Geochemical Statistics and Data Analysis
Spring 2025

GEOL 4315 (CRN 27128), GEOL 5315 (CRN 27129), GEOL 6315 (27130)
MW 1:30-2:50PM
Geology Building, Room 302

Instructor: Dr. Mark Engle (he/him) maengle@utep.edu - UTEP email is my preferred communication; please don't message through Blackboard
Office: Geology Building 302A/101
Office hours: Monday, 8:00 - 9:00 AM in my office. Otherwise, email me.

Course Description
This course introduces types of geochemical data (concentrations, spectral, time series, spatial etc.) and provides statistical and data analysis tools to interrogate and model these types of datasets. Particular focus will be paid to common issues with geochemical data including handing of censored (e.g., less than detection limits) and missing data and identification of compositional data and application of proper analysis techniques. Relevant software will be explored (e.g. CoDaWork and SGemS), but the focus will be on the R, a statistical programming language.

Learning outcomes
At the end of the course, students will:
- Understand various types of data and what kinds of approaches should be applied
- Strong knowledge of compositional data
- Know how to organize and import data into R
- Manipulate and adjust data using dplyr and other packages in R
- Perform and become familiar with basic univariate, bivariate, and multivariate statistics, exploratory data analysis, and modeling
- Be able to compile corresponding information into a report and/or presentation

Required textbook
Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., and Gilroy, E.J., 2020, Statistical methods in water resources: U.S. Geological Survey Techniques and Methods, book 4, chap. A3, 458 p., https://doi.org/10.3133/tm4a3. (You can download a pdf version **for free** or purchase a hard copy for about $20).

Recommended textbooks
Filzmoser, P., Hron, K., and Templ, M. *Applied Compositional Data Analysis with Worked Examples in R.* Springer Series in Statistics, Springer Nature Switzerland AG, Cham, Switzerland, 2018. ISBN: 978-3-319-96422-5.

Olea, R.A., 2018, A practical primer on geostatistics (ver. 1.4, December 2018): U.S. Geological Survey Open-File Report 2009–1103, 346 p., https://doi.org/10.3133/ofr20091103 (You can download a pdf version **for free)**

Class Project: Each person will be responsible for presenting his or her own individual data analysis of a data analysis of interest. This data set has the following requirements: > 30 data points, more samples than elements, location data. It can be your thesis data but doesn't have to be. Selection and discussion of your data set will be completed by the due date listed below.

Class project required deadlines (full details to be described in blackboard):
1) Monday of Week 4: Present research data set and question
2) Monday of Week 7: Results from univariate exploratory data analysis
3) Monday of Week 10: Results from bivariate exploratory data analysis
4) Monday of Week 13: Results from multivariate exploratory data analysis
5) Last day of class: Final Report due
6) Final: Class presentation

Structure of Class:
Monday: discuss readings + lecture
Wednesday, class activity on computer: Statistical programming with R-Studio, SGemS, or CoDaWork

Course Policies: (Grading, attendance, academic integrity, etc.)

Grading:
20% - Class participation
30% - Assignments
10% - In-class quizzes
20% - final presentation
20% - final report

Grading scale: A-100-85%; B- 84-75%; C-74-65%, D- 64-55%; F-below 55%.

Late Assignments, etc.: Assignments will be reduced 10% of the maximum assignment value per 24-hour period they are late. They will not be accepted 5 days beyond the due date. Late quizzes or presentations will not be accepted without prior approval.

Attendance: Attendance, which is key to success in this class, is mandatory and reflected in the course grading. If you have a legitimate reason for missing class (e.g., medical appointment or illness, military service, or official University activities), please notify the instructor prior to the absence (as possible) and provide documentation. In the case of poor class attendance, students may be dropped from the class.

Getting dropped from class: If you are severely failing the class (grade <55%) at the Drop Date or miss 3 or more assignments/lectures in a row, I may drop you from the course. If you would like to remain in the class and are under one or more of those conditions, please contact me prior to the drop date.

Academic Integrity: Academic dishonesty will be not tolerated in this class (please refer to the student conduct code handbook for details regarding university policy and definitions).

Dishonesty includes, but is not limited to, plagiarism on term papers, unauthorized notes brought into an exam; copying answers from another student or letting another student copy your answers. Academic dishonesty will be reported to the UTEP Office of Community Standards.

Graduate student expectations/requirements: Graduate students enrolled in the course must complete all work listed in this document. In addition, M.S. students in course must include at least 1 kind of statistical of mathematical modeling in their final project. Ph.D. students in the 6000-level of this course are required to explore a data analysis method not covered in this course, use that method in their final project, and provide a presentation to the rest of the class on this method.

Course Statements and Resources:

Students with Disabilities: If you have a disability or if you are experiencing learning difficulties, please contact the Center for Accommodations and Support Services (CASS). You may contact them Monday through Friday 8:00a.m.-5:00p.m. Phone:(915) 747-5148. Union Building East Room 106 cass@utep.edu. They provide any necessary accommodations. You should also meet with me in order to facilitate your needs. You are expected to provide documentation of your disability in order to make special arrangements in this class.

Student Success Help Desk (SSHD): Students experiencing challenges or obstacles to academic success including registration, financial, food, housing, and transposition resources my submit a ticket request assistance to studentsuccess@utep.edu

Military Student Success Center: Assists personnel in any branch of service to reach their educational goals.

