Spatial prediction and assessment of Soil Organic Carbon

- *Edited by*: T. Hengl (ISRIC), B. Kempen (ISRIC), and J. Sanderman (Woods Hole Research Center)

Prepared as a supplementary material for the following research articles:

- Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilibarda M, Blagotić A, et al. (2017) **SoilGrids250m: Global gridded soil information based on machine learning**. PLoS ONE 12(2): e0169748. doi:10.1371/journal.pone.0169748
- Sanderman, J., Hengl, T., Fiske, G., (2017) **The soil carbon debt of 12,000 years of human land use**. PNAS, doi:10.1073/pnas.1706103114

This step-by-step tutorial explains how to map Soil Organic Carbon Stocks (OCS) using soil samples (point data). We demonstrate derivation of values both at site level (per profile) and by using raster calculus (per pixel). We also show how to estimate total OCS for an area of interest (which can be a field plot, farm and/or administrative region). The R script you can download from **github**. Instructions on how to install and setup all software used in this example you can find here. For an introduction to soil mapping using Machine Learning Algorithms refer to this tutorial. To download global soil organic carbon (content, density and stock) maps at 250 m resolution visit ftp.soilgrids.org/data/recent/. To access ISRIC's global compilation of soil profiles please refer to: http://www.isric.org/explore/wosis.

Measurement and derivation of soil organic carbon

Carbon below ground can be organic and non-organic or mineral (usually carbonates and bicarbonates) i.e. $CaCO_3$ in the rocks. Organic carbon stock below ground (0–2 m) in terrestrial ecosystems consists of two major components:

1. Living organism biomass i.e. mainly:
   - Plant roots,
   - Microbial biomass (Xu et al., 2012),
2. Plant and animal residues at various stages of decomposition (organic matter).

Xu et al. (2013) have estimated that the global microbial biomass is about 17 Pg C, which is only about 2% of the total organic matter, hence amount of C in microbial biomass can be neglected in comparison to the total stock, although if one would include all living organism and especially tree roots, then the portion of the C in the living organism could be more significant, especially in areas under dense forests.

Soil Organic Carbon Stock (**OCS**) is the mass of soil organic carbon per standard area and for a specific depth interval, usually expressed in kg/m^2 or t/ha. It can be derived using (laboratory and/or field) measurement of soil organic carbon content (ORC; expressed in % or g/kg of <2mm mineral earth), taking into account bulk density (BLD), thickness of the soil layer, and volume percentage of coarse fragments (CRF) (Nelson and Sommers, 1982; Poeplau et al. 2017):

$$\textbf{OCS [kg/m}^2\textbf{] = ORC [\%]/100} \times \textbf{BLD [kg/m}^3\textbf{]} \times \textbf{(1-CRF[\%]/100)} \times \textbf{HOT [m]}$$

Note that if one has soil organic carbon content measured in g/kg then one should divide by 1000 instead of 100. The correction for gravel content is necessary because only material less than 2 mm is analyzed for ORC concentration. Ignoring the gravel content migh result in an overestimation of the organic carbon stock. Note also that OCS always refers to a specific depth interval or horizon thickness (HOT), e.g.:

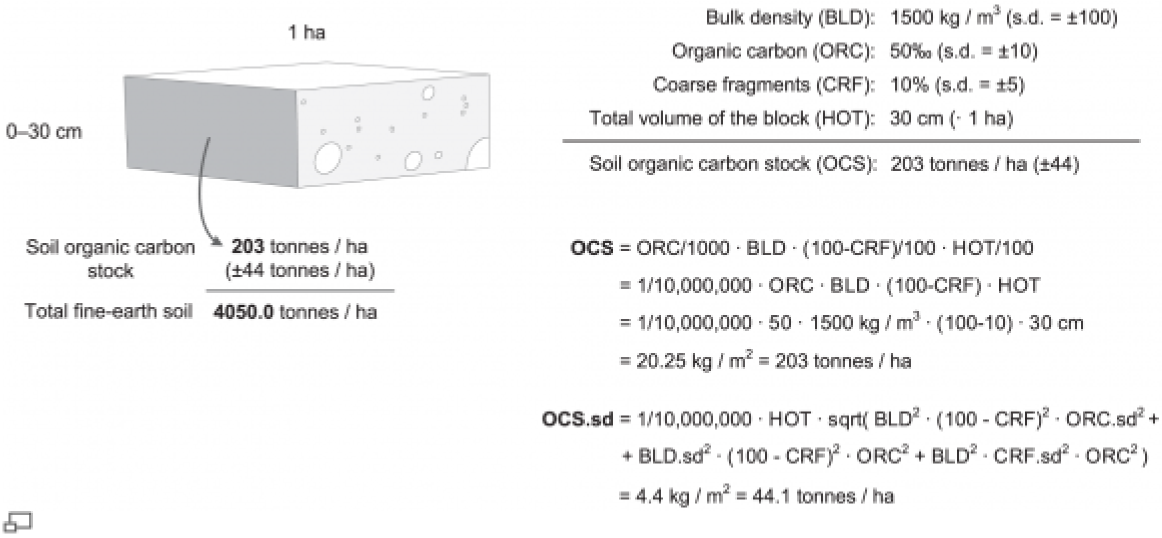
$$\text{kg/m}^2 \text{ for depth 0–30 cm (IPCC; Berhongaray and Alvarez, 2013),}$$

Values of OCS in kg/m^2 can also be expressed in tons/ha units, in which case simple conversion formula can be applied:

$$1 \times \text{kg/m}^2 = 10 \times \text{tons/ha}$$

Total OCS for an area of interest can be derived by multiplying OCS by total area e.g.:

$$120 \text{ tons/ha} \times 1 \text{ km}^2 = 120 \times 100 = 12{,}000 \text{ tons}$$

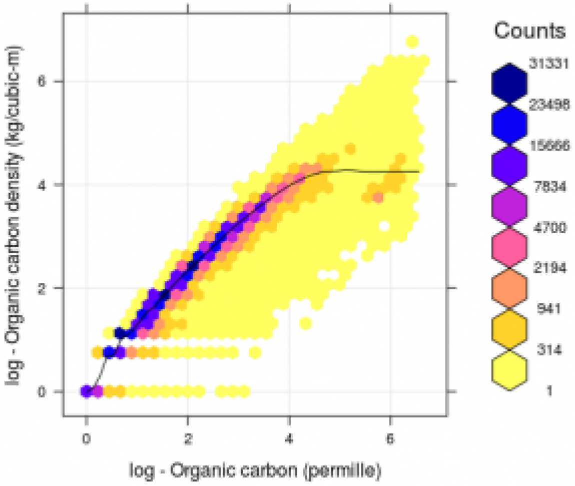

1 ha
0–30 cm
Soil organic carbon stock
203 tonnes / ha
(±44 tonnes / ha)
Total fine-earth soil
4050.0 tonnes / ha
Bulk density (BLD): 1500 kg / m^3 (s.d. = ±100)
Organic carbon (ORC): 50‰ (s.d. = ±10)
Coarse fragments (CRF): 10% (s.d. = ±5)
Total volume of the block (HOT): 30 cm (· 1 ha)
Soil organic carbon stock (OCS): 203 tonnes / ha (±44)
OCS = ORC/1000 · BLD · (100-CRF)/100 · HOT/100
= 1/10,000,000 · ORC · BLD · (100-CRF) · HOT
= 1/10,000,000 · 50 · 1500 kg / m^3 · (100-10) · 30 cm
= 20.25 kg / m^2 = 203 tonnes / ha
OCS.sd = 1/10,000,000 · HOT · sqrt(BLD^2 · (100 - CRF)^2 · ORC.sd^2 +
+ BLD.sd^2 · (100 - CRF)^2 · ORC^2 + BLD^2 · CRF.sd^2 · ORC^2)
= 4.4 kg / m^2 = 44.1 tonnes / ha

Example of OCS calculus and OCS propagated error for a given volume of soil using organic carbon content (ORC), bulk density (BLD), thickness of horizon (HOT), and percentage of coarse fragments (CRF). After Hengl et al. (2014). OCSKGM function available via the GSIF package.

Another way to express soil organic carbon is through **soil organic carbon density** (**OCD** in kg/m^3), which is in fact equivalent to OCS divided by the horizon thickness:

$$\mathbf{OCD\ [kg/m^3] = ORC\ [\%]/100 \times BLD\ [kg/m^3] \times (1\text{-}CRF[\%]/100) = OCS\ /\ HOT}$$

While OCS is a summary measure of SOC always associated with specific depth interval, OCD is a relative measure of soil organic carbon distribution and can be associated to any support size i.e. to arbitrary depth. In principle, OCD (kg/m^3) is strongly correlated with ORC (g/kg) as indicated in the figure below, however, depending on soil mineralogy and coarse fragment content, OCD can be lower or higher than what the smoothed line indicates (notice the range of values around the smoothed line is relatively wide). It is important to understand however, that, as long as ORC, BLD and CRF are known, one can convert the values from ORC to OCD and OCS and vice versa, without loosing any information about the soil organic carbon stock.



Correlation between soil organic carbon density and soil organic carbon content (displayed on a log-scale) created using a global compilations of soil profile data. Values 1, 2, 3, 4, 5 and 6 in the plot (log scale) correspond to values 2, 6, 19, 54, 147 and 402. Note that for ORC »12%, the OCD line flattens, which means that, organic carbon density practically stops to increase with the increase of ORC content.

In summary, there are four main variables to represent soil organic carbon:

1. **Soil Organic Carbon fraction or content** (ORC) in g/kg (permille) or dg/kg (percent),
2. **Soil Organic Carbon Density** (OCD) in kg/m^3,
3. **Soil Organic Carbon Stock** (OCS) in kg/m^2 or in tons/ha and for the given soil depth interval,
4. **Total Soil Organic Carbon Stock** (TOCS) in million tonnes or Pg i.e. OCS multiplied by surface area,

Global estimates of the total soil organic carbon stock are highly variable (Scharlemann et al. 2014): the current estimates of the current total soil organic carbon stock range between 800–2100 Pg C (for 0–100 cm), with the median estimate of about 1500 Pg C (for 0–100 cm). This means that the average OCS for 0–100 cm depth interval for the land mask (148,940,000 km^2) is about 11 kg/m^2 or 110 tons/ha, and that average soil organic carbon density (OCD) is about 11 kg/m^3 (compare to the standard bulk density of fine earth of 1250 kg/m^3); standard OCS for 0–30 cm depth interval is 7 kg/m^2 i.e. the average OCD is about 13 kg/m^3.

The distribution of soil organic carbon in the world is, however, highly patchy with large areas with OCS « 100 tons/ha, and then some 'pockets' of accumulated organic material i.e. organic soil types (histosols) with OCS up to 850 tons/ha (for 0–30 cm depth interval). The world's soil organic matter accumulation areas are usually the following biomes / land cover classes: wetlands and peatlands, mangroves, tundras and taigas.

Land use and agriculture in particular have led to dramatic decreases in soil carbon stocks in last 200+ years (agricultural and industrial revolutions). Lal (2004) estimated that approximately 54 Pg C have been added to the atmosphere due to agricultural activities with another 26 Pg C being lost from soils due to erosion. Wei et al. (2014) have estimated that, in average, conversion from forests to various agricultural land results to 30–50% decrease of SOCS. Modelling and monitoring of soil organic carbon dynamics is therefore of increasing importance (see e.g. FAO report "Unlocking the Potential of Soil Organic Carbon").

Derivation of OCS and OCD using soil profile data

As mentioned previously, OCS stock is most commonly derived from measurements of the organic carbon (ORC) content, soil bulk density (BLD) and the volume fraction of gravel (CRF). These are usually sampled either per soil layers or soil **horizons** (a sequence of horizons makes a soil profile), which can refer to variable soil depth intervals i.e. are non-standard. That means that, before one can determine OCS for standard fixed depth intervals (e.g. 0–30 cm or 0–100 cm), values of ORC, BLD and CRF need to be standardized so they refer to common depth intervals.

