OCNG 655 Experimental Design and Analysis in Oceanography
Fall2020
Instructor: Dr. Darren Henrichs

Homework Problem Set 2

Homework 2 consists of 4 problems. Please provide all graphics, codes, and worksheets that you used to do your assignment. Be sure that your name, Course number, UIN, Assignment number, problem number, and date are all ALL sheets that are turned in. You may work together to discuss the problems, but whatever you turn in MUST be your own work.

Total: 100 points

1. (50 points) The Palmer Drought Severity Index (PDSI) is a very common indicator for drought severity in the US. The National Oceanic and Atmospheric Administration (NOAA) posts these data on the web.

   https://www.ncdc.noaa.gov/temp-and-precip/drought/historical-palmers/

   a. (3 points) In your own words describe what the CMI, PDSI and SPI are and how they are determined and how they are different. (Use: https://www.ncdc.noaa.gov/sotc/drought/201403)

   b. (3 points) Go to the National Climate Data Center Online site:
   http://www7.ncdc.noaa.gov/CDO/cdo

   Click on the "Data Set/Product" link.
   Select "Climate Indices", then click "Access Data/Products"
   ***It's an older page so the icon images appear broken but the data access is still available.
   First, under the "Region" tab, select the "South" (the region that includes Texas). Retrieve **ten** years of data by selecting the Start and End periods; use 1/2006 through 12/2015. Use the "Text Output, Space Delimited" option. You can preview the data by selecting "Static Graphs" and the Palmer Drought Severity Index option. For the text output, after clicking on "Submit", you will be directed to a webpage containing a file ending with a '*.txt' extension. Download the data to your computer either by right clicking the file link or by clicking the link and copy/paste to a file. Name the file: "south_2006_2015.txt". There is also a document file on the website (just under the .txt filename) called "Data format documentation"; read it. It will give you valuable information about the data file.

   Next, click on the "State" tab, select "Texas" and retrieve ten years of data as above. Name the data file "texas_2006_2015.txt".

Lastly, click the “Division” tab, select “South Central” and again retrieve ten years of data as above. Name this file “div_bryan_2006_2015.txt. This file contains the data for the Brazos Valley (College Station area).

You should now have three text files with data: South Region, Texas State, South-Central Division. The files should look like the following (if you opened it in WordPad, Word, TextEdit, Notepad, Notepad++ or similar program):

texas_2006_2015.txt

```
StateCode Division YearMonth     PCP     TMP    PDSI    PHDI    ZNDX    PMDI CDD HDD    SP01    SP02    SP03    SP06    SP09    SP12    SP24
       41       00    200401    2.18    48.9   -1.51   -1.51    -.33   -1.51  19 446     .73    -.28    -.38     .03    -.08    -.31     .13
41       00    200402    2.92    47.4      .5    -.85    1.49     .02   7 434    1.38    1.25     .45     .43      .6    -.21     .34      41
00    200403    2.34    61.9     .52    -.69     .21      .3 100 142     .68    1.27     1.2      .3      .4    -.01     .37      41      00
200404       4    64.4    1.29    1.29    2.49    1.29 116  60    1.23    1.19    1.46     .73     .74     .56     .62      41      00
200405    2.44    73.7     .77     .77   -1.18    -.01 299   3     -.6     .43      .6     .62     .55     .72      .7      41      00
200406    6.51    78.7    2.26    2.26    4.71    2.26 429   0    1.99    1.19    1.47    1.63     .99    1.02    1.14      41      00
200407    2.44    80.7    2.81    2.81    2.34    2.81 530   0     .18     1.6    1.04     1.5    1.04     .98     .89      41      00
200408     3.2    79.2    3.41    3.41    2.69    3.41 496   0     .86      .6    1.81    1.46    1.31    1.12    1.02      41      00
200409    2.52      75    2.92    2.92    -.42    2.69 393   1    -.21     .32     .23    1.21    1.43     .92     .99      41      00
200410    4.07    69.4    3.12    3.12    1.51    3.12 276  20     .93     .53     .72    1.14    1.54    1.17     .74
       41       00    200411    6.44    55.6    5.24    5.24    7.31    5.24  46 225     2.6    2.18    1.79    2.43    2.04    1.97    1.26
```

c. (4 points) Using Python, plot the PDSI (column 6) versus date. Remember to label your plots and include a title (be sure to include in the title where the data came from, i.e., region, division, date. If you use Python, your code should look **something** like:

```
fname = 'path/to/file/texas_2006_2015.txt'
t_data = pd.read_csv(fname, header=0, delim_whitespace=True)
t_data.PDSI.plot()
plt.xlabel('Month from Jan 2006')
plt.ylabel('PDSI')
plt.title('PDSI: Jan 2006 - Dec 2015')
plt.savefig('path/where/to/save/file/hw2_1c.pdf')
```

NOTE: Count the columns of the data file. Does it match what was read?

