EXCELLENCE IN AGRONOMY
For Sustainable Intensification
& Climate Change Adaptation

Validation Analytics Guideline

2024

Contents

Introduction

The analytics guidelines document is designed to offer a understanding of diverse analytics methodologies, data preparation techniques, and the visualization of validation data. It equips users with the knowledge and tools/ parameters needed to derive meaningful insights from their data and effectively measure Agronomic gain key performance indicators (KPIs), which are designed to monitor, evaluate and measure the impact of changes in agronomic practices in the CGIAR Excellence in Agronomy initiative (EiA).

The use case is expected to run the analysis parallel to ascertain the validity of the data provided.

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.

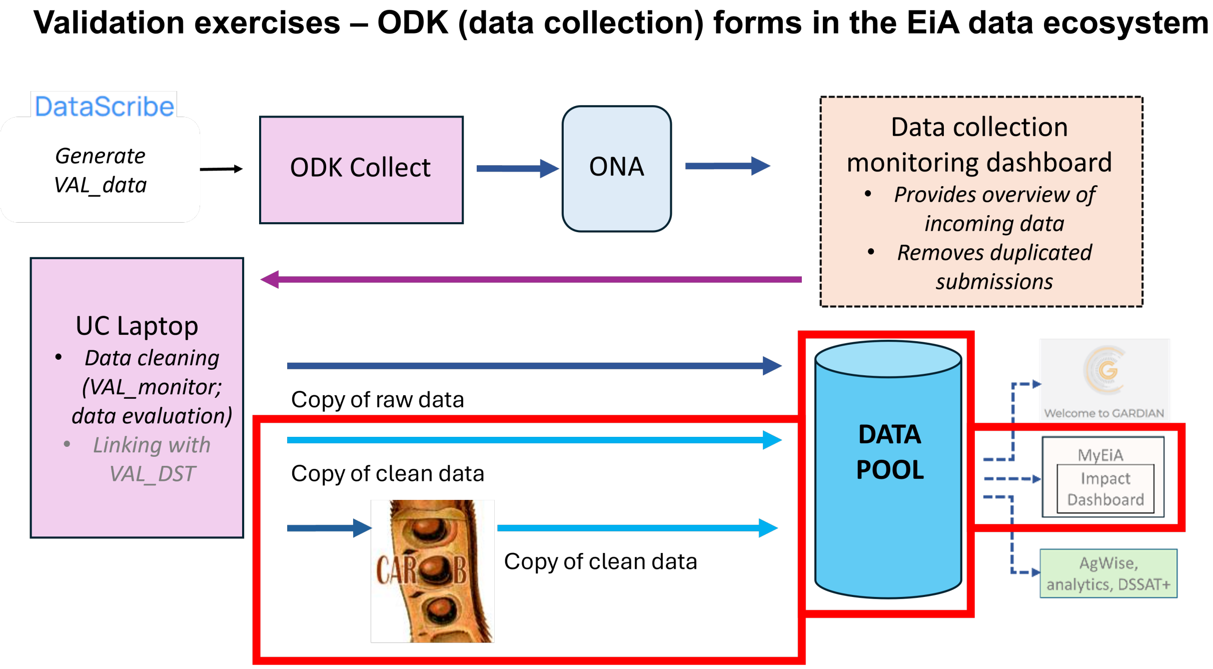
Validation exercises – ODK (data collection) forms in the EiA data ecosystem
DataScribe
Generate VAL_data
ODK Collect
ONA
Data collection monitoring dashboard
Provides overview of incoming data
Removes duplicated submissions
UC Laptop
Data cleaning (VAL_monitor; data evaluation)
Linking with VAL_DST
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 exercises - ODK (data collection) forms) in the EiA data ecosystem

Validation Analysis

Use Cases must perform a validation exercise to test the benefits of their Minimum Viable Product (MVP). 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

List of main KPIs considered in EiA Use Cases:

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. (Kreye et al., 2023).

