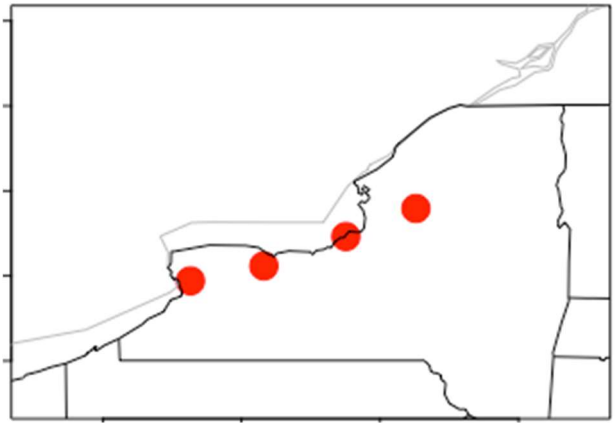
BEE 4310/6310: Multivariate Statistics for Environmental Applications
Assignment #3 (10 points)

Include responses to everything below in bold, and make sure your final assignment is well organized (i.e., numbered responses, full sentences, code attached at the end). This makes it easier to grade (and easier to give partial credit).

In this exercise, you will explore and model how annual precipitation totals co-vary across 4 rain gages along the southern shore of Lake Ontario in New York State. Annual precipitation in inches between 1980 and 2010 for each of these gages is provided in the file "Ontario.prcp" on Canvas under the folder Assignment 3. Note that some of these data are missing.



1. Assess Normality

a) Create a matrix (annual.prcp) only containing the precipitation data at the four gages (not the years). Convert this matrix from a data.frame object (what read.table() returns) into a data.matrix object. Data matrices are better suited for matrix multiplication (which you will use below).

Check Marginal Distributions:

b) **(1 pt) In a 2x2 panel figure, plot probability plots of the data for all four gages. Comment on whether the marginal distributions of these data appear normal.** You can use the pnorm() function to calculate the cdf of the normal distribution. You can use the sort() function to order a vector of data from smallest to largest, and set the argument na.last=NA to drop the NAs.

Check Multivariate Structure

c) Calculate the covariance matrix of the precipitation data matrix using the cov() function. It is important that you set use= "pairwise.complete" in the function to ignore missing values. Save this covariance matrix in a variable S. Also calculate the mean vector mu using the apply function, making sure to set na.rm=T to remove the NAs from the calculation of the mean.

d) Write a for loop to calculate the Manahalobis distance between each of the 31 observations and the mean vector. Save these 31 values in a vector D.sq.

e) **(1 pt) Create a probability plot for the vector D.sq.** You can use the pchisq() function to calculate the cdf of the chi-squared distribution.

f) **(1 pt) Based on your results from both parts above, please comment on the use of the multivariate normal distribution to model the stochastic behavior of annual precipitation at these 4 gages.**

2. Gap-Filling Missing Data (Imputation)

g) **(4 pts)** Use the property of conditional normality for multivariate normal random variables to estimate precipitation data for each year and gage with missing data using all available observations for that year. **Report estimates of the conditional mean and 95% confidence bounds for each missing data point (8 in total).** Cooperation is strongly encouraged for this problem.

h) Create a new matrix of gap-filled data for all 31 years, called annual.prcp.fill, replacing each missing data point with its conditional mean estimate.

---

3. Multivariate Inference

Finally, we will determine whether there is evidence that mean precipitation at gages along the southern shore of Lake Ontario has shifted over the past 31 years.

i) **(1 pt) First, plot the gap-filled data for each gage through time. Also conduct separate regressions against years for each gage, and report the p-value for the coefficient on year for each gage. Comment on the trends you see.**

j) Next, partition the data into two segments, one composed of the first 15 years of data, and the second composed of the next 16 years of data.

k) Calculate the multivariate mean for each of these subsets.

l) Calculate the covariance matrix for each of these subsets.

m) Estimate the pooled covariance matrix based on the two covariance matrices above.

n) **(1 pt) Calculate and report the Hoteling-T statistic that compares the two means of these different subsets. Then calculate and report the F-statistic used in the F-test to compare these mean vectors (a scaled version the Hoteling-T statistic).**

o) **(1 pt)** Compare this statistic to a 95% critical value of the F-distribution with appropriate degrees of freedom. **Comment on whether there is evidence to suggest that the vector of means of annual precipitation at the four gages has changed across the southern shore of Lake Ontario. How does this compare against the 4 individual trend tests you conducted above?**