Replace NAs and restructurate calibration data

Marcos Angelini

April 19, 2016

```
## Loading required package: boot
```

Replace NAs and restructurate calibration data

Calibration data for second paper has to be structured in such a form that we have all the variables per location in one row. They are three soil properties (CEC, OC and clay) in three major horizons (A, B and C) and the covariate values that belong to each site. If one row has any missing value, all the row (sample) is lost. We have 344 soil profiles, however they have many NAs. Here, I explain the criteria to replace some NAs and get a calibration data set suitable for our study.

First, a summary of soil properties: CEC, OC and Clay at A, B and C

```
round(stat.desc(m[,2:10],norm = T),3)
```

```
##                 CEC.A    CEC.B    CEC.C    OC.A    OC.B    OC.C   clay.A
## nbr.val       328.000  330.000  261.000 325.000 330.000 263.000  326.000
## nbr.null        0.000    0.000    0.000   0.000   0.000   0.000    0.000
## nbr.na          6.000    4.000   73.000   9.000   4.000  71.000    8.000
## min            11.200   14.900    8.400   0.220   0.121   0.010    9.079
## max            33.700   49.122   52.000   3.019   1.240   0.580   49.200
## range          22.500   34.222   43.600   2.799   1.119   0.570   40.121
## sum          7179.389 9349.561 5834.999 614.493 158.461  31.580 8052.417
## median         22.117   28.466   21.200   1.880   0.472   0.110   25.036
## mean           21.888   28.332   22.356   1.891   0.480   0.120   24.701
## SE.mean         0.199    0.317    0.425   0.026   0.008   0.004    0.248
## CI.mean.0.95    0.392    0.623    0.836   0.052   0.015   0.007    0.488
## var            13.013   33.145   47.034   0.227   0.020   0.003   20.062
## std.dev         3.607    5.757    6.858   0.476   0.140   0.058    4.479
## coef.var        0.165    0.203    0.307   0.252   0.291   0.481    0.181
## skewness        0.205    0.227    1.041  -0.331   0.940   2.772    0.361
## skew.2SE        0.763    0.844    3.451  -1.225   3.501   9.229    1.337
## kurtosis        0.522    0.330    1.720   0.633   2.824  16.881    3.099
## kurt.2SE        0.972    0.617    2.863   1.174   5.275  28.203    5.754
## normtest.W      0.990    0.991    0.945   0.983   0.958   0.820    0.964
## normtest.p      0.024    0.038    0.000   0.001   0.000   0.000    0.000
##                 clay.B   clay.C
## nbr.val        331.000  253.000
## nbr.null         0.000    0.000
## nbr.na           3.000   81.000
## min             12.575    4.800
## max             59.300   43.000
## range           46.725   38.200
## sum          12494.296 4749.323
## median          38.111   18.000
## mean            37.747   18.772
## SE.mean          0.454    0.487
## CI.mean.0.95     0.893    0.960
## var             68.149   60.104
```

```
## std.dev          8.255    7.753
## coef.var         0.219    0.413
## skewness         0.018    0.644
## skew.2SE         0.069    2.104
## kurtosis        -0.161    0.028
## kurt.2SE        -0.302    0.046
## normtest.W       0.998    0.964
## normtest.p       0.944    0.000
```

It can be seen that C horizons have the highest NAs. It is partly because several profiles do not have C horizon. Let see it

```
length(unique(sp$id.p)) # number of soil profiles
```

```
## [1] 334
```

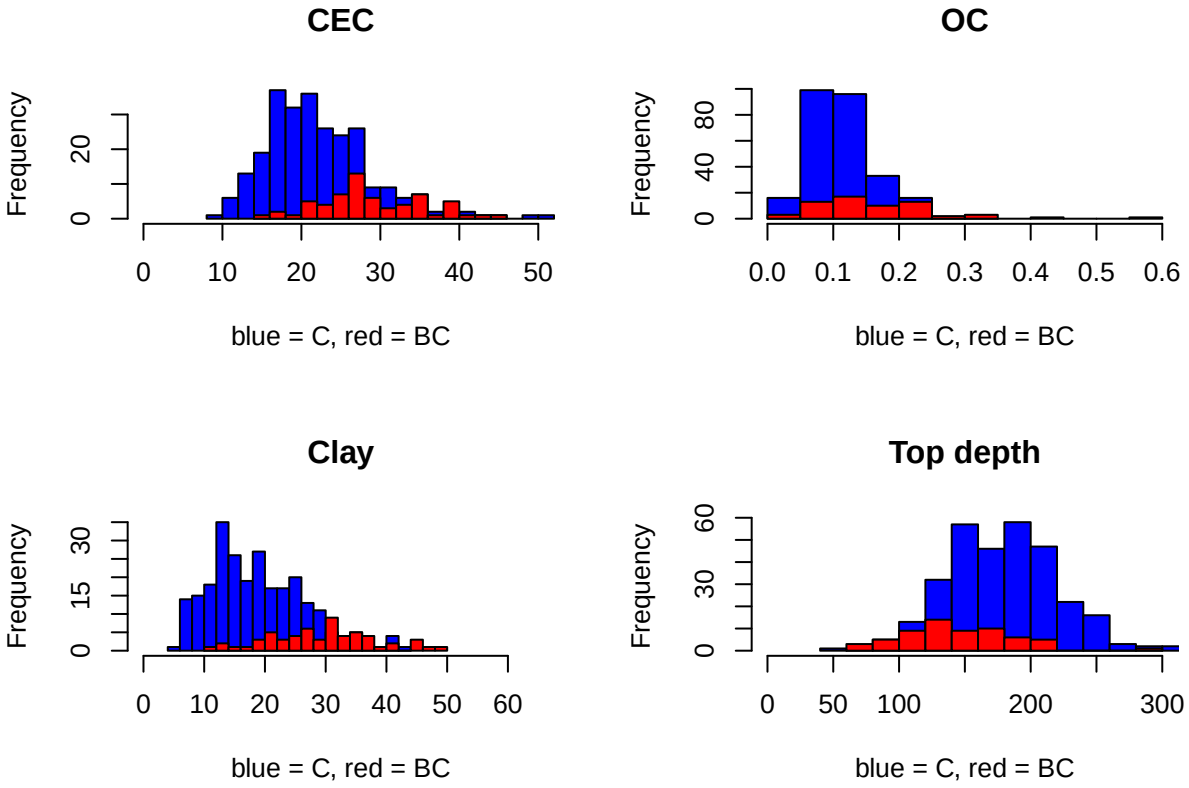
```
length(unique(sp$id.p[sp$hor=="C"])) # number of soil profiles with C horizons
```

```
## [1] 268
```

```
length(m$id.p[complete.cases(m)]) # number of complete cases
```

```
## [1] 233
```

