EXCELLENCE IN AGRONOMY
For Sustainable Intensification
& Climate Change Adaptation

Validation Analytics Guideline

2024

Contents

Introduction

We designed this document to assist the Use Case teams in understanding the validation data set and the related standard analytics including, data preparation techniques, and the visualization of validation data. We pay specific attention to the parameters measuring effectively Agronomic gain key performance indicators (KPIs). The KPIs of each Use Case will be integrated into Excellence in Agronomy (EiA) Impact Dashboard.

While TRANSFORM is ready to analyze the validation data for display on the impact dashboard it is vital that the Use Case team engages in data cleaning, i.e. verification of valid field set-ups and data submissions. Nevertheless, we strongly encourage Use Case to run the analysis in parallel for their own insights and to ascertain the validity of the results.

The KPI documentation (Saito et al., 2022) is accessible via this link. The KPI guide provides a description of various KPIs and how to calculate them. They cover land productivity and its stability, resource use efficiency and soil health and are used across geographies, farming systems, and research and development (R&D) stages (like validation and piloting stages) (Saito et al., 2022)

The graphic below shows the various stages of the data flow during validation exercise. The highlighted section (in red) is what will be highlighted in detail in this analytics document.

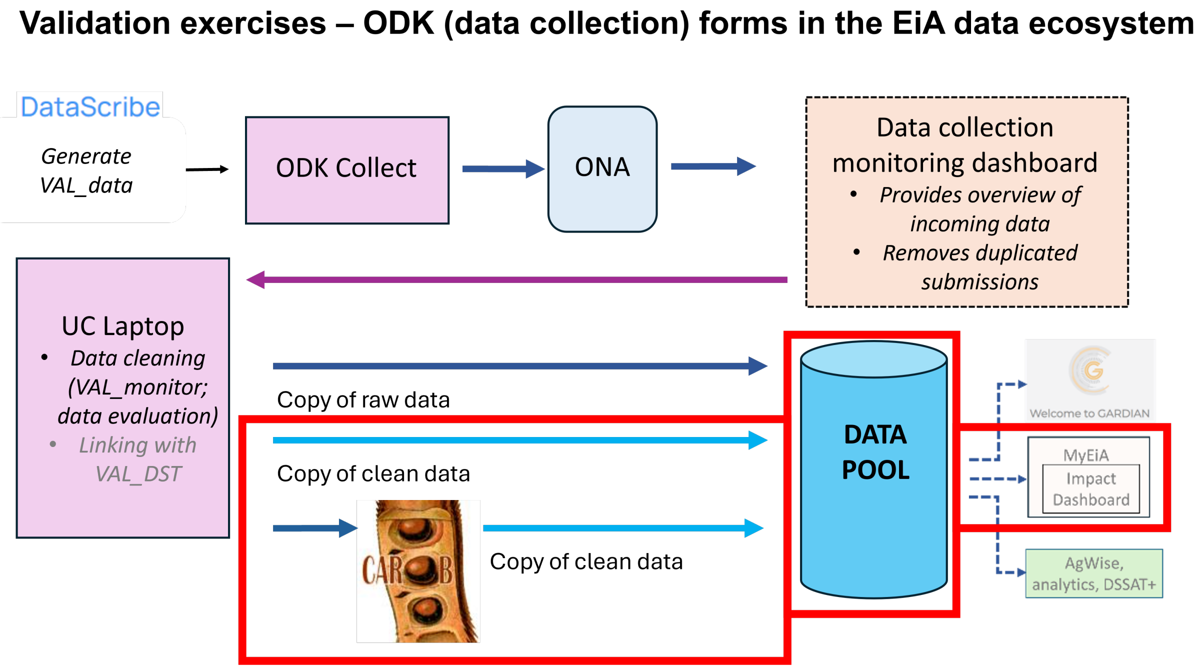
Validation exercises – ODK (data collection) forms in the EiA data ecosystem
DataScribe
Generate VAL_data
ODK Collect
ONA
Data collection monitoring dashboard
Provides overview of incoming data
Removes duplicated submissions
UC Laptop
Data cleaning (VAL_monitor; data evaluation)
Linking with VAL_DST
Copy of raw data
Copy of clean data
CAROB
Copy of clean data
DATA POOL
Welcome to GARDIAN
MyEiA Impact Dashboard
AgWise, analytics, DSSAT+

Figure 1: Validation data flow stages in the EiA ecosystem: Collected validation data is cleaned and transformed then passed to the data pool and finaly impact dashboard for visualization

Validation Analysis

Use Cases perform a validation exercise to test whether their Minimum Viable Product (MVP) meets the requirements (performance metric) that the Use Case postulated for each KPI. For more information on the protocol for validation exercises see Kreye et al., 2023.