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", "TRT1_Yplot", "plotL1_BRR", "plotW1_BRR", "TRT2_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)

# Here maybe we need to add an example of cleaning... Not sure what would be a
# good example.
```

Use the above snippet as an example qand 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 Transformations

The parameters required for the KPI metrics might be included in the raw data or not. Therefore, it is needed to make sure the use case has collected the necessary data in advance. If the necessary information is collected, there might be intermediate necessary steps to calculate the final KPI metrics. For example, a use case might have already collected the yield of the **primary** product in kg/ha, but it could be that the use case collects in one hand the fresh weight (kg), the plot length and width (m), and also the moisture content (%), in which case it would be probably interesting to calculate the dry-weight **primary product harvested yield** in dry-weight as:

`Please review this formula` $Yield = \frac{FreshWeight*(Moisture/100)}{(PlotLength*PlotWidth)/10000}$

In another use case, they might want to calculate the **primary product harvested yield** as fresh-weight, in which case:

`Please review this formula` $Yield = \frac{FreshWeight}{(PlotLength*PlotWidth)/10000}$

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 it into the final KPI calculations.

```
# Calculate Primary Productivity Harvested Yield (PPHY)
yield.vars <- c("HHID", "EAID", "Region", "TRT1_Yplot", "plotL1_BRR", "plotW1_BRR",
    "riceSystem", "moistureBRR")
pphy.data <- clean.data %>%
    select(yield.vars)

# Generate Primary Productivity Harvested Yield (PPHY)
pphy <- (pphy.data$TRT1_Yplot * (pphy.data$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 Excellence in Agronomy Impact Dashboard.

Visualization

Description of terms used:

Treatment 1 (TRT1): used to refer to the site-specific recommendation

Treatment 2 (TRT2): used to refer to control /local c/ farmer practice for which the site specific is compared with.

Treatment 3 (TRT3): in this case used to refer to other blanket recommendation. Also compared with site-specific

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

Description of how these treatments are implemented is detailed in Kreye et. al. 2023

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. 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","TRT2_Yha","TRT1_Yha","TRT3_Yha",
                            "eSSR","incrSSR" )]
subset_df<-distinct(subset_df)

# Group by region and calculate the averages
averages_by_state <- subset_df %>%
  group_by(Region) %>%
  summarise(
    TRT2 = mean(TRT2_Yha, na.rm = TRUE),
    TRT1 = mean(TRT1_Yha, na.rm = TRUE),
    TRT3 = mean(TRT3_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("TRT2" = "#004080", "TRT1" = "#4caf50", "TRT3" = "#c26e60")) +
  #theme as defined at start
  them2
```

General Explanation

```
#Pie plot to show  difference (positive vs negative)
#incrSSR  is the change variable (TRT1 - TRT2)

#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> (TRT1 yield - TRT2 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> (TRT1 yield - TRT2 yield)",
                    "Yield Difference <br> (TRT1 yield - TRT2 yield)"),
```

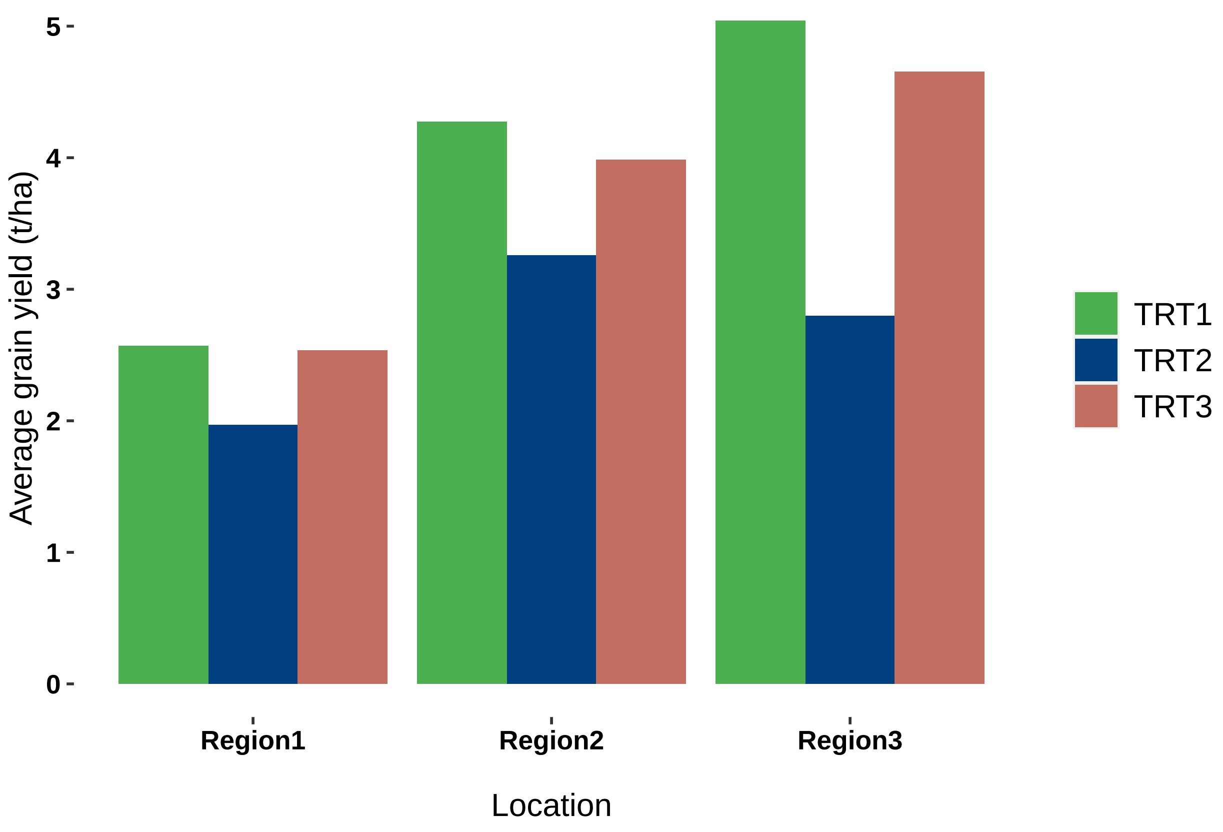

Figure 2: Yield distribution