Plot the PDSI for all three files in one graph. Use color or symbols to distinguish the different files, use **legend** to label the different datasets on the graph.

```
t_data.PDSI.plot()
r_data.PDSI.plot()
s_data.PDSI.plot()
plt.xlabel('Month from Jan 2006')
plt.ylabel('PDSI')
plt.title('PDSI: Jan 2006 - Dec 2015')
plt.legend()
plt.savefig('path/where/to/save/file/hw2_1c.pdf')
```

Note: you will have to use the **read_csv** command three times, i.e., once for each file. Use a different variable name each time and make sure you change the variables names above to match whatever you chose. You will have to modify the **read_csv** command so that each file will have different variable names. In the example code, the Texas data was called t_data.

d. (3 points) Use the **hist** command to create a histogram of the PDSI with 10 bins: **hist(**PDSI,10**)**. Save the histogram and describe the distribution for each of the three data files.
   Close the previous figure before plotting this histogram.

```
t_data.PDSI.hist(bins=10, grid=False)
plt.savefig('path/where/to/save/file/hw2_1d.pdf')
```

e. (3 points) For each data set, use the **std**, **mean**, **mode**, and **median** command to get estimates of the standard deviation, average, mode, and median values of the PDSI, temperature (TMP), and precipitation (PCP). List these values in a table (use Excel or a word processor to make the table). Discuss the relative magnitudes of each of the parameters, how does it relate to the distribution? Remember these functions from HW1? The only new one should be mode. The nice thing about the pandas library in Python is that it has a lot of these functions built-in! We saw this with plotting above.

```
t_data.PDSI.mean()
```

f. (3 points) Based on the values shown in the table created in 1e, is your distribution normal or not normal? Defend your answer statistically (use the results of part 1e).

g. (3 points) Describe the plot generated in 1c. Was there excessive, mild, or severe drought during the chosen time period in the region? In Texas? In the South-Central Division? How long did the drought last?

h. (4 points) Plot the South region PDSI (x-axis) versus the Texas State PDSI (y-axis).

```
plt.scatter(r_data.PDSI, t_data.PDSI); # variable r_data has South Region data
```

Don't forget to change the variable names to whatever you called your region data and state data.

Label the graph appropriately. Describe the relationship, if any, between the South and Texas PDSI data sets. **Hold** the plot and add a second **scatter** plot for the South PDSI (s6) versus the South-Central Division PDSI (b6).

```
plt.scatter(south.PDSI, t_data.PDSI, color='red')
```

Describe the relationship between the variables shown in this graph.

i. (4 points) Download a fourth file from the NOAA webpage. This time download ten years of data from the "Central Region". Load this file in by modifying the code from earlier. **Scatter** plot the Central Region PDSI versus the South Region PDSI. Describe the relationship between the two variables.

j. (10 points) We will now quantify the relationship between the PDSI's using the **corrcoef** command. Use Python help to see how this command is used.

Compare the South PDSI to the Texas PDSI using corrcoef().

```
np.corrcoef(s_data.PDSI,t_data.PDSI)
```

```
alternatively:
```

```
s_data.PDSI.corr(t_data.PDSI)
```

Compare the Texas PDSI to the South-Central Division PDSI using corrcoef(t6,b6).

```
np.corrcoef(t_data.PDSI,r_data.PDSI)
```

Compare the South PDSI to the Central Region PDSI using corrcoef().

Describe the outputs of the **corrcoef** command in the three cases above. Relate the correlation coefficient to the scatter plots of 1h and 1i.

k. (10 points) **Scatter** plot the Cooling Degree Days (CDD) versus Heating Degree Days (HDD) for the Texas data set. Use appropriate labels and title. Is the relationship linear or something else? Compare the CDD and HDD using **corrcoef**. Is **corrcoef** useful to quantify the relationship between these two variables?

