PENNSYLVANIA DEPARTMENT OF TRANSPORTATION
ARCHAEOLOGICAL PREDICTIVE MODEL SET

TASK 4: Study Regions 1, 2, and 3

CONTRACT #355I01

ARCHAEOLOGICAL PREDICTIVE MODEL SET

Category #05 – Environmental Research

July 2014

URS

PENNSYLVANIA DEPARTMENT OF TRANSPORTATION
ARCHAEOLOGICAL PREDICTIVE MODEL SET
TASK 4: STUDY REGIONS 1, 2, AND 3

CONTRACT #355I01

Prepared for

Pennsylvania Department of Transportation
Bureau of Planning and Research
Keystone Building
400 North Street, 6th Floor, J-East
Harrisburg, PA 17120-0064

Prepared by

Matthew D. Harris, Principal Investigator
Susan Landis
and
Andrew R. Sewell, Hardlines Design Company

URS Corporation
437 High Street
Burlington, NJ 08016-4514

July 2014

ABSTRACT

This report is the documentation for Task 4 of the Statewide Archaeological Predictive Model Set project sponsored by the Pennsylvania Department of Transportation (PennDOT). This project was solicited under Contract #355I01, Transportation Research, Education, and Technology Transfer ITQ, Category #05 – Environmental Research. The goal of this project is to develop a set of statewide predictive models to assist the planning of transportation projects. PennDOT is developing tools to streamline individual projects and facilitate Linking Planning and NEPA, a federal initiative requiring that NEPA activities be integrated into the planning phases for transportation projects. The purpose of Linking Planning and NEPA is to enhance the ability of planners to predict project schedules and budgets by providing better environmental and cultural resources data and analyses. To that end, PennDOT is sponsoring research to develop a statewide set of predictive models for archaeological resources to help project planners more accurately estimate the need for archaeological studies.

The objective of Task 4, discussed in the following report, is to create a series of archaeological predictive models for Regions 1, 2, and 3 of Pennsylvania. In total, this area covers 17,677 square miles which is 38-percent of the state. These three regions cover much of western Pennsylvania and the Appalachian Plateaus physiographic provinces. A total of 8,126 prehistoric archaeological components were incorporated into this modeling effort. Thirty spatially separate models were created to cover these three regions. The final model ensemble composed of these 30 separate models correctly classifies 100-percent of the known archaeological sites within a high and moderate sensitivity area of 30.8-percent of the three regions with an average RMSE prediction error of 0.124.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1
INTRODUCTION

The purpose of this project is to use the existing Pennsylvania Archaeological Site Survey file database (PASS) to produce a baseline model for the sensitivity of prehistoric site-presence throughout the entire Commonwealth. The resulting assessments of archaeological sensitivity will be used by transportation, planning, and other Cultural Resource Management (CRM) practitioners to make better-informed and more consistent assessments of prehistoric archaeological sensitivity, with the ultimate goal of saving time, money, and sparing cultural resources.

Building off of the previous task in this project, the creation of a pilot model for central Pennsylvania, this report documents the first in a series of three tasks that apply the modeling methodology to the entire state. This report details the creation, findings, and conclusions of predictive models created for Regions 1, 2, and 3 (Figure 1). These regions comprise a total of 17,677 square miles, 38-percent of the entire state. Covering almost the entirety of western Pennsylvania, this process involved creating 30 individual models from a dataset of over 8,000 prehistoric archaeological sites.

The process reported on below developed three statistical models (logistic regression, adaptive regression splines, and random forest) for each of the 30 subareas. Each of these three model types is discussed and detailed in the previous Task 3 report. Ultimately, the random forest models were used to represent each of the 30 subareas because of their accuracy and ability to discriminate archaeological site locations. The end result of this process is the classification of a high, moderate, and low sensitivity model that covers the entirety of each of the three regions. The report below documents the model building process, as well as, the breadth of previous modeling attempts in the regions, the prehistoric context of the area, an assessment of PASS data quality, and special topics of concern for the modeling process.

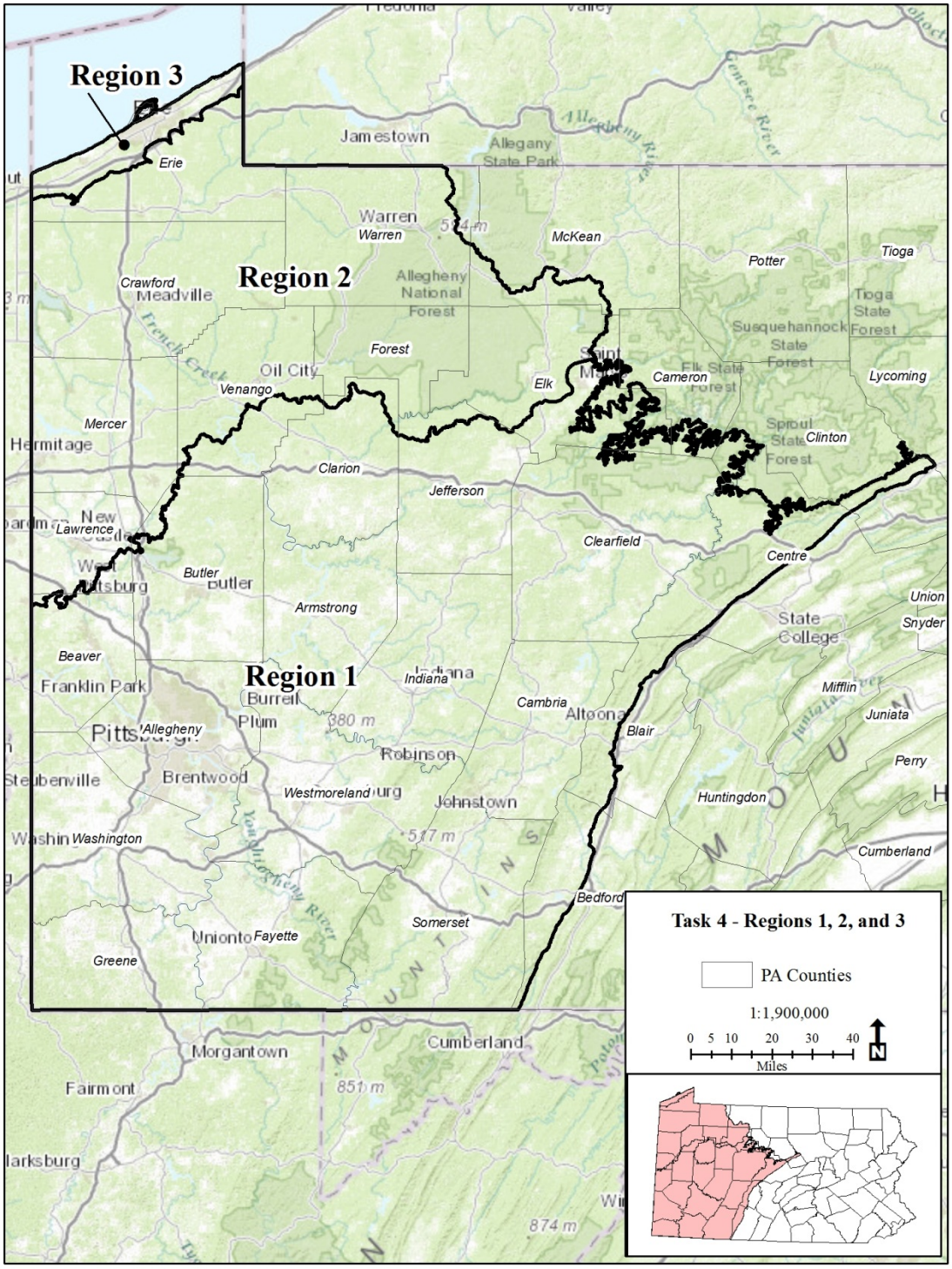


Figure 1 - Overview of Regions 1, 2, and 3

PREDICTIVE MODELING IN REGIONS 1, 2, AND 3

Region 1

Numerous archaeological predictive model (APM) studies have been undertaken within Region 1, many for compliance-related projects. Because of this association, the models often focused on an area determined by the location of the project, and not generated to answer specific questions about settlement patterns. Several predictive models generated in Region 1 did not attempt to predict anything beyond a general archaeological sensitivity for prehistoric resources (for example, Michael and Herbstritt 1980; Herbstritt 1981; Means et al. 1998; Baublitz et al. 2003; Coppock et el. 2003), mainly due to issues with the resolution of environmental data and concerns about the accuracy of PASS data. Coppock et al. (2003:8) noted in particular that much of the site data in the PASS files was generated from interviews of collectors and submittals by avocational archaeologists, and thus the level of detail about site location, function, and structure was considered insufficient to predict site locations by type and temporal association with the accuracy required of an effective predictive model. One example of a non-site type or temporal period-specific model is one developed by Coppock and Heberling (2001), which divided an upland study area into five zones of probability: a rock shelter zone, a high-probability stratified site zone (greater than 1 m depth below surface), a high-probability shallow site zone (less than 1 m depth below surface), a moderate-probability shallow site zone, and a low site potential zone (corresponding to areas that had been surface mined). Duncan (2002) found that within a region in Centre and Clearfield counties that had a low site density, it was useful to collapse the prehistoric temporal periods into three broad categories of Early Hunter/Gatherer (corresponding to the Paleoindian and Early Archaic periods), Archaic Hunter/Gatherer (corresponding to the Middle Archaic through Middle Woodland periods), and Agriculturalist (corresponding to the Late Woodland period).

Some predictive models produced useful generalizations about landform preferences by prehistoric groups during different temporal periods, which may be applicable across Region 1 as a whole. Cowin (1981) in particular produced a detailed predictive model based on a survey of 289 sites within Region 1. She found that the majority of sites were located on terraces across all time periods, although Late Woodland sites showed a large drop in percentage of sites located on that landform. The model produced by Cowin (1981:62) featured six key predictions for site locations:

1. Flood plains and terraces are to be considered as having high archaeological potential, especially when they coincide with confluences.
2. Nearly all the sites in the sample occurred on well-drained soil types.
3. Sites located on hilltops, benches, saddles, and hillslopes were all in proximity to water sources.
4. Upland sites are found in locations with conditions to maximize positive weather effects (sunlight) and minimize negative weather effects (rain, prevailing winds); this translates to sites being located below the highest elevation and on south and east faces of slopes.

5. Large villages tend to be found on terraces and saddles or benches with less than 8 degree slopes.
6. Medium to large sites will be found in locations with a high diversity of resources and close to lithic sources.

Duncan et al. (1996) evaluated a previously generated predictive model for the Crooked Creek watershed in Allegheny and Washington counties, and improved the model using GIS and revisions to the size of the sample quadrants. They found that the majority of sites in their study area were from the Woodland period, especially the Late Woodland, with a good number of Late Archaic sites as well. These sites showed a preference for upland settings, with lowland sites generally limited to terraces, although the authors noted that a lack of flood plain settings within the survey area may have biased the apparent preference for upland settings. Using the same GIS model architecture, Duncan and Schilling (1999) examined a project area in Fayette and Washington Counties. They noted that 86% of the sites in the site population used in their model were located in uplands, and most were unidentifiable as to function, followed by camps and villages (Duncan and Schilling 1999:11). Katz et al. (2002), in a study performed in Allegheny and Butler Counties, found that proximity to water was the primary variable in predicting site locations. Flood plains and terraces along major rivers, especially at stream junctions and locations of historically known Indian trail crossings, were considered high probability for archaeological sites. Upland sites tended to be located mostly on landforms situated below a topographical high point, such as saddles and benches. Katz et al. (2002) found that soil drainage, while important, was not as valuable an indicator of site probability as one may have expected; rather, the degree of slope was much more important, with the majority of sites found on slopes of 8 degrees or less.

Region 2

One study using predictive modeling was located for the Region 2 study area. This study was produced to predict archaeologically sensitive areas within two units of the Erie National Wildlife Refuge (Glenn 2010). The model used known site locations and environmental factors including slope, cost-distance to water, cost-distance to confluences, cost-distance to prime farmland, and hydric soils, and assigned rankings to each factor. A GIS was used to rasterize the study areas into 30-m-square grids, and each factor was mapped according to its ranking in each grid square. The combined ranking of factors was used to identify areas with high, medium, low, and unlikely sensitivity for archaeological sites. Data from the population of known sites used in the study indicated that most would be short-term camps, with 40% located on flood plains and only 12% in uplands. Fully 89% of sites were located on slopes of 10% or less. In addition, the cost-distance analysis showed that the majority of sites were located at the first natural break in the data, indicating that sites are most likely to be found at locations with the least path of resistance to water.