```
      type = "sunburst",
      branchvalues = 'total',
      textinfo = "label+percent entry",
      hoverinfo = "text",
      hovertext = paste("% of farmers experiencing<br>",
                        tolower(ds$labels), "from TRT1")) %>%
layout(title = 'Effects on grain yield of TRT1 vs TRT2')
```

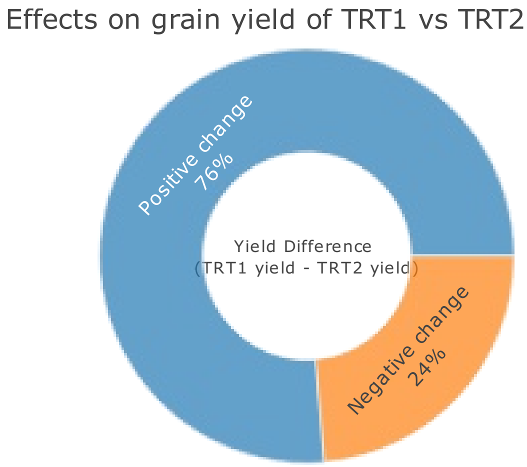

Figure 3: Effects on grain yield of TRT1 vs TRT2

General Explanation

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$TRT2_Yplot),]

#plot the data
ggplot(
  dataHm.pp1,
  aes(
    TRT2_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$TRT2_Yplot, na.rm = TRUE),
      max(dataHm.pp1$TRT2_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 (TRT2)") +
  ylab(" Grain yield (TRT1 and TRT3) (t/ha)") +
  labs(title = "Yield distribution") +
   #theme as defined above
  them2
```

