Land-use data harmonization

Vincent Pellissier

8 November 2018

This report explains the steps taken to correct the land-use data. The masterfile used is the version 1.8, as found on November 8th 2018. The sheet in Grassplot 1.8.xlsx containing the LU information is the sheet 'datasets' and was read and saved as a .rds file to ease the process (faster loading)

```
df <- readRDS(file.path(path_grassplot, 'Grassplot 1.8_Data.rds'))
df <- df %>%
        mutate_at(.vars = c(95:104), funs(ifelse(. %in% c('NA', '[NA]'), NA, .)))
```

The columns 98 (mowing frequency) and 100 (grazing intensity) are renamed:

```
names(df)[c(98,100)] <- c('mowing_frequency', 'grazing_intensity')
```

Correction of the land use column

The file lookup_table_LU.xlsx contains two sheets:

- land_use_detail: lookup table between the column 96 and other column. This needs to be discussed further before I finalize the table.
- global_land_use: lookuptable matching the values in column 95 with a new classification. This new classification is a set of binary columns (allowing for mixed land-use) plus one declarative column. Each binary column correspond to a management practice (grazed, mown, burnt, fertilized, abandoned, natural_grassland), the declarative column stores additional information that cannot be put in one of the 6 columns (*i.e* trampled) and that will be used afterward. **NOTE** LAS, P, L were noted as land-use for the dataset PL_C and N was noted as land-use for the dataset RU_J. I could not find what these stand for. I assumed that these plots don't have a land-use but the info is still in the column 'other'.

```
lut <- read_excel(file.path(path_grassplot, "lookup_table_LU.xlsx" ),
                  sheet = 'global_land_use')
```

The global_land_use lookup table is used to harmonize the column 95 with the new binary columns:

```
df <- df %>%
    left_join(lut, by = c("Land use (5 standard categories: mown, grazed, abandoned, natural grassland,
```

Matching new binary columns and intensity columns

In this step, we identify datasets with discrepancies between the new binary columns and the matching intensity column to:

- Manually check in the original datasets or publications
- Correct the discrepancies.

Mowing and mowing intensity

Datasets containing plots for which mowing intensity > 0 & mown != 1 (here and after, we refer to the new binary columns)

```
## [1] "EU_K" "CZ_J"
```

- EU_K contains resp. 4 and 4 plots classified as natural_grassland with a mowing frequency of resp. 0.03 and 4. Only the plots with a frequency >= 0.2 will be classified as mown (abandoned less than 5 years).
- EU_K contains resp. 12, 84, 30, 4 and 12 plots classified as abandoned with a mowing frequency of resp. 0.03, 0.05, 0.1, 0.2, 0.3. Only the plots with a frequency >= 0.2 will be classified as mown (abandoned less than 5 years).

```
df[df$`Dataset ID` == 'EU_K' & df$mowing_frequency %in% c('4', '1', '0.5', '0.3', '0.2'),] %<>%
  mutate(mown = 1)
```

- CZ_J contains plot classified as abandoned which have a mowing frequency == 0.05. These will not be classified as mown. (I could not find the original land-use information in CZ_J.xls)

Grazing and Grazing intensity

Datasets containing plots for which grazing intensity > 0 & grazed != 1:

```
## [1] "IR_A" "PL_D" "TR_B" "UA_G" "EU_K"
```

- IR_A contains 34 plots classified as mown which have a grazing_intensity == 0.1. These plots will be classified as grazed (and also remain mown as indicated in the original database IR_A.xls). **Note for Idoia** In the original DB, the grazing intensity is noted as 1, 2, 3. Is it correct that it stands for low, medium and high intensity, and hence translate as 0.1, 0.5, 1 in the master file?

```
df[df$`Dataset ID` == 'IR_A',] %<>%
  mutate(grazed = ifelse(grazing_intensity != '0', 1, 0))
```

- PL_D contains 39 plots classified as mown which have a grazing_intensity == 0.5. These plots will be classified as grazed. (No composition data, so I could not check the original dataset)

```
df[df$`Dataset ID` == 'PL_D' & df$grazing_intensity  %in% '0.5',] %<>%
  mutate(grazed = 1)
```

