



NPTEL ONLINE CERTIFICATION COURSES

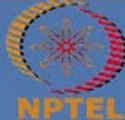
Machine Learning for Soil and Crop Management
Prof. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology Kharagpur

**Week 11: DIGITAL SOIL MAPPING WITH
CONTINUOUS VARIABLES**

LECTURE 51

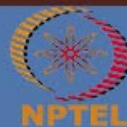
CONCEPTS COVERED

- Handling data for DSM
- Continuous Models
- GIS with R-overview



KEYWORDS


- GIS
- rgdal
- EPSG
- KML
- Google Earth



EPSG CODE

- EPSG stands for European Petroleum Survey Group and is an organization that maintains a geodetic parameter database with standard codes, the EPSG codes, for coordinate systems, and datums

<http://spatialreference.org/>

spatial reference list

[Home](#) | [Upload Your Own](#) | [List user-contributed references](#) | [List all references](#)

Search References:

[Next Page](#)

- EPSG:2000: Anguilla 1957 / British West Indies Grid
- EPSG:2001: Antigua 1943 / British West Indies Grid
- EPSG:2002: Dominica 1945 / British West Indies Grid
- EPSG:2003: Grenada 1953 / British West Indies Grid
- EPSG:2004: Montserrat 1958 / British West Indies Grid
- EPSG:2005: St. Kitts 1955 / British West Indies Grid
- EPSG:2006: St. Lucia 1955 / British West Indies Grid
- EPSG:2007: St. Vincent 45 / British West Indies Grid
- EPSG:2008: NAD27(CGQ77) / SCoPQ zone 2
- EPSG:2009: NAD27(CGQ77) / SCoPQ zone 3
- EPSG:2010: NAD27(CGQ77) / SCoPQ zone 4
- EPSG:2011: NAD27(CGQ77) / SCoPQ zone 5
- EPSG:2012: NAD27(CGQ77) / SCoPQ zone 6
- EPSG:2013: NAD27(CGQ77) / SCoPQ zone 7
- EPSG:2014: NAD27(CGQ77) / SCoPQ zone 8
- EPSG:2015: NAD27(CGQ77) / SCoPQ zone 9
- EPSG:2016: NAD27(CGQ77) / SCoPQ zone 10
- EPSG:2017: NAD27(76) / MTM zone 8
- EPSG:2018: NAD27(76) / MTM zone 9
- EPSG:2019: NAD27(76) / MTM zone 10
- EPSG:2020: NAD27(76) / MTM zone 11
- EPSG:2021: NAD27(76) / MTM zone 12
- EPSG:2022: NAD27(76) / MTM zone 13
- EPSG:2023: NAD27(76) / MTM zone 14
- EPSG:2024: NAD27(76) / MTM zone 15
- EPSG:2025: NAD27(76) / MTM zone 16
- EPSG:2026: NAD27(76) / MTM zone 17
- EPSG:2027: NAD27(76) / UTM zone 15N
- EPSG:2028: NAD27(76) / UTM zone 16N
- EPSG:2029: NAD27(76) / UTM zone 17N
- EPSG:2030: NAD27(76) / UTM zone 18N
- EPSG:2031: NAD27(CGQ77) / UTM zone 17N
- EPSG:2032: NAD27(CGQ77) / UTM zone 18N
- EPSG:2033: NAD27(CGQ77) / UTM zone 19N
- EPSG:2034: NAD27(CGQ77) / UTM zone 20N
- EPSG:2035: NAD27(CGQ77) / UTM zone 21N
- EPSG:2036: NAD83(CSR598) / New Brunswick Stereo
- EPSG:2037: NAD83(CSR598) / UTM zone 19N
- EPSG:2038: NAD83(CSR598) / UTM zone 20N
- EPSG:2039: Israel / Israeli TM Grid
- EPSG:2040: Locodjo 1965 / UTM zone 30N
- EPSG:2041: Abidjan 1987 / UTM zone 30N
- EPSG:2042: Locodjo 1965 / UTM zone 29N
- EPSG:2043: Abidjan 1987 / UTM zone 29N
- EPSG:2044: Hanoi 1972 / Gauss-Kruger zone 18
- EPSG:2045: Hanoi 1972 / Gauss-Kruger zone 19
- EPSG:2046: Hartebeesthoek94 / Lo15
- EPSG:2047: Hartebeesthoek94 / Lo17
- EPSG:2048: Hartebeesthoek94 / Lo19
- EPSG:2049: Hartebeesthoek94 / Lo21

