ECO 221: Project (2025)

You are provided the data for crop production and inputs for major crops across Indian districts for 2017 (ECO221_Project_2025_Final.csv). As a first step, please check the group assignment from the 'Group Allocation' sheet. Each group is assigned one crop. Details of the variables included in the dataset:

Variable Name	Description
crop	Crop Name
area1000hectares	Area under the Crop (in '000 hectares) per district
production1000tonnes	Production of Crop (in '000 tonnes) per district
irrigatedarea1000hectares	Irrigated Area for the Crop (in '000 hectares) per district
districtcode	District Code
districtname	Name of the District
year	Year = 2017
statecode	State Code
statename	Name of the State
nitrogenconsumptiontonnes	Nitrogen Application (in tonnes) per district
phosphateconsumptiontonnes	Phosphate Application (in tonnes) per district
potashconsumptiontonnes	Potash Application (in tonnes) per district

Please note:

- Requests for changes in the group assignment will not be entertained.
- Cleaning district names across the different datasets to ensure you do not lose data when merging them is important.
- If you end up losing data for some reason, document it systematically.
- The final submission for the project will include your slides, dataset, and your code file (Python or R). More details on the slides will be released closer to the date of presentations.

Task 1:

Merge the district level crop production and inputs data with the corresponding district-level data on rainfall. Note that the rainfall dataset provides district-wise monthly rainfall for the year 2017. In order to merge with your district-level file, you will need to construct a district-wise rainfall summary for the year 2017. When doing this exercise, pay attention to the fact that the crop you are assigned may only be grown in one of the seasons in the year. In that case, constructing a season-specific rainfall summary is (likely) more informative.

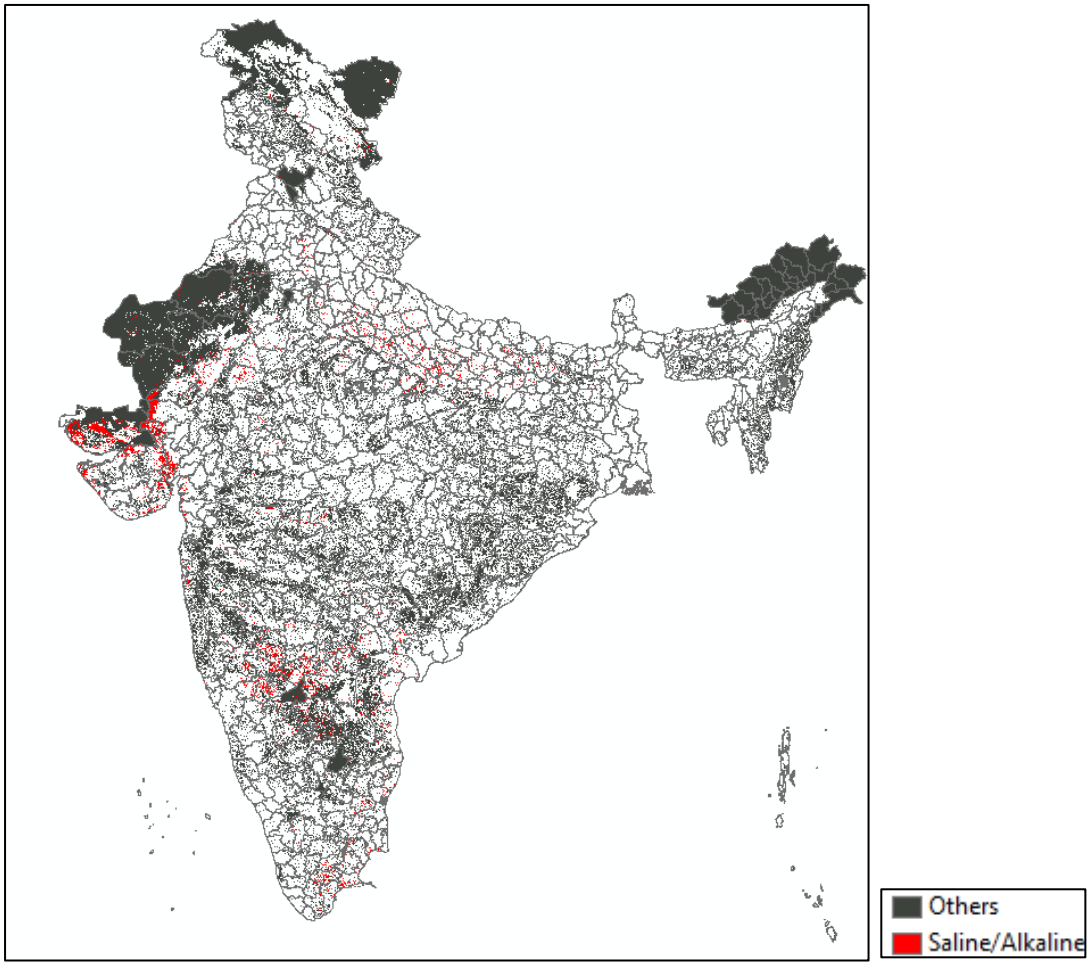
File Name: RF_DistrictWise_ECO221_2025.csv

