AMERICAN MUSEUM OF NATURAL HISTORY
CENTER FOR BIODIVERSITY
AND CONSERVATION

October, 2011

Training Guide for adjusting random forests regression images

Introduction

This guide explains how to adjust values predicted using random forest regression models. When used for regression, the random forests algorithm will overestimate low values and underestimate high values. The CorrectRF_Result.R script adjusts this effect by calculating regression coefficients using the response variable from a input training data file and the corresponding predictor variables from the image output using X/Y coordinates also provided by the input file. A new adjusted image image is output after applying gain and offset (slope and intercept) values to the original predicted image values. The user can select the type of regression to apply. To facilitate the choice, a graph of a scatter plot using the response and predictor variables is displayed and the regression lines for three different regression algorithms plotted. The solid line is the regression line for the Theil-Sen Siegel repeated medians algorithm, the dashed line is for a linear regression , and the dotted line is for a linear regression line with an intercept of 0. The user can also set the minimum and maximum values for the output image. This is useful when processing percent cover images when the valid range is between 0 and 1.

Before running the scripts it is important that R is installed on your computer and you have downloaded the necessary packages. Instructions for installing R and the necessary packages can be found on the Biodiversity Informatics website. In addition to the packages listed in the instructions for installing R and the necessary packages you will need to install the "mblm" package.

1. Edit and run the script

This script requires that a number of variables be set before running it. The variables are located in the "SET VARIABLES HERE" section of the script. Below is an explanation of each variable.

fileType: This is the type of file that contains the X/Y coordinates and response variables. Enter a 1 if is is a dbf file or 2 if it is a CSV file with a header.
pointData: Path and name for the CSV or DBF file containing X, Y, and response variable (i.e., biomass, % cover...) data.
inImage: Path and name for the input predicted image.
outImage: Path and name for the output adjusted image.
nd: No-data value for the input image
x_coord: The the name (case sensitive and in quotes) or the column number of the field containing X coordinates

y_coord: The the name (case sensitive and in quotes) or the column number of the field containing Y coordinates
responseVar: The the name (case sensitive and in quotes) or the column number of the field containing response variable from the training data file
minValue: Minimum valid output value
maxValue: Maximum valid output value
numSamps: Number of points to be randomly sampled. Enter -1 to use all sample points. It may be necessary to use a subset of the sample points to avoid memory problems.
regType: Type of regression to be applied: 1 = Theil-Sen Siegel repeated medians, 2 = linear, 3 = linear with intercept of 0
dispGraphs: Logical value to specify if the regression graphs should be plotted. Use "TRUE" or "FALSE" (without quotes).

Notes: When specifying the directory path in R for a Windows computer it is necessary to use a double-backslash ("\\") instead of a single backslash ("\") or you can use a forward slash ("/") which will also work on Apple and Linux operating systems. For example the directory path: C:\Data would be typed C:\\Data or C:/Data.

R is case sensitive. In other words "a" is different from "A". Make sure the directory path and file name is exactly as it appears when you list the directory contents. Also, note that in R variable or file names should not start with a number or most special characters. It is best to start variable names with an upper or lower-case letter.

Other parts of the script can also be modified but you will need to have a good understanding of how the different commands work. The script file must be saved as an ASCII text file.

To run the script in Windows using the interface that comes with R click on the window where the script is displayed to make the window active and then from the R menu select Edit => Run all. This will execute the script and progress can be monitored in the R Console window. Another option (required when using Linux) is to launch the script from the R Console using this syntax: source("path and filename"). For example, type in the following text: source("C:\\IPY\\R_scripts\\rf_percent_cover_dual.R").

Remember to use double backslash instead of single backslash in the directory path unless you are running the script in Linux in which case you would use a single forward slash.

Appendices:

A – Citations and license information

If you cite this document we ask that you include the following information:
Horning, N. 2011. Training Guide for adjusting random forests regression images. American Museum of Natural History - v1, Center for Biodiversity and Conservation. Available from http://biodiversityinformatics.amnh.org/. (accessed on *the date*).

B – Acknowledgements

We would also like to thank the John D. and Catherine T. MacArthur Foundation for supporting this effort.