[Next Page](#)

[About](#)



ESRI SHAPE FILE

- A shapefile is an Esri vector data storage format for storing the location, shape, and attributes of geographic features. It is stored as a set of related files and contains one feature class.

*.shp: contains the feature geometries.

*.dbf: contains feature attribute data, as a table.

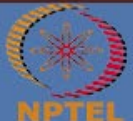
*.shx: indexation data for iterations accross the features.

*.prj: the coordinate reference system represented as text.

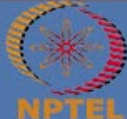


REFERENCE

Malone, B. P., Minasny, B., & McBratney, A. B. (2017). *Using R for digital soil mapping* (Vol. 35). Cham, Switzerland: Springer International Publishing.



*Thank
you*





NPTEL ONLINE CERTIFICATION COURSES

Machine Learning for Soil and Crop Management
Prof. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology Kharagpur

**Week 11: DIGITAL SOIL MAPPING WITH
CONTINUOUS VARIABLES**

LECTURE 52

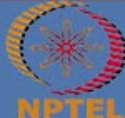
CONCEPTS COVERED

- Basic GIS operations
- Some useful packages in DSM
- Mass Preserving Spline
- Covariates
- Intersecting Covariates



KEYWORDS

- Mass preserving spline
- GlobalSoilMap.net
- Normality
- Covariates
- OneProfile



SOME USEFUL PACKAGES

- **Soil science and pedometrics**
- **aqp:** Algorithms for quantitative pedology. <http://cran.r-project.org/web/packages/aqp/index.html>. A collection of algorithms related to modeling of soil resources, soil classification, soil profile aggregation, and visualization.
- **GSIF:** Global soil information facility. <http://cran.r-project.org/web/packages/GSIF/index.html>. Tools, functions and sample datasets for digital soil mapping.



SOME USEFUL PACKAGES

- **GIS**
- **sp**: <http://cran.r-project.org/web/packages/sp/index.html>. A package that provides classes and methods for spatial data. The classes document where the spatial location information resides, for 2D or 3D data. Utility functions are provided, e.g. for plotting data as maps, spatial selection, as well as methods for retrieving coordinates, for sub-setting, print, summary, etc.
- **raster**: <http://cran.r-project.org/web/packages/raster/index.html>. Reading, writing, manipulating, analyzing and modeling of gridded spatial data. The package implements basic and high-level functions and processing of very large files is supported.



SOME USEFUL PACKAGES

- **GIS**
- **rgdal**: <http://cran.r-project.org/web/packages/rgdal/index.html>. Provides access to projection/transformation operations from the PROJ.4 library. Both GDAL raster and OGR vector map data can be imported into R, and GDAL raster data and OGR vector data exported. Use is made of classes defined in the sp package.
- **RSAGA**: <http://cran.r-project.org/web/packages/RSAGA/index.html>. RSAGA provides access to geocomputing and terrain analysis functions of SAGA GIS <http://www.saga-gis.org/en/index.html> from within R by running the command line version of SAGA.



SOME USEFUL PACKAGES

- **Modeling**
- caret: Extensive range of functions for training and plotting classification and regression models. See the caret website for more detailed information <http://topepo.github.io/caret/index.html>.
- Cubist: Regression modeling using rules with added instance-based corrections. Cubist models were developed by Ross Quinlan. Further information can be found at Rulequest <https://www.rulequest.com/>
- C5.0: C5.0 decision trees and rule-based models for pattern recognition. Another model structure developed by Ross Quinlan.
- gam: Functions for fitting and working with generalized additive models.
- nnet: Software for feed-forward neural networks with a single hidden layer, and for multinomial log-linear models.
- gstat: Variogram modelling; simple, ordinary and universal point or block (co)kriging, sequential Gaussian or indicator (co)simulation; variogram and variogram map plotting utility functions. A related and useful package is automap (<http://cran.r-project.org/web/packages/automap/index.html>), which performs an automatic interpolation by automatically estimating the variogram and then calling gstat.