Consider, for example, the following two real life examples of soil profile data for a standard agricultural soil and an organic soil. In the first example, profile from Australia, the soil profile data shows:

upper limit (cm)	lower limit (cm)	organic carbon content (g / kg)	bulk density (kg / m-cubic)	CF (%)	SOCS (kg / m-square)
0	10	8.2	1340*	6*	1.1
10	20	7.5	1367*	6*	1
20	55	6.1	1382*	7*	3
55	90	3.3	1433*	8*	1.7
90	116	1.6	1465*	8*	0.6

Note that BLD variable was not available for described horizons (the original soil profile description / laboratory data indicates that no BLD has been observed for this profile), hence we can at least use the BLD estimated using SoilGrids250m data. It (unfortunately) commonly happens that soil profile observations miss BLD measurements, and hence BLD needs to be generated using a Pedo-Transfer function or extracted from soil maps.

To determine OCS for standard depth intervals 0-30, 0–100 and 0–200 cm, we first fit a mass-preserving spline:

```
> library(GSIF)
> library(aqp)
> library(plyr)
> lon = 149.73; lat = -30.09;
> id = "399_EDGEROI_ed079"; TIMESTRR = "1987-01-05"
> top = c(, 10, 20, 55, 90)
> bottom = c(10, 20, 55, 90, 116)
> ORC = c(8.2, 7.5, 6.1, 3.3, 1.6)
> BLD = c(1340, 1367, 1382, 1433, 1465)
> CRF = c(6, 6, 7, 8, 8)
> #OCS = OCSKGM(ORC, BLD, CRF, HSIZE=bottom-top)
> prof1 <- join(data.frame(id, top, bottom, ORC, BLD, CRF),
+               data.frame(id, lon, lat, TIMESTRR), type='inner')
```

```
Joining by: id
```

```
> depths(prof1) <- id ~ top + bottom
> site(prof1) <- ~ lon + lat + TIMESTRR
> coordinates(prof1) <- ~ lon + lat
> proj4string(prof1) <- CRS("+proj=longlat +datum=WGS84")
> ORC.s <- mpspline(prof1, var.name="ORC", d=t(c(,30,100,200)), vhigh =
2200)
```

```
Fitting mass preserving splines per profile...
|===========================================================================
==========| 100%
```

```
> BLD.s <- mpspline(prof1, var.name="BLD", d=t(c(,30,100,200)), vhigh =
2200)
```

```
Fitting mass preserving splines per profile...
|===========================================================================
==========| 100%
```

```
> CRF.s <- mpspline(prof1, var.name="CRF", d=t(c(,30,100,200)), vhigh =
2200)
```

```
Fitting mass preserving splines per profile...
|===========================================================================
==========| 100%
```

now we can derive OCS per each centimeter by using:

```
> OCSKGM(ORC.s$var.std$`-30 cm`, BLD.s$var.std$`-30 cm`, CRF.s$var.std$`-30
cm`, HSIZE=30)
```

```
[1] 2.875408
attr(,"measurementError")
[1] 3.84
attr(,"units")
[1] "kilograms per square-meter"
```

```
> OCSKGM(ORC.s$var.std$`30-100 cm`, BLD.s$var.std$`30-100 cm`,
CRF.s$var.std$`30-100 cm`, HSIZE=70)
```

```
[1] 3.616302
attr(,"measurementError")
[1] 9.18
attr(,"units")
[1] "kilograms per square-meter"
```

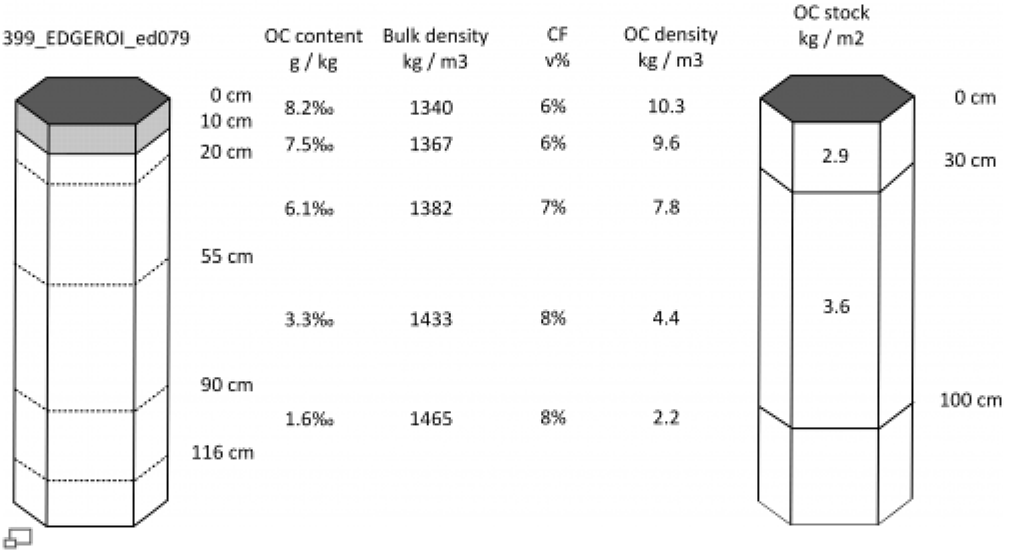
Note that the OCSKGM function requires soil organic carbon content in g/kg. If one has contents measured in % then first multiply the values by 10. Bulk density data should be provided in kg/m3, gravel content in %, and layer depth in cm. Running the OCSKGM function for the Edgeroi profile gives the following estimates of OCS for standard depth intervals:

- 0–30 cm: 2.9 kg / m-square
- 0–100 cm: 6.5 kg / m-square

- 0–200 cm: 8.5 kg / m-square (85 tonnes / ha)

Value of OCS between 5–35 kg / m-square for 0–100 cm are most common for a variety of mineral soils with e.g. 1–3% of soil organic carbon.

Note that the measurement error is computed from default uncertainty values (expressed in standard deviations) for organic carbon (10 g/kg), bulk density (100 kg/m3) and coarse fraction content (5%). When these are not provided by the user. The outcome should thus be interpreted with care.



Determination of soil organic carbon density and stock for standard depth intervals: example of a mineral soil profile from Australia.

In the second example we look at a profile from Canada (a histosol with >40% of organic carbon):

upper limit (cm)	lower limit (cm)	organic carbon content (g / kg)	bulk density (kg / m-cubic)	CF (%)	SOCS (kg / m-square)
0	31	472	185*	5*	25.7
31	61	492	172*	6*	23.9
61	91	487	175*	6*	24.1
91	122	502	166*	6*	24.3
122	130	59	830*	6*	3.7

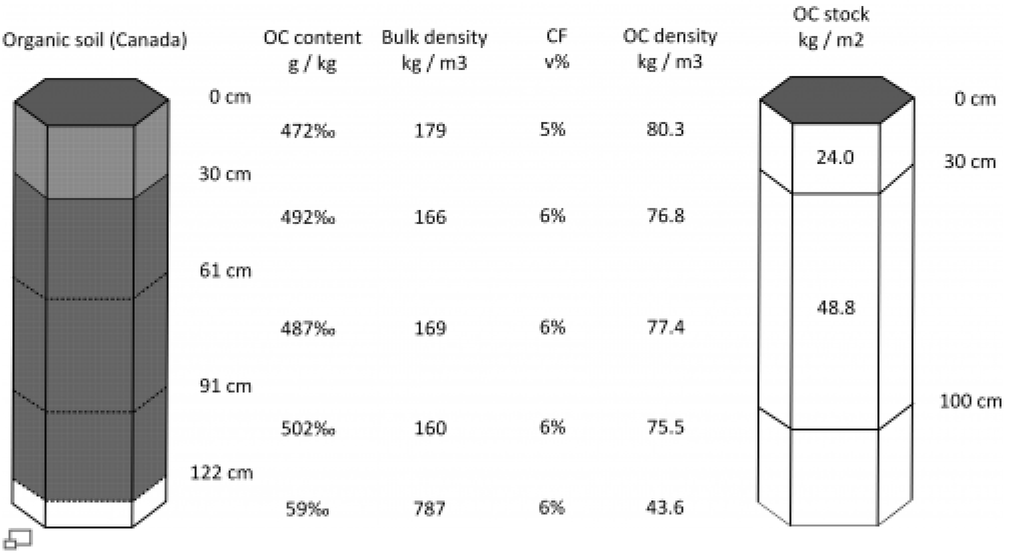
Here also BLD values were missing hence need to be estimated. For this we can use the simple Pedo-Transfer rule e.g. from Köchy et al. (2015):

$$BLD.f = (-0.31 * \log(ORC/10) + 1.38)*1000$$

We divide the organic carbon content here by 10 to convert the organic carbon content from g/kg to % that the PTF requires. Note that one might want to use different PTFs for different soil layers. For mineral soils the bulk density of subsoil layers often is somewhat higher than for topsoil layers. For organic soils this typically is the other way around. For instance, Köchy et al. (2015) propose the following PTF for the subsoil [for layers with SOC > 3%]: -0.32 * log(ORC[%]) + 1.38, which gives slightly lower bulk density values. Another useful source for PTFs for organic soils is work by Hossain et al. (2015). For illustrative purposes, we have here used only one PTF for all soil layers.

We can again fit mass-preserving splines and determine OCS for standard depth intervals by using the functions applied to the profile 1. This finally gives the following estimates:

- 0–30 cm: 24.8 kg / m-square
- 0–100 cm: 75.3 kg / m-square
- 0–200 cm: 114.5 kg / m-square (1145 tonnes / ha)



Determination of soil organic carbon density and stock for standard depth intervals: example of an organic soil profile from Canada.

Note that only 3–4% of the total soil profiles in the world have organic carbon content above 8% (soils with ORC >12% are often classified as organic soils or histosols in USDA and/or WRB classification and are even less frequent), hence soil-depth functions of organic carbon content and derivation of OCS for organic soils specific to patches of organic soils. On the other hand, organic soils carry much more total OCS. Precise processing and mapping of organic soils is often crucial for accurate estimation of total OCS for large areas, and hence it is fairly important to use a good PTF to fill in missing values for BLD for organic soils. As a rule of thumb, organic soil will rarely have density above some number e.g. 120 kg/m^3 because even though SOC content can be >50%, bulk density of such soil gets proportionally lower and bulk density is physically bound with how is material organized in soil (unless soils is artificially compacted). Also, getting the correct estimates of coarse fragments is important as otherwise (if CRF is ignored) total stock can be over-estimated >100% (Poeplau et al. 2017).

