Data Exploration

13.05 - 21.05

Elizaveta Shcherbinina

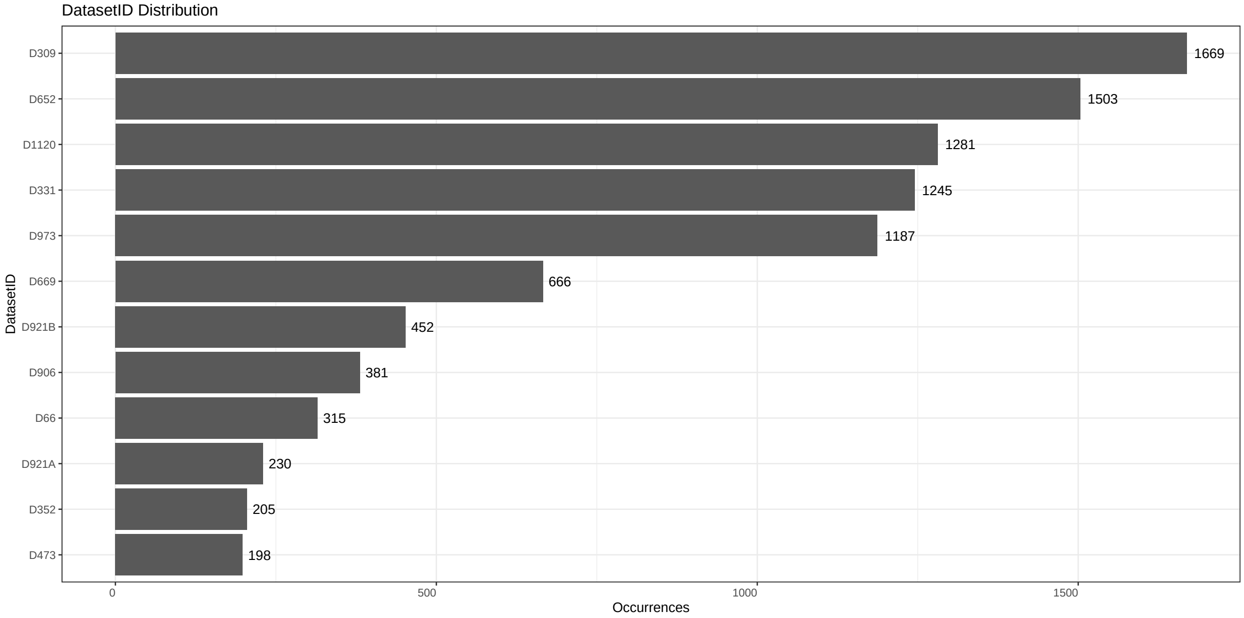
2024-05-15

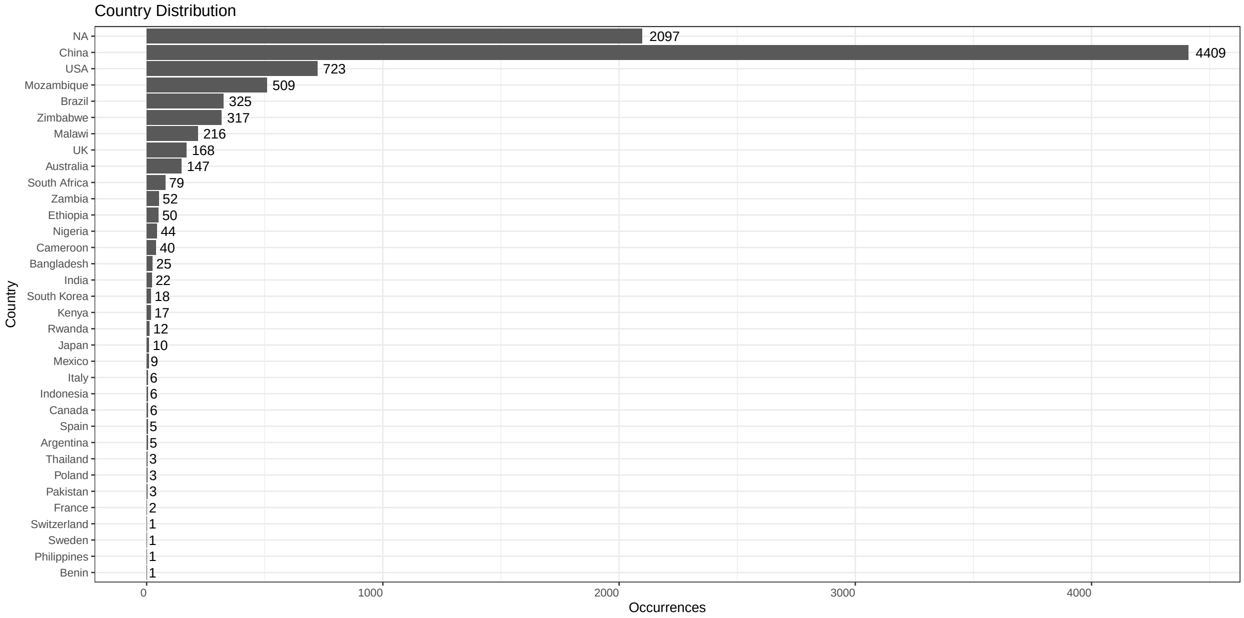
##Data Structure

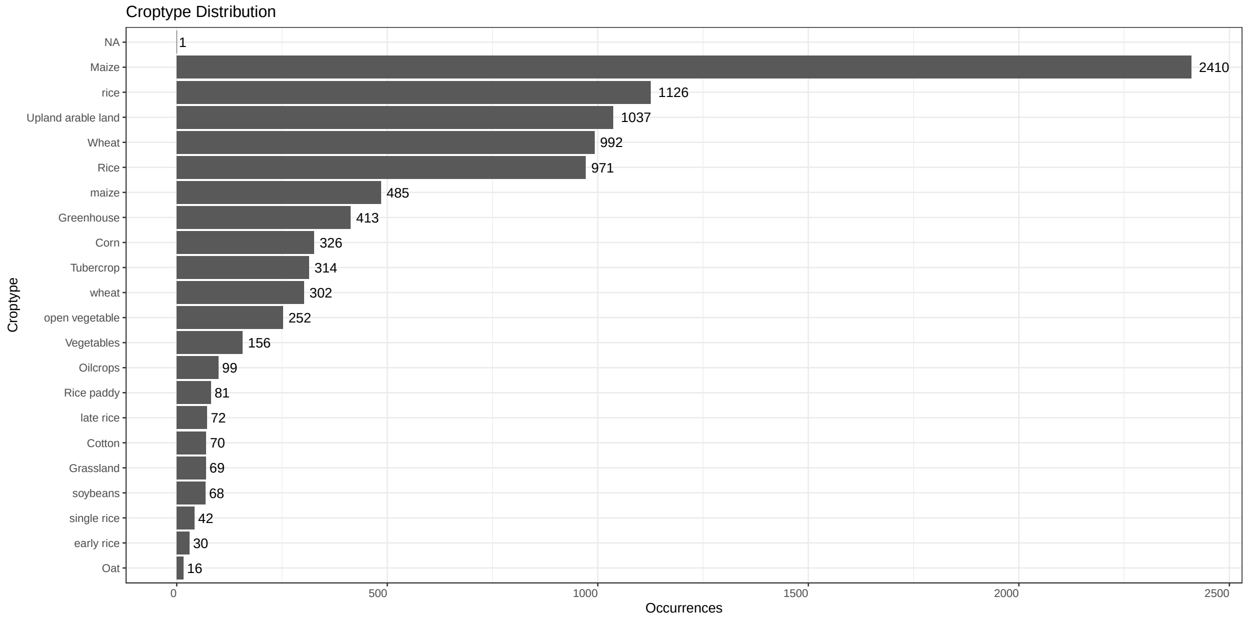
```
## Rows: 9,332
## Columns: 14
## $ No.                 <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ~
## $ DatasetID           <chr> "D66", "D66", "D66", "D66", "D66", "D66", "D66", "D~
## $ Source              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ Country             <chr> "China", "China", "China", "China", "China", "China~
## $ SiteRegion          <chr> "Shaanxi", "Shaanxi", "Shaanxi", "Shaanxi", "Shaanx~
## $ Latitude            <dbl> 35.41861, 35.41861, 35.41861, 35.41861, 35.41861, 3~
## $ Longitude           <dbl> 107.9286, 107.9286, 107.9286, 107.9286, 107.9286, 1~
## $ Croptype            <chr> "wheat", "wheat", "wheat", "wheat", "wheat", "wheat~
## $ yield_control       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 407~
## $ yield_treatm        <dbl> 3158, 2922, 2651, 3238, 3169, 3144, 4427, 3790, 383~
## $ yield_measure       <chr> "kg/hm^2", "kg/hm^2", "kg/hm^2", "kg/hm^2", "kg/hm^~
## $ Treatment           <chr> "Legume Green Manure", "Legume Green Manure", "Legu~
## $ QualityOfDataPoint  <chr> "Control and Treatment not paired", "Control and Tr~
## $ Coordinate_format   <chr> "Degr-Min-Sec", "Degr-Min-Sec", "Degr-Min-Sec", "De~
```