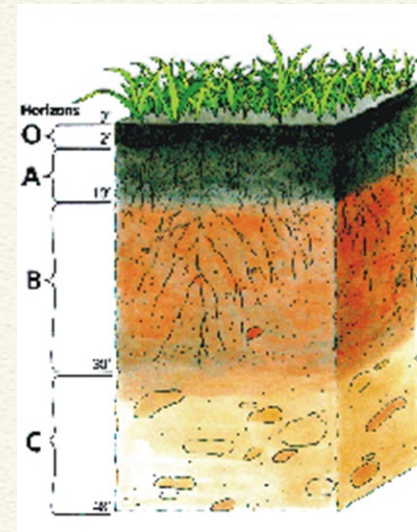
SOME USEFUL PACKAGES

- **Plotting**
- **ggplot2**: This package is an implementation of the grammar of graphics in R. It combines the advantages of both base and lattice graphics: conditioning and shared axes are handled automatically, and you can still build up a plot step by step from multiple data sources. It also implements a sophisticated multidimensional conditioning system and a consistent interface to map data to aesthetic attributes. See the ggplot2 website for more information, documentation and examples (<http://ggplot2.org>).



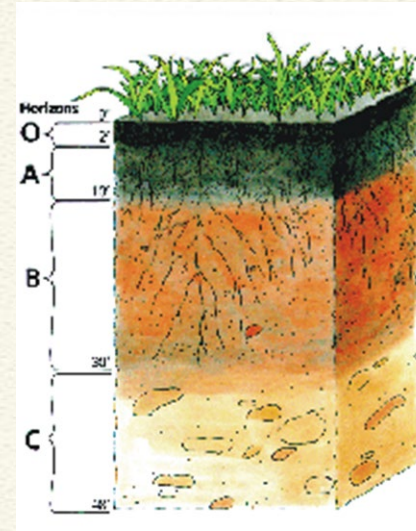
DSM ISSUES

- All soils are not measured universally at the same depths.
 - Some soils are sampled per horizon or at regular depths.
 - Some soil studies examine only the topsoil, while others sample to the bedrock depth
 - Then different soil attributes are measured at some locations and depths, but not at others.
- Thus, a number of preprocessing steps are needed to fulfill the requirements of DSM



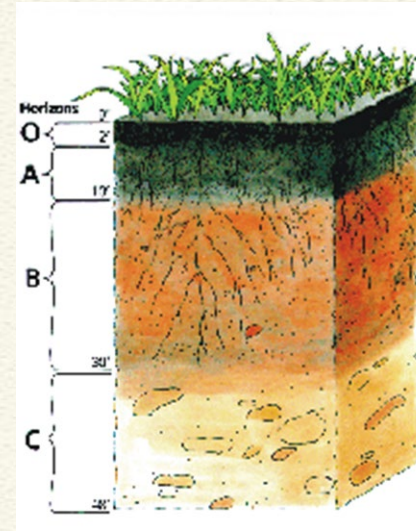
SOIL DEPTH FUNCTION

- The traditional method of sampling soil involves dividing a soil profile into horizons.
- The number of horizons and the position of each are generally based on attributes easily observed in the field, such as morphological soil properties
- From each horizon, a bulk sample is taken and it is assumed to represent the average value for a soil attribute over the depth interval from which it is sampled.



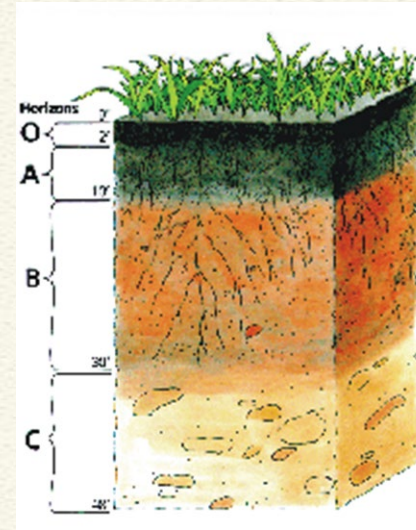
THEN WHAT IS THE PROBLEM?