Two other reports had information useful for consideration in predictive model building, but did not attempt to either create or apply a predictive model as part of the work that was documented. In

1980–1982, the PHMC regional archaeologist recorded 62 prehistoric sites during a site survey in what was then called Survey Region IV (now including part of Region 2). As a result of the study, the regional archaeologist determined that multi-component or intensive utilization sites in Crawford County around the Pymatuning March and Conneaut Lake were mainly located on low rises, while upland areas were sparsely used (Johnson 1981:33). In 2004 Christine Davis Consultants, Inc., conducted a 530-acre archaeological survey on flood plains and kame terraces along the Shenango River in Lawrence County. Their work suggests that prehistoric sites within their project area focused on confluences with the Shenango River, and that people exploited resources associated with wetlands in the area (Davis et al. 2004).

Region 3

The small number of sites identified within Region 3 makes the creation of generalizations about site types and landform preference difficult to generate from PASS data. The difficulty of this analysis is increased as a result of the lack of large-scale surveys containing synthetic statements regarding settlement patterns. Hart (1994), however, created a general prehistoric site predictive model for the Lake Plain and Glacial Escarpment physiographic regions within Erie County, using data from the same physiographic regions in neighboring Ohio and New York. Hart determined that prehistoric "sites tend to be located on relatively level, well-drained landforms that are near streams, and at high elevations above streams and stream confluences" (Hart 1994:8) within the Lake Plain region; in other words, a mix of lowland and upland settings close to water sources, which is borne out by a review of the PASS data. Hart's study, however, did not attempt to predict site locations by time period or cultural phase, or by site type, which limits its application to the current study.

2
STUDY AREA – REGIONS 1, 2, AND 3

PHYSICAL CHARACTER

All of Regions 1 and 2 are located within the Appalachian Plateaus physiographic province, which occupies much of the western and northern portions of Pennsylvania on the western side of the Appalachian Mountain formation. Four sections of the Appalachian Plateaus fall within Region 1 (Allegheny Mountain, Allegheny Front, Waynesburg Hills, and Pittsburg Low Plateau), while two sections are within Region 2 (High Plateau and Northwestern Glaciated Plateau). Region 3 is located within the Central Lowlands Province on the southeast shore of Lake Erie and includes just one section (Eastern Lake) (Table 1; Figure 2).

Table 1 - Physiographic Provinces and Sections for Modeling Regions 1, 2, and 3

Modeling Region	Physiographic Province	Physiographic Section
1	Appalachian Plateaus	Allegheny Mountain
		Allegheny Front
		Waynesburg Hills
		Pittsburgh Low Plateau
2	Appalachian Plateaus	High Plateau
		Northwestern Glaciated Plateau
3	Central Lowlands	Eastern Lake

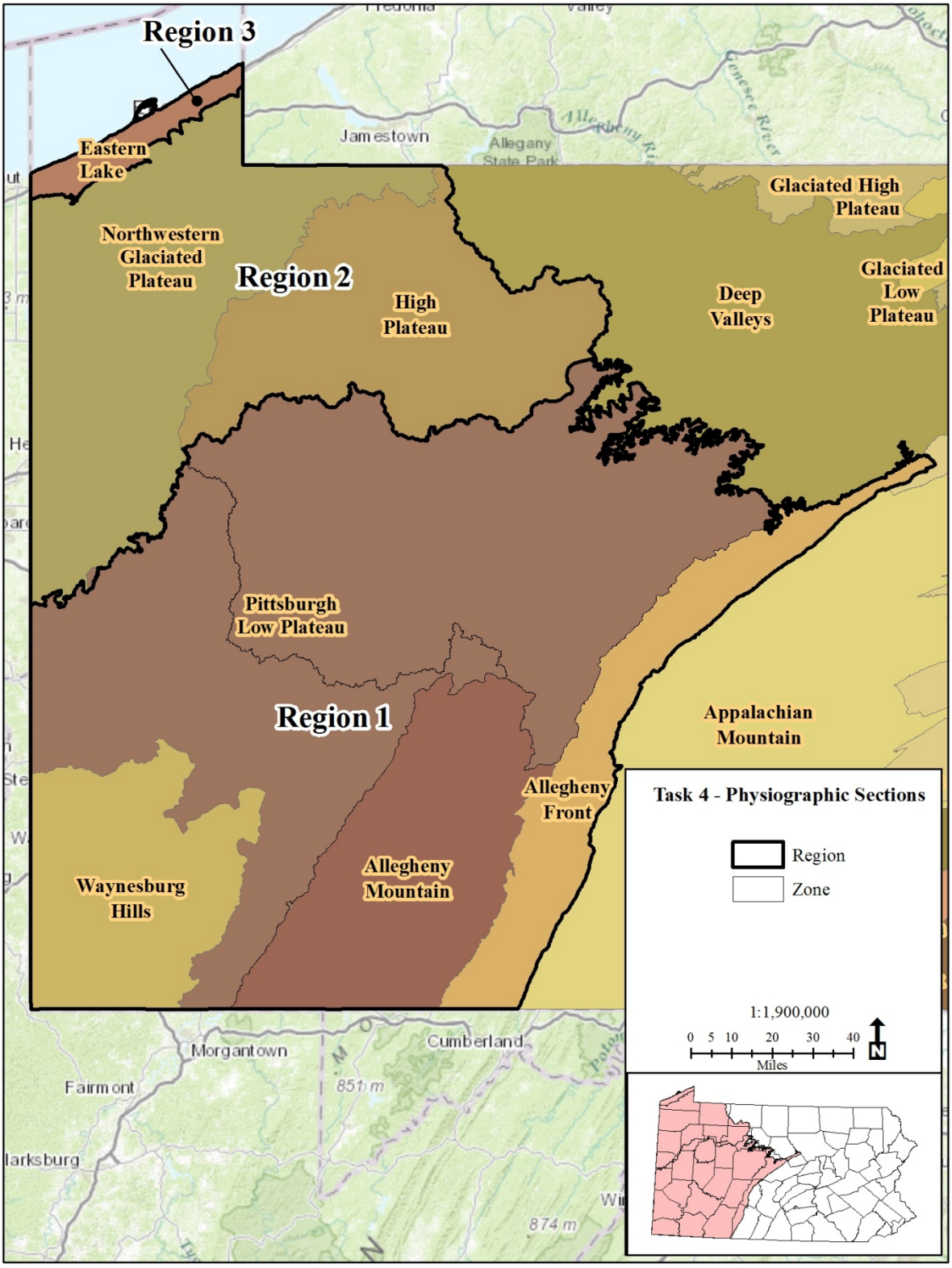


Figure 2 - Regions 1, 2, and 3 physiographic sections.

Appalachian Plateaus

Allegheny Mountain

The Allegheny Mountain section is surrounded on two sides (west and north) by the Pittsburgh Low Plateau section and is bordered on the east by the Allegheny Front section. The eastern boundary of the section is drawn between coal and non-coal producing areas. The western boundary follows the base of the west flank of Chestnut Ridge. The northern extent of the section is defined by the approximate northeast terminus of the area's large amplitude and open folds that make up the section's geologic structure. The underlying rock types typically found in the Allegheny Mountain section are sandstone, siltstone, shale, and conglomerate stone. Also in the area, much like the bordering Pittsburgh Low Plateau section, is some limestone and coal inclusions. The dominant topographic form of the section is made up of wide ridges separated by broad valleys. As the ridges stretch further north, the elevation decreases; overall, the section has a minimum elevation of 775 feet amsl and a maximum of 3,210 feet amsl. The local relief of the section is moderate to high (601 to >1,000 feet). The land formations, waterways, and carved surface that define the Allegheny Mountain section were created by the occurrence of fluvial erosion and some periglacial mass wasting. This means the area is subject to perennially frozen ground or permafrost as well as a seasonal thaw. With the massive amount of annual water movement caused by seasonal changes, the drainage pattern of the section created branch-like waterways called dendritic because of the way they mimic oak or maple trees.

Allegheny Front

The Allegheny Front section constitutes the eastern boundary of the Appalachian Plateaus province. The topography of the section is dominated by hills that were created by a combination of fluvial (river/stream) erosion and periglacial mass wasting (e.g., freeze/thaw creep, slow flow of soils above the permafrost resulting in events such as landslides, rock fall, etc.). The eastern half of the section is characterized by rounded to linear hills that gradually rise in stepped elevation as they approach the escarpment and the beginning of the Appalachian Mountain section to the east. The eastern boundary is delineated by a stream that runs along the base of the escarpment. These hills are crisscrossed by narrow valleys that separate the individual hilltops and often appear stepped as they form escarpments/cliffs as the elevation increases. The effect is of a series of hilly stairs rising to the east to meet with the Appalachian Mountains. The western half of the section slopes away to the west in a series of undulating hills. The underlying bedrock throughout this section is dominated by shale, siltstone, and sandstone. The geologic structure of the section is fairly uniform, characterized by beds having a low northwest dip, with the occasional fault making an appearance. The exception to this is in the southernmost portion of the section, which can be differentiated by the transition to geologic structure typified by broad folds. The elevation throughout the section is highly variable, ranging from just 540 feet amsl to as much as 2,980 feet amsl. This variance is due in large part to the

undulating land surface and deep cut valleys between hills and escarpments. The drainage pattern for this area is classified as both parallel and trellis. While the eastern edge of this section is defined by an escarpment, the western edge is largely arbitrary but is loosely delineated by the boundaries of coal fields.

Waynesburg Hills

The Waynesburg Hills section is located in the southwest corner of Pennsylvania and spills into West Virginia over the western and southern border. The boundaries of the section are arbitrary lines that follow the change of topography along the borders of the Pittsburgh Low Plateau section to the north and east. The topography is made up of very hilly terrain and narrow hilltops. The section also includes narrow valleys with steep slopes and moderate local relief. The elevation of the section has a much smaller range than that of the surrounding Pittsburgh Low Plateau section, with elevations between a 848 and 1,638 feet amsl. The origin of the section and formation of the topography was not only caused by fluvial erosion but also included the occurrence of landslides and shifting surfaces over time. The drainage pattern of the section is dendritic and reflects that of the Pittsburgh Low Plateau section, being directly attributable to fluvial erosion. The underlying rock types within the section are similar to the surrounding areas with sandstone, shale, red beds, and limestone. The absence of coal deposits in the section and the geologic structure of horizontal beds contribute to the arbitrary boundaries.

Pittsburgh Low Plateau

The Pittsburgh Low Plateau section covers an immense land area located to the west of the Appalachian Mountains and north of the Allegheny Mountains. It continues all the way to the western edge of the state and into Ohio and West Virginia and is defined to the northwest by the southern extent of glaciation. The dominant topography of the area is variable. Some areas are relatively smooth while other areas are characterized by a more undulating ground surface. Shallow, narrow valleys are littered across the landscape. These small valleys form a dendritic drainage network that is typical of the section. The topography of the section was created by a combination of fluvial erosion and periglacial mass wasting. It is noteworthy that due to human activity and centuries of mining there are many areas that are composed of strip mines or reclaimed mine land. Human modification of the environment has had an important geologic and topographic impact on the terrain and topography of this section and is in part what defines it. While the underlying rock contains much of the same sandstone, siltstone, and shale found in the Allegheny Front section, the Pittsburgh Low Plateau also includes the notable additions of limestone and coal. The geologic structure of the section is variable but in general is typified by moderate to low amplitude open folds, which are responsible for the undulating appearance of the land surface. In the northern portion of the section the appearance of these open folds decrease and the land surface levels out. While the section is as a whole far flatter than the Allegheny Front section, it too ranges in elevation from 600 to 2,340 feet amsl.

