DEPARTMENT OF TRANSPORTATION

Preparing Data for Modeling

MnModel Phase 4 User Guide: Appendix B

Hobbs, Elizabeth, Andrew Brown, Alexander Anton, Jeffrey Walsh, Carson Smith, and Luke Burds

June 24, 2019

MnModel was financed with Transportation Enhancement and State Planning and Research funds from the Federal Highway Administration and a Minnesota Department of Transportation match.

Contents

Introduction

This Appendix to the MnModel Phase 4 User Guide documents the steps necessary to prepare the archaeological and environmental data for statistical modeling in R. These procedures are necessary to ensure data used in the modeling process are not only accurate and understandable, but are also in the required format for the modeling programs.

Archaeological Data

Site Polygons

Digitize Site Polygons

MnDOT began collecting archaeological site shapefiles from Contractors in 2000. In 2012, for MnModel Phase 4, Contractors digitized sites mapped to USGS topo sheets by the State Historic Preservation Office (SHPO) from site forms. This was followed by a project to digitize all remaining site boundaries from site forms, most of which were on file at SHPO and the remainder were found at the Office of the State Archaeologist (OSA). U.S. Forest Service (USFS) sites data from both Chippewa National Forest (CNF) and Superior National Forest (SNF) were subsequently appended to the SHPO/OSA site database. Specific modeling fields pertaining to site description, function, and archaeological tradition were populated based on the recorded database information.

Perform Quality Control on Site Polygons

Prior to sampling, perform quality control procedures on the sites database to remove duplicate and overlapping polygons. Also remove site types that are likely to introduce 'noise' into the model. For MnModel Phase 4, rock art, rock alignment, rock shelter, and quarry sites were not used in modeling as their location depends on lithic sources, for which we have no geospatial data. Single artifact sites were excluded because their more random nature may lead to the introduction of noise in the model. Sites lacking a prehistoric component were not used because this is a model of prehistoric site distributions. Historic documents provide better information about human activity and locations in the Historic period than do the variables used for the predictive model. We also excluded locations ('Alpha sites') not yet officially confirmed as archeological sites by the Office of the State Archaeologist.

Prior to modeling, check all of the sites that were excluded from your sample and make sure they are not there by mistake. Automated procedures may have erroneously classified some sites as anomalies if they had historic era trading posts or trails in addition to prehistoric artifacts.

First flag all sites with duplicate polygons. To do this, add a field DUPLICATE to the sites shapefile/feature class. In the site attribute table, sort sites by site number and mark each duplicate polygon record with a '1' in the DUPLICATE field. This is not strictly necessary but helps keep track of the duplicate polygons that need to be addressed. Once all duplicate polygons are flagged, use either the **Merge Selected Sites** or **Delete Selected Sites** tools (Brown et al. 2019). Delete or merge polygons until only one remains. The remaining polygon should most closely represent the site as mapped and described on its corresponding state site form.

After quality control for archeological sites is complete, copy the sites feature class to \MNMODEL4\STATE\SAMPLE\SAMPLE.GDB\SITESAMP. This feature class will be clipped by each modeling region boundary during the next step.

Prepare Regional Datasets

The **Sample Region** tool (Brown et al. 2019) will clip the statewide archaeological sites polygon feature class (\MNMODEL4\STATE\SAMPLE\SAMPLE.gdb\SITESAMP) by the region's boundary (not the buffered boundary) to create the regional feature class (\MNMODEL4\REGIONS*REG*\SAMPLE\SAMPLE_*REG*.gdb\SITESAMP).). All fields in SITESAMP besides SITENUM will be deleted by the Sample Region tool before sampling. The sampling procedure is documented later in this report.

Survey Polygons

Digitize Survey Polygons

MnDOT began collecting survey boundary shapefiles from Contractors in 2000. These are the most accurate survey polygons available. From 2012 to 2014, for MnModel Phase 4, Contractors evaluated survey reports in the SHPO and OSA offices. If the reports used modern survey standards and if the mapping of the survey boundaries was good, these were digitized into this shapefile. For these latter surveys, attribute data were obtained from SHPO's archaeological survey report database. Beginning in 2014, for MnModel Phase 4, Contractors digitized surveys from various hardcopy sources with more concern about recording a large number of surveys and less concern about those surveys meeting modern survey standards. This database incorporates information from records at both the Minnesota Historic Preservation Office and the Minnesota Office of the State Archaeologist. USFS survey data from both Chippewa National Forest (CNF) and Superior National Forest (SNF) were subsequently appended to the SHPO/OSA survey polygons database. For both CNF and SNF survey lines, a 10-meter buffer was applied based on the maximum width of line transects reported by MnDOT contractors.

MnDOT coded surveys by levels of confidence. Confidence ratings were based on a combination of the confidence that the survey boundary was accurately mapped and confidence that the entire survey polygon was actually surveyed using modern survey standards. These confidence ratings were used to determine which survey polygon to keep when polygons overlap. They could also be used in the future to create models of only the surveys with the highest confidence levels.

Remove Water Bodies from Survey Polygons

Survey polygons are often generalized depictions of where surveyors performed field work. Some are simply boundaries of project areas, rather than the actual area surveyed. Although these polygons often overlap lakes, rivers, and wetlands, it is unlikely that these features were surveyed for archaeological sites. We recommend removing surface water features from the survey polygons prior to modeling. Unfortunately, this step was missed in the MnModel Phase 4 procedures.

