Continuous soil attribute modeling and mapping: Regression Kriging (2)

Soil Security Laboratory

2018

1 Regression kriging

In the previous sections we looked at a few soil spatial prediction functions which at the most fundamental level, target the the correlation between the target soil variable and the available covariate information. We fitted a number of models which included simple linear functions to non-linear functions such as regression trees to other more complicated data mining techniques (Cubist and Random Forest). In this section we will extend upon this DSM approach from what are called deterministic models to also include the spatially correlated residuals that result from fitting these models.

The approach we will now concentrate is a hybrid approach to modelling, whereby the predictions of the target variable are made via a deterministic method (regression model with covariate information) and a stochastic method where we determine the spatial auto-correlation of the model residuals with a variogram. The deterministic model essentially "detrends" the data, leaving behind the residuals for which we need to investigate whether there is additional spatial structure which could be added to the regression model predictions. These residuals are the random component of the $scorpan + emodel$. This method is described as regression kriging and has formally been described in Odeh et al. (1995) and is synonymous with universal kriging (Hengl et al., 2007), which is the formal linear model procedure to this soil spatial modeling approach. The purpose of this exercise is to introduce some basic concepts of regression kriging. You will have already had some experience in regression models. We have also investigated briefly the fundamental concepts of kriging for which the variogram is central to.

1.1 Regression kriging with Cubist models

In the first example the universal kriging model was introduced. Here we generalise the regression kriging. The following example will provide the steps one would use to perform regression kriging that incorporates a complex model structure such as a data mining algorithm. Here we will use the Cubist model that was used earlier. Lets start from the beginning.

First get the data and perform the covariate data intersection

```
library(ithir)
library(raster)
library(rgdal)
library(sp)
library(gstat)

# point data
data(HV_subsoilpH)

# Start afresh round pH data to 2 decimal places
HV_subsoilpH$pH60_100cm <- round(HV_subsoilpH$pH60_100cm, 2)

# remove already intersected data
HV_subsoilpH <- HV_subsoilpH[, 1:3]

# add an id column
HV_subsoilpH$id <- seq(1, nrow(HV_subsoilpH), by = 1)

# re-arrange order of columns
HV_subsoilpH <- HV_subsoilpH[, c(4, 1, 2, 3)]

# Change names of coordinate columns
names(HV_subsoilpH)[2:3] <- c("x", "y")


# grids (covariate raster)
data(hunterCovariates_sub)
```

Perform the covariate intersection.

```
coordinates(HV_subsoilpH) <- ~x + y

# extract
DSM_data <- extract(hunterCovariates_sub, HV_subsoilpH, sp = 1, method = "simple")
DSM_data <- as.data.frame(DSM_data)
str(DSM_data)
```

```
## 'data.frame': 506 obs. of  15 variables:
##  $ id                      : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ x                       : num  340386 340345 340559 340483 340734 ...
##  $ y                       : num  6368690 6368491 6369168 6368740 6368964 ...
##  $ pH60_100cm              : num  4.47 5.42 6.26 8.03 8.86 7.28 4.95 5.61 5.39 3.44 ...
##  $ Terrain_Ruggedness_Index: num  1.34 1.42 1.64 1.04 1.27 ...
##  $ AACN                    : num  1.619 0.281 2.301 1.74 3.114 ...
##  $ Landsat_Band1           : num  57 47 59 52 62 53 47 52 53 63 ...
##  $ Elevation               : num  103.1 103.7 99.9 101.9 99.8 ...
##  $ Hillshading             : num  1.849 1.428 0.934 1.517 1.652 ...
##  $ Light_insolation        : num  1689 1701 1722 1688 1735 ...
##  $ Mid_Slope_Positon       : num  0.876 0.914 0.844 0.848 0.833 ...
```

```
##  $ MRVBF                   : num  3.85 3.31 3.66 3.92 3.89 ...
##  $ NDVI                    : num  -0.143 -0.386 -0.197 -0.14 -0.15 ...
##  $ TWI                     : num  17.5 18.2 18.8 18 17.8 ...
##  $ Slope                   : num  1.79 1.42 1.01 1.49 1.83 ...
```

Often it is handy to check to see whether there are missing values both in the target variable and of the covariates. It is possible that a point location does not fit within the extent of the available covariates. In these cases the data should be excluded. A quick way to assess whether there are missing or NA values in the data is to use the `complete.cases` function.

```
which(!complete.cases(DSM_data))
```

```
## integer(0)
```

```
DSM_data <- DSM_data[complete.cases(DSM_data), ]
```

Now lets begin the regression kriging modeling

```
library(Cubist)
set.seed(875)
training <- sample(nrow(DSM_data), 0.7 * nrow(DSM_data))
mDat <- DSM_data[training, ]

# fit the model
hv.cub.Exp <- cubist(x = mDat[, c("AACN", "Landsat_Band1", "Elevation", "Hillshading",
    "Mid_Slope_Positon", "MRVBF", "NDVI", "TWI")], y = mDat$pH60_100cm,
     cubistControl(rules = 100, extrapolation = 15), committees = 1)
```

Now derive the model residual which is the model prediction subtracted from the residual.

```
mDat$residual <- mDat$pH60_100cm - predict(hv.cub.Exp, newdata = mDat)
mean(mDat$residual)
```

```
## [1] 0.1572845
```