- Issue 1: from the pedological perspective soil generally varies continuously with depth; however, representing the soil attribute value as the average over the depth interval of horizons leads to discontinuous or stepped profile representations.
 - What if one wants to know the value of an attribute at a specified depth?



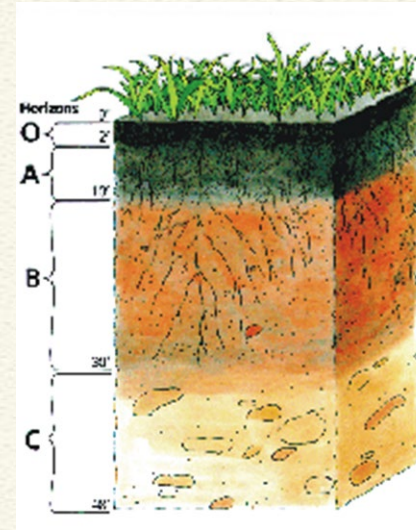
THEN WHAT IS THE PROBLEM?

- Issue 2: difficult for DSM modeling: observations at each horizon for each profile will rarely be the same between any two profiles



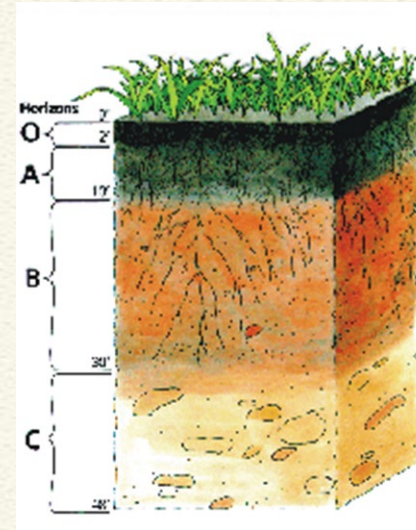
SOLUTION

- We can not ignore legacy data
- need to derive a continuous function using the available horizon data as some input.
 - Polynomials and exponential decay type depth functions.
 - Continuous depth function like the equal-area quadratic spline function.



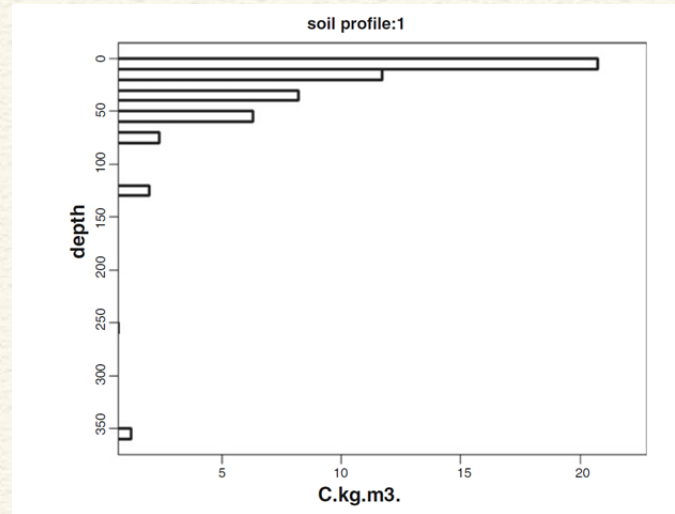
WHY “MASS-PRESERVING”

- The original data is preserved and can be retrieved again via integration of the continuous spline
- The spline parameters are the values of the soil attribute at the standard depths that are specified by the user
 - Harmonize soil profile data and model at a specified depths



GlobalSoilMap.net project

- A global consortium has been formed that aims to make a new digital soil map of the world using state-of-the-art and emerging technologies for soil mapping and predicting soil properties at fine resolution.
- Depths
 - 0-5 cm
 - 5-15 cm
 - 15-30 cm
 - 30-60 cm
 - 60-100 cm
 - 100-200 cm



