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')
```

PRELIMINARY NOTE: there 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

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,
```

Correction of the intensity/frequency columns

These two columns (mowing_frequency and grazing_intensity) are presently coded as text columns, because they contains odd values.
The column mowing_frequency contains 0 non-numeric values .
The column grazing_intensity contains 4 non-numeric values : low, middle, high, overgrazing.

Mowing frequency

The mowing_frequency does not have any odd values.
The values in the columns are left unchanged but the column is converted to numerical.

```
df$mowing_frequency <- as.numeric(df$mowing_frequency)
```

Grazing intensity

The grazing_intensity contains a mix of verbal description and numerical values. The verbal descriptors low, middle, high, overgrazing are replaced by numerical values and the column is converted to numerical:

- `overgrazing` and `high` are replaced by 1
- `middle` is replaced by 0.5
- `low` is replaced by 0.1

```
df[df$grazing_intensity %in% c('overgrazing', 'high'),]%<>%
  mutate(grazing_intensity = 1)

df[df$grazing_intensity %in% c('middle'),]%<>%
  mutate(grazing_intensity = 0.5)

df[df$grazing_intensity %in% c('low'),]%<>%
  mutate(grazing_intensity = 0.1)

df$grazing_intensity <- as.numeric(df$grazing_intensity)
```

Matching new binary columns and intensity or duration columns (yet to be automatized)

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.

NOTE in the following, I assumed that if the intensity/frequency/duration column was non empty, but the binary was, the intensity/frequency/duration column takes precedence (*i.e* if the frequency/intensity/duration, the binary column HAS to be == 1)

Mowing and mowing intensity

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

- The datasets EU_K, CZ_J contain 225 plots having a mowing intensity but classified as mown == 0. These plots have been manually checked the 8/11/2018 (masterfile v1.8) and contains not errors or additional information. They are thus reclassified as mown == 1.
  NOTE If necessary, we can discuss a frequency threshold the classify a plot as mown or not.

```
df[df$mowing_frequency != 0 & !is.na(df$mowing_frequency) & df$mown == 0,] %<>%
  mutate(mown = 1)
```

Post reclassification check: number of plots with a discrepancy - number of plots reclassified (should be 0) = 0

Grazing and Grazing intensity

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

- The datasets IR_A, PL_D, TR_B, UA_G, EU_K contain 164 plots having a mowing intensity but classified as grazed == 0. These plots have been manually checked the 8/11/2018 (masterfile v1.8) and contains not errors or additional information. They are thus reclassified as grazed == 1.

```
df[df$grazing_intensity != 0 & !is.na(df$grazing_intensity) & df$grazed == 0,] %<>%
  mutate(grazed = 1)
```

Post reclassification check: number of plots with a discrepancy - number of plots reclassified (should be 0) = -5 **Manual note: the difference found here is due to duplicates**

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. Here, we consider that mown == 1 or grazed == 1 is always correct, since it is based on the above correction (based either on the broad land-use or on the mowing / grazing intensity). Thus, we only identify (and potentialy correct) plots for which the new binary column is 0 or NA and the new one is not 0 or NA.

Mowing

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

Mowing (1/0)	0	1	<NA>
?	NA	NA	6
0	6042	213	22
1	NA	620	NA
NA	150296	6810	16068

Datasets with discrepancies between mown and Mowing (1/0) (automatic report):

```
## [1] "The dataset(s) PL_A contain(s) 6 plots with mown = NA and Mowing (1/0) = ?"
```

Manual verification of discrepancies (has to be updated manually if necessary for each version of the masterfile)

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

The following plots in PL_A have mown = NA and Mowing (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"
```

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

Check

Table 2: Contingency table of columns mown and Mowing (1/0) after reclassification (check)

Mowing (1/0)	0	1	<NA>
?	NA	NA	6
0	6042	213	22
1	NA	620	NA
NA	150296	6810	16068

Grazing

Table 3: Contingency table of columns grazed and Grazing (1/0)

Grazing (1/0)	0	1	<NA>
?	NA	7	6
0	3221	NA	17
1	6	5424	NA
2	NA	1	NA
probably	6	NA	NA
NA	126093	29223	16073

Datasets with discrepancies between grazed and Grazing (1/0) (automatic report):

```
## [1] "The dataset(s) PL_A contain(s) 6 plots with grazed = NA and Grazing (1/0) = ?"
## [1] "The dataset(s) PL_A contain(s) 6 plots with grazed == 0 and Grazing (1/0) = probably"
## [1] "The dataset(s) PL_A contain(s) 6 plots with grazed == 0 and Grazing (1/0) = 1"
```

Manual verification of discrepancies (has to be updated manually if necessary for each version of the masterfile)

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

The following plots in PL_A have grazed = NA and 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"
```

No additional information could be found, so the plots are left as grazed = NA

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

The following plots in PL_A have grazed == 0 and 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"
```

No additional information could be found, so the plots are left as grazed == 0.

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

The following plots in PL_A have grazed == 0 and 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)
```

Check

Table 4: Contingency table of columns grazed and Grazing (1/0) after reclassification (check