The following steps are required before any analysis and visualization can be performed on the validation data.

KPI Parameters

Revisit your protocol and the your KPIs, Which KPIs did you select and what are the performance metrices that you expect your MVP to meet for each of these? The list of main KPIs considered in EiA Use Cases comprises Yield, Yield stability, Profit, Nutrient-use efficiency, NUE for N, NUE for P, NUE for K, Water productivity, Labor productivity, Soil organic carbon, yield-scaled GHGs, Product quality.

Below are required parameters for different KPI calculations (Saito et al., 2021).

Table 1: Table of Required KPI Variables

KPI	Detailed Indicator	Unit	Required Variables Description
Land productivity and its stability	Primary product harvested yield (referred to as yield)	kg/ha	Weight of primary harvested crop product Area of the plot where the trial was conducted
	Secondary product harvested yield	kg/ha	Weight of secondary harvested product Area of the plot where the trial was conducted
	Profit or cost–benefit balance	US$/ha or Local currency/ha	Gross revenue Total production cost
Resource use efficiency	Nutrient-use efficiency (e.g. nitrogen, phosphorus)	kg (yield)/kg (nutrient input) or kg (nutrient in yield)/kg (nutrient input)	Weight of primary harvested crop product Area of the plot where the trial was conducted Nutrient applied to a crop via inorganic fertilizer Nutrient applied to the crop via organic input
	Water productivity	kg (yield)/m3 (water input [rainfall + irrigation])	Weight of primary harvested crop product Area of the plot where the trial was conducted Total amount of irrigated water to the plot Total amount of rainfall water to the plot
	Labor productivity	kg (yield)/ work-day	Weight of primary harvested crop product Area of the plot where the trial was conducted Total number of person-day dedicated to the trial

Data Cleaning

Real data from the use cases will come in slightly different formats due to adjustments from the standard fieldbooks or ODK forms. Data cleaning entails the process of identifying and correcting errors, inconsistencies, or duplicates in a data set. In this guide we will not focus on the data cleaning steps, as these can differ from case to case depending on the purpose and the interest. However, it is important to notice that different cleaning purposes might result in different final results. Below is an example of data cleaning as performed in the Data Collection Monitoring Tool. Use cases that are collecting their data via the tool and downloading their data from there for analysis are already running some cleaning on their data. Below are examples steps.

An initial data cleaning step would be to subset the data to relevant variables. For this and the following examples we will make use of the tidyverse framework.

```r
# Load the necessary R packages
suppressPackageStartupMessages(library(tidyverse))

# Read the data
file <- "./samples/sampledata/sample_yield_data.csv"
raw.data <- read.csv(file)

# Variables to be used
vars <- c("HHID", "EAID", "Region", "T1_Yplot", "plotL1_BRR", "plotW1_BRR", "T2_Yplot",
    "plotL1_SSR", "plotW1_SSR", "riceSystem", "moistureBRR", "moistureSSR")

sub.data <- raw.data %>%
    select(vars)

# Remove duplicated enumerator entries, keeping the last entry by date of
# duplicated records
clean.data <- sub.data %>%
    distinct(EAID, .keep_all = TRUE)

# OR drop duplicated values
clean.data <- clean.data[!duplicated(clean.data), ]

#


# Here maybe we need to add an example of cleaning... Not sure what would be a
# good example- Providing a link to SAnDMan repo below for user to be guided on
# any specific issues.
```

For further detailed cleaning code details please see an example script on this Github link.

Use the above snippet as an example and substitute as per your own specific needs. For example, substitute the "…" with the list of your own variables that are not needed from your data set. In any case, the use case team must perform data cleaning to calculate the required standard parameters.

Data cleaning – what to look at:

The obvious: are there any outliers? In general, applies to all input data. Review your data collection form and check all parameters that are relevant. Yield will be a first choice, but there are also other parameters that help understanding whether the output KPI data are good enough for the validation analysis. Among these are (list is not complete…): number of plants per plot, plot dimensions, spacing, confirmations of inputs, plot management (like weeding) as well as the evaluation of the side-by-side comparison by the EA for stresses (group name) can help the assessment whether or not to include the data of a specific farm. However, do make an effort to scrutinize carefully whether data can be rightfully excluded. Keep a log with rules for exclusion.