Edit Survey Polygons

Survey polygons are often very large, and some surveys overlap with others. Both of these characteristics were potentially problems for sampling and modeling. To adequately represent local environments within surveys, very large polygons (.4,000,000 m^2) were split to represent a sampling interval of 2,000 m. Where survey polygons overlapped, only the polygon with the highest confidence value was kept. Both of these tasks are accomplished by running the tool **Split Large Survey Polygons** tool (Brown et al. 2019). This creates the regional \\MNMODEL4\STATE\SAMPLE\SAMPLE.gdb\SURVSAMP feature class. Note that the **Create Statewide Fishnet** tool must be run prior to splitting polygons.

Predictor Points

To apply model values to all of the cells in a modeling region, you will need a set of 'prediction points.' The **Generate State Prediction Points** tool (Brown et al. 2019) will generate a 30-m resolution point grid with x,y coordinates for the entire state. The **Sample Regions** tool (Brown et al. 2019) will clip the statewide prediction points by the region boundary, sample each variable at each point, and export the data to a text file. Users may also run the **Generate Regional Prediction Points** tool before sampling regions but this is not strictly necessary. The sample text file will be imported into R and, when the statistical model has been completed, R will attach model 'predictions' to the records. These predictions will then be exported as a text file and imported into ArcGIS as a raster, thanks to the x, y coordinates in the file. This raster becomes the GIS version of the predictive model.

Environmental Data

Terrain

Terrain data are necessary for predictive modeling. MnModel Phase 4 obtained high quality statewide digital terrain data from LiDAR. Since none of the infrastructure and disturbance features detectable with these data had any bearing on archaeological site locations, we attempted to minimize the potential negative effects of these features on our terrain variables. To achieve this, we created a ten meter resolution 'conditioned' DTM from the LiDAR data with the goal of restoring as much as possible of the pre-disturbance land surface (Hobbs, 2019b; Hobbs, Walsh, and Hudak 2019).

Soils

All soil variables for MnModel Phase 4 were extracted from 2017 gSSURGO data. These data are available for most of Minnesota from the Natural Resources Conservation Service (NRCS). Even where soils data are present, there are many gaps in coverage. These include missing variable values within water bodies, disturbed areas (e.g. gravel pits or mines), and urban areas. Some variables simply were not reported for all map units. In some cases, missing data can be extracted from map unit names or other text fields. However, the extent of missing attribute data affected which variables we could use for modeling. We supplemented the gSSURGO data with drainage and productivity indices provided by Michigan State University (Schaetzl et al. 2009).

The gSSURGO database provides a mapunit table that aggregates selected soil attributes by soil mapunit. Many more attributes are not aggregated, but are presented in tables by soil components and soil horizons that require many-to-one joins to the mapunit table (and hence to the GIS data). We developed Python tools (Brown et al. 2019) to aggregate these data by determining the values occupying the largest percentage of the mapunit.

Geomorphology

Digital geomorphic data for parts of Minnesota are available from several sources at scales ranging from 1:24,000 to 1:100,000. Prior to MnModel Phase 4, only the most generalized data were available for the entire state. The MnModel Landscape Model is the result of the MnModel Phase 4 project's reclassification and mosaicking of Minnesota Department of Natural Resources (MnDNR), Minnesota Geological Survey (MGS), and MnDOT derived regional and local surficial geology and geomorphic data. The Landscape Model provides the highest resolution data available for any given area as well as a consistent hierarchical classification (Hobbs, 2019b; Hobbs, Walsh, and Hudak 2019).

Hydrography

Hydrographic variables for MnModel Phase 4 were derived from the MnModel Phase 4 Historic/Prehistoric Hydrographic Model (Hobbs et al. 2019). This model is based primarily on Public Land Survey Data for the historic period, though these historic data were augmented as necessary by modern data. For the prehistoric period, the Public Land Survey and soils data are the key sources.

Public Land Survey Data

Public Land Survey plant maps were scanned and georeferenced by the Minnesota Geospatial Information Office and mosaicked by MnDOT for this project. MnDOT digitized hydrographic features from the digital mosaic..

Lakes and major rivers from these maps were used directly in the Hydrographic Model after reviewing and selecting those that best represented the historic features. A ‘USE’ field was added to the feature attribute table to code this selection. In many cases, it was necessary to separate river polygons from lake polygons. Rivers were used if they appeared to follow a natural path, were within the historic floodplain, or were consistent with visible incised meanders in the terrain model. PLS rivers were preferred if their course met these criteria and was different or narrower than that of rivers found in the modern surface hydrography dataset, particularly if the modern river was coded as an impounded or altered stream. Rivers that are wider in the modern data are typically reservoirs. In cases where rivers from the PLS dataset were nearly identical to those in the modern hydrography dataset, modern rivers were chosen due to the more precise delineation of their boundaries. PLS features were most often used for larger rivers, such as the Minnesota and Mississippi Rivers, as surveyors mapped the banks of larger rivers more accurately than smaller rivers and streams.

Wetlands were taken directly from the MnModel Phase 4 Historic Vegetation Model (Hobbs 2019a), as wetland boundaries on the plat maps are more imaginary than realistic

Modern Surface Water

Phase 3 of MnModel clearly suffered from the use of modern hydrographic data that fails to represent the many historic lakes and wetlands that have been drained. For MnModel Phase 4, modern lakes and rivers were used only as necessary to supplement the historic data from the Public Land Survey.