OneProfile data



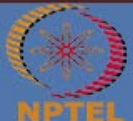
Intersecting Soil Point Observations with Environmental Covariates

- In order to carry out DSM in terms of evaluating the significance of environmental variables in explaining the spatial variation of the target soil variable under investigation, we need to link both sets of data together and extract the values of the covariates at the locations of the soil point data.
- The first task is to bring in to our working environment some soil point data

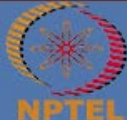


REFERENCE

Malone, B. P., Minasny, B., & McBratney, A. B. (2017). *Using R for digital soil mapping* (Vol. 35). Cham, Switzerland: Springer International Publishing.



*Thank
you*





NPTEL ONLINE CERTIFICATION COURSES

Machine Learning for Soil and Crop Management
Prof. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology Kharagpur

**Week 11: DIGITAL SOIL MAPPING WITH
CONTINUOUS VARIABLES**

LECTURE 53

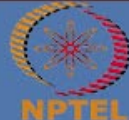
CONCEPTS COVERED

- Exploratory data analysis with R
- Kriging with R



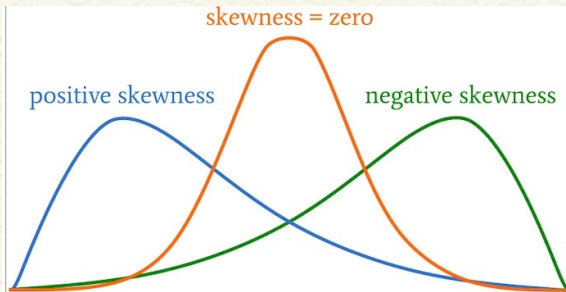
KEYWORDS

- Kriging
- Normality test
- Skewness
- Kurtosis
- Variogram



SKEWNESS

Distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data.



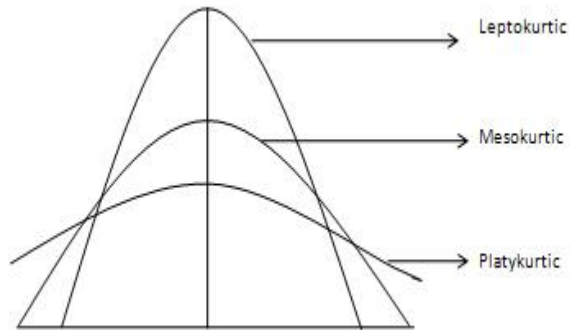
Skewness value: >1 or <-1 indicates a highly skewed distribution.

Value between 0.5 to 1 or -0.5 to -1 is moderately skewed.

Value between -0.5 and 0.5 indicates that the distribution is fairly symmetrical.

KURTOSIS

The sharpness of the peak of a frequency-distribution curve.



kurtosis identifies whether the tails of a given distribution contain extreme value.

If the value is greater than +1, the distribution is too peaked. Likewise, a kurtosis of less than -1 indicates a distribution that is too flat.

SEMIVARIANCE

The average variance between any pair of sampling points (calculated as the semi-variance) for a soil property S at any point of distance h apart.

$$\gamma(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} \{s(x_i) - s(x_i + h)\}^2$$

- $\gamma(h)$ = average semi-variance,
m = the number of pairs of sampling points
s = value of the attribute under investigation,
x = coordinates of the point
h = lag (separation distance of point pairs)

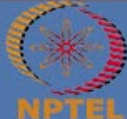


REFERENCE

Malone, B. P., Minasny, B., & McBratney, A. B. (2017). *Using R for digital soil mapping* (Vol. 35). Cham, Switzerland: Springer International Publishing.



*Thank
you*





NPTEL ONLINE CERTIFICATION COURSES

Machine Learning for Soil and Crop Management
Prof. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology Kharagpur

**Week 11: DIGITAL SOIL MAPPING WITH
CONTINUOUS VARIABLES**

LECTURE 54

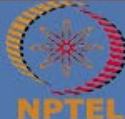
CONCEPTS COVERED

- Model validation
- SLR and MLR based mapping
- Stepwise regression based mapping
- Decision Tree based mapping



KEYWORDS

- SLR
- MLR
- Stepwise regression
- Decision Tree
- goof





Model Validation

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^n (obs_i - pred_i)^2}{n}}$$

where *obs*=the observed soil property

pred =predicted soil property from a given model

n =the number of observations *i*.

