me by 12:00 midnight the day before your presentation. The final presentation is worth 15% of your grade.

In addition to presenting your own talk you will be in charge of grading two others using a rubric I provide. Please take this seriously as your feedback will be provided to the students you review. Thoughtful completion of these rubrics is worth 3% of your grade.

COURSE POLICIES

COVID-19: This course will meet synchronously on-line every Tu/Th. This is not an in-person or hybrid class so there are not specific personal protective equipment (PPE) requirements for attendance. This said, appropriate out-of-class behaviors including social distancing and washing hands frequently and properly are strongly encouraged. These measures are stated in Virginia Tech's full list of wellness principles which can be found at: https://vt.edu/content/dam/vt_edu/covid-19/ready/wellness-commitment-

8.5x11_VT.pdf. Please do not meet in person to work on class assignments or to study course material. I strongly encourage meeting over Zoom instead to reduce the risk of transmitting COVID-19.

Principles of Community: The Virginia Tech Principles of Community are intended to increase access and inclusion and to create a community that nurtures learning and growth for all of its members. They are defined at: inclusive.vt.edu

Honor Code: All students must adhere to the Honor Code Policies of Virginia Tech. For information about the Graduate Honor System of Virginia tech, please visit graduateschool.vt.edu/academics/expectations/graduate-honor-system.html. Any suspected violations of the Honor Code (plagiarizing published work, copying another student's work, cheating on exams, etc) will be promptly reported to the honor system. Honesty in your academic work will develop into professional integrity. The faculty and students of Virginia Tech will not tolerate any form of academic dishonesty.

Attendance Policy: Daily online attendance is expected. Participation during lecture is encouraged, and incentivized. Completing and turning in in-class coding exercises is worth up to 5% extra course credit.

Accommodations: Students are encouraged to address any special needs or accommodations with me during the first two weeks of the semester, or as soon as you become aware of your needs. Those seeking accommodations based on disabilities are required to obtain a Faculty Letter from the Services for Students with Disabilities office in Lavery Hall (www.ssd.vt.edu/).

If you have emergency medical information to share with me, please make an appointment with me during the first two weeks of the semester or as soon thereafter as you become aware of the need.