Task 2:

Next merge your data with district-level soil quality data. The data contains an index on extent of land degradation (as percentage of land area) due to excess salinity/ alkalinity in

soils, constructed using a spatially-delineated dataset on land degradation provided by the National Remote Sensing Centre linked here (screenshot below).

File Name: Salinity_Alkalinity_ECO221_2025.csv



Question 1:

Please provide a detailed summary of all the variables in your dataset.

Part A: Cobb-Douglas Production Function

A Cobb-Douglas production function models output (Y) as a function of inputs ($X_k; k = 1,2, \dots K$) such that $Y = AX_1^{\alpha_1}X_2^{\alpha_2} \dots X_K^{\alpha_K}$. Here A represents the quality of the overall technology of the producer (farmer), and α_k represents the effectiveness of the k^{th} input in influencing output. To estimate this using a regression model, we take a log of the equation and estimate a log-linear model.

Please note:

- While environmental controls like weather and soils are included as part of the input vector, these are outside of a farmer's control.
- When running a log-linear model, you will need to account for zeroes in the $irrigated\ area$ variable, you can do this by redefining $irrigated\ area\ (new) = irrigated\ area + 1$ and including the new variable instead of the old one.
- Further, define a new variable *Unirrigated Area = Total Cropped Area - Irrigated Area (New)* and include both irrigated and unirrigated area as part of the input vector.
- Please include the environmental variables without taking a log.

Question 2:

Using this data, estimate a Cobb-Douglas production function which models output (production in tonnes) as a function of inputs (e.g., area, irrigated area, fertilizer) and controls for the production environment (e.g., weather). Interpret your results paying careful attention to the units.

Question 3:
Are there any outliers and/or influential observations? What are the collinearity diagnostics? How do they influence the estimation and interpretation?

Question 4:
Visualize the model residuals on a plot having the *crop production* on Y-axis and *crop area* on the X-axis. Now, construct a second plot having the residuals on the Y-axis and *crop area* on x-axis. Are the plots, what you would expect? Explain. Plot a histogram of residuals and verify that the sum of the residuals is zero.

Question 5:
Based on your estimates, what can you say about the returns to scale for the production function? Test the hypothesis that inputs within the farmer's control (i.e., excluding controls for production environment) exhibit constant returns to scale. Clearly state the null and alternate hypotheses.

Part B: Quadratic Production Function

Another popular production function is a quadratic production function which models output (Y) as a function of inputs ($X_k; k = 1,2, \dots K$) such that $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_2 X_2 + \beta_3 X_2^2$ .... In this case both the dependent and independent variables are in levels with an inclusion of squared input terms to account for the diminishing effectiveness of inputs in influencing output beyond a certain threshold. You may assume that input use is always positive.

Question 6:

As mentioned above, we expect that the impact of an additional unit of area/ input(s) on output is lower as the area/ input(s) levels increase beyond a threshold (this is known as diminishing marginal product of inputs). How would you test this hypothesis? Clearly state the null and alternate hypotheses for each input and augment the model to test the same. Summarize your findings.

Question 7:

Finally, irrigation and fertilizer application are such that additional fertilizer application is more effective with irrigation than without (this is known as input complementarity). How would you test whether this hypothesis holds for your crop?

Question 8:

Enhance the model in parts A and B to test if there are differences in the estimated production function across regions within India. The definition of the states belonging to each region given by the Reserve Bank of India are linked here.

Question 9:

Of the two models in Part A and Part B (i.e., the Cobb-Douglas and Quadratic models), which would you prefer and why?