An alternative option to fill these gaps is to take the deepest BC horizons. The following histogram that shows CEC, OC and clay in C (blue) and BC (red) horizons



If we include these horizons as C horizons, the statistics will change as follows:

```
round(stat.desc(n[,2:10],norm = T),3)
```

```
##                  CEC.A    CEC.B    CEC.C    OC.A    OC.B    OC.C   clay.A
## nbr.val        328.000  330.000  323.000 325.000 330.000 324.000  326.000
## nbr.null         0.000    0.000    0.000   0.000   0.000   0.000    0.000
## nbr.na           6.000    4.000   11.000   9.000   4.000  10.000    8.000
## min             11.200   14.900    8.400   0.220   0.121   0.010    9.079
## max             33.700   49.122   52.000   3.019   1.240   0.580   49.200
## range           22.500   34.222   43.600   2.799   1.119   0.570   40.121
## sum           7179.389 9349.561 7629.599 614.493 158.461  41.160 8052.417
## median          22.117   28.466   22.600   1.880   0.472   0.120   25.036
## mean            21.888   28.332   23.621   1.891   0.480   0.127   24.701
## SE.mean          0.199    0.317    0.406   0.026   0.008   0.003    0.248
## CI.mean.0.95     0.392    0.623    0.800   0.052   0.015   0.007    0.488
## var             13.013   33.145   53.372   0.227   0.020   0.004   20.062
## std.dev          3.607    5.757    7.306   0.476   0.140   0.063    4.479
## coef.var         0.165    0.203    0.309   0.252   0.291   0.493    0.181
## skewness         0.205    0.227    0.801  -0.331   0.940   2.040    0.361
## skew.2SE         0.763    0.844    2.951  -1.225   3.501   7.530    1.337
## kurtosis         0.522    0.330    0.688   0.633   2.824   9.478    3.099
## kurt.2SE         0.972    0.617    1.272   1.174   5.275  17.546    5.754
## normtest.W       0.990    0.991    0.961   0.983   0.958   0.870    0.964
## normtest.p       0.024    0.038    0.000   0.001   0.000   0.000    0.000
##                  clay.B   clay.C
## nbr.val        331.000  312.000
## nbr.null         0.000    0.000
## nbr.na           3.000   22.000
## min             12.575    4.800
## max             59.300   49.000
## range           46.725   44.200
## sum          12494.296 6484.823
## median          38.111   19.500
## mean            37.747   20.785
## SE.mean          0.454    0.506
## CI.mean.0.95     0.893    0.996
## var             68.149   79.956
## std.dev          8.255    8.942
## coef.var         0.219    0.430
## skewness         0.018    0.610
## skew.2SE         0.069    2.209
## kurtosis        -0.161   -0.108
## kurt.2SE        -0.302   -0.196
## normtest.W       0.998    0.966
## normtest.p       0.944    0.000
```

```
# number of soil profiles (the difference belongs to soil profiles without
# data in any of the three soil properties)
length(unique(n$id.p))
```

```
## [1] 334
```

```
length(n$id.p[complete.cases(n)]) # number of complete cases
```

```
## [1] 289
```