High Plateau

The High Plateau section is set to east and south of the Northwestern Glaciated Plateau section, to the west of the Deep Valleys section, and to the north of the Pittsburgh Low Plateau section. The boundaries of the section include a glacial border to the northwest and margins of deep valleys to the northeast. The southern border is an arbitrary drainage divided between coal and non-coal deposits. The geologic structure of the section consists of low-amplitude open folds including rock types of sandstone, siltstone, shale, conglomerate stones, and some locations containing coal. The dominant topographic forms of the section include broad, rounded to flat uplands having deep angular valleys with moderate to high local relief. With the topography across the area incorporating high uplands and deep valleys, the approximate elevations range from 980 to 2,360 feet amsl. The origins of the topography are from fluvial erosion and periglacial mass wasting including perennially frozen ground and seasonally thawed ground. Due to the heavy movement and energy that formed the section, a dendritic drainage pattern was formed.

Northwestern Glaciated Plateau

The Northwestern Glaciated Plateau section covers the northwestern area of Pennsylvania, located to the south of Lake Erie, and continues into Ohio as well as New York. The section abuts the High Plateau section, the Eastern Lake section, and the Pittsburgh Low Plateau section. To the northwest, the boundary is along the base of an escarpment, and the boundary to the southeast is a glacial border. The dominant topography of the area consists of a broad, rounded upland and deep, steep-sided, linear valleys partly filled with glacial deposits. The topography of the section was formed by fluvial and glacial erosion with glacial deposition throughout the section. The glacial deposition refers to the sediment and minerals that are shed by the glacier while moving, or once melting begins to take place. Due to the shaping of the topography by fluvial and glacial erosion, a dendritic drainage pattern was formed throughout the section. The underlying rock type of the section comprises shale, siltstone, and sandstone creating a geologic structure of sub-horizontal beds. The elevation of the Northwestern Glaciated Plateau section varies between 900 and 2,200 feet amsl. The local relief of the section or radical elevation increases is defined as very low to moderate (0–600 feet).

Central Lowlands

Eastern Lake

The Eastern Lake section is located at the most northeastern corner of Pennsylvania and abuts the Northwestern Glaciated Plateau section to the south. The established boundaries that delineate the section are Lake Erie to the northwest and the base of escarpment to the southeast. The origin or sculpting of the Eastern Lake section was done by glacial, lake, and fluvial deposition. This

deposition occurred when the glaciers or lake ice transported large amounts of stone and sediment before shedding it in the section when melting ensued. The drainage pattern throughout the section is a parallel waterway system whereby the waterways flow parallel or sub-parallel to one another over a considerable area. The high energy water movement and shaping of the land created topographic forms with the dominant land form consisting of a northwest sloping hill, lake parallel, and low relief ridges. The approximate elevation of the section ranges from 570 to 1,000 feet amsl, and the local relief (drastic elevation changes) of the section is categorized as very low to low (0–300 feet). The underlying geologic structure incorporates horizontal or low southern dip beds. The most common rock types found throughout the section are shale and siltstone.

PREHISTORIC BACKGROUND

The following review of regional contextual studies is organized by the cultural-historical divisions used by the PHMC. Namely, these are the Paleoindian, Early Archaic, Middle Archaic, Late Archaic, Terminal Archaic, Early Woodland, Middle Woodland, and Late Woodland periods. It is acknowledged that different areas within the Commonwealth and the Middle Atlantic region often utilize variations of these periods or recognize different cultural-historical periods altogether. For example, the use of Ohio Valley temporal periods such as Adena and Fort Ancient may be common in some areas of western Pennsylvania, but not in the central or eastern part of the state. Regionally specific terms for temporal periods will be cited where appropriate within this overview.

The Peopling of the Americas and the Paleoindian Period

The first people likely reached North America no earlier than 30,000 years ago. The chronology of the Paleoindian period in Pennsylvania includes a Pre-Clovis era dating from about 14,000 to 9,500 B.C (Quinn et al. 1994), which is largely supported through the extensive research performed at Meadowcroft Rockshelter. Meadowcroft Rockshelter has a minimum early date of 9300 B.C., although Carr and Adovasio (2002:7) argue that the average date of the deepest deposits point to a Pre-Clovis occupation by 13,950 B.C. The Pre-Clovis material is marked by a distinct prismatic blade industry at Meadowcroft Rockshelter (Quinn et al. 1994).

Most evidence of early human occupation in eastern North America is associated with the Paleoindian period (9500 B.C. to 8000 B.C.), which is characterized primarily by its lithic assemblages. Fluted projectile points, usually produced from high-quality chert, are generally considered the diagnostic marker of the time period. In Pennsylvania, the Clovis point is the most commonly recovered Paleoindian point type, followed in lesser frequency by Gainey, Barnes, Crowfield, Holcombe, Beach, and Plano types (Carr and Adovasio 2002:17).

Boyd et al. (2000:38) note that Paleoindians in the eastern United States likely had a settlement pattern in which a small group would be highly mobile through part of the year, and then practice a semi-sedentary lifestyle the rest of the year, according to the specific seasonally available resources

that were the focus of subsistence at any particular time. This pattern results in two basic types of Paleoindian sites within Region 1: base camps and short-term resource procurement camps. The short-term camps subsume other specialized site types, such as hunting stations, quarries, and isolated point finds. Boyd et al. (2000:43) also use the same site types for the subsequent Early Archaic period.

The Archaic Period

The Archaic period is the longest documented temporal segment of prehistory in eastern North America. In Pennsylvania, it is typically divided into the three periods of Early Archaic (8500–6000 B.C.), Middle Archaic (6000–4000 B.C.), Late Archaic (4000–1800 B.C.), and Terminal Archaic (1800–1000 BC), based on the marked differences in subsistence and settlement patterns (Quinn et al. 1994).

Small bands of Early Archaic hunter-gatherers appear to have been highly mobile and may have traveled across large territorial ranges and a variety of landforms (Jefferies 1990:150). Raber et al. (1998:121) note that Early Archaic lifeways show a high degree of continuity with the preceding Paleoindian period. In a recent study, Purtill (2009:569) suggests that seven distinct horizons are visible within the Early Archaic period based on projectile point usage patterns. These horizons include morphologically similar hafted bifaces that were contemporary in use: Early Archaic Side-Notched, Charleston, Thebes, Kirk/Palmer, Kirk Stemmed, Large Bifurcate, and Small Bifurcate (Purtill 2009:569). While settlement data is scarce, one notable site within Region 2 is 36ME105 in Mercer County, which yielded a postmold pattern associated with a hearth radiocarbon-dated to the Early Archaic, one of the earliest such structures to be identified in northwestern Pennsylvania (Koetje 1998:35).

By the Middle Archaic, populations had shifted their movement strategies from high mobility to reduced mobility (Stafford 1994). The appearance of ground stone tools and the related implication of increased plant usage also support the idea that Middle Archaic populations were somewhat more sedentary than those living in the region before them. Several technological innovations took place between the Early and Middle Archaic periods. Projectile point types of this time period in Pennsylvania include MacCorkle, LeCroy, St. Albans, Kanawha, Neville, Otter Creek, and Stanly (Justice 1995; Carr 1998:80); the bifurcated base is typically seen as first occurring in the early Middle Archaic. Ground stone tools such as axes, pitted stones, pestles, and grinding stones first appeared at this time (Jefferies 1996:48). In addition, archaeological evidence indicates that Middle Archaic people were also familiar with the atlatl, or spear thrower (Jefferies 1996:48). Middle Archaic sites are characterized by Boyd et al. (2000:50) as represented by the same two basic site types as the preceding periods (base camps and short-term camps), but display a tendency to exploit a wider range of topographic settings, with an increase in use of upland habitats. This expansion into the uplands is likely related to a correlated expansion of oak/hemlock forests into the same areas.

Trends first seen in the Middle Archaic, such as the increased use of plant resources, increased sedentism, and the use of cemeteries, continued into the Late Archaic period. The Late Archaic lithic assemblage is dominated by a variety of side-notched and corner-notched point types, such as the Brewerton group, as well as hafted scrapers and ground stone tools, including celts and adzes (Prufer and Long 1986; Dragoo 1976). Some evidence from sites in the southeastern United States indicates that Late Archaic populations began to experiment with fired clay (Sassaman 1993; Milanich 1994).

Chiarulli (2001) identified two Late Archaic site types, base camps and short-term camps, in her study of the Conemaugh River-Blacklick Creek within Region 1. Base camps are typically located on flood plains or near rivers, while the short-term camps are usually found in upland settings. The settings of Late Archaic sites examined by Boyd et al. (2000) show that the typical topographic settings expected for the site types are not set in stone, as one of the two base camps was located on an upland saddle, while 16 short-term camps were found in lowland settings.

The Terminal Archaic, also known as the Transitional period as evidence shows an accumulation of Woodland-like traits with a continuation of basic Archaic lifeways, is thought to be linked with a climatic change that resulted in warmer and dryer conditions (Custer 1996:187). Sites associated with the Terminal Archaic include evidence of an increase in sedentary lifestyles, with base camps occupied for longer periods. Boyd et al. (2000) note there is an apparent shift from an upland focus during the Late Archaic to a riverine focus during the Terminal Archaic. Diagnostic artifacts associated with the Terminal Archaic include the Broadspear type projectile points, such as Lehigh Broad, Susquehanna Broad, and Perkiomen Broad points (Quinn et al. 1994). Other types associated with the Transitional Archaic include the Genesee type and Snook Hill type of the Genesee cluster (Justice 1987:159). Transitional Archaic sites are often characterized by high densities of fire-cracked rock, suggesting intensive cooking techniques. Steatite bowls first appear in this period. Stewart (2003:6) notes that some Terminal Archaic sites in Pennsylvania include early ceramic use as well. In Region 1, the transition from the Late Archaic to Terminal Archaic is not distinct. Fire-cracked rock densities are high, but the diagnostic broadspears and steatite bowls are less common.

The Woodland Period

The Woodland period is generally associated with increased sedentary lifestyles and the introduction and widespread use of ceramic vessels. In Pennsylvania, the Woodland Period is usually divided into three temporal units: the Early Woodland (1000–100 B.C.), the Middle Woodland (100 B.C.–A.D. 1000), and the Late Woodland (A.D. 1000–1620). Raber (2003) notes that in Pennsylvania, especially in the east, there is difficulty in identifying and dating Early and Middle Woodland sites, due in part to scarce evidence for the distinctive Adena and Hopewell cultural traits in Pennsylvania, and to continuity with preceding Archaic lifeways. Davis et al. (2004) note that Woodland period villages and base camps were mainly located on terraces, at least those located in their study area in Lawrence County. In Regions 1, 2, and 3, the Woodland Period is marked by an apparent population

increase, as indicated by higher numbers of Woodland sites in comparison to the preceding Archaic period.

In Pennsylvania, Early Woodland settlement patterns resembled those of the Late Archaic and Terminal Archaic periods, with larger base camps situated in flood plain settings. Seasonal movement between summer base camps located on larger flood plains to upland winter camps may also have occurred (Yerkes 1988:319). Evidence for use of domesticated plants is found during the Early Woodland period, but the timing of this slight increase in domestication varies regionally and does not occur in some areas until after A.D. 100. Toward the end of the Early Woodland period, ca. 500–150 B.C., the Adena people of the central Ohio Valley directed their surplus energy into building numerous mounds, some with burials and others without burials that possibly functioned as territorial markers or aggregation loci (Yerkes 1988:317). Sites associated with the Adena culture are found in western Pennsylvania, with five sites in Region 3 yielding Adena Stemmed projectile points and one burial mound that may be Adena or Adena-influenced (36ER0136, the Green Horse mound).

Early Woodland ceramics are generally thick walled and either cordmarked, plain, or fabric-impressed. Half-Moon, Adena Plain, and Vinette I ceramics associated with the Early Woodland have all been found in Regions 1 and 2, according to the PASS database. Half-Moon cordmarked vessels associated with the Early Woodland are found in Region 1 at sites such as Meadowcroft Rockshelter (Adovasio et al. 2003:72). Marcey Creek Plain is another Early Woodland pottery type. Diagnostic projectile points include Adena Ovate Base, Robbins Stemmed, and Cresap Stemmed styles.

