Project

Submission

You will work in pairs, so find a partner and tell us who you are working with.

Deadline

By 17.00 on 10 January 2020 *at the latest*, you should submit on Moodle a PDF file containing your report, entitled `SurnameA_SurnameB_Report.pdf`, and a text file, entitled `SurnameA_SurnameB_Code.R`, (saved in ASCII format) containing the `R` code. The code should be functional (that is, your plots and analysis should be reproducible). It will be used to check that your results are honest, and that there has been no sharing of code.

   You may discuss your work with other students, but the report and the code should be all your own work. Any signs of plagiarism or code-sharing will be heavily penalized.

   After 18 December you may find it hard to find us if you want help or advice, so

Don't leave your analysis to the last minute!

Problem

The dry, high-speed Santa Ana winds occur in Southern California from autumn to spring, and are caused by dry desert air masses in the high-altitude, interior Great Basin region that warm adiabatically as they sink to lower altitudes and accelerate as they pass through the Sierra Nevada mountain corridors out to the Pacific coast. These winds are of interest because of their tendency to knock down utility poles, causing electrical arcing that can ignite wildfires, a threat exacerbated by the watershed management and the fire suppression strategies of the 20th century, making Southern California one of the environments most susceptible to wildfires in the world.

   Santa Ana winds also spread wildfires, due to a dry environment with ample fuel sources consisting of low-lying shrubs and grasses, making fires during the Santa Ana wind season especially difficult to contain. Between 1990 and 2009, fires during Santa Ana wind seasons were responsible for an estimated $3.1 billion in losses.

   Because of the potential that these winds have for igniting and spreading fires, modelling the dependence of these extreme winds may be useful for utility pole design and fire safety regulation, as well as understanding Santa Ana fire diffusion.

   The data available are from the Hadley Centre and concern the relation between daily maximum windspeed (miles per hour) and the Fosberg fire weather index (FFWI) in Southern California. The latter is a simple index for evaluating the potential influence of weather on a wildland fire based on regional climate model output of temperature, relative humidity and wind speed, so calibrated that it equals 100 when the wind speed is 30 miles per hour and the moisture content of the air (a function of temperature and relative humidity) equals zero. Values of FFWI higher than 100 are (in principle) truncated back to 100. The windspeeds were kindly extracted from the HadISD database (https://www.metoffice.gov.uk/hadobs/hadisd/) by Professor Ben Shaby, and they have undergone some corrections for bad data, conversion from metres per second to miles per hour, and may have been homogenised to deal with instrument drift and other issues. As a result the wind speeds are rounded to some peculiar values. Dividing the given wind speeds by 2.23694 will give speeds in m/s.

Each dataset consists of two columns containing daily maximum windspeeds and the FFWI from 31 March 1973 to 31 May 2015. Different pairs of students have data from different sites in Southern California (many of them airfields or airports in the greater LA region).

Goal

The goal of the project is to use tools of extreme value theory to estimate the upper tail of the probability distributions for high winds, for high FFWI, and for both. There are a number of potential complications:

- missing data in one or more of the series (some of which are quite incomplete);

- the discreteness of the windspeeds may need to be taken into account (it may help to `jitter` them for some plots);

- inhomogeneity of the series, visible as obvious shifts in the location and/or scale;

- any overall trend;

- seasonal variation, which could be dealt with (a) by fitting a time series model to get a stationary series of residuals, whose extremes you can consider, (b) by including seasonal factors in a regression model for the extremes, or (c) by restricting the analysis to the period of the year when the winds are strongest, and assuming (having checked!) that the winds are stationary within this period;

- the type of dependence between the two variables, and between successive values of the same variable; and

- persistence of high winds or high FFWI (i.e., the number of successive days with high values), which may affect the probabilities of extreme values of the variables.