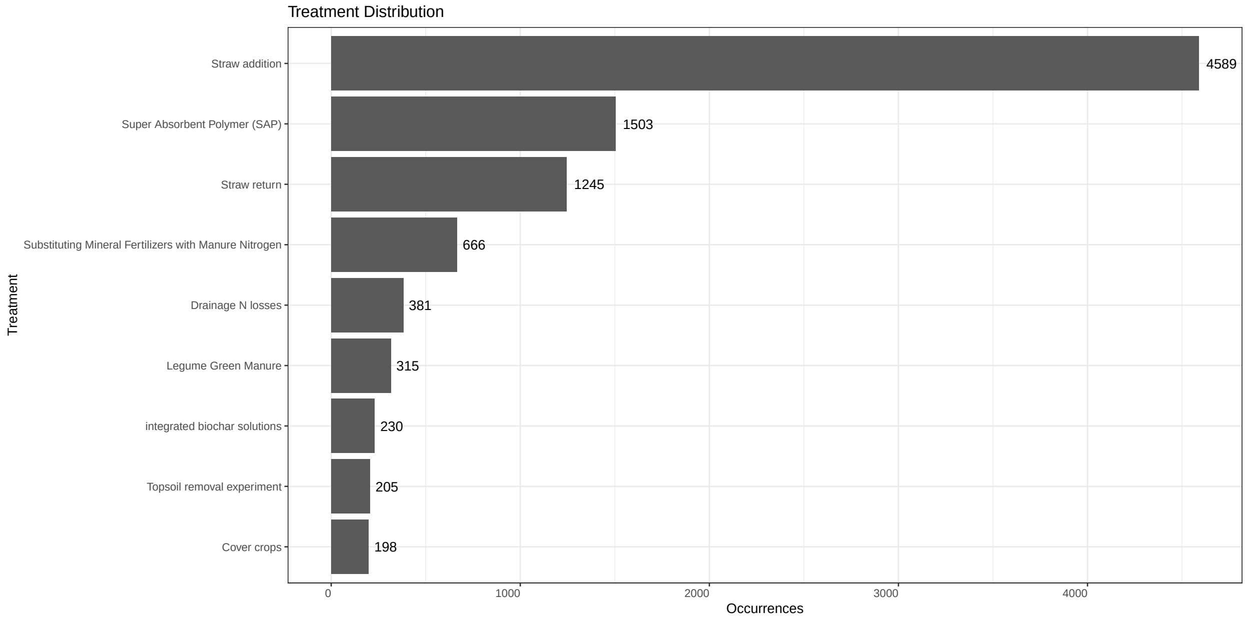
Writing a function to depict distributions of variables automatically:

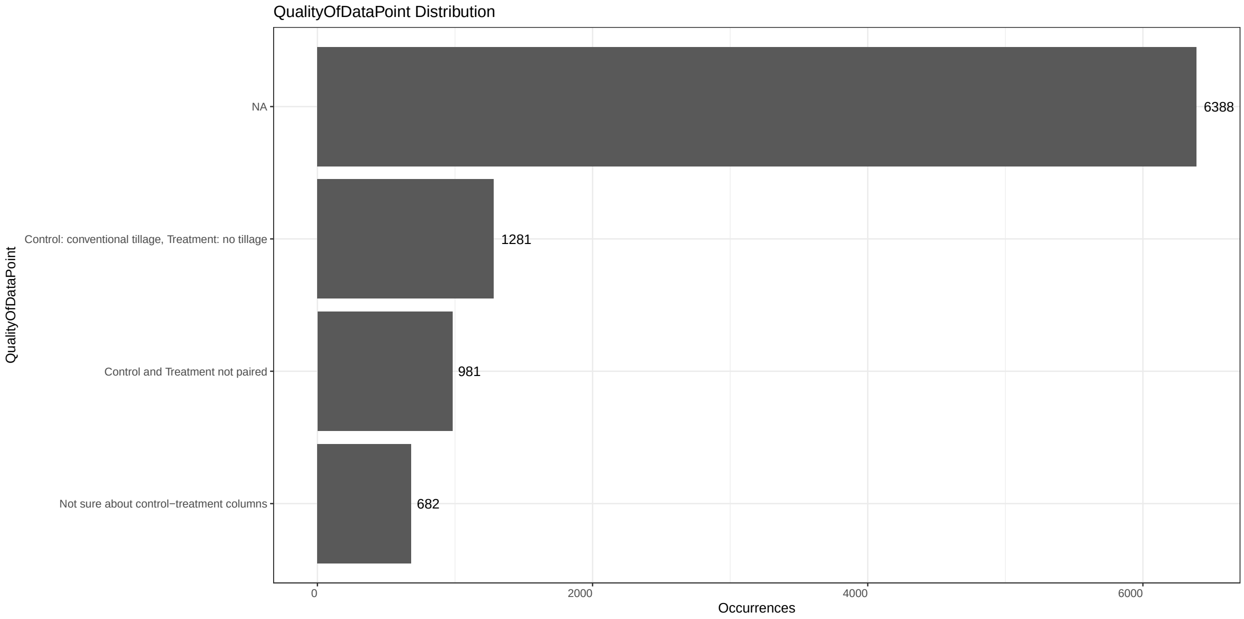
```r
plot_occurrences <- function(data_pr, parameter) {

  parameter <- rlang::sym(parameter)  # convert parameter to a symbol for use in dplyr

  data_pr %>%
    group_by(!!parameter) %>%
    summarise(occurences = n()) %>% # calculate the occurrences of values of attributes
    arrange(-desc(occurences)) %>%
    mutate(!!parameter := factor(!!parameter, levels = unique(!!parameter))) %>% # relevel the dataset
    ggplot(aes(x = !!parameter, y = occurences)) +
    geom_bar(stat = "identity") +
    scale_fill_brewer(palette = "Set2") +
    geom_text(aes(label = occurences), hjust = -0.25) +
    labs(x = parameter, y = "Occurrences",
         title = paste(as_label(parameter), "Distribution")) +
    theme_bw() +
    theme(axis.text.x = element_text(hjust = 1)) +
    coord_flip()

  }
```


DatasetID Distribution
DatasetID
Occurrences
D309
1669
D652
1503
D1120
1281
D331
1245
D973
1187
D669
666
D921B
452
D906
381
D66
315
D921A
230
D352
205
D473
198
0
500
1000
1500


Country Distribution
Country
Occurrences
NA
2097
China
4409
USA
723
Mozambique
509
Brazil
325
Zimbabwe
317
Malawi
216
UK
168
Australia
147
South Africa
79
Zambia
52
Ethiopia
50
Nigeria
44
Cameroon
40
Bangladesh
25
India
22
South Korea
18
Kenya
17
Rwanda
12
Japan
10
Mexico
9
Italy
6
Indonesia
6
Canada
6
Spain
5
Argentina
5
Thailand
3
Poland
3
Pakistan
3
France
2
Switzerland
1
Sweden
1
Philippines
1
Benin
1
0
1000
2000
3000
4000


Croptype Distribution
Croptype
NA
Maize
rice
Upland arable land
Wheat
Rice
maize
Greenhouse
Corn
Tubercrop
wheat
open vegetable
Vegetables
Oilcrops
Rice paddy
late rice
Cotton
Grassland
soybeans
single rice
early rice
Oat
1
2410
1126
1037
992
971
485
413
326
314
302
252
156
99
81
72
70
69
68
42
30
16
0
500
1000
1500
2000
2500
Occurrences


Treatment Distribution
Treatment
Straw addition
Super Absorbent Polymer (SAP)
Straw return
Substituting Mineral Fertilizers with Manure Nitrogen
Drainage N losses
Legume Green Manure
integrated biochar solutions
Topsoil removal experiment
Cover crops
4589
1503
1245
666
381
315
230
205
198
0
1000
2000
3000
4000
Occurrences



Results:
1. DatasetID:
* 5 data sets have the largest influence on the data composition with more than 1000 enteties:
+ D309 + D652 + D1120
+ D331 + D973
2. Country: * 2097 entities come from unknown sources * 4409 entities from China * the rest of the countries scoring under 1000 entities

Tuning the data set

Unify the measurement metrics

```
## # A tibble: 6 x 2
##   yield_measure    occurences
##   <fct>                 <int>
## 1 kg/ha                  4071
## 2 kg/hm^2                1984
## 3 Mg dry matter/ha       1187
## 4 Mg/ha                   579
## 5 t/ha                    230
## 6 <NA>                   1281
```

Results:
1. 1281 NA values
2. Mg dry matter/ha and Mg/ha implies that some studies were accounting for dry and some for fresh biomass. Therefore those values will not be comparable. -> lets introduce a column if the study worked with dry or fresh matter.
* There might be a change that Mg/ha were also working with the dry biomass, since the metric is too small for the fresh biomass.