Several sources of modern hydrographic data were combined to create MnModel Phase 4's modern water feature class. A Minnesota Surface Water Inventory (SWI) derived from the National Wetlands Inventory (NWI) data were acquired from USFWS and reclassified to wetland types consistent with Aaseng (1993) to the extent possible. No other available data sources provided sufficient attribute information to support this level of classification, so NWI was considered the most authoritative source. These made up about 91 percent of the final feature class. However, other data sources mapped bodies of water, some of them quite significant, that were missing from NWI. To make sure that we were working with the most comprehensive data possible, MnDNR water and wetlands (about two percent of the total) and National Hydrography Dataset (NHD) polygons (about six percent of the total) were combined and used to supplement NWI. The data sources were found to be inconsistent with each other in their classification of water and wetlands. Some useful attributes were extracted from gSSURGO and used to incorporate additional information into the dataset to support classification of non-NWI features. Linear streams data from DNR were also used. Finally, MnDNR's Native Plant Communities dataset (NPC) was found to contain wetland areas not mapped by other sources, so it was also incorporated into the feature class. For this model, modern water polygons or portions thereof were eliminated if they were artificial or if they exhibited a combination of soil drainage and topographic characteristics implying they were in improbable locations for historic hydrographic features.

Because the lakes and rivers depicted on the PLS plat map were neither a complete census of the features present historically, nor were they completely accurate, it was necessary to use some modern hydrographic data as model input. The challenge was to determine which portions of the modern hydrographic database were useful. A minimum area of 12,141 square meters was established. Any polygon smaller than this minimum was either merged into an adjacent, larger polygon, or deleted.

A 'USE' field was added to the feature attribute table to code features that were needed as model input. Natural, unaltered lakes and rivers were selected for use:

- When the terrain data clearly showed that they were more accurately mapped than the same PLS lake.
- They were not represented in the PLS data, particularly if they did not intersect a section line.

Streams and rivers were used if they were not impounded, diverted, or ditched and were better representations of the river course than the corresponding PLS river or stream. Decision rules used for coding rivers and streams are documented in Table 1. In addition, rivers were selected from the MnModel Phase 4 Landscape Model if they met the same criteria and provided a better representation of the historic river than all other sources.

Table 1: Decision Rules for Coding Rivers for Use in Hydrographic Model

River Type	Characteristics	USE Value
Reservoir River	Looks like a lake; STREAMS data STREAM LIKE '%IMPOUNDED' OR STREAM LIKE '%CONNECTOR%'	0
Diverted River	Very different course from GLO river; may be straighter than GLO river; STREAMS data STREAM LIKE '%ALTERED%	0
Ditched River	Very straight, may be different course than GLO; STREAMS data STREAM LIKE '%CHANNELED%'	0
Natural River	Looks like a river; course very close to GLO river, especially where crossing section lines; may be narrower than GLO river; within deeply incised meanders; STREAMS data STREAM LIKE '%NATURAL% OR STREAM LIKE '%CENTERLINE%'	1

Vegetation

Historic vegetation was modeled for MnModel Phase 4 using Public Land Survey data as the source (Hobbs 2019a). Surveyors' vegetation observations and bearing trees were extracted to section and quarter section corner points and published by MnDNR in 1997. In a separate effort, John Almendinger (MnDNR) extracted survey notes to section lines for the northern half of the state and made these available to MnDOT. In 2013, MnDOT created a statewide mosaic of the scanned and georeferenced Public Land Survey plat maps and digitized polygons of hydrographic and vegetation features.

To create the historic vegetation dataset used for modeling, MnDOT reclassified the DNR point data, using the bearing trees, line notes, and plat maps as supporting evidence. These reclassified data points were modeled in R using much the same procedures as the archaeological predictive model (Landrum and Hobbs 2019).

Pedestrian Transportation

MnModel Phase 4 From Everywhere To Everywhere (FETE) Model (Hobbs 2019b) was developed by Devin White at Sandia National Laboratories for MnModel Phase 4. The model is based on White and Barber's (2012) models of pedestrian transportation networks in Mexico. Values in the model indicate the number of paths that cross each cell when least-cost paths are calculated from every cell to every other cell. Paths suggested by this model were grouped into major, intermediate, and minor path categories based on their weights (Hobbs, Walsh and Hudak 2019).

Modeling Mask

Masks are used in ArcGIS to restrict procedures to specified portions of the total geographic extent. For MnModel, statistical procedures will not run on data with NULL values. We use a mask to denote where soil variables are most likely to have NULL values. The statistical analyses for the predictive models will not run if there are any NULL values in the data. This will require us to run one version of each model, using soil variables, for locations that have no NULL values and a separate version without soil variables for the entire region. The final model will be a composite of the two models.

The statewide modeling mask for MnModel was created from gSSURGO data using the custom **Create MODMASK** tool (Brown et al. 2019). It is found at: \MNMODEL4\STATE\DATA\BOUNDARIES\BOUNDARIES.gdb\MODMASK (raster) and MODMASKPL (polygons). Both masks should be clipped for each region using the region's buffered boundary. MODMASK values are defined in Table 2. The regional feature classes will be \MNMODEL4\REGIONS*REG*\DATA\BOUNDARIES\BOUNDARIES_*REG*.gdb\MODMASK (raster) and MODMASKPL (polygons).

Table 2: MODMASK Values

VALUE	Definition
-999	NULL (not masked)
1	No Soils Data
2	Modern Water Body
3	Intermittent Water
4	Constructed Pond
5	Cut/Fill/Disturbed
6	Mine/Pit
7	Dump/Stockpile
8	Built Feature
9	Urban Land

Predictor Variables

Predictor variables are aspects of the environment that are thought to be important to prehistoric human location decisions. These are derived from the source data and organized into geodatabases so that they can be sampled for modeling. Predictor variables are more completely documented in Hobbs, Walsh and Hudak (2019).

Create Regional Variable Geodatabases

Variables for MnModel Phase 4 are derived for each modeling region by several custom Python tools (Brown et al. 2019) and are organized by topic within geodatabases. These variable geodatabases should all reside in the \MNMODEL4\REGIONS*REG*\VARIABLES folder. All variable grids extend to the 10 km buffer around the region's boundary to avoid NULL values when sampled by points near the region's boundary.