The Middle Woodland period is characterized by a dramatic increase in mound construction, including burial mounds and large geometric earthworks in the Ohio River Valley. In the past, researchers have equated the Middle Woodland period with the Hopewellian Interaction Sphere, a name given to the trade network of the Hopewell (Caldwell 1964). Distinctive markers of the Hopewell culture include bladelet technology, exotic artifacts in burial contexts, special purpose ceramics, and cordmarked and stamped, surface-treated ceramics (Asch and Asch 1985). True Hopewell culture sites are not known to be present in Pennsylvania, although Hopewell influence is seen at some Middle Woodland sites in western Pennsylvania, in artifacts and burial ritualism (Weed 2004:162). Diagnostic artifacts of the Middle Woodland period in western Pennsylvania include Raccoon Notched, Snyders, Levanna, and Jack's Reef projectile point types. Middle Woodland ceramics are rarely found, however, and tend to lack distinguishing characteristics (Adovasio et al. 2003:72), although ceramics from Region 2 include Mahoning Ware, a grit-tempered pottery type.

In western Pennsylvania, burial mounds similar to Hopewell burials with stone-lined cists and covered burials are found, and materials linked to the Hopewell Interaction Sphere occur on Middle Woodland sites dating between A.D. 300 and 500. Unlike Ohio Hopewell sites to the west, Middle Woodland sites in western Pennsylvania do not seem to include hamlets and other, similar sedentary site types; instead, Middle Woodland sites largely resemble those of the preceding Early Woodland

period. In addition, the large geometric earthworks of western Middle Woodland groups in Ohio, Kentucky, and Indiana are not present in western Pennsylvania.

The Late Woodland period is marked by a move toward nucleated, fortified settlements and the emergence of maize-based agricultural groups (Griffin 1967). In southwestern Pennsylvania, this new cultural phase is known as the Monongahela culture. Some of these communities were located in defensible topographic settings and were surrounded by ditches and stockades. Houses were small, arranged in a circular or semi-circular arrangement with a central plaza, and covered storage pits are frequently associated with the houses (Means 2008:8). Boyd et al. (2000:173) note that Monongahela villages were typically located in uplands at fifth-order drainage divides, apparently choosing these locations for their strategic control over the widest variety of resources within those territories.

In Region 2, the Monongahela culture is not well-represented, instead the early Late Woodland period includes Meade Island and Mahoning cultures, which were replaced or evolved into by the McFate and French Creek complexes around A.D. 1200 (Weed 2004:177). A Monongahela component was identified at the Wilson Shutes village site (36CW0005), but otherwise, the culture is sparsely documented in the PASS database for Region 2.

In Region 3, the early Late Woodland is represented by Glen Meyer complex-related groups, and the middle Late Woodland by the McFate complex (Brose 2000:99). By the end of the Late Woodland period, Region 3 was occupied by groups belonging to the Eastwall Complex, an Iroquoian or Iroquoian-influenced group (Brose 2000:96).

By the end of the Late Woodland period, villages consisted of concentric circles of houses with a large central building in the center. The Late Woodland groups in Pennsylvania had dispersed by the time of European contact. Diagnostic Late Woodland artifacts include small, triangular projectile points and grit-tempered pottery; late in the period, however, shell tempers appeared in Monongahela ceramics.

REGION 1 SITES

Paleoindian

Within Region 1, there have been 105 sites identified with Paleoindian components, according to the PASS database (Table 2). Paleoindian sites in Region 1 are largely found in topographic settings that are close to water sources, with 77 sites (73.3%) in lowland settings. Twenty-six (24.8%) Paleoindian sites were found in upland settings. Of the single component Paleoindian sites, the site types that may represent resource extraction camps (Open habitation, prehistoric; Open prehistoric site, unknown function; and Unknown function open site, greater than 20 m radius) occur primarily in locations that are typically adjacent to water sources, such as flood plains, terraces, rises in flood plains, and stream benches. This pattern also appears to apply to multi-component sites with Paleoindian material.

Table 2 - Region 1, Paleoindian Site Types by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Isolated find	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	3
Isolated fluted point locus	0	0	0	0	1	2	0	1	0	0	0	0	0	0	0	0	4
Open habitation, prehistoric	0	11	2	0	5	6	1	1	0	0	0	0	2	0	0	0	28
Open prehistoric site, unknown function	0	2	1	0	1	2	2	0	0	1	0	2	0	0	0	0	11
Unknown function open site greater than 20 m radius	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Unknown function surface scatter less than 20 m radius	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
(blank)	0	0	0	0	0	3	0	0	1	0	1	0	0	0	0	1	6
Part of multi-component site	0	16	4	1	4	12	1	0	2	1	3	1	1	1	2	1	50
Total	0	32	7	1	12	25	4	2	3	2	6	3	3	1	2	2	105

Early Archaic

The PASS database records 282 sites with Early Archaic components in Region 1 (Table 3). Early Archaic sites in Region 1 are largely found in topographic settings that are close to water sources. The single component Early Archaic sites in the PASS data that probably represent some form of resource extraction camp include Open habitation, prehistoric; Open prehistoric site, unknown function; and Unknown function, open site greater than 20 m radius. These three site types appear evenly distributed between lowland and upland settings according to the PASS data. Multi-component sites designated by landform with Early Archaic material, however, show a distribution markedly focused on settings associated with close distances to water, with 142 (64.8%) such sites on flood plains, stream benches, terraces, and similar settings, compared to 77 (35.2%) sites with Early Archaic components in upland settings.

Table 3 - Region 1, Early Archaic Site by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Isolated find	0	0	0	0	1	1	0	0	1	0	0	0	0	1	0	0	4
Lithic reduction	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Open habitation, prehistoric	0	3	1	1	4	5	1	6	2	0	0	0	4	0	0	2	29
Open prehistoric site, unknown function	0	2	0	0	2	0	1	0	1	1	0	2	0	0	2	0	11
Unknown function open site greater than 20 m radius	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3
Unknown function surface scatter less than 20 m radius	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3
(blank)	0	1	0	0	3	2	0	0	0	0	0	1	0	0	0	1	8
Part of multi-component site	1	45	5	0	36	55	9	7	14	4	3	8	21	7	4	4	223
Total	1	55	6	1	46	64	11	13	18	5	4	11	25	8	6	8	282

Middle Archaic

The PASS database includes 571 sites with Middle Archaic components in Region 1 (Table 4). Six Middle Archaic single component site types are identified in the PASS data for Region 1, including a Petroglyph/pictogram site, the Circle Rock Petroglyph (36BV0013). The Circle Rock Petroglyph is apparently associated with an adjacent find of a St. Albans point, which may be why it is classified as a Middle Archaic site. Middle Archaic sites in Region 1 are fairly evenly divided between upland and lowland topographic settings. Of the 137 single component sites classified by landform, 71 (51.8%) were found in upland settings, and 66 (48.2%) were in lowland settings. When multi-component sites with Middle Archaic components are considered, however, there appears to be more of a preference for lowland sites, with 276 (65.6%) multi-component Middle Archaic sites in lowland settings compared to 145 (34.4%) multi-component sites in upland settings. When single component Middle Archaic site types are considered, 56.1% of Open habitation, prehistoric sites are identified in lowlands; this site type represents the likeliest candidate for base camp sites. The Open prehistoric site, unknown function site type, which may represent short-term resource extraction camps, shows 61.2% of all such sites in an upland setting. Raber et al. (1998) noted that Middle Archaic resource

exploitation camps were to be found in upland settings, while base camps were located on post-Pleistocene terraces; the data for Region 1 appears to fit this observation.

Table 4 - Region 1, Middle Archaic Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Isolated find	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	3
Lithic reduction	0	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0	4
Open habitation, prehistoric	0	15	2	0	11	9	8	6	2	0	0	8	3	1	1	0	66
Open prehistoric site, unknown function	0	5	0	0	5	9	11	0	2	3	3	4	6	1	0	0	49
Petroglyph/pictograph	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Unknown function surface scatter less than 20 m radius	0	1	0	0	0	4	0	0	1	0	0	0	1	0	0	0	7
(blank)	0	0	0	0	0	0	2	2	0	1	1	0	0	1	0	2	9
Part of multi-component site	0	99	4	1	63	109	21	17	26	9	9	17	28	10	8	11	432
Total	0	121	6	1	80	134	43	25	32	13	13	30	38	13	9	13	571

Late Archaic

The PASS database includes 1,251 sites with Late Archaic components in Region 1 (Table 5). Late Archaic sites in Region 1 show a focus toward lowland topographic settings. Of the 423 single component sites with landform data, 250 (59.1%) were found in lowland settings, and 173 (40.9%) were in upland settings. The preference for lowland settings is even stronger when multi-component sites with Late Archaic components are considered, with 545 (68.4%) multi-component Late Archaic sites in lowland settings compared to 252 (31.6%) multi-component sites in upland settings. Sites with Late Archaic components are most commonly found on flood plains (n = 300, 24.6%) and terraces (n = 289, 23.7%), followed by stream benches (n = 185, 15.2%); all other landforms each have less than 10% of the total population of identified Late Archaic site types.

Table 5 - Region 1, Late Archaic Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Isolated find	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	4
Lithic reduction	0	0	1	0	4	1	1	0	1	0	0	0	1	0	0	0	9
Open habitation, prehistoric	0	60	2	0	42	49	16	21	8	5	5	11	14	6	1	5	245
Open prehistoric site, unknown function	0	13	1	0	19	11	3	3	9	6	7	7	7	4	4	2	96
Quarry	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Rock shelter/cave	0	0	0	0	2	0	0	1	0	1	1	0	0	0	0	0	5
Unknown function open site greater than 20 m radius	0	5	0	0	2	4	2	2	2	0	0	1	0	0	3	0	21
Unknown function surface scatter less than 20 m radius	0	8	0	0	4	11	3	0	0	0	0	0	0	0		0	26
(blank)	1	3	0	0	3	2	2	3	1	3	3	0	1	0	1	0	23
Part of multi-component site	1	211	15	0	108	210	31	44	39	21	17	34	34	25	7	24	821
Total	2	300	19	0	185	289	59	74	62	36	33	53	57	35	16	31	1251

Single component Late Archaic site types that may represent the likeliest candidates for seasonal occupation sites, such as base camps and short-term resource extraction camps, are Open habitation, prehistoric; Open prehistoric site, unknown function; and Unknown function open site, greater than 20 m radius. The landforms that contain the greatest number of Open habitation, prehistoric sites, which likely includes a number of base camps, are typically lowland settings, with upland landforms possessing lesser numbers of this site type. The Open prehistoric site, unknown function site type has a larger number of sites found on upland landforms, with 50 sites in that topographic setting compared to 44 in the lowlands of this site type, which may represent short-term resource extraction camps. The Unknown function open site, greater than 20 m radius site type occurs in both upland and lowland settings, with slightly more on lowland landforms (52.4%). The upland landforms associated with this site type appear to be the higher elevation types; in these topographic settings, it seems likely that this site type also represents short-term resource extraction camps.

Terminal Archaic

The PASS database includes only 324 sites with Terminal Archaic components in Region 1, perhaps reflective of the Late Archaic continuity in western Pennsylvania The majority of the multi-component sites with Terminal Archaic components also contained Late Archaic and/or Early Woodland components (n = 230), excluding Middle Woodland and Late Woodland villages that also had a Terminal Archaic component (n = 18). The fact that Terminal Archaic site components are strongly associated with preceding Late Archaic and subsequent Early Woodland components suggests group continuity within Region 1 between the Late Archaic and Early Woodland periods (Table 6). The majority of the multi-component sites with Terminal Archaic components also contained Late Archaic and/or Early Woodland components (*n* = 230), excluding Middle Woodland and Late Woodland villages that also had a Terminal Archaic component (*n* = 18). The fact that Terminal Archaic site components are strongly associated with preceding Late Archaic and subsequent Early Woodland components suggests group continuity within Region 1 between the Late Archaic and Early Woodland periods.