Estimation of Bulk Density using a globally-calibrated PTF

In the case bulk density is missing and no local PTF exists, WoSIS points (global compilation of soil profiles) can be used to fit a PTF that can fill-in the gaps in bulk density measurements globally. A regression matrix extracted on 15th of May 2017 (and which contains harmonized values for BD, organic carbon content, pH, sand and clay content, depth of horizon and USDA soil type at some 20,000 soil profiles world-wide), can be fitted using a random forest model (Ramcharan et al. 2017):

```
> dfs_tbl = readRDS("wosis_tbl.rds")
> ind.tax = readRDS("ov_taxousda.rds")
> library(ranger)
> fm.BLD = as.formula(paste("BLD ~ ORCDRC + CLYPPT + SNDPPT + PHIHOX +
DEPTH.f +", paste(names(ind.tax), collapse="+")))
```

```
> m.BLD_PTF <- ranger(fm.BLD, dfs_tbl, num.trees = 85,
importance='impurity')
> m.BLD_PTF
```

```
...
Type:                             Regression
Number of trees:                  85
Sample size:                      98650
Number of independent variables:  70
Mtry:                             8
Target node size:                 5
Variable importance mode:         impurity
OOB prediction error:             32782.78
R squared:                        0.5431644
```

This shows somewhat lower accuracy i.e. an RMSE of ±180 kg/m^3, but still probably better than dropping totally observations without bulk density from SOC assessment. A disadvantage of this model is that, in order to predict BD for new locations, we need to also have measurements of texture fractions, pH and organic carbon of course. For example, an Udalf (TAXOUSDA84) with 1.1% organic carbon, 22% clay, pH of 6.5, sand content of 35% and at depth of 5 cm would result in bulk density of:

```
> ind.tax.new = ind.tax[which(ind.tax$TAXOUSDA84==1)[1],]
> predict(m.BLD_PTF, cbind(data.frame(ORCDRC=11, CLYPPT=22, PHIHOX=6.5,
SNDPPT=35, DEPTH.f=5), ind.tax.new))$predictions
```

```
[1] 1532.635
```

Note also that the PTF from above needs USDA suborder values per point location following the SoilGrids legend, and formatted as in the `ind.tax` object. Unfortunately, the model from above probably over-estimates bulk density for organic soils as these are usually under-represented i.e. often not available (consider a Saprist with 32% organic carbon):

```
> ind.tax.new = ind.tax[which(ind.tax$TAXOUSDA13==1)[1],]
> predict(m.BLD_PTF, cbind(data.frame(ORCDRC=320, CLYPPT=8, PHIHOX=5.5,
SNDPPT=45, DEPTH.f=10), ind.tax.new))$predictions
```

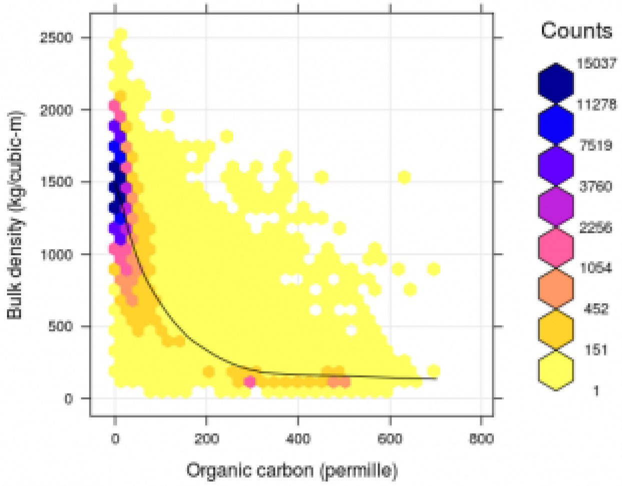
```
[1] 766.1135
```

An alternative to estimating BLD is to just use ORC values, e.g. (see plot below):

```
> m.BLD_ls = loess(BLD ~ ORCDRC, ovA, span=1/18)
> predict(m.BLD_ls, data.frame(ORCDRC=220))
```

```
       1
329.2059
```

This gives almost 2 times lower value than the random forest-based PTF from above. Over-estimating BLD would also result in two times higher OCS, hence clearly accurate information on BLD can be crucial for any OCS monitoring project. The PTF fitted using random forest above is likely over-estimating BLD values, mainly because there are not enough training points in organic soils that have both measurements of ORC, BLD, soil pH and texture fractions (if ANY of the calibration measurements are missing, the whole horizons are taken out of calibration and hence different ranges of BLD could be completely misrepresented).



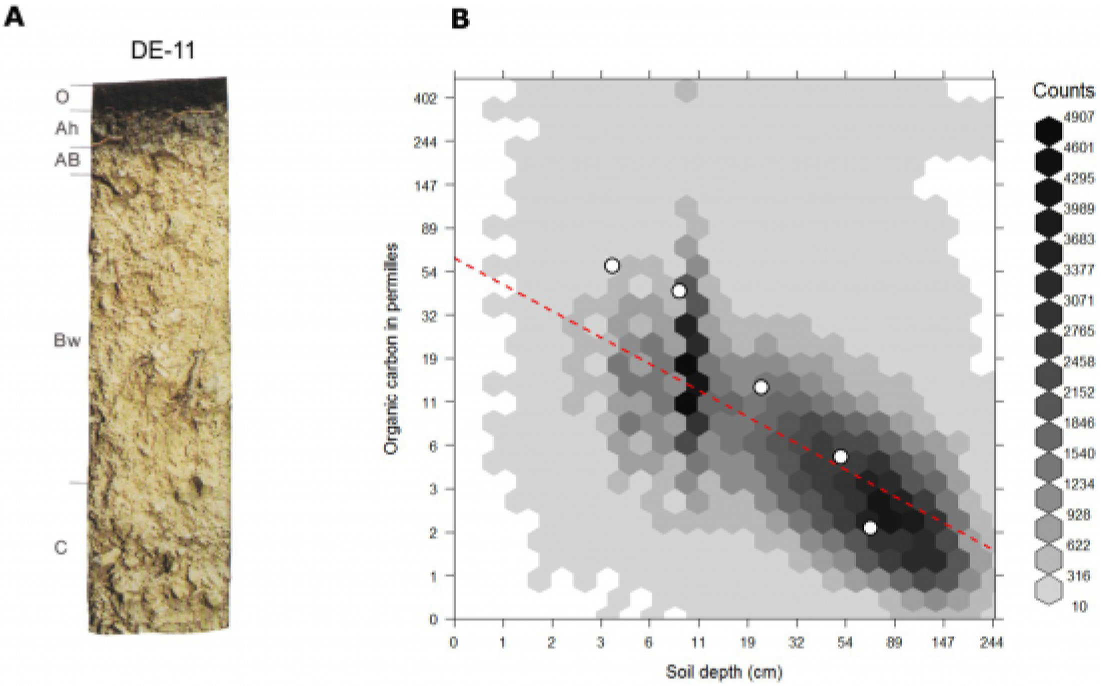
Correlation plot between soil organic carbon density and bulk density (fine earth), created using the global compilations of soil profile data. Black line indicates fitted loess polynomial surface. Note that there is still quite some scatter around the fitted line: many combinations of BLD and ORC, that do not fall close to the correlation line, can still be observed.

To fill-in missing values for BLD, SoilGrids project uses a combination of the two global Pedo-Transfer functions: (1) PTF fitted using random forest model that locally predicts BLD as a function of organic carbon content, clay and sand content, pH and coarse fragments, and (2) simpler model that predicts BLD just based on ORC. The average RMSE of these PTFs for BLD is about ±150 kg/m^3.

For mineral soils relationship between soil organic carbon and soil depth follows a log-log relationship which can be also approximated with the following (global) model (R-square: 0.36; see figure below):

$$\text{ORC (depth)} = \exp[\,4.1517 - 0.60934 * \log(\text{depth})\,]$$

This also illustrates that any organic carbon spatial prediction model can significantly profit from including depth into the statistical modelling.



Globally fitted regression model for predicting soil organic carbon using depth only (log-log regression) and (a) individual soil profile from the ISRIC soil monolith collection. After Hengl et al. (2014). Dots shown on the plot on the right are the actual ORC values for the profile on the left.

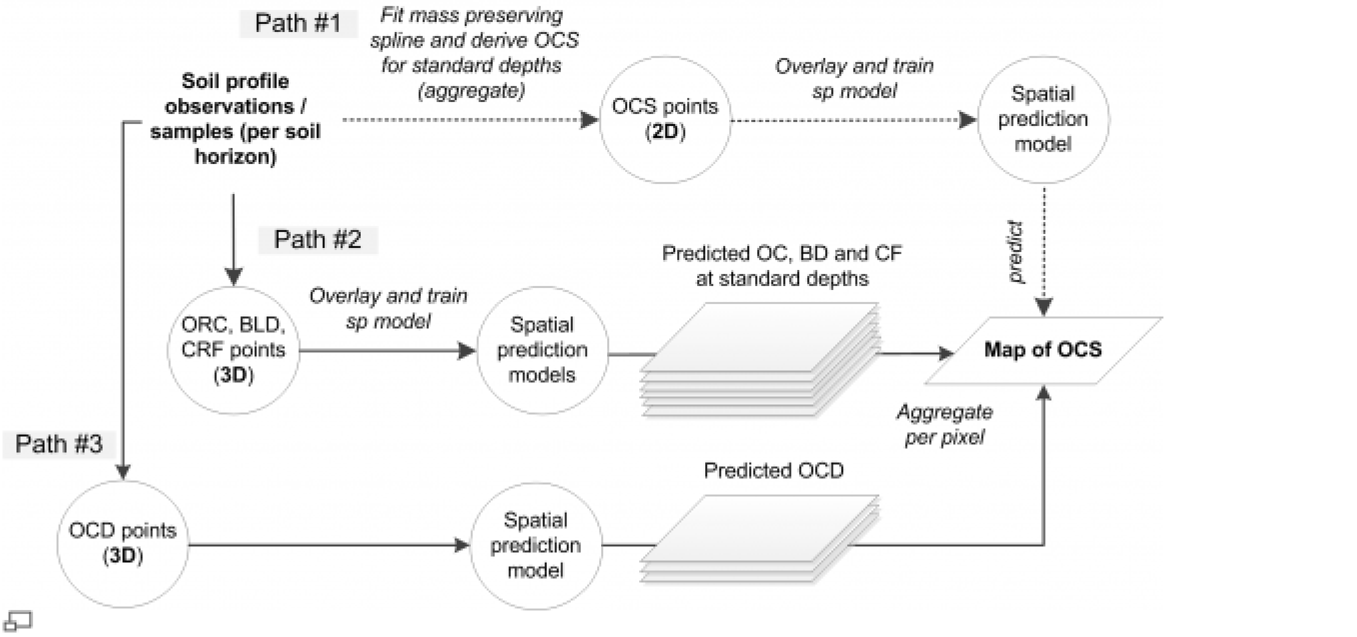
In summary, PTFs can be efficiently used to fill in gaps in BLD values (BLD is usually highly correlated with organic carbon content and depth, texture fractions, soil classification and soil pH can also help improve accuracy of the PTFs), however, for organic soils there is in general less calibration data and hence the errors are potentially higher. Mistakes in estimating BLD can result in systematic and significant over/under-estimations of the actual stock; on the other hand, removing all soil horizons from OCS assessment that do not have BLD measurements leads also to poorer accuracy as less points are included in training of the spatial prediction models. Especially for organic soils (>12% organic carbon), there is no easy solution for filling-in missing values for BLD and collecting additional (local) calibration points might unavoidable. Lobsey and Viscarra Rossel (2016) have recently proposed a method that combines gamma-ray attenuation and visible–near infrared (vis–NIR) spectroscopy to measure ex situ the bulk density using samples that are sampled freshly, wet and under field conditions. Hopefully BLD measurements (or their complete lack of) will be less and less problem in the future.

Generating maps of OCS

Most of projects focused on monitoring OCS require that an estimate of OCS is provided for the whole area of interest so that the user can also visually explore spatial patterns of OCS. In this tutorial we demonstrate how to generate maps of OCS using point samples and RS based covariates. The output of this process is usually a gridded map (`SpatialPixelsDataFrame`) covering the area of interest (plot, farm, administrative unit or similar). Once OCS is mapped, we can multiply OCS densities with area of each pixel and sum up all numbers we can compute the total OCS in total tonnes using the formula from above. Predicted OCS values can also be aggregated per land cover group or similar. If series of OCS maps are produced for the same area of interest (time-series of OCS), these can be used to derive OCS change per pixel.

