NaPPI Data Preparation

Elise

2024-06-09

Set the right working directory.

```
setwd("C:/Users/elise/Documents/Mémoire/Main/Data/Drive/NaPPI")
```

Packages importation

1. Data importation

The first step in this data preparation process involves importing all the pertinent datasets listed in the Google Sheets “Variables template” document. Fist we find the files, then import them.

```
##  [1] "2020_Maize_DA02_RGB1_All_Rounds.xlsx"
##  [2] "2020_NaPPI_DM.txt"
##  [3] "2020_NaPPI_env_multisensor.txt"
##  [4] "2020_NaPPI_FW.txt"
##  [5] "2020_NaPPI_RGB1.txt"
##  [6] "ISA_EPPN2020_NaPPI.xlsx"
##  [7] "NaPPI_Data-Preparation.html"
##  [8] "NaPPI_Data Preparation.Rmd"
##  [9] "NaPPI_Initial Code Draft"
## [10] "NaPPI_Template"
## [11] "NaPPI_Templatemod.gif"
## [12] "NaPPI_Templatetest.gif"
```

We can extract the coordinates of each plant with the ISA_EPPN.xlsx dataset, using a made-up function “coordinates_isaTAB”.

```
# Get the coordinates
isaTAB <- read_excel("ISA_EPPN2020_NaPPI.xlsx", sheet = "s_exp")
```

```
## New names:
## • `Unit` -> `Unit...9`
## • `Term Source REF` -> `Term Source REF...10`
## • `Term Accession Number` -> `Term Accession Number...11`
## • `Unit` -> `Unit...13`
## • `Term Source REF` -> `Term Source REF...14`
## • `Term Accession Number` -> `Term Accession Number...15`
## • `Unit` -> `Unit...22`
## • `Term Source REF` -> `Term Source REF...23`
## • `Term Accession Number` -> `Term Accession Number...24`
## • `Unit` -> `Unit...26`
## • `Term Source REF` -> `Term Source REF...27`
## • `Term Accession Number` -> `Term Accession Number...28`
```

```
coordinates <- coordinates_isaTAB(isaTAB)
```

A. Datasets structures

We can take a quick look at all the datasets.

- coordinates
- data_FW
- data_DM
- data_imaging
- data_imaging_raw
- data_environment

```
head(coordinates)
```

```
##     Sample.Name nrow ncol rep
## 1 2020_Maize001    1    1   1
## 2 2020_Maize002    1    2   1
## 3 2020_Maize003    1    3   1
## 4 2020_Maize004    1    4   1
## 5 2020_Maize005    1    5   2
## 6 2020_Maize006    1    6   2
```

```
head(data_FW)
```

```
##        Plant.ID Plant.Info    Plant.Name shoot_fresh_biomass_scale_gram  X
## 1 2020_Maize001       S1_1 UHEL_EPPN20_T                         167.78 NA
## 2 2020_Maize002       S1_1 UHEL_EPPN06_H                         219.75 NA
## 3 2020_Maize003       S2_1 UHEL_EPPN08_H                         205.48 NA
## 4 2020_Maize004       S2_1 UHEL_EPPN10_L                         103.12 NA
## 5 2020_Maize005       S2_2 UHEL_EPPN05_H                         135.96 NA
## 6 2020_Maize006       S2_2 UHEL_EPPN11_H                         118.15 NA
```

```
head(data_DM)
```

```
##        Plant.ID Plant.Info    Plant.Name shoot_dry_biomass_scale_gram
## 1 2020_Maize001       S1_1 UHEL_EPPN20_T                       35.95
## 2 2020_Maize002       S1_1 UHEL_EPPN06_H                       41.49
## 3 2020_Maize003       S2_1 UHEL_EPPN08_H                       35.77
## 4 2020_Maize004       S2_1 UHEL_EPPN10_L                       40.11
## 5 2020_Maize005       S2_2 UHEL_EPPN05_H                       27.15
## 6 2020_Maize006       S2_2 UHEL_EPPN11_H                       31.60
```

```
head(data_imaging)
```

```
## # A tibble: 6 × 23
##    rank `Measuring Date`    `Measuring Time`    `Experiment ID` `Round Order`
##   <dbl> <dttm>              <dttm>                        <dbl>         <dbl>
## 1     1 2020-06-17 08:02:12 2020-06-17 08:02:12              80             2
## 2     2 2020-06-17 08:02:51 2020-06-17 08:02:51              80             2
## 3     3 2020-06-17 08:03:29 2020-06-17 08:03:29              80             2
## 4     4 2020-06-17 08:05:03 2020-06-17 08:05:03              80             2
## 5     5 2020-06-17 08:05:42 2020-06-17 08:05:42              80             2
## 6     6 2020-06-17 08:06:20 2020-06-17 08:06:20              80             2
## # ℹ 18 more variables: `Tray ID` <chr>, `Tray Info` <lgl>, `Plant ID` <chr>,
## #   Position <chr>, `Plant Name` <chr>, `Plant Info` <chr>, PID <chr>,
## #   Angle <dbl>, `Camera Position` <dbl>, AREA_PX <dbl>, AREA_MM <dbl>,
## #   PERIMETER_PX <dbl>, PERIMETER_MM <dbl>, COMPACTNESS <dbl>, WIDTH_PX <dbl>,
## #   WIDTH_MM <dbl>, HEIGHT_PX <dbl>, HEIGHT_MM <dbl>
```

```
head(data_environment)
```

```
##             Timestamp air_temp_degree_celcius_buffer
## 1 16/06/2020 14:49:38                           34.2
## 2 16/06/2020 14:50:38                           34.2
## 3 16/06/2020 14:51:38                           34.2
## 4 16/06/2020 14:52:38                           34.3
## 5 16/06/2020 14:53:38                           34.3
## 6 16/06/2020 14:54:38                           34.4
##   air_temp_degree_celcius_tunnel air_relative_humidity_pct_buffer
## 1                           35.0                              35.6
## 2                           35.0                              35.6
## 3                           34.9                              35.6
## 4                           34.8                              35.6
## 5                           34.7                              35.7
## 6                           34.7                              35.6
##   air_relative_humidity_pct_tunnel PAR_ppfd
## 1                             35.3      123
## 2                             34.9      126
## 3                             35.6      148
## 4                             35.1      117
## 5                             35.5      144
## 6                             35.2      175
```

