3. Experimental Approach

The experimental analysis was conducted using the Keras framework (version 2.3) with TensorFlow (version 1.14) on a desktop computer equipped with an Intel Core i9-7900X Skylake processor, an NVIDIA GTX 1080 Ti 11GB graphics card, and 64GB of RAM. The methodology for this analysis began with Exploratory Data Analysis (EDA), which was performed using advanced statistical tools from the scikit-learn library in Python. This initial step involved a thorough examination and analysis of the dataset, focusing on understanding the relationships between soil properties, crop treatments, and crop yield. Following the EDA, statistical tests such as ANOVA and t-tests were conducted to evaluate the significance of mean differences between various groups within the data. To present the findings, diverse graphical representations were utilized, including bar charts, histograms, violin plots, and box plots. These visualizations effectively conveyed the results and insights derived from the analysis.

3.1 Description of Dataset

The dataset used in this research was collected from a farmland in Oro, Kwara State, Nigeria. It is structured into four main segments:

i. Treatment: This segment details the different crop treatments applied.
ii. Yield Data: This includes measurements of grain yield weight.
iii. Plant Growth Data: This encompasses metrics such as pod length and the number of pods.
iv. Pest Infestation Data: This covers the presence of specific pests, namely **Megalurothrips sjostedti** and **Aphis craccivora**.

Table 1 summarizes the crop yield dataset, outlining key parameters and variables crucial for subsequent analysis and modeling. To enhance clarity and facilitate interpretation, abbreviated codes were used for longer variable names. Additionally, the dataset was supplemented with meteorological data obtained from the local weather station's website for the farm region.

3.2 Experimental Pre-processing of Data

Preprocessing is essential for ensuring data quality and preventing errors during analysis and model training. The process began by merging the soil nutrient data with the soil particle data using the `SampleID` as a key. During data cleaning, it was discovered that the soil nutrient data had missing values. These were addressed using the method proposed by Newgard and Lewis (2015), which involved replacing missing values with the average value from each column. Additionally, the cowpea yield dataset had 10 missing data points, which were removed to prevent anomalies and bias in model deployment.

The dataset includes several categorical variables, such as Treatment, Soil Classes, and Sample Specimen, as outlined in Table 1. To prepare these categorical variables for analysis and machine learning training, numerical values were assigned using the `LabelEncoder` class from scikit-learn. This step facilitates the encoding of categorical data for further processing. Key features that significantly influence yield outcomes were identified from the merged and cleaned data. Spearman's rank correlation coefficient (Eq. 2) was used to ensure data consistency during feature selection. Features with absolute correlation coefficients greater than 0.35 were selected as crucial for model training. To facilitate analysis and comparison, the dataset was segmented by treatments and replications. Data normalization was then performed to improve data quality and consistency. The `MinMaxScaler` function from scikit-learn was employed for normalization, scaling values to a range between 0 and 1. This step is critical for accurate data evaluation and interpretation, as detailed in Eq. (1).

$$A_{normalized} = \frac{A - A_{minimum}}{A_{minimum} - A_{maximum}} \quad (1)$$

Where A is the attribute data considered, and the respective minimum and maximum value for each attribute data considered.

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(M_i - \bar{M})^2}} \quad (2)$$

wherein 'r' symbolizes the coefficient outcomes. 'y_i' represents the actual occupancy values pertinent to the prediction tasks, while 'M_i' corresponds to the values of each feature. Further clarification stems from $\bar{y}$ and $\overline{\mathrm{M}}$, signifying the average values of 'y_i' and 'M_i', respectively. It is

noteworthy that the proximity of the absolute value of r to 1 or -1 indicates a heightened degree of correlation.

Table 1 Description of dataset variable

Category	Variable Name	Code_ABB	Data Type
Treatment	A	Treatment	object
	B		
	C		
	D		
	E		
POD	Pod length at 65 days after planting (cm)	POD_LENGTH_65	float64
	Number of pod per plant at 60 days after planting	POD_PLANT_60	float64
	Pod load scoring	POD_SCORE	int64
Observations/Infestations	Maruca vitrata % of flower infestation and damage	MVF_INF	int64
	Megalurothrips sjostedti observation 1st observation	MSJ_OBS_1ST	float64
	Megalurothrips sjostedti observation 2nd observation	MSJ_OBS_2ND	float64
	Megalurothrips sjostedti observation 3rd observation	MSJ_OBS_3RD	float64
	Megalurothrips sjostedti observation 4th observation	MSJ_OBS_4TH	float64
	Megalurothrips sjostedti observation 12th observation	MSJ_OBS_12TH	float64
	Aphis craccivora infestation observation 1st	ACINF_OBS_1ST	float64
	Aphis craccivora infestation 2nd observation	ACINF_OBS_2ND	float64
	Aphis craccivora infestation 3rd observation	ACINF_OBS_3RD	float64
	Aphis craccivora infestation 4th observation	ACINF_OBS_4TH	float64
	Aphis craccivora infestation 5th observation	ACINF_OBS_5TH	float64
	Aphis craccivora infestation 6th observation	ACINF_OBS_6TH	float64
Grain Yield	Weight of grain yield per hill of 20 plants (g)	GY_PER_20	float64
	Net plot grain weight (g)	NP_Grain (g)	int64

3.3 Data Interpretation

- Descriptive Statistics:
  - **Table 2**: Entails the yield statistic based on the treatment categories. The yield statistics indicate that Treatment C produced the highest average yield (417.5 units), although with significant variability (Std = 112.27 units), suggesting a trade-off between high yield and consistency. Treatment B, while yielding lower on average (327.5 units), exhibited the least variability (Std = 49.76 units), making it the most consistent and reliable option. Treatment D showed relatively high yields (Mean = 356.0 units) but with considerable variability and a skew towards lower yields. Treatment A presented moderate yields with low variability, while Treatment E had the lowest yield (136.5 units) and moderate variability, indicating it was the least effective among the treatments in terms of yield output.

- **Table 3**: The soil class distribution shows that most of the samples are Loamy Sand (13 counts), followed by Sand (4 counts) and Sandy Loam (3 counts). The mixed class of Sand/Loamy Sand and Sandy Clay Loam were less common, each with 3 and 2 counts respectively. This distribution indicates a predominance of loamy sand, with fewer instances of pure sand and other sandy soil types.
- **Table 4**: The one-way ANOVA results indicate a statistically significant difference between the treatment group means, as evidenced by an F-statistic of 5.9653 and a p-value of 0.0044. Since the p-value is below the conventional threshold of 0.05, we can reject the null hypothesis, suggesting that at least one group's mean is significantly different from the others.
- **Table 5**: The pairwise t-test results reveal that most comparisons between treatments do not show statistically significant differences, as indicated by p-values greater than 0.05. Specifically, comparisons between Treatments A and B, A and C, A and D, B and C, B and D, and C and D all have p-values above 0.05, suggesting no significant differences in mean yields between these pairs. However, significant differences were found in the comparisons of Treatments A vs. E ($p = 0.0038$), B vs. E ($p = 0.001$), C vs. E ($p = 0.0033$), and D vs. E ($p = 0.0185$), indicating that Treatment E's mean yield is significantly different from all other treatments. This suggests that Treatment E is statistically distinct in its effect on yield compared to the other treatments.