Counseling and Psychological Services: Provides a variety of counseling services including individual, couples, and group sessions as well as career and disability assessments.

UTEP Food Pantry: Non-perishable food items are available to students who are currently enrolled in classes. Bring a Miner Gold Card to Memorial Gym, Room 105, Monday through Friday, 10 a.m. to 2 p.m.

Tentative Schedule: (List of topics to be covered by specified timeline. Special target deadlines, such as final project deadlines, last day to withdraw without penalty, and date and time of final presentation also shown.)

Week	Lecture	Lab	Notes	Reading
1 (Jan22)	Course Overview	None		None
2 (Jan 27, 29)	Scales of measure, sig digs, notation	Spreadsheet basics; data preprocessing		None

3 (Feb 3, 5)	QA/QC and data preparation	Intro to R		Geboy and Engle, 2011
4 (Feb 10,12)	Summarizing univariate data	dplyr & univariate statistics	Deadline 1 due on Monday	Helsel et al., Ch. 1
5 (Feb 17, 19)	Univariate EDA	univariate EDA plotting		Helsel et al., Sec. 2.1, 2.2
6 (Feb 24,26)	Bivariate EDA and mapping	bivariate EDA in R		Helsel et al., Sec. 2.3, 2.4.4
7 (Mar 3,5)	Data Transformations and Hypothesis Testing	Transformations and Statistical Testing in R	Deadline 2 due on Monday	Helsel et al., Ch. 3,4
8 (Mar 17,19)	Robust and traditional methods for correlation	Correlation in R		Helsel et al., Ch. 8
March 10-14 Spring Break				
9 (Mar 24, 26)	Compositional Data Analysis 1	CoDA univariate and bivariate EDA in R		Geboy et al., 2013
10 (Mar 31, Apr 2)	Regression/Censored data	Regression and dealing with censored data in R	Deadline 3 due on Monday; April 4: Drop day	Helsel et al., Ch. 9, 10
11 (Apr 7,9)	Multivariate EDA	PCA, cluster analysis, etc. in R		To be assigned
12 (Apr 14,16)	Compositional Data Analysis 2	CoDA Multivariate EDA		Daunis-i-Estadella et al., 2006
13 (Apr 21, 23)	Spatial statistics	Kriging in R	Deadline 4 due on Monday	Srivastava, 2013
14 (Apr 28, 30)	Multivariate modeling	Multivariate Regression and machine learning prediction in R		To be assigned
15 (May 5,7)	Compositional Data Analysis	Linear modeling with CoDa	Final report due on Wednesday	To be assigned
Final Presentations: Wednesday, May 14th from 4:00 PM-6:45 PM				

Suggested and Assigned Readings:

Astel, A., S. Tsakovski, P. Barbieri, and V. Simeonov (2007), Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets, Water Res., 41(19), 4566–4578.

Geboy, N. J., and M. A. Engle (2011), Quality Assurance and Quality Control of Geochemical Data: A Primer for the Research Scientist, U.S. Geological Survey Open-File Report 2011-1187, 28 p.

Filzmoser, P., K. Hron, and C. Reimann (2009), Univariate statistical analysis of environmental (compositional) data: Problems and possibilities, Sci. Total Environ., 407(23), 6100–6108.

Helsel, D. R. (2005), More than obvious: Better methods for interpreting nondetect data, Environ. Sci. Technol., 39(20), 419A–423A.

Filzmoser, P., R. Garrett, and C. Reimann (2005), Multivariate outlier detection in exploration geochemistry, Comput. Geosci., 31(5), 579–587.

Filzmoser, P., K. Hron, and C. Reimann (2009a), Univariate statistical analysis of environmental (compositional) data: Problems and possibilities, Sci. Total Environ., 407(23), 6100–6108.

Filzmoser, P., K. Hron, C. Reimann, and R. Garrett (2009b), Robust factor analysis for compositional data, Comput. Geosci., 35(9), 1854–1861.

Filzmoser, P., K. Hron, and C. Reimann (2010), The bivariate statistical analysis of environmental (compositional) data, Sci. Total Environ., 408(19), 4230–4238.

Pawlowsky-Glahn, V., and J. J. Egozcue (2006), Compositional data and their analysis: an introduction, Geological Society, London, Special Publications, 264(1), 1–10.

Puig, R., R. Tolosana-Delgado, N. Otero, and A. Folch (2011), Mixing isotopic and compositional data: A discrimination of regions prone to nitrate pollution, in Compositional Data Analysis: Theory and Applications, edited by V. Pawlowsky-Glahn and A. Buccianti, pp. 302–317, Wiley.

Reimann, C., and R. G. Garrett (2005), Geochemical background—concept and reality, Sci. Total Environ., 350(1–3), 12–27.

Srivastava, R. M. (2013), Geostatistics: a toolkit for data analysis, spatial prediction and risk management in the coal industry, Int. J. Coal Geol., 112, 2-13.

Templ, M., P. Filzmoser, and C. Reimann (2008), Cluster analysis applied to regional geochemical data: Problems and possibilities, Appl. Geochem., 23(8), 2198–2213.

Tolosana-Delgado, R., and H. von Eynatten (2010), Simplifying compositional multiple regression: Application to grain size controls on sediment geochemistry, Comput. Geosci., 36(5), 577–589.