TERRVARS_*REG*.gdb

TERRVARS_*REG*.gdb contains a region's terrain variables (Table 3). The variables and the geodatabase are created by running the **Create Terrain Variables** tool (Brown et al. 2019). This tool clips ELEV, TWI, and VISIBLE from statewide variable rasters that reside in the \MNMODEL4\STATE\VARIABLES\TERRVARS.gdb geodatabase and creates the remainder of the variables for the region only.

Table 3: Terrain Variables in Each Region's TERRVARS_*REG*.gdb

VARIABLE	DEFINITION
ASP_RNG	Aspect classified by range breaks
ASPECT	Aspect (this is needed to create ASP_RNG)
CURV	Surface Curvature
ELEV	Elevation
REL	Relative Elevation
REL90	Relative Elevation within 90 meters
RGH	Surface Roughness
RGH90	Surface Roughness within 90 meters
SHELTER	Shelter Index
SLOPE	Percent Slope
TPI1000	Topographic Position Index within 1000 meters

VARIABLE	DEFINITION
TPI1MI	Topographic Position Index within one Mile
TPI250	Topographic Position Index within 250 meters
TPI5MI	Topographic Position Index within five miles
TPI90	Topographic Position Index within 90 meters
TWI	Topographic Wetness Index
VISIBLE	Visibility

SOILVARS_*REG*.gdb

SOILVARS_*REG*.gdb contains the region's soil variables (Table 4). The variables and the geodatabase are created by running the **Create Soil Variables** tool (Brown et al. 2019). Soil variables reside in various tables that the tool joins to the gSSURGO mapunit polygon feature class, then clips and converts to 30 m rasters. Note that some of these variables are not on the list of predictor variables for the archaeological predicted models (Table 9 below) as they are used only for developing the Historic Vegetation Model (Hobbs 2019a).

Table 4: Soil Variables in Each Region's SOILVARS_*REG*.gdb

VARIABLE	DEFINITION
AWS150	Available Water Storage (cm) in the top 0-150 cm of the soil column (weighted average)
CACO3	Percent calcium carbonate in the surface horizon
CEC7	Cation Exchange Capacity of the surface horizon
CLAY	Percent clay in the surface horizon
DI	Drainage Index (Schaetzl et al. 2009)
DRAIN	Soil drainage

VARIABLE	DEFINITION
FFD_R	Frost-free days
FLDFRQD	Flood frequency
GRTGRP	Soil Great Group taxonomic class
GYPSUM	Percent gypsum in the surface horizon
HYDGRPDCD	Hydric Group (dominant condition)
HYDPRS	Hydric soil presence (percentage of the mapunit that is hydric)
HZDEP	Depth of surface soil horizon
OM	Percent organic matter in the surface horizon
PI	Productivity Index
REG_RICH	Ecological habitat richness (Schaetzl et al. 2009)
REG_WET	Ecological habitat moisture regime (Schaetzl et al. 2009)
SAND	Percent sand in the surface horizon
SILT	Percent silt in the surface horizon
WETSOIL	On a wetland soil

LANDVARS_*REG*.gdb

Geomorphic variables are contained in LANDVARS_*REG*.gdb (Table 5), which is created by the **Create Landscape Variables** tool (Brown et al. 2019). The tool extracts two of these variables, LFORM (landform) and LSCAPE (landscape), from the MnModel Phase 4 Landscape Model (Hobbs 2019b). Watershed sizes are measured from the MnDNR major and minor watershed boundaries. These same boundaries are used to represented 'ridges' for the path distance variables. Finally, the 'islands' feature class used was extracted from hydrographic data derived from NWI and published by MnDNR.

Table 5: Geomorphic Variables in Each Region's LANDVARS_*REG*.gdb

VARIABLE	DEFINITION
CP_MAJRIDGE	Path distance to nearest major ridge or divide
CP_MINRIDGE	Path distance to nearest minor ridge or divide
ISLAND	On an island
LFORM	Landform
LSCAPE	Landscape
MAJ_SIZE	Size of major watershed
MIN_SIZE	Size of minor watershed

VEGVARS_*REG*.gdb

There are three types of vegetation variables (Table 6): the dominant type of vegetation at the location of a site, the diversity of vegetation types within one, five, and ten-kilometer distances from a site, and the cost-path distance to wild rice resources. In addition, there is a resistance variable associated with vegetation type that is used in all least-cost path calculations. All vegetation variables except distance to wild rice are derived from the MnModel Phase 4 Historic Vegetation Model (Hobbs 2019a), and all are created by the **Create Vegetation Variables** tool (Brown et al. 2019).

Table 6: Vegetation Variables in Each Region's VEGVARS_*REG*.gdb

VARIABLE	DEFINITION
CP_RICE	Path distance to nearest wild rice location
RESIST	The estimated impedence or 'cost' of walking through a vegetation type. This value is used as input to the cost-path variables.
VEGDIV10K	Vegetation diversity within ten km

VARIABLE	DEFINITION
VEGDIV1K	Vegetation diversity within one km
VEGDIV5K	Vegetation diversity within five km
VEGMOD	Historic vegetation type

HYDVARS_*REG*.gdb

Hydrographic variables are contained in the HYDVARS_*REG.*gdb (Table 7). Except for Order of Nearest Stream (ORD_STRM), Path Distance to Nearest Intermittent Stream (CP_INT) and Path Distance to Nearest Perennial Stream (CP_PEREN) all hydrographic variables are last-cost path distances to various types of surface hydrographic features, either historic or prehistoric, from the MnModel Phase 4 Hydrographic Model (Hobbs et al. 2019). The hydrographic variables are created by the **Create Hydrographic Variables** tool (Brown et al. 2019).