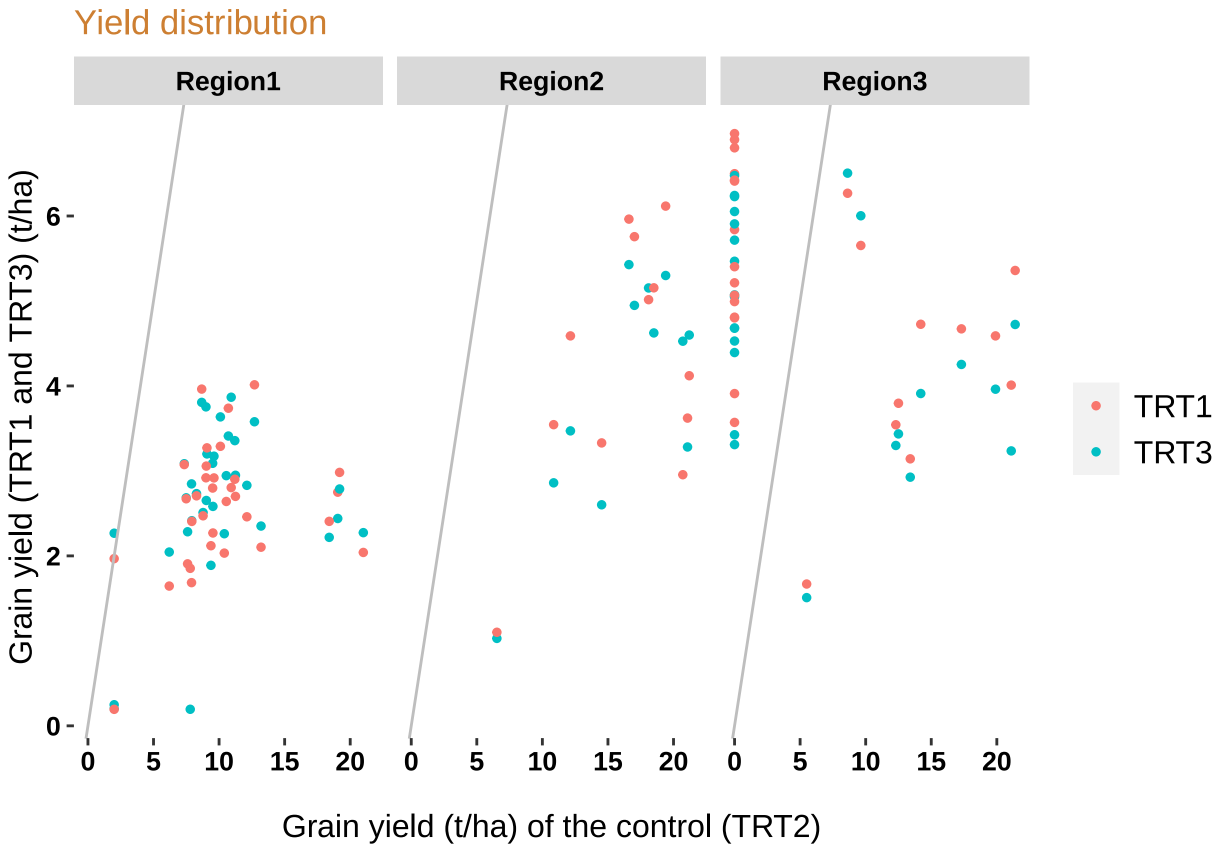

Figure 4: Yield distribution

General Explanation

Visualizing difference in primary yield, secondary yield or profit.

```
# Detail-cumulative distribution

#eSSR is the difference in yield between the treatments (trt1-trt2)
#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 (TRT1 - TRT2) (t/ha)") +
  labs(title="Comparison of yields for site-specific (TRT1) and control (TRT2)")+
  ylab("Cumulative probability") +
  facet_wrap(~Region)+
  them2
```

General Explanation

II. 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)
```

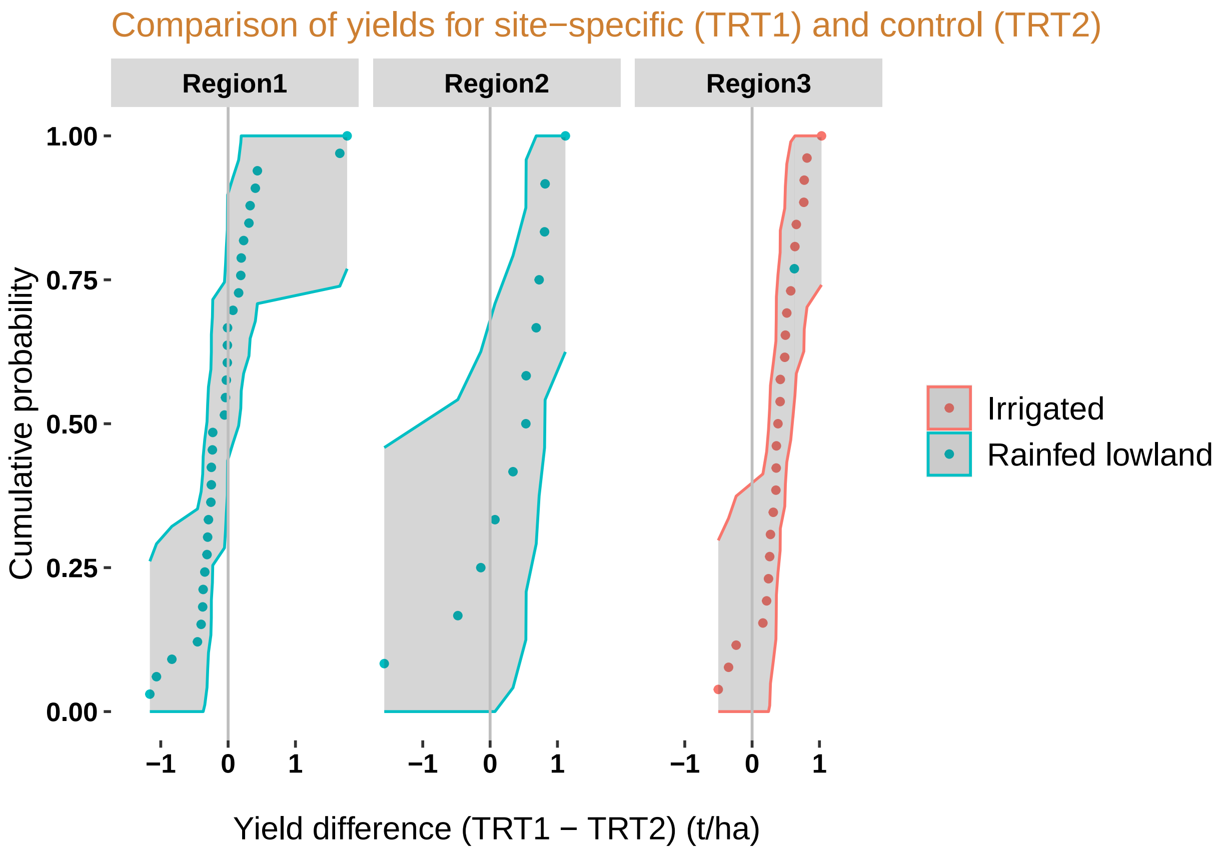

Figure 5: Comparison of yields for site-specific (TRT1) and control (TRT2