Data Transformations

Very likely, you will need to transform your data to be able to report on your KPIs. For example, the KPI is dry matter primary yield in kg per ha and the use case has recorded the fresh yield of the product in kg), the plot length and width (m), and also the moisture content (%). From this the use case can calculate the KPI as:

$Yield = \frac{FreshWeight(kg)*(Moisture/100)}{(PlotLength(m)*PlotWidth(m))/10000}$

In another use case, the required KPI may be primary fresh yield in kg per ha following this calculation:

$Yield = \frac{FreshWeight(kg)}{(PlotLength(m)*PlotWidth(m))/10000}$

In essence, different KPIs require different variables (explained in table 1) which require specific unit conversions and data transformations. On your data, this will require some additional steps of organizing (and transforming) it to get to the final KPI calculations.

```r
# Calculate Primary Productivity Harvested Yield (PPHY)
yield.vars <- c("HHID", "EAID", "Region", "T1_Yplot", "plotL1_BRR", "plotW1_BRR",
    "riceSystem", "moistureBRR")
pphy.data <- clean.data %>%
    select(yield.vars)

# Generate Primary Productivity Harvested Yield (PPHY)
pphy <- (pphy.data$T1_Yplot * (pphy.data$moistureBRR/100))/((pphy.data$plotL1_BRR *
    pphy.data$plotW1_BRR)/10000)

# Create table with transformation
kpi.data <- cbind(pphy.data, pphy) %>%
    select("HHID", "Region", "riceSystem", "pphy")
```

The example above can be implemented for all other KPI metrics as needed by the use case. Following sections will focus on demonstrating visualization examples as produced in the EiA's impact dashboard.

Visualization

Description of terms used:

`Treatment 1 (T1): site-specific recommendation`

`Treatment 2 (T2): control`

`Treatment 3 (T3): blanket recommendation`

`Change: change in yield, profit or other KPI being measured. (Treatment 1 - Treatment 2)`

General view indicates the simple descriptive analysis of the data to show distribution (via bar chart) and difference in the different treatments (via pie chart).

The Detailed view delves into more statistical analysis to compare the distribution using a scatter plot and also show difference in the various treatments using cumulative distribution plots and boxplots to visualize distribution of nutrient-use efficiency and water-use efficiency.

I. KPI Land Productivity and its Stability

The following code illustrates how to visualize Primary Product Harvested Yield, Secondary Product Harvested Yield, and Profit.

General View

```
# Define the path to the file location
file <- "./samples/sampledata/sample_yield_data.csv"
# Read the data
dataHm.pp1 <- read.csv(file)
#Variables of interest
subset_df <- dataHm.pp1[, c("HHID", "Region","riceSystem","T2_Yha","T1_Yha","T3_Yha",
                            "eSSR","incrSSR" )]
subset_df<-distinct(subset_df)

# Group by region and calculate the averages
averages_by_state <- subset_df %>%
  group_by(Region) %>%
  summarise(
    T2 = mean(T2_Yha, na.rm = TRUE),
    T1 = mean(T1_Yha, na.rm = TRUE),
    T3 = mean(T3_Yha, na.rm = TRUE)
  )
averages_by_state <- averages_by_state %>%
  pivot_longer(!Region, names_to = "Treatment", values_to = "Average")

#plot data
ggplot(
  averages_by_state,
  aes(
  fill=Treatment,
  y=Average, x=Region
  )) +
  geom_bar(position="dodge", stat="identity") +
  xlab("Location") +
  ylab(" Average grain yield (t/ha)") +
  scale_fill_manual(values = c("T2" = "#004080", "T1" = "#4caf50", "T3" = "#c26e60")) +
  #theme as defined at start
  them2
```