Table 6 - Region 1, Terminal Archaic Site by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Lithic reduction	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2
Open habitation, prehistoric	0	9	0	0	5	9	3	0	0	0	0	0	1	0	0	0	27
Open prehistoric site, unknown function	0	2	0	0	0	1	1	0	0	2	0	0	0	1	0	0	7
Unknown function open site greater than 20 m radius	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3
Unknown function surface scatter less than 20 m radius	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
(blank)	0	2	0	0	0	1	1	1	0	1	0	0	0	0	0	1	7
Part of multi-component site	0	95	4	0	28	79	10	11	8	6	5	6	8	12	1	3	276
Total	0	112	4	0	33	92	16	12	8	9	5	6	9	13	1	4	324

Terminal Archaic sites in Region 1 show a marked focus toward lowland topographic settings. Of the single component sites with landform data, 35 (74.5%) were found in lowland settings, and 12 (25.5%) were in upland settings. Considering multi-component sites with landform data, there are

206 (75.5%) multi-component Terminal Archaic sites in lowland settings compared to 67 (24.5%) multi-component sites in upland settings. Interestingly, the preference for lowland landforms rises significantly at sites with Terminal Archaic and Early Woodland components, with 82.1% of all such multi-component sites located in lowlands. Single component Terminal Archaic site types that may represent the likeliest candidates for seasonal occupation sites, such as base camps and short-term resource extraction camps, are Open habitation, prehistoric and Open prehistoric site, unknown function. Open habitation, prehistoric sites, which likely include a number of base camps, are almost exclusively found in lowland settings. The Open prehistoric site, unknown function site type, which may represent short-term resource extraction camps, are almost evenly divided between upland (n = 4) and lowland landforms (n = 3).

Early Woodland

The PASS database includes 665 sites with Early Woodland components in Region 1 (Table 7). There are 355 Early Woodland multi-component sites possessing Terminal Archaic and Middle Woodland components, representing 53.4% of the total population of Early Woodland sites. The fact that Early Woodland site components are strongly associated with preceding Terminal Archaic and subsequent Middle Woodland components suggests group continuity within Region 1 between the Terminal Archaic and Middle Woodland periods. Early Woodland site types include base camps and short-term resource extraction camps, similar to the preceding Late and Terminal Archaic periods, but also can include burial mounds and small hamlets, such as the Mayview Depot site (36AL124; as described by Robertson et al. 2008:123).

Early Woodland sites in Region 1 show a marked focus toward lowland topographic settings. Of the 137 single component sites with landform data, 82 (59.9%) were found in lowland settings, and 55 (40.1%) were in upland settings. There are 366 (71.5%) multi-component Early Woodland sites in lowland settings compared to 146 (28.5%) multi-component sites in upland settings. Single component Early Woodland site types that may represent the likeliest candidates for seasonal occupation sites, such as base camps and short-term resource extraction camps, are Open habitation, prehistoric and open prehistoric site, unknown function. Open habitation, prehistoric sites, which likely include a number of base camps, are primarily found in lowland settings (59.4%), with much greater relative numbers of this site type found on upland landforms than for either the Late Archaic or Terminal Archaic. The Open prehistoric site, unknown function site type, which may represent short-term resource extraction camps, shows nearly the same distribution as the Open habitation, prehistoric site type, with 58.8% of that site type in lowland settings.

Table 7 - Region 1, Early Woodland Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Burial mound	0	2	0	0	2	0	0	0	0	0	0	3	1	0	0	0	8
Earthwork	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Isolated find	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3
Open habitation, prehistoric	0	13	1	0	13	11	5	5	6	1	1	1	6	1	0	2	66
Open prehistoric site, unknown function	0	3	0	0	4	13	3	1	0	2	0	1	3	2	1	1	34
Other specialized aboriginal site	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0	4
Rock shelter/cave	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1	4
Unknown function open site greater than 20 m radius	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	4
Unknown function surface scatter less than 20 m radius	0	4	0	0	0	2	1	0	2	0	0	0	0	0	0	0	9
(blank)	0	1	0	0	1	3	1	0	0	0	0	1	1	0	0	0	8
Part of multi-component site	1	144	10	1	75	135	19	18	26	12	11	20	22	9	9	12	524
Total	1	172	12	1	97	165	29	25	35	16	13	26	33	14	10	16	665

Middle Woodland

The PASS database includes 776 sites with Middle Woodland components in Region 1 (Table 8). There are 466 Middle Woodland multi-component sites possessing Early and Late Woodland components, representing 60.0% of the total population of Middle Woodland sites. The fact that Middle Woodland site components are strongly associated with preceding Early Woodland and subsequent Late Woodland components suggests group continuity within Region 1 between the three Woodland periods. One single-component Middle Woodland mound is present in Region 1, the Meadows Mound (36WH0276); there are five other mounds that include a Middle Woodland component among their temporal associations. The lone Middle Woodland earthwork in Region 1 is the Stone House Mound (36BV0269).

Table 8 - Region 1, Middle Woodland Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Burial mound	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Earthwork	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Isolated find	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	3
Lithic Reduction	0	0	1	0	0	1	1	0	1	0	0	0	0	0	0	0	4
Open habitation, prehistoric	1	24	2	0	22	24	6	1	6	2	3	2	3	2	1	0	99
Open prehistoric site, unknown function	1	9	1	0	12	12	10	0	2	2	5	7	6	1	3	1	72
Other specialized aboriginal site	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Rock shelter/cave	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1	4
Unknown function open site greater than 20 m radius	0	0	0	0	1	2	1	1	0	0	0	1	0	1	0	0	7
Unknown function surface scatter less than 20 m radius	0	0	0	0	2	4	0	0	1	1	0	0	1	0	0	0	9
Village	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
(blank)	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	1	4
Part of multi-component site	0	150	8	0	79	154	27	18	21	15	18	24	24	14	9	8	569
Total	2	188	12	0	118	199	45	20	32	22	26	35	34	18	14	11	776

Middle Woodland sites in Region 1 show a marked focus toward lowland topographic settings. Of the 204 single component sites with landform data, 128 (62.7%) were found in lowland settings, and 76 (37.3%) were in upland settings. There are 391 (69.7%) multi-component Middle Woodland sites in lowland settings compared to 170 (30.3%) multi-component sites in upland settings. Year-round occupations are indicated at Middle Woodland sites identified as villages, including the single component Courson site (36CL0054), and five other multi-component village sites where the Middle Woodland component is the latest occupation, suggesting that the village designation likely applies to this time period. Ceremonial sites appear rare; earthworks and mounds account for only 0.9% of all single component Middle Woodland site types. Single component Middle Woodland site types that may represent the likeliest candidates for seasonal occupation sites, such as base camps and short-term resource extraction camps are Open habitation, prehistoric and Open prehistoric site,

unknown function. Open habitation, prehistoric sites are primarily found in lowland settings (73.7%). This site type likely includes a number of base camp sites. The Open prehistoric site, unknown function site type, which may represent short-term resource extraction camps, shows 49.3% of this site type occurring in lowland settings.

Late Woodland

The PASS data for Region 1 includes 1,050 sites with Late Woodland components in Region 1 (Table 9). Late Woodland sites with Middle Woodland components represent 31.6% of the total population of Late Woodland sites, which suggests some degree of group continuity within Region 1 between these two Woodland periods.

Late Woodland sites in Region 1 show a general focus toward lowland topographic settings, with some exceptions. Some Late Woodland site types identified in Region 1 include villages, short-term base camps, short-term resource extraction sites, and hamlets (Chiarulli 2001). Village sites are perhaps the defining site type for the Late Woodland period. Although the two landforms with the greatest number of Late Woodland villages in the PASS data as defined above are flood plains (n = 35) and terraces (n = 28), villages are also frequently found on hilltops (n = 26), ridgetops (n = 24), and saddles (n = 20), reflecting the introduction of the need for defense into residential site selection during this period. Single component Late Woodland site types that may represent the likeliest candidates for seasonal occupation sites, such as base camps and short-term resource extraction camps, are Open habitation, prehistoric and Open prehistoric site, unknown function. Finally, rock shelters/caves should be mentioned. There are far greater numbers of single component rock shelter/cave sites for the Late Woodland in Region 1 than for any other time period (n = 34). This may reflect a preference for rock shelters or caves as short-term resource extraction camps or base camps over open upland locations during the Late Woodland.

Table 9 - Region 1, Late Woodland Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Burial mound	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Earthwork	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Isolated find	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	3
Lithic reduction	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	2
Open habitation, prehistoric	1	47	4	2	16	39	6	9	10	4	3	16	6	2	4	1	170
Open prehistoric site, unknown function	0	11	1	0	6	9	4	0	9	1	1	2	2	0	0	0	46
Other specialized aboriginal site	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Quarry	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
Rock shelter/cave	0	1	0	0	3	2	0	9	1	6	4	0	0	0	6	2	34
Unknown function open site greater than 20 m radius	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	3
Unknown function surface scatter less than 20 m radius	0	4	1	0	0	0	1	0	0	0	0	0	0	0	0	0	6
Village, single component	0	24	0	0	1	24	13	2	24	0	1	21	18	2	0	0	130
Village, multi-component without Early or Middle Woodland components	0	11	0	0	2	4	0	0	2	1	0	3	2	0	0	0	25
(blank)	0	5	0	0	0	4	0	1	2	0	4	0	1	0	0	2	19
Part of multi-component site	0	178	2	2	81	161	32	20	23	15	18	12	29	16	6	12	607
Total	1	283	8	4	112	244	57	41	72	27	32	54	58	23	16	18	1050

REGION 2 SITES

Paleoindian

Within Region 2, there have been 48 sites identified with Paleoindian components, according to the PASS database (Table 10). Paleoindian sites in Region 2 are largely found in topographic settings that are close to water sources, with 41 sites (89.1%) found on flood plains, rises on flood plains, terraces, stream benches, and beaches. Only five Paleoindian sites were found in upland settings.

Table 10 - Region 2, Paleoindian Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Isolated find	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Isolated fluted point locus	0	3	0	0	1	3	0	0	0	0	0	0	0	0	0	0	7
Open habitation, prehistoric	0	0	0	0	0	3	0	0	0	0	0	1	0	0	1	0	5
Open prehistoric site, unknown function	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	4
Unknown function open site greater than 20 m radius	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	4
Part of multi-component site	0	13	0	0	0	11	0	0	0	0	0	0	1	0	1	1	27
Total	1	18	1	0	2	19	0	0	0	0	0	1	1	1	2	2	48

Of the single component Paleoindian sites, the site types that may represent resource extraction camps (Open habitation, prehistoric; Open prehistoric site, unknown function; and Unknown function open site, greater than 20 m radius) occur primarily in lowland settings. This pattern also appears even stronger with multi-component sites with Paleoindian material; all but two of the multi-component sites are found in lowland settings. The lowland setting of these camps, in proximity to water, fit what has been described as the typical Paleoindian site for western Pennsylvania (Meyers and Moses Meyers 2014).

Early Archaic

The PASS database records 66 sites with Early Archaic components in Region 2 (Table 11). Early Archaic sites in Region 2 are largely found in lowland topographic settings, with 49 sites (74.2%) found in lowland settings.

Table 11 - Region 2, Early Archaic Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Isolated find	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	2
Lithic reduction	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	2
Open habitation, prehistoric	0	0	0	0	2	1	1	0	0	0	0	0	0	0	0	0	4
Open prehistoric site, unknown function	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	2
Unknown function open site greater than 20 m radius	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
(blank)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Part of multi-component site	0	19	1	0	4	20	1	0	2	0	0	1	1	1	2	2	54
Total	0	19	1	0	6	23	3	1	3	1	1	1	1	1	2	3	66

The single component Early Archaic sites in the PASS data that probably represent some form of resource extraction camp include Open habitation, prehistoric; Open prehistoric site, unknown function; and Unknown function, open site greater than 20 m radius. These three site types appear fairly evenly distributed between lowland and upland settings according to the PASS data. Multi-component sites with Early Archaic material in Region 2, however, show a distribution markedly focused on settings associated with close distances to water, with 44 such sites in lowland settings, compared to 8 sites with Early Archaic components in upland settings.

Middle Archaic

The PASS database includes 106 sites with Middle Archaic components in Region 2 (Table 12). Middle Archaic sites in Region 2 show a general preference for lowland topographic settings. Nine sites with Middle Archaic components did not have a landform recorded in the PASS data. Of the 28 single component sites with landform data, 23 (82.1%) were found in lowland settings and 5 (17.9%) were in upland settings. Similarly, when multi-component sites with Middle Archaic components are

considered, there appears to be a stronger preference for lowland sites, with 55 (79.7%) multi-component Middle Archaic sites in lowland settings compared to 14 (20.3%) multi-component sites in upland settings.