Table 2 Yield statistics by treatment

Treatments	Mean	Median	Std	Min	Max
A	296.5	304.0	56.7	224	354
B	327.5	342.0	49.76	260	366
C	417.5	426.0	112.27	278	540
D	356.0	393.0	130.61	176	462
E	136.5	119.0	41.29	110	198

Table 3 Soil class distribution

Soil Class	count
LOAMY SAND	13
SAND	4
SANDY LOAM	3
SAND/LOAMY SAND	3
SANDY CLAY LOAM	2

Table 4 One-way ANOVA results

Metrics	Value
F-statistic	5.9653
p-value	0.0044

Table 5 Pairwise t-test results

Treatment 1	Treatment 2	t-statistic	p-value
A	B	-0.8219	0.4425
A	C	-1.9242	0.1027
A	D	-0.8358	0.4353
A	E	4.5624	0.0038
B	C	-1.4658	0.1931
B	D	-0.4078	0.6976
B	E	5.908	0.001
C	D	0.7142	0.5019
C	E	4.6983	0.0033
D	E	3.2048	0.0185

- EDA Interpretation:
  - **Figure 1:** Presents the Spearman's rank correlation heatmap for all variables, identifying the strength and direction of the relationships between them. Notable correlations include a strong positive relationship between "GY_PER_20" and "NP_Grain (g)" (0.59),

"POD_LENGTH_65" and "POD_PLANT_60" (0.67), as well as a negative correlation between "POD_SCORE" and both "NP_Grain (g)" (-0.53) and "POD_LENGTH_65" (-0.60). The matrix highlights crucial features with absolute correlation coefficients greater than 0.35, which are critical for model training. For instance, "MVF_INF" and "POD_SCORE" (0.78) also show a significant positive correlation.

- **Figure 2:** Focuses on the correlations among the POD variables. "POD_LENGTH_65" and "POD_PLANT_60" are positively correlated (0.67), indicating that plants with longer pods tend to have more pods per plant. Conversely, "POD_SCORE" is negatively correlated with "POD_LENGTH_65" (-0.60), "POD_PLANT_60" (-0.37), and "NP_Grain (g)" (-0.53), suggesting that a higher pod score might be associated with shorter pod length and fewer pods per plant, as well as lower grain weight. These correlations, especially those with absolute values greater than 0.35, highlight important variables for predictive modeling in agricultural studies.

- **Figure 3:** Figure 3's boxplot illustrates the distribution of the various crop yield variable, highlighting differences in central tendencies and variability. Metrics like "GY_PER_20" and "ACINF_OBS_6TH" show higher median values, suggesting higher central values, while others, such as "POD_LENGTH_65" and "ACINF_OBS_2ND," display outliers, indicating variability or anomalies. The varying box widths reflect different levels of data dispersion, with "ACINF_OBS_1ST" showing less variability and metrics like "POD_SCORE" and "MSL_OBS_1ST" exhibiting greater spread. Overall, the boxplot effectively summarizes the distribution and variability across these metrics.

- **Figure 4:** The histograms illustrate the distribution of the crop-related metrics. Each subplot represents a specific metric, showing the frequency of data points across different ranges. The x-axis represents the metric values, while the y-axis indicates the count of observations within each bin. The distributions vary considerably among the metrics: some, like "POD_PLANT_60" and "ACINF_OBS_1ST," exhibit a skewed distribution with most

values clustered towards one end, while others, such as "GY_PER_20" and "MVF_INF," display a more uniform or bimodal distribution. The variation in shapes and spreads across the histograms suggests diverse patterns and characteristics in the data, reflecting different aspects of the crop's attributes and their measurements.

- **Figure 5:** The boxplots in the image illustrate the influence of different treatments (A, B, C, D, and E) on various observations related to plant growth and pest infestation. For the yield-related metrics like GY_PER_20, POD_LENGTH_65, POD_PLANT_60, and MVF_INF, Treatment E consistently shows distinct outcomes, often with lower variability and different median values compared to other treatments, indicating that Treatment E may have a unique impact on these parameters. Conversely, Treatments A, B, C, and D tend to show more overlap, particularly in their effects on pod length and plant metrics, suggesting similar influences on these aspects of plant growth.

  Regarding pest infestation observations (MSJ_OBS and ACINF_OBS series), the variability across treatments is more pronounced. For example, the first infestation observations (ACINF_OBS_1ST) show stark differences, with Treatment B having a notably high median and little variability, while others like Treatment E show much lower infestation levels. Over time, as seen in later observations like ACINF_OBS_4TH to 6TH, the differences between treatments appear to lessen, with most treatments converging towards similar infestation levels, though some treatments still show distinct patterns, indicating a varying degree of effectiveness in pest management across treatments.

- **Figure 6:** The first graph illustrates the progression of **Aphis craccivora** infestation over time across five treatments. Initially, all treatments show a similar increase in infestation, but over time, differences emerge. Treatments A, B, C, and D exhibit a steady rise in infestation levels, with Treatment E showing a slightly steeper increase initially, then stabilizing. The shaded regions indicate variability, with some overlap among treatments, but overall, infestation tends to rise across all treatments, with no clear outlier in terms of effectiveness in reducing infestation.

  The second graph tracks **Megalurothrips sjostedti** infestation over a longer observation period. Unlike the first graph, this one show more fluctuation in the early observations,

particularly with Treatment D, which experiences a sharp initial increase followed by a decrease, then stabilization. Most treatments show a more stable trend over time, with slight increases or relatively flat patterns after the initial fluctuation. The variability among treatments appears greater, especially in early observations, but over time, the differences between treatments diminish, indicating similar levels of infestation in the later stages of observation.