Unification code:

```
data_pr <-
data_pr %>%
  mutate(yield_control = case_when(
    yield_measure == "Mg dry matter/ha" ~ yield_control*10^(-6),
    yield_measure == "Mg/ha" ~ yield_control*10^(-6),
    yield_measure == NA_character_ ~ 0, # since we do not know this metric we cannot utilize this data;
    yield_measure == NA_character_ ~ yield_control*1000,
    TRUE ~ yield_control # the rest can stay the same (kg/hm2 is the same as kg/ha)
  )) %>%
  mutate(yield_control = ifelse(yield_control == 0, NA, yield_control)) %>%
  mutate(yield_treatm = case_when(
    yield_measure == "Mg dry matter/ha" ~ yield_treatm*10^(-6),
    yield_measure == "Mg/ha" ~ yield_treatm*10^(-6),
    yield_measure == NA_character_ ~ 0, # since we do not know this metric we cannot utilize this data;
    yield_measure == NA_character_ ~ yield_treatm*1000,
    TRUE ~ yield_treatm # the rest can stay the same (kg/hm2 is the same as kg/ha)
  )) %>%
  mutate(yield_treatm = ifelse(yield_treatm == 0, NA, yield_treatm)) %>%
  mutate(yield_measure_unified = case_when(
    yield_measure == NA_character_ ~ NA_character_,
    TRUE ~ "kg/ha"
  )) %>%
  rename(yield_measure_original = yield_measure) %>%
  select(No., DatasetID, Source, Country, SiteRegion, Latitude, Longitude, Croptype, yield_control, yiel
```

How many NAs does the processed dataset have

```
## # A tibble: 1 x 15
##   NA_No. NA_DatasetID NA_Source NA_Country NA_SiteRegion NA_Latitude
##    <int>        <int>     <int>      <int>         <int>       <int>
## 1      0            0       315       2097          3837         154
## # i 9 more variables: NA_Longitude <int>, NA_Croptype <int>,
## #   NA_yield_control <int>, NA_yield_treatm <int>,
## #   NA_yield_measure_unified <int>, NA_Treatment <int>,
## #   NA_QualityOfDataPoint <int>, NA_Coordinate_format <int>,
## #   NA_yield_measure_original <int>
```

compared to the original data:

```
## # A tibble: 1 x 14
##   NA_No. NA_DatasetID NA_Source NA_Country NA_SiteRegion NA_Latitude
##    <int>        <int>     <int>      <int>         <int>       <int>
## 1      0            0       315       2097          3837         154
## # i 8 more variables: NA_Longitude <int>, NA_Croptype <int>,
## #   NA_yield_control <int>, NA_yield_treatm <int>, NA_yield_measure <int>,
## #   NA_Treatment <int>, NA_QualityOfDataPoint <int>, NA_Coordinate_format <int>
```

Results:
1. The difference only in the NA_yield_treatment variable. So all the mg/ha were in that column. Most likely it was dry biomass. * In some cases the Latitude is missing where the Longitude is present

Investigate the Croptype variable

```
## # A tibble: 22 x 1
##    Croptype
##    <fct>
##  1 wheat
##  2 rice
##  3 maize
##  4 Maize
##  5 Wheat
##  6 Rice
##  7 soybeans
##  8 Corn
##  9 Tubercrop
## 10 Oilcrops
## # i 12 more rows
```

Results:
1. "wheat", "Wheat" AND "maize", "Maize" (maybe also "Corn") AND "rice", "Rice", "Rice paddy", "single rice", "early rice", "late rice"
2. "Vegetables" and "open vegetable"
3. "Greenhouse" what is greenhouse as a crop type
4. What is "Grassland" as a crop type
5. Exclude the 1 NA?

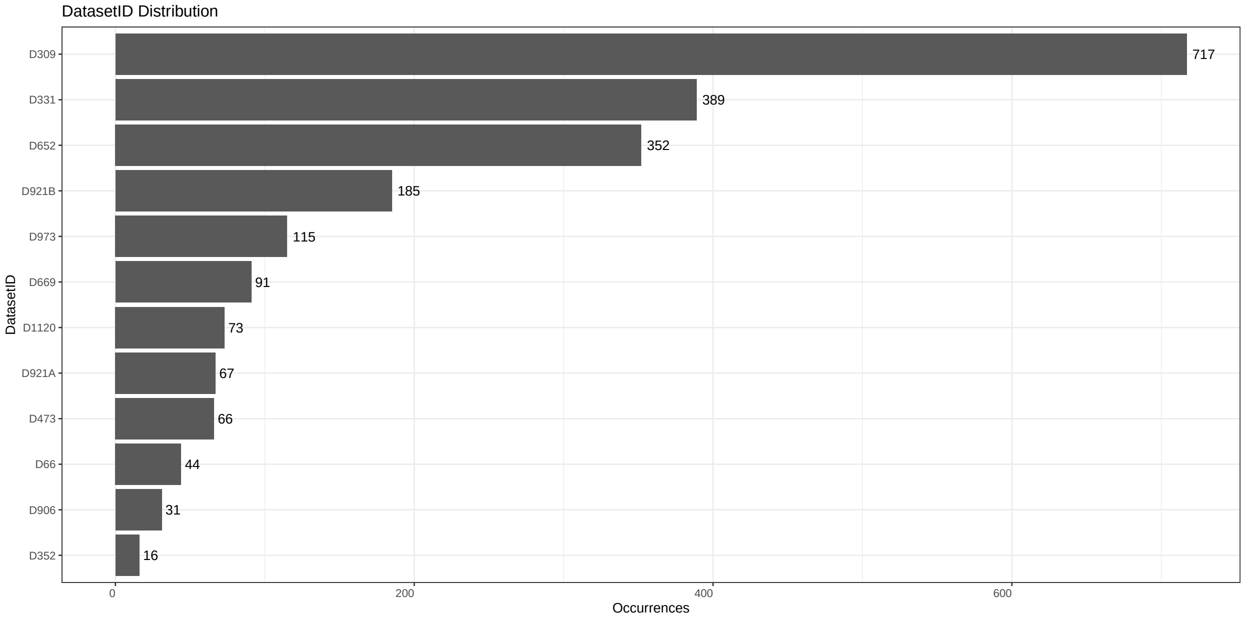
Suggested modification: 1. everything written with the first Capital