Grazing (1/0)	0	1	<NA>
?	NA	7	6
0	3221	NA	17
1	NA	5430	NA
2	NA	1	NA
probably	6	NA	NA
NA	126093	29223	16073

Burning

Table 5: Contingency table of columns burnt and Burning (1/0)

Burning (1/0)	0	1	<NA>
?	6	NA	NA
0	6363	NA	62
1	134	NA	3
NA	157439	39	16031

Datasets with discrepancies between burnt and Burning (1/0) (automatic report):

```
## [1] "The dataset(s) BY_A, IR_A, RS_A, RU_J, TJ_A, UA_G, UA_I, UA_J, CH_E, RU_L contain(s) 134 plots w
## [1] "The dataset(s) BY_A contain(s) 3 plots with burnt = NA and Burning (1/0) = 1"
## [1] "The dataset(s) PL_A contain(s) 6 plots with burnt == 0 and Burning (1/0) = ?"
```

Manual verification of discrepancies (has to be updated manually if necessary for each version of the masterfile)

Datasets with burnt == 0 and Burning (1/0) == 1

The following plots have burnt = 0 and Burning (1/0) == '1':

```
##   [1] "BY_A_P003"          "BY_A_P004"          "BY_A_P005"
##   [4] "IR_A_N004_0.0001a" "IR_A_N004_0.0001b" "IR_A_N004_0.001a"
##   [7] "IR_A_N004_0.001b"  "IR_A_N004_0.01a"   "IR_A_N004_0.01b"
##  [10] "IR_A_N004_0.1a"    "IR_A_N004_0.1b"    "IR_A_N004_1000"
##  [13] "IR_A_N004_100a"    "IR_A_N004_100b"    "IR_A_N004_10a"
##  [16] "IR_A_N004_10b"     "IR_A_N004_1a"      "IR_A_N004_1b"
##  [19] "IR_A_N004_25a"     "IR_A_N004_25b"     "RS_A_N010_100"
##  [22] "RS_A_N011_0.0001a" "RS_A_N011_0.0001b" "RS_A_N011_0.001a"
##  [25] "RS_A_N011_0.001b"  "RS_A_N011_0.01a"   "RS_A_N011_0.01b"
##  [28] "RS_A_N011_0.1a"    "RS_A_N011_0.1b"    "RS_A_N011_100"
##  [31] "RS_A_N011_10a"     "RS_A_N011_10b"     "RS_A_N011_1a"
##  [34] "RS_A_N011_1b"      "RU_J_P020"          "RU_J_P021"
##  [37] "TJ_A_N004_0.001a"  "TJ_A_N004_0.001b"  "TJ_A_N004_0.01a"
##  [40] "TJ_A_N004_0.01b"   "TJ_A_N004_0.1a"    "TJ_A_N004_0.1b"
##  [43] "TJ_A_N005_0.001a"  "TJ_A_N005_0.001b"  "TJ_A_N005_0.01a"
##  [46] "TJ_A_N005_0.01b"   "TJ_A_N005_0.1a"    "TJ_A_N005_0.1b"
##  [49] "TJ_A_N011_0.001a"  "TJ_A_N011_0.001b"  "TJ_A_N011_0.01a"
##  [52] "TJ_A_N011_0.01b"   "TJ_A_N011_0.1a"    "TJ_A_N011_0.1b"
##  [55] "TJ_A_N011_100"     "TJ_A_N011_10a"     "TJ_A_N011_10b"
##  [58] "TJ_A_N011_1a"      "TJ_A_N011_1b"      "TJ_A_N012_0.0001a"
```

```
##  [61] "TJ_A_N012_0.0001b" "UA_G_N008_100"     "UA_G_N009_1"
##  [64] "UA_G_N009_10"      "UA_G_N009_100"     "UA_G_N010_1"
##  [67] "UA_G_N010_10"      "UA_G_N010_100"     "UA_G_N011_1"
##  [70] "UA_G_N011_10"      "UA_G_N011_100"     "UA_G_N012_1"
##  [73] "UA_G_N012_10"      "UA_G_N012_100"     "UA_I_N005_0.001a"
##  [76] "UA_I_N005_0.001b"  "UA_I_N005_0.01a"   "UA_I_N005_0.01b"
##  [79] "UA_I_N005_0.1a"    "UA_I_N005_0.1b"    "UA_I_N005_100"
##  [82] "UA_I_N005_10a"     "UA_I_N005_10b"     "UA_I_N005_1a"
##  [85] "UA_I_N005_1b"      "UA_I_N006_0.0001a" "UA_I_N006_0.0001b"
##  [88] "UA_J_N005_0.01a"   "UA_J_N005_0.1a"    "UA_J_N005_1a"
##  [91] "UA_J_N005_10a"     "UA_J_N005_0.0001b" "UA_J_N005_0.001b"
##  [94] "UA_J_N005_0.01b"   "UA_J_N005_0.1b"    "UA_J_N005_1b"
##  [97] "UA_J_N005_10b"     "UA_J_N005_100"     "UA_J_N006_0.0001a"
## [100] "UA_J_N006_0.001a"  "UA_J_N006_0.01a"   "UA_J_N006_0.1a"
## [103] "UA_J_N006_1a"      "UA_J_N006_10a"     "UA_J_N006_0.0001b"
## [106] "UA_J_N006_0.001b"  "UA_J_N006_0.01b"   "UA_J_N006_0.1b"
## [109] "UA_J_N006_1b"      "UA_J_N006_10b"     "UA_J_N006_100"
## [112] "UA_J_N007_0.0001a" "UA_J_N007_0.001a"  "UA_J_N009_1b"
## [115] "UA_J_N009_10b"     "UA_J_N010_0.1b"    "UA_J_N010_10b"
## [118] "UA_J_N011_0.01a"   "UA_J_P010"         "UA_J_P046"
## [121] "UA_J_P058"         "CH_E_N003_0.01a"   "CH_E_N003_0.1a"
## [124] "CH_E_N003_1a"      "CH_E_N003_10a"     "CH_E_N003_0.0001b"
## [127] "CH_E_N003_0.001b"  "CH_E_N003_0.01b"   "CH_E_N003_0.1b"
## [130] "CH_E_N003_1b"      "CH_E_N003_10b"     "CH_E_N003_100"
## [133] "RU_L_N001_0.0001a" "RU_L_N001_0.0001b"
```