There are several profiles that have not A, B or C horizons. They are removed in the next step

```
t <- wt.mean.properties(data = sp, properties = c("CEC", "OC","clay"))
t <- as.matrix(t)
t[is.nan(t)] <- NA
t <- as.data.frame(t)
sp[which(sp$id.p %in% t[is.na(t$OC.A),]$id.p),] # profiles without A hz
```

```
##     id.p hor top bottom thick   tb  CEC  phw phkcl  resist   OC clay
## 53   356   A   0     18    18 14.6 16.4  6.8   6.1  555.00   NA 17.2
## 54   356   B  18     32    14   NA 20.1  9.0   8.0  209.00   NA 15.5
## 55   356   B  32     54    22   NA 38.5  8.9   7.9  154.00   NA 31.4
## 56   356   C  54     74    20   NA 35.1  8.9   7.7  222.00   NA 28.7
## 116  379   A   0     16    16 21.8 19.1  8.0   6.6  767.00   NA   NA
## 117  379   B  16     48    32   NA 20.2  9.4   7.7  441.00   NA   NA
## 118  379   B  48     80    32   NA 30.6  9.6   7.7  383.00 0.19 49.1
## 119  379   C 160    180    20 27.3 23.0  9.2   6.8  843.00 0.07 28.2
## 172  417   A   0     20    20 21.1 16.7  8.6   7.0 1840.00   NA   NA
## 173  417   B  20     43    23 31.5 28.2  9.2   7.4 1169.00   NA   NA
## 174  417   B  43     58    15 38.2 36.5  8.9   7.1 1015.00 0.31 59.3
## 175  417   C  58     75    17 35.3 33.7  8.4   6.5 1246.00 0.13 39.9
## 184  421   A   0     18    18 18.8 16.4  9.5   7.7  920.00   NA   NA
## 185  421   B  18     46    28   NA 29.4  9.7   7.8    4.60   NA   NA
## 186  421   B  46     72    26   NA 33.6  9.2   7.5    6.12 0.27 54.0
## 187  421   C  72     92    20 32.1 28.0  8.9   7.0 1000.00 0.12 31.2
## 215  433   B  14     48    34   NA 27.4  9.9   7.7  468.00 0.39 38.8
## 216  433   B  48     80    32   NA 32.6 10.0   7.6  377.00 0.25 39.1
## 366  502   B  33     64    31 25.2 26.0  8.9   7.1  631.00 0.45 31.8
## 367  502   B  64     90    26   NA 34.3  9.1   7.4  735.00 0.34 39.3
## 368  502   C 150    170    20 24.9 24.2  8.3   6.7 1277.00 0.13 16.3
## 384  508   B  29     55    26   NA 34.1  9.1   7.6  734.00 0.45 44.2
## 385  508   B  55     80    25   NA 35.5  8.9   7.3  757.00 0.38 45.9
## 386  508   C 105    130    25 24.0 23.5  8.6   6.9 1378.00 0.12 13.6
## 506  539   B  16     40    24 23.6 39.2  9.1   7.4  951.00 0.65 38.5
## 507  539   C  60     90    30   NA 26.2  9.2   7.5 1338.00 0.11 20.1
## 508  539   C  90    120    30   NA 20.5  9.2   7.5 1647.00 0.03 17.7
## 509  539   C 120    140    20 20.8 17.9  8.2   6.1 2456.00 0.05 12.9
## 675  601   B  13     38    25   NA 29.4  9.1   7.2  771.00 0.61 45.7
## 676  601   B  38     64    26   NA 29.8  9.1   7.2  792.00 0.30 27.8
## 677  601   C 114    140    26 18.4 18.3  8.8   6.3 2210.00 0.07 17.5
##     silt20 sand.mf
## 53    31.8     8.4
## 54    32.6     8.1
## 55    29.4     3.7
## 56    28.7     5.0
## 116     NA      NA
## 117     NA      NA
## 118   18.9     4.3
```

```
## 119   32.3     6.5
## 172     NA      NA
## 173     NA      NA
## 174   17.4     3.6
## 175   31.9     3.9
## 184     NA      NA
## 185     NA      NA
## 186   22.9     3.5
## 187   33.1     5.5
## 215   24.9    13.2
## 216   23.1    14.9
## 366   23.2     9.3
## 367   12.6     8.6
## 368   21.4    17.7
## 384   17.1     5.9
## 385   16.2     5.6
## 386   27.1    10.9
## 506   26.6     8.2
## 507   31.2    15.6
## 508   27.3    15.0
## 509   25.5    16.6
## 675   14.9    16.8
## 676   24.9    16.0
## 677   19.0    29.1
```

```
sp[which(sp$id.p %in% t[is.na(t$OC.B),]$id.p),] # profiles without B hz
```

```
##     id.p hor top bottom thick   tb  CEC phw phkcl resist   OC clay silt20
## 53   356   A   0     18    18 14.6 16.4 6.8   6.1    555   NA 17.2   31.8
## 54   356   B  18     32    14   NA 20.1 9.0   8.0    209   NA 15.5   32.6
## 55   356   B  32     54    22   NA 38.5 8.9   7.9    154   NA 31.4   29.4
## 56   356   C  54     74    20   NA 35.1 8.9   7.7    222   NA 28.7   28.7
## 649  592   A   0     21    21  9.8 12.9 5.7   4.9     NA 0.87 11.2    6.1
## 650  592   A  21     35    14 11.6 14.4 6.2   5.1     NA 0.71 13.2    5.6
## 651  592   C  90    130    40  9.7 11.8 7.1   5.4     NA 0.12  8.5    4.2
## 652  592   C 130    150    20 10.1 12.3 7.0   5.4     NA 0.10  8.3    3.6
## 692  608   A   0     16    16 12.5 16.7 6.6   5.1   4850 1.38 16.4   12.7
## 693  608   A  16     38    22 14.6 18.3 6.8   5.1   7760 1.12 19.6   13.4
## 694  608   C 130    150    20 11.5 11.8 7.6   5.6   9700 0.09 15.0   11.8
## 695  609   A   0     16    16 11.3 12.9 6.1   5.0   3891 0.84  8.5    4.6
## 696  609   A  16     38    22  9.5 13.1 6.7   5.0   9231 0.69  9.5    3.4
## 697  609   C  73    140    67  9.7 10.5 7.5   5.8   9955 0.09  8.0    2.6
##     sand.mf
## 53      8.4
## 54      8.1
## 55      3.7
## 56      5.0
## 649    56.2
## 650    46.9
## 651    34.6
## 652    23.2
## 692    50.5
## 693    46.6
## 694    51.7
```

```
## 695    67.7
## 696    66.4
## 697    69.4
```

```
sp[which(sp$id.p %in% t[is.na(t$OC.C),]$id.p),] # profiles without C hz
```

```
