Name: Bhargav Narasimha Shandilya
Bayesian Statistics Project

Multi-cropping and Soil Quality in Southeast Asia - Regression

In several Southeast Asian countries, rice is one of the most popular food crops. Rice is also the staple diet in several parts of countries such as Bangladesh, India, Indonesia, and China. To meet the immense demand for this grain, farmers often exhaust soil nutrients and moisture very quickly. Unfortunately, rice is a very water-intensive crop that depends on heavy availability of water and soil nutrients. To make matters worse, farmers sometimes choose to grow more than one crop in a year in order to maximize land usage and profits. For instance, a farmer might choose to grow corn for 5 months and then wheat for 3 months. This is known as multi-cropping. Within a few years of intensive multi-cropping, soil quality becomes extremely poor.

Currently, due to increased rice production in the Indian state of Punjab, the soil water index has reached an all-time low[1]. The government is now encouraging the cultivation of other crops.

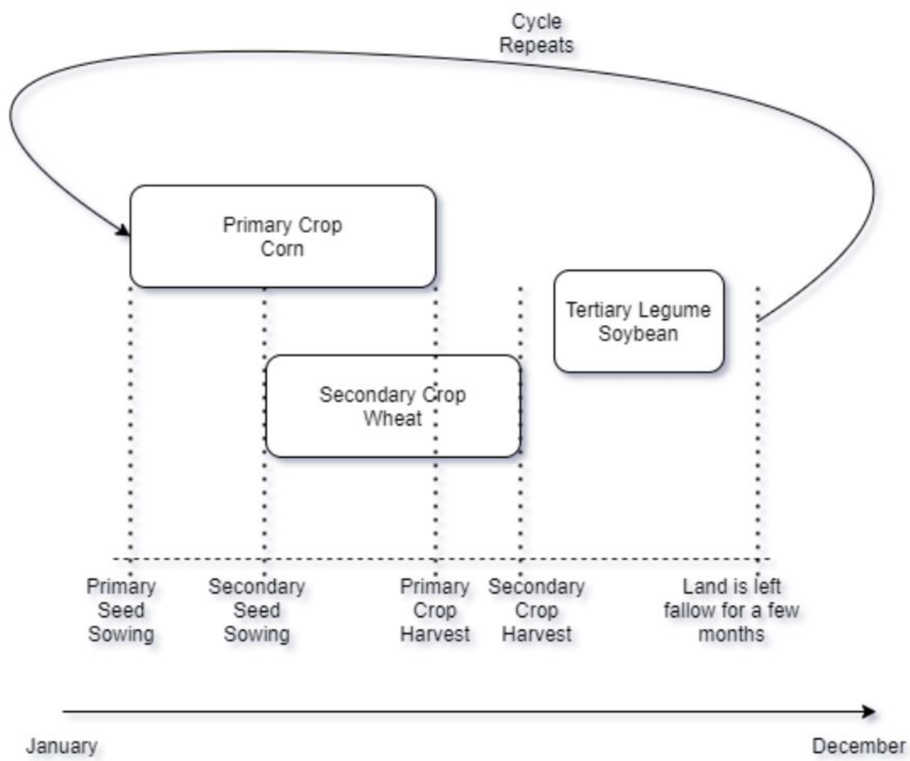
To avoid soil deterioration, farmers are encouraged to grow rotation crops, also known as legumes. These legumes can be crops such as peas, lentils, and beans. Besides having a short growing cycle, legumes add manure and rejuvenate the soil quality to some extent. The effect of legumes on soil quality is well-documented as shown in [2].

Problem Statement:

However, the specific impact on Punjab’s unique soil composition still requires more experimentation and research. In this analysis, I will attempt to measure the effect of multi-cropping and legume cultivation on soil quality. Since some of my data was missing for soil quality, I will also attempt to impute the response in certain scenarios. Lastly, I will also be predicting the impact of a certain growth cycle on soil quality.