In principle, there are three main approaches to estimating total OCS for an area of interest:

- By directly predicting OCS, here called the **the 2D approach to OCS mapping** (this often requires vertical aggregation / modeling of soil variable depth curves as indicated above),
- By predicting ORC, BLD and CRF, and then deriving OCS per layer, here called **the 3D approach to OCS mapping with ORC, BLD and CRF mapped separately**,
- By deriving OCD (organic carbon density) and then directly predicting OCD and converting it to OCS, here called **the 3D approach to OCS mapping via direct modeling of OCD**,


Path #1
Fit mass preserving spline and derive OCS for standard depths (aggregate)
Soil profile observations / samples (per soil horizon)
OCS points (2D)
Overlay and train sp model
Spatial prediction model
predict
Path #2
ORC, BLD, CRF points (3D)
Overlay and train sp model
Spatial prediction models
Predicted OC, BD and CF at standard depths
Map of OCS
Aggregate per pixel
Path #3
OCD points (3D)
Spatial prediction model
Predicted OCD

Three main computational paths (2D and 3D) to producing maps of organic carbon stock. In this tutorial we in general recommend using the path 3 for modeling organic carbon stock.

Although 2D prediction of OCS from point data seems to be more straightforward, many soil profiles contain measurements at non-standard depth intervals (varying support sizes also) and hence 2D modeling of OCS can often be a cumbersome. In most of situations where legacy soil profile data is used, 3D modeling of OCD is probably the most elegant solution to mapping OCS because:

- No vertical aggregation of values via spline fitting or similar is needed to standardize values per standard depths,
- No additional uncertainty is introduced (in the case of the 2D approach splines likely introduce some extra uncertainty in the model),
- Predictions of OCD/OCS can be generated for any depth interval using the same model (i.e. predictions are based on a single 3D model),

A disadvantage of doing 3D modeling of OCD is, however, that correlation with covariate layers could be less clear than if separate models are build for ORC, BLD and CRF: because OCD is a composite variable, it can often be difficult to distinguish whether the values are lower or higher due to differences in ORC, BLD or CRF. We leave it to the users to compare various approaches to OCS mapping and then select the method that achieves best accuracy and/or is most fit for use for their applications.

Selecting spatial prediction models for SOC

The purpose of spatial prediction is to (a) produce a map showing spatial distribution of the variable of interest for the area of interest, and (b) to do this in an unbiased way. A comprehensive path to evaluating spatial predictions is the caret approach (Kuhn and Johnson, 2013), which wraps up many of the standard processes such as model training and validation, method comparison and visualization. Consider for example the meuse data set, often used to demonstrate geostatistical modeling steps, that contains 155 measurements of organic matter % in topsoil. We can quickly compare performance of using GLM vs random forest vs no model for predicting organic matter (om):

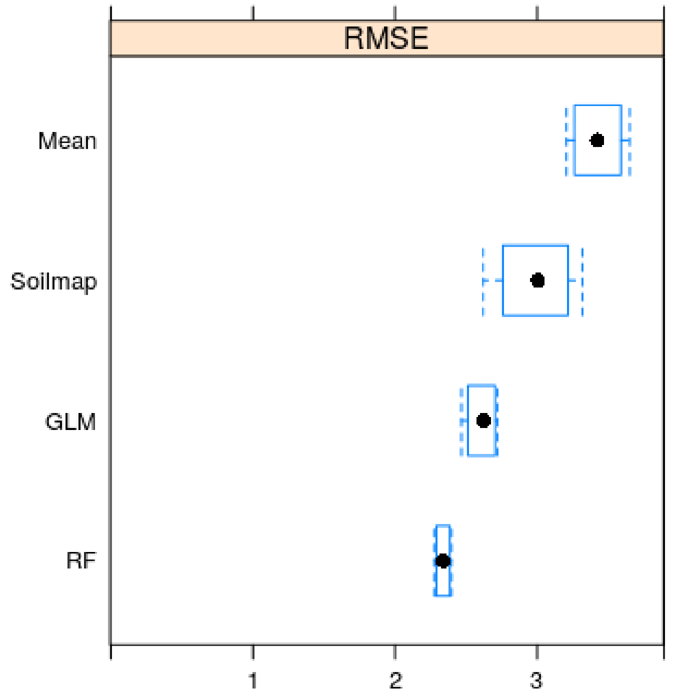
```
> library(caret); library(rgdal)
> demo(meuse, echo=FALSE)
> fitControl <- trainControl(method="repeatedcv", number=2, repeats=2)
> meuse.ov <- cbind(over(meuse, meuse.grid), meuse@data)
> mFit0 <- train(om~1, data=meuse.ov, method="glm",
family=gaussian(link=log), trControl=fitControl, na.action=na.omit)
```

```
Warning message:
In nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInf\o,  :
  There were missing values in resampled performance measures.
```

```
> mFit1 <- train(om~soil, data=meuse.ov, method="glm",
family=gaussian(link=log), trControl=fitControl, na.action=na.omit)
> mFit2 <- train(om~dist+soil+ffreq, data=meuse.ov, method="glm",
family=gaussian(link=log), trControl=fitControl, na.action=na.omit)
> mFit3 <- train(om~dist+soil+ffreq, data=meuse.ov, method="ranger",
trControl=fitControl, na.action=na.omit)
```

so that we can compare performance of the three models by using:

```
> resamps <- resamples(list(Mean=mFit0, Soilmap=mFit1, GLM=mFit2, RF=mFit3))
> bwplot(resamps, layout = c(3, 1))
```



Comparison of spatial prediction accuracy (RMSE at cross-validation points) for simple averaging (Mean), GLM with only soil map as covariate (Soilmap), GLM and random forest (RF) models with all possible covariates. Error bars indicate range of RMSE values for repeated CV.

In the case above, it seems that random forest (ranger package) helps decrease mean RMSE of predicting organic matter for about 32%:

```
> round((1-min(mFit3$results$RMSE)/min(mFit0$results$RMSE))*100)
```

```
[1] 32
```

In the case above, there is certainly added value in using spatial covariates (in the case above: distance to water and flooding frequency maps) and in using machine learning for spatial prediction, even with smaller data sets.

Note also that the assessment of spatial prediction accuracy for the three models based on the train function above is model-free, i.e. cross-validation of the models is independent from the models used because at each cross-validation subset fitting of the model is repeated and validation points are kept away from model training. Subsetting point samples is not always trivial however: in order to consider cross-validation as completely reliable, the samples ought to be representative of the study area and preferably collected using objective sampling such as simple random sampling or similar (Brus et al., 2011). In the case the sampling locations are clustered in geographical space i.e. if some parts of the study area are completely omitted from sampling, then also the results of cross-validation will reflect that sampling bias / poor representation. In all the following examples we will assume that cross-validation gives a reliable measure of mapping accuracy and we will use it as the basis of accuracy assessment i.e. mapping efficiency. In reality, cross-validation might be tricky to implement and could often lead to somewhat over-optimistic results if either sampling bias exists or/and if there are too little points for model validation. For example, in the case of soil profile data, it is highly recommended that whole profiles are taken out from CV because soil horizons are too strongly correlated (as discussed in detail in Gasch et al., 2015).

The whole process of spatial prediction of soil properties could be summarized in 5 steps:

1. Initial model comparison (comparison of prediction accuracy and computing time).
2. Selection of applicable model(s) and estimation of model parameters i.e. model fitting.
3. Predictions i.e. generation of maps for all areas of interest.
4. Objective accuracy assessment using independent (cross-)validation.
5. Export and sharing of maps and summary documentation explaining all processing steps.

Studying the **caret package tutorial** is highly recommended for anyone looking for a systematic introduction to predictive modelling.

Soil covariate layers for OCS mapping (30–100 m resolution)

As we have shown in the previous example, adding relevant covariates that can explain distribution of soil organic carbon increases accuracy of spatial predictions. Hence prior to generating predictions of OCS, it is a good idea to invest into preparing a list of Remote Sensing (RS), geomorphological/lithologic and DEM-based covariates that could potentially help explain spatial distribution of OCS. Since 2016, there are many high resolution (30–250 m) covariates with global coverage, and that are publicly available without restrictions. Both spatial detail, accessibility and accuracy of RS-based products has been growing exponentially and there is now evidence that that trend is going to slow down in the coming decades (Herold et al. 2016). The most relevant publicly available remote sensing-based covariates that can be downloaded and used to improve predictive soil mapping at high spatial resolutions are, for example:

- SRTM and/or ALOS W3D Digital Elevation Model (DEM) at 30 m and MERIT DEM at 100 m (these can be used to derive some 8–12 DEM derivatives from which some could be crucial for mapping of soil organic carbon);
- Landsat 7, 8 satellite images, either available from USGS's GloVis / EarthExplorer, or from the GlobalForestChange project repository (Hansen et al. 2013);
- Landsat-based Global Surface Water (GSW) dynamics images at 30 m resolution for period 1985–2016 (Pekel et al. 2016);
- Global Land Cover (GLC) maps based on the GLC30 project at 30 m resolution for 2000 and 2010 (Jun et al. 2014) and similar land cover projects (Herold et al. 2016);
- USGS's global bare surface images at 30 m resolution;

Note that the download time for 30 m global RS data could be significant if the data is needed for a larger area (hence you might consider using some RS data processing hub such as Sentinel hub, Google Earth Engine and/or Amazon Web Services instead of trying to download large mosaics yourself). Number of covariates used for generating SoilGrids can also be accessed from Geonode.isric.org.

Predicting OCS from point data (the 2D approach)

