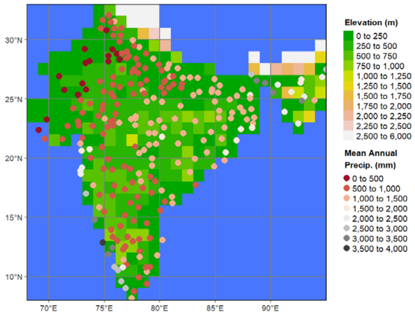
University of Colorado
Department of Civil, Environmental and Architectural Engineering
Advanced Data Analysis Techniques (Statistical Learning Techniques for Engineering and Science)
CVEN 6833
Homework 1
Due: 02/13/2025

Topics: Surface Fitting – *Linear, GLM, Local Polynomials, Linear and Local Polynomial GLMs, GAM, Spatial Models, Hierarchical and Bayesian methods*

Please present your work neatly. Organization of R-commands, functions will fetch 15% of points.

1. Derive the link function, Fischer score and Information matrix for Poisson distribution. As a bonus do the same for Gamma distribution.

2. Indian monsoon related data for the following problems is assembled at this location http://civil.colorado.edu/~balajir/CVEN6833/HWs/HW-1 .



- Annual average precipitation at 236 station locations covering India in the file climatol-ann.txt. The columns are longitude, latitude, elevation (meters) and annual precipitation average (millimeters)
- The file india-grid-topo.txt provides information of longitude, latitude and elevation on a high-resolution DEM grid covering India and surrounding land mass.

You may create a grid covering just India based on the 1^{o} x 1^{0}grid in Rajeevan-grid.txt

(a) Display the annual average precipitation along with the topography as a spatial map

(b) Compute the distance from Arabian Sea and Bay of Bengal for each station location. Thus, for each station the covariates are – Longitude, Latitude, Elevation, Distance to Arabian Sea and Distance to Bay of Bengal.

(c) You are interested in obtaining the underlying precipitation surface from these sparse spatial observations – as this will be used in a variety of applications (e.g., flood plain management, natural hazard mitigation etc.). To this end, the first objective is to perform the following:

i. Fit a *best* linear regression model (use one of the objective functions – AIC or BIC; you can also try both to see any differences). This entails fitting the model with all possible combinations of covariates (Latitude, Longitude, Elevation, distance to Arabian Sea and distance to Bay of Bengal) and selecting the model with the minimum objective function.

ii. Show the scatterplot of observed and modeled precipitation along with the 1:1 line.

iii. Perform ANOVA (i.e. model significance) and model diagnostics (i.e., check the assumptions of the residuals – Normality, independence, homoskedasticity).

iv. Compute drop-one cross-validated estimates from the best model and scatterplot them against the observed values with the 1:1 line. This and the scatterplot in (ii) above, is to visually see how the model performs in a fitting and cross-validated mode.

v. Drop 10% of observations, fit the model (i.e., the 'best' model from i. above) to the rest of the data and predict the dropped points. Compute RMSE and correlation and show them as boxplots.

vi. Spatially map the model estimates and the standard error from the *best model* on the high-resolution grid

vii. Briefly discuss what you find [bullet points are fine]

3. Repeat 2c. by fitting a ***GLM*** with appropriate link function.

4. Repeat 2c with Local Polynomial method but using the appropriate link function (i.e., '***Local GLM***').
[For the Local Polynomial approach, the 'best model' involves fitting the best subset of predictors and the smoothing parameter, alpha. You can also compare the GCV from these four different methods.]
Briefly discuss the results from the local polynomial approach and compare them to linear regression.

5. Repeat 2c by fitting a ***Generalized Additive Model*** (GAM) and compare with the GAM fitted in a local polynomial framework.

6. Perform a ***Hierarchical Spatial Model. [Fit a best linear model as in problem 2 and perform Kriging on the residuals]***

7. Repeat 6. with a ***Bayesian Hierarchical Spatial Model*** (see Verdin et al. 2015, for tips)
[In the Bayesian models, plot the posterior historgram/PDF of the parameters, spatial maps of posterior mean and standard error]

8. I want you to reflect on the suite of analyses performed above on the spatial precipitation data – in particular, the relative performance of the methods, their advantages and disadvantages and potential application of these methods on a problem/data set of your interest. Keep this short and crisp (bullet points are fine).