Preliminary report

Intro

This is a quick and dirty report from the preliminary data analysis, built with RMarkdown language. It is just meant for internal reporting, to collect together the results for discussion and to explain them briefly with captions.

I modify slightly the data: when there are two altitudes reported in the same cell, I use the average of the two

The first step is to normalize the C stocks and sequestration rates by depth. I do so using the calibrated equation 1 from te following reference: Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D., & Van Meirvenne, M. (2009). Predicting Soil Organic Carbon Stock Using Profile Depth Distribution Functions and Ordinary Kriging. Soil Science Society of America Journal, 73(2), 614–621. https://doi.org/10.2136/sssaj2007.0410
I calculate the integral of such equation over 20 cm:
$C_{0,Z} = \int_0^Z ae^- bZ$
and the integral over each depth considered in the table, and then calculate the ratio:
$r = \frac{\int_0^{20} ae^{-b \cdot 20}}{\int_0^Z ae^{-bZ}}$
Which I then use to multiply all the values involvinc C stocks (both stocks and sequestration rates).

Random forest model calibration and optimization

The RF models are trained with the Caret package

```
#Random Forests analysis
model1<-randomForest(SOC_Storage_rate_t_ha_yr ~ ., data=cardinael_data_subset[,-13])
model2<-randomForest(AFS_Stock_t_ha ~ ., data=cardinael_data_subset[,-13])
```

The following are the selected variables for the plots after some discussion between me and Remi:

```
selected_names
```

```
##  [1] "Longitude"                          "Latitude"
##  [3] "Region"                             "Mean_annual_rainfall"
##  [5] "Mean_annual_temperature"            "IPCC_Climate"
##  [7] "Agroforestry_classification"        "Previous_land_use"
##  [9] "Age_yrs"                            "SOC_Storage_rate_t_ha_yr_normalized"
## [11] "Control_Stock_t_ha_normalized"      "AFS_Stock_t_ha_normalized"
## [13] "Soil_type"                          "Total_tree_density"
## [15] "Altitude_masl"
```

`model1` considers all the variables we decided to include except `AFS_Stock_t_ha` in order to predict the SOC sequestration rate `SOC_Storage_rate_t_ha_yr`.
`model2` considers all the variables we decided to include (including of course in this case `AFS_Stock_t_ha`) in order to predict the SOC stocks after agroforestry LUC `AFS_Stock_t_ha`.

Random forest results of comprehensive models

We then proceed to exclude the previous SOC stocks `Control_Stock_t_ha` , which are of course a rather important determinant of the SOC situation and possibly also of the emissions at a certain time, and we proceed to create all version x.2 of the models.