Using **corrcoef**, estimate the correlations between PDSI and PCP, PDSI and PHDI, PDSI and ZNDX for the Texas data and for the Central Region data. Make a table of these correlation coefficients and describe the results found in this table, i.e., explain which variables are correlated and which are not.

2. (20 points) This question entails data collected on the Texas Continental Shelf during Hurricane Ike in 2008. Go to the TABS website http://tabs.gerg.tamu.edu/Tglo/

Select one of the buoys by clicking on the map or the buoy letter on the right hand side of the page. Select the "Search TABS Database" on the left hand side of the page. Use the data query to select 30 days of velocity data in "Data Table" format. Be sure that your selection includes 10-18 September 2008.

a. (3 points) Download the selected data to your computer.

b. (3 points) Load the data into Python (or similar program) and plot east velocity versus date. Overplot north velocity (i.e. plot two lines on same plot). Label x and y axes and provide a title. Describe the plot.

The code should look <u>something</u> like this:

```
fname = 'path/to/your/file/TABSdata.txt'
tabs = pd.read_csv(fname, header=None, skiprows=3, delim_whitespace=True)
tabs.columns = ['Date', 'Time', 'East', 'North', 'Speed', 'Direction',
'WaterTemp']
tabs.index = pd.to_datetime(tabs.Date + ' ' + tabs.Time)
tabs.East.plot()
tabs.North.plot()
```

c. (3 points) Plot temperature in the same way as 2.b. Describe the plot.

d. (3 points) Plot speed versus date (as in 2.b). Speed is the square-root of the sum of the squares of the individual velocity components.

```
spd = np.sqrt(tabs.East**2 + tabs.North**2)
```

e. (4 points) Use the **hist** command to create a histogram of the speed with 100 bins: *hist()*. Print the histogram and describe the distribution for your data.

```
spd.hist(bins=100)
```

f. (4 points) Create a table of basic statistics (mean, mode, median, maximum, minimum, standard deviation) of speed, east velocity, and north velocity. Describe how the basic statistics of each parameter compare. What is largest speed observed at this buoy?

3. (20 points) Download data from two (2) TABS buoy using the procedure used above in question 2. Download 60 days of data from the time period between September 2015 and September 2019. Choose any two buoys but be certain the data exists for the full duration of your chosen times
    a. (5 points) Plot the calculated speed from the two buoys on the same graph. Describe any similarities or differences between the two series.

```
buoy1.Speed.plot()
buoy2.Speed.plot()
```

Note: the data from the two buoys have been placed in separate variables in the code above: e.g., *buoy1* is from the first buoy and *buoy2* is for the second buoy.

    b. (5 points) Plot the histogram of the speed from the two buoys. Describe any similarities or differences between these two histograms.

    c. (5 points) Create a table of basic statistics (mean, mode, median, maximum, minimum, standard deviation) of speed, east velocity, and north velocity. Be sure the table is formatted legibly, with clear descriptions of rows and columns. Describe how the basic statistics of each parameter compare to those found in 2.f. What is largest speed observed at these buoys, i.e., use the **max** command?

```
spd.max() #replace spd with whatever your variable is named
```

d. (5 points) Use the **corrcoef** to estimate the correlation between velocity and temperatures from each of the two buoys as given below. Report the coefficients in a table. Calculate and discuss the correlation coefficient for:
    i. temperature from the two buoys;
    ii. the *u* (east) velocities from the two buoys;
    iii. the *v* (north) velocities from the two buoys;
    iv. the *u* and *v* each buoy pair; and
    v. temperature and *u* velocity from each buoy pair.
    vi. Verify and comment on the correlation coefficients by creating a **scatter** plot for each of the pairs in 3.d.*i*-*v* above. Use the subplot command so that the seven scatter plots are all on the same page. Ensure all plots have appropriate and correct labeling.
    Note: to do this correctly the input variables to **corrcoef** must have the same number of points, so trim records as needed.

4. (10 points) Using the output of the histogram for east velocity, determine the expectation value and expected variance for hurricane data and problem 3 buoy1 and buoy2 east/west velocities and temperature. Do the same for north velocity for each buoy. How do your expected values compare to the values calculated and reported in 2f and 3c?