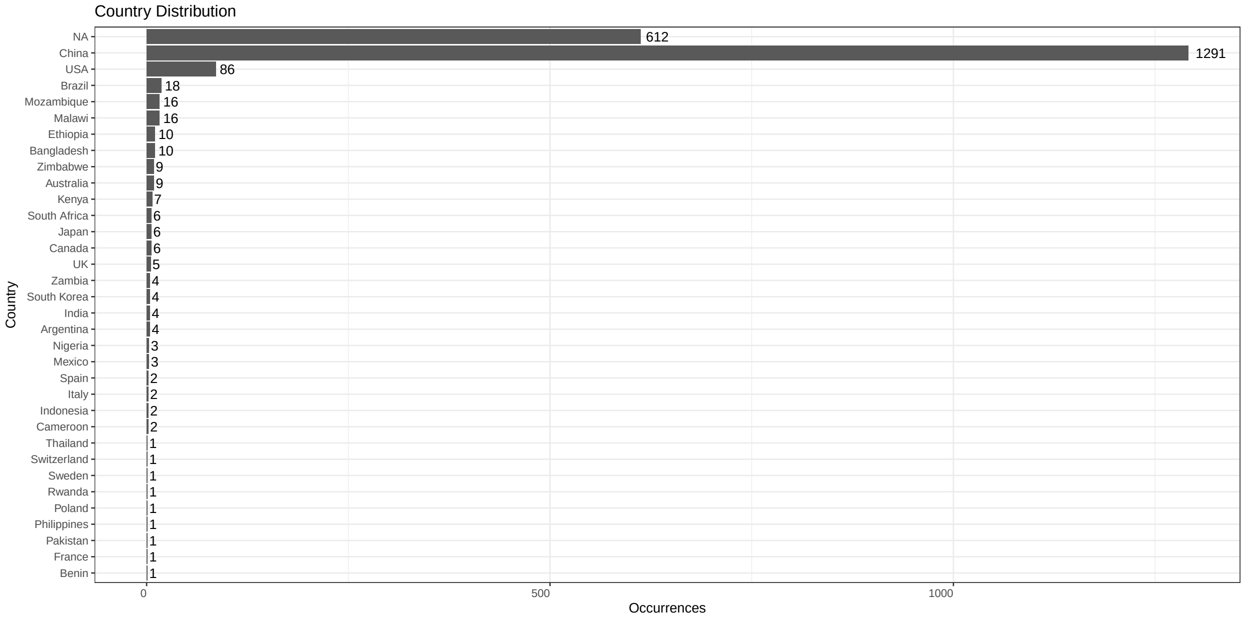
```
## # A tibble: 5 x 17
##   DatasetID Source      Country SiteRegion Latitude Longitude Croptype Treatment
##   <fct>     <chr>       <fct>   <fct>         <dbl>     <dbl> <chr>    <fct>
## 1 D1120     Barrios et~ Argent~ Esteban E~    -34.8     -58.5 Maize    Straw ad~
## 2 D1120     Bono et al~ Argent~ Anguil        -36.5     -64.0 Maize    Straw ad~
## 3 D1120     Cassel & W~ USA     Salisbury      35.7     -80.6 Maize    Straw ad~
## 4 D1120     Diaz-Zorit~ Argent~ Drabble       -34.9     -63.7 Maize    Straw ad~
## 5 D1120     Franzluebb~ USA     Watkinsvi~     33.9     -83.4 Maize    Straw ad~
## # i 9 more variables: QualityOfDataPoint <fct>, Coordinate_format <chr>,
## #   yield_control_mean <dbl>, yield_control_median <dbl>,
## #   yield_control_se <dbl>, yield_treatm_mean <dbl>, yield_treatm_median <dbl>,
## #   yield_treatm_se <dbl>, ObjectID <int>
```

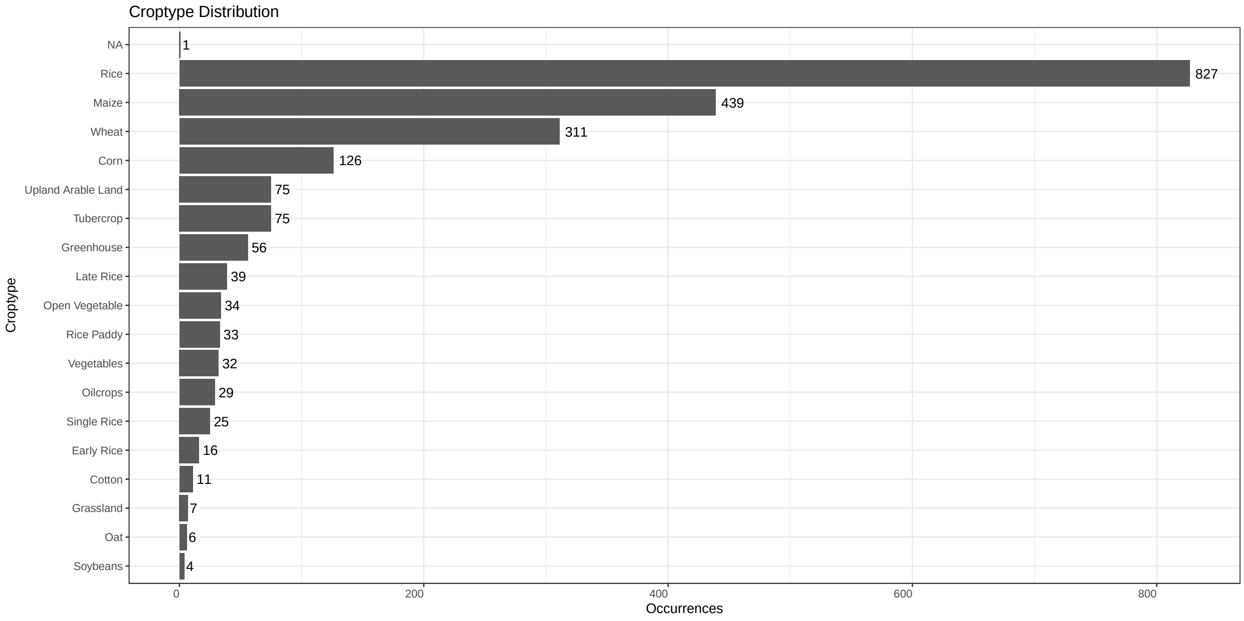
Results:
1. the number of observations goes down to 2146 (wide format) or 4292 (long format)

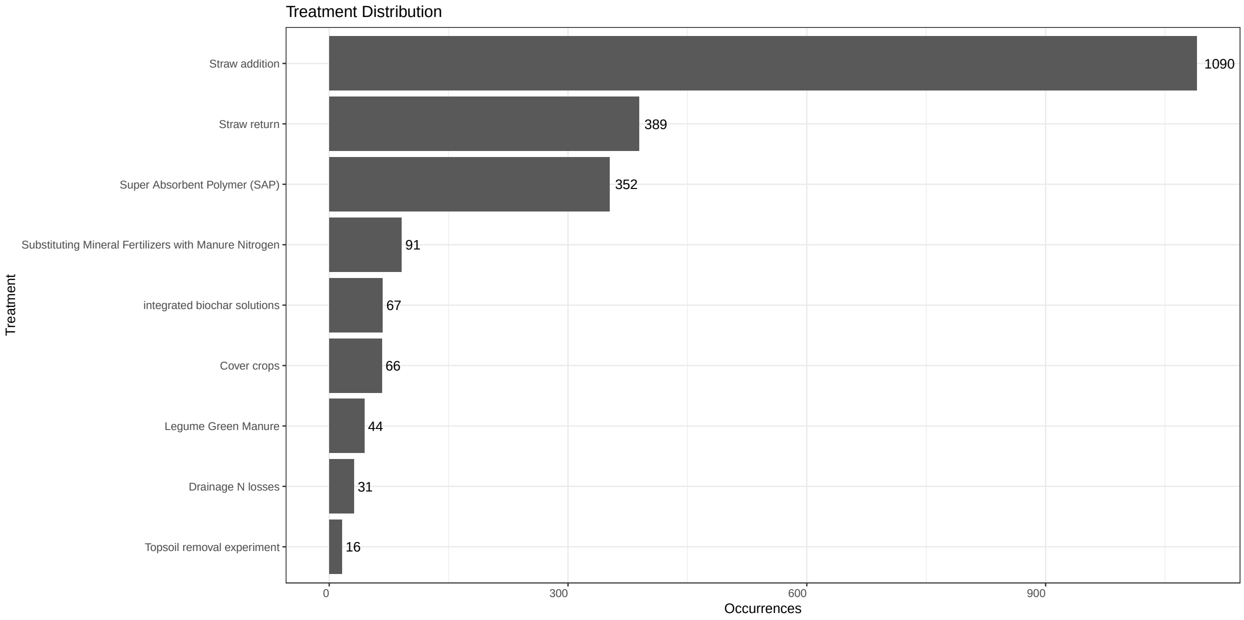
Data structure of the sumarised data set