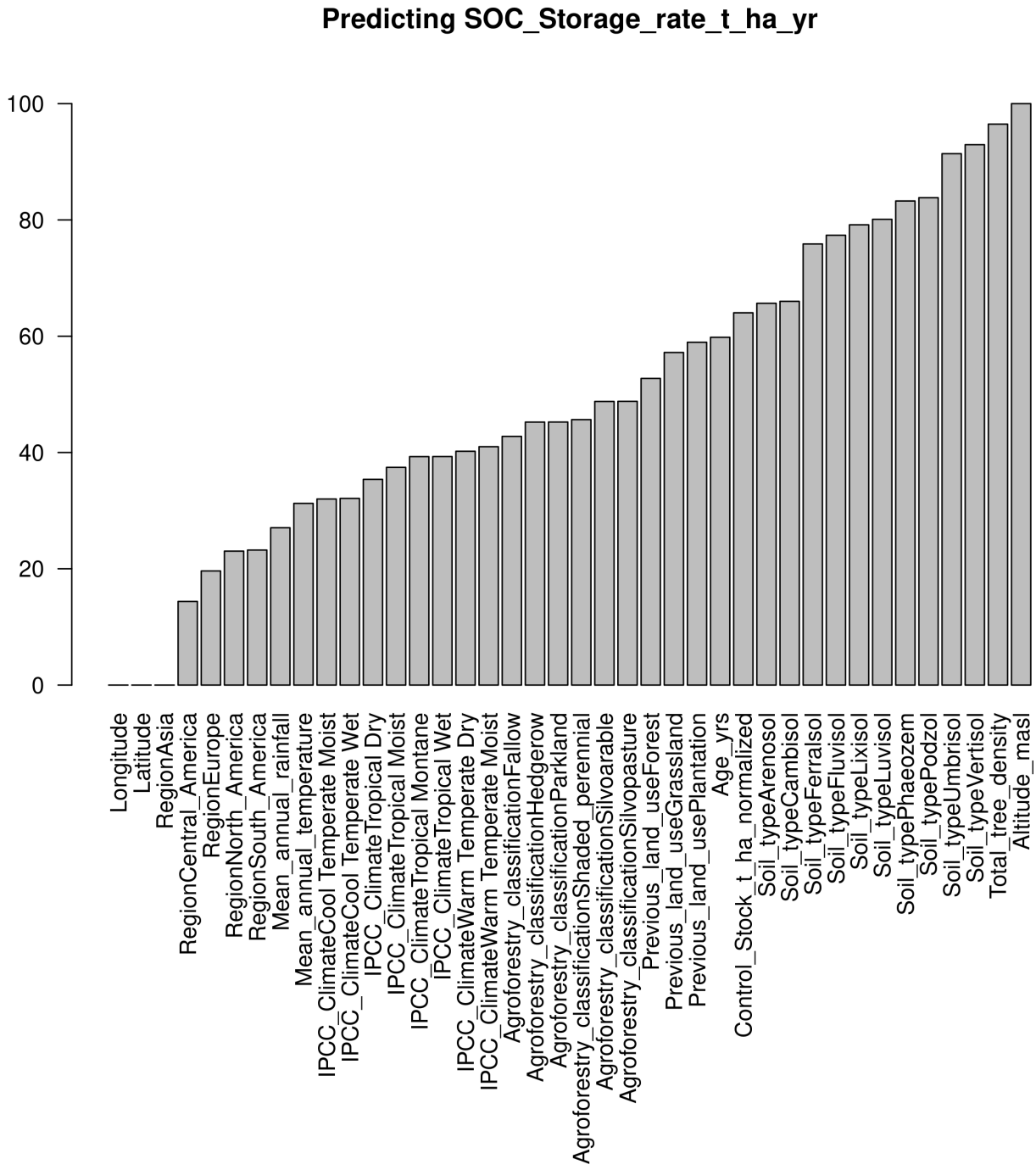


Figure 1: *Comprehensive model to predict C accumulation rates, decomposition of explained variance.*

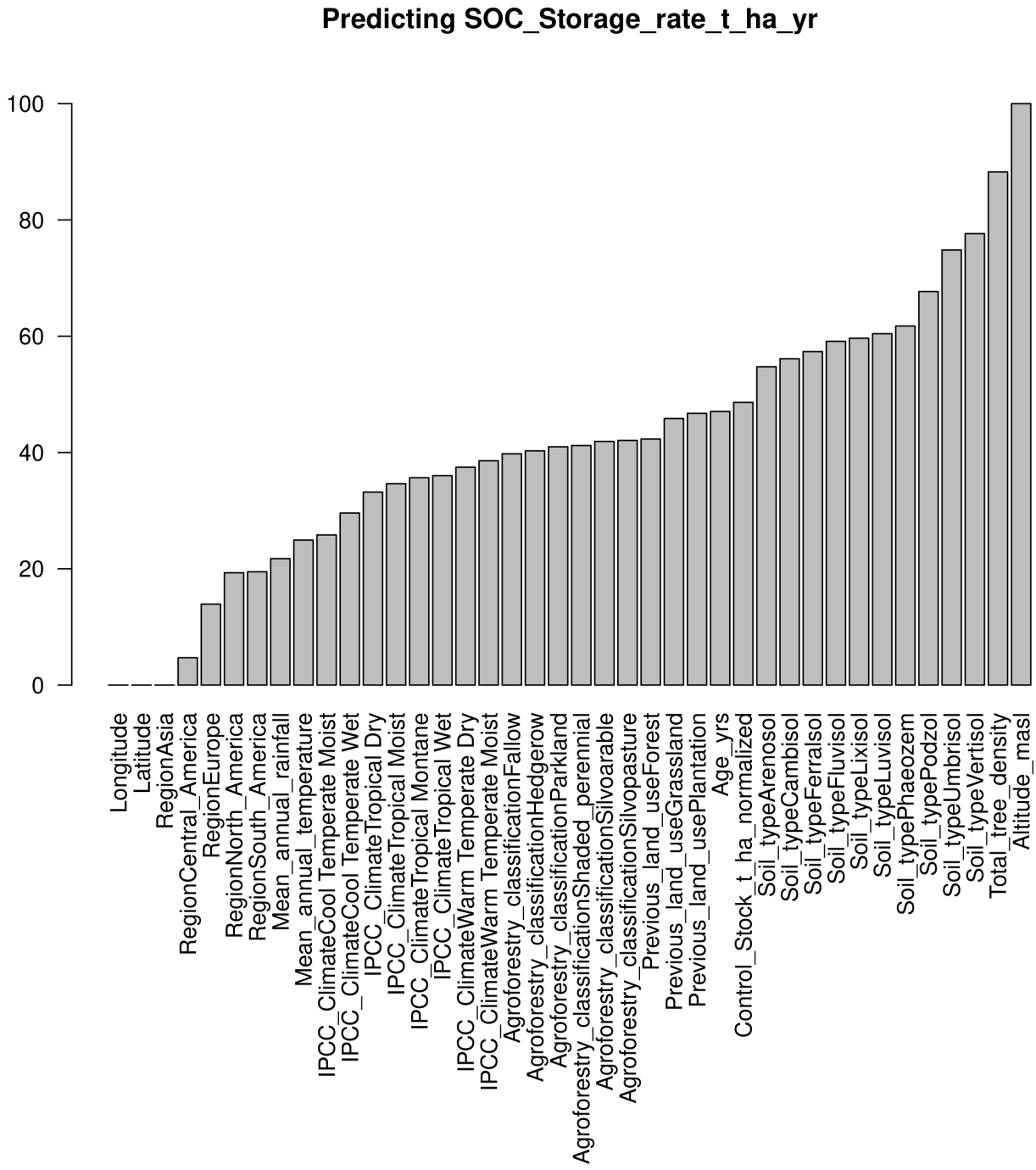


Figure 2: *Comprehensive model to predict C stocks, decomposition of explained variance.*

```
#subsetting the dataset for escluding previous SOC
which_exclude_2<-which(names(cardinael_data_subset) %in% c("Control_Stock_t_ha"))
cardinael_data_subset_2<-cardinael_data_subset[,-which_exclude_2]
#Random Forests analysis
model1.2<-randomForest(SOC_Storage_rate_t_ha_yr ~ ., data=cardinael_data_subset_2)
model2.2<-randomForest(AFS_Stock_t_ha ~ ., data=cardinael_data_subset_2)
```