##      id.p hor top bottom thick   tb  CEC  phw phkcl resist   OC clay
## 53    356   A   0     18    18 14.6 16.4  6.8   6.1    555   NA 17.2
## 54    356   B  18     32    14   NA 20.1  9.0   8.0    209   NA 15.5
## 55    356   B  32     54    22   NA 38.5  8.9   7.9    154   NA 31.4
## 56    356   C  54     74    20   NA 35.1  8.9   7.7    222   NA 28.7
## 70    363   A  42     60    18   NA 22.0  8.7   7.5   1165 0.64 24.6
## 71    363   A  60     84    24   NA 22.0  8.5   7.3   1831 0.64 27.7
## 72    363   B  84    104    20   NA 24.3  8.2   7.0   1290 0.29 39.6
## 145   404   A   0     20    20   NA 32.8  9.5   8.1    358 0.75 25.3
## 146   404   B  20     40    20   NA 52.4  9.4   7.8    304 0.33 59.0
## 147   404   B  40     65    25   NA 46.5  9.1   7.3    519 0.23 35.7
## 215   433   B  14     48    34   NA 27.4  9.9   7.7    468 0.39 38.8
## 216   433   B  48     80    32   NA 32.6 10.0   7.6    377 0.25 39.1
## 217   441   A   0     16    16 20.0 25.7  6.7   5.4   4782 2.18 26.5
## 218   441   A  16     23     7 17.6 20.6  7.0   5.4   5602 1.29 29.3
## 219   441   B  34     52    18 20.1 21.9  7.0   5.3   2581 0.63 38.6
## 220   441   B  52     74    22 22.6 24.2  7.6   5.4   2541 0.44 37.8
## 221   441   C 115    135    20 16.9 16.2  7.4   5.6   4113   NA 20.2
## 250   465   A   0     15    15 16.9 23.3  6.1   5.0   2614 2.55 22.5
## 251   465   A  15     28    13 16.6 19.5  6.8   5.5   3940 1.54 25.0
## 252   465   B  28     48    20 16.4 17.4  7.0   5.5   3644 0.72 25.5
## 253   465   B  48     72    24 14.8 16.4  7.2   5.7   3704 0.38 23.4
## 254   465   C 135    155    20 13.8 14.4  7.8   5.7   5910   NA 18.1
## 861   670   A   0      5     5 14.4 17.7  6.3   5.0   2610 3.06 18.1
## 862   670   A   5     15    10 19.7 18.9  8.5   6.6   1469 1.41 24.0
## 863   670   B  15     35    20 32.6 33.6  9.2   7.2    578 0.82 47.0
## 864   670   B  35     60    25 28.1 29.9  9.4   7.3    743 0.37 31.0
## 865   670   C 130    150    20 31.8 21.8  9.0   6.8   1692   NA 18.5
## 977   701   A   0     10    10 18.7 17.5  8.8   7.5    238 1.54 19.3
## 978   701   B  20     40    20   NA 47.7  8.8   7.5    178 0.91 40.1
## 979   701   B  40     56    16   NA 33.7  8.7   7.4    193 0.70 37.2
## 1154  742   A   0     20    20 14.5 16.5  6.6   5.3   4944 1.83 22.5
## 1155  742   B  27     55    28   NA 27.5  8.9   6.9    721 0.54 45.7
## 1156  742   B  55     80    25   NA 27.2  9.0   7.2    453 0.38 28.0
## 1157  742   C 130    175    45   NA 21.6  9.3   7.4    906   NA 20.6
## 1289  774   A   0     18    18 21.7 23.2  6.5   5.4   4592 2.21 26.4
## 1290  774   A  18     29    11 22.3 23.3  6.7   5.5   4592 1.67 30.7
## 1291  774   B  40     67    27 33.1 35.2  6.5   4.5   4945 0.28 53.5
## 1292  774   B  67     97    30 29.8 30.3  6.7   4.8   4945   NA 40.2
## 1293  774   B  97    140    43 35.1 34.2  6.9   4.9   4945   NA 39.9
## 1294  774   C 210    255    45   NA 32.0  8.0   6.6   4239   NA 33.9
##      silt20 sand.mf
## 53     31.8     8.4
## 54     32.6     8.1
## 55     29.4     3.7
## 56     28.7     5.0
## 70     34.1     3.3
## 71     33.2     3.1
```

```
## 72     22.1     4.2
## 145    30.2    10.5
## 146    13.7     4.8
## 147    21.3     8.0
## 215    24.9    13.2
## 216    23.1    14.9
## 217    28.2    16.7
## 218    25.3    16.0
## 219    21.2    15.8
## 220    16.7    21.4
## 221    18.2    28.1
## 250    19.3    35.3
## 251    19.9    30.9
## 252    16.4    35.3
## 253    14.8    38.7
## 254    11.5    44.0
## 861    26.5    19.0
## 862    21.6    18.0
## 863    20.8     8.8
## 864    25.8    13.3
## 865    25.9    17.5
## 977    31.7    12.2
## 978    25.6     9.3
## 979    22.9    10.9
## 1154   31.0    12.8
## 1155   25.0     7.4
## 1156   30.4    10.4
## 1157   37.9     8.8
## 1289   31.1     4.0
## 1290   30.7     7.1
## 1291   21.0     5.9
## 1292   25.6     7.1
## 1293   27.8     5.4
## 1294   34.5     5.6
```

```r
sp <- sp[sp$id.p!=433,] # no A, no C
sp <- sp[sp$id.p!=502,] # no A
sp <- sp[sp$id.p!=508,] # no A
sp <- sp[sp$id.p!=539,] # no A
sp <- sp[sp$id.p!=601,] # no A
sp <- sp[sp$id.p!=592,] # no B
sp <- sp[sp$id.p!=608,] # no B
sp <- sp[sp$id.p!=609,] # no B
sp <- sp[sp$id.p!=363,] # no C
sp <- sp[sp$id.p!=404,] # no C
sp <- sp[sp$id.p!=701,] # no C
sp <- sp[sp$id.p!=749,] # no CEC at any hz.
t <- wt.mean.properties(data = sp, properties = c("CEC", "OC","clay"))
round(stat.desc(t[,2:10]),3)
```

```
##                CEC.A    CEC.B    CEC.C    OC.A    OC.B    OC.C   clay.A
## nbr.val       322.000  322.000  316.000 318.000 321.000 316.000  319.000
## nbr.null        0.000    0.000    0.000   0.000   0.000   0.000    0.000
## nbr.na          0.000    0.000    6.000   4.000   1.000   6.000    3.000
```

```
## min            11.200   14.900    8.400   0.220   0.121   0.010   13.400
## max            33.700   47.339   52.000   3.019   1.240   0.580   49.200
## range          22.500   32.439   43.600   2.799   1.119   0.570   35.800
## sum          7062.947 9071.363 7507.345 607.885 154.495  40.372 7925.514
## median         22.181   28.302   22.662   1.888   0.477   0.120   25.138
## mean           21.935   28.172   23.757   1.912   0.481   0.128   24.845
## SE.mean         0.195    0.313    0.410   0.026   0.008   0.004    0.243
## CI.mean.0.95    0.385    0.615    0.806   0.050   0.015   0.007    0.477
## var            12.303   31.493   53.018   0.209   0.019   0.004   18.769
## std.dev         3.507    5.612    7.281   0.458   0.139   0.063    4.332
## coef.var        0.160    0.199    0.306   0.239   0.288   0.495    0.174
##                 clay.B   clay.C
## nbr.val        322.000  304.000
## nbr.null         0.000    0.000
## nbr.na           0.000   18.000
## min             12.575    4.800
## max             59.300   49.000
## range           46.725   44.200
## sum          12148.296 6374.089
## median          38.038   19.759
## mean            37.728   20.967
## SE.mean          0.464    0.514
## CI.mean.0.95     0.913    1.012
## var             69.292   80.439
## std.dev          8.324    8.969
## coef.var         0.221    0.428
```

