DEPARTMENT OF TRANSPORTATION

Environmental Variables

MnModel Phase 4

Elizabeth Hobbs, Jeffrey Walsh, and Curtis Hudak…

June 10, 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

Archaeological predictive modeling assumes a relationship between archeological site locations and a suite of environmental 'predictor' variables. These variables are chosen subjectively by the modelers because they are assumed to have some bearing on why prehistoric peoples chose to live or spend time in specific locations. The choice of variables is limited by available environmental data.

This document defines and evaluates the environmental variables used as predictors for the MnModel Phase 4 archaeological site and survey locational models. Specific instructions for deriving these variables are documented in the MnModel Phase 4 *Tools Handbook* (Brown et al. 2019).

Regionalization

MnModel is a statewide archaeological predictive model for Minnesota. As Minnesota is a large state with a variety of environmental zones, the model is a mosaic of twenty models for individual regions. Regions are defined based on the Ecological Classification System (Cleland et al. 1997; Hanson and Hargrave 1996). More information about the MnModel Phase 4 regions can be found in Hobbs (2019b). Most variables were derived by regions using customized tools, as this was usually most efficient procedure. However, several variables were derived for the entire state. These included:

- The geomorphic variables *Landform* ('LFORM') and *Landscape* ('LSCAPE'), which were simply converted to raster format from attributes in the MnModel Phase 4 Landscape Model.
- *Elevation* ('ELEV'), which was extracted from the statewide conditioned digital terrain model (DTM).
- *Topographic Wetness Index* ('TWI'), which was derived for the entire state by our contractor using TauDEM software.
- *Visibility* ('VISIBLE'), which was also derived by our contractor in smaller processing extents, because of the computational demands, then mosaicked into a statewide raster.

Buffers

Some variables require information from outside of the region or state. For example, if we want to know the distance from a point to the nearest lake, that lake may be in an adjoining region. Likewise, to measure vegetation diversity of a point near a region's boundary, we must also count the vegetation types in the adjacent region. To insure that variables for locations near each region's boundary were accurately measured, we included a ten km buffer zone around each region. The buffered regional boundary was used both for deriving variables and for modeling.

To populate the buffer zone where it extended outside of Minnesota, we acquired terrain data for all surrounding states and Canada. We also acquired soils data for the surrounding states, but were unable to find comparable data for Canada. Entire lakes and rivers for surrounding states were available from National Wetlands Inventory and gSSURGO soils data (Natural Resource Conservation Service). Watershed boundaries

from Minnesota Department of Natural Resources (MnDNR) extend well beyond state boundaries. MnDNR also maps some Minnesota lakes where they extend into Canada. We did not have vegetation data for surrounding states or Canada, but when we built our vegetation models we included the buffer so that the modeled historic vegetation extends into adjacent states. The lack of geomorphic data for surrounding states and Canada did not pose a problem for the predictive models because the variables derived from the landscape model did not require measurements of distance or diversity.

Path Distances

Many of MnModel's predictor variables are based on distances to key resources, such as water bodies. In MnModel Phase 3, we used Euclidean distance measures. With improvements in ArcGIS distance functions and computing power, we were able to use cost-path distances for MnModel Phase 4.

Cost-path distances require the use of a weighted cost surface to calculate impedance values, providing an estimate of the relative difficulty of traversing a given landscape. The **Path Distance** tool in ArcGIS's Spatial Analyst toolbox finds the least costly or most efficient path from each cell in the raster surface to the nearest feature of the designated type (lake, stream, etc.).

Our cost surface was based on terrain, using our ten meter resolution conditioned digital terrain model (see section on terrain variables below) and impedance values assigned to vegetation types from our historic vegetation model (Hobbs 2019a). Weighting factors for vegetation types were estimated based on published literature (Soule and Goldman 1972; Givoni and Goldman 1971; Herzog 2014). A compilation of published and interpreted impedance values is summarized in Table 1.

Table 1: Multipliers for Different Land Covers or Soils Derived from Published Studies

Surface	Multiplier
Blacktop road, Barren or Sparsely Vegetated[1]	1.0
Dirt road or grass[1]	1.1
Light brush[1], Prairie Grass[4] and Major River Downstream[4]	1.2
Ploughed field[2]	1.3
Heavy brush[1]	1.5
Forests – Evergreen, Deciduous, and Mixed[1]	1.5
Hard Snow/Ice[2]	1.6

Surface	Multiplier
Lake[4]	1.7
Swampy Bog[1], Permanent Wetlands[1], Great Lake[4], and Major River Upstream[4]	1.8
Loose Sand[3]	2.0

[1]Soule and Goldman 1972; [2]Givoni and Goldman 1971; [3]compiled data from multiple sources and interpreted by I. Herzog 2014; [4]interpreted for MnModel Phase 4.

Our interpretations (Table 1) were based on the following assumptions:

- Walking across Prairie Grass would take about the same effort as traversing Light Brush.
- Paddling or floating downstream would take about the same effort as walking in Light Brush or Prairie Grass.
- Howey (2011) found that having low costs for Lakes caused puddle jumping in her least cost paths, so she reassigned a higher cost to avoid this impractical path. We assigned a 1.7 multiplier, assuming it would take more energy to build a canoe and paddle than it would to traverse Hard Snow and Ice, but not as much energy as canoeing a Great Lake or traversing a Swampy Bog.
- We assigned a 1.8 multiplier to Major River Upstream, assuming that it would take more energy to build a canoe and paddle up-current than to traverse Loose Sand and more energy than building a canoe and paddling a relatively calm Lake.
- The Great Lake (i.e. Lake Superior) both stores and releases a great amount of energy, which makes it more difficult to navigate. We equated canoeing the Great Lake to paddling upstream on a major river or traversing a swampy bog or wetland.

Translating these estimates to the values in our historic vegetation model required additional interpretations and adjustments. Resistance values assigned to our vegetation model classes are summarized in Table 2. Since travel on rivers could be either upstream or downstream, we split the difference and assigned an intermediate value (1.5) to all river travel. The 'Wet Land' category consists of areas predicted by the model to be lakes or rivers that were not mapped as such by the Public Land Survey. These were usually adjacent to lakes or on floodplains. We suspect these areas were wetlands or intermittent water and assigned a value higher than shallow seasonal wetlands (wet meadows and floodplain forest) but lower than deep wetlands (marshes and swamps). There are obvious conditions that we cannot take into account in our model. Aside from direction of travel on rivers, considerations such as travel upslope or down and travel at different times of the year are beyond our abilities to incorporate into this model.

Table 2: Multipliers Assigned to MnModel Phase 4 Historic Vegetation Model Classes

Historic Vegetation Model Class	Multiplier
Coniferous Savanna, Deciduous Savanna, Brush-Prairie, Prairie	1.2
Coniferous Forest, Deciduous Forest, Mixed Coniferous-Deciduous Forest	1.4
River	1.5
Floodplain Forest, Wet Meadow/Fen	1.5
Coniferous Woodland, Deciduous Woodland	1.5
Wet Land	1.6
Lake	1.7
Bog, Conifer Swamp, Hardwood Swamp, Shrub Swamp, Marsh,	1.8

Missing Data

Statistical modeling requires that there be valid values for every variable for each cell in the model region. Unfortunately, we were missing valid values for soils data for large parts of our model. First, there are no gSSURGO data for Canada. Second, gSSURGO mapping is not yet complete for several Minnesota counties. Finally, even where county mapping is complete, gSSURGO does not record values for most attributes within water bodies, urban land, and disturbed features such as mines and gravel pits.

The missing soils data did not affect the derivation of soil variables, but it did require us to build two separate models for each region. The first model included no soil variables, so was complete for the region. The second model included soil variables, but had NULL values where soils data were missing. We then made composites of the models so that values from the first model could replace the NULL values in the second model (Hobbs 2019b). This is mentioned here only to illustrate that it is possible to use imperfect datasets, when necessary, for modeling.

Environmental Variables

Ideally, environmental variables used as predictors for statistical model should meet several criteria. First, they must be in raster format and, therefore, must be numeric. For MnModel Phase 4, we prefer to use only integer grids, as floating point grids take too much hard drive space, require too much RAM, and take too long to

process. Second, they should be high resolution. Resolution is a function of the source data, and this varies from one data type to another. We prefer data at a scale of 1:24,000, but achieving this is not always possible. Finally, they should be reflective of the environment at the time that archaeological resources were deposited. This is the most difficult criterion to meet. For MnModel Phase 4, we have attempted to approach this by creating models of historic elevation (see section on Terrain, below), historic and prehistoric surface hydrography (Hobbs et al. 2019a), and historic vegetation (Hobbs 2019a). The following sections discuss how we developed each of the predictor variables used for the MnModel Phase 4 archaeological predictive models.

Terrain Variables

Minnesota is fortunate to have high quality statewide digital terrain data from LiDAR. The downside of this, for MnModel, is that infrastructure, anthropogenic features and surface disturbance are blatantly apparent in the data. Since none of these features had any bearing on archaeological site locations, except by indicating where sites may have been disturbed or destroyed, our challenge was to try 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.

Conditioned DTM

LiDAR data were available for all of Minnesota. For a 15-mile buffer around the state, LiDAR data were used where available and, where LiDAR was not available, older USGS NED 10 meter DEM data were used. The surface elevation vertical accuracy for the LiDAR data is plus or minus two feet.

The first step in the process was to down-sample the original three meter resolution county LiDAR data to ten meter resolution. In addition, all elevation values were converted from meters to feet, and all county datasets were mosaicked into a seamless statewide LiDAR dataset. At the county boundaries, small strips of 'NoData' cells were replaced with a 3x3 focal mean value. This created a DTM, called DTM10_ORIG, with a source scale of 1:20,000.

The next step was to replace level lake planes in DTM10_ORIG with rasterized bathymetric lake contours acquired from the Minnesota DNR. We did this because a number of Minnesota's lakes have been enlarged by damming. The bathymetric data provides us with an approximation of the historic terrain below modern reservoir levels. Bathymetric contours were available from MnDNR for approximately 1,830 lakes. Additional bathymetric data were acquired for Lake of the Woods, Upper and Lower Red Lake, and the North Shore of Lake Superior. For Lake of the Woods, depth contours were heads-up digitized from 1:24,000 scale topographic maps. For Upper and Lower Red Lake, depth contours were derived from a 2011 report published by the Red Lake Nation, which included depth survey contours conducted for their Integrated Resource Management Plan. For Lake Superior, near shore 10-meter LiDAR was obtained from the Minnesota DNR and older 90-meter sonar based bathymetry data were obtained from NOAA to fill-in the remainder of the 15-mile buffer zone. All bathymetric data were converted to rasterized negative depth values, then subtracted from the LiDAR-derived elevation of each lake surface. Vertical accuracy of the bathymetric data is unknown and likely varies.

Topographic data from 1899, digitized by Minnesota Geological Survey (MGS) were used to replace a portion of the Mesabi Iron Range, restoring the pit mines to a more natural surface (Lively et al. 2002).

The DTM with the bathymetric data and Mesabi Range topography was further processed to remove man-made features to the greatest extent possible. Mapped features, such as roads, ditches, railroads, and airports, were buffered, merged, and dissolved. This composite buffer was then used to clip out the modern elevation values from the LiDAR-based DTM. A set of custom processing tools was developed in Python to search for "NoData" cells and if found, replace these cells with a dynamic cut-fill process that referenced the existing terrain using multiple, iterative passes to fill in the clipped areas one row of cells per pass starting along the outermost edge. The purpose of this procedure was to raise ditches and lower road crowns by calculating a local mean elevation to approximate the original terrain surface as close as possible. A secondary goal was to reduce the slopes within the replacement zones to less than 15 degrees since terrain variables used in MnModel Phase 3 were found to have a sensitivity to slopes of 15 degrees or greater. These procedures produced the DTM10COND raster.