Table 12 - Region 2, Middle Archaic Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Isolated find	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Open habitation, prehistoric	0	1	2	0	0	10	0	0	0	0	1	0	1	0	0	2	17
Rock shelter/cave	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
Unknown function open site greater than 20 m radius	0	2	0	0	0	2	1	1	0	0	0	0	0	0	0	0	6
Unknown function surface scatter less than 20 m radius	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	3
(blank)	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Part of multi-component site	0	23	2	0	7	23	2	2	4	0	0	0	3	1	2	7	76
Total	0	29	4	0	8	37	3	3	4	0	1	1	4	1	2	9	106

Raber et al. (1998) noted that Middle Archaic resource exploitation camps were to be found in upland settings, while base camps were located on post-Pleistocene terraces. However, the sites types most likely to represent such camps in Region 2 (Open habitation, prehistoric; Rock shelter/cave; and Unknown function open site, greater than 20 m radius) follow the overall trend for Region 2 Middle Archaic sites and are mostly found in lowland settings.

Late Archaic

The PASS database includes 211 sites with Late Archaic components in Region 2 (Table 13). Late Archaic sites in Region 2 show a preference toward lowland topographic settings. Of the 56 single component sites with landform data, 37 (66.1%) were found in lowland settings and 19 (33.9%) were in upland settings. The preference for lowland settings is even stronger when multi-component sites with Late Archaic components designated by landform are considered, with 104 (80.0%) multi-component Late Archaic sites in lowland settings compared to 26 (20.0%) multi-component sites in upland settings.

Table 13 - Region 2, Late Archaic Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Isolated find	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	3
Lithic reduction	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	3
Open habitation, prehistoric	1	0	0	0	2	16	1	1	0	1	0	0	0	0	0	4	26
Open prehistoric site, unknown function	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Rock shelter/cave	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	3
Unknown function open site greater than 20 m radius	0	3	0	0	1	5	0	2	2	0	0	0	0	0	0	0	13
Unknown function surface scatter less than 20 m radius	0	1	1	0	1	3	3	0	0	0	0	0	1	0	0	4	14
(blank)	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	3
Part of multi-component site	0	42	2	0	9	51	5	2	4	3	1	3	1	1	6	15	145
Total	1	46	3	0	15	76	10	9	8	4	1	3	3	1	6	25	211

Single component Late Archaic site types that may represent the likeliest candidates for seasonal occupation sites, such as base camps and short-term resource extraction camps, are Open habitation, prehistoric; Open prehistoric site, unknown function; Rock shelter/cave; and Unknown function open site, greater than 20 m radius. Late Archaic Open habitation, prehistoric sites are typically found in lowland settings (n = 19), with upland landforms possessing only three examples of this site type, which likely includes a number of base camp sites. The Open prehistoric site, unknown function site type has only one site classified as such in the PASS data for Region 2, and it is located in a lowland setting. All three Late Archaic Rock shelter/cave sites are found in uplands. Unknown function open site, greater than 20 m radius occur in both upland and lowland settings, but mainly on lowland landforms (69.2% of all such Late Archaic sites).

Terminal Archaic

The PASS database includes 101 sites with Terminal Archaic components in Region 2 (Table 14). The majority of the multi-component sites with Terminal Archaic components also contain Late Archaic and/or Early Woodland components (*n* = 78). The fact that Terminal Archaic site components are strongly associated with preceding Late Archaic and subsequent Early Woodland components suggests group continuity within Region 2 between the Late Archaic and Early Woodland periods. There are two burial mounds associated with Terminal Archaic occupations in Region 2, include the Z1 site (36WA0139) and the Biebel site (36ER0231). Both were radiocarbon-dated to the Terminal Archaic according to the PASS data, and represent early mound building activity in northwestern Pennsylvania.

Table 14 - Region 2, Terminal Archaic Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Burial mound	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	2
Open habitation, prehistoric	0	2	0	0	2	4	0	0	0	0	0	0	1	0	0	0	9
Other specialized aboriginal site	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Unknown function surface scatter less than 20 m radius	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2
(blank)	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Part of multi-component site	1	25	3	0	5	24	5	4	3	2	2	1	0	2	5	4	86
Total	1	27	3	0	8	32	5	4	3	2	2	1	1	3	5	4	101

Terminal Archaic sites in Region 2 show a marked focus toward lowland topographic settings. Of the 15 single component sites, nearly all were found in lowland settings (n = 13, 86.7%). There are 58 (70.7%) multi-component Terminal Archaic sites in lowland settings compared to 24 (29.3%) multi-component sites in upland settings. Interestingly, sites that have Terminal Archaic and Early Woodland components appear to have a preference for upland locations, as 56.5% of Terminal Archaic sites with Early Woodland material present are found in uplands in Region 2.

Only one single component Terminal Archaic site type may represent a likely candidate for representing seasonal occupation sites, such as base camps and short-term resource extraction camps: Open habitation, prehistoric. Open habitation, prehistoric sites are almost exclusively found in

lowland settings (n = 8), with only one example of this site type found on upland landforms. This site type likely includes a number of base camp sites.

Early Woodland

The PASS database includes 181 sites with Early Woodland components in Region 2 (Table 15). There are 113 Early Woodland multi-component sites possessing Terminal Archaic and/or Middle Woodland components, representing 62.4% of the total population of Early Woodland sites. The fact that Early Woodland site components are strongly associated with preceding Terminal Archaic and subsequent Middle Woodland components suggests group continuity within Region 2 between the Terminal Archaic and Middle Woodland periods, especially between the Early and Middle Woodland periods Dragoo (1989[1963]:139) notes that burial mounds decreased significantly in frequency with distance from the Ohio River Valley. Accordingly, there are only four ceremonial single-component sites within Region 2. The three Early Woodland burial mounds are all found in Warren County (36WA0117, 36WA0140, and 36WA0206), while the single earthwork is in Erie County (36ER0242, the Albion Park site). Early Woodland sites in Region 2 show a marked focus toward lowland topographic settings. Of the 30 single component sites with landform data, 22 (73.3%) were found in lowland settings, and 8 (26.7%) were in upland settings. There are 93 (68.4%) multi-component Early Woodland sites in lowland settings compared to 43 (31.6%) multi-component sites in upland settings.

Single component Early Woodland site types that may represent the likeliest candidates for seasonal occupation sites, such as base camps and short-term resource extraction camps, are Open habitation, prehistoric; Open prehistoric site, unknown function, and Unknown function open site, greater than 20 m radius. Open habitation, prehistoric sites are mainly found in lowland settings (66.7%). This site type likely includes a number of base camp sites. The Open prehistoric site, unknown function, and the Unknown function open site, greater than 20 m radius site types, which may represent short-term resource extraction camps, occur in far fewer numbers than the Open habitation, prehistoric site type, but the preference holds for lowland settings.

Table 15 - Region 2, Early Woodland Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Burial mound	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	3
Earthwork	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Isolated find	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Open habitation, prehistoric	0	0	0	0	1	9	1	0	0	1	0	0	2	0	1	1	16
Open prehistoric site, unknown function	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Rock shelter/cave	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	3
Unknown function open site greater than 20 m radius	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3
Village	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
(blank)	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
Part of multi-component site	1	37	4	0	7	44	5	5	3	4	4	3	2	3	14	14	150
Total	1	42	4	0	9	59	6	5	4	5	4	4	4	3	16	15	181

Middle Woodland

The PASS database includes 209 sites with Middle Woodland components in Region 2 (Table 16). There are 128 Middle Woodland multi-component sites possessing Early and/or Late Woodland components, representing 61.2% of the total population of Middle Woodland sites. The fact that Middle Woodland site components are strongly associated with preceding Early Woodland and subsequent Late Woodland components suggests group continuity within Region 2 between the three Woodland periods. Middle Woodland sites in Region 2 show a marked focus toward lowland topographic settings. Of the 57 single component sites with landform data, 45 (78.9%) were found in lowland settings, and 12 (21.1%) were in upland settings. There are 101 (77.1%) multi-component Middle Woodland sites in lowland settings compared to 30 (22.9%) multi-component sites in upland settings. Interestingly, the preference for lowland landforms drops somewhat at sites with Early Woodland and/or Late Woodland components, with 67.2% of all such multi-component sites located in lowlands.

Table 16 - Region 2, Middle Woodland Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Burial mound	0	8	0	0	2	2	0	0	0	0	0	0	0	0	1	1	14
Earthwork	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Isolated find	0	0	0	0	2	3	0	0	1	0	0	0	0	0	0	0	6
Open habitation, prehistoric	1	4	1	0	2	10	1	2	2	0	0	0	1	0	0	2	26
Open prehistoric site, unknown function	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Rock shelter/cave	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	3
Unknown function open site greater than 20 m radius	0	0	1	0	1	3	0	1	0	0	0	0	0	0	0	0	6
Unknown function surface scatter less than 20 m radius	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Village	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
(blank)	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Part of multi-component site	1	48	0	0	5	47	4	6	1	3	4	0	2	1	9	17	148
Total	2	64	2	0	12	66	5	9	4	3	5	0	3	1	12	21	209

Year-round occupations are indicated at Middle Woodland sites identified as villages, including the two single component Middle Woodland village sites: the Stewart Farm site (36ME0051) and the Ziegler Site (36WA0080); there are also two multi-component village sites where the Middle Woodland component is the latest occupation, suggesting that the village designation likely applies to this time period (36WA0087, the John Harrington No. II Site, and 36ME0016, the Hitchcock Site). Single component sites representing ceremonial activities, including earthworks and mounds, account for 7.2% of all Middle Woodland sites. There are 15 ceremonial sites in Region 2, an increase over the preceding Early Woodland. Interestingly, single component Middle Woodland mounds are the norm, with only two multi-component mound sites in Region 2. The lone Middle Woodland earthwork in Region 2 is the Lindstrom I site (36WA0126). The increase in mound building activity may reflect the influence of Hopewell cultures in the Ohio River Valley to the west. Single component Middle Woodland site types that may represent the likeliest candidates for seasonal occupation sites, such as base camps and short-term resource extraction camps, are Open habitation, prehistoric; Open prehistoric site, unknown function; and Unknown function open site,

greater than 20 m radius. Open habitation, prehistoric sites are mainly found in lowland settings (75.0%). This site type likely includes a number of base camp sites. The Open prehistoric site, unknown function site type, which may represent short-term resource extraction camps, only occurs as a single site, and its landform was not recorded. Unknown function open site, greater than 20 m radius sites may also represent short-term resource extraction camps, and they tend to occur almost entirely in lowland settings (83.3%).

Late Woodland

The PASS data for Region 2 includes 265 sites with Late Woodland components in Region 2 (Table 17). There are 94 Late Woodland multi-component sites also possessing Middle Woodland components, representing 35.4% of the total population of Late Woodland sites. The fact that Late Woodland site components are well associated with preceding Middle Woodland components suggests some degree of group continuity within Region 2 between these two Woodland periods.

Late Woodland sites in Region 2 show a strong focus toward lowland topographic settings, with some exceptions. Of the 113 single component sites with landform data, 66 (58.4%) were found in lowland settings and 47 (41.6%) were in upland settings. There are 93 (72.1%) multi-component Late Woodland sites in lowland settings compared to 36 (27.9%) multi-component sites in upland settings. Sites with Late Woodland components defined by landform are most commonly found on flood plains (n = 82, 33.9%) and terraces (n = 60; 24.8%); all other landforms each have 10% or less of the total population of identified Late Woodland site types. Village sites are perhaps the defining site type for the Late Woodland. The landform with the greatest number of Late Woodland villages in the PASS data as defined above is flood plain (n = 16), representing nearly half of all villages in Region 2. Only eight village sites are found in upland settings. Single component Late Woodland site types that may represent the likeliest candidates for seasonal occupation sites, such as base camps and short-term resource extraction camps, are Open habitation, prehistoric; Open prehistoric site, unknown function; Rock shelter/cave; and Unknown function open site, greater than 20 m radius. Open habitation, prehistoric sites are mainly found in lowland settings (79.2%). This site type likely includes a number of base camp sites. The Open prehistoric site, unknown function site type, which may represent short-term resource extraction camps, shows both examples of this site type with landform data occurring in a lowland setting. Unknown function open site, greater than 20 m radius sites may represent either a base camp or a short-term resource extraction camp, and are also found largely in the lowlands (78.6%). Finally, rock shelters/caves should be mentioned. There are far greater numbers of single component rock shelter/cave sites for the Late Woodland in Region 2 than for any other time period (n = 30), 90% of which are in upland settings. This may reflect a preference for rock shelters or caves as short-term resource extraction camps or base camps over open upland locations during the Late Woodland period.