Bias, also called the mean error of prediction and is defined as:

$$bias = \frac{\sum_{i=1}^n pred_i - obs_i}{n}$$

Pearson's correlation coefficient

$$r = \frac{\sum_{i=1}^n (obs_i - \overline{obs})(pred_i - \overline{pred})}{\sqrt[2]{\sum_{i=1}^n (obs_i - \overline{obs})^2} \sqrt[2]{\sum_{i=1}^n (pred_i - \overline{pred})^2}}$$

R²= coefficient of determination

Lin's concordance correlation coefficient (Lin 1989):is a single statistic that both evaluates the accuracy and precision of the relationship. It is often referred to as the goodness of fit along a 45 degree line. Thus it is probably a more useful statistic than the R² alone

$$\rho_c = \frac{2\rho\sigma_{pred}\sigma_{obs}}{\sigma_{pred}^2 + \sigma_{obs}^2 + (\mu_{pred} - \mu_{obs})^2}$$

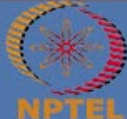
where μ_{pred} and μ_{obs} are the means of the predicted and observed values respectively. σ^2_{pred} and σ^2_{obs} are the corresponding variances. ρ is the correlation coefficient between the predictions and observations.

REFERENCE

Malone, B. P., Minasny, B., & McBratney, A. B. (2017). *Using R for digital soil mapping* (Vol. 35). Cham, Switzerland: Springer International Publishing.



*Thank
you*





NPTEL ONLINE CERTIFICATION COURSES

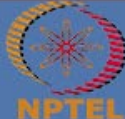
Machine Learning for Soil and Crop Management
Prof. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology Kharagpur

**Week 11: DIGITAL SOIL MAPPING WITH
CONTINUOUS VARIABLES**

LECTURE 55

CONCEPTS COVERED

- CUBIST
- RANDOM FOREST
- UNIVERSAL KRIGING



KEYWORDS

- Cubist
- Random Forest
- ntree
- cubistControl
- Universal Kriging



CUBIST

- A very popular model structure used within the DSM community.
- Its popularity is due to its ability to “mine” non-linear relationships in data, but does not have the issues of finite predictions that occur for other decision and regression tree models

CUBIST

- Based on the M5 algorithm of Quinlan (1992)
- The Cubist model first partitions the data into subsets within which their characteristics are similar with respect to the target variable and the covariates.
- A series of rules (a decision tree structure may also be defined if requested) defines the partitions, and these rules are arranged in a hierarchy. Each rule takes the form:
 - **if [condition is true]**
 - **then [regress]**
 - **else [apply the next rule]**

CUBIST

- The condition may be a simple one based on one covariate or, more often, it comprises a number of covariates. If a condition results in being true then the next step is the prediction of the soil property of interest by ordinary least-squares regression from the covariates within that partition. If the condition is not true then the rule defines the next node in the tree, and the sequence of if, then, else is repeated.
- The result is that the regression equations, though general in form, are local to the partitions and their errors smaller than they would otherwise be.

CUBIST

- Luckily, fitting a Cubist model in R is not too difficult—although it will be useful to spend some time playing around with many of the controllable parameters the function has.
- In the example we will try today we can control the number of potential rules that could potentially partition the data (note this limits the number of possible rules, and does not necessarily mean that those number of rules will actually be realized i.e. the outcome is internally optimised).
- We can also limit the extrapolation of the model predictions, which is a useful model constraint feature. These various control parameters plus others can be adjusted within the `cubistControl` parameter.
- Does not unnecessarily overfits the data

REFERENCE

Malone, B. P., Minasny, B., & McBratney, A. B. (2017). *Using R for digital soil mapping* (Vol. 35). Cham, Switzerland: Springer International Publishing.



*Thank
you*