- TR_B contains 32 plots classified as abandoned which have a grazing_intensity == 0.1. These plots will be classified as grazed (no further info could be found in TR_B.xls)

```
df[df$`Dataset ID` == 'TR_B' & df$grazing_intensity  %in% '0.1',] %<>%
  mutate(grazed = 1)
```

- UA_G contains reps. 12, 15, 30 and 3 plots classified as natural_grassland which have a grazing_intensity == ‘high’, ‘low’, ‘middle’ or ‘overgrazing’. These plots will be classified as grazed (I could not find the original land-use information in UA_G.xls)

```
df[df$`Dataset ID` == 'UA_G' & df$grazing_intensity %in%
     c('high', 'low', 'middle', 'overgrazing'),] %<>%
  mutate(grazed = 1)
```

- EU_K contains 4 plots classified as natural_grassland which have a grazing_intensity == 0.5. These plots will be classified as grazed (no further info could be found in EU_K.xls)

```
df[df$`Dataset ID` == 'EU_K' & df$grazing_intensity %in% '0.5',] %<>%
  mutate(grazed = 1)
```

Matching new and old binary columns

In this step, we identify and correct discrepancies between the newly created and corrected binary column, and the former ones

Table 1: Contingency table of columns mown and Mowing (1/0)

Mowing (1/0)	0	1	<NA>
?	NA	NA	6
0	6247	8	22
1	NA	620	NA
NA	150296	6810	16068

Mowing

There are:

- 8 plot with mown == 1 & Mowing (1/0) == 0
- 0 plots with mown == 0 & Mowing (1/0) == 1
- 6810 plots with mown == 1 & Mowing (1/0) = NA
- 6 plots with mown = NA & Mowing (1/0) == '?'

Datasets containing plots with mown == 1 & Mowing (1/0) == 0 (new vs former column):

```
## [1] "EU_K"
```

These 8 plots corresponds to 4 plots reclassified as mown == 1 earlier in EU_K and 4 plots classified as mown == 1 but misclassified in the former Mowing (1/0) column. The classifiaction of these plots is left untouched

Datasets containing plots with mown = NA & Mowing (1/0) == ? (new vs former column):

```
## [1] "PL_A"
```

No additional information could be found, so the plots are left with mown = NA

Datasets containing plots with mown == 1 & Mowing (1/0) = NA (new vs former column):

```
##  [1] "AT_A" "BG_A" "CH_A" "CH_B" "CH_C" "CZ_A" "CZ_B" "CZ_C" "CZ_H" "CZ_I"
## [11] "DE_A" "DE_E" "DE_F" "DE_H" "DE_I" "DE_J" "DE_L" "DE_N" "DE_O" "DE_P"
## [21] "ES_A" "ES_C" "EU_A" "EU_B" "EU_C" "EU_E" "EU_G" "EU_J" "FR_A" "HR_A"
## [31] "HU_A" "HU_C" "IT_H" "IT_K" "JP_A" "LV_A" "NL_A" "PL_B" "PL_C" "PL_D"
## [41] "RO_A" "RO_B" "RS_A" "RU_A" "TJ_A" "UA_C" "UA_D"
```

These plots are all classified as mown (or combination of mown other land-use) in the original land use column of the masterfile and are left classified as mown == 1.

Grazing

There are:

- 6 plots with grazed == 0 & Grazing (1/0) == '1'
- 6 plots with grazed == 0 & Grazing (1/0) == 'probably'
- 7 plots with grazed == 1 & Grazing (1/0) == '?'
- 1 plots with grazed == 1 & Grazing (1/0) == '2'
- 29219 plots with grazed == 1 & Grazing (1/0) = NA
- 6 plots with grazed = NA & Grazing (1/0) == '?'