DTM10CONDPR is a pit-removed version of DTM10COND that was processed with the TauDEM (Terrain Analysis Using Digital Elevation Models) software **Pit Removal** tool to fill-in all sinks. This allowed us to perform statewide surface hydrology calculations using other TauDEM tools. TauDEM is a collection of surface hydrology processing tools created by David Tarboton and is available from Utah State University.

Local Variables

Elevation (ELEV)

ELEV consists of the elevation values in feet from the conditioned DTM (DTM10COND). A statewide ELEV raster was created by clipping DTM10CON with the project's ten kilometer buffer.

Percent Slope (SLOPE)

A percent slope raster was generated from DTM10COND using the ArcGIS Spatial Analyst **SLOPE** tool, using 'DEGREE' as the output option.

Aspect Range (ASP_RNG)

ASPECT was created by processing DTM10COND with the ArcGIS Spatial Analyst **ASPECT** tool. ASP_RNG is a classified version of aspect, using the range breaks specified in Table 3. This classification insures that the most exposed, sunniest locations have the highest values and that north-facing slopes always have the lowest values.

Table 3: Aspect Range Breaks

ASPECT Value	Direction	ASP_RNG Value
-1	Flat	9
0-22.5	North	1
22.5-67.5	Northeast	2
67.5-112.5	East	4
112.5-157.5	Southeast	6
157.5-202.5	South	8
202.5-247.5	Southwest	7
247.5-292.5	West	5
292.5-337.5	Northwest	3
337.6-360	North	1

Surface Curvature (CURV)

CURV is a 10-meter curvature raster generated from DTM10COND using the ArcGIS Spatial Analyst **CURVATURE** tool with the default settings. Positive values indicate that the land surface is upwardly convex at the cell. Negative values indicate that the surface is concave at the cell.

Measures of Relative Topographic Position

A location's relative position within its surroundings may have bearing on how desirable it is for habitation and thus its suitability for archaeological sites. For MnModel Phase 4 we experimented with several different measures of relative topographic position applied to different sized neighborhoods.

Relative Elevation within Five Kilometers (REL)

Relative elevation is a measure of a cell's height above the lowest point within a 10,000 meter diameter circle (Hammer 1993). If the source cell is the lowest cell within the radius, a value of "0" is output. REL was created by processing DTM10COND with a combination of the ArcGIS Spatial Analyst **ZONAL STATISTICS**, **FOCAL**

STATISTICS, and **CON** tools within a custom Python script (Brown et al. 2019). The output represents the relative elevation at each cell above the lowest elevation within a 5,000 meter radius. If the source cell is the lowest cell within a 5,000 meter radius, a value of "0" is output.

Relative Elevation with 90 Meters (REL90)

REL90 is a measure of a cell's height above the lowest point within 90 meters. REL90 was created by processing DTM10COND with a combination of the ArcGIS Spatial Analyst **FOCAL STATISTICS** and **MINUS** tools within a custom Python script (Brown et al. 2019).

Topographic Position Index (TPIxx)

The Topographic Position Index (TPI) is intended to elucidate whether a terrain cell is situated on a ridge, within a valley, or on a side-slope. Calculations are based on a method developed by Guisan et al. (1999). Positive TPI values indicate locations higher than their neighborhood surroundings; near zero values indicate flat areas or areas of constant slope; and negative values are lower than their surroundings. The neighborhood size is variable, so TPI values are strongly scale-dependent. For this reason, we calculated TPI at six scales.

TPI1MI was created by processing ELEV with the ArcGIS Spatial Analyst **FOCAL STATISTICS** and **MINUS** tools, using 90 m (TPI90), 250 m (TPI250), 500 m (TPI500), 1000 m (TPI1000), 1-mile (TPI1MI), and 5-mile (TPI5MI) search distances, then calculating the amount each cell in the input raster is above or below the mean elevation given the search distance.

Measures of Roughness (RGH, RGH90)

Roughness considers the variability of the terrain surrounding each cell. Hammer (1993) suggested the formula (RGH = ((ELEV * 0.3048) + (SLOPE * 6) + (REL * 0.6096)) / 2) for calculating roughness. Because REL is an input to formula, roughness is scale dependent. We calculated roughness at two scales.

Surface Roughness (RGH)

RGH is a 10-meter surface roughness raster calculated using REL, so is a measure of roughness within a 5,000-meter radius.

Surface Roughness within 90 Meters (RGH90)

RGH90 is a 10-meter surface roughness raster calculated using REL90, so is based on the relative elevation above the lowest point on the ground within a 90-meter radius.

Complex Indices

The terrain variables discussed so far have no obvious direct relationship to human activities. They may be considered proxy variables (Kamermans 2011; Kohler and Parker 1986), as they may have multiple meanings to people looking to select a site for habitation or another use. Higher relative elevation, for example, may be important for keeping one's feet dry, for allowing a view of approaching animal herds, or for a myriad of other reasons. Several indices have been suggested that attempt to more directly measure aspects of terrain that may

be important to humans. As such, their contribution to models may be more readily 'interpretable' than other terrain variables.

Shelter Index (SHELTER)

The Shelter index is designed to measure how 'sheltered' or 'exposed' a cell is with respect to the surrounding landscape. The 'shelter index' is conceptualized by placing a cylinder with a 300 meter radius placed over a cell (Kvamme and Kohler 1988). The volume of the DTM surrounding the cell is calculated and subtracted from the volume of the cylinder. If the index value is high, the cell is on a hilltop and is exposed to the elements. If the index value is small, the cell is in a valley bottom or depression and is sheltered. The Shelter Index variable, SHELTER, is a 10-meter raster calculated in ArcGIS Spatial Analyst using the **FOCAL STATISTICS**, **TIMES**, **MINUS**, and **CON** tools within a custom Python script (Brown et al. 2019).

Topographic Wetness Index (TWI)

The Topographic Wetness Index (TWI) is a function of slope and the upstream contributing area orthogonal to the flow direction. TWI is used to quantify the topographic component in hydrological processes. Values are estimate of water accumulation and will be high in flat or depressed areas and low on slopes. For MnModel Phase 4, TWI was calculated as a 10-meter statewide raster using TauDEM. Calculations used both a Specific Catchment Area grid and Flow Direction grid as inputs. Both inputs were derived from the pit-removed version of the 10 m resolution conditioned DTM developed for MnModel Phase 4. Flat areas, typically lake beds, that received 'No Data' values in the analysis were assigned the maximum wetness value of '28'.

Visibility (VISIBLE)

Visibility, the ability to see and be seen, is important to humans for a variety of reasons, including safety and food procurement. The goal of the visibility analysis was to calculate the number of 'observer' cell locations within 3 miles that were visible to each cell in the statewide 10-meter DTM. Selecting the best approach required extensive testing using both the ArcGIS Spatial Analyst **VISIBILITY** and **VIEWSHED 2** tools.

Initial testing with the **VISIBILITY** tool resulted in lengthy processing times. It took 18 hours to process a 15 mile x 15 mile area, for example. We determined that the tool was far too slow for generating a statewide visibility layer. The **VIEWSHED 2** tool is graphics card accelerated (leveraging both Graphics Processing Unit (GPU) and Video RAM (VRAM) resources). The main constraints with the **VIEWSHED 2** tool are the density of the observer points used, the processing extents, and the outer radius specified. Each of these parameters has significant impacts on both processing times and the stability of the tool. The key to running this tool is finding a balance between observer point density, processing extent, and maximum outer radius to adequately achieve the desired output without crashing the program. Table 4 displays the results of testing a range of input extents, observer point densities, and outer radii.

Table 3: Results of Testing Viewshed 2 Tool

Observer Spacing (m)	Extent - Side (m)	Total Observers	Outer Radius (m)	Total Input Raster Cells	Observers x Total Cells	Total Process Time (min)	Process Time /mi^2(sec)	Notes
100	1609	259	1609	25888810	7	2	120	1 x 1 Mile
100	9654	9320	1609	258888100	2413	9	15	6 x 6 Mile
100	16090	25889	1609	631686964	16354	26	16	10 x 10 Mile (1 mile outer radius)
100	16090	25889	3218	766308776	19839	100	60	10 x 10 Mile (2 mile outer radius)
500	16090	1036	8045	1294440500	1340	32	19	10 x 10 Mile (mile outer radius)
500	16090	1036	3218	766308776	794	4	2	10 x 10 Mile (2 mile outer radius)
500	16090	1036	4827	921641636	954	10	6	10 x 10 Mile (3 mile outer radius)
500	16090	1036	6436	1097685544	1137	17	10	10 x 10 Mile (4 mile outer radius)
500	16090	1036	8045	1294440500	1340	24	14	10 x 10 Mile (5 mile outer radius)

Observer Spacing (m)	Extent - Side (m)	Total Observers	Outer Radius (m)	Total Input Raster Cells	Observers x Total Cells	Total Process Time (min)	Process Time /mi^2(sec)	Notes
500	16090	1036	9654	1511906504	1566	FAIL	-	10 x 10 Mile (6 mile outer radius)
1000	12872	166	14481	1915771940	317	FAIL	-	8 x 8 Mile (9 mile outer radius)
1000	16090	259	9654	1511906504	391	10	6	10 x 10 Mile (6 mile outer radius)
1000	16090	259	11263	1750083556	453	13	8	10 x 10 Mile (7 mile outer radius)
1000	16090	259	12872	2008971656	520	FAIL	-	10 x 10 Mile (8 mile outer radius)
1000	16090	259	12872	2008971656	520	FAIL	-	10 x 10 Mile (8 mile outer radius)
1000	16090	259	14481	2288570804	592	FAIL	-	10 x 10 Mile (9 mile outer radius)
1200	16090	180	14481	2288570804	411	14	8	10 x 10 Mile (9 mile outer radius)
1200	16090	180	16090	2588881000	465	FAIL	-	10 x 10 Mile (10 mile outer radius)

From this testing, we determined that the following input parameters would be most suitable for processing a statewide visibility layer using the conditioned 10-meter LiDAR-based DTM as the input source:

- Input raster: statewide integer 10-meter DTM
- Observer points: 1 kilometer spacing
- Surface Offset: +1.3 meters
- Observer Offset: 0 meters
- Outer Radius: 3 miles (4,828 meters)

Using these parameters, a total of 1,130 ten mile x ten mile processing extents successfully completed in approximately ten full days of processing. Initially, a five mile outer radius was used, but was found to cause the tool to occasionally throw errors, so the outer radius was reduced to three miles, which proved to be a very stable setting with no adverse effects on the output. The outer radius of 3 miles created a 3-mile overlap in all directions between the processing extents. Once processing of all 1,130 mile x ten mile extents was completed, the output files were combined into a single mosaic using the ArcGIS **MOSAIC to NEW RASTER** tool with the 'MAXIMUM' operator. The VISIBILITY raster values are the total count of observer points visible in all directions from each raster cell.

Note that the **Viewshed 2** tool, when running with a large number of observer points, calculates the count of all observer points visible for each raster cell location in all directions within the outer radius limit. Thus the output values are simply the visible observer point count totals per ten meter cell. Also, given the large processing extents and directionality of visibility counts in overlap areas, including a buffer zone greater than the width of the processing extents was essential to avoid decreased observer counts near the state boundary.

Geomorphology

Variables from MnModel Landscape Model

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. No single data source covered the entire state at a scale (<1:100,000) meaningful to the MnModel project. The Landscape Model provides the highest resolution data available for any given area as well as a consistent hierarchical classification scheme identifying Region, Region Name, Subregion, Subregion Name, Landscape, Landform, and Mantle. Source data scales range from 1:24,000 to 1:100,000. Two variables were extracted directly from this model.