B. Data manipulation

This next step standardizes diverse datasets by renaming variables for consistency, converting data into appropriate units, adding necessary columns, and merging the datasets.

```
##############################################################################
# COORDINATES
##############################################################################
# Unit.ID
coordinates$Unit.ID <- seq_len(nrow(coordinates))
# Reference for Sample.Name et Unit.ID
reference <- coordinates[, c("Sample.Name", "Unit.ID")]
## We can then copy dataset2$Unit.ID <- reference$Unit.ID[match(dataset2$Sample.Name, r
eference$Sample.Name)]

##############################################################################
# DATA_FW and DATA_DM
##############################################################################
# Time, Date and Timestamp
data_FW$Date <- as.Date("2020-07-05")
data_DM$Date <- as.Date("2020-07-05")

# Name of the platform
data_FW$Platform <- "NaPPI"
data_DM$Platform <- "NaPPI"

# Unit.ID
data_FW$Unit.ID <- reference$Unit.ID[match(data_FW$Plant.ID, reference$Sample.Name)]
data_DM$Unit.ID <- reference$Unit.ID[match(data_DM$Plant.ID, reference$Sample.Name)]

# Soil
data_FW$Soil <- sapply(strsplit(as.character(data_FW$Plant.Info), split = "_"), '[', 1)
data_DM$Soil <- sapply(strsplit(as.character(data_DM$Plant.Info), split = "_"), '[', 1)

# Genotype
data_FW$Genotype <- substr(as.character(data_FW$Plant.Name), start = nchar(as.character
(data_FW$Plant.Name)) - 7, stop = nchar(as.character(data_FW$Plant.Name)))
data_DM$Genotype <- substr(as.character(data_DM$Plant.Name), start = nchar(as.character
(data_DM$Plant.Name)) - 7, stop = nchar(as.character(data_DM$Plant.Name)))


# Rename the columns for the template
data_FW$FW_shoot_g <- data_FW$shoot_fresh_biomass_scale_gram
data_DM$DW_shoot_g <- data_DM$shoot_dry_biomass_scale_gram

##############################################################################
# DATA_IMAGING
##############################################################################
# Time, Date and Timestamp
data_imaging$Timestamp <- data_imaging$`Measuring Time`

data_imaging$Date <- sapply(strsplit(as.character(data_imaging$Timestamp), split = "
"), '[', 1)
data_imaging$Time <- sapply(strsplit(as.character(data_imaging$Timestamp), split = "
"), '[', 2)

# Name of the platform
data_imaging$Platform <- "NaPPI"

# Unit.ID
```

```
data_imaging$Unit.ID <- reference$Unit.ID[match(data_imaging$`Plant ID`, reference$Samp
le.Name)]

# Rename the columns for the template
data_imaging <- rename(data_imaging,
                           S_Area_pixel = AREA_PX,
                           S_Area_mm_squared = AREA_MM,
                           S_Perimeter_pixel = PERIMETER_PX,
                           S_Perimeter_mm = PERIMETER_MM,
                           S_Compactness = COMPACTNESS,
                           S_Width_pixel = WIDTH_PX,
                           S_Width_mm = WIDTH_MM,
                           S_Height_pixel = HEIGHT_PX,
                           S_Height_mm = HEIGHT_MM
                           )
```

