Biostatistics 6640
Python and R in Data Science
Data Analysis Project: due December 12 by 5:00 PM

In this project, you will be analyzing weather and (simulated) malaria incidence and (simulated) intervention data from Mozambique. The objective of this project is to describe the temporal and spatial variation in these data and to draw conclusions about the relationships between the variables. Information about malaria can be found on the Centers for Disease Control and Prevention's website (`https://www.cdc.gov/malaria/`).

The data for this project can be found in two places. The weather data are here (`http://rap.ucar.edu/staff/monaghan/colborn/mozambique/daily/v2/`) and all other data are in the folder "FinalProject" on Canvas. The "incidence.csv" data contain incidence data and other details of the districts. The "intervention.csv" data contain the intervention start times for various districts across the years of observation. All data need to be merged together. Finally, the shapefile ending in .shp can be used to map attribute data.

The climate data are:

`year` is the year
`mo` is the month
`day` is the day
`raint(mm/d)` is the daily rainfall in mm
`tavg(C)` is the daily average temperature in Celcius
`rh(%)` is the relative humidity in %
`sd(mmHg)` is the saturation vapor pressure deficit in mm of mercury (another measure of humidity)
`psfc(hPa)` is the surface barometric pressure (a general indicator of large-scale weather activity and exhibits a strong seasonal cycle

The district names are in the url's for each climate file. Those need to be saved so that you can create a district name variable that will allow you to merge these data to the other data sets.

The intervention data can be assumed to be protective for specific periods of time. For example, we can assume that the IRS (indoor residual spraying) variable has 75% protection 6 months after the start date. The ITNs (insecticide treated bednets) are thought to be

60% protective 24 months after the start date. When merging these data, these coverages need to be applied to the subsequent weeks after the start week. For simplicity, you can assume a constant decrease per week to achieve the protection described above, where we can assume that IRS and ITNs start with 100% effectiveness at the start week. If you want to be more specific than this, here is a paper to help guide your assumptions `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4644290/`.

Malaria tends to be related to weather (increased rainfall and warmer temperatures) in a lagged fashion. This is because there is a 7-14 day incubation period between exposure to an infective bite by a mosquito and the onset of symptoms. The incidence data today are likely related to exposure up to 14 days prior and the effects of weather and temperature, etc, are likely related to exposure at an uncertain time before that. This time is typically thought to be 2, 4 or 8 weeks from the day the person showed up in the health center. You are expected to create the lagged variables and explore their relationships with malaria incidence.

Some tips:

1. Given that this is a class on R and Python, we expect the project to be done in R and/or Python.

2. As part of the project, we would like to you to set up a github repository (`https://github.com/`) to store the code you used for your project.

3. We expect you to produce exploratory analysis figures (histograms, splines, etc), maps of incidence (by district, across years, etc) and explore, at the very least, basic relationships between the independent variables and the outcome (malaria incidence)

4. Malaria case counts are provided and population data are provided by district, so to get incidence, you need to divide cases by the population, and incidence is typically reported as cases per 1,000 population

Write-up: This project should be completed by all individuals in the class. You may work in groups, but if you do, each person needs to submit their own write-up and please include the names of the individuals with whom you worked. The write-up should include a background section with a literature review and citations (at least one page), a description of the problem and data (approximately one page), results (no page expectation, but there should be 4-5 figures and at least one should be a map), conclusions, references and any supplemental material you choose to include.