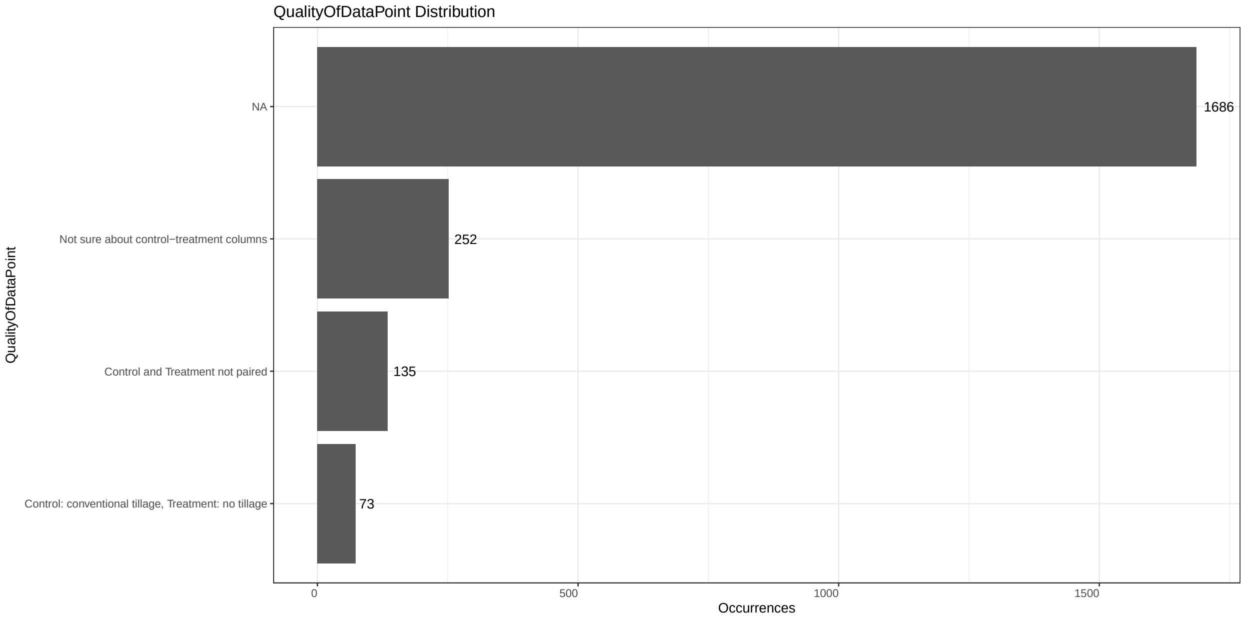
Variables' frequencies





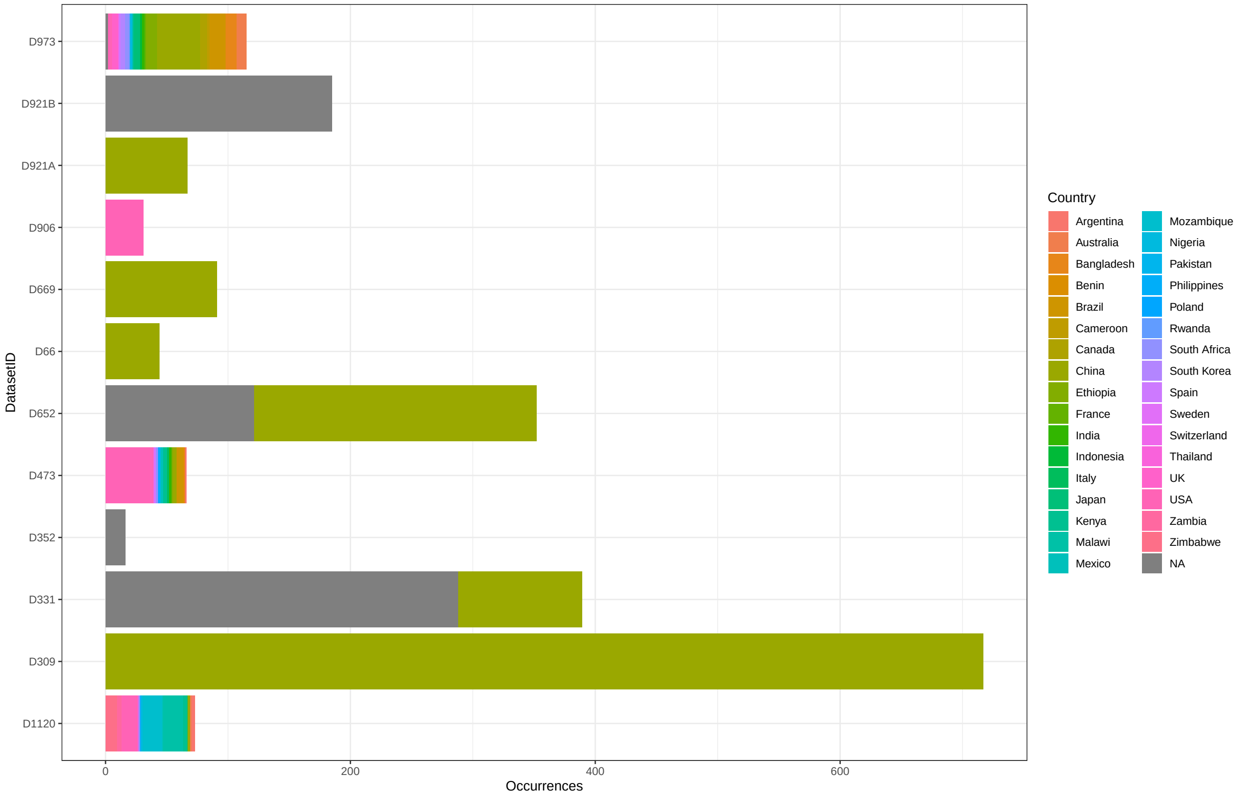

Croptype Distribution
Croptype
NA 1
Rice 827
Maize 439
Wheat 311
Corn 126
Upland Arable Land 75
Tubercrop 75
Greenhouse 56
Late Rice 39
Open Vegetable 34
Rice Paddy 33
Vegetables 32
Oilcrops 29
Single Rice 25
Early Rice 16
Cotton 11
Grassland 7
Oat 6
Soybeans 4
0
200
400
600
800
Occurrences


Treatment Distribution
Treatment
Straw addition 1090
Straw return 389
Super Absorbent Polymer (SAP) 352
Substituting Mineral Fertilizers with Manure Nitrogen 91
integrated biochar solutions 67
Cover crops 66
Legume Green Manure 44
Drainage N losses 31
Topsoil removal experiment 16
0
300
600
900
Occurrences

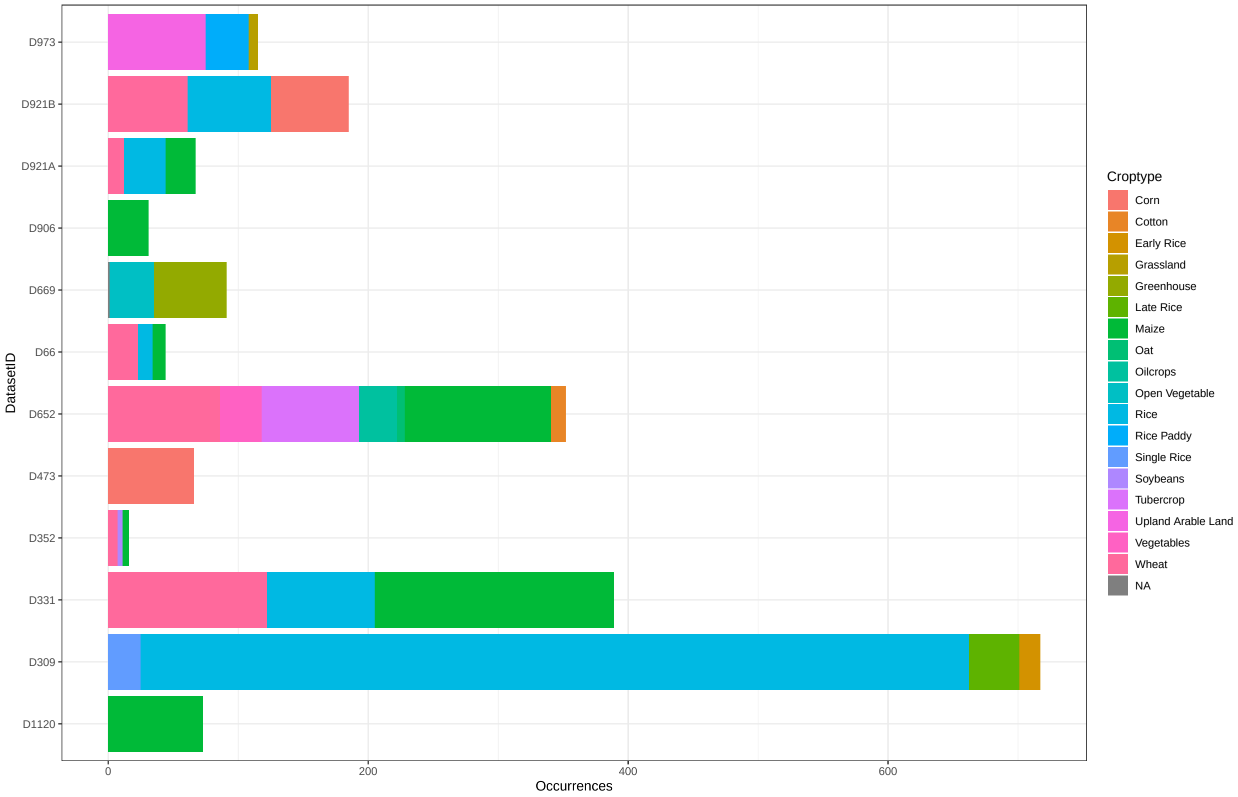


What countries were investigated in which data sets?

