Cornell STSCI 4520/5520 - Statistical Computing

Final Project Spring 2024 - USA Climate Trends (a.k.a. “Do your own climate research”)

Before doing anything, sit down and read this entire document from start to finish.

Introduction

You will be creating an R package related to the US Climate Reference Network (USCRN), which is a set of high quality weather stations maintained by the National Centers for Environmental Information (NCEI) within the National Oceanic and Atmospheric Administration (NOAA), which is interestingly part of the US Department of Commerce.

You will then use your package to conduct some statistical analyses of the data.

You already have some experience with these data. We have used the hourly weather data from the Ithaca weather station. For this project, we will be using all of the weather stations within the contiguous USA, and we will be using daily data instead of hourly data, due to the sheer volume of data.

The data is available as a .zip file in the “files” section on canvas and can also be accessed here:

https://www.ncei.noaa.gov/pub/data/uscrn/products/daily01/

Project Parameters

The project is due on **Tuesday, May 7** on canvas.

No extensions will be granted. You have over 3 weeks to complete the project, so plan your efforts accordingly. Getting the package to pass checks and install properly can be tricky, so give yourself several days to iron out final details.

You may work individually or in teams of 2. If you choose to work with a team member, you must notify both of the TAs by Friday, April 19.

You may discuss the project with other teams, but you are not allowed to show your code to any students not on your team.

Make sure your code is readable. This means

1. including comments, giving enough, but not too much information about what the code is doing,
2. using informative variable names,
3. using functions to avoid repetitive code,
4. using indentation where appropriate (4 spaces preferred),
5. keeping line widths below 80 spaces,
6. using whitespace to break up long blocks of code.

Products

R package

You will turn in an R package in tar.gz format. Your R package must pass all automatic checks.

Part of your grade will depend on whether we can install and run the code in your R package with minimal effort, so make sure you test that it works!

Excessive use of external R packages will be penalized. Feel free to make judicious use of a handful of packages (e.g. `GpGp`, `maps`, and `fields`). A file that loads 10 different packages will be viewed unfavorably.

Each function and dataset in your package should be documented with Roxygen comments.

Each function in your package should have a meaningful unit test, written with `testthat`, and all of the tests should pass. For plotting functions, the test should simply check that running the function does not produce an error.

In addition to all the necessary parts of an R package, your package should contain the following components.

1. A memorable package name.
2. A dataset with one row for each station, providing information about the station identifier, station name, state, longitude, and latitude.
3. The full daily dataset with the following columns:

    - WBANNO, state, station name, LST_DATE, CRX_VN, LONGITUDE, LATITUDE, T_DAILY_MAX, T_DAILY_MIN, T_DAILY_MEAN, T_DAILY_AVG, P_DAILY_CALC, SOLARAD_DAILY

    Convert missing value codes into NAs. The date column should be in R's date format. Be sure to document your dataset, explaining what each column means.

    You should create the datasets with an R script. This R script should be turned in with your submission, but it should not be part of the R package. Turn it in as a separate filed called `create_datasets.R`.
4. A function for extracting the time series for a specific station by station id. It should have optional arguments for the starting date and ending date of the time series.
5. A function for estimating the yearly cycle for one station. A yearly cycle is simply the expected temperature on each day of the year. The function should return a data frame with row for each day, a column for day number (1-365), and a column for the expected average temperature on each day.
6. A function for estimating the trend of temperatures over time, in units of degrees Celsius per year.
7. A function for creating a grid of points that fall within the contiguous USA. You may consider using external packages to get map data and find points inside a polygon. Your function should have argument(s) for controlling the resolution of the grid.
8. A function for interpolating data from the stations to a grid points within the contiguous USA.
9. A function for plotting the gridded interpolations on a map.

Analysis - Rmarkdown vignette

You should use your package to conduct an analysis of the dataset, addressing the prompts below. This analysis should go in the vignettes directory of your package. In addition, turn in a pdf version of your vignette called `vignette.pdf`.

Use the T_DAILY_AVG temperature variable in your analyses.

1. Make a map of the average temperature at each station for the month of March 2024.
2. Fit a spatial model and plot an interpolated map of average temperatures for March 2024. Consider including elevation in your model.
3. Estimate the warmest and coldest day of the year for each station, and plot those days on two maps. Think carefully about how to represent the days numerically.

    In your report, describe the statistical analysis that you used for estimating the warmest and coldest days at each station, including writing down any statistical models in mathematical notation. Be sure to define all your symbols and assumptions.

Interpolate maps of the warmest and coldest days, and plot the interpolated maps of warmest and coldest days.

4. Make a single plot of the estimated yearly cycles for 10 different stations, highlighting a diversity of climates around the contiguous USA. Your plot should clearly indicate which cycle is from which station.

5. Estimate the trend over the years for each station, in units of degrees Fahrenheit per year, and plot the trend values on a map. Indicate visually on your map which of the trends are statistically significant.

   In your report, write the statistical model that you used in mathematical notation. Be sure to define all your symbols and assumptions.

   Interpolate the estimated trends to a grid, and plot them on a map. For the interpolations, you may consider using only the trend estimates whose standard errors are sufficiently small.

6. Find a reputable source for the average temperature trend in the continguous USA over the past 20 years, and compare your results to the source's.