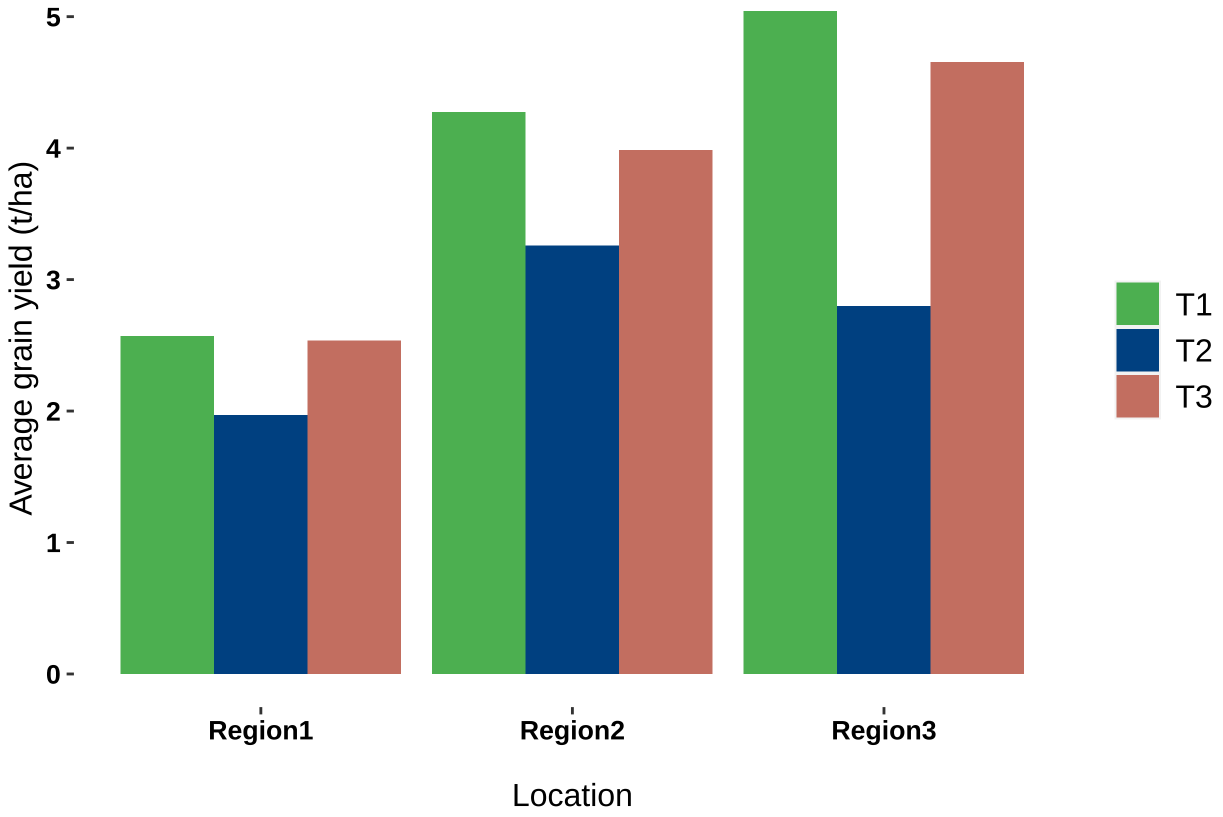

Figure 2: Yield distribution

```
#Pie plot to show  difference (positive vs negative)
#incrSSR  is the change variable (T1 - T2)

#calculate negative and positive change percentage values
x <- (subset_df[! is.na(subset_df$incrSSR),] )$incrSSR
xi <- x[x<0]
xj <- x[x>0]
pos <- (length(xj)/length(x))*100
neg <- (length(xi)/length(x))*100
ds <- data.frame(labels = c("Yield Difference <br> (T1 yield - T2 yield)",
                            "Positive change", "Negative change"),
                 values = c(NA, pos, neg))

#plot the data
plot_ly(data = ds,
        labels = ~labels,
        values = ~values,
        parents = c("", "Yield Difference <br> (T1 yield - T2 yield)",
                    "Yield Difference <br> (T1 yield - T2 yield)"),
        type = "sunburst",
        branchvalues = 'total',
        textinfo = "label+percent entry",
        hoverinfo = "text",
        hovertext = paste("% of farmers experiencing<br>",
                          tolower(ds$labels), "from T1")) %>%
  layout(title = 'Effects on grain yield of T1 vs T2')
```