Landscape (LSCAPE)

Both the Region and Subregion features from the Landscape Model were deemed to be large to display sufficient variation within a modeling region to be useful variables for modeling. The largest unit used as a variable was landscape field (LSCAPE), which was extracted from the statewide Landscape Model (LANDMOD) using the **LANDMODVARS** tool in the MnModel Phase 4 **LANDVARS toolbox**, then clipped for buffered regions using the **Clip LANDVARS** tool in same toolbox. There are eighteen unique values for LSCAPE (Table 5). The

most extensive are the Stagnant Ice and Glaciolacustrine landscapes, while the rarest are the Tributary Fans and Meltwater Trough Fans.

Table 5: Landscape (LSCAPE) Values and their Extent in the MnModel Landscape Model

Landscape	Area (sq. km)
Active Ice	37880.80
Catastrophic Flood	2848.45
Collapsed Meltwater Trough	4661.39
Collapsed Outwash Plain	9359.94
Collapsed Sand Plain	1645.29
Dissected Bedrock Uplands	4579.73
Eolian	372.99
Floodplain	7684.83
Glaciofluvial	14518.90
Glaciolacustrine	52968.80
Ice Contact	3010.94
Lacustrine	600.42
Meltwater Trough Fan	83.33
Peatland	7764.65
Stagnant Ice	67210.40
Tributary Fan	50.81
Valley Margin	3004.20
Valley Terrace	3034.91

The LSCAPE variable indicates which landscape encompasses the majority of the archaeological site polygon. LSCAPE is one of only a handful of categorical variables used for modeling. The landscape values were converted to numeric codes so that they could be interpreted by the statistical software.

Landform (LFORM)

Landform (LFORM) values were also extracted directly from the MnModel Landscape Model using the **LANDMODVARS** tool in the MnModel Phase 4 **LANDVARS toolbox**, then clipped for buffered regions using the **Clip LANDVARS** tool in same toolbox. Landforms are the smallest geomorphic unit mapped. There are 89 unique landforms defined by the Landscape Model. 'Plain' is the most extensive. The 'Plain' landform can be found in a variety of landscapes. Because of the mapping scale of the landforms in river valleys, derived from

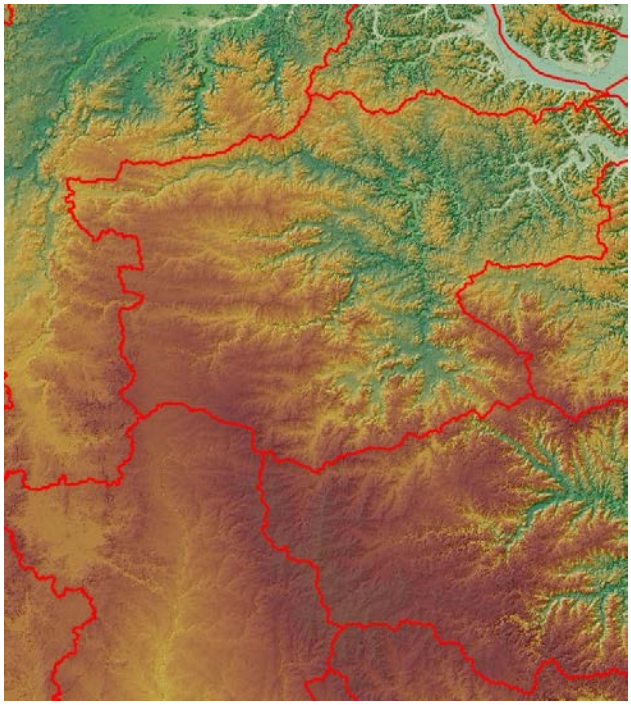
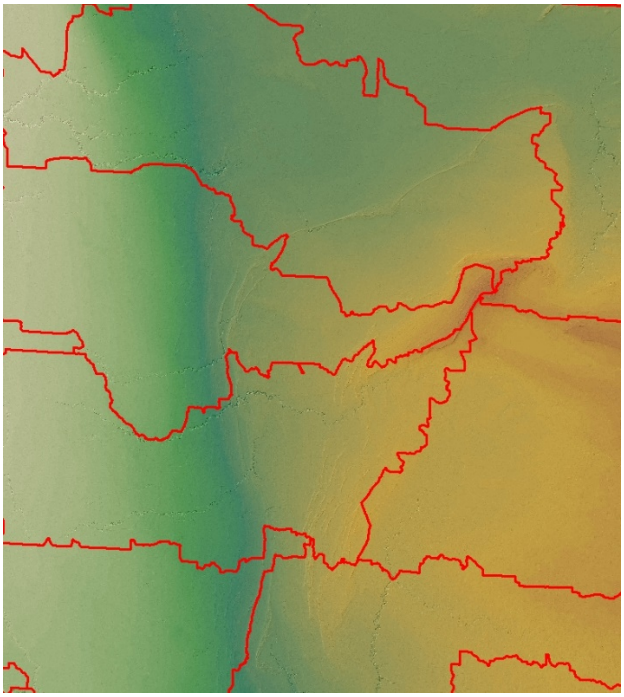
MnDOT's Landforms Sediment Assemblage (LfSA) mapping, the smallest and least extensive landforms are mapped in river valleys. These include Crevasse Splay Meander Belt, Nivation Hollow Ramp, Crevasse Splay Distributary Mouth Bar (Distal), and Spit.

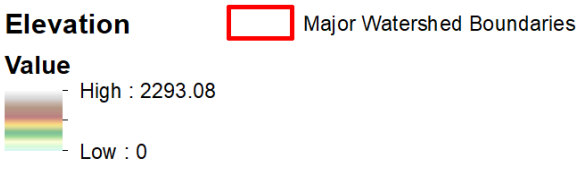
Rare categorical values were problematic for modeling. The statistical analysis requires that the prediction points have the same values for each categorical variable as the sample data. If values are present in the prediction points that were not present in the sample data, the model cannot predict the outcome of the cell. Because prediction points are much more numerous and closely spaced than either the archaeological sites or background points in the sample data, they are more likely to intercept the rare landforms. This required reclassifying rare landforms on the fly to combine them with more common landforms. For future modeling, it would be preferable to take this step prior to modeling to minimize the amount of reclassification necessary when running the models.

Variables from MnDNR Watershed Boundaries

MnDNR watershed boundaries were used as proxies for ridges, which may be important as transportation corridors or as viewpoints. Watersheds also may have meaning as units for resource acquisition, as travel within a watershed may be easier than travel between watersheds. MnDNR maps a two-level nested hierarchy of watersheds: major and minor watersheds. The watershed boundaries are necessarily the highest points between drainages. While they may define obvious ridges in the hilly southeastern part of the state (Figure 1a), they do not necessarily occupy particularly high elevations or what most people would consider to be 'ridges' in the flatter regions (Figure 1b).

Figure 1: Major Watershed Boundaries in Two Topographic Situations

1a: Southeastern Minnesota	1b: Northwestern Minnesota
	



Path Distance to Nearest Major Ridge or Divide (CP_MAJRIDGE)

The variable CP_MAJRIDGE measures the least-cost path distance from each cell to the nearest major watershed boundary using the ArcGIS Spatial Analyst **Path Distance** tool. Low values indicate cells near the boundaries of a watershed while high values are found in the interior of the watershed.

Path Distance to Nearest Minor Ridge or Divide (CP_MINRIDGE)

The variable CP_MINRIDGE measures the least-cost path distance from each cell to the nearest minor watershed boundary using the ArcGIS Spatial Analyst **Path Distance** tool.

Size of Major Watershed (MAJ_SIZE)

Assuming that pedestrian or even water travel will be easier within a watershed than between watersheds, the size of a watershed may be a surrogate for the range and quantity of resources that are readily available. The variable MAJ_SIZE measures the size of the major watershed in which each cell is located.

Size of Minor Watershed (MIN_SIZE)

The variable MIN_SIZE measures the size of the minor watershed in which each cell is located.

Islands

Minnesota archaeologists consider islands to be very desirable locations for archaeological sites. This may be for a variety of reasons. Islands may be defensible. They are surrounded by water, so would have both terrestrial and aquatic resources. The surrounding water provides protection from prairie and forest fires. Islands were extracted from the National Wetlands Inventory data for Minnesota. A total of 8,873 island polygons were mapped in both lakes and rivers. After modeling was completed, 89 of these were found to be erroneously classified as islands. All but one of the misclassified polygons formed a contiguous mass within an Aitkin County swamp, so could have affected only a single modeling region (STTA).

On an Island (ISLAND)

The variable ISLAND indicates simply whether a cell is on an island (VALUE = 1) or not on an island (VALUE = 0).

Historic Vegetation

Variables from MnModel Historic Vegetation Model

Vegetation directly provides a variety of critical resources for humans, including food, shelter, and food. Vegetation also serves as habitat for animal species that are important sources of food. To reconstruct natural vegetation in the historic and recent prehistoric past, we developed the MnModel Phase 4 Historic Vegetation Model. It is an approximation of potential natural vegetation at the time of the Public Land Survey based on a statistical analysis of surveyors' observations and a suite of environmental variables (Hobbs 2019a). This model identifies 36 distinct vegetation types in Minnesota (Table 6). Model accuracy is highest for the dominant vegetation types and becomes very low for rare vegetation types. Prairie is the most extensive vegetation type mapped, covering nearly 77,000 square kilometers. The rarest vegetation types mapped are hardwood swamp (1.7 km^2), aspen openings (11.8 km^2), northern conifer woodland (35.9 km^2), and white pine-hardwood forest (53.6 km^2). As with the rare values for landforms, the rare vegetation types had to be reclassified on the fly when modeling. In the future, it would be best to perform the reclassification at an earlier stage.

Table 6: Historic Vegetation Types from the MnModel Phase 4 Vegetation Model

Vegetation Type	Area (sq. km)
LAKE	12395.0
WET LAND	1643.2
RIVER	918.2
BOG	2646.5
CONIFER SWAMP	38451.5
MARSH	10479.1
FLOODPLAIN FOREST	1818.4
HARDWOOD SWAMP	1.7
WET MEADOW/FEN	2458.9
PINE FOREST	404.5
JACK PINE FOREST	5583.0
RED PINE FOREST	3779.2
WHITE PINE FOREST	2391.5
SPRUCE-FIR FOREST	2116.5
BLACK SPRUCE-FEATHERMOSS FOREST	316.1
UPLAND WHITE CEDAR FOREST	272.5
PINE BARRENS	126.1
JACK PINE WOODLAND	103.2
NORTHERN CONIFER WOODLAND	35.9
BOREAL HARDWOOD-CONIFER FOREST	13553.9
MIXED PINE-HARDWOOD FOREST	839.5
NORTHERN HARDWOOD-CONIFER FOREST	273.1
WHITE PINE-HARDWOOD FOREST	53.6
ASPEN FOREST	4908.0
ASPEN-BIRCH FOREST	396.0
PAPER BIRCH FOREST	2526.8
LOWLAND HARDWOOD FOREST	2422.2

Vegetation Type	Area (sq. km)
MAPLE-BASSWOOD FOREST	8566.7
NORTHERN HARDWOOD FOREST	1299.2
OAK FOREST	7679.7
ASPEN OPENINGS	11.8
OAK SAVANNA	10652.2
BRUSH-PRAIRIE	436.9
PRAIRIE	76853.4

Two aspects of vegetation were used as variables for modeling. The vegetation type serves as a proxy for the types of resources available directly at a site or cell. For example, if the vegetation type is Deciduous Forest, resources might include acorns for food and wood for fuel and shelter. If the vegetation type is Prairie, the resources might include root crops and grains and populations of large grazing animals. Vegetation diversity, the number of different vegetation types within a specified radius, providers an indicator of the variety of resources available. Low vegetation diversity may indicate less variety of resources while high diversity may indicate more variety.

