

Homework 5: Model Predictions

Updated March 2016

Table of Contents

[Homework 5: Model Predictions](#)

[Table of Contents](#)

[Rasters](#)

[Homework Exercise](#)

[OBJECTIVE](#)

[Step 1 - Adding Covariates to the Mix](#)

[Step 2 - Final Model Selection](#)

[Step 3 - Predicting Carbon](#)

[Resources](#)

[Data](#)

[References](#)

[Citations](#)

[Keywords](#)

Rasters

In its simplest form, a raster consists of a matrix of cells (or pixels) organized into rows and columns (or a grid) where each cell contains a value representing information, such as temperature, elevation or precipitation. Rasters are digital aerial photographs, imagery from satellites, digital pictures, or even scanned maps [1]. (Figure from ESRI.com)

General characteristics of raster data

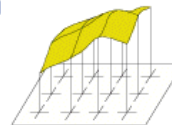
In raster datasets, each cell (which is also known as a pixel) has a value. The cell values represent the phenomenon portrayed by the raster dataset such as a category, magnitude, height, or spectral value. The category could be a land-use class such as grassland, forest, or road. A magnitude might represent gravity, noise pollution, or percent rainfall. Height (distance) could represent surface elevation above mean sea level, which can be used to derive slope, aspect, and watershed properties. Spectral values are used in satellite imagery and aerial photography to represent light reflectance and color.

Cell values can be either positive or negative, integer, or floating point. Integer values are best used to represent categorical (discrete) data, and floating-point values to represent continuous surfaces. For additional information on discrete and continuous data, see [Discrete and continuous data](#). Cells can also have a NoData value to represent the absence of data. For information on NoData, see [NoData in raster datasets](#).

Value applies to the center point of the cell

For certain types of data, the cell value represents a measured value at the center point of the cell. An example is a raster of elevation

315	319	321	323
317	323	328	326
313	318	325	329



Value applies to the whole area of the cell

For most data, the cell value represents a sampling of a phenomenon, and the value is presumed to represent the whole cell square.

50	45	40	35
35	40	35	25
20	25	30	20



Homework Exercise (3 Points total)

These exercises rely on the cleaned lab data produced from your previous exercises (re-uploaded to CATCOURSES for comparison under `riparian_HW5.csv`. These are the same data from last week's homework.

Your document should have the following sections, and provide written explanations formatted in RMarkdown that explains your code, output and graphics in the following format:

NAME
CLASS
DATE

Homework Assignment 5

Objective Statement: [What are you trying to accomplish?]

Methods: [In general terms, what analyses are you doing?]

Data: [What are the data and where did they come from?]

Code: [In specific terms, what is the code that was used to conduct the analysis?]

Results: [What do the results show? Numerical evidence and graphic evidence are required.]

Discussion: [What do the results mean?]

Limitations: [What are the limitations, caveats, and assumptions of the analysis?]

OBJECTIVE

You have been asked to finalize your analysis of riparian tree data by developing a predictive model to characterize the amount of biomass found in the riparian habitats of northern California. You have explored a dataset of field measurements and observations taken at various project sites throughout northern California that are intended to estimate aboveground carbon stocks in riparian areas. Based on your preliminary assessment, you realize that new data must be brought into the model to better allow for predictions in places not yet sampled, as your client has requested an assessment of standing carbon for a project site that has only tree diameters, not height. This will require prediction.

To finalize this investigation, you extend your analysis from last week (Homework 5) that developed a promising linear regression model. In order to make reasonable estimates of standing carbon in a new location, you decide to finalize a linear model that relates tree height as a function of its diameter (at breast height) or dbh and genus but with one additional covariate, such as temperature, precipitation, latitude, or elevation. In this exercise, you will

build a new report showing model results for the model of `ht ~ dbh * genus + parameter` where `parameter` is one of the covariates. In addition to defending the model -- based on performance and ecological justification -- use the model to predict the standing carbon (expressed as Mg of C per hectare) of a new project site.

Step 1 - Adding Covariates to the Mix

Load the following raster data: `DEM.tif` (digital elevation model), `tmean_8.tif` (mean temperature in August) and `precip_8.tif` (precipitation in August) raster data using the `rgdal` and `raster` libraries. Use the following as a guide for loading rasters:

```
#read data from file (2 steps)
gdal_grid = readGDAL("rasterfile.tif")
r = raster(gdal_grid) #use data as a projected raster
```

Visualize your data by plotting the DEM raster and add the project locations (`ProjLoc` from homework 3 using `aggregate`) as points (use the `plot` and `points` functions).