Random forest results of models not considering previous stocks

This model considers the following variables:

```
colnames(cardinael_data_subset_training_2)
```

```
##  [1] "Longitude"                          "Latitude"
##  [3] "Region"                             "Mean_annual_rainfall"
##  [5] "Mean_annual_temperature"            "IPCC_Climate"
##  [7] "Agroforestry_classification"        "Previous_land_use"
##  [9] "Age_yrs"                            "SOC_Storage_rate_t_ha_yr_normalized"
## [11] "AFS_Stock_t_ha_normalized"          "Soil_type"
## [13] "Total_tree_density"                 "Altitude_masl"
```

We then exclude (on top of revious SOC stocks `Control_Stock_t_ha`) also the age of the stocks `Age_yrs` , and we proceed to create all version x.3 of the models.

```
#subsetting the dataset for escluding age
which_exclude_3<-which(names(cardinael_data_subset_2) %in% c("Age_yrs"))
cardinael_data_subset_3<-cardinael_data_subset_2[,-which_exclude_3]
names(cardinael_data_subset_3)
#Random Forests analysis
model1.3<-randomForest(SOC_Storage_rate_t_ha_yr ~ ., data=cardinael_data_subset_3[,-11])
model2.3<-randomForest(AFS_Stock_t_ha ~ ., data=cardinael_data_subset_3)
```

Random forest results of models not considering previous C stocks nor age (time elapsed)

This model considers the following variables:

```
colnames(cardinael_data_subset_training_3)
```

```
##  [1] "Longitude"                          "Latitude"
##  [3] "Region"                             "Mean_annual_rainfall"
##  [5] "Mean_annual_temperature"            "IPCC_Climate"
##  [7] "Agroforestry_classification"        "Previous_land_use"
##  [9] "SOC_Storage_rate_t_ha_yr_normalized" "AFS_Stock_t_ha_normalized"
## [11] "Soil_type"                          "Total_tree_density"
## [13] "Altitude_masl"
```

Random forest results of minimal model

I also prepared a version of a predictive model with less parameters. This model considers the following variables:

```
colnames(cardinael_data_subset_training_4)
```

```
##  [1] "Region"                             "Mean_annual_rainfall"
##  [3] "Mean_annual_temperature"            "IPCC_Climate"
##  [5] "Agroforestry_classification"        "Previous_land_use"
```

