This README describes the general outline of the code accompanying the Blackford et al. paper "Digital Soil Mapping workflow for forest resource applications: A case study in the Hearst Forest, Ontario" (currently in preprint). While data sharing agreements prohibit the authors from sharing the data used in the Hearst case study, this code can hopefully be a useful guide for others who want to perform Digital Soil Mapping analysis. Many of the objects, arguments, variable names etc. are specific to our Hearst Forest case study and as such they will need to be modified if applying this code to other analyses.

The code is provided "as is", without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and noninfringement. In no event shall the authors or copyright holders be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of or in connection with the software or the use or other dealings in the code.

For questions or concerns, please contact christopher.blackford@canada.ca

Code structure:

The R code is numbered (1 – 8) and is meant to be run in that order. The R file names are meant to mirror the DSM workflow in Figure 1. If R files share the same number, that means it makes no difference which file is run first.

The R code assumes the user will place their environmental and soil data in a folder titled "Raw" and create subfolders within the "Raw" folder for the different data layers. For the Hearst case study, within the "Raw" folder, we had folders titled "Biota_Raw", "DEM_Raw", "Geology_Raw", "Hydrology_Raw", "Response_data", and a SAGA GIS folder for the biotic, elevation, geology, hydrology, soil data, and the SAGA GIS program respectively:
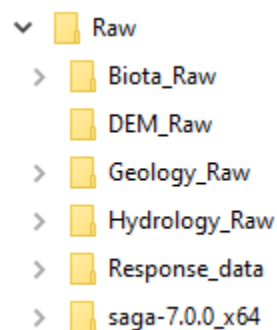
The directory structure for the raw data files

The Raw folder is located at the same directory level as all the code. Again, the user can alter their directory structure if they wish, this is explained here so that the directories in the code files make sense. The "Raw" folder (or it's equivalent) is the only folder the user will need to

setup. The R code provided will automatically create new folder directories that it will populate with DSM outputs.

The SAGA GIS program can be downloaded from "http://www.saga-gis.org/en/index.html" (see preprint manuscript for citation).

User inputs:

In every R file, the user will have to modify the beginning of the code. For example, the file name of the raw DEM we were using was called "10DEM", with "10" referring to the resolution of the DEM in metres. However, we had multiple DEM files in our folder so we needed to identify we were using the "10DEM" file by setting the object "Raw_DEM_reso" = 10. As mentioned above, this line may not be needed for other projects depending on the file directory, naming structure, and the types of analyses the user is interested in. Below is the list of common user inputs we used for this project and what each represent:

R object	Value we used	Explanation
Raw_DEM_reso	10	The resolution of the DEM we used
Filter_type	100	Filter window radius size, in raster cells
env	"./Raw/saga-7.0.0_x64"	The directory path to the SAGA GIS folder
DEM_subtraction_value	30	The elevation (in metres) we use to "burn-in" the DEM for hydrological metric calculations
Create_prob_raster	TRUE / FALSE	Boolean defining if probabilistic rasters should be created
MLM_abbrev	"RF" / "SVM" / "kNN"	Abbreviation of Random Forest, Support Vector Machine, and k-Nearest Neighbor to define which machine learning model output is being analysed.
Attribute	"M" / "T"	Abbreviation of Moisture Regime and Textural Class used to define which predicted soil attribute map should be created