Table 17 - Region 2, Late Woodland Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Earthwork	0	0	0	0	0	4	0	0	0	0	0	0	1	0	0	3	8
Lithic reduction	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
Open habitation, prehistoric	0	9	0	0	0	10	0	1	0	2	1	0	0	1	0	4	28
Open prehistoric site, unknown function	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3
Other specialized aboriginal site	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	2
Quarry	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Rock shelter/cave	0	0	0	0	1	1	2	2	1	1	9	3	0	0	9	1	30
Unknown function open site greater than 20 m radius	0	8	1	0	0	2	0	1	2	0	0	0	0	0	0	0	14
Unknown function surface scatter less than 20 m radius	0	3	2	0	1	0	1	0	0	0	0	0	0	0	0	0	7
Village	0	8	3	0	0	5	1	0	2	0	0	4	0	0	0	0	23
(blank)	0	2	0	0	0	2	0	0	0	0	0	0	0	2	0	0	6
Part of multi-component site	0	49	3	0	6	35	5	7	1	2	3	4	3	1	10	12	141
Total	0	82	9	0	8	60	9	12	6	5	13	11	4	4	19	23	265

REGION 3 SITES

Paleoindian

Within Region 3, there are no sites identified with Paleoindian components, according to the PASS database. Although the Laurentide ice sheet would have withdrawn from Region 3 around 11,800 B.C., this region may not have offered sufficient resources to attract the degree of activity from prehistoric groups that would have resulted in a visible archaeological record until the Archaic period. Paleoindian sites may very well be present, but simply have not been detected yet.

Early Archaic

The PASS database records only five sites with Early Archaic components in Region 3, all of which are multi-component sites with one or more time periods present at the site along with an Early Archaic component (Table 18). A sample size of five sites is too small to make any generalizations about Early Archaic site preference, especially as there are no single component Early Archaic sites in Region 3. Four of the five Early Archaic sites in Region 3 had landform data recorded in the PASS database; one site was located on a terrace and three were identified on an upland flat landform.

Table 18 - Region 3, Early Archaic Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Part of multi-component site	0	0	0	0	0	1	0	0	0	0	0	0	0	3	0	1	5
Total	0	0	0	0	0	1	0	0	0	0	0	0	0	3	0	1	5

Middle Archaic

The PASS database includes six sites with Middle Archaic components in Region 3, all but one of which are multi-component sites with one or more time periods also represented in the site assemblages (Table 19). The single component site is the Wiser C site (35ER0111), an Open habitation, prehistoric site located on a terrace. Three of the multi-component Middle Archaic sites are located on terraces as well, with the remaining two sites identified on the upland flat landform type. As with the Early Archaic period for Region 3, the small sample size is insufficient to draw any general observations about landform preference during the Middle Archaic period.

Table 19 - Region 3, Middle Archaic Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Open habitation, prehistoric	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Part of multi-component site	0	0	0	0	0	3	0	0	0	0	0	0	0	2	0	0	5
Total	0	0	0	0	0	4	0	0	0	0	0	0	0	2	0	0	6

Late Archaic

The PASS database includes 18 sites with Late Archaic components in Region 3, possibly demonstrating a population increase over the preceding Early and Middle Archaic periods (Table 20). Although the population of Late Archaic sites in Region 3 is probably too small to identify any preferences for a topographic setting, some observations about landform distribution can still be made. Of the seven single component sites with landform data, three sites were found in lowland settings, and four sites were in upland settings. Multi-component Late Archaic sites are evenly divided between lowlands and uplands in Region 3. The small population of single component Late Archaic sites hinders the identification of candidates for base camps, short-term resource extraction camps, and other functional site types, but it appears that the sites present in the PASS database are good candidates for a mix of both types.

Table 20 - Region 3, Late Archaic Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Open habitation, prehistoric	0	0	1	0	0	1	0	0	0	0	0	0	0	3	0	1	6
Unknown function surface scatter less than 20 m radius	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
(blank)	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	2
Part of multi-component site	0	0	0	0	0	4	0	0	0	0	0	0	0	4	0	1	9
Total	0	0	1	0	0	6	0	0	0	0	0	0	0	8	0	3	18

Terminal Archaic

The PASS database includes 17 sites with Terminal Archaic components in Region 3 (Table 21). The majority of the multi-component sites with Terminal Archaic components also contained Late Archaic and/or Early Woodland components (n = 10). The fact that Terminal Archaic site components are strongly associated with preceding Late Archaic and subsequent Early Woodland components suggests group continuity within Region 3 between the Late Archaic and Early Woodland periods.

Terminal Archaic sites with landform data in Region 3 occur in both lowland and upland topographic settings. The small population of Terminal Archaic sites in Region 3 makes extrapolation of settlement patterns unfeasible. The single component Terminal Archaic site type Open habitation,

prehistoric may represent a seasonal occupation site, such as a base camp or short-term resource extraction camp. There are only three such sites in Region 3: one located on a beach and one on a stream bench, with a third lacking landform data. Although a sample size of two sites is too small to draw generalizations from which to discuss settlement patterns, it seems probable that the two sites represent base camps, as they are located near water sources.

Table 21 - Region 3, Terminal Archaic Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Open habitation, prehistoric	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	3
(blank)	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	2
Part of multi-component site	2	1	0	0	0	3	0	0	0	0	0	0	0	5	0	1	12
Total	3	1	0	0	2	4	0	0	0	0	0	0	0	5	0	2	17

Early Woodland

The PASS database includes 27 sites with Early Woodland components in Region 3 (Table 22). There are 11 Early Woodland multi-component sites possessing Terminal Archaic and/or Middle Woodland components, representing 40.7% of the total population of Early Woodland sites. The fact that Early Woodland site components are associated with preceding Terminal Archaic and subsequent Middle Woodland components suggests a degree of group continuity within Region 3 between the Terminal Archaic and Middle Woodland periods.

There are too few Early Woodland sites in Region 3 to make predictions about topographic setting preferences, although some general observations can be made that may serve as a basis for generating future research questions. Both single component and multi-component sites are evenly split between upland and lowland settings, with slightly more single component sites (n = 7) than multi-component sites (n = 6) in lowland settings. Interestingly, multi-component sites that possess Terminal Archaic and/or Middle Woodland components along with Early Woodland components in Region 3 show a preference for upland settings, with six such sites in uplands and only three in lowlands. Single component Early Woodland site types that may represent the likeliest candidates for seasonal occupation sites, such as base camps and short-term resource extraction camps, are Open habitation, prehistoric and Open prehistoric site, unknown function. Following the general trend for Early Woodland sites, Open habitation, prehistoric sites are fairly evenly split between lowland settings (n = 5) and upland settings (n = 4). This site type likely includes a number of base camp

sites. The Open prehistoric site, unknown function site type, which may represent short-term resource extraction camps, only occurs once, in an upland setting.

Table 22 - Region 3, Early Woodland Sites by Landform

Site Type	Beach	Flood Plain	Island	Saddle	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Rise in Flood Plain	Upland Flat	Upper Slope	(Blank)	Total
Burial mound	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Open habitation, prehistoric	0	0	0	0	1	4	0	0	0	0	0	0	0	4	0	0	9
Open prehistoric site, unknown function	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Unknown function surface scatter less than 20 m radius	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
Village	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Part of multi-component site	1	1	0	0	1	3	0	0	0	0	0	0	0	6	0	2	14
Total	1	1	0	0	2	9	0	0	0	0	0	0	0	11	1	2	27

Middle Woodland

The PASS database includes only 12 sites with Middle Woodland components in Region 3 (Table 23). All of the Middle Woodland multi-component sites possess Early and/or Late Woodland components. The fact that Middle Woodland site components are associated with preceding Early Woodland and subsequent Late Woodland components suggests group continuity within Region 3 between the three Woodland periods, although the degree of continuity is uncertain. In fact, only one site has all three Woodland periods represented in its artifact assemblage. The dip in the number of sites from the Early Woodland and a substantial increase in the succeeding Late Woodland period may indicate a shift in group locations, with Late Woodland groups coming in to occupy a territory that had seen a decrease in utilization during the Middle Woodland. Alternately, the relative fewer numbers of Middle Woodland sites may represent groups from across the region coalescing at a small handful of sites, and then subsequently expanding during the Late Woodland due to population increases.

Table 23 - Region 3, Middle Woodland Sites by Lanform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Village	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
(blank)	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
Part of multi-component site	2	2	0	0	1	2	0	0	0	0	0	0	0	1	0	2	10
Total	2	2	1	0	1	2	0	0	0	0	0	1	0	1	0	2	12

Middle Woodland sites in Region 3 show a general trend toward lowland topographic settings, although with such a small sample, any such statements about landform preferences must be seen as very tenuous in nature. The two single component Middle Woodland sites include a village site (the Billings Site; 36ER0055) and a site without a site type recorded in the PASS database (the North East Access Site #2; 36ER0192). The village site is located on a rise in the floodplain, while the unknown site type is located on a ridgetop. Interestingly, there appears to be a focus toward uplands at sites with Early and Middle Woodland components, but lacking a Late Woodland occupation; sites with Early and Middle Woodland components only are nearly all found in uplands (n = 3) versus the lowlands (n = 1). Conversely, sites with only Middle and Late Woodland components are restricted to lowlands, with no such sites identified in an upland setting.

Year-round occupations are indicated at Middle Woodland sites identified as villages, including the one single component Middle Woodland village site identified above. Note that village sites may actually represent a hamlet rather than a village, as there is currently no way to distinguish a hamlet from a village using PASS data. In addition, there is one multi-component village site where the Middle Woodland component is the latest occupation, suggesting that the village designation likely applies to this time period (the McCord Site; 36ER0167). Middle Woodland seasonal occupation sites, such as base camps and short-term resource extraction camps, may be present, but cannot be teased out from the multi-component sites in the PASS database.

Late Woodland

The PASS data for Region 3 includes 29 sites with Late Woodland components (Table 24). There are five Late Woodland multi-component sites also possessing Middle Woodland components, representing 17.2% of the total population of Late Woodland sites. The fact that Late Woodland site components are not as well associated with preceding Middle Woodland components suggests a regional change in population, possibly due to an expansion of Late Woodland site types across Region 3 over that of the preceding Middle Woodland period.

Table 24 - Region 3, Late Woodland Sites by Landform

Site Type	Beach	Flood Plain	Rise in Flood Plain	Island	Stream Bench	Terrace	Hill Ridge /Toe	Hillslope	Hilltop	Lower Slope	Middle Slope	Ridgetop	Saddle	Upland Flat	Upper Slope	(Blank)	Total
Burial mound	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Earthwork	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Open habitation, prehistoric	1	1	0	0	1	3	1	0	0	0	0	0	0	4	0	0	11
Village	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	3
(blank)	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Part of multi-component site	1	1	0	0	2	2	0	0	0	0	0	0	0	4	0	2	12
Total	2	2	0	0	4	9	1	1	0	0	0	0	0	8	0	2	29

Late Woodland sites in Region 3 show a strong focus toward lowland topographic settings. A total of 11 (64.7%) single component sites with landform designations were found in lowland settings, and 6 (35.3%) sites were in upland settings. The same general preference for lowland sites holds with multi-component sites that included landform data: 6 (60.0%) are found in lowland settings compared to 4 (40.0%) in uplands. Village sites are perhaps the defining site type for the Late Woodland. Single component Late Woodland villages in Region 3 do not appear to focus on a particular landform; the three such sites are located on a hillslope, a stream bench, and a terrace. All three multi-component Late Woodland villages are found on upland flats. Single component Late Woodland site types that may represent the likeliest candidates for seasonal occupation sites, such as base camps and short-term resource extraction camps, are the Open habitation, prehistoric type, which are distributed fairly evenly between upland and lowland settings.