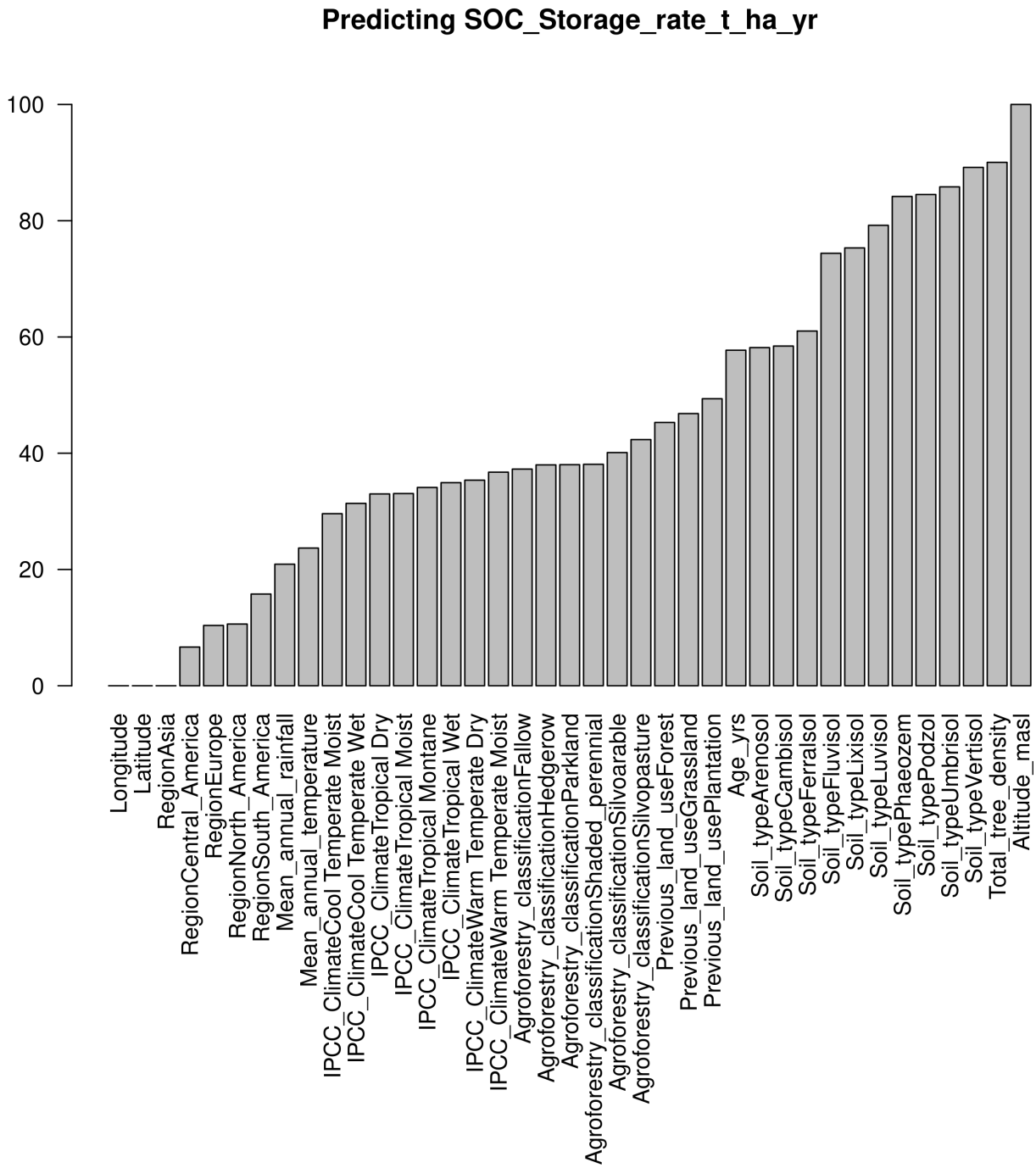


Figure 3: *Model to predict C accumulation rates not considering former C stocks in the control, decomposition of explained variance.*

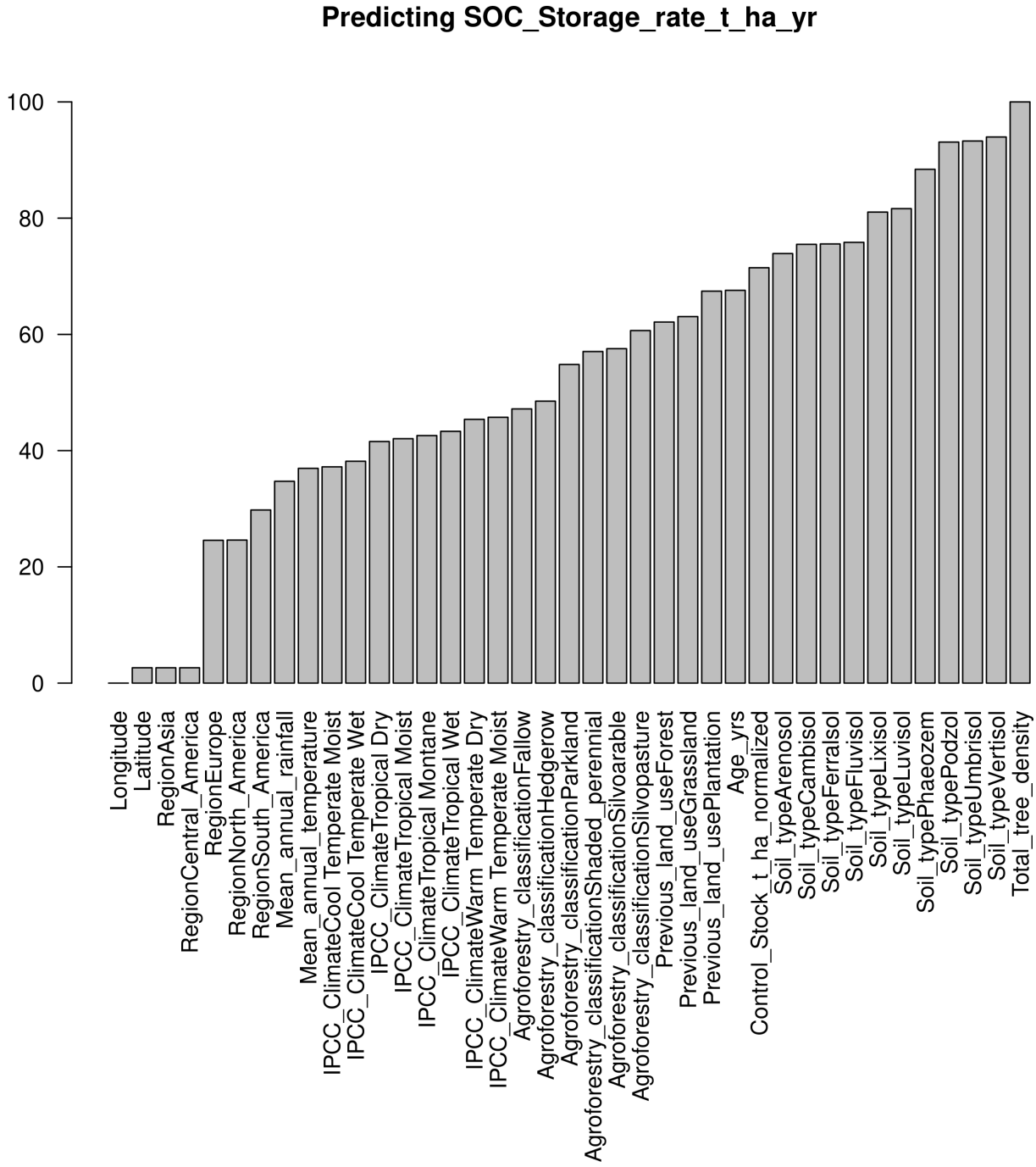


Figure 4: *Model to predict C stocks not considering former C stocks in the control, decomposition of explained variance.*