Based on the number of months for which each type of crop is grown in a year, can we build a linear regression model to predict the soil quality at the end of the year? We want to answer questions such as: if a farmer grows rice (primary crop) for 3 months, wheat (secondary crop) for 4 months, and a legume (rotation crop) for 2 months, what will the soil quality be at the end of the year?
The coefficients for the predictors will also tell us how each type of crop affects soil quality.

The image below explains a typical yearly cycle:



Data:

This is data that I personally collected a few years ago. Since I was looking to help farmers choose the most optimal crop cycle to maintain soil quality, I collected data from about 25 different farms growing rice as the primary crop, wheat as the secondary crop, and soybeans as the legume (tertiary rotation crop). The legume is expected to have a positive effect on soil quality, while the other two are expected to have a negative effect.

The first three columns represent the growth period (in months) of the primary, secondary, and rotation crop respectively. The fourth column is the soil quality (response). The soil quality index ranges from 0 to 1, with 0 indicating the worst possible soil quality, and 1 indicating natural (best) soil quality. Please note that these values between 0 and 1 are <u>not</u> probabilities. They are just a measure of quality.

I also did not have soil quality data for two of the farms. These are represented by 'NA', and can be imputed during the regression process.

	Primary Crop	Secondary Crop	Legume	Soil Quality
1	5.0	3.0	2.0	0.420
2	8.0	0.0	2.0	0.340
3	10.5	0.0	0.0	0.020
4	6.5	4.0	0.0	0.090
5	3.5	2.0	4.0	0.720
6	3.0	4.5	4.5	0.830
7	6.0	6.0	0.0	0.010
8	3.5	3.0	2.0	0.630
9	4.0	4.0	4.0	NA
10	7.0	3.0	2.0	0.230
11	11.0	0.0	0.0	0.010
12	0.0	9.0	3.0	0.590
13	0.0	0.0	10.0	0.980
14	3.0	7.0	2.0	0.494
15	4.0	6.0	2.0	NA
16	0.0	10.0	0.0	0.140
17	0.0	0.0	8.5	0.970
18	7.0	2.0	2.5	0.290
19	4.0	2.0	6.0	0.730
20	8.0	3.0	0.0	0.010
21	9.0	0.0	2.0	0.140
22	5.0	5.0	2.0	0.390
23	7.0	3.0	0.0	0.023
24	5.0	4.0	0.0	0.030
25	7.5	2.5	2.0	0.210

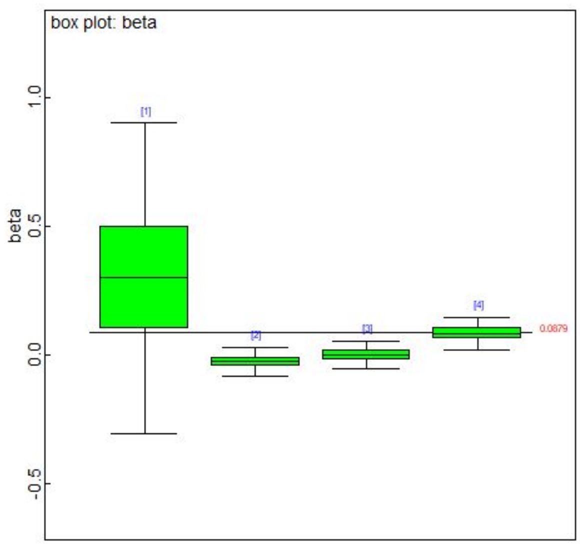
Special Note:
The values of the first three columns for a particular row do not have to add up to 12 months. A farmer might choose to grow all three crops within 10 months, and leave the land empty for the remaining months. Also, the growing cycles of two crops might overlap since two crops can be grown parallelly on the same plot of land. In that case, as well, the number of months for a row will not necessarily add up to 12.

Analysis of Results:

From the results above, it seems to be clear that the legume (beta[4]) growing period is quite important for soil quality. Although the means of beta[2] and beta[3] are negative, we cannot conclusively say that primary and secondary crops contribute to soil quality depletion since their credible sets contain '0'.

