BEE 4310/6310: Multivariate Statistics for Environmental Applications
Assignment #4 (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, we will explore the major modes of variability in a reconstruction of the Palmer Drought Severity Index (PDSI) across North America. The PDSI, originally developed by Palmer [1965], is a monthly measure of the balance between atmospheric moisture supply and demand and takes into account regional climatology and short-term hydrologic persistence. The Living Blended Drought Atlas (LBDA) is a paleoclimate reconstruction of average summer (June–August) PDSI over North America based on 1845 tree ring chronologies (see Figure 1) that is spatially complete over North America from 1473 to 2005 [Cook et al., 2010]. The LBDA is available in the RData file "LBDA.RData", which holds the matrix LBDA. This matrix has 4968 columns, associated with grid cells over North America, and 533 rows, each associated with a year. The longitude and latitude values associated with the columns of LBDA can be found in lon.lat.RData.

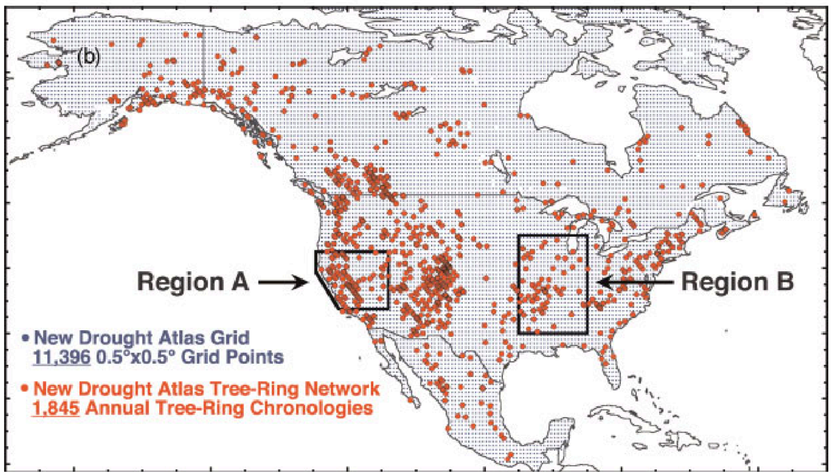


**Figure 1. The LBDA drought grid and tree ring network. From Cook et al., 2010 (Region A and Region B were focal areas in that study).**

a) Load in the LBDA.RData and lon.lat.RData files using the load() function. You can then run ls() in the console to see the variables that have been loaded into the R working environment.

1. Conduct a Principal Component Analysis

b) Conduct a PCA on the LBDA matrix using the prcomp() function and the correlation matrix. As part of this exercise, you should read through the documentation for this function and determine how to return the following:
   - The matrix of eigenvectors
   - The principal components
   - The eigenvalues (read carefully for this one)

c) **(1 pt) Calculate and report the variance explained by each of the first 10 EOFs.**

d) **(2 pts) Plot a scree plot and +/- 1 standard error for the first 10 eigenvalues. Be sure to label your axes. Use the shape of this plot and North's rule of thumb to determine how many EOFs to retain for further analysis. Briefly justify your choice.**

e) To visualize the spatial patterns associated with each PC you retain above, you will plot the loading patterns on a map of North America. You can use a relatively simple approach that colors each grid cell based on its loading value (note, you need to install and load the "maps" library in R). Before creating this map, you need to make this vector of colors, which can be done in 4 steps:

- First, create a variable mybreaks that defines a set of intervals:

  mybreaks <- c(seq(-0.04,-0.001,by=0.004),seq(0.001,0.04,by=0.004))

  (Note: the upper and lower bounds of these intervals, 0.04 and -0.04, were chosen based on the range of the loadings in the EOFs)

- Next, assign each of the elements of an EOF to one of the intervals

  mycut <- cut(cur_EOF,breaks=mybreaks,label=FALSE)

- Next, define a color pallete

  mycolpal <- colorRampPalette(c("red","white","blue"))

- Finally, use the pallete to assign a color to each EOF element based on its interval

  mycol <- mycolpal(length(mybreaks))[mycut]

f) **(1 pt) In a multi-panel figure, plot each EOF using the plot() function, the information in lon.lat, and the vector of colors in mycol. Add a map of countries by adding map("world",add=T) in the line following your plot() function call. You should have one map for each retained EOF.**

---

## 2. Create a Parsimonious Forecast Model

g) You can use the first several PCs that are retained to create a parsimonious model to forecast next year's summer PDSI (i.e., PDSI in 2006). You can do this by building simple time series models for each PC, forecasting each PC for the next year, and then synthesizing the original PDSI data using these PC forecasts.

First, for each of the PCs that you retain, build a simple autoregressive lag-1 time series model of that PC:

$$y_t = \alpha y_{t-1} + \varepsilon_t$$

This model suggests that there is memory in each PC, such that if a PC is above (or below) its mean in time t-1, it is more likely to be above (or below) its mean in time t (assuming α is positive).

To fit this model, you can fit a simple linear regression model between lead and lag versions of a given PC. For instance, for PC1, you can regress PC1[2:n] against PC1[1:(n-1)] to estimate $\alpha$ in the regression. Estimate separate models for each of the retained PCs.

h) Create a forecast for each of the retained PCs for the time step n+1 (i.e., the year 2006) using your fitted models and the value for each PC at time n. Here, n in the most recent observation in the dataset (i.e., the year 2005).

i) Place the forecasts for each PC into a vector. Then, using the synthesis equation and the vector of 1-step ahead forecasts, develop a 1-step ahead forecast for the PDSI for all of the grid cells.

j) **(3 pts) Using a similar approach as in (e) and (f), plot your forecast of the PDSI across North America. Have mybreaks range from -1 to 1 (since the PDSI is standardized, this indicated +/- 1 standard deviation). What does your forecast suggest about the PDSI in 2006?**

3. Rotate the EOFs and Derive Rotated Principal Components

k) You will now compare the EOFs to rotated EOFs (REOFs) using the varimax rotation. For consistency, you will rotate the EOFs that were retained in the previous analysis. To rotate the *M* EOFs you retained, you can use the varimax() function.

```
my.varimax <- varimax(EOFs[,1:M])
```

l) Calculate the rotated PCs by first rotating the original M EOFs by the rotation matrix returned by the varimax() function, and then multiplying the data LBDA by the new REOFs. This is a simple calculation in R using %*% for matrix multiplication. Be sure to order this matrix multiplication correctly.

m) **(1 pts) Report, compare, and discuss the correlation matrix of the rotated PCs with the correlation matrix of the original PC's. What is the largest difference in the correlation structure of the original PCs and the rotated PCs?**

n) **(2 pts) Similar to the analysis with the original EOFs, plot each rotated EOF on a map. Discuss the differences between these plots and those for the original EOFs.**

o) For those who are interested, read the Hannachi et al., [2007] paper that provides a good review of these techniques. This is a good paper to keep as a reference for future work.

Reference

Cook, E. R., R. Seager, R. R. Heim, R. S. Vose, C. Herweijer, and C. Woodhouse (2010), Megadroughts in North America: Placing IPCC projec- tions of hydroclimatic change in a long-term palaeoclimate context, J. Quat. Sci., 25(1), 48–61, doi:10.1002/jqs.1303.

Hannachi, A., Jolliffe, I.T., and Stephenson, D.B. (2007), Empirical orthogonal functions and related techniques in atmospheric science: A review, International Journal of Climatology, 27, 1119-1152.

Palmer, W. C. (1965), Meteorological drought, Weather Bur. Res. Pap. 45, pp. 58, U.S. Dep. of Commer., Washington, D. C