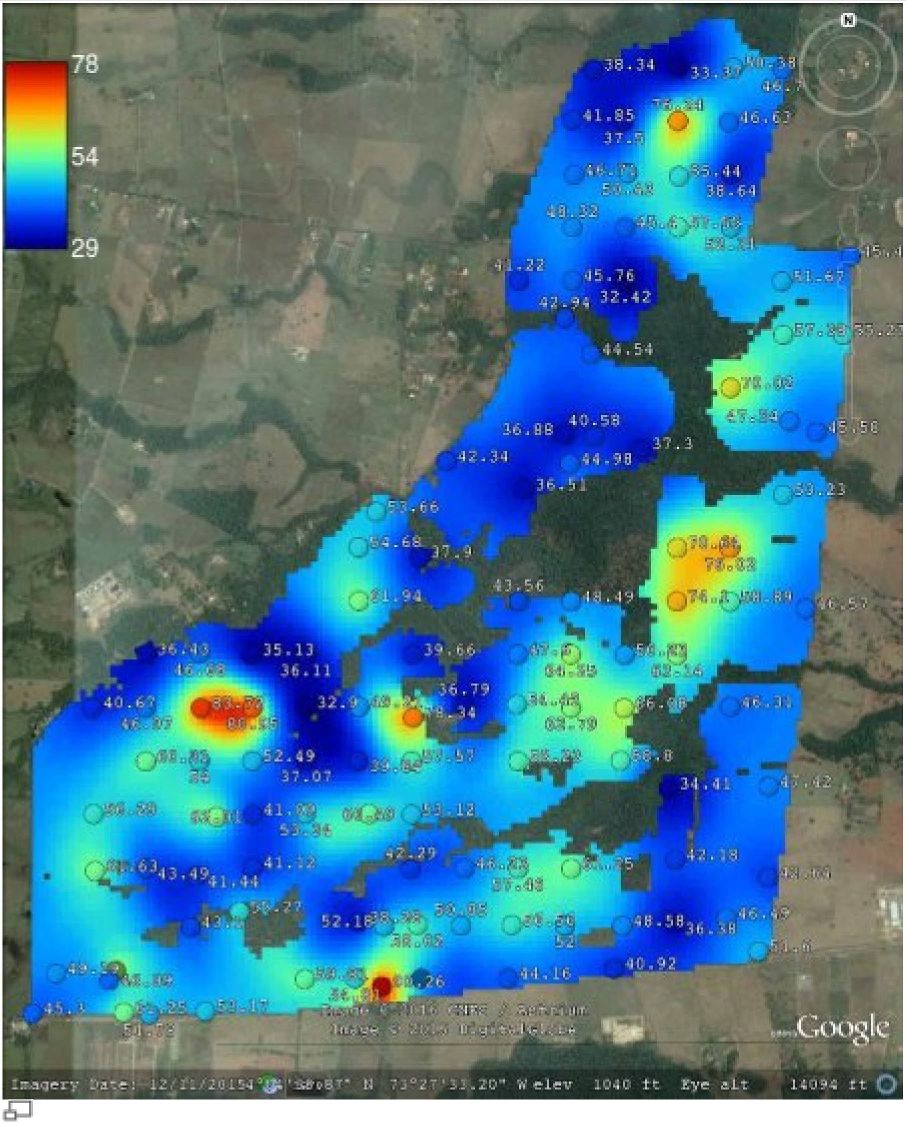
The geospt package contains 125 samples of OCS from Colombia already at standard depth intervals, hence this data set is ready for 2D mapping of OCS. The data sets consists of tabular values for points and a raster map containing the borders of the study area:

```
> load("COSha10.rda")
> load("COSha30.rda")
> str(COSha30)
```

```
'data.frame':    118 obs. of  10 variables:
 $ ID        : Factor w/ 118 levels "S1","S10","S100",..: 1 44 61 89 100 110
2 9 15 21 ...
 $ x         : int  669030 669330 670292 669709 671321 670881 670548 671340
671082 670862 ...
 $ y         : int  448722 448734 448697 448952 448700 448699 448700 448969
448966 448968 ...
 $ DA30      : num  1.65 1.6 1.5 1.32 1.41 1.39 1.51 1.39 1.55 1.63 ...
 $ CO30      : num  0.99 1.33 1.33 1.09 1.04 1.19 1.21 1.36 1.09 1.19 ...
 $ COB1r     : Factor w/ 6 levels "Az","Ci","Cpf",..: 5 5 2 5 2 5 2 2 2 5
...
 $ S_UDS     : Factor w/ 19 levels "BJa1","BQa1",..: 12 5 12 5 11 12 12 12
12 12 ...
 $ COSha30   : num  49.2 64 59.8 43.1 44.2 ...
 $ Cor4DAidep: num  43.3 56.3 54 37.9 39.9 ...
 $ CorT      : num  1.37 1.39 1.38 1.36 1.36 …
```

```
> load("COSha30map.rda")
> proj4string(COSha30map) = "+proj=utm +zone=18 +ellps=WGS84 +datum=WGS84
+units=m +no_defs"
```

COSha10 = 0–10 cm, COSha30 = 0–30 cm in tons / ha are values for OCS aggregated to standard soil depth intervals, so there is no need to do any spline fitting and/or vertical aggregation.



Example of a data set with OCS samples (for 2D prediction). Case study available via the geospt package (Colombia).

We can import a number of RS-based covariates to R by (these were derived from the global 30 m layers listed previously):

```
> covs30m = readRDS("covs30m.rds")
```

From the DEM layer, we can derive some 8–10 additional DEM derivatives (read more about how to prepare DEM derivatives using SAGA GIS) that could potentially help with mapping OCS.

We can also derive buffer distances from observations points:

```
> classes = cut(COSha30$COSha30, breaks=seq(, 100, length=10))
> covs30mdist = buffer.dist(COSha30["COSha30"], covs30m[1], classes)
```

and finally convert all these to Principal Components to help separate noise from the main signals:

```
> covs30m@data = cbind(covs30m@data, covs30mdist@data)
> fm.spc = as.formula(paste(" ~ ", paste(names(covs30m), collapse = "+")))
> fm.spc
```

```
~SRTMGL1_SRTMGL1.2_cprof + SRTMGL1_SRTMGL1.2_devmean +
SRTMGL1_SRTMGL1.2_openn +
    SRTMGL1_SRTMGL1.2_openp + SRTMGL1_SRTMGL1.2_slope +
SRTMGL1_SRTMGL1.2_twi +
    SRTMGL1_SRTMGL1.2_vbf + SRTMGL1_SRTMGL1.2_vdepth + SRTMGL1_SRTMGL1.2 +
    GlobalForestChange2000.2014_first_NIRL00 +
GlobalForestChange2000.2014_first_REDL00 +
    GlobalForestChange2000.2014_first_SW1L00 +
GlobalForestChange2000.2014_first_SW2L00 +
    GlobalForestChange2000.2014_treecover2000 + layer.1 + layer.2 +
    layer.3 + layer.4 + layer.5 + layer.6 + layer.7 + layer.8
```

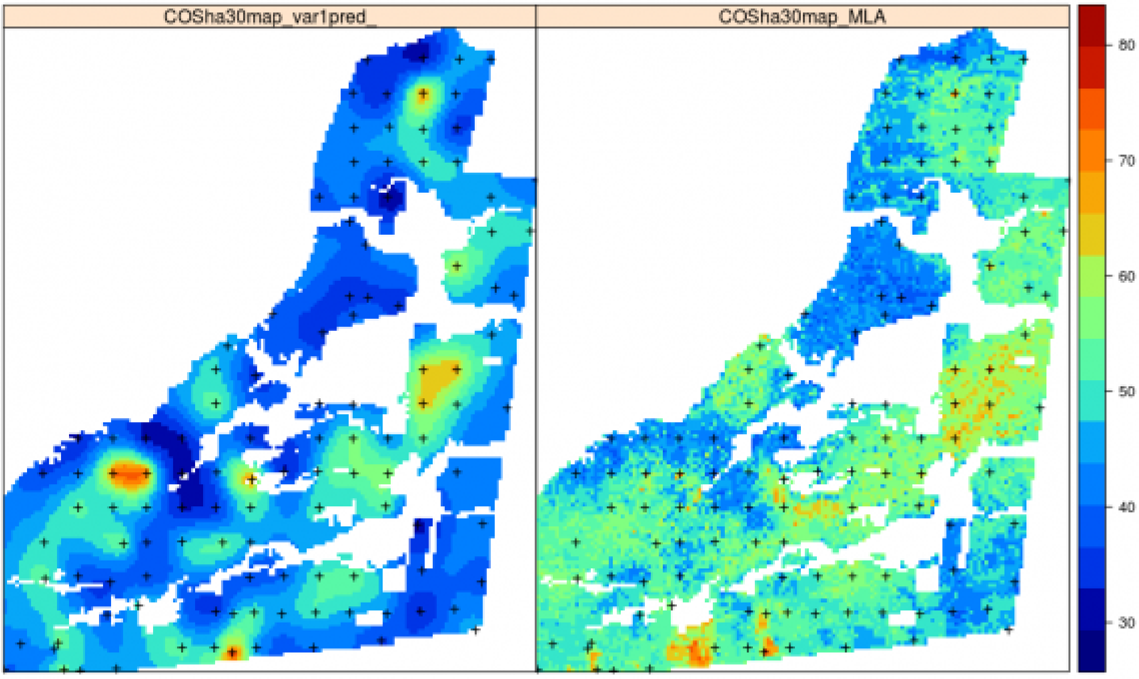
By using the above listed of covariates, we can fit a spatial prediction 2D model using some available models such as ranger, xgboost and gamboost. We model the target variable as a function of PCs:

```
> fm.COSha30
```

```
COSha30 ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 +
    PC10 + PC11 + PC12 + PC13 + PC14 + PC15 + PC16 + PC17 + PC18 +
    PC19 + PC20 + PC21
```

Comparison of random forest GLMboost and xgboost shows that none of the models are (unfortunately) distinct with the amount of variation that can be explained ranging between 10–20%. It is very common for soil mapping projects that the amount of variation that models explain are low and hence the average error of prediction and/or prediction intervals are wide. This could happen because the measurement errors were high, and/or because there are missing covariates, but it could also happen because natural complexity of soils in the area is simply high.

Note that our predictions of OCS are somewhat different from the predictions produced by the geospt package authors, although the main patterns are comparable.



Comparison of predictions generated using ordinary kriging (left) and machine learning with 30 m resolution covariates (right).

Deriving OCS from soil profile data (the 3D approach)

In the following example we will demonstrate, using a knonw data set (Edgeroi, from Australia) and which has been well documented in the literature (Malone et al, 2010), how to derive OCS in t/ha using soil profile data and 3D approach to spatial prediction based on mapping the Organic Carbon Density (OCD) in kg/m-cubic. The Edgeroi data set is a typical case of a soil profile data set that is relatively comprehensive, but still missing BLD measurements.

The Edgeroi data set can be loaded from the GSIF package:

```
> library(GSIF)
> data(edgeroi)
> edgeroi.sp = edgeroi$sites
> coordinates(edgeroi.sp) <- ~ LONGDA94 + LATGDA94
> proj4string(edgeroi.sp) <- CRS("+proj=longlat +ellps=GRS80
+towgs84=0,0,0,0,0,0,0 +no_defs")
> edgeroi.sp <- spTransform(edgeroi.sp, CRS("+init=epsg:28355"))
```

This data set comes with a list of covariate layers which can be used to explain distribution of soil organic carbon:

```
> con <-
url("http://gsif.isric.org/lib/exe/fetch.php?media=edgeroi.grids.rda")
```

```
> load(con)
> gridded(edgeroi.grids) <- ~x+y
> proj4string(edgeroi.grids) <- CRS("+init=epsg:28355")
```

Because some of the covariate layers are factors e.g. PMTGEO5 (parent material map) and because random forest requires numeric covariates, we can convert factors to numeric PCs by using:

```
> edgeroi.spc = spc(edgeroi.grids,
~DEMSRT5+TWISRT5+PMTGEO5+EV1MOD5+EV2MOD5+EV3MOD5)
```

```
Converting PMTGEO5 to indicators...
Converting covariates to principal components...
```

Edgeroi data set: locations of soil profiles and Australian soil classification codes.

Note that Edgeroi completely misses BLD values, hence before we can compute OCD values, we need to estimate BLD values for each corresponding horizon. Here the easiest option is probably to use the BLD values from the SoilGrids250m predictions (and which you can dowload from the SoilGrids FTP). Matching between the irregularly distributed soil horizons and SoilGrids BLD at standard depths can be implemented in three steps. First, we overlay the points and SoilGrids250m GeoTIFFs to get the BLD values at standard depths:

```
> ov.edgeroi.BLD =
raster::extract(stack(paste0("/mnt/cartman/ftp.soilgrids.org/data/recent/BLD
FIE_M_sl",1:7,"_250m_ll.tif")), spTransform(edgeroi.sp, CRS("+proj=longlat
+datum=WGS84")))
```

Second, we derive averaged estimates of BLD for standard depth intervals:

```
> ov.edgeroi.BLDm =
data.frame(BLD.f=as.vector(sapply(2:ncol(ov.edgeroi.BLD),
function(i){rowMeans(ov.edgeroi.BLD[,c(i-1,i)])})),
DEPTH.c=as.vector(sapply(1:6, function(i){rep(paste0("sd",i),
nrow(edgeroi$sites))})), SOURCEID=rep(edgeroi$sites$SOURCEID, 6))
```

```
> str(ov.edgeroi.BLDm)
```

```
'data.frame':    2154 obs. of  3 variables:
 $ BLD.f    : num  1270 1175 1296 1286 1278 ...
 $ DEPTH.c : Factor w/ 6 levels "sd1","sd2","sd3",..: 1 1 1 1 1 1 1 1 1 1
...
 $ SOURCEID: Factor w/ 359 levels "199_CAN_CP111_1",..: 1 2 3 4 5 6 7 8 9 10
...
```

Third, we match BLD values by matching horizon depths (center of horizon) with the standard depth intervals sd1 to sd6:

```
> edgeroi$horizons$DEPTH = edgeroi$horizons$UHDICM +
(edgeroi$horizons$LHDICM - edgeroi$horizons$UHDICM)/2
> edgeroi$horizons$DEPTH.c = cut(edgeroi$horizons$DEPTH,
include.lowest=TRUE, breaks=c(,5,15,30,60,100,1000),
labels=paste0("sd",1:6))
> summary(edgeroi$horizons$DEPTH.c)
```

```
sd1 sd2 sd3 sd4 sd5 sd6
100 314 356 408 391 769
```

```
> edgeroi$horizons$BLD.f =
plyr::join(edgeroi$horizons[,c("SOURCEID","DEPTH.c")],
ov.edgeroi.BLDm)$BLD.f
Joining by: SOURCEID, DEPTH.c
```

Now that we have an estimate of the bulk density for all horizons, we can derive OCD in kg/m-cubic by using:

```
> edgeroi$horizons$OCD = edgeroi$horizons$ORCDRC/1000 *
edgeroi$horizons$BLD.f
> summary(edgeroi$horizons$OCD)
<code rd>
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
  0.1402   2.5380   7.2950   9.5040  13.2100 110.0000      297
```

This shows that OCD values range from 0–110 kg/m-cubic, with an average of 9.5 kg/m-cubic (this corresponds to the average organic carbon content of about 0.8%).