Table 2: Contingency table of columns grazed and Grazing (1/0)

Grazing (1/0)	0	1	<NA>
?	NA	7	6
0	3238	NA	NA
1	6	5424	NA
2	NA	1	NA
probably	6	NA	NA
NA	126093	29223	16073

NOTE: the contingency table and figures reported above do not match because there is duplicated ID in the column `Grassplot ID of plot`. The following ID are duplicated PL_C_N198_10, TR_B_P01, DK_A_N001_0.25D, DK_A_N002_0.25D, DK_A_N003_0.25D, DK_A_N004_0.25D

##### *Datasets plots with grazed == 0 & Grazing (1/0) == '1':*

```
## [1] "PL_A_N005_0.0001b" "PL_A_N005_0.001b"  "PL_A_N005_0.01b"
## [4] "PL_A_N005_0.1b"    "PL_A_N005_10b"     "PL_A_N005_1b"
```

- PL_A contains 6 plots (PL_A_N005_xxxxb) noted as abandoned with 'occasionally grazed or trampled' in the detailed column (column 96). These plots will be classified as grazed in the new binary column."

```
df[df$`grazed` %in% '0' & df$`Grazing (1/0)` %in% '1',]%<>%
    mutate(grazed = 1)
```

##### *Datasets plots with grazed == 0 & Grazing (1/0) == 'probably':*

```
## [1] "PL_A_N004_0.0001a" "PL_A_N004_0.001a"  "PL_A_N004_0.01a"
## [4] "PL_A_N004_0.1a"    "PL_A_N004_10a"     "PL_A_N004_1a"
```

- PL_A contains 6 plots (PL_A_N004_xxxxa) noted as mown in column 95 with Grazing = 'probably'. Without further information, these plots are left as grazed = 0. ##### *Datasets plots with grazed == 1 & Grazing (1/0) == '?':*

```
## [1] "PL_A_N003_0.0001b" "PL_A_N003_0.001b"  "PL_A_N003_0.01b"
## [4] "PL_A_N003_0.1b"    "PL_A_N003_100"     "PL_A_N003_10b"
## [7] "PL_A_N003_1b"
```

```
* PL_A contains 7 plots (PL_A_N003_xxxxb) noted as mown/grazed in column 95. These plots are left with g
```

##### *Datasets plots with grazed == 1 & Grazing (1/0) == '2':*

```
## [1] "PL_A_N006_0.0001a"
```

- PL_A contains 1 plots (PL_A_N006_0.0001a) with Grazing (1/0) == 2. This plot is left as grazed == 1 as this is likely a typo

##### *Datasets plots with grazed == 1 & Grazing (1/0) = NA:*

```
##  [1] "AM_A" "BG_A" "BY_A" "CH_B" "CN_A" "CN_B" "DE_B" "DE_C" "DE_D" "DE_E"
## [11] "DE_G" "DE_H" "DE_I" "DE_L" "DE_N" "DE_O" "DE_P" "DE_Q" "EE_B" "ES_A"
## [21] "ES_B" "ES_C" "ES_E" "ES_F" "EU_D" "EU_E" "EU_G" "EU_I" "EU_J" "FR_A"
## [31] "GR_A" "HR_A" "HU_C" "HU_D" "IL_A" "IT_A" "IT_B" "IT_C" "IT_D" "IT_E"
## [41] "IT_F" "IT_G" "IT_K" "IT_Q" "LV_A" "MN_A" "MN_B" "MN_C" "PL_C" "RO_A"
## [51] "RO_B" "RO_C" "RS_A" "RU_A" "SE_A" "SE_B" "SE_C" "SE_D" "SI_A" "UA_A"
## [61] "UA_D" "UA_E" "EU_K" "CH_E" "IT_R" "DK_A"
```

- There is 29219 plots with grazed == 1 and Grazing (1/0) = NA. These plots are left with grazed == 1

Datasets plots with grazed = NA & Grazing (1/0) == '?':

```
## [1] "PL_A_N003_0.0001a" "PL_A_N003_0.001a"  "PL_A_N003_0.01a"
## [4] "PL_A_N003_0.1a"    "PL_A_N003_10a"     "PL_A_N003_1a"
```

- PL_A contains 6 plots (PL_A_N003_xxxxa) with Grazing (1/0) = '?'. Without further information, these plots are left as grazed = NA