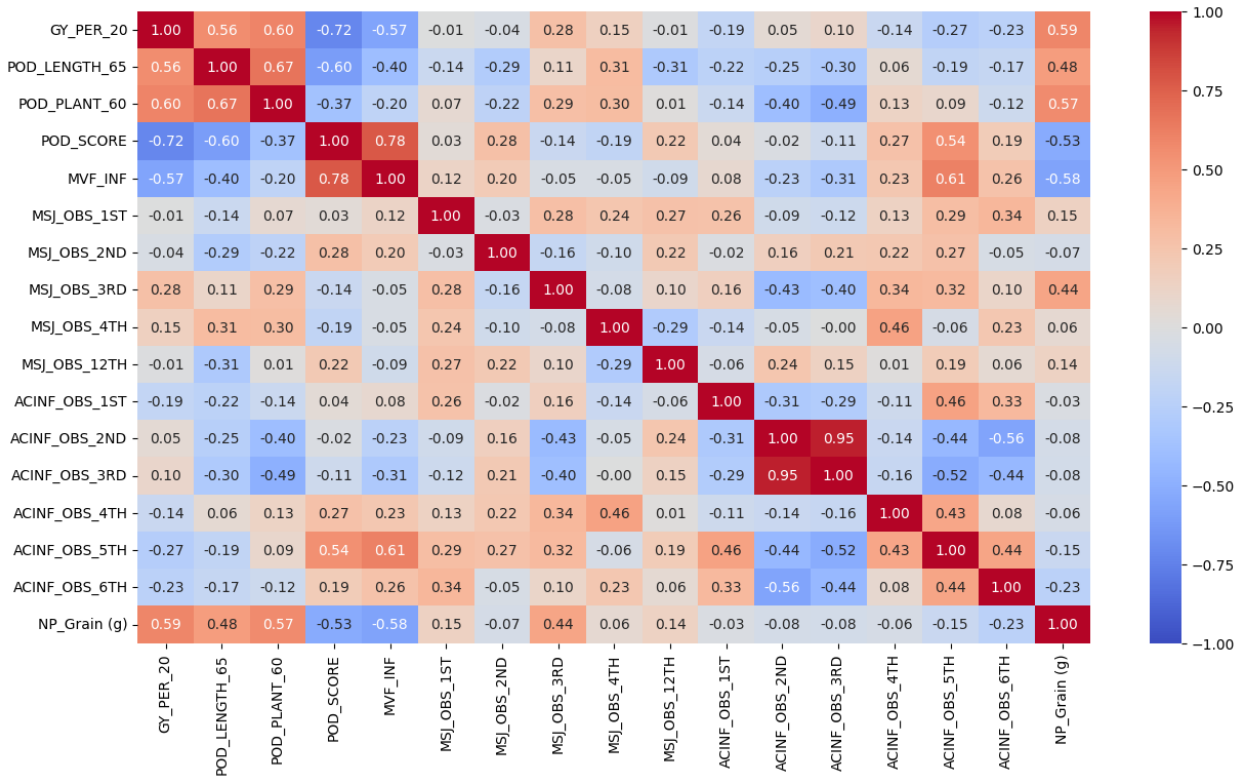

Fig. 1 Correlation heatmap of all variables

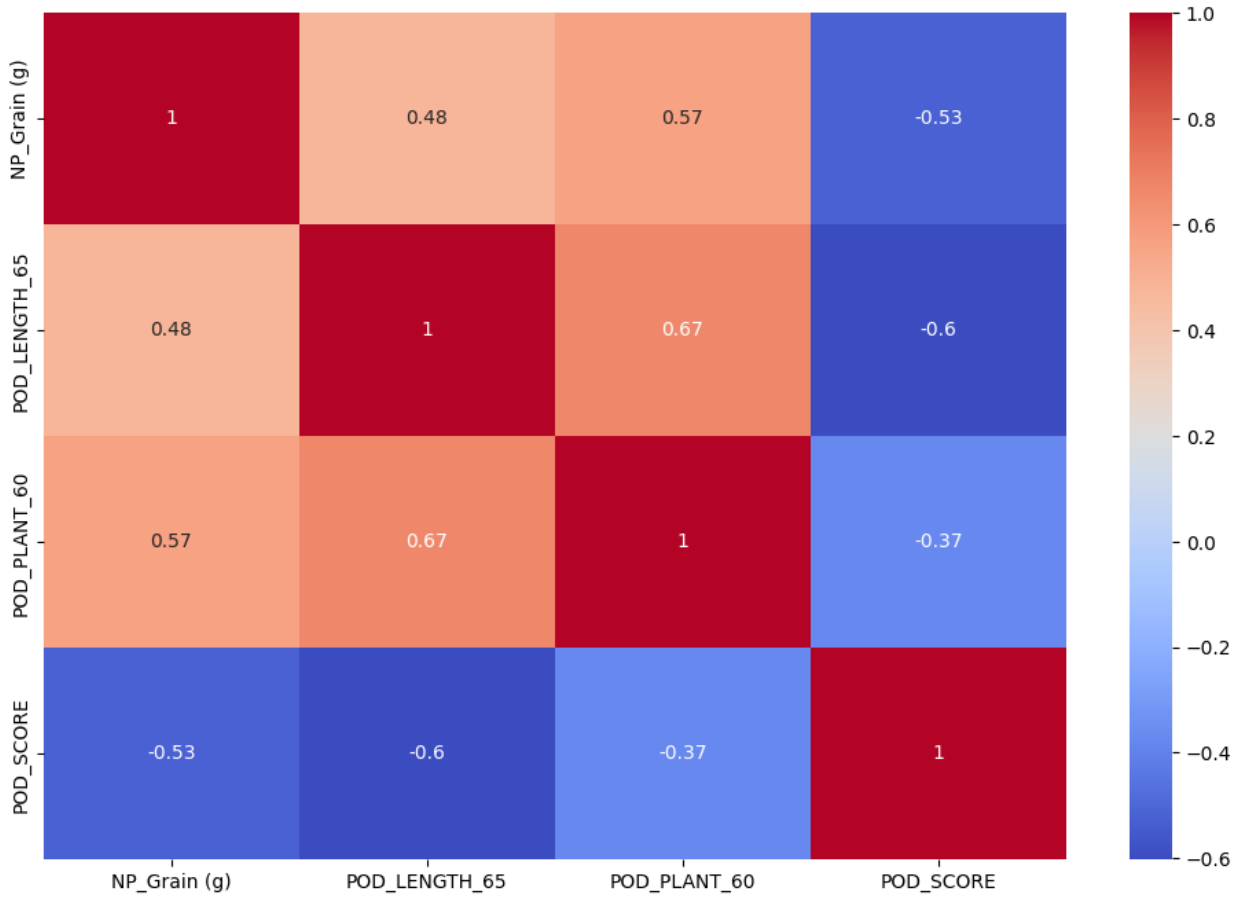

Fig. 2 Correlation score of POD variables