For further 3D spatial prediction of OCD we use the ranger package, which fits a random forest model to this 3D data. We start by overlaying points and rasters so that we can create a regression matrix:

```
> ov2 <- over(edgeroi.sp, edgeroi.spc@predicted)
> ov2$SOURCEID = edgeroi.sp$SOURCEID
> h2 = hor2xyd(edgeroi$horizons)
> m2 <- plyr::join_all(dfs = list(edgeroi$sites, h2, ov2))
```

```
Joining by: SOURCEID
Joining by: SOURCEID
```

The spatial prediction model can be fitted using:

```
> fm.OCD = as.formula(paste0("OCD ~ DEPTH + ",
paste(names(edgeroi.spc@predicted), collapse = "+")))
> fm.OCD
```

```
OCD ~ DEPTH + PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 +
    PC9 + PC10 + PC11 + PC12
```

```
> m.OCD <- ranger(fm.OCD, m2[complete.cases(m2[,all.vars(fm.OCD)]),],
keep.inbag = TRUE, importance = "impurity")
> m.OCD
```

```
Ranger result

Call:
 ranger(fm.OCD, m2[complete.cases(m2[, all.vars(fm.OCD)]), ],
keep.inbag = TRUE\, importance = "impurity")

Type:                             Regression
Number of trees:                  500
Sample size:                      4858
Number of independent variables:  13
Mtry:                             3
Target node size:                 5
Variable importance mode:         impurity
OOB prediction error (MSE):       17.28167
R squared (OOB):                  0.7031539
```

Which shows that the average error with Out-of-bag training points is ±4.2 kg/m-cubic. Note that setting `keep.inbag = TRUE` allows us to derive also a map of the prediction errors, following the method of Wager et al. (2014) and which is shown in Figure below.

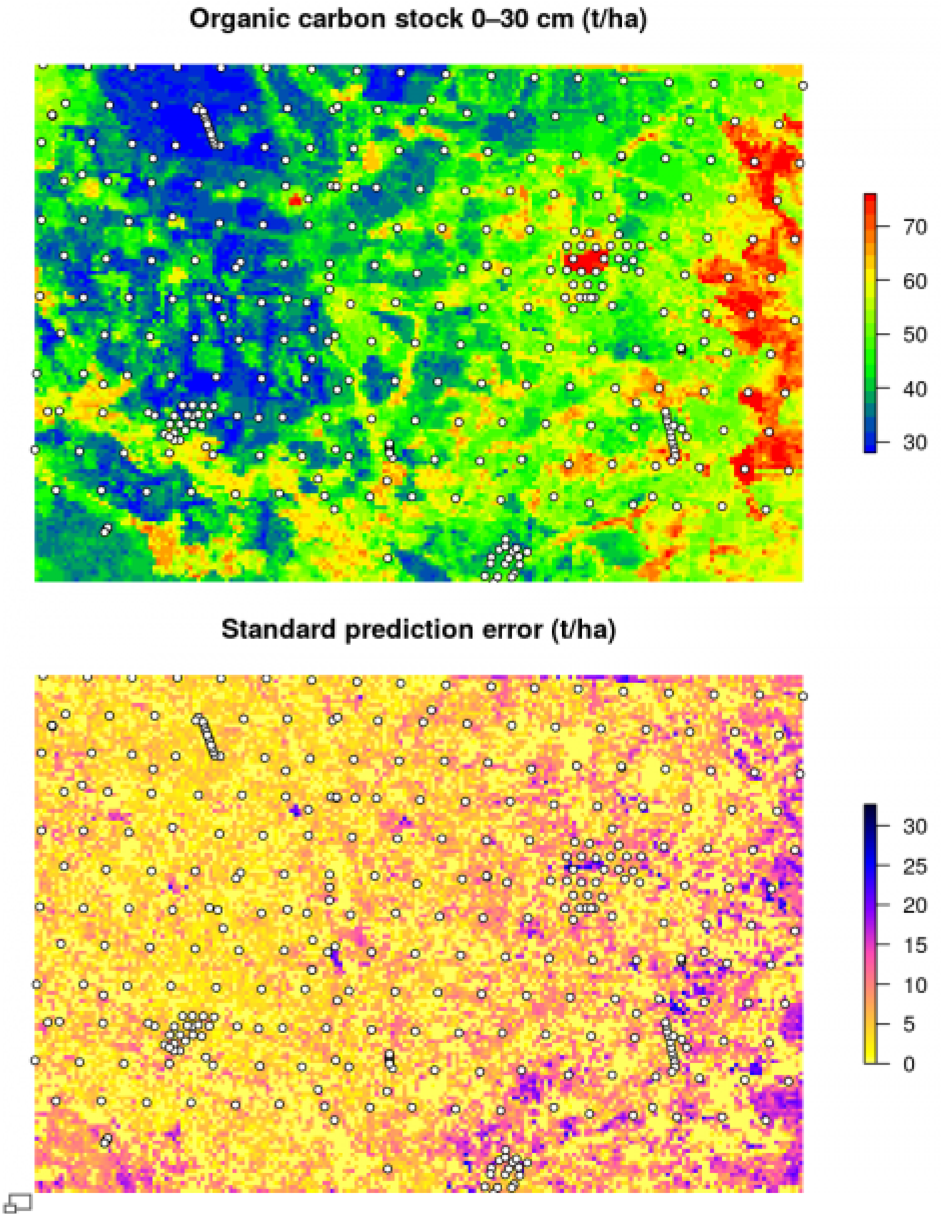
To derive OCS in tons/ha we can derive OCD at two depths (0 and 30 cm) and then take the mean value to produce a more representative value:

```
> for(i in c(,30)){
+   edgeroi.spc@predicted$DEPTH = i
+   OCD.rf <- predict(m.OCD, edgeroi.spc@predicted@data)
+   edgeroi.grids@data[,paste0("OCD.", i, "cm")] = OCD.rf$predictions
+ }
```

so that the final Organic carbon stocks in t/ha is (see formula above):

```
> edgeroi.grids$OCS.30cm = rowMeans(edgeroi.grids@data[,paste0("OCD.",
c(,30), "cm")]) * 0.3 * 10
```

Note that deriving the error map in the ranger package can be computationally intensive, especially if the number of covariates is high and is not yet recommended for large rasters.


Organic carbon stock 0–30 cm (t/ha)
70
60
50
40
30
Standard prediction error (t/ha)
30
25
20
15
10
5
0

Predicted organic carbon stock for 0–30 cm depth and error map for the Edgeroi data set. All values expressed in tons/ha.

Next, we can derive the total soil organic carbon stock per land use class (2007). For this we can use the aggregation function from the plyr package:

```
> edgeroi.grids$LandUse = readGDAL("edgeroi_LandUse.sdat")$band1
```

```
edgeroi_LandUse.sdat has GDAL driver SAGA
and has 128 rows and 190 columns
```

```
> lu.leg = read.csv("LandUse.csv")
> edgeroi.grids$LandUseClass =
paste(join(data.frame(LandUse=edgeroi.grids$LandUse), lu.leg,
match="first")$LU_NSWDeta)
```

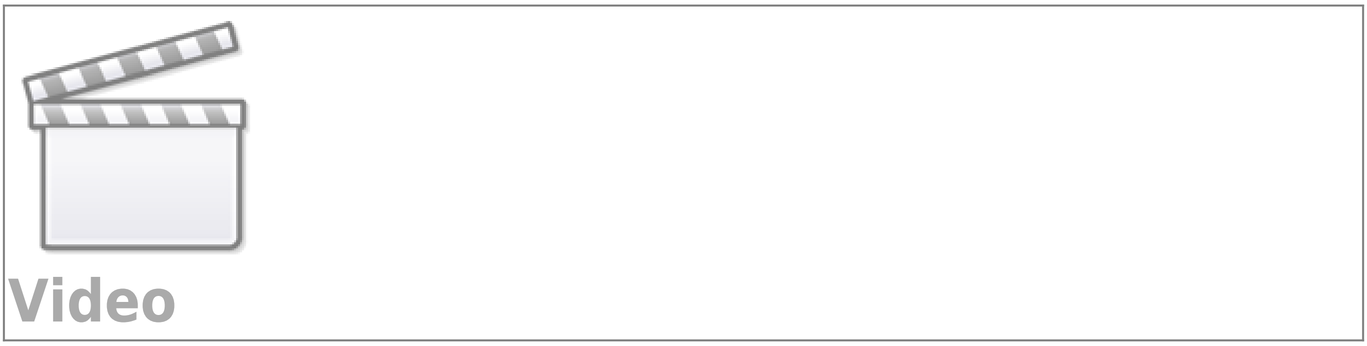
```
Joining by: LandUse
```

```
> OCS_agg.lu <- plyr::ddply(edgeroi.grids@data, .(LandUseClass), summarize,
Total_OCS_kt=round(sum(OCS.30cm*250^2/1e4, na.rm=TRUE)/1e3),
Area_km2=round(sum(!is.na(OCS.30cm))*250^2/1e6))
> OCS_agg.lu$LandUseClass.f = strtrim(OCS_agg.lu$LandUseClass, 34)
> OCS_agg.lu$OCH_t_ha_M =
round(OCS_agg.lu$Total_OCS_kt*1000/(OCS_agg.lu$Area_km2*100))
>
OCS_agg.lu[OCS_agg.lu$Area_km2>5,c("LandUseClass.f","Total_OCS_kt","Area_km2
","OCH_t_ha_M")]
```

```
                       LandUseClass.f Total_OCS_kt Area_km2 OCH_t_ha_M
2  Constructed grass waterway for wat           53       11         48
3                              Cotton           40        8         50
4                  Cotton - irrigated          800      203         39
5   Cropping - continuous or rotation         1724      402         43
6  Cropping - continuous or rotation           229       59         39
10                           Farm dam           52       10         52
11 Farm infrastructure - house, machi           87       18         48
12 Grazing - Residual strips (block o           48       10         48
13 Grazing of native vegetation. Graz          665      129         52
14 Grazing of native vegetation. Graz           66       13         51
16                     Irrigation dam           64       16         40
21                      Native forest          218       37         59
26                  Research facility           38        9         42
27 River, creek or other incised drai           67       11         61
28               Road or road reserve          113       23         49
29                       State forest          416       83         50
32 Volunteer, naturalised, native or          1341      238         56
33 Volunteer, naturalised, native or            62       16         39
34 Volunteer, naturalised, native or            74       14         53
35 Volunteer, naturalised, native or           457       99         46
37 Wide road reserve or TSR, with som          453       90         50
```

Which shows that, for the “Cropping - continuous or rotation”, which is the dominant land use class in the area, the average OCS is 43 tons/ha for 0–30 cm depth. In this case, the total soil organic carbon stock for the whole area (for all land use classes) is ca 7154 thousand tons of C. There seems not to be large differences in OCS between the natural vegetation and croplands.