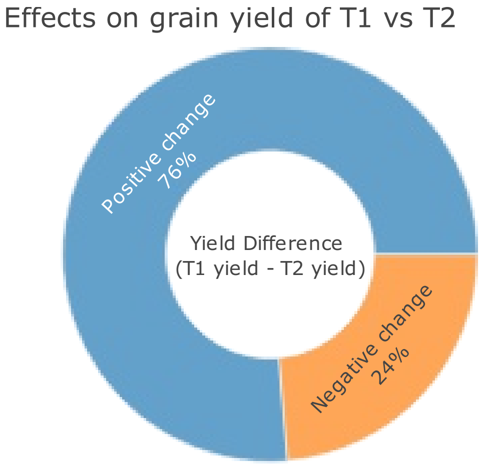

Figure 3: Effects on grain yield of T1 vs T2

Detailed View

Visualizing primary yield, secondary yield or profit distribution

```
# Detail-scatter plot

# Define the path to the file location
file <- "./samples/sampledata/sample_yield_data.csv"
# Read the data
dataHm.pp1 <- read.csv(file)
#transform data accordingly
dataHm.pp1 <- dataHm.pp1 %>%
  mutate(
    riceSystem = case_when(
      riceSystem == "rainfedLowland" ~ "Rainfed lowland",
      riceSystem == "irrigated" ~ "Irrigated",
      TRUE ~ riceSystem
    )
  )
dataHm.pp1$Region <- toTitleCase(dataHm.pp1$Region)
dataHm.pp1 <- dataHm.pp1[!is.na(dataHm.pp1$T2_Yplot),]

#plot the data
ggplot(
  dataHm.pp1,
  aes(
    T2_Yplot, yield,
    colour = plot
    )
  ) +
  geom_point(size = 1) +
  geom_abline(slope = 1, intercept = 0, size = 0.5, colour = "grey") +
  scale_x_continuous(
    minor_breaks = seq(
      min(dataHm.pp1$T2_Yplot, na.rm = TRUE),
      max(dataHm.pp1$T2_Yplot, na.rm = TRUE),
      by = 0.5
    )
  ) +
  scale_y_continuous(
    minor_breaks = seq(
      min(dataHm.pp1$yield, na.rm = TRUE),
      max(dataHm.pp1$yield, na.rm = TRUE),
      by = 0.5
    )
  ) +
  facet_wrap(~Region) +
  xlab("Grain yield (t/ha) of the control (T2)") +
  ylab(" Grain yield (T1 and T3) (t/ha)") +
  labs(title = "Yield distribution") +
   #theme as defined above
  them2
```