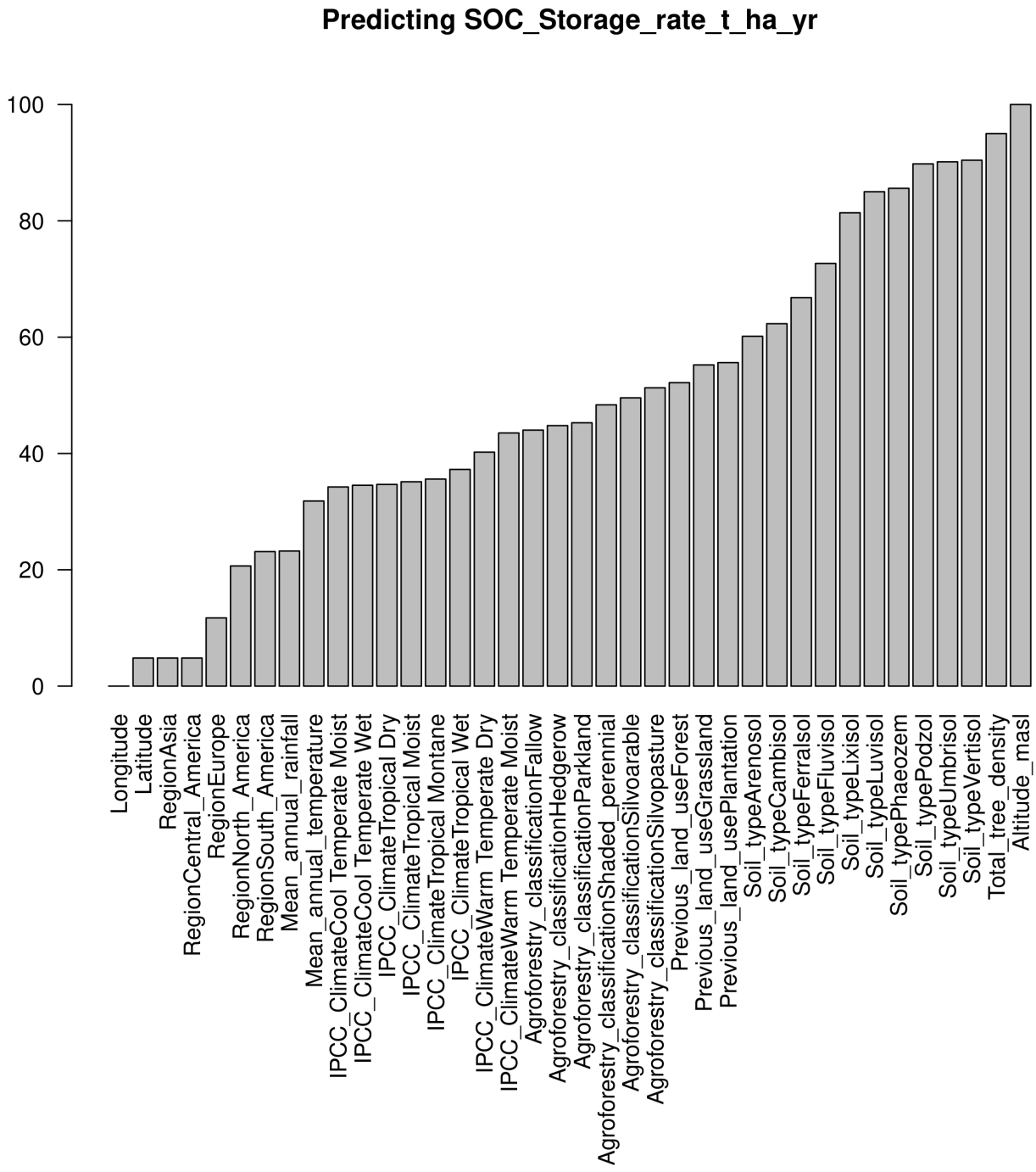


Figure 5: *Model to predict C accumulation rates not considering former C stocks in the control nor time elapsed in the experiment, decomposition of explained variance.*