Camera angles

For the NaPPI platform, the variables in the data_imaging are measured form 3 different camera angles. It is neccessary to consolidate theses measurements into a single value for each variable for the data analysis steps. In this code block, it is done by either taking the maximum of the 3 values or by taking the mean of the 3 values.

Depending on the variable, the mean or the maximum is taken. The result is stocked in the dataset data_imaging_2.

Variable	Mean or maximum
Height	Mean
Area	Maximum
Perimeter	Maximum
Width	Maximum
Convex hull	Maximum
Solidity	Maximum
Compactness	Maximum

```
# Data frame containing the results
data_imaging_2 <- data.frame()

for (i in seq(1, nrow(data_imaging), by = 3)) {
  if (i + 1 <= nrow(data_imaging)) {
    row1 <- data_imaging[i, ]
    row2 <- data_imaging[i + 1, ]
    row3 <- data_imaging[i + 2, ]

    # Compute the mean or the maximum of the 3 camera angles values
    mean_and_max_row <- data.frame(
      Date = row3$Date, # We keep the important columns
      Time = row3$Time, # We keep the important columns
      Unit.ID = row3$Unit.ID, # We keep the important columns
      Timestamp = row3$Timestamp, # We keep the important columns
      Platform = row3$Platform, # We keep the important columns

      S_Area_mm_squared = max(c(as.numeric(row1$S_Area_mm_squared), as.numeric(row2$S_A
rea_mm_squared), as.numeric(row3$S_Area_mm_squared))),
      S_Area_pixel = max(c(as.numeric(row1$S_Area_pixel), as.numeric(row2$S_Area_pixe
l), as.numeric(row3$S_Area_pixel))),

      S_Perimeter_mm = max(c(as.numeric(row1$S_Perimeter_mm), as.numeric(row2$S_Perimet
er_mm), as.numeric(row3$S_Perimeter_mm))),
      S_Perimeter_pixel = max(c(as.numeric(row1$S_Perimeter_pixel), as.numeric(row2$S_P
erimeter_pixel), as.numeric(row3$S_Perimeter_pixel))),


      S_Compactness = max(c(as.numeric(row1$S_Compactness), as.numeric(row2$S_Compactne
ss), as.numeric(row3$S_Compactness))),

      S_Width_mm = max(c(as.numeric(row1$S_Width_mm), as.numeric(row2$S_Width_mm), as.n
umeric(row3$S_Width_mm))),
      S_Width_pixel = max(c(as.numeric(row1$S_Width_pixel), as.numeric(row2$S_Width_pix
el), as.numeric(row3$S_Width_pixel))),


      S_Height_mm = mean(c(as.numeric(row1$S_Height_mm), as.numeric(row2$S_Height_mm),
as.numeric(row3$S_Height_mm))),
      S_Height_pixel = mean(c(as.numeric(row1$S_Height_pixel), as.numeric(row2$S_Height
_pixel), as.numeric(row3$S_Height_pixel)))
    )

    # Ajouter les résultats au dataframe final
    data_imaging_2 <- rbind(data_imaging_2, mean_and_max_row)
  }
}

# Afficher le dataframe final
head(data_imaging_2)
```

```
##         Date     Time Unit.ID           Timestamp Platform S_Area_mm_squared
## 1 2020-06-17 08:03:29       1 2020-06-17 08:03:29    NaPPI          4295.914
## 2 2020-06-17 08:06:20       2 2020-06-17 08:06:20    NaPPI         14468.041
## 3 2020-06-17 08:09:10       3 2020-06-17 08:09:10    NaPPI          6907.980
## 4 2020-06-17 08:11:58       4 2020-06-17 08:11:58    NaPPI          9650.059
## 5 2020-06-17 08:14:45       5 2020-06-17 08:14:45    NaPPI          8383.577
## 6 2020-06-17 08:17:32       6 2020-06-17 08:17:32    NaPPI          6479.683
##   S_Area_pixel S_Perimeter_mm S_Perimeter_pixel S_Compactness S_Width_mm
## 1        14935       1743.909          3251.611    0.01053193   830.2257
## 2        50299       3653.811          6812.726    0.19957863   536.3215
## 3        24016       1736.060          3236.976    0.35945639   243.4900
## 4        33549       2162.799          4032.654    0.01622185   742.2689
## 5        29146       2201.717          4105.219    0.34150800   262.7975
## 6        22527       1971.969          3676.841    0.03201772   539.0031
##   S_Width_pixel S_Height_mm S_Height_pixel
## 1          1548   1518.3261      2831.0000
## 2          1000    732.4364      1365.6667
## 3           454    357.7264       667.0000
## 4          1384   1559.8016      2908.3333
## 5           490    393.8387       734.3333
## 6          1005   1565.5224      2919.0000
```