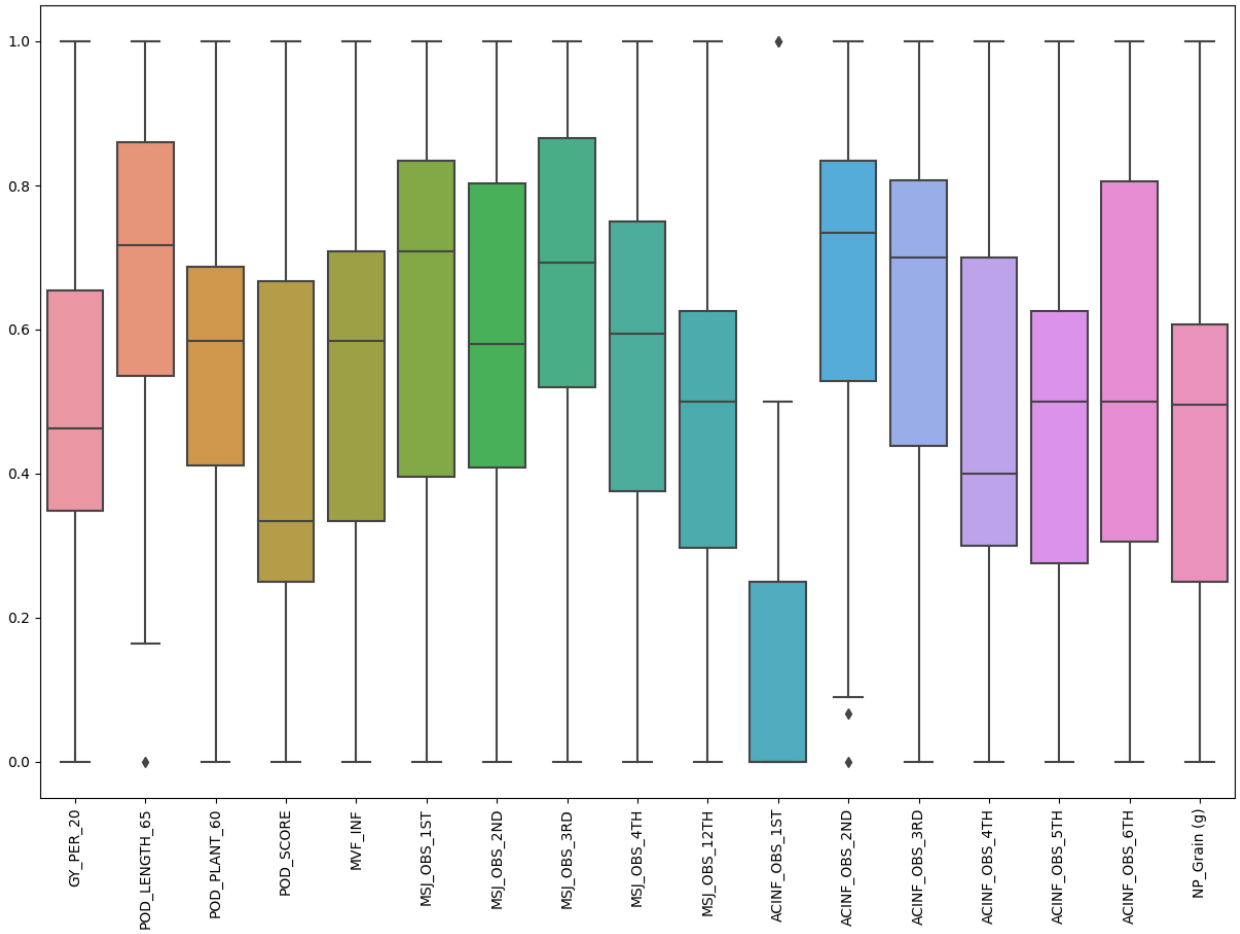

Fig. 3 Relationship Boxplot

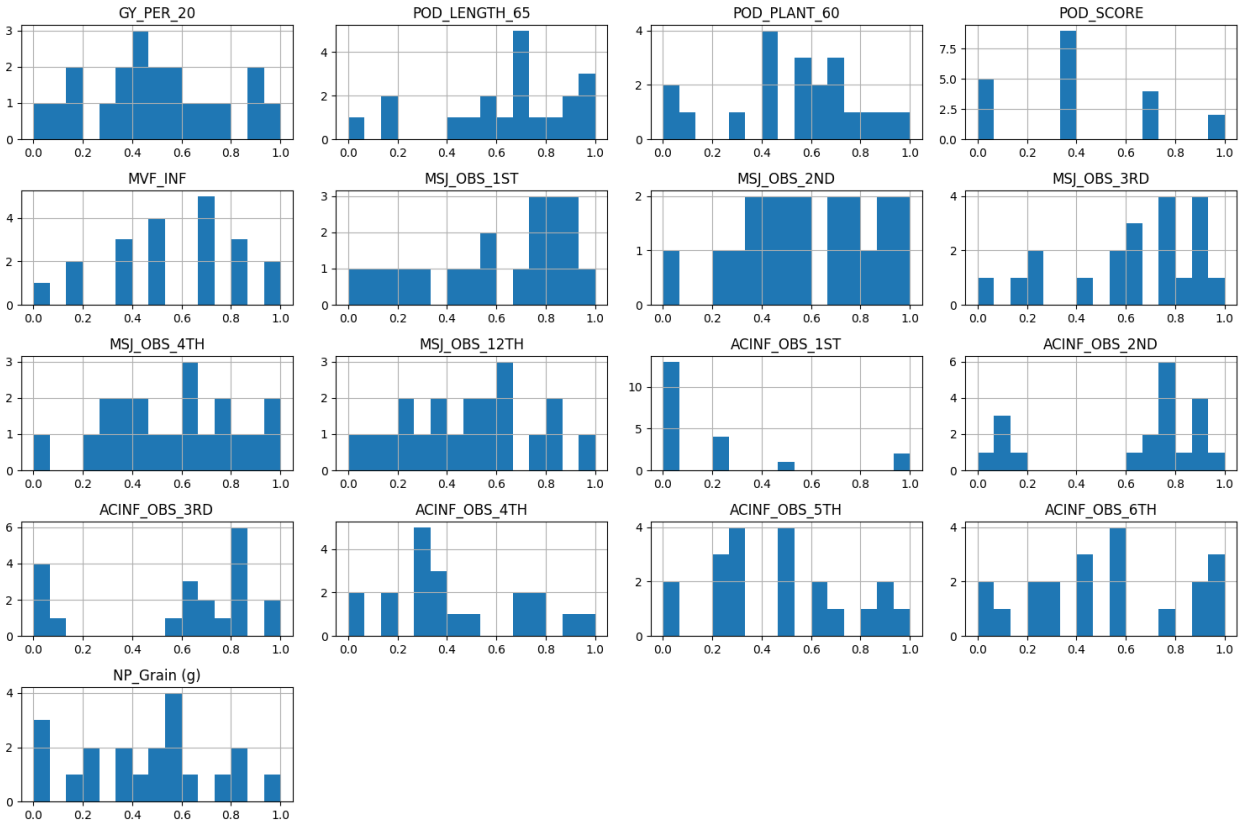

Fig. 4 Histogram plot

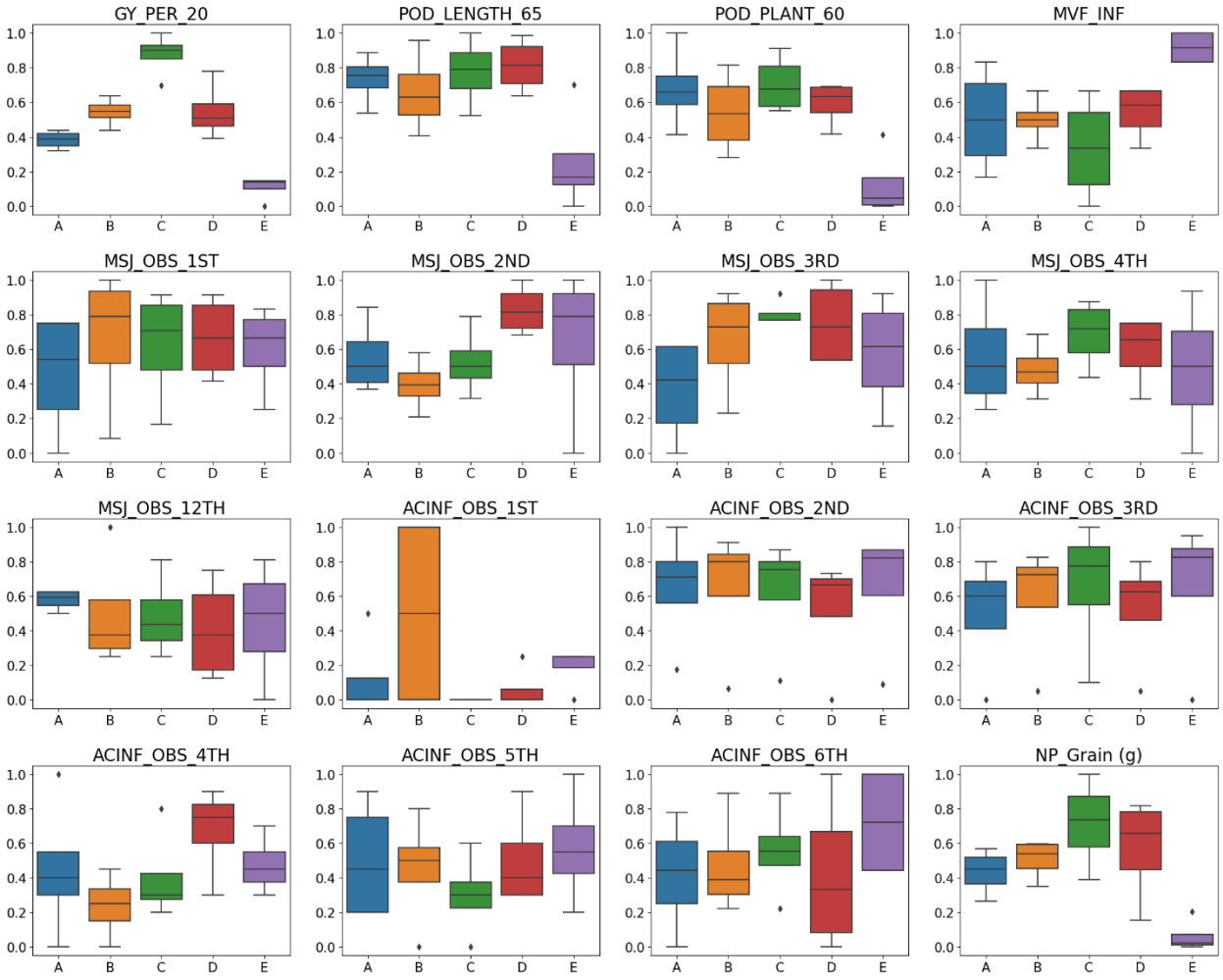

Fig. 5 Influence of Treatments class on each observation