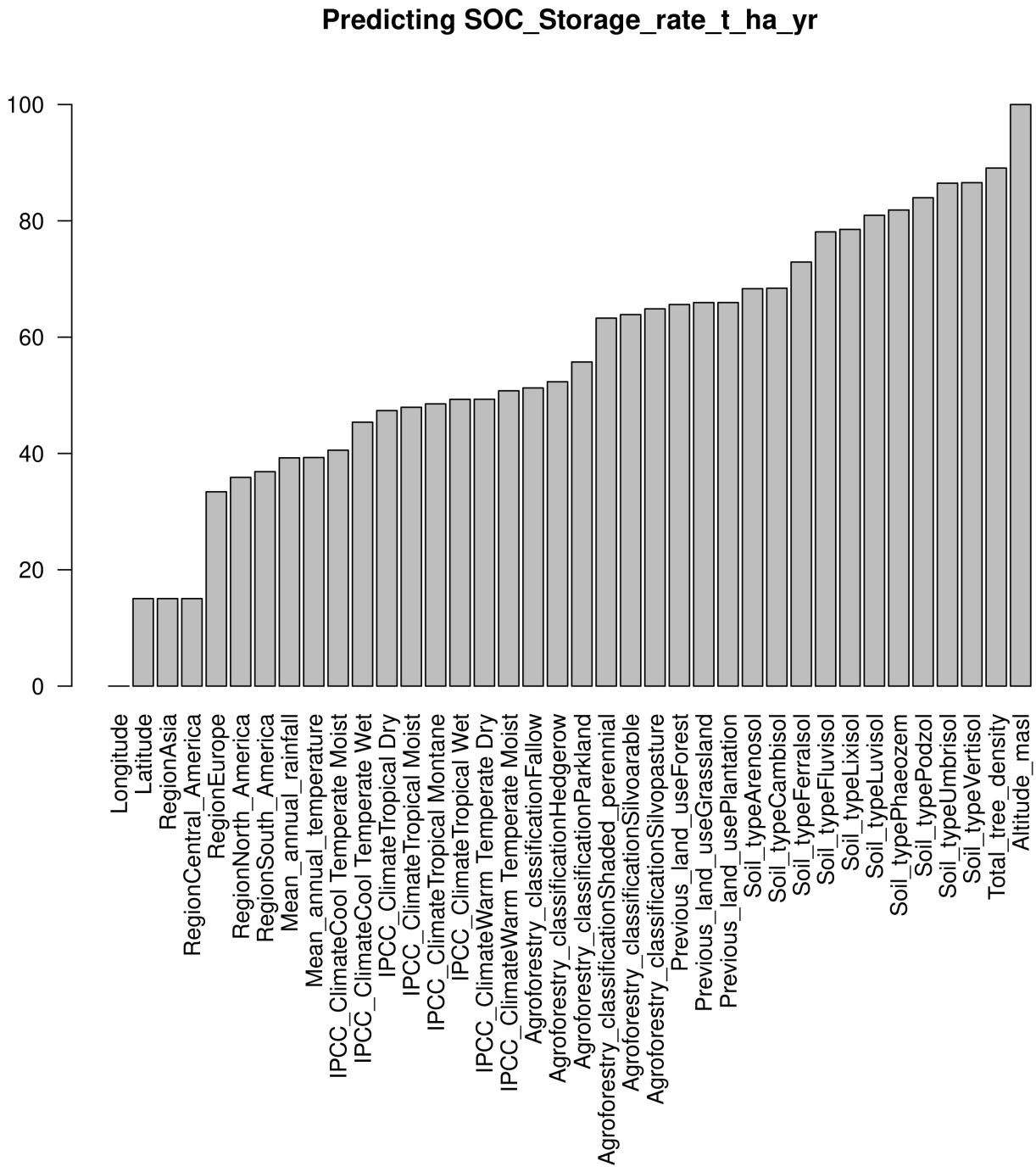


Figure 6: *Model to predict C stocks not considering former C stocks in the control nor time elapsed in the experiment, decomposition of explained variance.*

```
##  [7] "Age_yrs"                          "SOC_Storage_rate_t_ha_yr_normalized"
##  [9] "AFS_Stock_t_ha_normalized"        "Soil_type"
## [11] "Total_tree_density"               "Altitude_masl"
```