Table 7: Hydrographic Variables in Each Region's HYDVARS_*REG*.gdb

VARIABLE	DEFINITION
CP_BOG	Path distance to nearest historic bog
CP_FLOOD	Path distance to nearest historic floodplain
CP_INT	Path distance to nearest intermittent stream
CP_LAKE	Path distance to nearest historic lake
CP_LLK	Path distance to nearest large historic lake
CP_MARSH	Path distance to nearest historic marsh
CP_MEADOW	Path distance to nearest historic wet meadow or fen
CP_PEREN	Path distance to nearest perennial stream
CP_PFLOOD	Path distance to nearest prehistoric floodplain

VARIABLE	DEFINITION
CP_PLAKE	Path distance to nearest prehistoric lake
CP_PLLK	Path distance to nearest large prehistoric lake
CP_PWET	Path distance to nearest prehistoric wetland
CP_RIVER	Path distance to nearest river
CP_SWAMP	Path distance to nearest historic swamp
CP_WAT	Path distance to nearest historic surface water (of all types)
CP_WET	Path distance to nearest historic 'wet' land
CP_WETLAND	Path distance to nearest historic wetland (of any type)
ORD_STRM	Order of nearest stream

FETEVARS_*REG*.gdb

FETEVARS_*REG*.gdb (Table 8) contains pedestrian transportation variables, derived from the FETE model (Hobbs 2019b). These include cost-path distances to each of these categories of paths plus the 'order' or weight of the nearest path. These are created by the **Create FETE Variables** tool (Brown et al. 2019).

Table 8: Pedestrian Transportation Variables in Each Region's FETEVARS_*REG*.gdb

VARIABLE	DEFINITION
CP_MAJPATH	Path distance to nearest major pedestrian transportation route
CP_MEDPATH	Path distance to nearest medium pedestrian transportation route
CP_MINPATH	Path distance to nearest minor pedestrian transportation route
FETE	Classified version of the FETE model, with the following codes: • 1 = Minor path

VARIABLE	DEFINITION
	• 2 = Medium path • 3 = Major path
PATH_ORD	Order of nearest pedestrian transportation route

Check Variables For Region

Prior to modeling, perform quality control on the variables needed for the model you are running:

1. Check each variable geodatabase for your region to determine the following:

    a. Are all the variables present in your region? Even if a variable contains only NULL values, it should have been created. If not, run the appropriate variable generation tool. Variables with all NULL values will be deleted before exporting the data to R. Be sure to check if each variable raster exists in the appropriate geodatabase (Tables 3-8 above)**.**

    b. Do any variable rasters contain large numbers of NULL cell values?

        i. Valid NULL values in the soils data should have been coded as '-999'. The modeling mask should take care of most of these instances, but there may still be some variables in some counties for which soil scientists did not record values for all polygons. If they are present, they can be handled in the R script. No other variables should contain -999 values (with the exception of SHELTER, for which this may be a valid value).

        ii. There should be no cells within any dataset of NULL values that have not been coded as '-999'. If you find variables in your region with NULL values, especially if they are not on the region's border, there is a problem with your data. Determine what caused the problem when the variable raster was created and create a corrected version.

    c. Is the variable an integer grid? Note that there should be no floating point raster values, as these increase both storage and processing requirements.

    d. Does your variable raster contain valid values as documented in Table 9. The valid values will be discrete numeric values for text variables and a range of integer values for numeric variables. If the raster values do not match the valid value criterion, then the raster is not ready for modeling. Also, make sure the variable naming conventions and file paths are consistent with those expected by the sampling tools (Brown et al. 2019).

2. There are two variable lists for modeling (Table 9): ALLARCHLIST contains variables that are complete for the entire region and SOILARCHLIST contains the same variables plus soils variables, which have NULL values for some cells in every region (Hobbs, Walsh and Hudak 2019). The **Sample Region** tool (Brown et al. 2019) pulls the variables directly out of TERRVARS_*REG*.gdb, SOILVARS_*REG.*gdb, LANDVARS_*REG*.gdb,

etc. and creates sample files for each version, using the 'ALL' and 'SOIL' identifiers to distinguish between them.

a. ALLARCHLIST: Variables that are sampled for the entire region, including areas where soils data are absent. This list includes variables that can be sampled for ALL predictive points in the region. There can be no NULL or -999 values.

   - Variables include: ASP_RNG, CP_BOG, CP_FLOOD, CP_INT, CP_LAKE, CP_LLK, CP_MAJPATH, CP_MAJRIDGE, CP_MARSH, CP_MEADOW, CP_MEDPATH, CP_MINPATH, CP_MINRIDGE, CP_PEREN, CP_PFLOOD, CP_PLAKE, CP_PLLK, CP_PWET, CP_RICE, CP_RIVER, CP_SWAMP, CP_WAT, CP_WET, CP_WETLAND, CURV, ELEV, ISLAND, LFORM, LSCAPE, MAJ_SIZE, MIN_SIZE, ORD_STRM, PATH_ORD, REL, REL90, RGH, RGH90, SHELTER, SLOPE, TPI1000, TPI1MI, TPI250, TPI500, TPI5MI, TPI90, TWI, VEGDIV10K, VEGDIV1K, VEGDIV5K, VEGMOD, VISIBLE, WETSOIL
   - Sample files created include: ALLPREDARCH, ALLSITE, ALLSURV

b. SOILARCHLIST: All available variables, including those soil variables with NULL values in some places. These must be sampled only for portions of the region where soils data are not absent (areas where MODMASK IS NULL).