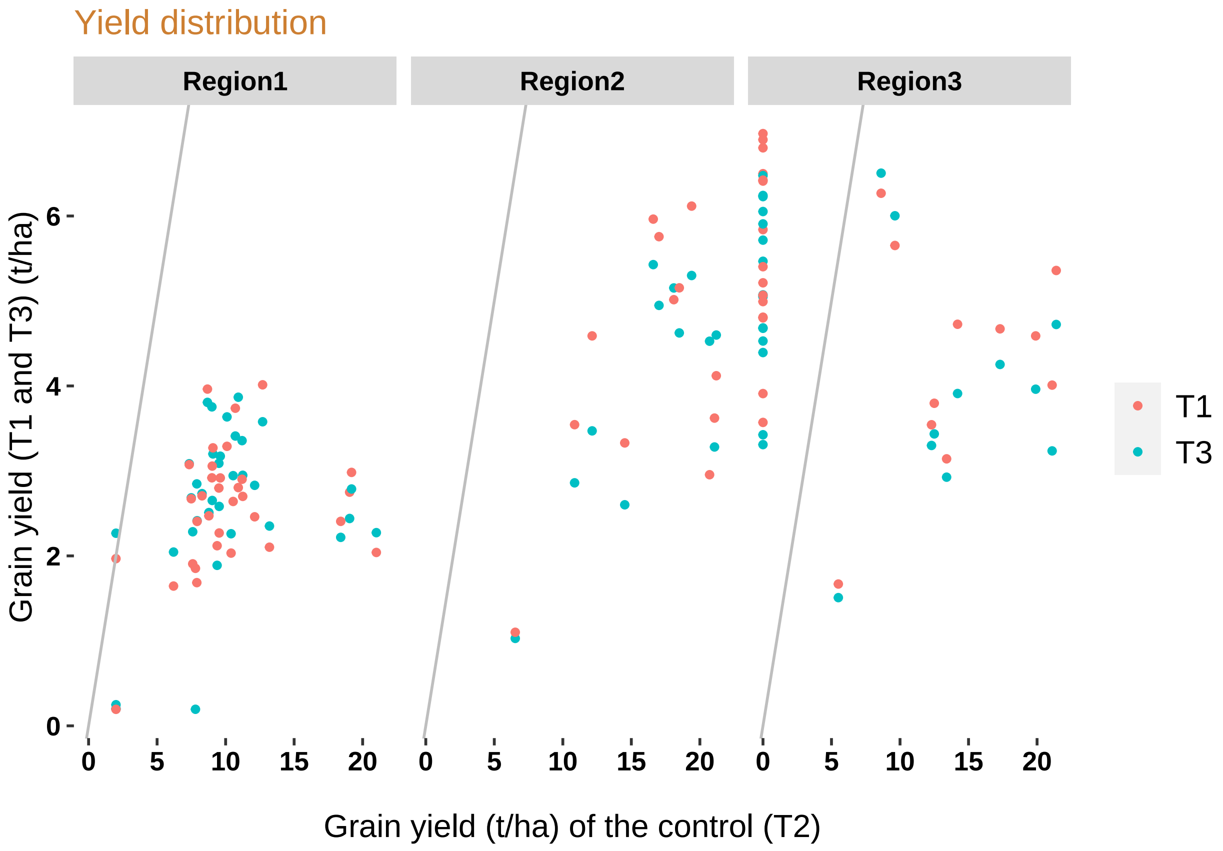

Figure 4: Yield distribution

Visualizing difference in primary yield, secondary yield or profit.

```
# Detail-cumulative distribution

#eSSR is the difference in yield between the treatments (T1-T2)
#riceSystem used to compare different attributes can also be landscapes.

# Define the path to the file location
file <- "./samples/sampledata/sample_Yieldiff.csv"
# Read the data
yieldiff <- read.csv(file)

#plot the data
p1<-ggplot(yieldiff, aes(eSSR, ecdf, colour=riceSystem))
p1+geom_point(size=1)+
  geom_ribbon(aes(ymin = lower,
                  ymax = upper,
  ),
  alpha=.2)+
  geom_vline(xintercept = 0, size = 0.5, colour = "grey")+
  xlab("Yield difference (T1 - T2) (t/ha)") +
  labs(title="Comparison of yields for site-specific (T1) and control (T2)")+
  ylab("Cumulative probability") +
  facet_wrap(~Region)+
  them2
```

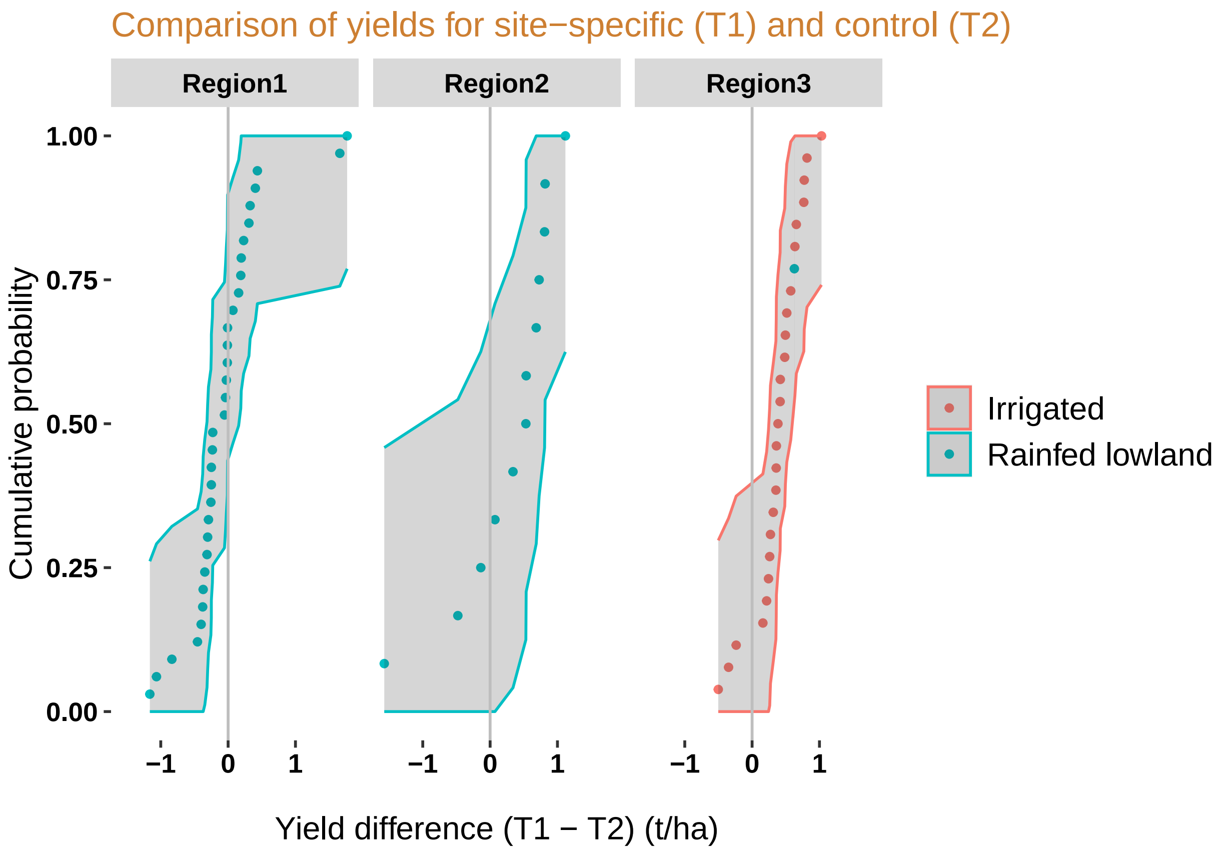

Figure 5: Comparison of yields for site-specific (T1) and control (T2

II. KPI Resource Use Efficiency

The following code illustrates how to visualize Nutrient-Use Efficiency, Water-Use Efficiency and Labor Productivity