Historic Vegetation Type (VEGMOD)

The VEGMOD variable indicates the historic vegetation type at each cell for background and prediction points and the dominant vegetation within an archaeological site polygon. The vegetation types are those listed in Table 6.

Vegetation Diversity within One KM (VEGDIV1K)

VEGDIV1K is a count of the number of unique vegetation types within a one kilometer radius of each cell. For the archaeological site polygons, the value is the average of the VEGDIV1K values within the polygon. The one kilometer radius was selected to indicate the variety of resources that are very closely at hand.

Vegetation Diversity within Five KM (VEGDIV5k)

VEGDIV5K is a count of the number of unique vegetation types within a five kilometer radius of each cell. For the archaeological site polygons, the value is the average of the VEGDIV5K values within the polygon. The five kilometer radius was selected to measure the variety of resources that are not in the immediate vicinity but are easily obtained in less than a day.

Vegetation Diversity within Ten KM (VEGDIV10k)

VEGDIV10K is a count of the number of unique vegetation types within a 10 kilometer radius of each cell. For the archaeological site polygons, the value is the average of the VEGDIV10K values within the polygon. The 10 kilometer radius was selected as this is considered by team archaeologists to be a reasonable distance for foraging within a day.

Wild Rice

Wild rice is an important food source for native peoples in Minnesota. We compiled a feature class of wild rice locations as mapped by the Minnesota Department of Natural Resources, the General Land Office Surveys, and archaeological sites with evidence of ricing activity. There may be other locations, modern, historic, or prehistoric, where wild rice grows or grew but that have not been documented. Missing data are most likely in the southwest part of the state where lakes and wetlands have been drained for agriculture and historic data from GLO land surveys are sparse because line notes have not been digitized.

Attribute accuracy varies by source. MnDNR wild rice locations are points and tend to be located in the centers of water bodies that are associated with wild rice, though whether rice ever grew in the center of the feature would be a function of its depth. Locations digitized from General Land Office Survey plat maps are also in the centers of the water bodies, in this case those labeled on the plat maps as rice marshes or lakes. Points from GLO line notes are along lines that crossed lakes or wetlands described as containing wild rice. They may be at the centers of these water bodies or at the edges where the surveyors entered or left the water body. Locations derived from archaeological data are centroids of site polygons that are recorded as having evidence of ricing activity, such as processing rice. As such, they are near, but not within, the water bodies containing the wild rice.

Polygon data (archaeological sites and GLO Plat map polygons) were converted to point (centroids). GLO line data were verified to make certain the mentions of 'rice' in the line notes were not simply place names. The field [DATA_SOURCE] was added to each attribute table and populated. All sources were appended, with only the attribute [DATA_SOURCE] maintained.

Path Distance to Nearest Wild Rice Location (CP_RICE)

The variable CP_RICE measures the least-cost path distance from each cell to the nearest wild rice location using the ArcGIS Spatial Analyst **Path Distance** tool. This variable is not calculated for three regions for which there are no documented wild rice locations (COTM, ICOT, PLAT).

Historic and Prehistoric Surface Hydrography

Minnesota has experienced a significant reduction in surface water since the introduction of Euro-American agricultural practices in the late 19th century. 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. The MnModel Phase 4 Historic and Prehistoric Hydrographic Model (Stark et al. 2008; Hobbs et al. 2019a) is an attempt to map these features for use as predictive variables.

The historic model (HISTHYD) represents approximate potential surface hydrographic features at the time of the Public Land Survey in Minnesota (1848-1907). The prehistoric model (PREHYD) represents the assumed total extent of surface water and wetlands based on the historic model, geomorphic data, and soils data. The model of prehistoric surface hydrography represents features that may have been present from about 10,000 BP to the time of the Public Land Survey.

All water

Path Distance to Nearest Historic Surface Water (CP_WAT)

The variable CP_WAT measures the least-cost path distance from each cell to the nearest historic permanent standing water from HISTHYD using the ArcGIS Spatial Analyst **Path Distance** tool. Lakes, 'wet land', rivers, bogs, swamps, and marshes are included as permanent standing water. Wet meadows/fens and floodplain forests are considered seasonal and are not included as source cells for this variable.

Lakes

Historic lakes are mostly accurate, as these were mapped on the Public Land Survey plat maps. Some small modern lakes were used to supplement the historic data where the lake was not likely to have been observed by surveyors. Prehistoric lakes include all historic lakes plus lake beds extracted from the geomorphic and soils data.

Path Distance to Nearest Historic Lake (CP_LAKE)

The variable CP_LAKE measures the least-cost path distance from each cell to the nearest historic lake from HISTHYD using the ArcGIS Spatial Analyst **Path Distance** tool. All historic lakes are used as source cells.

Path Distance to Nearest Large Historic Lake (CP_LLK)

The variable CP_LLK measures the least-cost path distance from each cell to the nearest large historic lake from HISTHYD using the ArcGIS Spatial Analyst **Path Distance** tool. Only historic lakes larger than 485,640 m^2 are used as source cells.

Path Distance to Nearest Prehistoric Lake (CP_PLAKE)

The variable CP_PLAKE measures the least-cost path distance from each cell to the nearest prehistoric lake from PREHYD using the ArcGIS Spatial Analyst **Path Distance** tool. All prehistoric lakes are used as source cells.

Path Distance to Nearest Large Prehistoric Lake (CP_PLLK)

The variable CP_PLLK measures the least-cost path distance from each cell to the nearest large prehistoric lake from PREHYD using the ArcGIS Spatial Analyst **Path Distance** tool. Only prehistoric lakes larger than 485,640 m^2 are used as source cells.

Rivers and Streams

Although the historic courses of some major rivers were surveyed and are reasonably represented on the Public Land Survey plat maps, most rivers and streams are drawn imaginatively and are not useful for our purposes. Additional major river courses, as polygons, were extracted from the National Wetlands Inventory for HISTHYD. We have also used modern streams data from MnDNR and the National Hydrography Dataset to represent perennial and intermittent stream courses as lines, though these are not incorporated into HISTHYD. We selected line data only for streams that were not coded as artificial, channeled, impounded, connectors, or having no definable channel. We did include lines coded as 'superseded natural channels.' These streams are extracted from the source data by the hydrographic modeling tool. As stream order is likely to determine whether there is sufficient flow in a stream for travel by canoe or to support certain species of fish, we developed a stream order grid from the 10-m DTM using TauDEM software.

Order of Nearest Stream (ORD_STRM)

The variable ORD_STRM documents the order of the stream nearest to each cell. Only streams of order '6' or higher are used as source cells as these orders seem to capture most of our mapped intermittent and perennial streams. Higher values of this variable indicate that cells are closest to larger streams.

Path Distance to Nearest Intermittent Stream (CP_INT)

The variable CP_INT measures the least-cost path distance from each cell to the nearest intermittent stream from the line data using the ArcGIS Spatial Analyst **Path Distance** tool. Only natural intermittent streams and superseded natural channels coded as intermittent are used as source cells.

Path Distance to Nearest Perennial Stream (CP_PEREN)

The variable CP_PEREN measures the least-cost path distance from each cell to the nearest perennial stream from the line data using the ArcGIS Spatial Analyst **Path Distance** tool. Only natural perennial streams, superseded natural channels coded as perennial, and river centerlines are used as source cells.

Path Distance to Nearest River (CP_RIVER)

The variable CP_RIVER measures the least-cost path distance from each cell to the nearest historic major river from HYDMOD using the ArcGIS Spatial Analyst **Path Distance** tool. Source cells include rivers large enough to have been mapped as polygons and buffered river centerlines that did not intersect the polygon rivers.

Floodplains

River channels change constantly. Floodplains are likely to be better indicators of river activity over time. Historic floodplains were extracted from geomorphic and soils data and incorporated into HYDMOD only where surface water or wetlands did not occupy the floodplain surface. Prehistoric floodplains were extracted from geomorphic data, and all historic floodplains were assumed to have been floodplains prehistorically as well.

Path Distance to Nearest Historic Floodplain (CP_FLOOD)

The variable CP_FLOOD measures the least-cost path distance from each cell to the nearest historic floodplain using the ArcGIS Spatial Analyst **Path Distance** tool. Source cells are from the HISTFLOODPLAINS feature class created as a by-product of creating the Historic Hydrographic Model (Hobbs et al. 2019a).

Path Distance to Nearest Prehistoric Floodplain (CP_PFLOOD)

The variable CP_PFLOOD measures the least-cost path distance from each cell to the nearest prehistoric floodplain using the ArcGIS Spatial Analyst **Path Distance** tool. Source cells are from the PREFLOODPLAINS feature class created as a by-product of creating the Prehistoric Hydrographic Model (Hobbs et al. 2019a).

Wetlands

Historic wetlands are taken from the Historic Vegetation Model (VEGMOD). Because this is a statistical model using data from surveyor's notes, it is more accurate for dominant types (for example, marshes in southern Minnesota and conifer swamps in northern Minnesota) than for less dominant types. For this reason, wetlands have been generalized to larger categories for creating predictive variables. Also, the terms surveyors used were often imprecise and may be the source of much error.

Where gSSURGO data were present, they were used to identify prehistoric wetlands using rules developed by Stark et al. (2008). Where gSSURGO data were not available, wetlands from HISTHYD were used in PREHYD.

Path Distance to Nearest Historic Bog (CP_BOG)

The variable CP_BOG measures the least-cost path distance from each cell to the nearest historic bog from HISTHYD using the ArcGIS Spatial Analyst **Path Distance** tool. All historic bogs are used as source cells in regions where they are present. No bogs were present in the ANOK, BGWD, BLUF, COTM, ICOT, MNRP, OSAV, PLAT, or STPB modeling regions. Bogs may have been more extensive historically, but surveyors may have called them 'swamps.' Where available, bearing tree data were used to help with vegetation classification. However, this cannot help distinguish conifer swamps from bogs.

Path Distance to Nearest Historic Marsh (CP_MARSH)

The variable CP_MARSH measures the least-cost path distance from each cell to the nearest historic marsh from HISTHYD using the ArcGIS Spatial Analyst **Path Distance** tool. All historic marshes are used as source cells. Marshes are mapped in all regions and are by far the dominant wetland type reported in southern Minnesota. They may be over-reported, as surveyors' often use the adjective 'marshy' as well as ambiguous terms such as 'swampy marsh' or 'marshy swamp.'

Path Distance to Nearest Historic Wet Meadow or Fen (CP_MEADOW)

The variable CP_MEADOW measures the least-cost path distance from each cell to the nearest historic wet meadow, wet prairie, or fen from HISTHYD using the ArcGIS Spatial Analyst **Path Distance** tool. All historic wet

meadows, wet prairies, and fens are used as source cells in regions where they are present. None were present in the BDLK, NSHH, or NSUP modeling regions.

Path Distance to Nearest Prehistoric Wetland (CP_PWET)

The variable CP_PWET measures the least-cost path distance from each cell to the nearest prehistoric wetland from PREHYD using the ArcGIS Spatial Analyst **Path Distance** tool. All prehistoric wetlands are used as source cells.

Path Distance to Nearest Historic Swamp (CP_SWAMP)

The variable CP_SWAMP measures the least-cost path distance from each cell to the nearest historic swamp from HISTHYD using the ArcGIS Spatial Analyst **Path Distance** tool. All historic swamps are used as source cells in regions where they are present. These include hardwood swamps, conifer swamps, and shrub swamps. Hardwood were most common in southern Minnesota, while conifer swamps dominated northern Minnesota. No swamps were mapped in the BLUF, COTM, ICOT, or PLAT modeling regions.