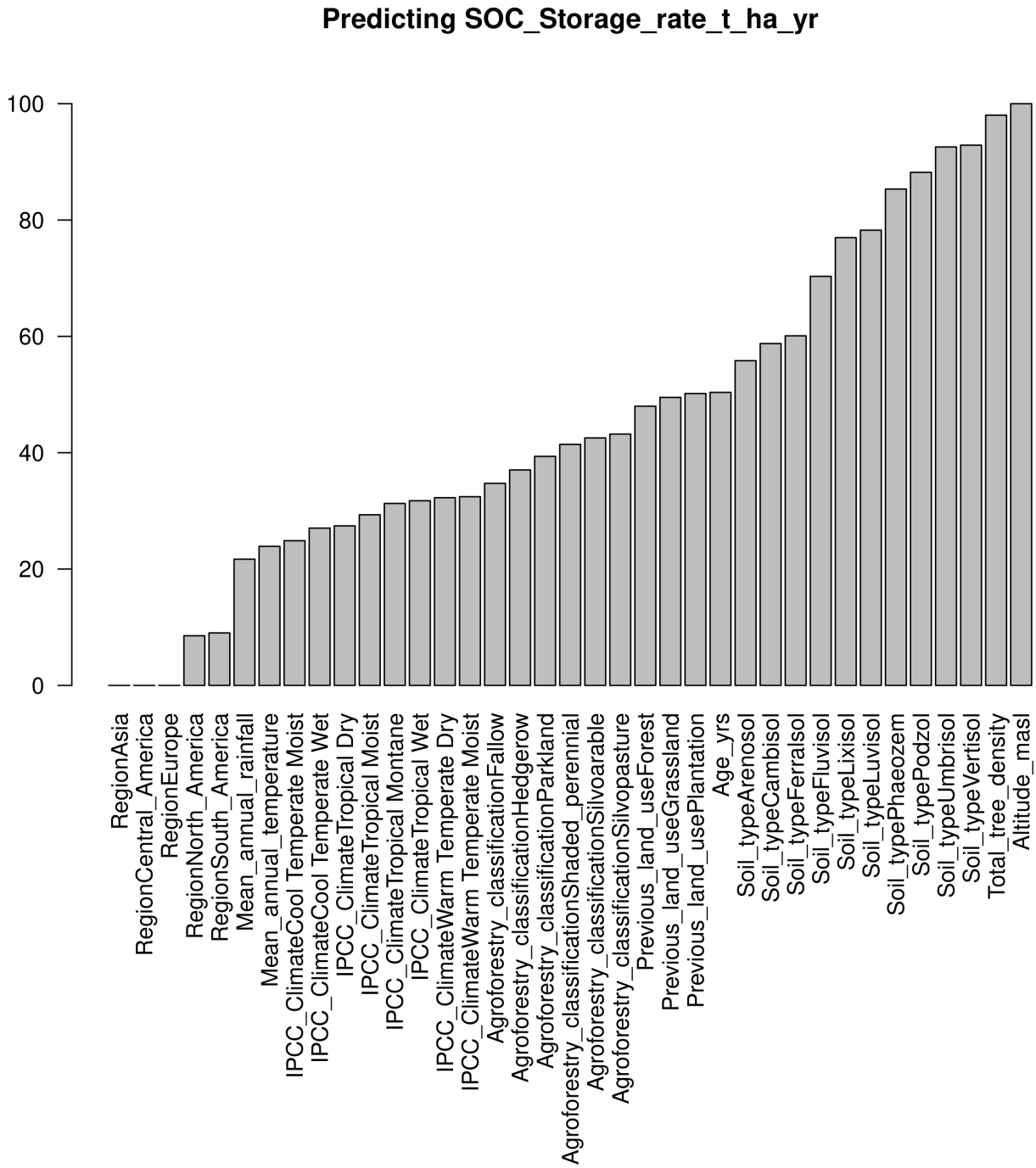


Figure 7: *Minimal model to predict C accumulation rates, decomposition of explained variance.*