Now, it can be seen that there are still several NAs sparsed in the variables. To replace some of them we could: 1) define constant value or 2) predict value using other soil properties as predictors. OC.C has 9 NAs. However, the amount of OC in C horizon is negligible. Values arround 0.15 may have low signal-to-noise ratio. For this reason replacing them for the median (which is not affected by extreme values) should not have high impact in the modelling step.

```
# replace NA at OC.C with a constant value
sp$OC[sp$hor=="C" & is.na(sp$OC)] <- median(t$OC.C, na.rm = TRUE)
t <- wt.mean.properties(data = sp, properties = c("CEC", "OC","clay"))
length(t$id.p[complete.cases(t)]) # number of complete cases
```

```
## [1] 294
```

This could be the calibration data.

```
write.csv(t[complete.cases(t),], "/mnt/L0135974_DATA/UserData/BaseARG/2_Calibration/calib.data-5.0.csv")
##############################################################################################
```

For OC at A and B horizon, let us first to analise the NA.

```
# replace NA at OC.C with a constant value
sp$OC[sp$hor=="C" & is.na(sp$OC)] <- median(t$OC.C, na.rm = TRUE)

a <- sp$id.p[sp$hor=="A" & is.na(sp$OC) & !is.na(sp$phw)] # OC.A is NA
b <- sp$id.p[sp$hor=="B" & is.na(sp$OC) & !is.na(sp$phw)] # OC.A is NA
```

```r
# profile 356 has clay and CEC values, but profiles 379, 417 and 421 have only CEC
sp[which(sp$id.p %in% a),-14] # profiles with OC.A = NA
```

```
##     id.p hor top bottom thick   tb  CEC phw phkcl  resist   OC clay silt20
## 53   356   A   0     18    18 14.6 16.4 6.8   6.1  555.00   NA 17.2   31.8
## 54   356   B  18     32    14   NA 20.1 9.0   8.0  209.00   NA 15.5   32.6
## 55   356   B  32     54    22   NA 38.5 8.9   7.9  154.00   NA 31.4   29.4
## 56   356   C  54     74    20   NA 35.1 8.9   7.7  222.00 0.12 28.7   28.7
## 116  379   A   0     16    16 21.8 19.1 8.0   6.6  767.00   NA   NA     NA
## 117  379   B  16     48    32   NA 20.2 9.4   7.7  441.00   NA   NA     NA
## 118  379   B  48     80    32   NA 30.6 9.6   7.7  383.00 0.19 49.1   18.9
## 119  379   C 160    180    20 27.3 23.0 9.2   6.8  843.00 0.07 28.2   32.3
## 172  417   A   0     20    20 21.1 16.7 8.6   7.0 1840.00   NA   NA     NA
## 173  417   B  20     43    23 31.5 28.2 9.2   7.4 1169.00   NA   NA     NA
## 174  417   B  43     58    15 38.2 36.5 8.9   7.1 1015.00 0.31 59.3   17.4
## 175  417   C  58     75    17 35.3 33.7 8.4   6.5 1246.00 0.13 39.9   31.9
## 184  421   A   0     18    18 18.8 16.4 9.5   7.7  920.00   NA   NA     NA
## 185  421   B  18     46    28   NA 29.4 9.7   7.8    4.60   NA   NA     NA
## 186  421   B  46     72    26   NA 33.6 9.2   7.5    6.12 0.27 54.0   22.9
## 187  421   C  72     92    20 32.1 28.0 8.9   7.0 1000.00 0.12 31.2   33.1
```

```r
# Again, rofile 356, 379, 417 and 421 are in the list.
sp[which(sp$id.p %in% b),-14] # profiles with OC.A = NA
```

```
##      id.p hor top bottom thick   tb  CEC phw phkcl  resist   OC clay
## 53    356   A   0     18    18 14.6 16.4 6.8   6.1  555.00   NA 17.2
## 54    356   B  18     32    14   NA 20.1 9.0   8.0  209.00   NA 15.5
## 55    356   B  32     54    22   NA 38.5 8.9   7.9  154.00   NA 31.4
## 56    356   C  54     74    20   NA 35.1 8.9   7.7  222.00 0.12 28.7
## 116   379   A   0     16    16 21.8 19.1 8.0   6.6  767.00   NA   NA
## 117   379   B  16     48    32   NA 20.2 9.4   7.7  441.00   NA   NA
## 118   379   B  48     80    32   NA 30.6 9.6   7.7  383.00 0.19 49.1
## 119   379   C 160    180    20 27.3 23.0 9.2   6.8  843.00 0.07 28.2
## 172   417   A   0     20    20 21.1 16.7 8.6   7.0 1840.00   NA   NA
## 173   417   B  20     43    23 31.5 28.2 9.2   7.4 1169.00   NA   NA
## 174   417   B  43     58    15 38.2 36.5 8.9   7.1 1015.00 0.31 59.3
## 175   417   C  58     75    17 35.3 33.7 8.4   6.5 1246.00 0.13 39.9
## 184   421   A   0     18    18 18.8 16.4 9.5   7.7  920.00   NA   NA
## 185   421   B  18     46    28   NA 29.4 9.7   7.8    4.60   NA   NA
## 186   421   B  46     72    26   NA 33.6 9.2   7.5    6.12 0.27 54.0
## 187   421   C  72     92    20 32.1 28.0 8.9   7.0 1000.00 0.12 31.2
## 1289  774   A   0     18    18 21.7 23.2 6.5   5.4 4592.00 2.21 26.4
## 1290  774   A  18     29    11 22.3 23.3 6.7   5.5 4592.00 1.67 30.7
## 1291  774   B  40     67    27 33.1 35.2 6.5   4.5 4945.00 0.28 53.5
## 1292  774   B  67     97    30 29.8 30.3 6.7   4.8 4945.00   NA 40.2
## 1293  774   B  97    140    43 35.1 34.2 6.9   4.9 4945.00   NA 39.9
## 1294  774   C 210    255    45   NA 32.0 8.0   6.6 4239.00 0.12 33.9
##      silt20
## 53     31.8
## 54     32.6
## 55     29.4
## 56     28.7
## 116      NA
```

```
## 117      NA
## 118    18.9
## 119    32.3
## 172      NA
## 173      NA
## 174    17.4
## 175    31.9
## 184      NA
## 185      NA
## 186    22.9
## 187    33.1
## 1289   31.1
## 1290   30.7
## 1291   21.0
## 1292   25.6
## 1293   27.8
## 1294   34.5
```