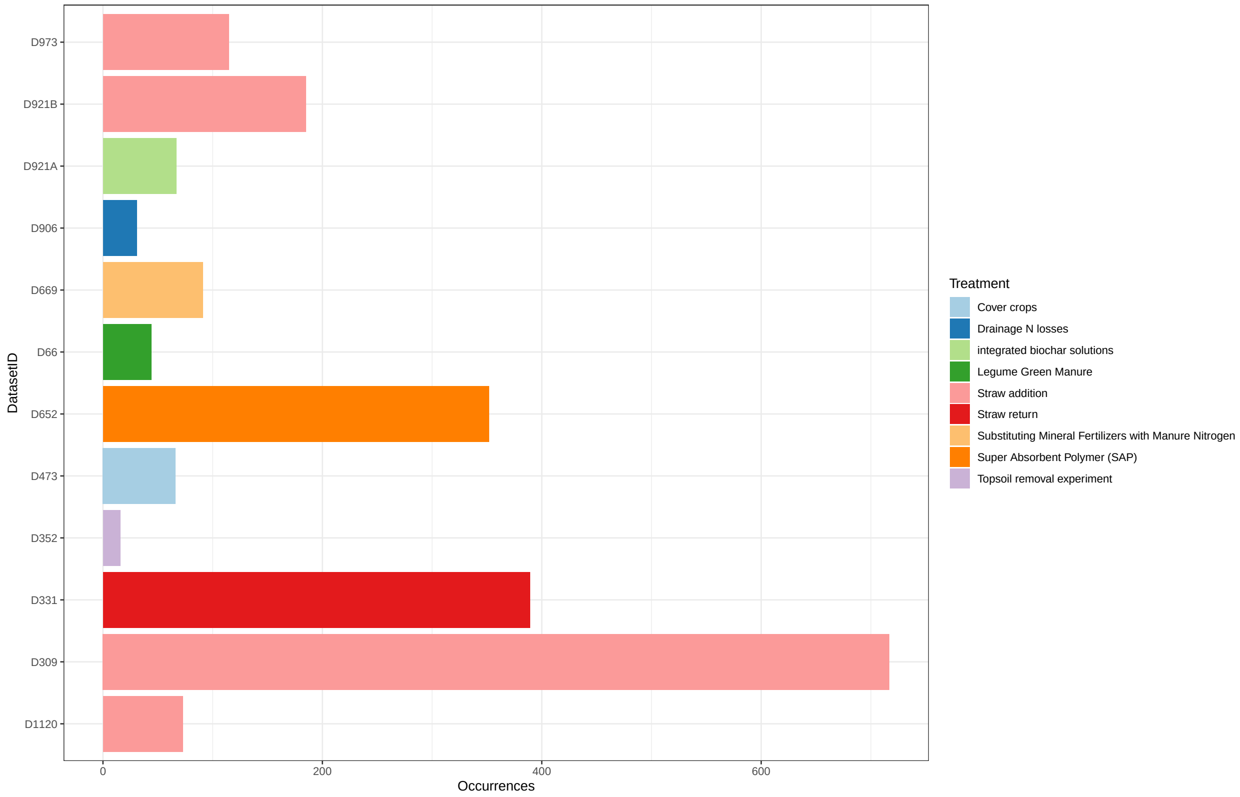
Some countries might be not well represented, because the data on them will come from a single source. Also this graph helps us to see in which data sets the NAs occure



What croptypes were investigated by which data sets?



What treatments were included in which data sets?



Looking on the Yield values across datasets

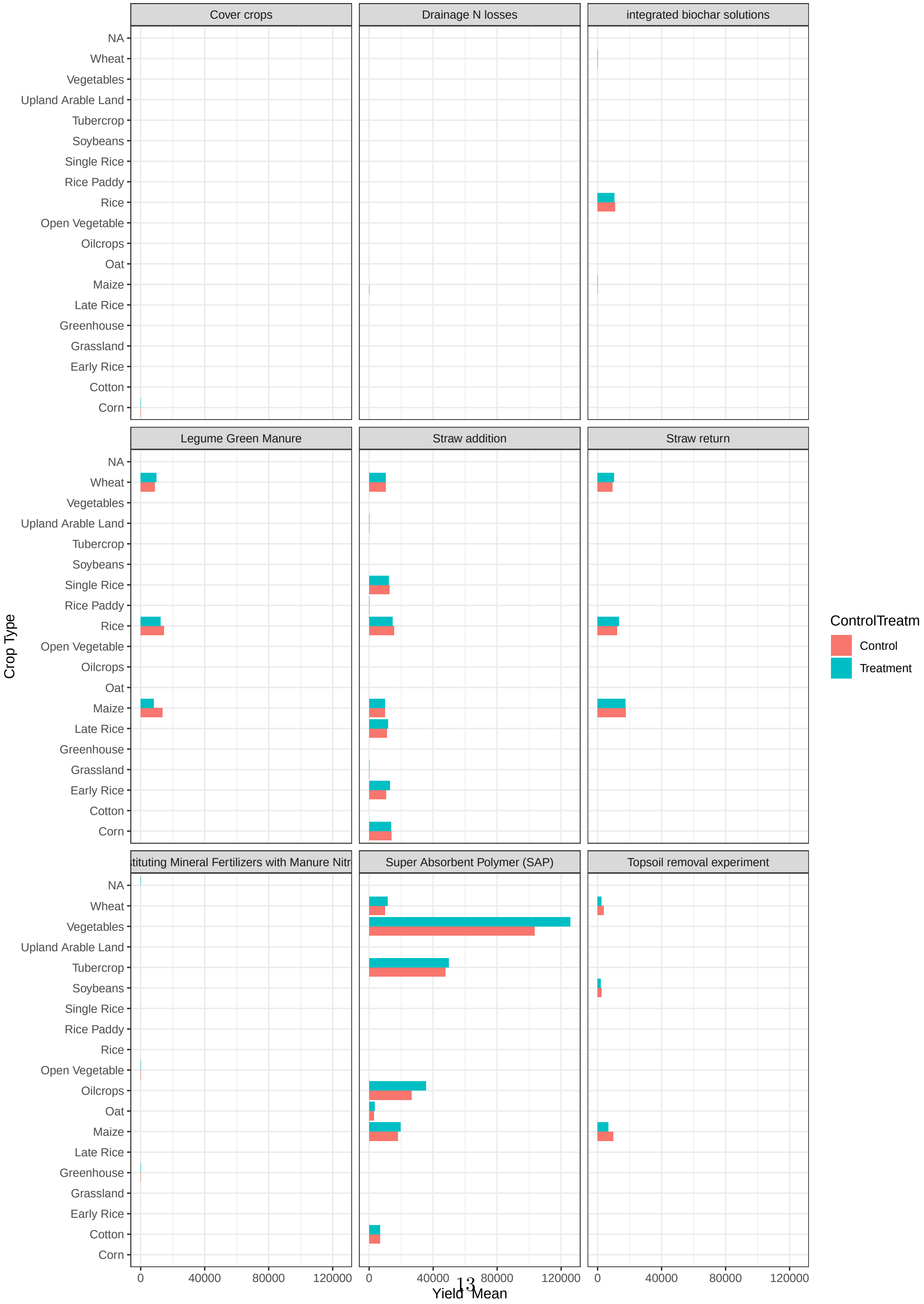
Tuning the data into the long format

```
## # A tibble: 5 x 15
##   No.   DatasetID Source Country SiteRegion Latitude Longitude Croptype
##   <fct> <fct>     <chr>  <fct>   <fct>         <dbl>     <dbl> <chr>
## 1 1     D66       <NA>   China   Shaanxi        35.4      108. Wheat
## 2 1     D66       <NA>   China   Shaanxi        35.4      108. Wheat
## 3 2     D66       <NA>   China   Shaanxi        35.4      108. Wheat
## 4 2     D66       <NA>   China   Shaanxi        35.4      108. Wheat
## 5 3     D66       <NA>   China   Shaanxi        35.4      108. Wheat
## # i 7 more variables: yield_measure_unified <chr>, Treatment <fct>,
## #   QualityOfDataPoint <fct>, Coordinate_format <chr>,
## #   yield_measure_original <fct>, ControlTreatm <fct>, Yield <dbl>
```

and summarizing it

```
## # A tibble: 5 x 15
##   DatasetID Source      Country SiteRegion Latitude Longitude Croptype Treatment
##   <fct>     <chr>       <fct>   <fct>         <dbl>     <dbl> <chr>    <fct>
## 1 D1120     Barrios et~ Argent~ Esteban E~    -34.8     -58.5 Maize    Straw ad~
## 2 D1120     Barrios et~ Argent~ Esteban E~    -34.8     -58.5 Maize    Straw ad~
## 3 D1120     Bono et al~ Argent~ Anguil        -36.5     -64.0 Maize    Straw ad~
## 4 D1120     Bono et al~ Argent~ Anguil        -36.5     -64.0 Maize    Straw ad~
## 5 D1120     Cassel & W~ USA     Salisbury      35.7     -80.6 Maize    Straw ad~
## # i 7 more variables: QualityOfDataPoint <fct>, Coordinate_format <chr>,
## #   ControlTreatm <fct>, Yield_mean <dbl>, Yield_median <dbl>, Yield_se <dbl>,
## #   ObjectID <int>
```