Path Distance to Nearest Historic 'Wet' Land (CP_WET)

The variable CP_WET measures the least-cost path distance from each cell to the nearest historic 'wet land' from HISTHYD using the ArcGIS Spatial Analyst **Path Distance** tool. The 'wet land' category is an artifact of the vegetation modeling procedures (Hobbs 2019a). Lakes and rivers were represented in the point data used to develop the statistical model. Consequently, the model predicted lake and river cells. However, we felt it important to ensure that lakes and rivers were as accurate as possible. For this reason, we inserted them from the source data for the historic hydrographic model (Public Land Survey and some modern lakes and rivers). This left areas classified by the model as lakes or rivers that were not mapped as such by surveyors. These tend to occur on floodplains, in reservoir basins, and in depressions. We classified these as 'wet land.' They may have been occupied by standing or intermittent water in the past, but were not mapped as such on the PLS plat maps. . In the future, it would be useful to revise the vegetation model to remove 'lake' and 'river' as options, since we have these documented on maps, then to update the hydrographic model accordingly.

Path Distance to Nearest Historic Wetland (CP_WETLAND)

The variable CP_WETLAND measures the least-cost path distance from each cell to the nearest wetland vegetation type from HISTHYD using the ArcGIS Spatial Analyst **Path Distance** tool. All historic bogs, marshes, swamps, floodplain forests, and wet meadow/fens are used as source cells.

Pedestrian Transportation

Pedestrian Transportation Model

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. We subjectively selected

cutoff values to define major, intermediate, and minor level paths from the model. We also determined, subjectively, that 8900 would be the minimum model value to be considered a path for our purposes.

Transportation Variables

Path Distance to Nearest Major Pedestrian Transportation Route (CP_MAJPATH)

The variable CP_MAJPATH measures the least-cost path distance from each cell to the nearest major pedestrian path using the ArcGIS Spatial Analyst **Path Distance** tool. Path cells with values between 138,408 and 569,235 were considered to be major paths.

Path Distance to Nearest Medium Pedestrian Transportation Route (CP_MEDPATH)

The variable CP_MEDPATH measures the least-cost path distance from each cell to the nearest medium or intermediate pedestrian path using the ArcGIS Spatial Analyst **Path Distance** tool. Path cells with values between 49,000 and 138,400 were considered to be intermediate paths.

Path Distance to Nearest Minor Pedestrian Transportation Route (CP_MINPATH)

The variable CP_MINPATH measures the least-cost path distance from each cell to the nearest minor pedestrian path using the ArcGIS Spatial Analyst **Path Distance** tool. Path cells with values between 8,900 and 49,000 were considered to be minor paths.

Order of Nearest Pedestrian Transportation Route (PATH_ORD)

The variable PATH_STRM documents the order of the path nearest to each cell. Only paths with values of 8,900 or higher are used as source cells. Higher values of this variable indicate that cells are closest to more heavily traveled paths.

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 to aggregate these data by determining the values occupying the largest percentage of the mapunit.

Drainage and Soil Water

Soil Drainage (DRAIN)

The variable DRAIN is an indication of soil drainage. DRAIN is was based primarily on the 'drclassdcd' (drainage class – dominant condition) variable from the gSSURGO mapunit table. Where 'drclassdcd' values were not available, the 'drclasswettest' (drainage class – wettest condition) field values from the mapunit's most extensive component were used (extracted from the component table). Numeric values were assigned to the drainage classes with a low value of '1' indicating very poorly drained soil and a high value of '7' indicating excessively drained soil.

Drainage Index (DI)

The variable DI is the drainage index developed by Michigan State University and the US Department of Agriculture. Values range from '0' (rocky, boulder escarpments) to 99 (water). DI values were linked to soil component keys, then aggregated to mapunits using customized Python scripts.

Flood Frequency (FLDFRQD)

The variable FLDFRQD is the 'flodfreqdcd' (flooding frequency – dominant condition) variable from the gSSURGO mapunit table. Numeric values were assigned to the classes ranging from '0' (None) to '5' (Very frequent).

Hydric Group (HYDGRPDCD)

The variable HYDGRPDCD is the 'hydgrpdcd' (hydrologic group – dominant condition) variable from the gSSURGO mapunit table. Numeric values were assigned to the classes as follows:

- 1 = A: Soils that have low runoff potential when thoroughly wet.
- 2 = A/D: Soils that have low runoff potential when thoroughly wet if the area can be adequately drained, otherwise have high runoff potential when thoroughly wet.
- 3 = B: Soils that have moderately low runoff potential when thoroughly wet.
- 4 = B/D: Soils that have moderately low runoff potential when thoroughly wet if the area can be adequately drained, otherwise have high runoff potential when thoroughly wet.
- 5 = C: Soils that have moderately high runoff potential when thoroughly wet.
- 6 = C/D: Soils that have moderately high runoff potential when thoroughly wet if the area can be adequately drained, otherwise have high runoff potential when thoroughly wet.
- 7 = D: Soils that have high runoff potential when thoroughly wet.

Hydric Soil Presence (HYDPRS)

The variable HYDPRS is the 'hydclprs' (hydric classification - presence) variable from the gSSURGO mapunit table. Numeric values range from 0 to 100, indicating the percentage of the mapunit occupied by hydric soils.

On a Wetland Soil (WETSOIL)

The variable WETSOIL is an indication of the presence of wetland soils as interpreted from gSSURGO soil taxonomy. Taxonomic indicators for soils that have been saturated for extended periods in the past are discussed by Stark et al. (2008). These are the same soils that were used to indicate the presence of prehistoric wetlands in the prehistoric hydrographic model (Hobbs et al. 2019a). Variable values are simply 0 (not a wetland soil) and 1 (wetland soil).

Other Variables

Frost-Free Days (FFD_R)

The variable FFD_R is the 'ffd_r' (frost-free days – representative value) variable from the gSSURGO component table. Component data were aggregated to mapunits using customized Python scripts. Numeric values range from 85 to 100 for Minnesota.

Depth of Surface Soil Horizon (HZDEP)

The variable HZDEP is the 'hzdepb_r' (depth of surface horizon) variable from the gSSURGO component table. Component data were aggregated to mapunits using customized Python scripts. Numeric values range from 2 to 203 for Minnesota.

Productivity Index (PI)

The variable PI is the productivity index developed by Michigan State University and the US Department of Agriculture. Values range from '0' (water, rocks, pits, urban land) to 18 (very rich mesic mollisols). PI values were linked to soil component keys, then aggregated to mapunits using customized Python scripts.

Preparation for Modeling

Variable Lists for Archaeological Predictive Modeling

In all, four different predictive models were developed for each modeling region. Two models were developed to predict site locations and two models to 'predict' survey locations (Hobbs 2019b). Sites and surveys were modeled using the same predictor variables, but each had to be modeled once without soils data and a second time with soils data. This was necessary because the statistical procedures cannot compensate for the NULL values in the soils data. Thus the models without soils data were the only models with no NULL values in the output. Values from these were then used to replace the NULL values in the models developed using soils data.

To guide the development of the two types of models, two variable lists were defined, one with soils data and one without. Sampling procedures (Brown et al. 2019) were designed to create sample files for each version, using the 'ALL' and 'SOIL' identifiers to distinguish between them.

ALLARCHLIST

Predictor variable rasters in ALLARCHLIST provide data for all cells in each modeling region by excluding all soil variables except WETSOIL. Because we could assign WETSOIL values to all nearly all cells we could use it as a variable in every region except where gSSURGO data are missing altogether for large areas (BDLK, MLAC, NSHH, NSUP, and STTA).

SOILARCHLIST

The predictor variable rasters in SOILARCHLIST include all variables but will have NULL values for most soil variables even when soils data are otherwise complete for a region. Consequently the site and survey models developed using these variables will contain NULL values that must be filled in with values from the models using the ALLARCHLIST variables.

Variable Performance

After variables are sampled by both the sample data (archaeological sites or surveys and background points) and the predictive points (Brown et al. 2019), the attribute tables from the resultant point feature classes are exported to .csv format for analysis in R statistical software. The statistical procedures include data cleaning, exploratory data analysis, modeling, and model evaluation. These procedures are fully documented by Landrum et al. (2019). Model results are documented by Hobbs (2019b). This section focuses on how well each variable performed as measured by the model evaluation procedures.

The contributions of predictor variables to the final site models are summarized in the Tables 7-17. In all of these tables, performance is measured by the increase in the mean square error of the model if the variable is removed. Keep in mind that there 20 regional models. If a variable appears in fewer than 20 models, it was removed for one of several reasons: it was absent from the region, it exhibited near zero variance, it failed to distinguish between sites and non-sites, or it was strongly correlated with another variable. More details on the variable performance measures can be found in Hobbs (2019b) and Landrum et al. (2019).

Overall Performance

Modeling regions were selected to minimize environmental diversity within regions. However, there is considerable diversity between the regions, so the predictor variables perform differently from region to region. On average, a variable is used in only about half the models. R reports the percent increase in mean square error (%IncMSE) for each variable in each model. This is interpreted as the increase in the mean square error of the model if that variable is removed. The higher the value, the more important the variable is to the model. From this one can rank the importance of each variable to each model. Overall, the highest average rank achieved by a variable is 6.5 for site models (where the most important variable has a value of 1) and 7.5 for survey models. Few variables achieve this level of performance across the state. When averaged across regions, the average %IncMSE is 16.8 for site model variables and 17.0 for survey model variables.

Predictor variables for MnModel Phase 4 fall into six categories (Table 7). The hydrographic variables, as a group, outperform all others. Vegetation, geomorphology, and path variables also performed well. A number of

terrain variables performed well, but several proved to be redundant with others and had to be removed from the analysis. Soil variables, on the whole, were disappointing.

Table 7. Performance by Categories of Predictor Variables

Category	Number of Variables	Number of Variables Ranking in Top Five of Any Model
Terrain	17	11
Hydrography	18	18
Vegetation	5	4
Geomorphology	7	5
Transportation	4	3
Soils	8	0

Table 8 reports the most significant variables in the site models, the number of models in which they figured, their mean rank in these models, and the mean %IncMSE they contributed. Variables are ranked in order of importance, in terms of their mean rank in the models. It is clear from this table that the variable *Path Distance to Large Historic Lake* is the most important variable for site location, though distances to wetlands and prehistoric large lakes are also significant. Hydrographic variables dominate the table, with 15 of the 28 variables listed. *Elevation* is the top terrain variable.