No other information are found. These plots are reclassified with burnt == 1

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

Datasets with burnt = NA and Burning (1/0) == 1

The following plots have burnt = NA and Burning (1/0) == '?':

```
## [1] "BY_A_P006" "BY_A_P007" "BY_A_P009"
```

No other information are found. These plots are reclassified with burnt == 1

```
df[is.na(df$burnt) & df$`Burning (1/0)` %in% '1',]%<>%
    mutate(burnt = 1)
```

Datasets with burnt = NA and Burning (1/0) == ?

The following plots have burnt == 0 and Burning (1/0) == '?':

```
## [1] "PL_A_N005_0.0001a" "PL_A_N005_0.001a"  "PL_A_N005_0.01a"
## [4] "PL_A_N005_0.1a"    "PL_A_N005_10a"     "PL_A_N005_1a"
```

No additional information could be found, so the plots are left as burnt == 0.

Check

Table 6: Contingency table of columns burnt and Burning (1/0) after reclassification (check)

Burning (1/0)	0	1	<NA>
?	6	NA	NA
0	6363	NA	62
1	NA	137	NA
NA	157439	39	16031

Fertilization

Table 7: Contingency table of columns fertilized and Fertilized (1/0)

Fertilized (1/0)	0	<NA>
0	5112	37
0.3	24	NA
0.5	112	NA
1	196	NA
NA	158537	16059

This field seems to be a fertilization intensity rather than a binary variable. We should discuss it before making a decision.

Proposal for a new land-use template

The column 96 provide a lot of information that is so far not fully exploited. To make use of it, it requires to manually check each of the value in this column, to fill other columns. After the discussion with Anne and Monika, I propose the following columns:

Land-use columns

Already explained above, these are binary columns mown, grazed, burnt, fertilized, abandoned, natural (*i.e. no land-use past or present). There is one extra column that contains verbal information not found in these five columns.
NOTE This should be discussed, but these columns could also be viewed as management columns.

Land-cover columns

This columns (binary) could be used to provide a clearer view on the land-use. After having briefly browsed the column 96, a first proposal is: grassland, moorland, heathland, fallowland, meadow, pasture (not entirely sure about this one).
NOTE Monika was suggesting to find a way to separate primary and secondary grasslands. I'm not sure how this fits in this proposal, so it's open to discussion.

Time since abandonment

Self explanatory. I'm not quite sure wether this overlap with Ex arable years (years since last ploughing). Maybe that could be combined with an additional column former_land_use, so that will not be restricted to ex-arable plots.

Grazing animal

Type of animal grazing the land. So far, we have five categories in mind: cow, sheep, goat, horse, other. I am not sure wether grazing by wild animals (which is sometimes present in the land-use details column) should

be considered, as it is not necessarily linked with land-use.

Table 8: Summary of the proposed template (**to be discussed**)

Column type	Column name	Variable type	Possible values
Land use			
	• mown	• binary	• 0/1
	• grazed	• binary	• 0/1
	• burnt	• binary	• 0/1
	• fertilized	• binary	• 0/1
	• abandoned	• binary	• 0/1
	• natural	• binary	• 0/1
	• other	• text	• undefined
Land cover			
	• grassland	• binary	• 0/1
	• moorland	• binary	• 0/1
	• heathland	• binary	• 0/1
	• fallowland	• binary	• 0/1
	• meadow	• binary	• 0/1
	• pasture	• binary	• 0/1
Land-use intensity			
	• grazing intensity	• numeric	• 0 to 1
	• mowing frequency	• numeric	• undefined
	• fertilisation intensity	• numeric	• undefined
	• time since abandonement	• numeric	• undefined
Grazing details			
	• grazing animal	• text	• cow, sheep, goat, horse, other