A solution may be to predict the value of OC.A at profiles *356*, *379*, *417* and *421*, and OC.B at profile *356*. It is proposed a MLR using soil properties that will not be used in SEM. This method is analogous to pedotransfer functions and are implemented below.

```
# creating subsets
sp.A <- sp[sp$hor=="A",] #subset of A horizons
sp.B <- sp[sp$hor=="B",] #subset of B horizons
sp.C <- sp[sp$hor=="C",] #subset of C horizons
# MLR for OC at A horizon
lm.OC.A <- lm(OC~CEC+bottom+thick+phw+resist, sp.A) # MLR for A hz.
summary(lm.OC.A)
```

```
##
## Call:
## lm(formula = OC ~ CEC + bottom + thick + phw + resist, data = sp.A)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.20387 -0.23961 -0.02572  0.25063  2.09443
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.417e+00  2.114e-01   6.701 5.49e-11 ***
## CEC          8.058e-02  4.979e-03  16.185  < 2e-16 ***
## bottom      -2.767e-02  2.427e-03 -11.402  < 2e-16 ***
## thick        2.152e-02  4.147e-03   5.190 3.05e-07 ***
## phw         -1.710e-01  2.481e-02  -6.891 1.64e-11 ***
## resist       8.422e-06  1.226e-05   0.687    0.493
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3868 on 509 degrees of freedom
##   (39 observations deleted due to missingness)
## Multiple R-squared:  0.5005, Adjusted R-squared:  0.4956
## F-statistic:   102 on 5 and 509 DF,  p-value: < 2.2e-16
```

```r
lm.OC.B <- lm(OC~CEC+bottom+thick+phw+resist, sp.B) # MLR for B hz.
summary(lm.OC.B)
```

```
##
## Call:
## lm(formula = OC ~ CEC + bottom + thick + phw + resist, data = sp.B)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39527 -0.08019 -0.01094  0.06847  0.62237
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.134e+00  7.481e-02  15.152  < 2e-16 ***
## CEC          4.811e-03  1.162e-03   4.142 3.97e-05 ***
## bottom      -5.634e-03  2.668e-04 -21.115  < 2e-16 ***
## thick        3.072e-03  6.490e-04   4.734 2.79e-06 ***
## phw         -6.156e-02  6.804e-03  -9.047  < 2e-16 ***
## resist       2.176e-06  6.370e-06   0.342    0.733
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1392 on 563 degrees of freedom
##   (40 observations deleted due to missingness)
## Multiple R-squared:  0.4987, Adjusted R-squared:  0.4942
## F-statistic:   112 on 5 and 563 DF,  p-value: < 2.2e-16
```

```r
# prediction
sp$OC[which(sp$id.p %in% a & sp$hor=="A")] <-# sp where hz is A, OC is NA and
  # pH is not NA is predicted with lm.OC.A
  predict(lm.OC.A,sp[which(sp$id.p %in% a & sp$hor=="A"),])
sp$OC[which(sp$id.p %in% 356 & sp$hor=="B")] <-# sp where hz is B, OC is NA and
  # pH is not NA is predicted with lm.OC.B
  predict(lm.OC.B,sp[which(sp$id.p %in% 356 & sp$hor=="B"),])
# statistics
t <- wt.mean.properties(data = sp, properties = c("CEC", "OC","clay"))
t <- as.matrix(t);t[is.nan(t)] <- NA;t <- as.data.frame(t)
round(stat.desc(t[,2:10]),3)
```

```
##                 CEC.A    CEC.B    CEC.C    OC.A    OC.B    OC.C   clay.A
## nbr.val       322.000  322.000  316.000 322.000 322.000 322.000  319.000
## nbr.null        0.000    0.000    0.000   0.000   0.000   0.000    0.000
## nbr.na          0.000    0.000    6.000   0.000   0.000   0.000    3.000
## min            11.200   14.900    8.400   0.220   0.121   0.010   13.400
## max            33.700   47.339   52.000   3.019   1.240   0.580   49.200
## range          22.500   32.439   43.600   2.799   1.119   0.570   35.800
## sum          7062.947 9071.363 7507.345 613.045 155.031  41.100 7925.514
## median         22.181   28.302   22.662   1.883   0.477   0.120   25.138
## mean           21.935   28.172   23.757   1.904   0.481   0.128   24.845
## SE.mean         0.195    0.313    0.410   0.026   0.008   0.003    0.243
## CI.mean.0.95    0.385    0.615    0.806   0.050   0.015   0.007    0.477
## var            12.303   31.493   53.018   0.212   0.019   0.004   18.769
## std.dev         3.507    5.612    7.281   0.461   0.138   0.063    4.332
```

```
## coef.var        0.160    0.199    0.306   0.242   0.287   0.490    0.174
##                clay.B   clay.C
## nbr.val       322.000  304.000
## nbr.null        0.000    0.000
## nbr.na          0.000   18.000
## min            12.575    4.800
## max            59.300   49.000
## range          46.725   44.200
## sum         12148.296 6374.089
## median         38.038   19.759
## mean           37.728   20.967
## SE.mean         0.464    0.514
## CI.mean.0.95    0.913    1.012
## var            69.292   80.439
## std.dev         8.324    8.969
## coef.var        0.221    0.428
```

```
dim(t[complete.cases(t),])
```

```
## [1] 295  10
```