Table 8. Top Ranked Site Model Variables and Mean %IncMSE

Variable	# Models	Mean Rank	Mean %IncMSE
Path distance to nearest large historic lake	25	4.0	29.65
Path distance to nearest historic wetland	36	5.0	19.86
Path distance to nearest large prehistoric lake	18	6.3	19.44
Elevation	28	8.3	18.35
Path distance to nearest historic 'wet' land	40	8.7	15.31
Path distance to nearest historic river	20	9.7	14.36
Path distance to nearest historic lake	24	9.8	13.56
Relative elevation within 5 km	33	10.4	13.79

Variable	# Models	Mean Rank	Mean %IncMSE
Topographic Position Index within 500 m	21	12.2	14.68
Path distance to nearest perennial stream	34	12.7	11.94
Path distance to nearest historic surface water	8	13.1	8.18
Path distance to nearest wild rice	31	13.1	12.0
Path distance to nearest prehistoric floodplain	25	13.6	12.02
Path distance to nearest prehistoric wetland	36	13.6	12.24
Path distance to nearest major path	18	13.7	14.73
Path distance to nearest historic floodplain	35	14.0	11.57
Topographic Position Index within 250-meter radius	10	14.3	9.58
Path distance to nearest historic bog	16	14.3	8.62
Topographic Position Index raster within 5-mile radius	31	14.7	10.31
Stream order of nearest stream (>=5)	32	14.7	9.45
Topographic Position Index raster within 1-mile radius	19	14.8	11.17
Path distance to nearest prehistoric lake	16	14.9	10.25
Potential historic vegetation type	29	15.5	8.99
Number of vegetation types within 1 km	24	15.9	8.16
Landforms	38	15.9	10.93
Path distance to major ridge or divide	25	16.2	8.27
Path distance to nearest historic swamp	22	16.5	8.32
Size of minor watershed	24	18.8	8.58

Table 9 reports the most significant variables in the survey models, ranking the variables in order of their mean rank in the models. The variable *Path Distance to Nearest Historic Surface Water of All Types* showed the stronger performance, though it appeared in only two models. *Elevation* has nearly as high a mean %IncMSE value and appears in 30 models, so should be considered the best variable for predicting which types of landscapes have most likely been surveyed. *Path Distance to Nearest Large Historic Lake* performs nearly as well. Hydrographic variables are less dominant than in the site model, with only 11 variables listed.

Table 9. Top Ranked Survey Model Variables and Mean %IncMSE

Variable	# Models	Mean Rank	Mean %IncMSE
Path distance to nearest historic surface water of all types	2	1.5	36.57

Variable	# Models	Mean Rank	Mean %IncMSE
Elevation	30	5.1	36.09
Path distance to nearest large historic lake	30	6.5	34.15
Path distance to nearest major paths	20	7.1	30.53
Path distance to nearest historic intermittent stream	37	8.1	29.09
Path distance to nearest major ridge	40	9.8	27.37
Path distance to nearest intermediate path	36	9.9	27.16
Path distance to nearest prehistoric wetland	35	10.1	28.64
Path distance to nearest large prehistoric lake	22	10.4	27.51
Path distance to nearest wild rice location	31	10.6	26.54
Path distance to nearest historic perennial stream	35	13.1	23.74
Path distance to nearest historic major river	20	13.1	24.6
Landform	37	13.1	23.5
Slope of land surface, in degrees	28	14.2	19.73
Path distance to nearest historic wet meadow/wet prairie/fen	30	14.2	21.73
Aspect classified by range breaks	6	14.8	10.97
Path distance to nearest minor path	36	14.8	22.42
Size of major watershed	36	14.9	23.57
Path distance to nearest historic swamp	22	15.0	21.58
Path distance to nearest historic lake	21	16.0	20.15
Path distance to nearest historic 'wet' land	40	16.0	21.61
Path distance to nearest prehistoric lake	21	16.1	18.86
Landscape	25	16.1	17.93
Potential historic vegetation type	25	16.4	18.53
Surface curvature	30	16.5	20.17
Size of minor watershed	40	16.7	21.37
Relative elevation within a 5-kilometer radius.	38	17.1	20.66
Path distance to nearest historic floodplain	32	17.1	20.57

Terrain

Terrain is important for both site and survey locations. Except for *Elevation* and *Relative elevation within 5 km*, however, site and survey models tended to rely on different terrain variables (Tables 8 and 9). The *Topographic Position Index* at several scales figured into the site models, while survey models tended to utilize less complex measures such as *Slope* and *Aspect*.

Future modelers can reduce the number of terrain variables, as several were consistently redundant with others. *Surface roughness within 90 m* (RGH90), the *Shelter Index* (SHELTER), and *Topographic Position within 1000 m* (TPI1000) are the most expendable terrain variables. RGH90 is most often correlated with *Elevation* and may also be correlated with *Slope*. *Shelter Index* is usually redundant with the *Topographic Position Indices* at scales of 250, 500, and 1000 meters and has very low %IncMSE scores in the few site models where it was used. It did contribute to the survey models in North Shore Highlands and Nashwauk-Toimi-Laurentian Uplands with %IncMSE values ranging from 11.6 to 12.9. TPI1000 contributed to site models in only one region, ASPK, with a high %IncMSE value of 11.7. It is most frequently redundant with *Shelter Index*, *Topographic Position within 500 Meters (*TPI500), and *Topographic Position within One Mile* (TPI1MI).

On average, *Elevation* is the strongest terrain model with very high %IncMSE values (Table 10). Other variables appear in more models, though their average impact is less. *Shelter Index* appears to be the weakest terrain variable, but only because it was removed for all models for being correlated with several Topographic Position Index variables. Thus, the contribution of topographic position at 250, 500, and 1000 meters may be, at least in part, a measure of the shelter provided. Other readily 'interpretable' terrain variables, *Visibility* and *Topographic Wetness Index*, were also out-performed by standard, though more ambiguous, variables.

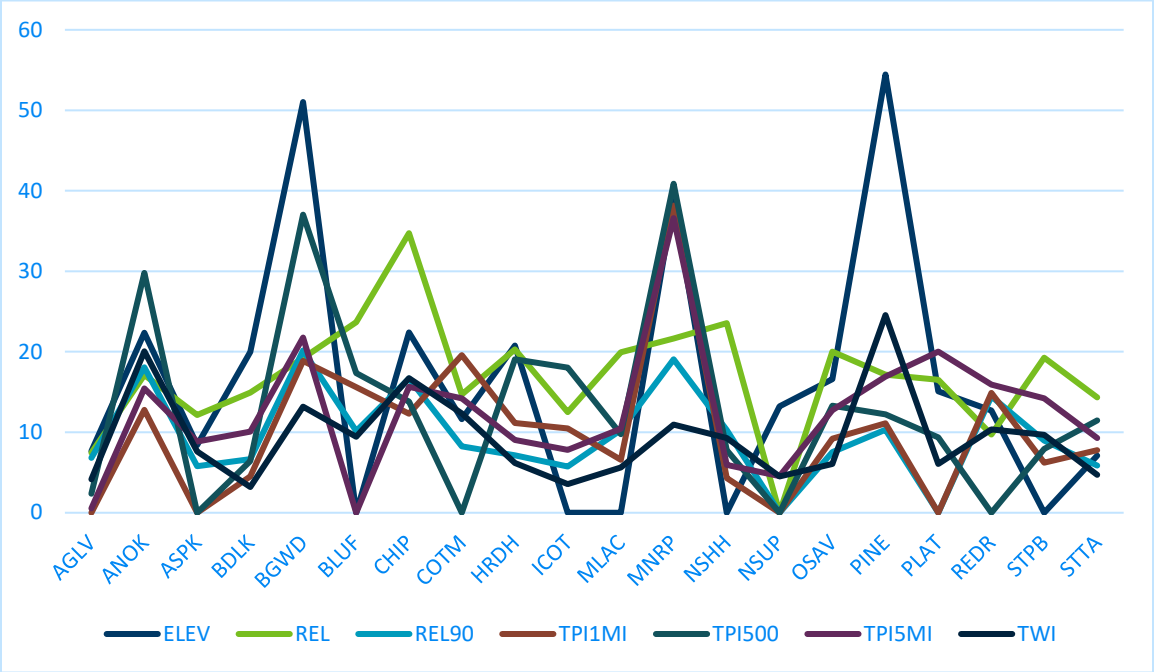
Table 10: Performance of Topographic Predictor Variables in Phase 4 Site Potential Models

Variable	No. Models	Minimum %IncMSE	Maximum %IncMSE	Mean %IncMSE
Aspect Range	10	0.9	20.6	4.4
Elevation	15	7.1	54.5	21.5
Relative Elevation within 5 km	19	7.4	34.7	17.8
Relative Elevation within 90 m	18	5.7	20.2	10.7
Shelter Index	0	0	0	0
Slope	17	2.6	13.2	7.7
Surface Curvature	17	2.0	20.0	8.8
Surface Roughness within 5 km radius	12	5.7	13.8	9.2
Topographic Position Index within 1 Mile	16	4.3	38.2	12.7

Variable	No. Models	Minimum %IncMSE	Maximum %IncMSE	Mean %IncMSE
Topographic Position Index within 1000 m	1	9.3	9.3	9.3
Topographic Position Index within 250 m	8	1.2	42.7	17.2
Topographic Position Index within 5 miles	19	0.5	36.6	13.2
Topographic Position Index within 500 m	16	2.4	40.9	16.0
Topographic Position Index within 90 m	18	0.9	15.8	6.3
Topographic Wetness Index	20	3.2	24.6	9.4
Visibility	13	0.9	16.1	8.2

The importance of terrain variables to the site models varies by region (Figure 2). Terrain variables are modest contributors to most models. The BGWD, CHIP, MNRP, and PINE regions stand out for having one or more terrain variables with %IncMSE values greater than 30. These variables are *Elevation* (ELEV) in BGWD, MRNP, and PINE; *Relative elevation within 5 km* (REL) in CHIP; *Topographic position index within 500 meters* (TPI500) in BGWD and MNRP; and *Topographic position index within 5 miles* in MNRP.

Figure 2: Performance (%IncMSE) of Terrain Variables[1] in Best Site Models, by Region



[1]This graph is limited to terrain variables present in more than half the best site models and contributing at least %IncMSE = 20 in one of the best site models.

Historic and Prehistoric Hydrography

Hydrographic variables dominate both the site and survey models. Prehistoric people needed sources of fresh water, so they located their activities near water bodies. Proximity to large lakes is particularly significant. Knowing this, archaeologists strongly prefer to survey near water. None of this is surprising.

Each hydrographic variable in the dataset was utilized in at least eight of the site models and two of the survey models. Their contributions to specific models can be quite high (Table 11). The %IncMSE of *Path distance to nearest large historic lake* was 35.3 in the Border Lakes site model. *Path distance to nearest historic wetland* was responsible for 39.6 %IncMSE in the Anoka Sand Plain site model.

Least-cost path distances to large historic lakes, prehistoric large lakes, historic 'wet' land, and historic wetlands are the dominant hydrographic variables. The several lake variables were often correlated with one another, so they may not all appear in the same models. Least-cost path distance to rivers, perennial streams, and intermittent streams all performed well. Individual types of wetlands (bogs, marshes, swamps, wet meadows) seemed to matter less than proximity to a basin filled with some type of water. However, *Path distance to historic surface water* (a variable that includes lakes, 'wet land', rivers, bogs, swamps, and marshes as source cells) did not perform well. It seems important to the models that distances to lakes, rivers, and wetlands be distinguished as these contribute different and useful information.