Deriving OCS using spatiotemporal models

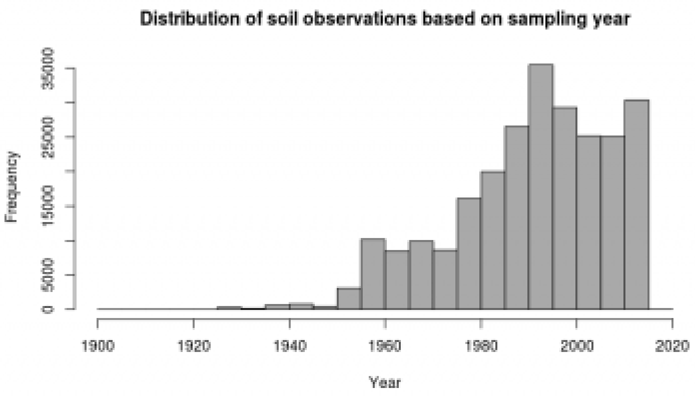


Assuming that measurements of ORC have been also temporally referenced (at least the year of sampling), points can be used to build spatiotemporal models of soil organic carbon. Consider for example the soil profile data available for conterminous USA:

```
> OCD_stN <- readRDS("usa48.OCD_spacetime_matrix.rds")
```

This data shows that there are actually enough observations spread in time (last 60+ years) to fit a spatiotemporal model:

```
> hist(OCD_stN$YEAR, main="Distribution of soil observations based on
sampling year", xlab="Year", col="darkgrey")
```



Histogram of sampling years for conterminous USA. Data source: the USA National Cooperative Soil Survey Soil Characterization database.

In fact, because the data set above represents values of OCD at variable depths, we can use this data to fit a full 3D+T spatiotemporal model in the form:

$$\mathbf{OCD(xydt) = d + X_1(xyt) + X_1(xyt) + \ldots + X_p(xyt)}$$

where d is the depth, $X_1 \ldots X_p$ are (static or dynamics) covariates, and xyt are spatiotemporal coordinates. Here we can assume that static covariates are mainly landform and lithology: these have probably not changed much in the last 100 years. Land cover, land use and climate, on the other hand, have probably changed drastically in the last 100 years and have to be represented with time-series of images. There are, indeed, several time-series data sets now available that can be used to represent land cover dynamics:

- HYDE 3.2 Historic land use data set (Klein et al. 2011): contains the distribution of main agricultural systems from 10,000 BC (pre-historic no landuse condition) to present time. 10 categories of land use have been represented: total cropping, total grazing, pasture (improved grazingland), rangeland (unimproved grazingland), total rainfed cropping, total irrigated cropping with further subdivisions for rice and non-rice cropping systems for both rainfed and irrigated cropping.
- CRU TS2.1 climatic surfaces for period 1960–1990 (Harris et al. 2014).
- UNEP-WCMC Generalised Original and Current Forest cover map.

All these are unfortunately available only at relatively coarse resolution of 10 km. Note also that, since these are time-series of images, spatiotemporal overlay can take time spatial overlay must be repeated for each time period. The spatiotemporal matrix file already contains results of overlay, so that we can focus directly on building spatiotemporal models of OCD e.g.:

```
> fm0.st <- as.formula(paste('OCDENS ~ DEPTH.f + ', paste(pr.lst,
collapse="+")))
> sel0.m = complete.cases(OCD_stN[,all.vars(fm0.st)])
> rf0.OCD_st <- ranger(fm0.st, data=OCD_stN[sel0.m,all.vars(fm0.st)],
importance="impurity", write.forest=TRUE, num.trees=120)
```

The model fitting result shows that model explains almost 60% of variation in OCD values:

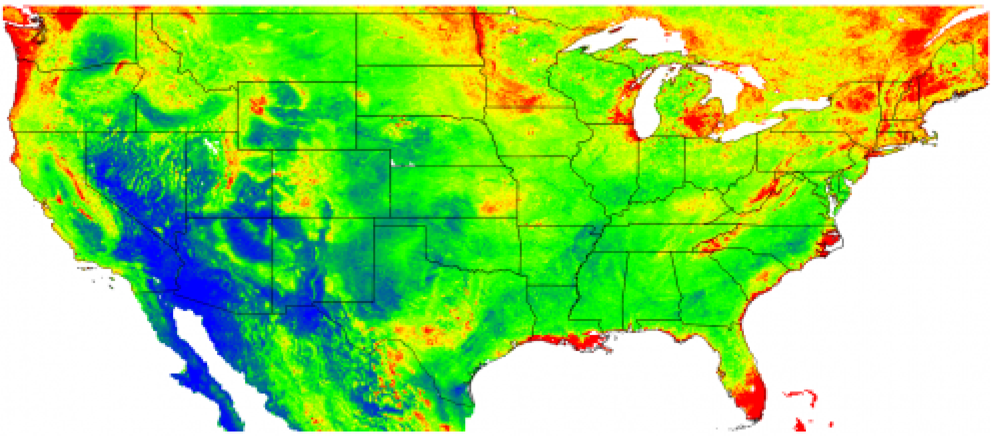
```
Type:                             Regression
Number of trees:                  120
Sample size:                      249025
Number of independent variables:  127
Mtry:                             11
Target node size:                 5
Variable importance mode:         impurity
OOB prediction error:             102.5703
R squared:                        0.5911552
```

the most important covariates being:

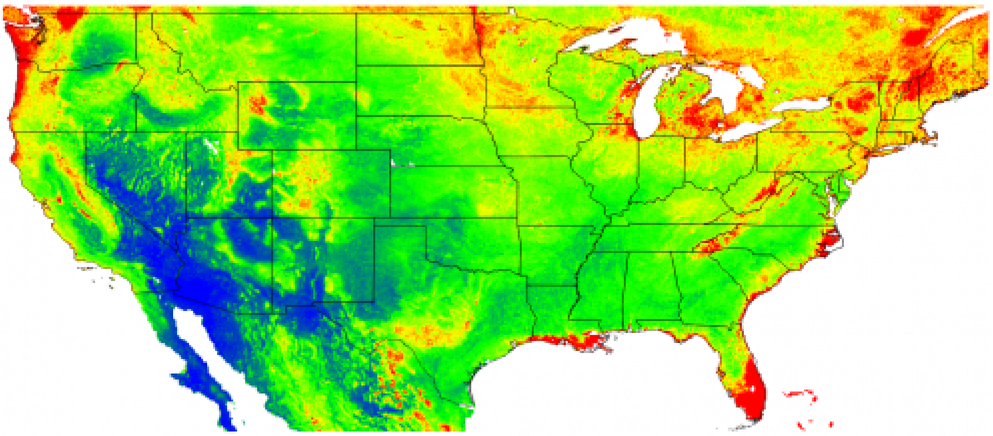
```
               [,1]
DEPTH.f  16242618.7
DEM        766114.8
GRAZING    716857.0
MODFC09    677867.9
VBF        663850.9
MODFC06    643100.1
MODFC03    619556.6
MODFC10    615712.3
CROPLAND   611457.7
MODFC07    610335.4
TWI        603085.5
MODFC08    599495.6
TRAINFED   590347.6
RFNORICE   584568.0
MODFC05    576126.9
```

which shows that the far the most important soil covariate is soil depth, followed by elevation, grazing, MODIS cloud fraction images, cropland and similar. For full description of codes please refer to this table.

Finally, based on this model, we can generate predictions for 3–4 specific time periods and for some arbitrary depth e.g. 10 cm. The maps below clearly show that ca 8% of the soil organic carbon has been lost in the last 90 years, most likely due to the increase of grazing and croplands. The maps also show, however, that some areas in the northern latitudes are experiencing an increase in SOC possibly due to higher rainfall i.e. based on the CRU data set.



Predicted OCD (in kg/m^3) at 10 cm depth for the year 2014. Blue colors indicate low values, red high values. Download map.



Predicted OCD (in kg/m^3) at 10 cm depth for the year 1925.Download map.

This demonstrates that, as long as there is enough training data spread through time, and as long as covariates are available for corresponding time range, machine learning can also be used to fit full 3D+T spatiotemporal prediction models (Gasch et al. 2015). Once we produce a time-series of images for some target soil variable of interest, the next step would be to implement time-series analysis methods to e.g. detect temporal trends and areas of highest soil degradation. An R package that is fairly useful for such analysis is the **greenbrown** package, primarily used to map and quantify degradation of land cover (Forkel et al. 2015).

We can focus on the time-series of predicted organic carbon density for Texas:

```
> library(greenbrown)
```

```
Loading required package: strucchange
Loading required package: zoo
```

```
Attaching package: ‘zoo’

The following objects are masked from ‘package:base’:

    as.Date, as.Date.numeric

Loading required package: sandwich
```

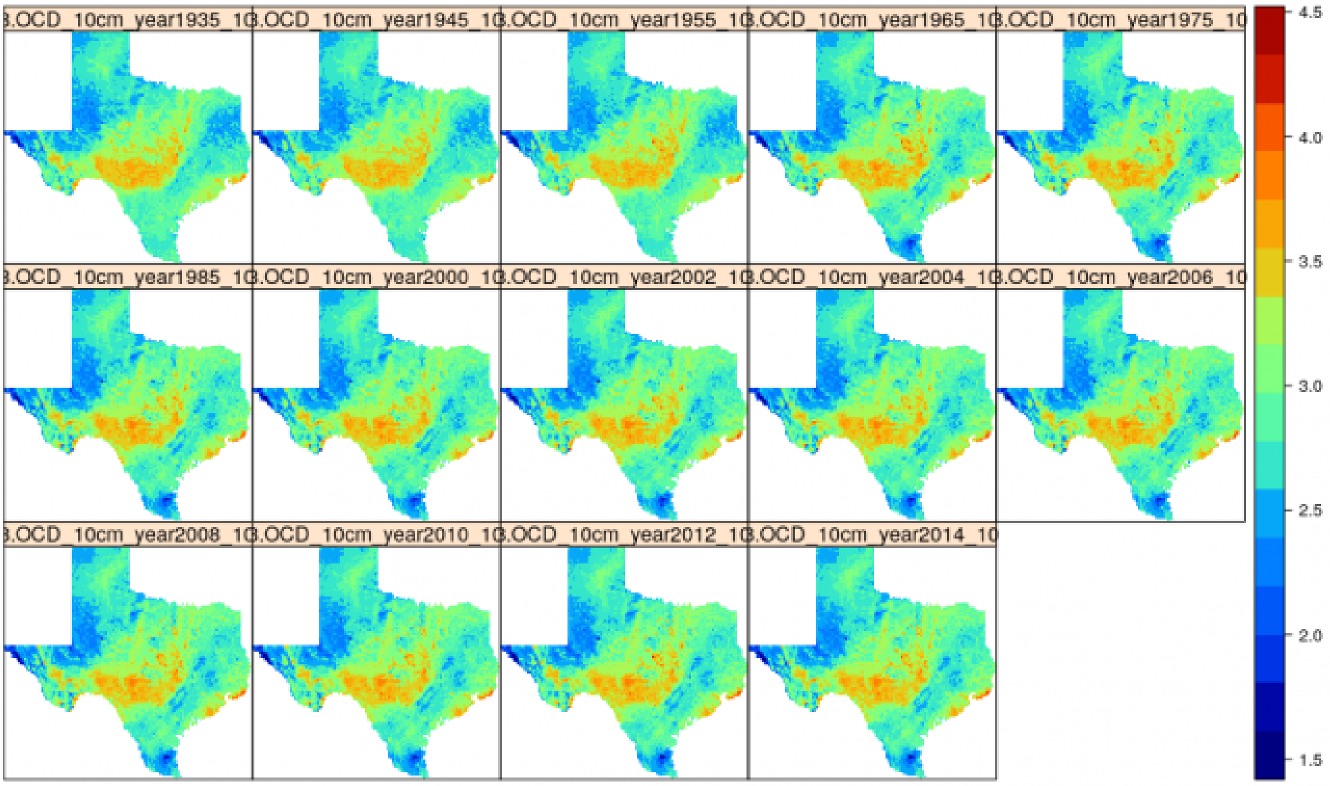
```
> library(raster)
> setwd("./USA48")
> tif.lst <- list.files(pattern="_10km.tif")
> g10km <- as(readGDAL(tif.lst[1]), "SpatialPixelsDataFrame")
```

```
usa48.OCD_10cm_year1935_10km.tif has GDAL driver GTiff
and has 2160 rows and 4320 columns
```

```
> for(i in 2:length(tif.lst)){ g10km@data[,i] = readGDAL(tif.lst[i],
silent=TRUE)$band1[g10km@grid.index] }
> names(g10km) = basename(tif.lst)
> g10km = as.data.frame(g10km)
> gridded(g10km) = ~x+y
> proj4string(g10km) = "+proj=longlat +datum=WGS84"
> library(maps)
> library(maptools)
```

```
Checking rgeos availability: TRUE\
```