```
#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("TRT2" = "#004080", "TRT1" = "#4caf50", "TRT3" = "#c26e60")) +
  them2
```

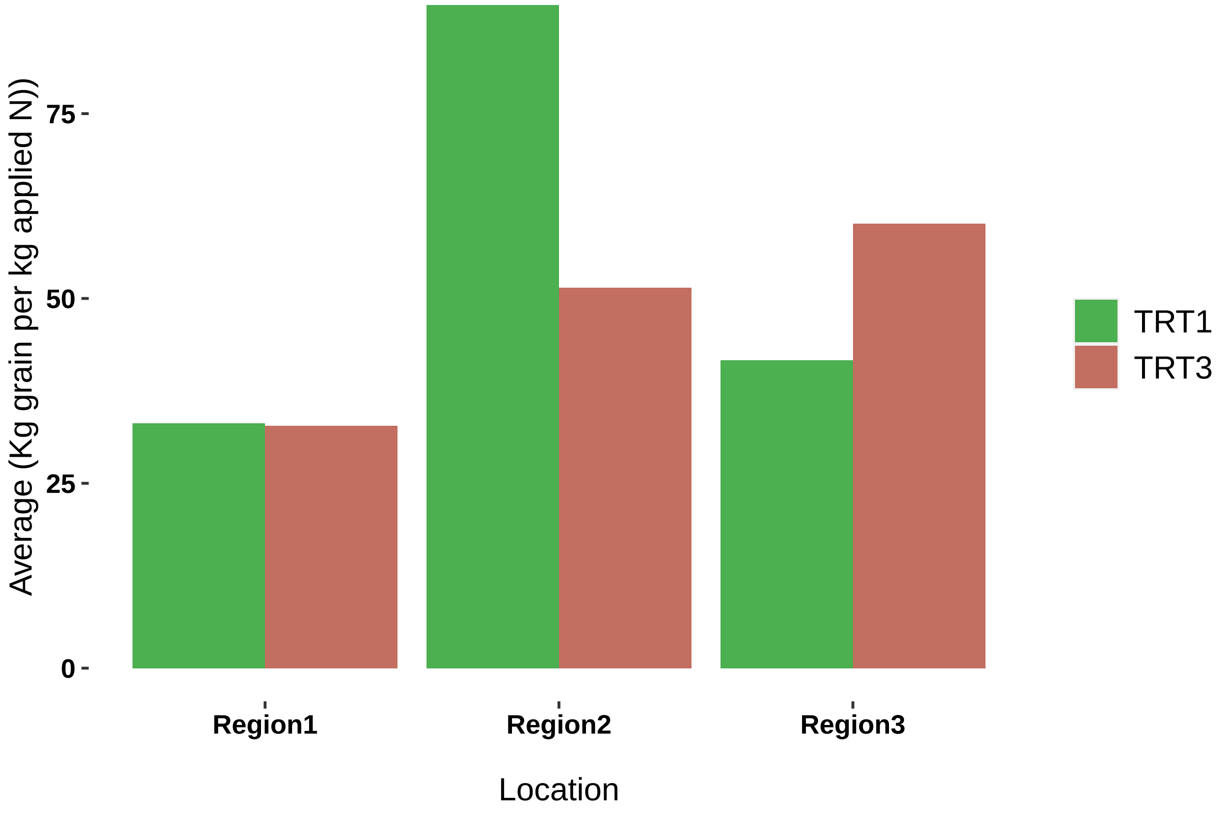

Figure 6: Nitrogen Use Efficiency

General Explanation

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
```