Table 11: Performance of Hydrographic Predictor Variables in Phase 4 Site Potential Models

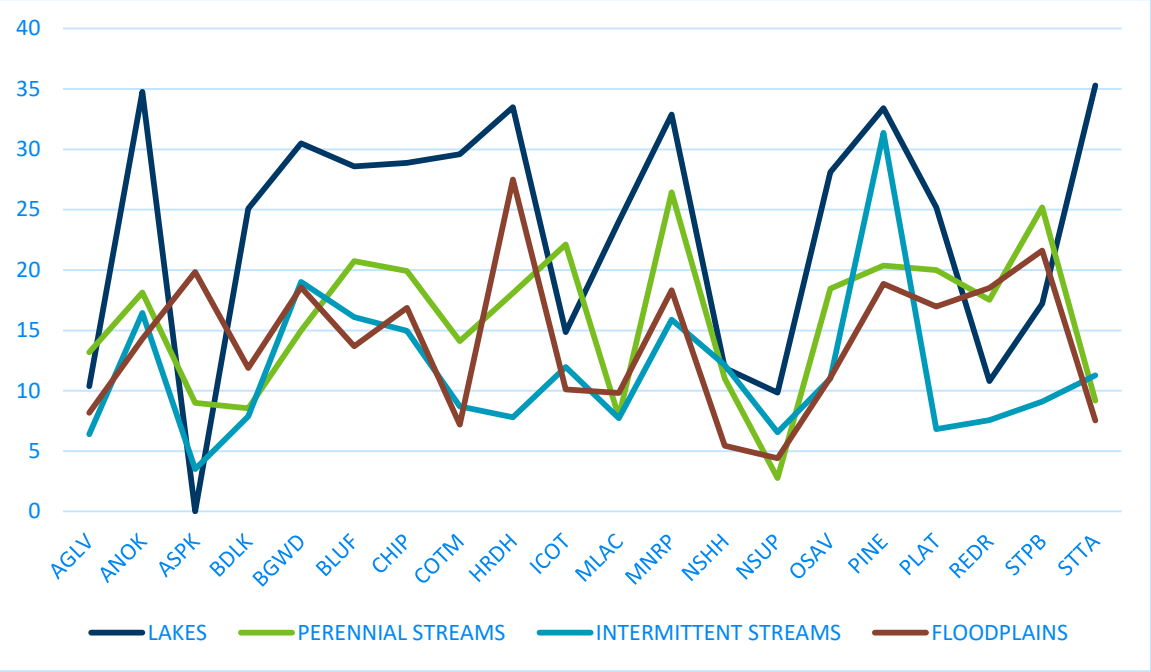
Variable	No. Models	Minimum %IncMSE	Maximum %IncMSE	Mean %IncMSE
Order of Nearest Stream	20	3.2	26.8	11.5
Path Distance to Bog	10	-0.2	19.4	8.8
Path Distance to Historic Floodplain	19	4.0	21.6	12.3
Path Distance to Historic Lake	11	9.4	28.1	16.3
Path Distance to Historic Surface Water *(all types of wetlands, lakes, rivers)*	1	16.2	16.2	16.2
Path Distance to Historic Swamp	14	1.9	17.7	9.8
Path Distance to Historic 'Wet' Land *(identified by the vegetation model as lakes, but not lakes on the historic map)*	20	6.2	31.0	18.4
Path Distance to Historic Wetlands *(all types of wetlands, but not lakes or rivers*	19	6.4	39.6	17.8
Path Distance to Intermittent Stream	20	3.5	31.4	11.6
Path Distance to Large Historic Lake	11	10.4	35.3	25.0
Path Distance to Marsh	17	4.0	20.2	10.2
Path Distance to Perennial Stream	16	2.8	25.2	14.0
Path Distance to Prehistoric Floodplain	15	5.5	27.5	13.8
Path Distance to Prehistoric Lake	13	5.9	25.2	14.0
Path Distance to Prehistoric Large Lake	12	7.4	33.4	19.9

Variable	No. Models	Minimum %IncMSE	Maximum %IncMSE	Mean %IncMSE
Path Distance to Prehistoric Wetland	19	5.1	33.1	12.7
Path Distance to River	10	9.2	26.4	17.6
Path Distance to Wet Meadow/Fen	15	1.0	17.6	10.1

The importance of lakes, both historic and prehistoric does vary geographically. Proximity to lakes clearly dominates models for most regions, even regions such as BLUF (The Blufflands) where lakes are rare.

It is clear that proximity to lakes plays no role in site locations in ASPK (Aspen Parklands), where all 'distance to lake' variables were removed from the models because they failed to distinguish between sites and background points. There are lakes in the ASPK region, but they are surrounded by wetlands. Consequently, lakeshores may not have provided suitable habitat historically or prehistorically. This may also be the reason 'distance to lake' variables receive relatively low scores in AGLV (Agassiz Lowlands/Littlefork Vermilion Uplands). In ICOT (Inner Coteau), all lakes were in the region's 10 km buffer zone, not within the region per se, and in the REDR (Red River Prairie) region lakes are small and concentrated only near the border with HRDH (Hardwood Hills).

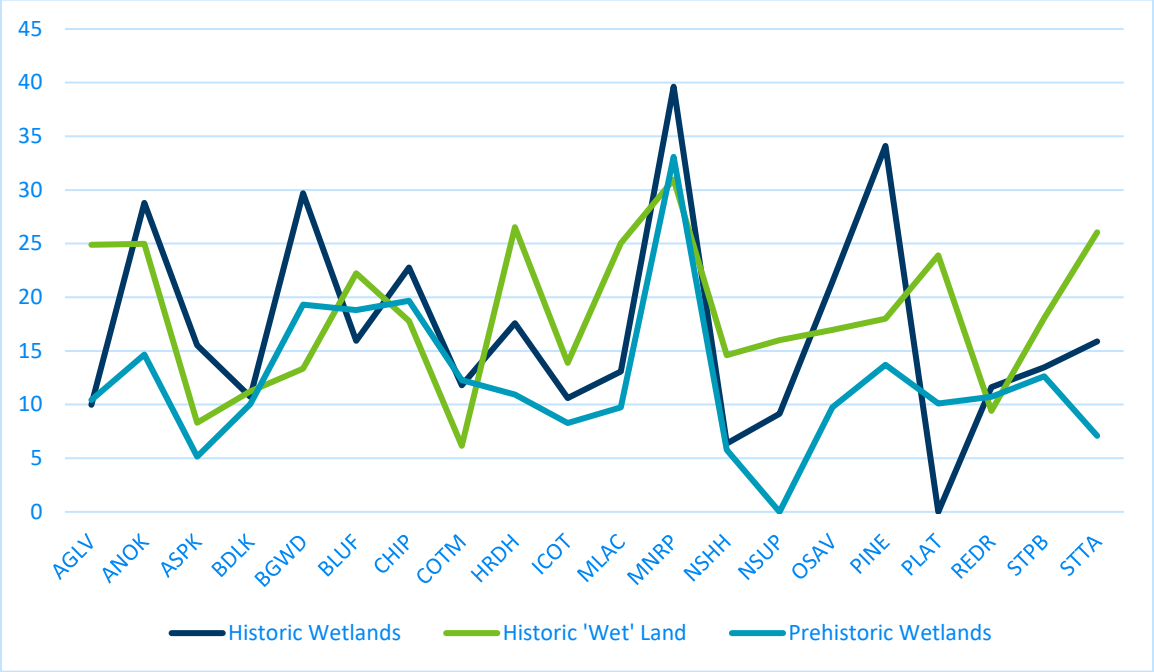
Figure 3: Performance (%IncMSE) of Surface Water Categories[1] in Best Site Models, by Region



[1]Variables were grouped into categories for this graph and redundant variables were removed to simplify the display. For the LAKES category, the highest model value for any of the four 'distance to lakes' variables was graphed. For the PERENNIAL STREAMS category, the highest model value for either *Path distance to nearest perennial stream* or *Path distance to nearest river* was used. For the FLOODPLAINS category, the highest model value for either of the two 'distance to floodplains' variables was used.

Since individual wetland types seemed to be less important to the models than the simple presence of wetlands, Figure 4 compares the performance of the three generalized 'distance to wetlands' variables by region. In most regions, compared to background points, sites tend to be closer to historic wetlands and 'wet' lands, but farther from prehistoric wetlands. Together *Path distance to historic wetlands* and *Path distance to historic 'wet' land* tend to hold more sway in models than *Path distance to prehistoric wetlands*. This may reflect the nature of the archaeological database, which is dominated by more recent archaeological sites. If the prehistoric hydrographic model is used to help determine future survey locations, archaeologists may be able to find additional ancient sites.

Figure 4: Performance (%IncMSE) of Three Wetland Variables in Best Site Models, by Region



Any future modeling should consider whether to reduce the number of hydrographic variables used. These variables are frequently correlated with each other to the extent that one of the correlated pair must be removed from the model. As one example, *Cost path distance to nearest large historic lake* was removed from eleven models because it was correlated with another variable, usually *Cost path distance to nearest large prehistoric lake*. *Cost path distance to nearest* river and *Cost path distance to nearest perennial river or stream* are also frequently redundant. Hydrographic variables (most commonly *Cost-path distance to nearest historic surface water)* were also removed from some models because they could not distinguish between sites or surveys and background points. *Cost-path distance to nearest historic surface water* was also removed from some models because it displayed near-zero variance.

Geomorphology

Geomorphic variables (Table 12) performed well with one exception. Although it is accepted wisdom in Minnesota that archaeological sites are found on islands, the variable *On an Island* was removed from every site model because it failed to distinguish between sites and non-sites. This is most likely because the number of

sites on islands in any given region is a very small portion of the total site population and, also, because many islands have not been surveyed for sites. The only model in which it performed was the survey model for Border Lakes. This is best explained by a more thorough survey of islands in the Border Lakes region than elsewhere in the state.

Table 12: Performance of Geomorphic Predictor Variables in Phase 4 Site Potential Models

Variable	No. Models	Minimum	Maximum	Mean
Path Distance to Major Ridge or Divide	18	3.5	25.9	12.4
Path Distance to Minor Ridge or Divide	12	2.4	13.4	6.8
On an Island	0	0	0	0
Landform	20	1.6	35.3	11.9
Landscape	17	3.0	22.3	10.7
Major Watershed Size	17	3.2	18.7	9.5
Minor Watershed Size	18	1.8	18.9	10.4

Landform and *Path distances to major ridge or divide* were the strongest geomorphic predictors. *Landform* was a top variable in the BLUF (The Blufflands) site model. Table 13 shows clearly that the majority of background points are located on the uplands while sites are concentrated on terraces, floodplains, colluvial slopes, and alluvial fans.

Table 13: Distribution of Sites and Background Points within Landforms in The Blufflands (BLUF) Modeling Region

Landform	% Sites	% Background Points
Alluvial Fan	3.0	0.2
Bar	6.5	0.3
Colluvial Slope	15.8	8.3
Floodplain & Terrace	3.6	0.6
Floodplain, Featureless	1.9	0.5

Landform	% Sites	% Background Points
Floodplain, Undifferentiated	12.3	3.6
Hillslope	4.9	18.9
Levee	1.2	0.5
Paleochannel	4.3	0.2
Plain	0.3	6.6
Summits & Hillslopes	6.6	52.8
Terrace	38.1	4.7

Historic Vegetation

All five vegetation variables figured in the top ten of at least one site model and one survey model. The best-performing vegetation variables for site prediction were *Path distance to nearest wild rice* and *Vegetation diversity within one kilometer* (Table 14). *Path distance to wild rice* is the only variable in our models that is specific to a single important food resource. High vegetation diversity within a catchment, which we modeled at several scales, implies access to a wider range of resources, but is not specific to how useful those resources might be. It appears from these models that vegetation diversity at the most local scale, one kilometer, is most significant for site locations. *Vegetation diversity within five kilometers* failed to achieve a better than average %IncMSE in any site model.

Table 14: Performance of Vegetation Predictor Variables in Phase 4 Site Potential Models

Variable	No. Models	Minimum	Maximum	Mean
Path Distance to Wild Rice	17	5.3	23.9	14.2
Vegetation Diversity within 10 km	16	1.8	15.6	9.4
Vegetation Diversity within 1 km	19	-0.5	32.1	9.7
Vegetation Diversity within 5 km	17	2.1	18.9	9.3
Vegetation Type	17	0.1	23.2	10.1

The dominant local *vegetation type* performed best in the Chippewa Plains (CHIP) region. Table 15 compares the distribution of sites and background points among the most common vegetation types of this region. Sites are more likely to be in 'wet' land or rivers than are background points. Since these are imprecisely mapped, it is safe to assume that sites are more likely to be close to 'wet' land (which may surround lakes) and historic rivers. Sites are also more likely to be in areas mapped as red pine forest and maple-basswood forest. Sites seem less likely to be within most other vegetation types, but particularly conifer swamps, white pine forests, boreal hardwood-conifer forest, and aspen forest. Much more analysis would be required to understand these observed distribution patterns.