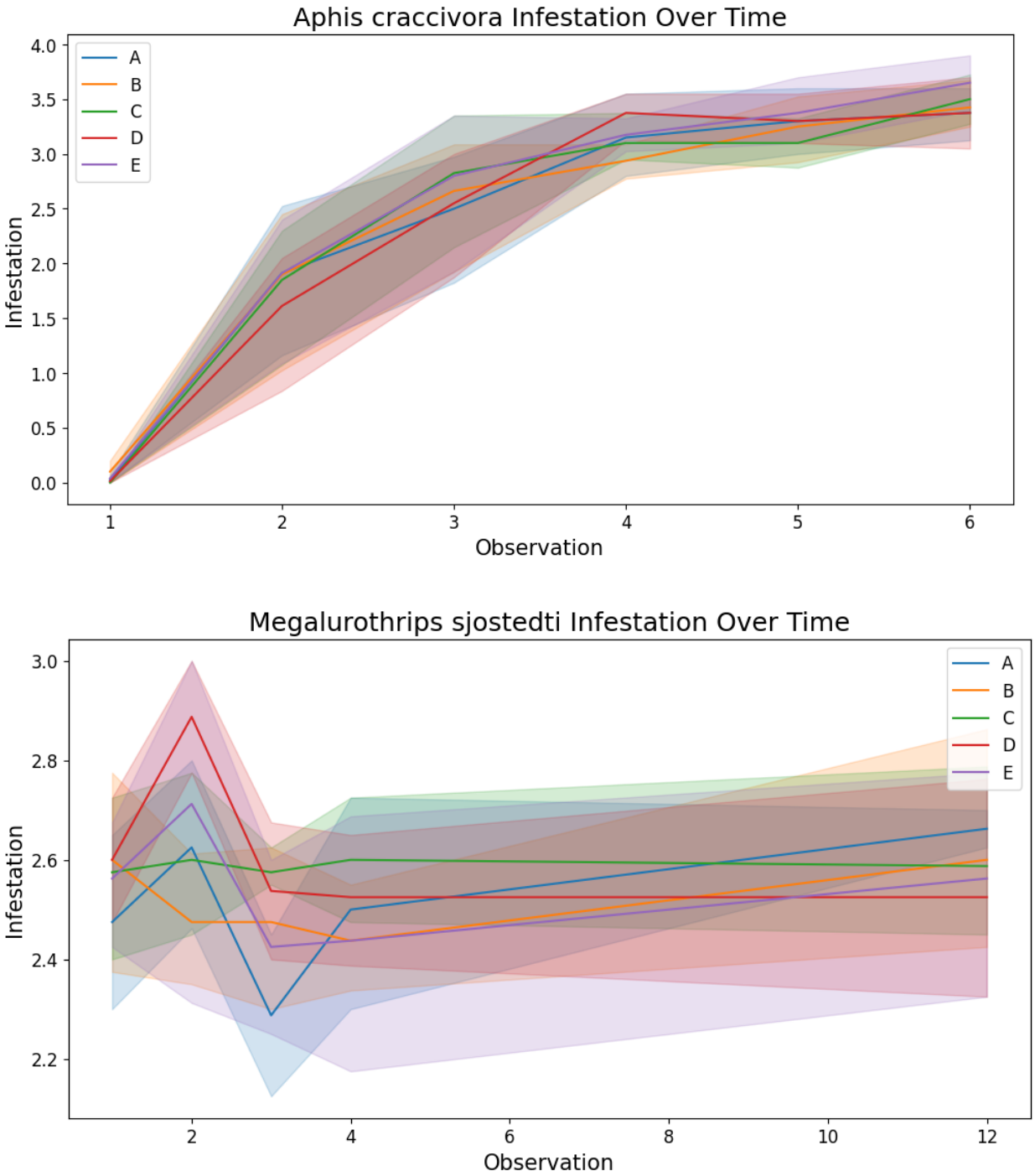

Fig. 6 Infestation observation over time based on tratment

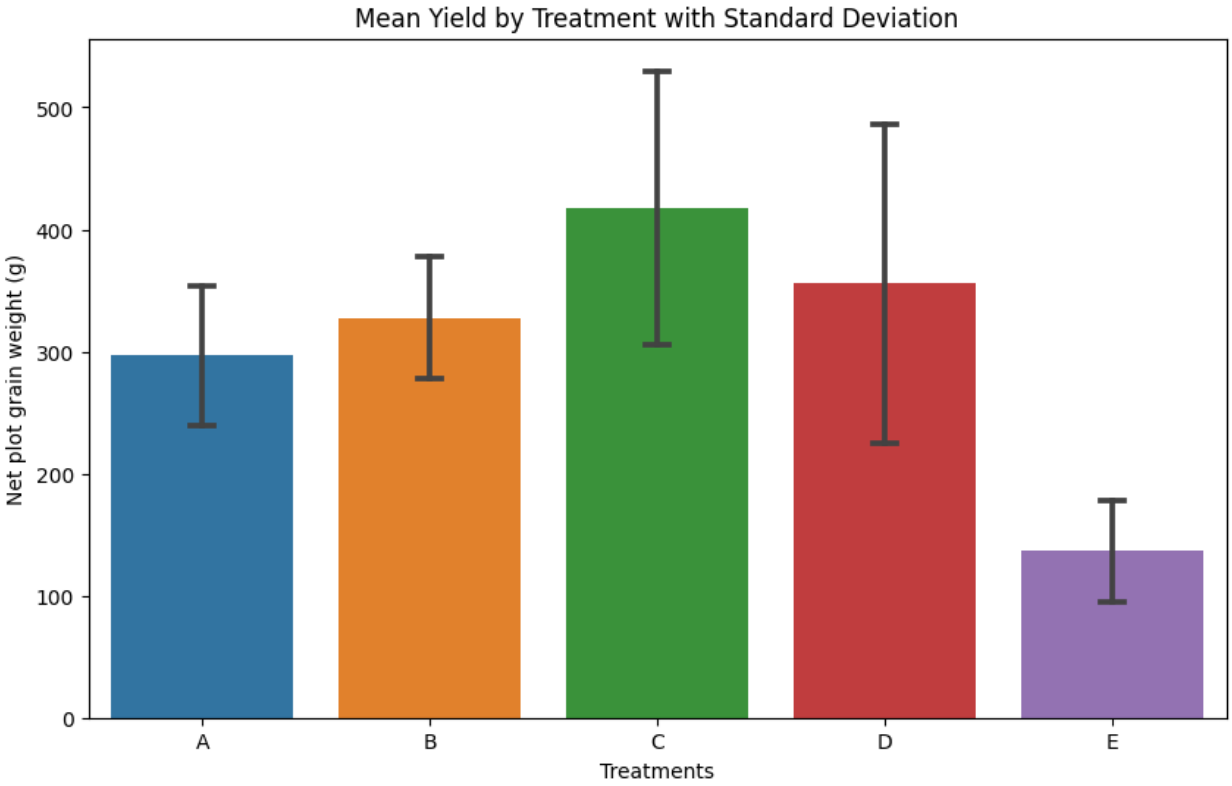