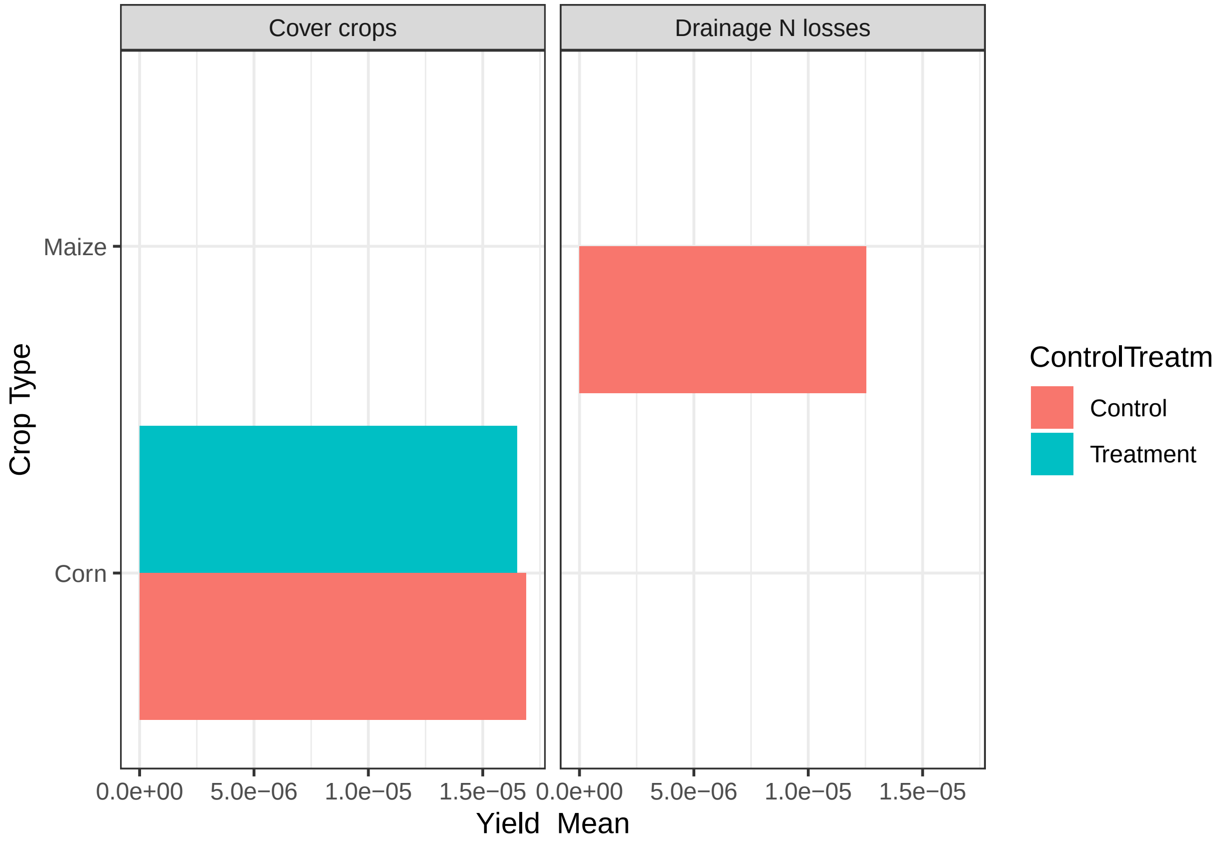
What treatments were appied to which crops and how did the treatment affect the crop yield?



Some treatments seem to have no observations. However, this is only the scaling problem, where the mg/ha measurements were unified to kg/ha and now they seem to be unproportionally small relative to the other observations. So, here we visualise those studies separately.

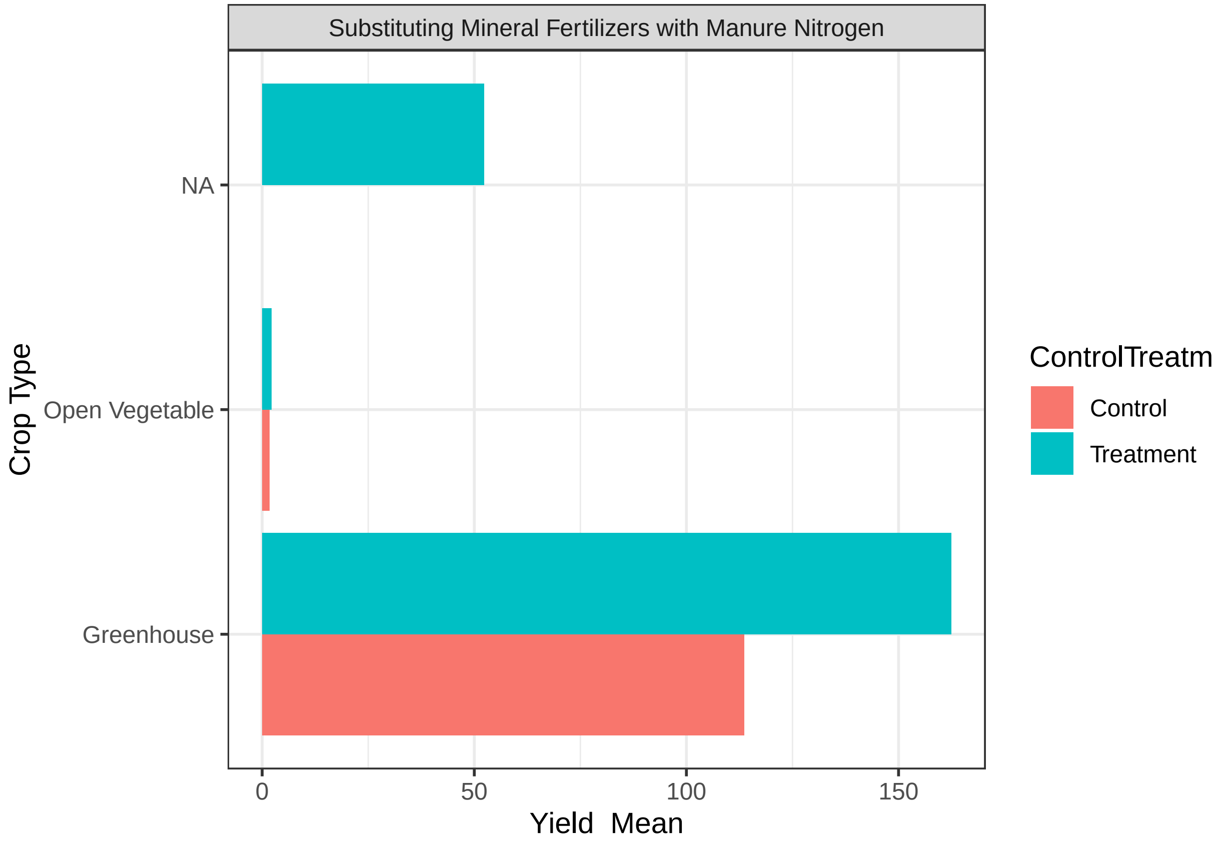
What treatments were appied to which crops and how did the treatment affect the crop yield?

in Cover crops and Drainage N lossess treatments



What treatments were appied to which crops and how did the treatment affect the crop yield?

in Substituting Mineral Fertilizers with Manure Nitrogen treatment



Moving on to the visualisaton on the map.

First, investigate the data set by its summary to check the latitudes and longitudes, whether they lie in appropriate ranges. *Longitude (-180; +180) and Latitude(-90;+90)*

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##  -139.00    29.05    32.44    38.01    37.44 13653.68
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## -122.34  106.97  113.89   98.38  118.60  150.52
```

The Latitude goes from -139 to 13653.68 and Longitude from -122 to 150. Clearly there are some issues with the Latitude.

After reading the file into coordinates in ArcGIS two points were taken out:

```
## # A tibble: 2 x 15
##   DatasetID Source      Country SiteRegion Latitude Longitude Croptype Treatment
##   <fct>     <chr>       <fct>   <fct>         <dbl>     <dbl> <chr>    <fct>
## 1 D669      Cao et al.~ China   Jiangsu      13654.      119. Open Ve~ Substitu~
## 2 D669      Cao et al.~ China   Jiangsu      13654.      119. Open Ve~ Substitu~
## # i 7 more variables: QualityOfDataPoint <fct>, Coordinate_format <chr>,
## #   ControlTreatm <fct>, Yield_mean <dbl>, Yield_median <dbl>, Yield_se <dbl>,
## #   ObjectID <int>
```

Also, some points again land in the sea, but most likely because their latitude and longitude are confused or converted wrongly. There is also mostly no information on the Quality of data points and no Country/SiteRegion information.