If you check the histogram of these residuals you will find that the mean is around zero and the data seems normally distributed. Now we can assess the residuals for any model structure.

```
coordinates(mDat) <- ~x + y
crs(mDat) <- "+proj=utm +zone=56 +south +ellps=WGS84
             + datum=WGS84 +units=m +no_defs"

vgm1 <- variogram(residual ~ 1, mDat, width = 200, cutoff = 3000)
mod <- vgm(psill = var(mDat$residual), "Sph", range = 3000, nugget = 0)
model_1 <- fit.variogram(vgm1, mod)
model_1
```

```
##   model     psill    range
## 1   Nug 0.6948288   0.0000
## 2   Sph 0.5827945 856.4408
```

```
# Residual kriging model
gRK <- gstat(NULL, "RKresidual", residual ~ 1, mDat, model = model_1)
```

With the two model components together, we can now compare the external validation statistics of using the Cubist model only and with the Cubist model and residual variogram together.

```
# Cubist model only
Cubist.pred.V <- predict(hv.cub.Exp, newdata = DSM_data[-training, ])

# Cubist model with residual variogram
vDat <- DSM_data[-training, ]
coordinates(vDat) <- ~x + y
crs(vDat) <- "+proj=utm +zone=56 +south +ellps=WGS84 + datum=WGS84 +units=m +no_defs"

# make the residual predictions
RK.preds.V <- as.data.frame(krige(residual ~ 1, mDat, model = model_1, newdata = vDat))

## [using ordinary kriging]

# Sum the two components together
RK.preds.fin <- Cubist.pred.V + RK.preds.V[, 3]

# validation cubist only
goof(observed = DSM_data$pH60_100cm[-training], predicted = Cubist.pred.V)

##          R2 concordance      MSE     RMSE       bias
## 1 0.2995264    0.460038 1.137435 1.066506 -0.1578419

# validation regression kriging with cubist model
goof(observed = DSM_data$pH60_100cm[-training], predicted = RK.preds.fin)

##          R2 concordance      MSE     RMSE        bias
## 1 0.3899789   0.6098103 1.025823 1.012829 -0.04931782
```

These results confirm that there to be some advantage in performing regression kriging with this particular data. In any case, to apply the regression kriging model here, it requires three steps: First apply the Cubist model, then apply the residual kriging, then finally add both maps together. The script below illustrates how this is done, and the resulting maps are shown on Figure 1.

```
par(mfrow = c(3, 1))
map.RK1 <- predict(hunterCovariates_sub, hv.cub.Exp,
filename = "soilpH_60_100_cubistRK.tif",
    format = "GTiff", datatype = "FLT4S", overwrite = TRUE)
plot(map.RK1, main = "Cubist model predicted soil pH")

map.RK2 <- interpolate(hunterCovariates_sub, gRK, xyOnly = TRUE, index = 1,
    filename = "soilpH_60_100_residualRK.tif", format = "GTiff", datatype = "FLT4S",
    overwrite = TRUE)
plot(map.RK2, main = "Kriged residual")
```

```
# Stack prediction and kriged residuals
pred.stack <- stack(map.RK1, map.RK2)

map.RK3 <- calc(pred.stack, fun = sum, filename = "soilpH_60_100_finalPredRK.tif",
    format = "GTiff", progress = "text", overwrite = T)
plot(map.RK3, main = "Regression kriging prediction")
```

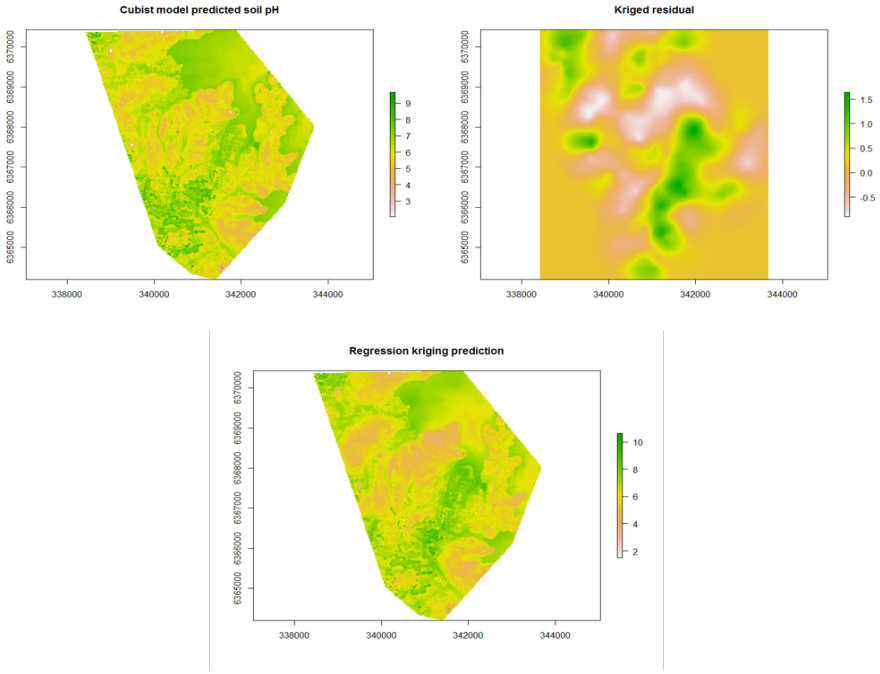

Figure 1: Regression kriging predictions with cubist models. Hunter Valley soil pH (60-100cm).

References

Hengl, T., G. B. M. Heuvelink, and D. G. Rossiter
2007. About regression kriging: From equations to case studies. *Computers & Geosciences*, 33:1301–1315.

Odeh, I. O. A., A. B. McBratney, and D. J. Chittleborough
1995. Further results on prediction of soil properties from terrain attributes: heterotopic co-kriging and regression kriging. *Geoderma*, 67:215–226.