   - Variables include: ASP_RNG, CP_BOG, CP_FLOOD, CP_INT, CP_LAKE, CP_LLK, CP_MAJPATH, CP_MAJRIDGE, CP_MARSH, CP_MEADOW, CP_MEDPATH, CP_MINPATH, CP_MINRIDGE, CP_PEREN, CP_PFLOOD, CP_PLAKE, CP_PLLK, CP_PWET, CP_RICE, CP_RIVER, CP_SWAMP, CP_WAT, CP_WET, CP_WETLAND, CURV, DI, DRAIN, ELEV, FFD_R, FLDFRQD, HYDGRPDCD, HZDEP, ISLAND, LFORM, LSCAPE, MAJ_SIZE, MIN_SIZE, ORD_STRM, PATH_ORD, PI, REL, REL90, RGH, RGH90, SHELTER, SLOPE, TPI1000, TPI1MI, TPI250, TPI500, TPI5MI, TPI90, TWI, VEGDIV10K, VEGDIV1K, VEGDIV5K, VEGMOD, VISIBLE, WETSOIL
   - Sample files created include: SOILPREDARCH, SOILSITE, SOILSURV

Note that the final tables for the sample points and predictive points for each model run must contain exactly the same variables, the names of the variables must match (R is case-sensitive), and it is also be necessary to keep the variables in the same order. Using the **Sample Region** tool should insure this.

Table 9: Complete List of MnModel Phase 4 Predictor Variables

Variable	Definition	Type	Valid Values	ALLARCHLIST	SOILARCHLIST
ASP_RNG	Aspect range	Categorical	1-9	X	X
CP_BOG	Path distance to nearest historic bog	Integer	Positive Integers	X	X
CP_FLOOD	Path distance to nearest historic floodplain	Integer	Positive Integers	X	X

Variable	Definition	Type	Valid Values	ALLARCHLIST	SOILARCHLIST
CP_INT	Path distance to nearest intermittent stream	Integer	Positive Integers	X	X
CP_LAKE	Path distance to nearest historic lake	Integer	Positive Integers	X	X
CP_LLK	Path distance to nearest large historic lake	Integer	Positive Integers	X	X
CP_MAJPATH	Path distance to nearest major pedestrian transportation route	Integer	Positive Integers	X	X
CP_MAJRIDGE	Path distance to nearest major ridge or divide	Integer	Positive Integers	X	X
CP_MARSH	Path distance to nearest historic marsh	Integer	Positive Integers	X	X
CP_MEADOW	Path distance to nearest historic wet meadow or fen	Integer	Positive Integers	X	X
CP_MEDPATH	Path distance to nearest medium pedestrian transportation route	Integer	Positive Integers	X	X
CP_MINPATH	Path distance to nearest minor pedestrian transportation route	Integer	Positive Integers	X	X
CP_MINRIDGE	Path distance to nearest minor ridge or divide	Integer	Positive Integers	X	X
CP_PEREN	Path distance to nearest perennial stream	Integer	Positive Integers	X	X
CP_PFLOOD	Path distance to nearest prehistoric floodplain	Integer	Positive Integers	X	X
CP_PLAKE	Path distance to nearest prehistoric lake	Integer	Positive Integers	X	X
CP_PLLK	Path distance to nearest large prehistoric lake	Integer	Positive Integers	X	X
CP_PWET	Path distance to nearest prehistoric wetland	Integer	Positive Integers	X	X
CP_RICE	Path distance to nearest wild rice location	Integer	Positive Integers	X	X
CP_RIVER	Path distance to nearest river	Integer	Positive Integers	X	X
CP_SWAMP	Path distance to nearest historic swamp	Integer	Positive Integers	X	X
CP_WAT	Path distance to nearest historic surface water (of all types)	Integer	Positive Integers	X	X
CP_WET	Path distance to nearest historic ‘wet’ land	Integer	Positive Integers	X	X
CP_WETLAND	Path distance to nearest historic wetland (of any type)	Integer	Positive Integers	X	X
CURV	Surface Curvature	Integer	Integers	X	X
DI	Drainage Index	Integer	0-99		X
DRAIN	Soil drainage	Categorical	1-8		X
ELEV	Elevation	Integer	Integers	X	X
FFD_R	Frost-free days	Integer	0-365		X
FLDFRQD	Flood frequency	Categorical	0-5		X

Variable	Definition	Type	Valid Values	ALLARCHLIST	SOILARCHLIST
HYDGRPDCD	Hydric Group (dominant condition)	Categorical	1-7		X
HZDEP	Depth of surface soil horizon	Integer	Positive Integers		X
ISLAND	On an island	Categorical	0-1	X	X
LFORM	Landform	Categorical	10-98	X	X
LSCAPE	Landscape	Categorical	10-27	X	X
MAJ_SIZE	Size of major watershed	Integer	Positive Integer	X	X
MIN_SIZE	Size of minor watershed	Integer	Positive Integer	X	X
ORD_STRM	Order of nearest stream	Categorical	1-14	X	X
PATH_ORD	Order of nearest pedestrian transportation route	Categorical	1-3	X	X
PI	Productivity Index	Categorical	0-18		X
REL	Relative Elevation	Integer	Integers	X	X
REL90	Relative Elevation within 90 meters	Integer	Integers	X	X
RGH	Surface Roughness	Integer	Integers	X	X
RGH90	Surface Roughness within 90 meters	Integer	Integers	X	X
SHELTER	Shelter Index	Integer	Integers	X	X
SLOPE	Percent Slope	Integer	Integers	X	X
TPI1000	Topographic Position Index within 1000 meters	Integer	Integers	X	X
TPI1MI	Topographic Position Index within one Mile	Integer	Integers	X	X
TPI250	Topographic Position Index within 250 meters	Integer	Integers	X	X
TPI5MI	Topographic Position Index within five miles	Integer	Integers	X	X
TPI90	Topographic Position Index within 90 meters	Integer	Integers	X	X
TWI	Topographic Wetness Index	Integer	Integers	X	X
VEGDIV10K	Vegetation diversity within ten km	Integer	Positive Integers	X	X
VEGDIV1K	Vegetation diversity within one km	Integer	Positive Integers	X	X
VEGDIV5K	Vegetation diversity within five km	Integer	Positive Integers	X	X
VEGMOD	Historic vegetation type	Categorical	See Table 10	X	X
VISIBLE	Visibility	Integer	Positive Integers	X	X
WETSOIL	On a wetland soil	Categorical	0-1	X	X