```
## # A tibble: 5 x 15
##   DatasetID Source      Country SiteRegion Latitude Longitude Croptype Treatment
##   <fct>     <chr>       <fct>   <fct>         <dbl>     <dbl> <chr>    <fct>
## 1 D352      Allen et a~ <NA>    <NA>           104.      47.8 Wheat    Topsoil ~
## 2 D352      Allen et a~ <NA>    <NA>           104.      47.8 Wheat    Topsoil ~
## 3 D352      Gao et al.~ <NA>    <NA>           125.      49.0 Soybeans Topsoil ~
## 4 D352      Gao et al.~ <NA>    <NA>           125.      49.0 Soybeans Topsoil ~
## 5 D352      Gorji et a~ <NA>    <NA>           125.      49.0 Maize    Topsoil ~
## # i 7 more variables: QualityOfDataPoint <fct>, Coordinate_format <chr>,
## #   ControlTreatm <fct>, Yield_mean <dbl>, Yield_median <dbl>, Yield_se <dbl>,
## #   ObjectID <int>
```

```
## # A tibble: 32 x 15
##    DatasetID Source     Country SiteRegion Latitude Longitude Croptype Treatment
##    <fct>     <chr>      <fct>   <fct>         <dbl>     <dbl> <chr>    <fct>
##  1 D352      Allen et ~ <NA>    <NA>           47.8      104. Wheat    Topsoil ~
##  2 D352      Allen et ~ <NA>    <NA>           47.8      104. Wheat    Topsoil ~
##  3 D352      Gao et al~ <NA>    <NA>           49.0      125. Soybeans Topsoil ~
##  4 D352      Gao et al~ <NA>    <NA>           49.0      125. Soybeans Topsoil ~
##  5 D352      Gorji et ~ <NA>    <NA>           49.0      125. Maize    Topsoil ~
##  6 D352      Gorji et ~ <NA>    <NA>           49.0      125. Maize    Topsoil ~
##  7 D352      Izaurrald~ <NA>    <NA>           53.4     -113. Wheat    Topsoil ~
##  8 D352      Izaurrald~ <NA>    <NA>           53.4     -113. Wheat    Topsoil ~
##  9 D352      Lamey et ~ <NA>    <NA>           53.6     -114. Wheat    Topsoil ~
## 10 D352      Lamey et ~ <NA>    <NA>           53.6     -114. Wheat    Topsoil ~
## # i 22 more rows
## # i 7 more variables: QualityOfDataPoint <fct>, Coordinate_format <chr>,
## #   ControlTreatm <fct>, Yield_mean <dbl>, Yield_median <dbl>, Yield_se <dbl>,
## #   ObjectID <int>
```

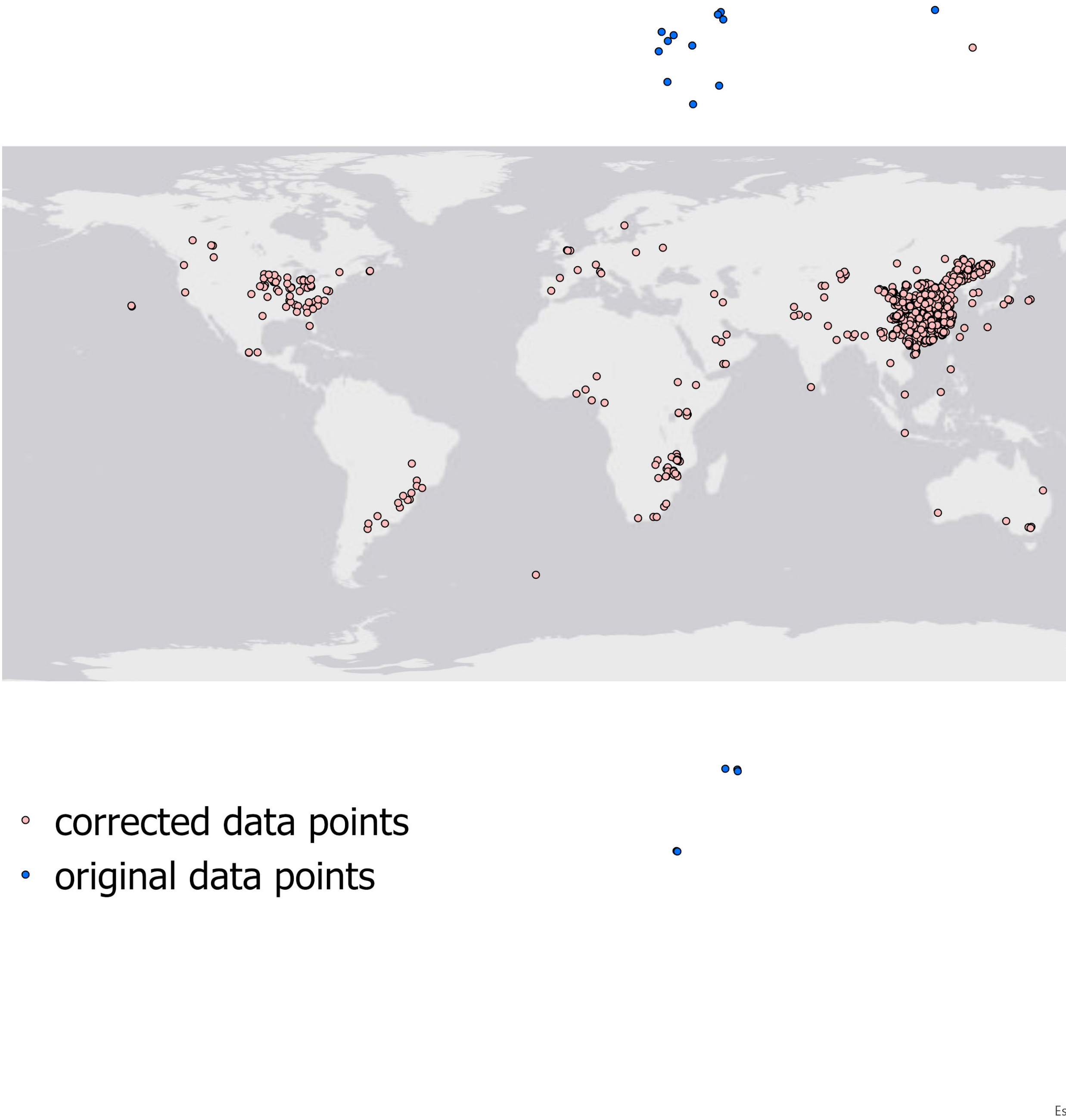
Also, for the studies in Marrabel, Australia the minus is misplaced. In the current dataset the coordinates are (lat: 34.14 long: -138.88) and should be (lat: -34.14 long: 138.88).

```
## # A tibble: 2 x 15
##   DatasetID Source      Country SiteRegion Latitude Longitude Croptype Treatment
##   <fct>     <chr>       <fct>   <fct>         <dbl>     <dbl> <chr>    <fct>
## 1 D973      Farhoodi e~ Austra~ Marrabel       34.1     -139. Upland ~ Straw ad~
## 2 D973      Farhoodi e~ Austra~ Marrabel       34.1     -139. Upland ~ Straw ad~
## # i 7 more variables: QualityOfDataPoint <fct>, Coordinate_format <chr>,
## #   ControlTreatm <fct>, Yield_mean <dbl>, Yield_median <dbl>, Yield_se <dbl>,
## #   ObjectID <int>
```

```
## # A tibble: 2 x 15
##   DatasetID Source      Country SiteRegion Latitude Longitude Croptype Treatment
##   <fct>     <chr>       <fct>   <fct>         <dbl>     <dbl> <chr>    <fct>
## 1 D973      Farhoodi e~ Austra~ Marrabel      -34.1      139. Upland ~ Straw ad~
## 2 D973      Farhoodi e~ Austra~ Marrabel      -34.1      139. Upland ~ Straw ad~
## # i 7 more variables: QualityOfDataPoint <fct>, Coordinate_format <chr>,
## #   ControlTreatm <fct>, Yield_mean <dbl>, Yield_median <dbl>, Yield_se <dbl>,
## #   ObjectID <int>
```

After correcting those issues we can import the data table into ArcGIS and visualise the points on the map. The result looks the follwoing way:

Study Map


○ corrected data points
● original data points
Es