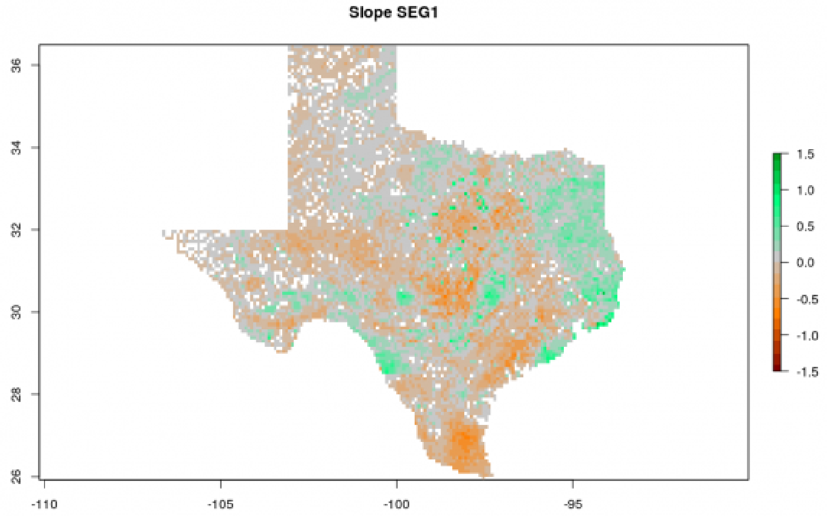
```
> states <- map('state', plot=FALSE, fill=TRUE)
> states = SpatialPolygonsDataFrame(map2SpatialPolygons(states,
IDs=1:length(states$names)), data.frame(names=states$names))
> proj4string(states) = "+proj=longlat +datum=WGS84"
> ov.g10km = over(y=states, x=g10km)
> txg10km = g10km[which(ov.g10km$names=="texas"),]
> txg10km = as.data.frame(txg10km)
> gridded(txg10km) = ~x+y
> proj4string(txg10km) = "+proj=longlat +datum=WGS84"
> spplot(log1p(stack(txg10km)), col.regions=SAGA_pal[[1]])
> g10km.b = raster::brick(txg10km)
```



Time-series of predictions of organic carbon density for Texas.

We can analyze this time-series data to see where is the decrease of organic carbon most significant, for example the slope of the change:

```
> trendmap <- TrendRaster(g10km.b, start=c(1935, 1), freq=1, breaks=1) ## can be computationally intensive
> plot(trendmap[["SlopeSEG1"]], col=rev(SAGA_pal[["SG_COLORS_GREEN_GREY_RED"]]), zlim=c(-1.5,1.5), main="Slope SEG1")
```



Predicted slope of change of soil organic carbon density for Texas for period 1935–2014. Negative values indicate loss of soil organic carbon.

which shows that loss of soil organic carbon is distinct especially in the southern part of Texas. The slope coefficient map is in average negative, which indicates that most of the state has lose organic carbon for the period of interest. Note that running such time-series analysis is not trivial as enough of observations in time (if possible: repetitions) are needed to be able to extract significant patterns. Also `TrendRaster` function can be quite computationally intensive, hence some careful planning of the processing steps / processing infrastructure is usually a good idea.

Summary points

Based on all the examples and discussion from above, the following key points can be emphasized:

1. OCS for an area of interest can be derived either using 2D or 3D approach. 3D approach typically includes modeling ORC, BLD and CRF separately (and then deriving OCS per pixel), or modeling OCD for standard depths and then converting to OCS.
2. Publicly available RS-based covariates (SRTM / ALOS DEM, Landsat, Sentinel satellites) are available for improving the mapping accuracy of OCS. Improving the accuracy of OCS maps is inexpensive given the increasing availability of RS data.
3. PT (Pedo-Transfer) rules can be used to fill in (impute) missing BLD values and to estimate ORC for deeper soil depths. Also SoilGrids250m global maps with predictions of BLD and CRF can be used to fill in the missing values.
4. Machine learning techniques such as Random Forest, neural nets, gradient boosting and similar, can be used to predict soil organic carbon in 2D, 3D and in spatiotemporal modeling frameworks. Accuracy of these predictions improves (in comparison to linear statistical models) especially where the relationship between soil organic carbon distribution and climatic, land cover, hydrological, relief and similar covariates is complex (i.e. non-linear).
5. Global estimates of ORC, BLD and CRF (SoilGrids.org) can be used as covariates so that consistent predictions can be produced (read more in: Ramcharan et al 2017).
6. By producing spatial predictions of OCS for specific time periods, one can derive estimates of OCS change (loss or gain).
7. Most of statistical / analytical tools required for running spatial analysis, time series analysis, export and visualization of soil carbon data is available in R, especially thanks to the contributed packages: aqp, caret, ranger, xgboost, GSIF, greenbrown and similar.

References

- Berhongaray, G. and Alvarez, R., (2013) The IPCC Tool for predicting soil organic carbon changes evaluated for the Pampas, Argentina. Agriculture, Ecosystems & Environment, 181, 241–245.
- Brus, D. J., Kempen, B., & Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. European Journal of Soil Science, 62(3), 394-407.
- Chen Jun, Yifan Ban, Songnian Li, (2014) China: Open access to Earth land-cover map, Nature, 514:434.
- Forkel, M., Migliavacca, M., Thonicke, K., Reichstein, M., Schaphoff, S., Weber, U., Carvalhais, N., (2015) Codominant water control on global interannual variability and trends in land surface phenology and greenness. Glob Change Biol 21, 3414–3435. doi:10.1111/gcb.12950
- Gasch, C. K., Hengl, T., Gräler, B., Meyer, H., Magney, T. S., & Brown, D. J. (2015). Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D+ T: The Cook Agronomy Farm data set. Spatial Statistics, 14, 70-90.

- Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., … & Kommareddy, A. (2013). High-resolution global maps of 21st-century forest cover change. Science, 342(6160), 850-853.
- Harris, I. P. D. J., Jones, P. D., Osborn, T. J., & Lister, D. H. (2014). Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset. International Journal of Climatology, 34(3), 623-642.
- Herold, M., See, L., Tsendbazar, N., & Fritz, S. (2016). Towards an Integrated Global Land Cover Monitoring and Mapping System. Remote Sensing, 8(12).
- Hossain, M.F., Chen, W., & Zhang, Y. (2015). Bulk density of mineral and organic soils in the Canada's arctic and subarctic. Information processing in Agriculture, 2, 183-190.
- Kempen, B., Brus, D. J., & Stoorvogel, J. J. (2011). Three-dimensional mapping of soil organic matter content using soil type–specific depth functions. Geoderma, 162(1), 107-123.
- Klein Goldewijk, K., Beusen, A., Van Drecht, G., & De Vos, M. (2011). The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years. Global Ecology and Biogeography, 20(1), 73-86.
- Köchy, M., Hiederer, R., & Freibauer, A. (2015). Global distribution of soil organic carbon–Part 1: Masses and frequency distributions of SOC stocks for the tropics, permafrost regions, wetlands, and the world. Soil, 1(1), 351-365.
- Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1-26.
- Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.
- Lal R (2004) Soil carbon sequestration impacts on global climate change and food security. Science, 304(5677), 1623-1627.
- Lobsey, C. R., & Viscarra Rossel, R. A. (2016). Sensing of soil bulk density for more accurate carbon accounting. European Journal of Soil Science, 67(4), 504–513.
- Malone, B.P., McBratney, A.B., Minasny, B. (2010). Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154, 138-152.
- Nelson, D.W., Sommers, L. (1982) Total carbon, organic carbon, and organic matter. In: Page, A., Miller, R., Keeney, D., editors, Methods of soil analysis, Part 2, Madison, WI: ASA and SSSA, Agron. Monogr. 9. 2nd edition, pp. 539–579.
- Ramcharan, A., Wills, S., Hengl, T., Beaudette, D. (2017) A Soil Bulk Density Pedotransfer Function Based on Machine Learning: A Case Study with the NCSS Soil Characterization Database. Soil Science Society of America Journal, doi: 10.2136/sssaj2016.12.0421 (download data)
- Ramcharan, A., Hengl, T., Nauman, T., Waltman, S., Brungard, C., Wills, S., Thomson, J. (2017) Soil Property and Class Maps of the Conterminous US at 100 meter Spatial Resolution based on a Compilation of National Soil Point Observations and Machine Learning. Soil Science Society of America Journal, in press.
- Scharlemann, J.P., Tanner, E.V., Hiederer, R., and Kapos, V. (2014). Global soil carbon: understanding and managing the largest terrestrial carbon pool. Carbon Management, 5 (1), 81–91.
- Pekel, J. F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. Nature 540, 418-422. (doi:10.1038/nature20584)
- Poeplau, C. and Vos, C. and Don, A., 2017? Soil organic carbon stocks are systematically overestimated by misuse of the parameters bulk density and rock fragment content. SOIL Discussions, 3 (1), 61–66.
- Wager, S., Hastie, T., & Efron, B. (2014). Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. Journal of Machine Learning Research, 15(1), 1625-1651.
- Wei, X., Shao, M., Gale, W., & Li, L. (2014). Global pattern of soil carbon losses due to the conversion of forests to agricultural land. Scientific reports, 4, 4062.
- Wright, M.N., Ziegler, A. (2016). ranger: A Fast Implementation of Random Forests for High

Dimensional Data in C++ and R. Journal of Statistical Software, v77(1).
- Xu, X., Thornton, P.E., and Post, W.M. (2012). A global analysis of soil microbial biomass carbon, nitrogen and phosphorus in terrestrial ecosystems. Global Ecology and Biogeography, 22 (6), 737–749.

---

wiki soil organic carbon

☒ from 0 votes (Details)

○ ○ ○ ○ ○ Rate

☒ 0 visitor votes
☒ 0 visitor votes
☒ 0 visitor votes
☒ 0 visitor votes
☒ 0 visitor votes

From:
http://gsif.isric.org/ - **GSIF (tutorials)**

Permanent link:
http://gsif.isric.org/doku.php/wiki:soil_organic_carbon

Last update: **2017/10/08 23:23**