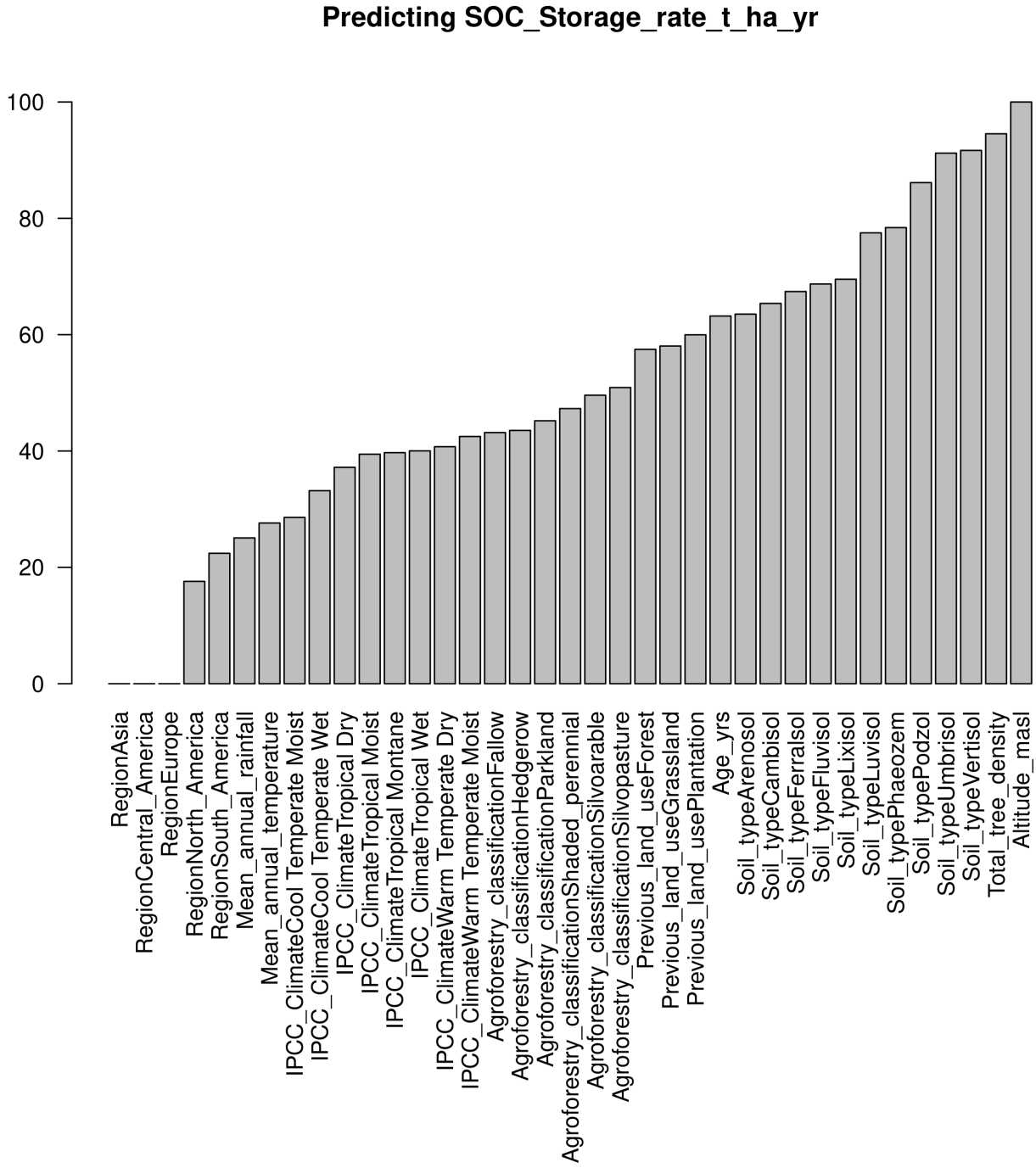
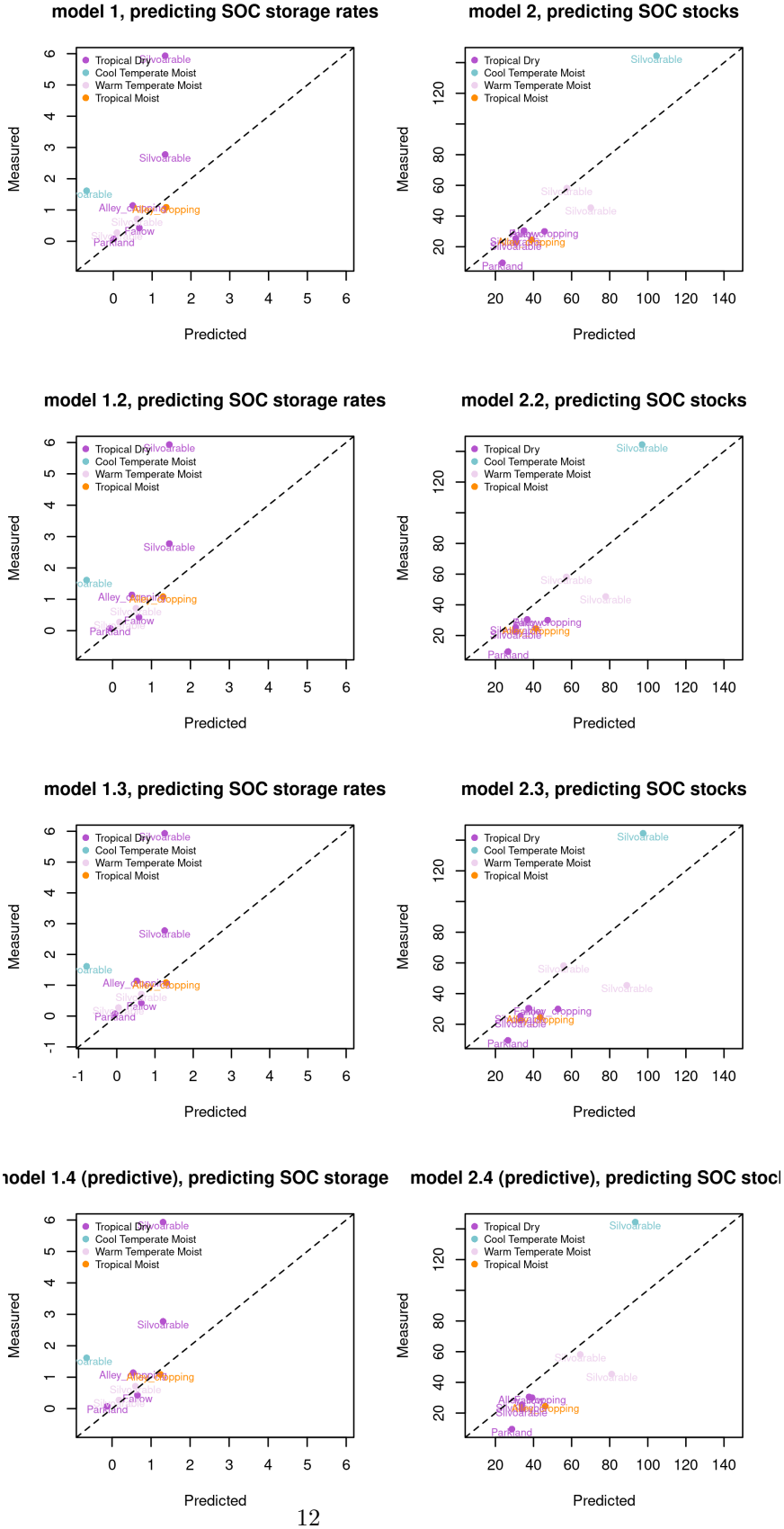


Figure 8: *Minimal model to predict C stocks not considering former C stocks, decomposition of explained variance.*

Predictions

I test the six developed model on a 10% of the sites, kept separated from the training dataset, for an independent validation:



The information in the data seems rather redundant since the RMSE of the models does not change much by reducing the selected parameters:

Table 1: The RMSE of the models considered here

	SOC sequestration	SOC stocks
Full	1.096688	33.07987
No previous stocks	1.077914	37.52626
No age	1.095410	37.56753
Minimal	1.099271	39.51941

```
getwd()
```

```
## [1] "/home/ilmenichetti/Documents/agroforestry"
```