Fig. 7 Mean yield per Treatments class on each observation and net grain

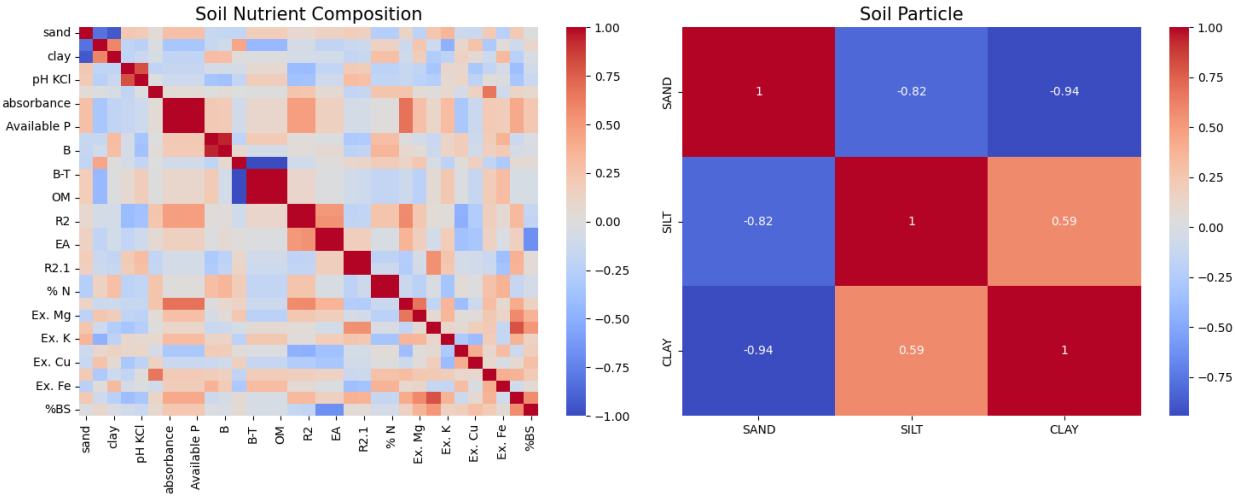

Fig. 8 Soil correlation heatmap

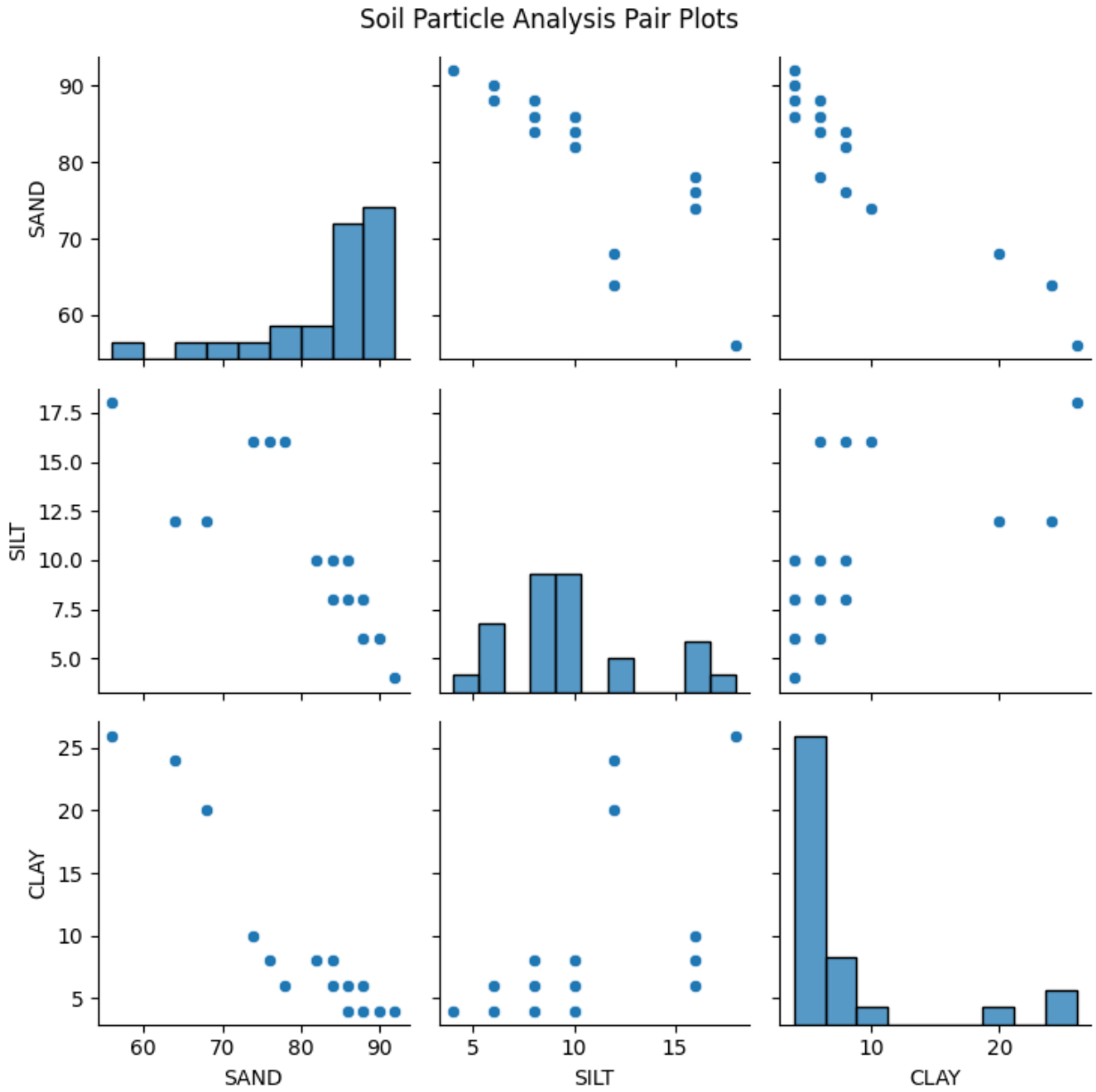

Fig. 9 Soil pair plots