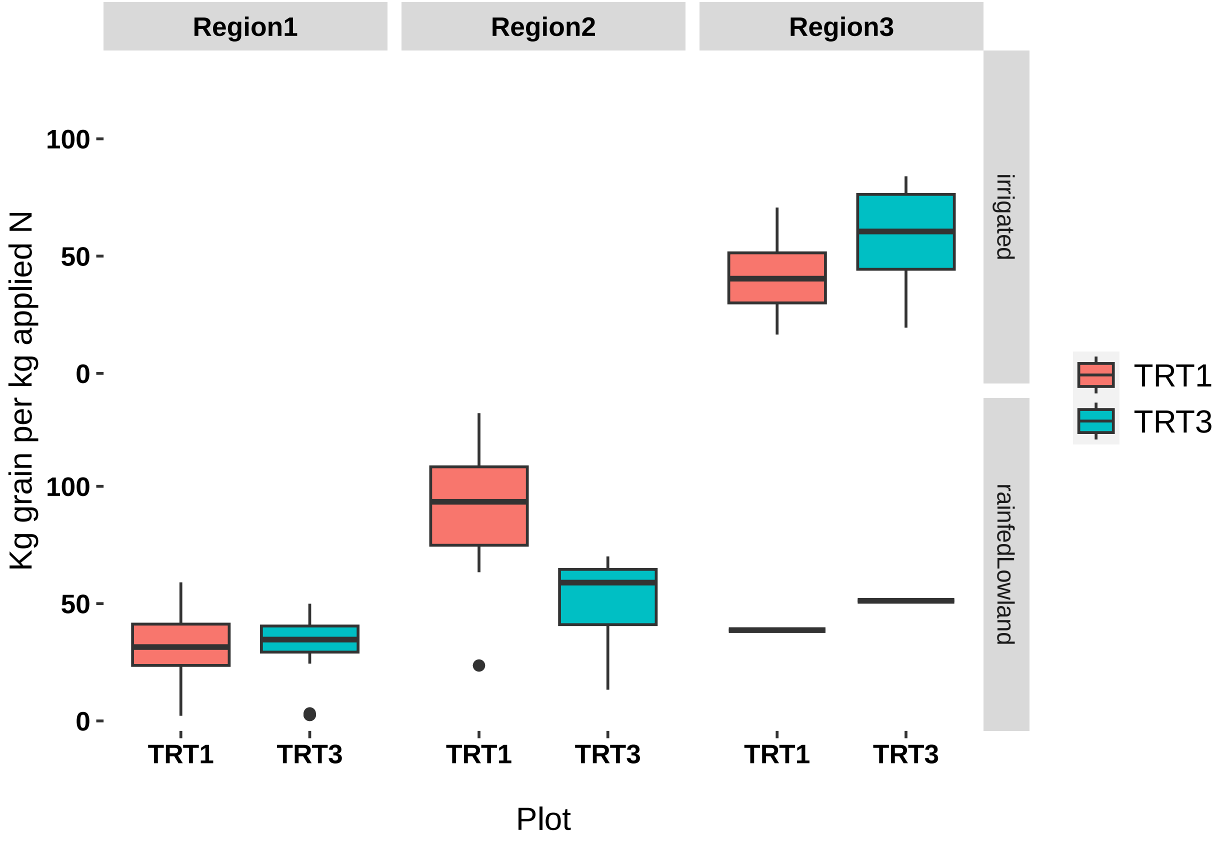

Figure 7: Nitrogen Use Efficiency

General Explanation

Add-ON Analysis

Coming soon…

Data Cleaning

The use case cleans the data and calculates standard parameters

Organising data

Add-ON Parameters

Below are required parameters for Add-on analysis

Table 2: Table of Required Add-On Variables

X…

##Visualization

Coming soon…

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:

Guidelines written by Regina Kilwenge, Eduardo Garcia and Christine Kreye.
April 2024
Excellence in Agronomy Initiative of CGIAR

Citation:

EXCELLENCE IN AGRONOMY
ADAPT INTENSIFY GROW