Extract the values from the rasters at the locations of each points and add each of them as a new column in your riparian dataframe. Name your columns as follows:

<code>\$Elevation</code>	<code>#extracted values from the DEM raster</code>
<code>\$Temp_aug</code>	<code>#extracted values from the tmean8 raster</code>
<code>\$Precp_aug</code>	<code>#extracted values from the precip8 raster</code>

Use the following as a guide for extracting values from rasters:

```
#Example lon and lat extraction
Longitude = c(-122,-121) #Just an example! Use the lon and lat
Latitude = c(36,38)      #from your data!

#x,y locations
xy = cbind(Longitude,Latitude)

#extract the values from raster r
vals = extract(r,xy)

#join columns to data frame using cbind() or similar method
#see melt() function in library(reshape) for more robust approach
```

Conduct exploratory data analysis on these new data. Are they normally distributed? Are they highly correlated?

Step 2 - Final Model Selection

From the previous homework, it was concluded that the model using interaction with genus seemed to perform the best ($htcm \sim DBH * Genus$). But we would like to test whether the model could be improved using the covariates from Step 1. Along with these new parameters, latitude is also thought to have a relationship with tree heights. From these covariates, find the best parameter to add to the model (if any!).

Build models for each new predictor variable by adding it to the interaction model (`lm.pred1 <- lm(htc~DBH*Genus + [predictor1])`, where `[predictor]` is the new parameter). Test all four models against each other and against the base interaction model to decide which is best. Considering p-values of parameters and R^2 of each model, plus AIC values (using the `AIC()` function) when comparing models, defend which model you would choose to predict new data.

Advanced users are encouraged to evaluate their models using `step()` and `stepAIC()`. Super advanced users are encouraged to develop a model that has a zero intercept.

Step 3 - Predicting Carbon

A new dataset has been given to you (`new_data.csv`) that represents dbh measurements taken at a new project location. Use your chosen (final) model from Step 2 to predict (`predict.lm()`) the heights of the unmeasured trees in the 10 m x 10 m plot and then apply the following conversion to the volume of the tree (this equation, adapted from Pilsbury (1984) and Cornwell (2006), is specific to Valley oaks (*Quercus lobata*), but use it for all trees in the plot). Assume that C is 50% of biomass volume (it typically varies 45-55%). Report your results as Mg of C per hectare for the new site.

$$\text{Above Ground Tree Volume} = 705 * (0.0000334750 * (\text{dbh in cm})^{2.33631} * (\text{ht in cm})^{0.74872})$$

Given the assumptions of density and the single allometric [2] equation, is your model likely to under or over-estimate your predicted C stocks? If you needed to estimate 95% confidence intervals around your fitted model, how might you do that? (See `help(predict.lm)` as a starting point.)

###NOTE###

#the way that predict works for new data requires that predictor
#variables have the same column names as in the model, for example:

```
lm1 = lm(slope~elev, data=dem) # create model
with(dem,plot(elev,slope)) # plot two variables
x = seq(0,2000) # create 'new' elevation values in sequence

### use predict() with new x values to get new y values ###
### newdata requires a data.frame or a list, not a vector ###
### column names need to match however, so check data first ###

lm1.pred.y <- predict(lm1,newdata=list(elev=x))

lines(x,lm1.pred.y,col="blue",lwd=3) # model predicted line on plot
```

Resources

Data

- riparian_HW5.csv on CATCOURSES (Resources→HOMEWORK→HOMEWORK 5)
- DEM.tif on CATCOURSES (Resources→HOMEWORK→HOMEWORK 6)
- precip_8.tif on CATCOURSES (Resources→HOMEWORK→HOMEWORK 6)
- tmean_8.tif on CATCOURSES (Resources→HOMEWORK→HOMEWORK 6)
- new_data.csv on CATCOURSES (Resources→HOMEWORK→HOMEWORK 6)

References

- See Crawley sections on model selection and AIC.

Citations

- [1] http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=What_is_raster_data%3F
[2] <http://en.wikipedia.org/wiki/Allometry>

Keywords

readGDAL(), raster(), cbind(), extract(), predict.lm()