Unit conversions

The data template is only in cm, cm² and g. This step converts the data in the right units.

For the NaPPI platform, 4 variables are in mm.

```
data_imaging_2$S_Height_cm <- 0.01 * data_imaging_2$S_Height_mm
data_imaging_2$S_Area_cmsquared <- 0.01 * 0.01 * data_imaging_2$S_Area_mm_squared
data_imaging_2$S_Perimeter_cm <- 0.01 * data_imaging_2$S_Perimeter_mm
data_imaging_2$S_Width_cm <- 0.01 * data_imaging_2$S_Width_mm
```

2. Data template

A. Data template: plant_info

This dataset contains information about the plant: Unit.ID, genotype, replication, row and column location in the greenhouse, and soil treatment.

B. Data template: endpoint

This datasets contains information of the end of the experiment (variables at harvest). It is then linked by the Unit.ID to the plant_info data template.

C. Data template: timeseries

This section in divided in three data templates:

- timeseries
- S_timeseries (variables computed from sideview imaging or image processing)
- T_timeseries (variables computed from topview imaging or image processing)

The time interval between data timestamps varies in each platform. They are then linked by the Unit.ID to the plant_info data template.

D. NaPPI data templates

- plant_info
- endpoint
- timeseries
- S_timeseries
- T_timeseries

```
##   Unit.ID Genotype Soil Replication Row Column Platform
## 1       1 EPPN20_T   S1           1   1      1    NaPPI
## 2       2 EPPN06_H   S1           1   1      2    NaPPI
## 3       3 EPPN08_H   S2           1   1      3    NaPPI
## 4       4 EPPN10_L   S2           1   1      4    NaPPI
## 5       5 EPPN05_H   S2           2   1      5    NaPPI
## 6       6 EPPN11_H   S2           2   1      6    NaPPI
```

```
##   Unit.ID Time       Date Timestamp DW_shoot_g FW_shoot_g DW_root_g FW_root_g
## 1       1   NA 2020-07-05        NA      35.95     167.78        NA        NA
## 2       2   NA 2020-07-05        NA      41.49     219.75        NA        NA
## 3       3   NA 2020-07-05        NA      35.77     205.48        NA        NA
## 4       4   NA 2020-07-05        NA      40.11     103.12        NA        NA
## 5       5   NA 2020-07-05        NA      27.15     135.96        NA        NA
## 6       6   NA 2020-07-05        NA      31.60     118.15        NA        NA
##   Leaf_number Plant_height_cm DW_plant_g Root_length_cm Root_number Root_angle
## 1          NA              NA         NA             NA          NA         NA
## 2          NA              NA         NA             NA          NA         NA
## 3          NA              NA         NA             NA          NA         NA
## 4          NA              NA         NA             NA          NA         NA
## 5          NA              NA         NA             NA          NA         NA
## 6          NA              NA         NA             NA          NA         NA
##   Total_wu DW_seed_g FW_seed_g Leaf_area_cmsquared Genotype Soil Replication
## 1       NA        NA        NA                  NA EPPN20_T   S1           1
## 2       NA        NA        NA                  NA EPPN06_H   S1           1
## 3       NA        NA        NA                  NA EPPN08_H   S2           1
## 4       NA        NA        NA                  NA EPPN10_L   S2           1
## 5       NA        NA        NA                  NA EPPN05_H   S2           2
## 6       NA        NA        NA                  NA EPPN11_H   S2           2
##   Row Column Platform
## 1   1      1    NaPPI
## 2   1      2    NaPPI
## 3   1      3    NaPPI
## 4   1      4    NaPPI
## 5   1      5    NaPPI
## 6   1      6    NaPPI
```

```
##   Unit.ID Time Date Timestamp Manual_Plant_height_cm Leaf_number Wue
## 1    <NA>   NA   NA        NA                     NA          NA  NA
##   Plant_biomass Ligulated_leaf_number Plant_emergence Plant_transpiration
## 1            NA                    NA              NA                  NA
##   Daily_wu Soil_water_potential Genotype Soil Replication  Row Column Platform
## 1       NA                   NA     <NA> <NA>        <NA> <NA>   <NA>     <NA>
```

```
##   Unit.ID           Timestamp       Date     Time S_Height_cm S_Height_pixel
## 1       1 2020-06-17 08:03:29 2020-06-17 08:03:29   15.183261      2831.0000
## 2       2 2020-06-17 08:06:20 2020-06-17 08:06:20    7.324364      1365.6667
## 3       3 2020-06-17 08:09:10 2020-06-17 08:09:10    3.577264       667.0000
## 4       4 2020-06-17 08:11:58 2020-06-17 08:11:58   15.598016      2908.3333
## 5       5 2020-06-17 08:14:45 2020-06-17 08:14:45    3.938387       734.3333
## 6       6 2020-06-17 08:17:32 2020-06-17 08:17:32   15.655224      2919.0000
##   S_Area_cmsquared S_Area_pixel S_Perimeter_cm S_Perimeter_pixel
## 1        0.4295914        14935       17.43909          3251.611
## 2        1.4468041        50299       36.53811          6812.726
## 3        0.6907980        24016       17.36060          3236.976
## 4        0.9650059        33549       21.62799          4032.654
## 5        0.8383577        29146       22.01717          4105.219
## 6        0.6479683        22527       19.71969          3676.841
##   S_Convex_hull_area_cmsquared S_Solidity S_Compactness S_Width_cm
## 1                           NA         NA    0.01053193   8.302257
## 2                           NA         NA    0.19957863   5.363215
## 3                           NA         NA    0.35945639   2.434900
## 4                           NA         NA    0.01622185   7.422689
## 5                           NA         NA    0.34150800   2.627975
## 6                           NA         NA    0.03201772   5.390031
##   S_Width_pixel S_Leaf_area_cmsquared Genotype Soil Replication Row Column
## 1          1548                    NA EPPN20_T   S1           1   1      1
## 2          1000                    NA EPPN06_H   S1           1   1      2
## 3           454                    NA EPPN08_H   S2           1   1      3
## 4          1384                    NA EPPN10_L   S2           1   1      4
## 5           490                    NA EPPN05_H   S2           2   1      5
## 6          1005                    NA EPPN11_H   S2           2   1      6
##   Platform
## 1    NaPPI
## 2    NaPPI
## 3    NaPPI
## 4    NaPPI
## 5    NaPPI
## 6    NaPPI
```

```
##   Unit.ID Time Date Timestamp T_Area_cm_squared T_Area_pixel T_Perimeter_cm
## 1    <NA>   NA   NA        NA                NA           NA             NA
##   T_Perimeter_pixel T_Convex_hull_area_cmsquared T_Solidity T_Compactness
## 1                NA                           NA         NA            NA
##   T_Roundness T_Roundness2 T_Isotropy T_Eccentricity T_Rms T_Sol Genotype Soil
## 1          NA           NA         NA             NA    NA    NA     <NA> <NA>
##   Replication  Row Column Platform
## 1        <NA> <NA>   <NA>     <NA>
```

3. Export the data templates in .txt

Stock the new data sets in a new folder.

```
setwd("C:/Users/elise/Documents/Mémoire/Main/Data/Templates/NaPPI")

write.table(plant_info, file = "plant_info.txt", sep = "\t", row.names = FALSE, quote = FALSE)
write.table(endpoint, file = "endpoint.txt", sep = "\t", row.names = FALSE, quote = FALSE)
write.table(timeseries, file = "timeseries.txt", sep = "\t", row.names = FALSE, quote = FALSE)
write.table(S_timeseries, file = "S_timeseries.txt", sep = "\t", row.names = FALSE, quote = FALSE)
write.table(T_timeseries, file = "T_timeseries.txt", sep = "\t", row.names = FALSE, quote = FALSE)
```