While the credible sets for beta[2], beta[3], and beta[4] are quite narrow, the credible set for the intercept beta[1] is quite large. This probably indicates that the state of the soil before crops are grown causes huge variations as compared to the growing periods of each type of crop. We also observe that the standard deviation for the intercept is an order of magnitude higher than the other coefficients. Since the current year's pre-sowing soil quality is dependent on the previous years' crop cycle, it might be better if we had cyclic data for several years instead of a single year.

The regression model fits quite well to the data as indicated by the adjusted R-squared value of 0.8268. The two missing values are also imputed to be 0.5402 and 0.3623. The new soil quality is predicted to be 0.6526.



	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
BR2	0.8448	0.05591	2.702E-4	0.7057	0.8558	0.9191	1001	200000
BR2adj	0.8226	0.0639	3.088E-4	0.6637	0.8352	0.9076	1001	200000
beta[1]	0.3005	0.2689	0.006485	-0.2237	0.2999	0.8361	1001	200000
beta[2]	-0.02811	0.02533	6.004E-4	-0.07862	-0.02808	0.02155	1001	200000
beta[3]	-3.687E-4	0.02459	5.647E-4	-0.0493	-3.437E-4	0.04776	1001	200000
beta[4]	0.08759	0.02796	6.447E-4	0.0319	0.08758	0.1423	1001	200000
soil_quality	0.6526	0.05258	7.991E-4	0.5482	0.6525	0.7565	1001	200000
soil_quality_s	0.6526	0.1426	8.134E-4	0.3701	0.6529	0.9358	1001	200000
y[9]	0.5402	0.1411	0.001115	0.2595	0.5395	0.8191	101001	100000
y[15]	0.3623	0.1408	9.569E-4	0.08366	0.3621	0.6419	101001	100000

Zellner’s Prior

I also attempted to use Zellner’s prior with g = 25. However, this yielded a lower adjusted R-squared value of 0.7578. The imputed values were also slightly different. The table below shows the results with Zellner’s prior.

	mean	sd	MC_error	val2.5pc	val5.0pc	median	val95.0pc	val97.5pc	start	sample
BR2	0.7881	0.0671	2.435E-4	0.6248	0.6637	0.8005	0.8721	0.8816	101001	100000
BR2adj	0.7578	0.07669	2.782E-4	0.5712	0.6157	0.772	0.8538	0.8647	101001	100000
beta[1]	0.2978	0.3054	6.662E-4	-0.3051	-0.2043	0.2986	0.7982	0.9007	501	200500
beta[2]	-0.02784	0.02876	6.281E-5	-0.08476	-0.07477	-0.02787	0.01936	0.02886	501	200500
beta[3]	-0.001448	0.02792	6.003E-5	-0.05657	-0.04713	-0.001469	0.04446	0.05363	501	200500
beta[4]	0.08312	0.0318	6.965E-5	0.02025	0.03107	0.08307	0.1352	0.1462	501	200500
y[9]	0.5133	0.1645	3.777E-4	0.1877	0.2442	0.5133	0.7831	0.8381	501	200500
y[15]	0.3431	0.164	3.734E-4	0.01858	0.07434	0.3429	0.6124	0.6684	501	200500

References:

[1] B. D. Dhawan. “Ground Water Depletion in Punjab.” Economic and Political Weekly, vol. 28, no. 44, 1993, pp. 2397–2401. JSTOR, www.jstor.org/stable/4400350. Accessed 5 Dec. 2020.

[2] Gogoi, Nirmali & Baruah, Kushal & Meena, Ram Swaroop. (2018). Grain Legumes: Impact on Soil Health and Agroecosystem. 10.1007/978-981-13-0253-4_16.

[3] Code: Course Material and Sample Code from https://www2.isye.gatech.edu/isye6420/
Specific Example Used as Reference: fatmulti.odc from the code_for_week7 folder