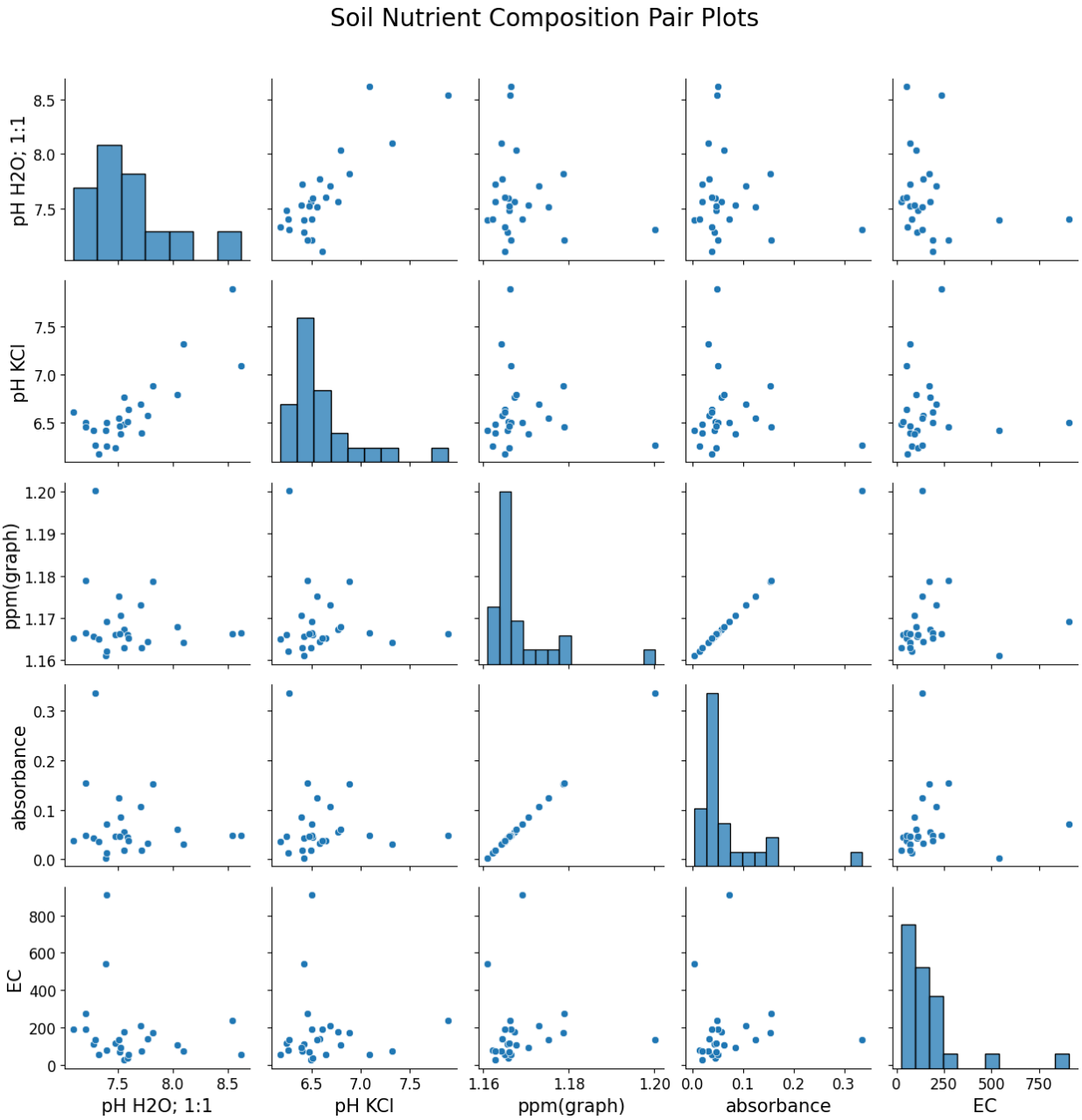

Fig. 10 Relationship pair plots of soil nutrient compositions

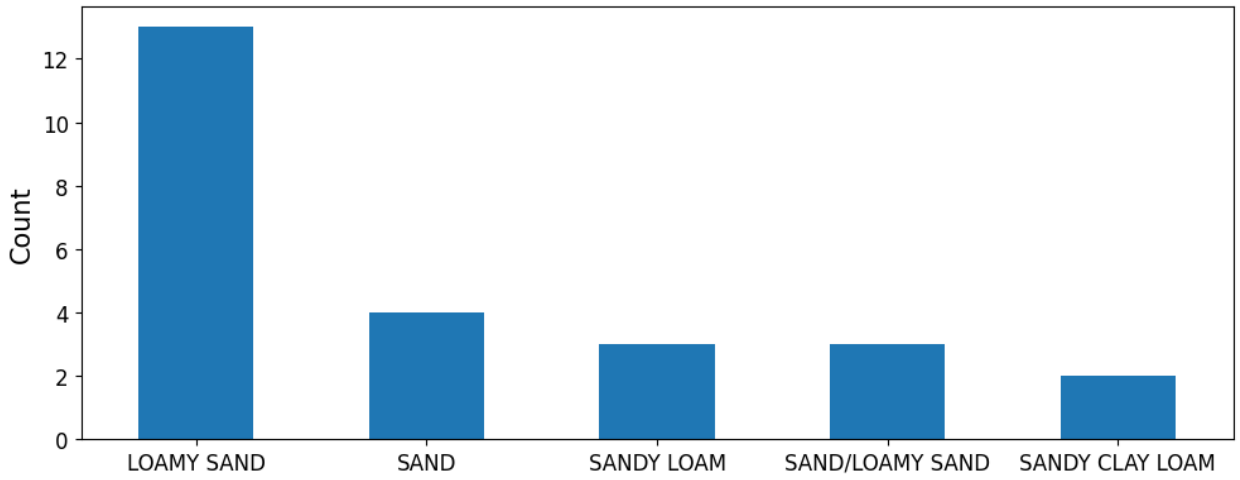

Fig. 11 Distribution of soil classes

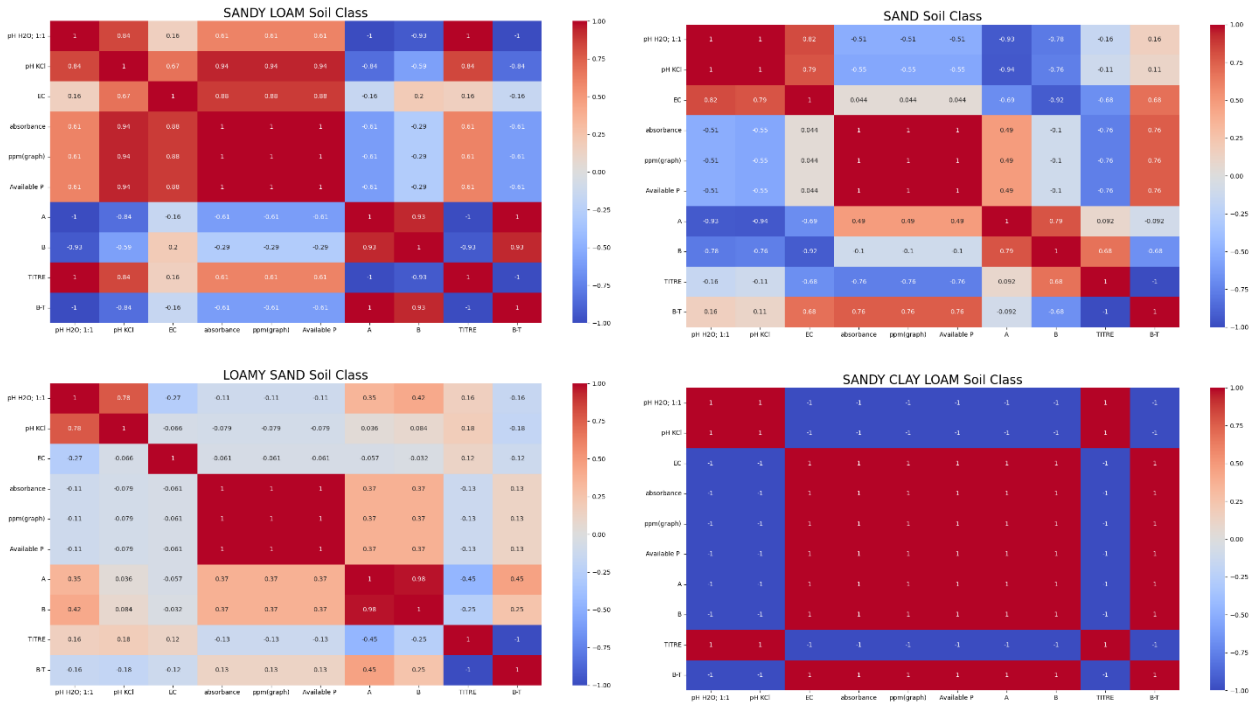

Fig. 12 Correlation matrices for each soil classes

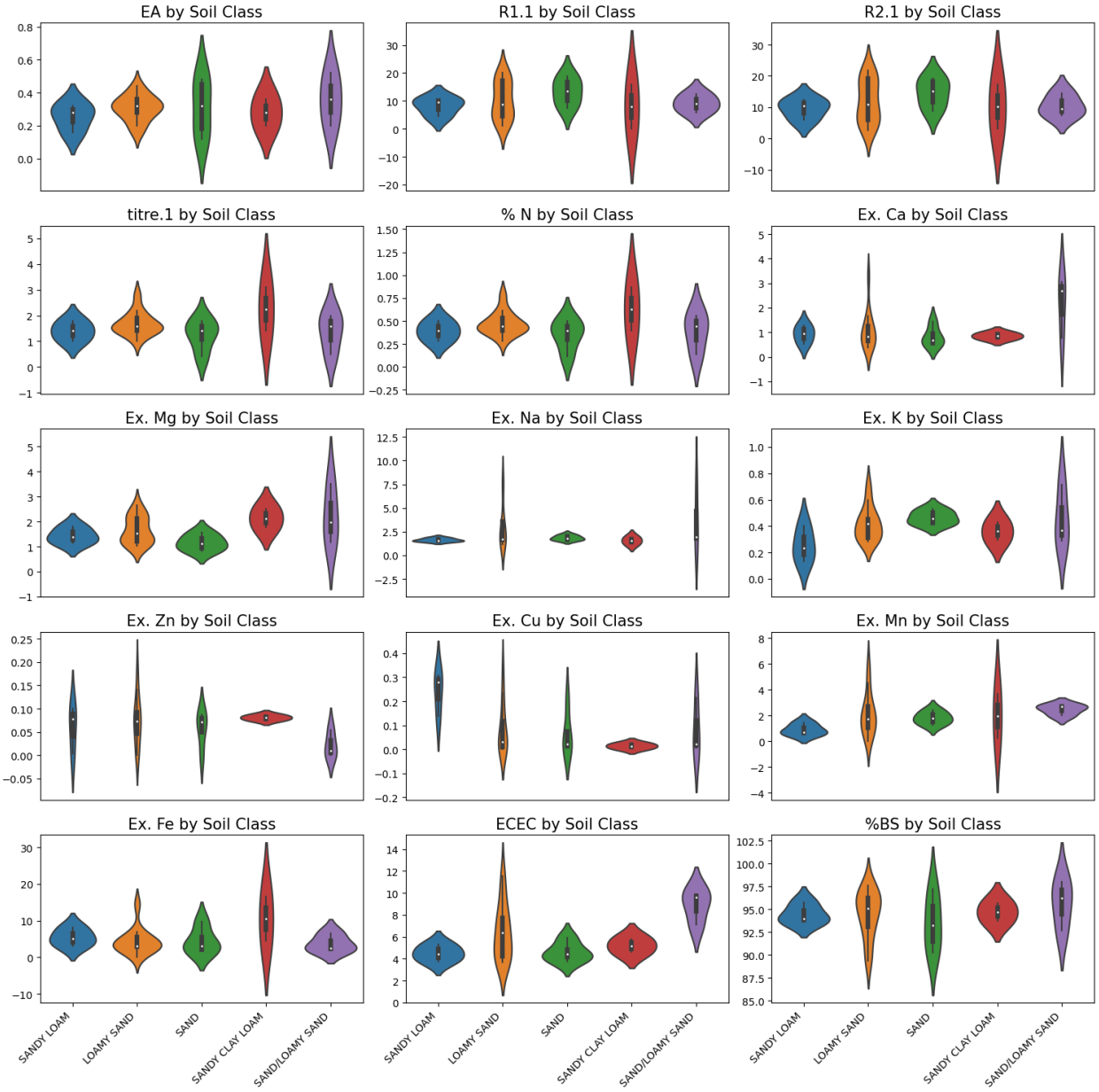

Fig. 13 Violin plot distributin of soil nutrient by the classes