General Overview

```
#Bar plot to show distribution for various treatments across regions

#this example shows distribution of nitrogen-use efficiency.

# Define the path to the file location
file <- "./samples/sampledata/sample_nue_data.csv"
# Read the data
dataHNm.nn1 <- read.csv(file)
#Variables of interest
nue_df <- dataHNm.nn1[, c( "Region", "plot"   , "useN" )]
#Transform data and calculate average by group
nue_df<-distinct(nue_df)
nue_by_state <- nue_df %>%
  group_by(Region, plot) %>%
  summarise(avg_useN = mean(useN, na.rm = TRUE))

#Plot the data
ggplot(
  nue_by_state,
  aes(fill=plot, y=avg_useN, x=Region)
) +
  geom_bar(position="dodge", stat="identity") +
  xlab("Location") +
  ylab("Average (Kg grain per kg applied N))") +
  scale_fill_manual(values = c("T2" = "#004080", "T1" = "#4caf50", "T3" = "#c26e60")) +
  them2
```

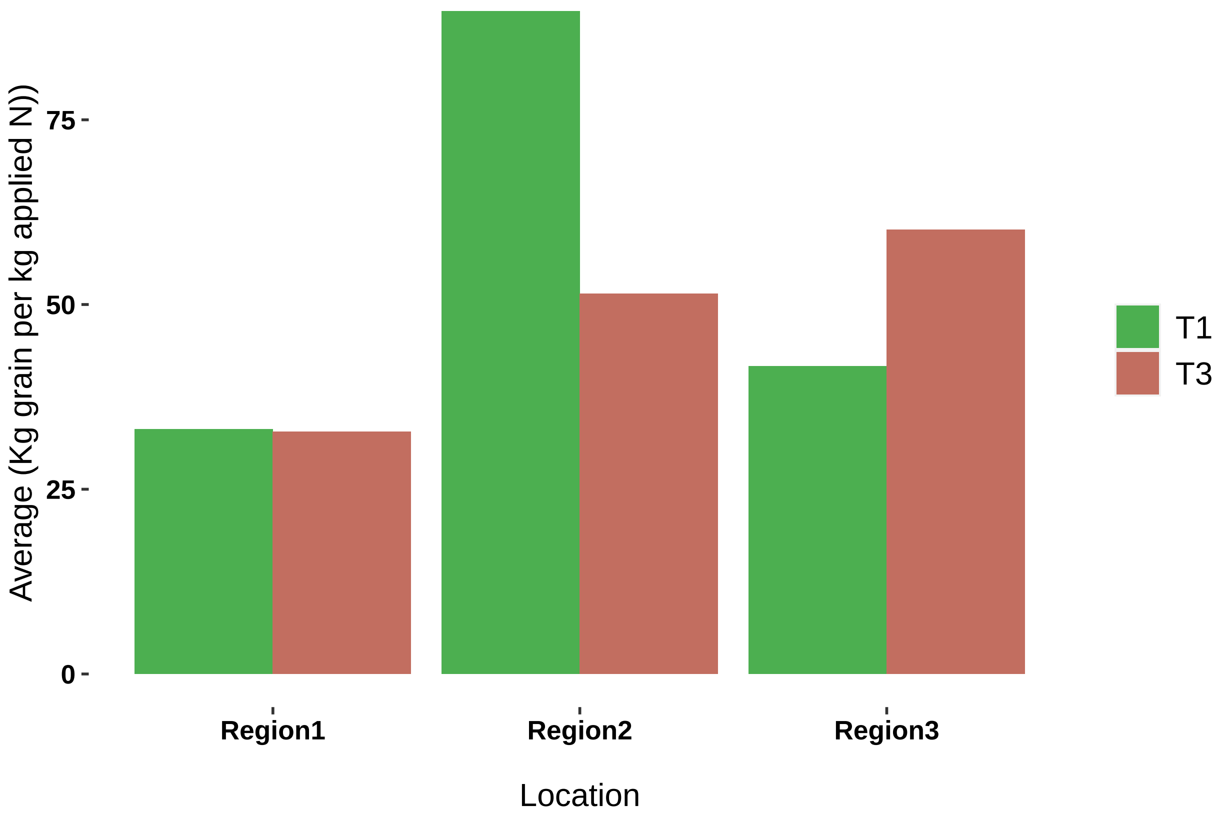

Figure 6: Nitrogen Use Efficiency

Detailed View

Distribution: Nutrient-Use Efficiency, Water-Use Efficiency and Labor Productivity

```
#Box plot- nitrogen use efficiency
ggplot(dataHNm.nn1, aes(plot, useN, fill = plot))+
  geom_boxplot()+
  facet_grid(riceSystem~Region)+
  xlab("Plot") +
  ylab("Kg grain per kg applied N") +
  them2
```