Table 15: Distribution of Sites and Background Points within Vegetation Types in the Chippewa Plains (CHIP) Modeling Region

Landform	% Sites	% Background Points
Lake	6.1	6.4
'Wet' Land	1.1	0.6
River	1.7	0.1
Conifer Swamp	28.5	34.3
Marsh	1.1	1.1
Jack Pine Forest	11.0	9.8
Red Pine Forest	35.8	9.1
White Pine Forest	1.0	5.5
Spruce-Fir Forest	0.5	1.1
Boreal Hardwood-Conifer Forest	5.4	11.0
Mixed Pine-Hardwood Forest	0.5	1.7
Aspen Forest	2.6	7.6
Paper Birch Forest	0.5	1.3
Lowland Hardwood Forest	0.1	0.7
Maple-Basswood Forest	1.8	0.6

Landform	% Sites	% Background Points
Northern Hardwood Forest	0.1	1.9
Oak Forest	0.1	1.5
Oak Savanna	0.4	1.2
Aspen Woodland	0.2	1.4

Pedestrian Transportation

Pedestrian transportation variables derived from the FETE model performed well (Table 16), with the exception of *Order of nearest path.* Major paths have the most impact on models in regions where they are present. For site prediction, *Path distance to nearest major path* was the most important variable in Chippewa Plains with a %IncMSE value of 35.5. Proximity to lower level paths are also important, especially where major paths are distant. *Path distance to nearest intermediate path* and *Path distance to nearest minor path* both achieved better than average %IncMSE values in at least one site model. *Path distance to nearest intermediate path* was the top variable in the survey model for St. Paul-Baldwin Plains & Moraines. The order of the nearest path seems to be less important than its proximity.

Table 16: Performance of Pedestrian Transportation Variables in Phase 4 Site Potential Models

Variable	No. Models	Minimum	Maximum	Mean
Path Distance to Major Path	8	9.7	35.5	20.0
Path Distance to Medium Path	14	2.2	24.1	13.0
Path Distance to Minor Path	20	2.5	21.4	9.8
Order of Nearest Path	3	3.0	4.9	3.8

Soils

Site models developed using soils data were the best models in 11 of the 20 regions modeled. Soils data performed better when predicting survey locations and were selected for use in 14 of the 20 regions. However, soil variables were not particularly strong predictors (Table 17). Only *Wetland Soils* was relatively strong, but that variable is redundant with prehistoric wetlands and, because it figured into the vegetation model, with historic wetlands. *Flooding frequency* was nearly always removed from the models as it failed to distinguish between sites and non-sites.

Table 17: Performance of Soil Variables in Phase 4 Site Potential Models

Variable	No. Models	Minimum	Maximum	Mean
Drainage Index	3	5.2	8.2	6.5
Soil Drainage	7	2.2	14.0	8.1
Frost-Free Days	8	1.8	13.1	6.2
Flooding Frequency	0	0	0	0
Hydric Group	9	-0.5	10.4	5.3
Horizon Depth	4	5.6	11.9	8.5
Productivity Index	11	2.6	12.9	5.7
Wetland Soils	14	1.9	27.9	7.2

Discussion

A thorough understanding of the role of each variable in the predictive models would require quite a bit of additional analysis. Not only does the performance of a variable vary between regions and between site and survey models, it may also vary between two models of the same type simply because different sample points are used.

This report has focused primarily on the performance of variables for predicting archaeological site locations. It would be informative to compare the performance of variables predicting actual site locations to that of variables predicting survey locations, as surveys tend to be in locations where archaeologists expect sites to be found. This may help elucidate the differences between archaeologists' mental models, based on the information they have traditionally used to determine where to survey, and the patterns detected by the statistical models based on the data used for this project. This analysis could identify patterns of site/environment relationships associated with site that have not previously been recognized by archaeologists. These patterns could be field tested with new surveys.

Conclusions

Even with the very coarse level of analysis presented here, we can make some general statements about the usefulness of variables for future modeling. First, several variables have shown to have little utility for modeling and may be discarded. These include *Surface roughness within 90 meters* (RGH90), *Shelter index* (SHELTER), *Topographic Position Index within 1000 meters* (TPI1000), *Path order* (PATH_ORD), *Path distance to nearest*

historic surface water (CP_WAT), *Drainage Index* (DI), *Flooding frequency* (FLDFRQD), and *On and island* (ISLAND). Second, it would be worthwhile to further examine variables that tend to be redundant with one another, such as the various 'distance to lake' variables, to determine if any can be discarded for simplicity. Third, careful consideration should be given to simplifying both the landscape and vegetation models so that rare landforms and vegetation types are minimized without blurring distinctions between categories that show significant relationships with site distributions.

It is unlikely that refining the variables would make much difference in the models at this time. The Phase 4 models are very precise and do an excellent job predicting the distribution of the archaeological sites in the current database. However, these recommendations should be considered when enough new site and/or survey data have been collected to make the effort to develop new models worthwhile.

References

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

Cleland, D.T., P.E. Avers, W.H. McNab, M.E. Jensen, R.G. Bailey, T. King, and W.E. Russell
1997 National Hierarchical Framework of Ecological Units. In *Ecosystem Management Applications for Sustainable Forest and Wildlife Resources,* edited by M.S. Boyse and A. Haney, pp. 181-200. Yale University Press, New Haven, CT.

Giovani, B. and R.F. Goldman
1971 Predicting metabolic energy cost. *Journal of Applied Physiology* 30:429-433.

Guisan, A., S.B., Weiss, and A.D. Weiss
1999. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology* 143: 107-122.

Hammer, John
1993 A New Predictive Site Location Model for Interior New York State. *Man in the Northeast* 45:39-76.

Hanson, D.H. and B.C. Hargrave
1996 Development of a Multilevel Ecological Classification System for the State of Minnesota. *Environmental Monitoring and Assessment* 39:75-84.

Herzog, Irmela
2014 Least-cost Paths – Some Methodological Issues. Internet Archaeology, Issue 36.

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
2019a *Historic/Prehistoric Hydrographic Models for Minnesota: MnModel Phase 4*. Minnesota Department of Transportation. St. Paul, MN.

Hobbs, Elizabeth, Andrew Brown, Alexander Anton, Jeffrey Walsh, Carson Smith, and Luke Burds
2019b Preparing Data for Modeling. Appendix B in *MnModel Phase 4 User Guide*, by Carla Landrum et al. Minnesota Department of Transportation. St. Paul, MN.

Kvamme, Kenneth L. and Timothy A. Kohler
1988 Geographic Information Systems: Technical Aids for Data Collection, Analysis and Display. In *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modeling,* edited by W. James Judge and Lynne Sebastian, 493-548. U.S. Government Printing Office, Washington, D.C.

Landrum, Carla, Elizabeth Hobbs, Alexander Anton, Andrew Brown, and Luke Burds
2019 *Archaeological Predictive Modeling Guide: MnModel Phase 4*. Minnesota Department of Transportation. St. Paul, MN.

Lively, R.S., G.B. Morey, and E.J. Bauer
2002 *One hundred years of mining: alterations to the physical and cultural geography of the western half of the Mesabi Iron Range, northern Minnesota*. MGS Miscellaneous Map Series, M-118, 4 pls. Scale 1:100,000. Paper plots. Pl. 1, land-surface topography; pl. 2, drainage and cultural features; pl. 3, topographic disturbance; pl. 4, bedrock geology. Minnesota Geological Survey. 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.

Soule, Roger G. and Ralph F. Goldman
1972 Terrain coefficients for energy cost prediction. *Journal of Applied Physiology* 32:706-708.

Stark, Stacey L., Patrice M. Farrell, and Susan C. Mulholland
2008 *Methods to Incorporate Historic Surface Hydrology Layer in Mn/Model [Phase 4] Using Existing Geographic Information System Data*. Minnesota Department of Transportation. St. Paul, MN.

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.*

Appendix A: Variable List

Table A1: Complete List of MnModel Phase 4 Predictor Variables

VARIABLE	DEFINITION	ALLARCHLIST	SOILARCHLIST
ASP_RNG	Aspect range	X	X
CP_BOG	Path distance to nearest historic bog	X	X
CP_FLOOD	Path distance to nearest historic floodplain	X	X
CP_INT	Path distance to nearest intermittent stream	X	X
CP_LAKE	Path distance to nearest historic lake	X	X
CP_LLK	Path distance to nearest large historic lake	X	X
CP_MAJPATH	Path distance to nearest major pedestrian transportation route	X	X
CP_MAJRIDGE	Path distance to nearest major ridge or divide	X	X
CP_MARSH	Path distance to nearest historic marsh	X	X
CP_MEADOW	Path distance to nearest historic wet meadow or fen	X	X
CP_MEDPATH	Path distance to nearest medium pedestrian transportation route	X	X
CP_MINPATH	Path distance to nearest minor pedestrian transportation route	X	X
CP_MINRIDGE	Path distance to nearest minor ridge or divide	X	X
CP_PEREN	Path distance to nearest perennial stream	X	X

VARIABLE	DEFINITION	ALLARCHLIST	SOILARCHLIST
CP_PFLOOD	Path distance to nearest prehistoric floodplain	X	X
CP_PLAKE	Path distance to nearest prehistoric lake	X	X
CP_PLLK	Path distance to nearest large prehistoric lake	X	X
CP_PWET	Path distance to nearest prehistoric wetland	X	X
CP_RICE	Path distance to nearest wild rice location	X	X
CP_RIVER	Path distance to nearest river	X	X
CP_SWAMP	Path distance to nearest historic swamp	X	X
CP_WAT	Path distance to nearest historic surface water (of all types)	X	X
CP_WET	Path distance to nearest historic 'wet' land	X	X
CP_WETLAND	Path distance to nearest historic wetland (of any type)	X	X
CURV	Surface Curvature	X	X
DI	Drainage Index		X
DRAIN	Soil drainage		X
ELEV	Elevation	X	X
FFD_R	Frost-free days		X
FLDFRQD	Flood frequency		X
HYDGRPDCD	Hydric Group (dominant condition)		X

VARIABLE	DEFINITION	ALLARCHLIST	SOILARCHLIST
HYDPRS	Hydric soil presence		X
HZDEP	Depth of surface soil horizon		X
ISLAND	On an island	X	X
LFORM	Landform	X	X
LSCAPE	Landscape	X	X
MAJ_SIZE	Size of major watershed	X	X
MIN_SIZE	Size of minor watershed	X	X
ORD_STRM	Order of nearest stream	X	X
PATH_ORD	Order of nearest pedestrian transportation route	X	X
PI	Productivity Index		X
REL	Relative Elevation	X	X
REL90	Relative Elevation within 90 meters	X	X
RGH	Surface Roughness	X	X
RGH90	Surface Roughness within 90 meters	X	X
SHELTER	Shelter Index	X	X
SLOPE	Percent Slope	X	X
TPI1000	Topographic Position Index within 1000 meters	X	X
TPI1MI	Topographic Position Index within one Mile	X	X

VARIABLE	DEFINITION	ALLARCHLIST	SOILARCHLIST
TPI250	Topographic Position Index within 250 meters	X	X
TPI5MI	Topographic Position Index within five miles	X	X
TPI90	Topographic Position Index within 90 meters	X	X
TWI	Topographic Wetness Index	X	X
VEGDIV10K	Vegetation diversity within ten km	X	X
VEGDIV1K	Vegetation diversity within one km	X	X
VEGDIV5K	Vegetation diversity within five km	X	X
VEGMOD	Historic vegetation type	X	X
VISIBLE	Visibility	X	X
WETSOIL	On a wetland soil	X	X