Table 10: Valid Values for VEGMOD

MODTYPE	VALUE
LAKE	100
WET LAND	150
RIVER	200
BOG	210
CONIFER/SHRUB SWAMP	220
MARSH	230
FLOODPLAIN FOREST	240
HARDWOOD SWAMP	250
WET MEADOW/FEN	270
PINE FOREST	310
JACK PINE FOREST	311
RED PINE FOREST	312
WHITE PINE FOREST	313
SPRUCE-FIR FOREST	321
BLACK SPRUCE-FEATHERMOSS FOREST	322
UPLAND WHITE CEDAR FOREST	323
PINE BARRENS	330

MODTYPE	VALUE
JACK PINE WOODLAND	341
NORTHERN CONIFER WOODLAND	342
BOREAL HARDWOOD-CONIFER FOREST	351
MIXED PINE-HARDWOOD FOREST	352
NORTHERN HARDWOOD-CONIFER FOREST	353
WHITE PINE-HARDWOOD FOREST	354
ASPEN FOREST	361
ASPEN-BIRCH FOREST	362
PAPER BIRCH FOREST	363
LOWLAND HARDWOOD FOREST	364
MAPLE-BASSWOOD FOREST	365
NORTHERN HARDWOOD FOREST	366
OAK FOREST	367
ASPEN OPENINGS	371
OAK SAVANNA	372
ASPEN WOODLAND	381
OAK WOODLAND	382
BRUSH-PRAIRIE	391

MODTYPE	VALUE
PRAIRIE	392
NO DATA	-999

Sample Variables

Sampling procedures are accomplished using the custom **Sample Region** tool (Brown et al. 2019). This tool performs all the necessary steps in the sampling process.

Sample Site/Survey Polygons

The predictor variable value for sites and surveys represents a majority count or mean (Table 11) of raster cell values for each predictor variable falling within the site or survey polygon footprint. As a general rule, variables from polygon predictor variables were sampled using majority. These included soil variables and watershed size, for example). Categorical variables, such as landform and vegetation type, were also sampled using majority. Measures based on small neighborhood variation (RGH90, REL90, TPI90) used majority. Finally, measures of surface characteristics (SLOPE, TWI, CURV) use majority. Only measures based on larger neighborhood variation (RGH, REL, TPI1000, SHELTER) and distance variables (CP_LAKE,CP_RIVER) were sampled using means. This is because the number of potential unique values for these variables could be so large that there may not be a majority value.

Table 11: Statistics Type Used for Sampling Variables

Variable	Definition	Sampling Method
ASP_RNG	Aspect classified by range breaks.	Majority
CP_BOG	Cost-path distance to nearest bog	Mean
CP_FLOOD	Cost-path distance to nearest historic floodplain	Mean
CP_INT	Cost-path distance to nearest historic intermittent stream	Mean
CP_LAKE	Cost-path distance to nearest historic lake	Mean
CP_LLK	Cost-path distance to nearest large historic lake (> 485,640 sq m)	Mean
CP_MAJPATH	Cost-path distance to nearest major path (FETE > 138,400)	Mean

Variable	Definition	Sampling Method
CP_MAJRIDGE	Distance to major ridge or divide	Mean
CP_MARSH	Cost-path distance to nearest historic marsh	Mean
CP_MEADOW	Cost-path distance to nearest historic wet meadow/wet prairie/fen	Mean
CP_MEDPATH	Cost-path distance to nearest medium-sized path (FETE 49,000-138,400)	Mean
CP_MINPATH	Cost-path distance to nearest minor path (FETE 8,900-49,000)	Mean
CP_MINRIDGE	Distance to minor ridge or divide	Mean
CP_PEREN	Cost-path distance to nearest historic perennial river or stream	Mean
CP_PFLOOD	Cost-path distance to nearest prehistoric floodplain	Mean
CP_PLAKE	Cost-path distance to nearest prehistoric lake	Mean
CP_PLLK	Cost-path distance to nearest large prehistoric lake (>485,640 sq m)	Mean
CP_PWET	Cost-path distance to nearest prehistoric wetland	Mean
CP_RICE	Cost-path distance to wild rice	Mean
CP_RIVER	Cost-path distance to nearest historic major river	Mean
CP_SWAMP	Cost-path distance to nearest historic swamp	Mean
CP_WETLAND	Cost-path distance to nearest historic wetland	Mean
CURV	Surface curvature, designating both convex and concave surfaces.	Majority
DI	Drainage Index	Majority
DRAIN	Drainage Class – Interpreted	Majority
ELEV	Surface elevation based on a terrain dataset with man-made features removed and bathymetric data added.	Mean
FFD_R	Frost Free Days - Representative Value	Majority