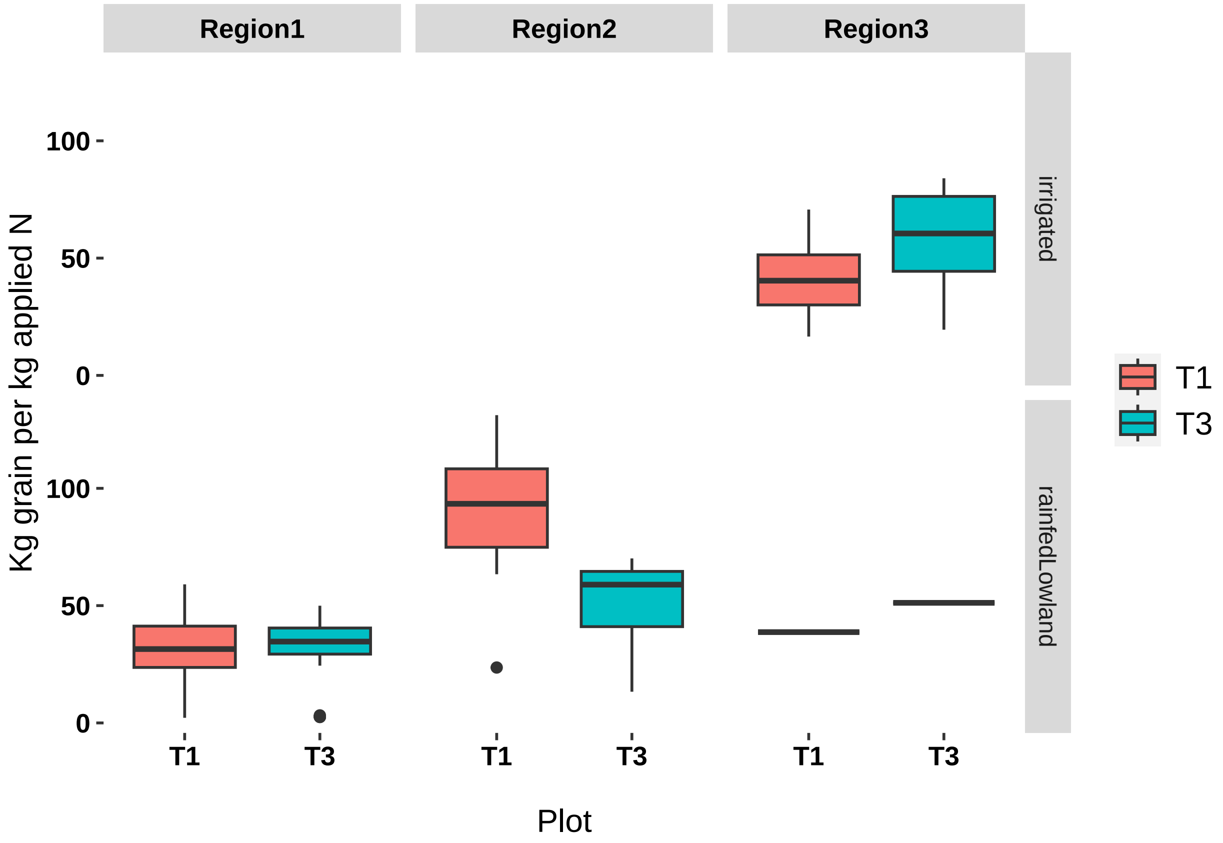

Figure 7: Nitrogen Use Efficiency

General Explanation:

References

C.Kreye, R.Flor, R.Manners, C.Aubert , 2023. Protocol for the validation exercise. Excellence in Agronomy Initiative, CGIAR, 21p.

Saito K, Johnson J-M, Hauser S, Corbeels M, Devkota M and Casimero M. 2022. Guideline for measuring agronomic gain key performance indicators in on-farm trials, v. 1. Africa Rice Center, Abidjan, Côte d'Ivoire.

Contributors

Regina Kilwenge (International Institute of Tropical Agriculture [IITA]), Eduardo Garcia (International Institute of Tropical Agriculture [IITA]) and Christine Kreye (International Institute of Tropical Agriculture [IITA])

This work was financially supported by the Excellence in Agronomy for Sustainable Intensification and Climate Change Adaptation Initiative.

Suggested Citation: