Fertilizer advice for farmers growning maize in Nigeria DST (Decission Support Tools)

Ivana ALEKSOVSKA

2024-06-09

Carob

Carob creates reproducible workflows that standardize primary agricultural research data from experiments and surveys. Standardization includes the use of a common file format, variable names, units and accepted values according to the terminag standard. Standardized data sets are aggregated into larger collections that can be used in further research. We do this by writing an *R* script for each individual dataset. See the website for more information.

Carob is an open access *Extract, Transform, and Load* (ETL) framework supported by CGIAR to support predictive analytics (machine learning, artifical intelligence) and other types of data analysis.

Contributions are welcome from anyone, and they can be made via pull-requests. Feel free to improve these scripts, or provide new ones. See the [Guidelines for contributors] for instructions on how to write a Carob script, and follow the steps described here. You can also raise issues on this github site. A good place to discover new data sets is the Gardian website or our to-do list.

Get the data

Compiled versions of the dataset can be downloaded from carob-data.org and some will eventually be made available on the carob dataverse.

You can also compile your own version by cloning the repo and running

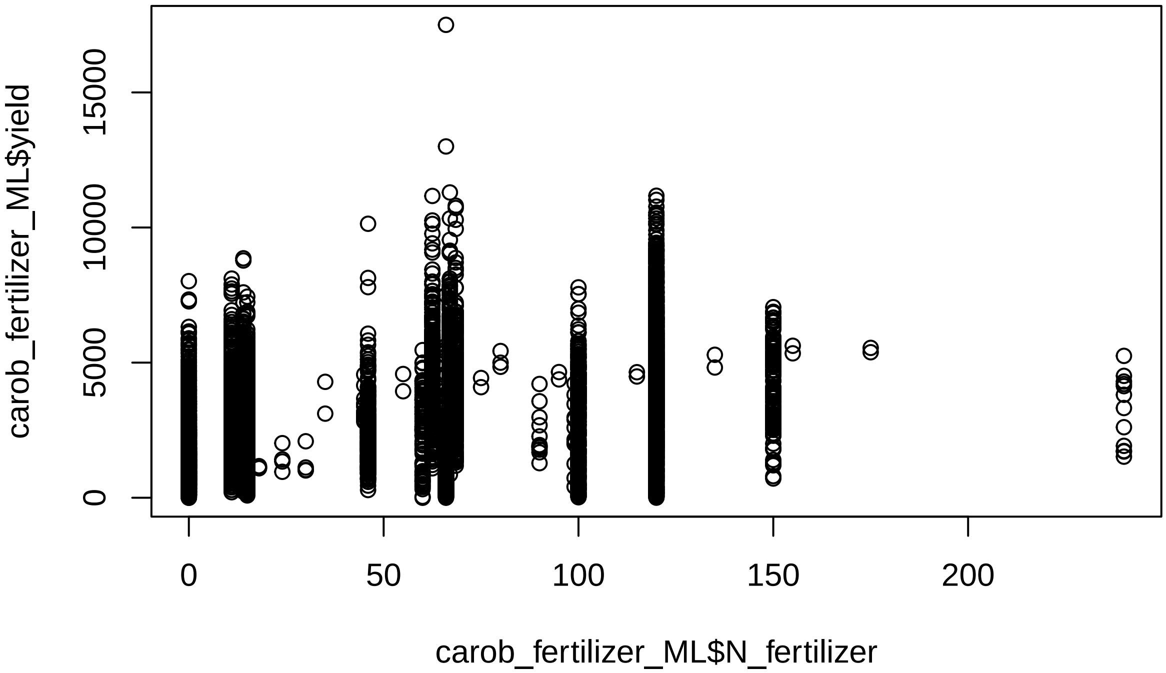
```
carob_fertilizer <- read_csv("data/compiled/carob_fertilizer.csv")
```

```
## Warning: One or more parsing issues, call ‘problems()‘ on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)

## Rows: 111653 Columns: 142
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (38): dataset_id, trial_id, country, adm1, adm2, adm3, adm4, adm5, loca...
## dbl  (89): record_id, longitude, latitude, elevation, rep, dmy_roots, dmy_st...
## lgl   (9): on_farm, is_survey, OM_used, inoculated, irrigated, plot_area, re...
## date  (6): date, emergence_date, transplanting_date, flowering_date, maturit...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
# SELECT THE DATA OF MAIZE THAT CORRESPONDS TO THE REGION OF NIGERIA
carob_fertilizer<- carob_fertilizer[ which(carob_fertilizer$country=="Nigeria" & carob_fertilizer$crop =

# Identify the important predictors X that influence the yield
predictors=c('latitude','longitude','N_fertilizer','P_fertilizer','K_fertilizer','yield')

# from the carob_fertilizer select only those columns that will be used into ML algo
carob_fertilizer_ML=carob_fertilizer[predictors]

# Before all clean the NA (to avoid missing values in the statistics)
# and the remove the duplicate values in the data frame
# (this comes usually from errors in savings, and it introduce bias since double)
carob_fertilizer_ML <-na.omit(carob_fertilizer_ML)
carob_fertilizer_ML <- carob_fertilizer_ML[!duplicated(carob_fertilizer_ML),]

# Let's see the main statistics in the data

summary(carob_fertilizer_ML)
```

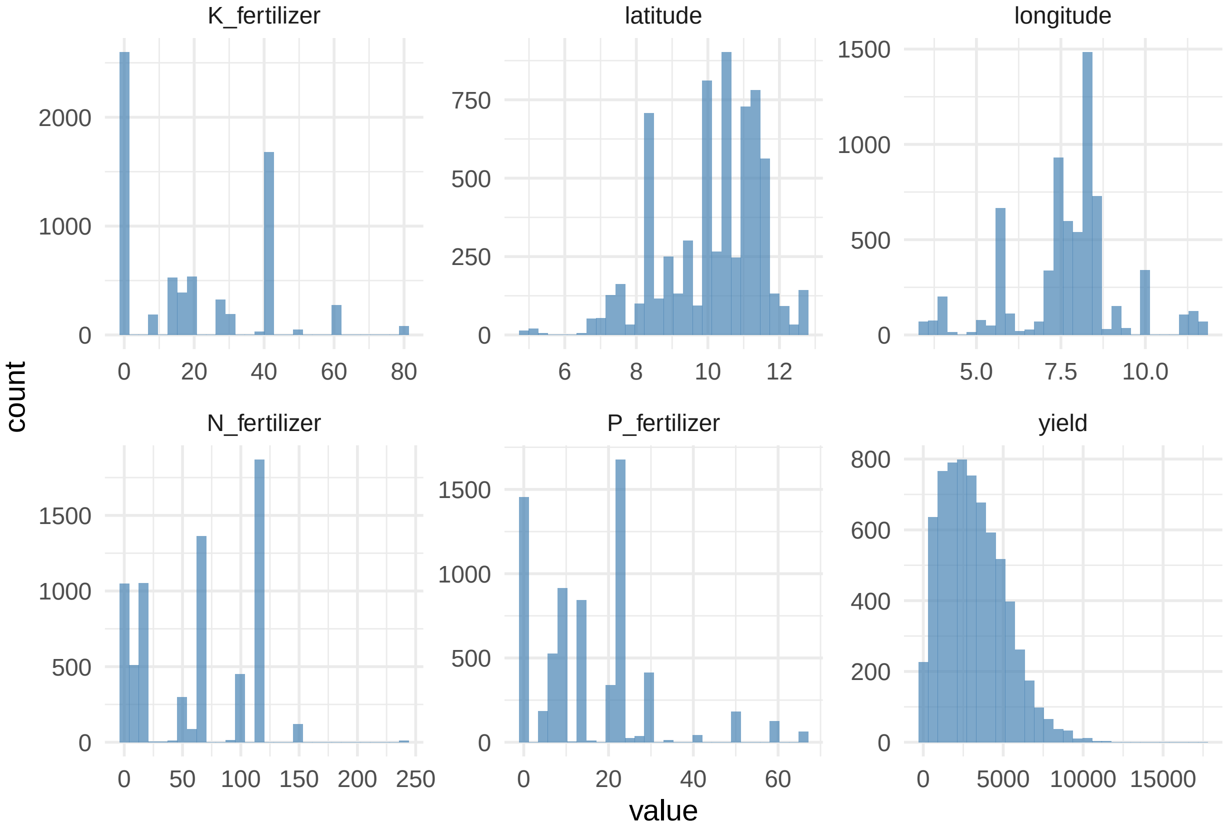
```
##     latitude        longitude       N_fertilizer     P_fertilizer
##  Min.   : 4.980   Min.   : 3.433   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 9.298   1st Qu.: 7.280   1st Qu.: 14.00   1st Qu.: 6.55
##  Median :10.518   Median : 7.900   Median : 66.00   Median :13.54
##  Mean   :10.135   Mean   : 7.690   Mean   : 61.74   Mean   :15.34
##  3rd Qu.:11.230   3rd Qu.: 8.370   3rd Qu.:120.00   3rd Qu.:21.85
##  Max.   :12.800   Max.   :11.715   Max.   :240.00   Max.   :66.00
##   K_fertilizer       yield
##  Min.   : 0.00   Min.   :    0
##  1st Qu.: 0.00   1st Qu.: 1584
##  Median :17.43   Median : 2889
##  Mean   :19.77   Mean   : 3133
##  3rd Qu.:41.50   3rd Qu.: 4405
##  Max.   :80.00   Max.   :17500
```

```r
plot(carob_fertilizer_ML$N_fertilizer, carob_fertilizer_ML$yield)
```



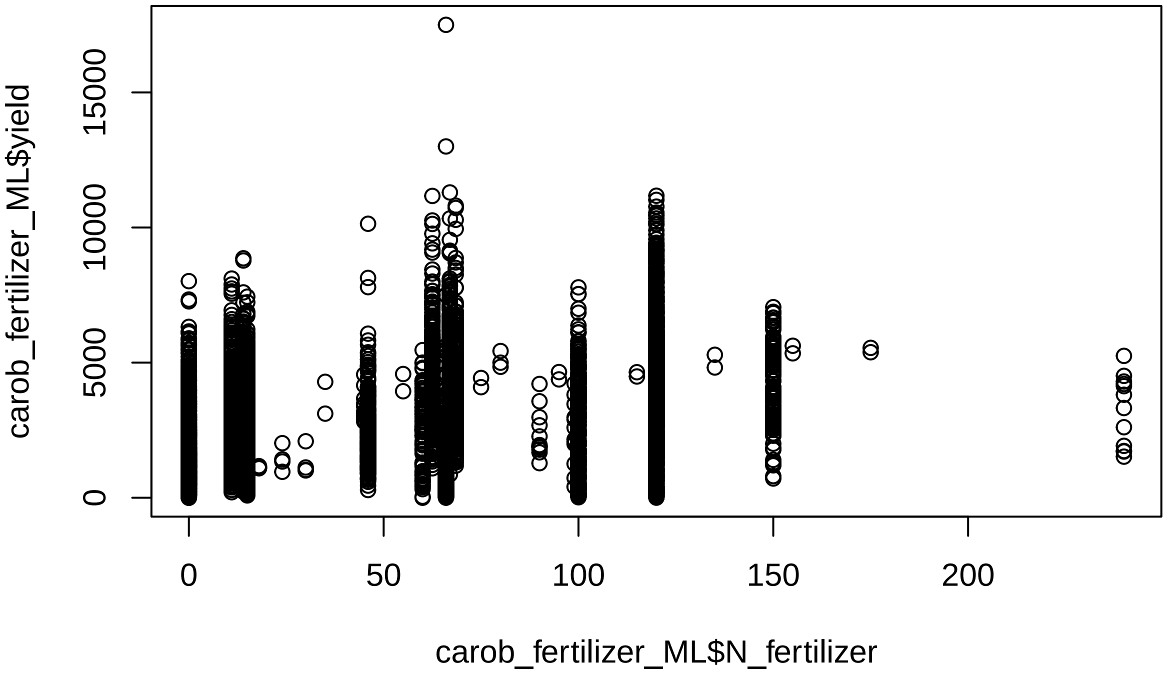
```
# See the histograms for every variable (column)
carob_fertilizer_ML %>% gather() %>%
  ggplot(aes(x=value)) +
  geom_histogram(fill="steelblue", alpha=.7) +
  theme_minimal() +
  facet_wrap(~key, scales="free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
remotes::install_github("reagro/carobiner")
ff <- carobiner::make_carob(path)
```

```
plot(carob_fertilizer_ML$N_fertilizer, carob_fertilizer_ML$yield)
```



where `path` is the folder of the cloned repo (e.g. `"d:/github/carob"`)

Use

if you use the aggregated data, you can run `carobiner::get_citations(data)` to get references (citations) to the orginal data sets used.