This could be the calibration data.

```
write.csv(t[complete.cases(t),], "/mnt/L0135974_DATA/UserData/BaseARG/2_Calibration/calib.data-5.1.csv")
#################################################################################################
```

Now, OC has not NAs. The larger amount of NAs remain in clay. We predict their values using MLR, as we did before.

```
# creating subsets
sp.A <- sp[sp$hor=="A",] #subset of A horizons
sp.B <- sp[sp$hor=="B",] #subset of B horizons
sp.C <- sp[sp$hor=="C",] #subset of C horizons

lm.clay.A <- lm(clay~CEC*OC+bottom+thick+phw+resist, sp.A) # MLR for A hz.
summary(lm.clay.A)
```

```
##
## Call:
## lm(formula = clay ~ CEC * OC + bottom + thick + phw + resist,
##     data = sp.A)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7427 -1.8621 -0.1299  1.6356 10.3694
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0160220  3.0304866   0.005   0.9958
## CEC          1.4424462  0.1095605  13.166  < 2e-16 ***
## OC           3.1191823  1.3561873   2.300   0.0219 *
## bottom       0.0443461  0.0211276   2.099   0.0363 *
```

```
## thick       -0.0313585  0.0329796  -0.951   0.3421
## phw         -0.5251550  0.2040200  -2.574   0.0103 *
## resist       0.0001319  0.0000949   1.390   0.1651
## CEC:OC      -0.2445822  0.0543059  -4.504 8.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.998 on 507 degrees of freedom
##   (39 observations deleted due to missingness)
## Multiple R-squared:  0.5137, Adjusted R-squared:  0.507
## F-statistic: 76.51 on 7 and 507 DF,  p-value: < 2.2e-16
```

```
lm.clay.C <- lm(clay~CEC+bottom+thick+phw+resist, sp.C) # MLR for A hz.
summary(lm.clay.C)
```

```
##
## Call:
## lm(formula = clay ~ CEC + bottom + thick + phw + resist, data = sp.C)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1054  -5.0213  -0.4454   4.2901  23.3543
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.0127644  5.4010997   2.039   0.0423 *
## CEC          0.7704470  0.0683644  11.270   <2e-16 ***
## bottom      -0.0142779  0.0099079  -1.441   0.1506
## thick        0.0061109  0.0190716   0.320   0.7489
## phw         -0.5657600  0.5112155  -1.107   0.2693
## resist      -0.0005805  0.0003359  -1.728   0.0850 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.788 on 306 degrees of freedom
##   (44 observations deleted due to missingness)
## Multiple R-squared:  0.4359, Adjusted R-squared:  0.4267
## F-statistic: 47.29 on 5 and 306 DF,  p-value: < 2.2e-16
```

```
a <- sp$id.p[sp$hor=="A" & is.na(sp$clay)] # clay.A is NA
c <- sp$id.p[sp$hor=="C" & is.na(sp$clay)] # clay.C is NA

# prediction
# clay where hz is A & clay is NA is predicted with lm.clay.A
sp$clay[which(sp$id.p %in% c(379,417,421) & sp$hor=="A")] <-
  predict(lm.clay.A,sp[which(sp$id.p %in% c(379,417,421) & sp$hor=="A"),])
# clay where hz is C & clay is NA is predicted with lm.clay.C
c <- c[c(-(4:7),-13,-19)]
sp$clay[which(sp$id.p %in% c & sp$hor=="C")] <-# sp where hz is B, OC is NA and
  # pH is not NA is predicted with lm.OC.B
  predict(lm.clay.C,sp[which(sp$id.p %in% c & sp$hor=="C"),])

t <- wt.mean.properties(data = sp, properties = c("CEC", "OC","clay"))
```

```
t <- as.matrix(t);t[is.nan(t)] <- NA;t <- as.data.frame(t)
round(stat.desc(t[,2:10]),3)
```

```
##                  CEC.A    CEC.B    CEC.C    OC.A    OC.B    OC.C   clay.A
## nbr.val        322.000  322.000  316.000 322.000 322.000 322.000  322.000
## nbr.null         0.000    0.000    0.000   0.000   0.000   0.000    0.000
## nbr.na           0.000    0.000    6.000   0.000   0.000   0.000    0.000
## min             11.200   14.900    8.400   0.220   0.121   0.010   13.400
## max             33.700   47.339   52.000   3.019   1.240   0.580   49.200
## range           22.500   32.439   43.600   2.799   1.119   0.570   35.800
## sum           7062.947 9071.363 7507.345 613.045 155.031  41.100 7983.951
## median          22.181   28.302   22.662   1.883   0.477   0.120   25.036
## mean            21.935   28.172   23.757   1.904   0.481   0.128   24.795
## SE.mean          0.195    0.313    0.410   0.026   0.008   0.003    0.242
## CI.mean.0.95     0.385    0.615    0.806   0.050   0.015   0.007    0.476
## var             12.303   31.493   53.018   0.212   0.019   0.004   18.878
## std.dev          3.507    5.612    7.281   0.461   0.138   0.063    4.345
## coef.var         0.160    0.199    0.306   0.242   0.287   0.490    0.175
##                  clay.B   clay.C
## nbr.val        322.000  322.000
## nbr.null         0.000    0.000
## nbr.na           0.000    0.000
## min             12.575    4.800
## max             59.300   49.000
## range           46.725   44.200
## sum          12148.296 6797.410
## median          38.038   20.350
## mean            37.728   21.110
## SE.mean          0.464    0.489
## CI.mean.0.95     0.913    0.961
## var             69.292   76.900
## std.dev          8.324    8.769
## coef.var         0.221    0.415
```

```
dim(t[complete.cases(t),])[1] #number of soil profiles
```

```
## [1] 316
```