Variable	Definition	Sampling Method
FLDFRQD	Flooding Frequency Dominant Condition	Majority
HYDGRPDCD	Hydrologic Group – Dominant	Majority
HZDEP	Depth of Surface Horizon	Majority
ISLAND	On an island	Majority
LFORM	Dominant landform	Majority
LSCAPE	Dominant landscapes	Majority
MAJ_SIZE	Size of major watershed	Majority
MIN_SIZE	Size of minor watershed	Majority
ORD_STRM	Stream order of nearest stream (order = 6 or higher)	Majority
PATH_ORD	Order of nearest path (FETE > 8,900)	Majority
PI	Productivity Index	Majority
REL	Relative elevation within a 5-kilometer radius.	Mean
REL90	Relative elevation within a 90-meter radius.	Majority
RGH	Surface roughness based on a 5 kilometer radius of analysis.	Mean
RGH90	Surface roughness based on a 90 meter radius of analysis.	Majority
SHELTER	Shelter index.	Mean
SLOPE	Slope of cell, in degrees.	Majority
TPI1000	Topographic Position Index based on 1-kilometer radius.	Mean
TPI1MI	Topographic Position Index based on 1-mile radius.	Mean
TPI250	Topographic Position Index based on 250-meter radius.	Mean
TPI500	Topographic Position Index based on 500-meter radius.	Mean
TPI5MI	Topographic Position Index based on 5-mile radius.	Mean
TPI90	Topographic Position Index based on 90-meter radius.	Majority
TWI	Topographic Wetness Index.	Majority
VEGMOD	Dominant historic vegetation type	Majority
VEGDIV1K	Number of vegetation types within 1 km	Majority
VEGDIV5K	Number of vegetation types within 5 km	Majority
VEGDIV10K	Number of vegetation types within 10 km	Majority

Variable	Definition	Sampling Method
VISIBLE	Visibility	Majority
WETSOIL	Wet Soils - Interpreted	Majority

After polygons are sampled, the **Sample Region** tool creates points representing the centroids of site and survey polygons and attaches the sampled mean or majority values to these points. If any points have values for all soil variables of '-999' (NULL), they are deleted from the database.

Create and Sample Background Points

Additional points, the 'background points,' are created by the **Sample Region** tool to represent areas where no sites or surveys are present. Separate sets of background points must be created for each region's site and survey models, as the number and spacing of predictor points are proportional to the number of site or survey points in the region. Background points are created as a regular grid with a target ratio of between 2-3:1 background points to sites/surveys. They are arranged in a regularly spaced grid of varying scales, depending on the number of points needed. Background points are not created within 1,500 m of site/survey polygons to make sure they are representing different environments. Sample points are clipped by the region's boundary. These background points are sampled simply as the variable value at the point.

Sample Prediction Points

The **Sample Region** tool also samples variables using the prediction points. Like the background points, prediction point variable values are simply the value of the variable at the point itself.

Find and Remove NULL Values

NULL values, whether they appear in the data tables as NULL or as '-999', cause problems in R. If they appear as NULL, the **Sample Region** tool will remove them prior to modeling, as we do not have reliable procedures for removing them in R. If they appear as '-999', there are R procedures for removing them, and we will rely on those as they are faster.

Combine Site/Survey and Background Points

Finally, the **Sample Region** tool will combine site/survey points with background points and produce several feature classes:

- ALLSITE: All archaeological site points and associated background points
- SOILSITE: Only archaeological site points in the areas where MODMASK indicates soils data are present and their background points.
- ALLSURV: All survey points and their associated background point.

- SOILSURV: Only archaeological survey points in the areas where MODMASK indicates soils data are present and their background points.
- ALLPREDARCH: All prediction points for the region.
- SOILPREDARCH: Only the prediction points in the areas where MODMASK indicates soils data are present.

In addition to these feature classes in the \MNMODEL4\REGIONS*REG*\SAMPLE\SAMPLE_*REG*.gdb, the tool will export corresponding .csv files for input to R. Once you have these files, you are ready to model.

References

Aaseng, Norman E. et al. 1993. Minnesota's Native Vegetation: A Key to Natural Communities. Version 1.5. Minnesota Department of Natural Resources, Natural Heritage Program. St. Paul, MN. 111 pp.

Brown, Andrew, Alexander Anton, Luke Burds, and Elizabeth Hobbs
2019 *Tool Handbook*. Appendix C in *MnModel Phase 4 User Guide*, by Carla Landrum et al. Minnesota Department of Transportation. St. Paul, MN.

Hobbs, Elizabeth
2019a *Historic Vegetation Model for Minnesota: MnModel Phase 4*. Minnesota Department of Transportation. St. Paul, MN.
2019b *MnModel Phase 4: Project Summary and Statewide Results.* Minnesota Department of Transportation. St. Paul, MN.

Hobbs, Elizabeth, Andrew Brown, Alexander Anton, and Luke Burds
2019 *Historic/Prehistoric Hydrographic Models for Minnesota: MnModel Phase 4*. Minnesota Department of Transportation. St. Paul, MN.

Hobbs, Elizabeth, Jeffrey Walsh and Curtis M. Hudak
2019 *Environmental Variables: MnModel Phase 4*. Minnesota Department of Transportation. St. Paul, MN.

Landrum, Carla and Elizabeth Hobbs
2019 *Vegetation Modeling User's Guide: MnModel Phase 4*. Minnesota Department of Transportation. St. Paul, MN.

Schaetzl, Randall J., Frank J. Krist, Jr., Kristine Stanley, and Christina M. Hupy
2009 The natural soil drainage index: an ordinal estimate of long-term soil wetness. *Physical Geography* 30:383-409.

White, Devin A. and Sarah B. Barber
2012 Geospatial modeling of pedestrian transportation networks: a case study from precolumbian Oaxaca, Mexico. *Journal of Archaeological Science 39: 2684-2696.*