This could be the calibration data.

```
write.csv(t[complete.cases(t),], "/mnt/L0135974_DATA/UserData/BaseARG/2_Calibration/calib.data-5.2.csv")
##############################################################################################
```

Finally, we predict CEC at C (6 NA).

```
# creating subsets
sp.C <- sp[sp$hor=="C",] #subset of C horizons

lm.CEC.C <- lm(CEC~clay+bottom+thick+phw, sp.C) # MLR for C hz. (resistance is not available)
summary(lm.CEC.C)
```

```
##
## Call:
## lm(formula = CEC ~ clay + bottom + thick + phw, data = sp.C)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9187  -3.3050  -0.2128   2.8279  22.1089
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.688438   2.938671   0.915 0.360922
## clay         0.526638   0.033582  15.682  < 2e-16 ***
## bottom       0.028280   0.007391   3.826 0.000155 ***
## thick       -0.020783   0.013969  -1.488 0.137746
## phw          0.715105   0.337680   2.118 0.034928 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.454 on 339 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.4474, Adjusted R-squared:  0.4409
## F-statistic: 68.62 on 4 and 339 DF,  p-value: < 2.2e-16
```

```
c <- sp$id.p[sp$hor=="C" & is.na(sp$CEC)] # clay.C is NA

# prediction
# CEC where hz is A & CEC is NA is predicted with lm.CEC.A
sp$CEC[which(sp$id.p %in% c & sp$hor=="C")] <-
  predict(lm.CEC.C,sp[which(sp$id.p %in% c & sp$hor=="C"),])

t <- wt.mean.properties(data = sp, properties = c("CEC", "OC","clay"))
t <- as.matrix(t);t[is.nan(t)] <- NA;t <- as.data.frame(t)
round(stat.desc(t[,2:10]),3)
```

```
##                CEC.A    CEC.B    CEC.C    OC.A    OC.B    OC.C   clay.A
## nbr.val       322.000  322.000  322.000 322.000 322.000 322.000  322.000
## nbr.null        0.000    0.000    0.000   0.000   0.000   0.000    0.000
## nbr.na          0.000    0.000    0.000   0.000   0.000   0.000    0.000
## min            11.200   14.900    8.400   0.220   0.121   0.010   13.400
## max            33.700   47.339   52.000   3.019   1.240   0.580   49.200
## range          22.500   32.439   43.600   2.799   1.119   0.570   35.800
## sum          7062.947 9071.363 7659.651 613.045 155.031  41.100 7983.951
## median         22.181   28.302   22.820   1.883   0.477   0.120   25.036
## mean           21.935   28.172   23.788   1.904   0.481   0.128   24.795
## SE.mean         0.195    0.313    0.403   0.026   0.008   0.003    0.242
## CI.mean.0.95    0.385    0.615    0.794   0.050   0.015   0.007    0.476
## var            12.303   31.493   52.422   0.212   0.019   0.004   18.878
## std.dev         3.507    5.612    7.240   0.461   0.138   0.063    4.345
## coef.var        0.160    0.199    0.304   0.242   0.287   0.490    0.175
##                clay.B   clay.C
## nbr.val       322.000  322.000
## nbr.null        0.000    0.000
## nbr.na          0.000    0.000
## min            12.575    4.800
```

```
## max             59.300   49.000
## range           46.725   44.200
## sum          12148.296 6797.410
## median          38.038   20.350
## mean            37.728   21.110
## SE.mean          0.464    0.489
## CI.mean.0.95     0.913    0.961
## var             69.292   76.900
## std.dev          8.324    8.769
## coef.var         0.221    0.415
```

```
dim(t[complete.cases(t),])[1] #number of soil profiles
```

```
## [1] 322
```

This could be the calibration data.

```
write.csv(t[complete.cases(t),], "/mnt/L0135974_DATA/UserData/BaseARG/2_Calibration/calib.data-5.3.csv")
##############################################################################################
```