3
DATA QUALITY – REGIONS 1, 2, AND 3

INTRODUCTION

PASS forms have been used by submitters to record archaeological site data for more than 65 years. When PASS forms are accurately filled out, they offer the PHMC vital information regarding location and artifact data. Over the past few decades PHMC has been working diligently to get the PASS form data into its CRGIS database, a map-based inventory of the historic and archaeological sites and surveys currently stored in the files of the Bureau for Historic Preservation (BHP). The CRGIS database is designed to include all information on the PASS forms, with the goal of obtaining as much accurate information as possible about Pennsylvania's archaeological and historic sites. Using roughly 23,000 completed PASS forms, PHMC has managed to accurately enter almost all known archaeological sites into the CRGIS database. The CRGIS database has become PHMC's primary tool when attempting to accurately record and map Pennsylvania's historic and prehistoric past.

In order to establish the validity of the data used for the predictive model set project, the CRGIS database and PASS form data were compared for a sample of Pennsylvania's 18,232 prehistoric archaeological sites. Archaeological site forms were analyzed and compared with the data included in the CRGIS data base. Site forms from all of Pennsylvania's 67 counties were considered and a 10% random sample was selected from each county. The following conclusions and data are the results of the 10% sample for the counties within Regions 1, 2, and 3.

METHODS

Within Regions 1, 2, and 3, PASS forms and CRGIS data were examined for 755 prehistoric archaeological sites. The following section presents the results of the analysis by region. Location accuracy, artifact data quality, and form completeness were rated for each of the selected sites using information from the PASS forms and CRGIS database. Ratings were assigned numerical values to facilitate comparison between the two data sources and across regions. Table 25 lists the criteria used to derive ratings for each category of data.

Location data were analyzed by manually comparing mapped locations within the CRGIS with maps provided in the original PASS forms. Artifact information was also manually compared between the PASS forms and the CRGIS data base. Discrepancies between the two data sets were categorized using the ranking outlined in Table 25.

Table 25 - Rating Criteria for Site Data

Rating	Criterion
	Location Accuracy, PASS Form
1	*No location information*. No location data are present on the site form.
2	*Coordinates only*. Location is documented only by coordinates with no physical description or landmarks.
3	*Poor accuracy*. The only location information is a hand-drawn map with low detail.
4	*Medium accuracy*. The form contains a USGS map with the site location indicated.
5	*High accuracy* The form contains a detailed map with reference points or an aerial photo and the site location is assumed to be accurate.
	How Well Location is Reflected in CRGIS
1	*Not mapped.* The site has not been mapped into the CRGIS system.
2	*Mapped, > 500 m.*The site location is mapped, but is more than 500 m away from the location indicated on the PASS form. Note that in some cases this reflects corrections to the location data in CRGIS, resulting in *increased* accuracy.
3	*Mapped, 250–500 m*. The site location is mapped, but is between 250 and 500 m away from the location indicated on the PASS form (see note above re: accuracy).
4	*Mapped, < 250m*. The site location is mapped less than 250 m away from the PASS form location.
5	*Mapped accurately*. The site location in CRGIS matches the location on the PASS form.
	Artifact Data Quality, PASS Form
1	*No artifacts.* The PASS form contains no artifact information, either because no artifacts were found or because they were not recorded.
2	*Artifacts poorly represented.* No artifacts are listed on the PASS form, but a note indicating that artifacts were found is included indicating that artifacts were found but not recorded.
3	*Poor quality recording.* The PASS form contains poorly hand-drawn artifacts and/or mislabeled items.
4	*Moderate recording.* Few artifacts are listed on the PASS form or only a small selection were drawn; the location of the collection is not indicated.
5	*Good recording.* All artifacts are listed on the form, which also includes high-quality hand-drawn images or photographs; the location of the collection is usually indicated.
	How Well Artifacts are Reflected in CRGIS
1	*No artifacts.* The CRGIS data base does not include any artifacts.
2	*Less artifacts.* Fewer artifacts than appear on the PASS form are included in the CRGIS data base.
3	*Moderate quality*. Artifacts are listed in the CRGIS data base, but not with any detail.
4	*Higher quality*. The CRGIS data base contains more artifacts than are listed on the PASS form.
5	*Accurate recording.* Artifacts listed in the CRGIS data base match those listed on the PASS form.
	PASS Form Completeness
1	*Name and/or location.* Only site name and/or location are included on the PASS form.
2	*< 25% completed.* The PASS form contains more than just name and location, but is missing at least 25% of data.
3	*25–75% completed.* The PASS form is mostly filled out and contains artifact and location data.
4	*> 75% completed.* The PASS form is filled out completely and contains all required information.
	PASS Form Type
1	*1950–1980 version.* This form has limited room for data; usually only location information and material culture information was collected.
2	*1981–2007 version.* This form has more space for documentation and includes a requirement for sketched images of artifacts.
3	*2008–present version.* This form is several pages in length; it requires artifacts to be categorized and location information to be detailed on attached maps.

REGION 1

PASS forms and CRGIS data were examined for a total of 573 sites within Region 1.

Location Accuracy

A total of 65% of the PASS forms in the site sample for Region 1 are referenced on highly detailed maps or on USGS maps, while 32% contain only coordinates or are mapped poorly. The remaining 3% of PASS forms contains no location data at all (Figure 3). By comparison, 92% of the same site sample have accurately mapped locations in the CRGIS data base, while 8% of the sites are unmapped or are mapped 250 m or more from the location indicated on the PASS forms (Figure 4). The latter is not necessarily an indication of incorrect mapping; in fact, it may reflect locations that were corrected in the migration from PASS form to CRGIS.

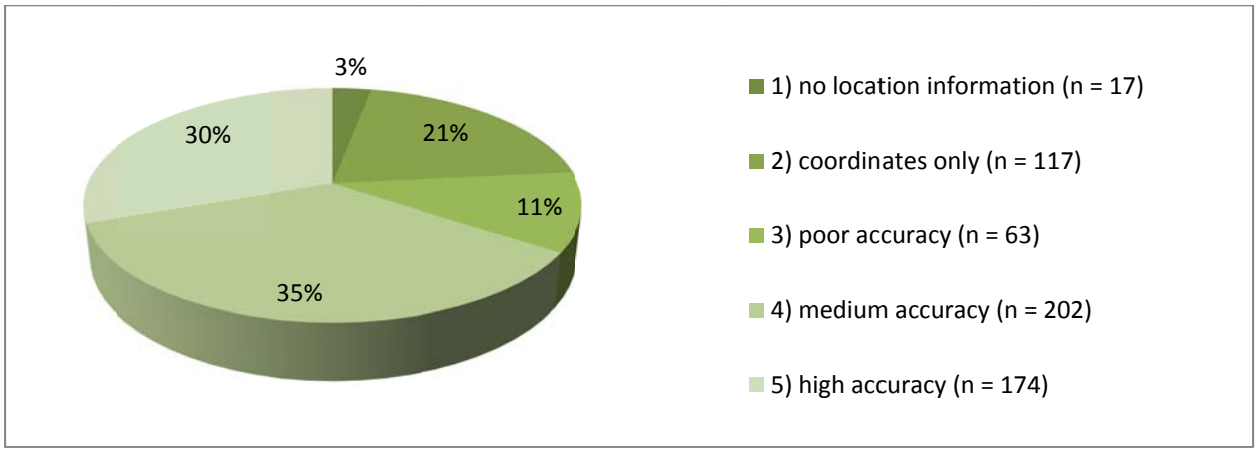


Figure 3 - Quality of location information on PASS forms within Region 1.

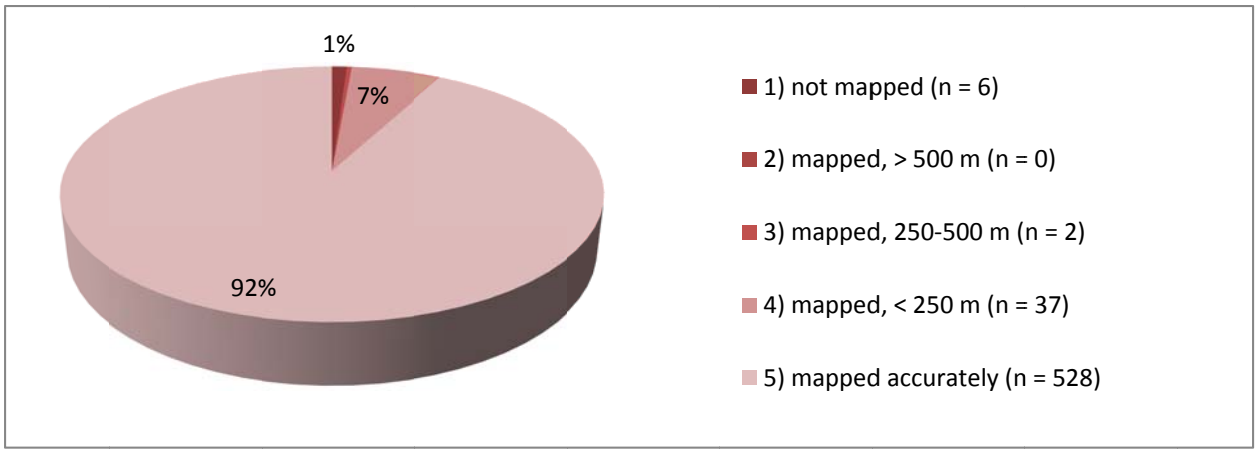


Figure 4 - Quality of location information reflected in CRGIS within Region 1.

Artifact Data

Almost two-thirds (65%) of the archaeological site sample of PASS forms within Region 1 contain artifact descriptions that can be described as good or moderate. The remaining 35% of sites contain poor descriptions or none at all (Figure 5). Just over half (51%) of the sites in the CRGIS database contain artifact information that accurately matches the information found in their PASS forms, while 15% contain more artifact data than the original PASS form, and another 7% have good quality artifact data (Figure 6). The percentage of sites with no artifact data in the CRGIS data base nearly matches that of the original PASS forms (23% and 24%, respectively), while only 2% of sites in the CRGIS data base contained less information than was found on the original PASS forms.

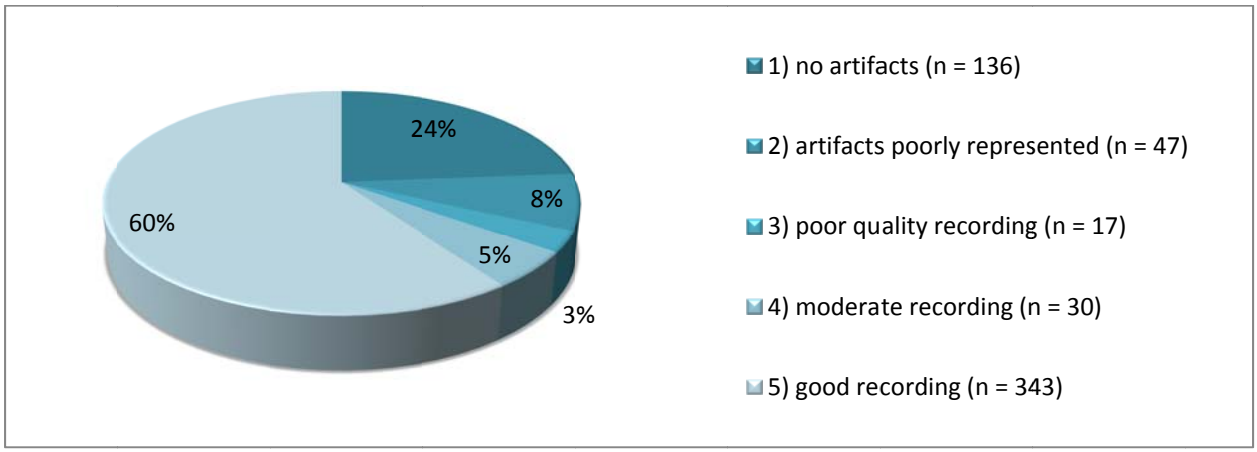


Figure 5 - Original artifact data recorded on PASS forms for Region 1.

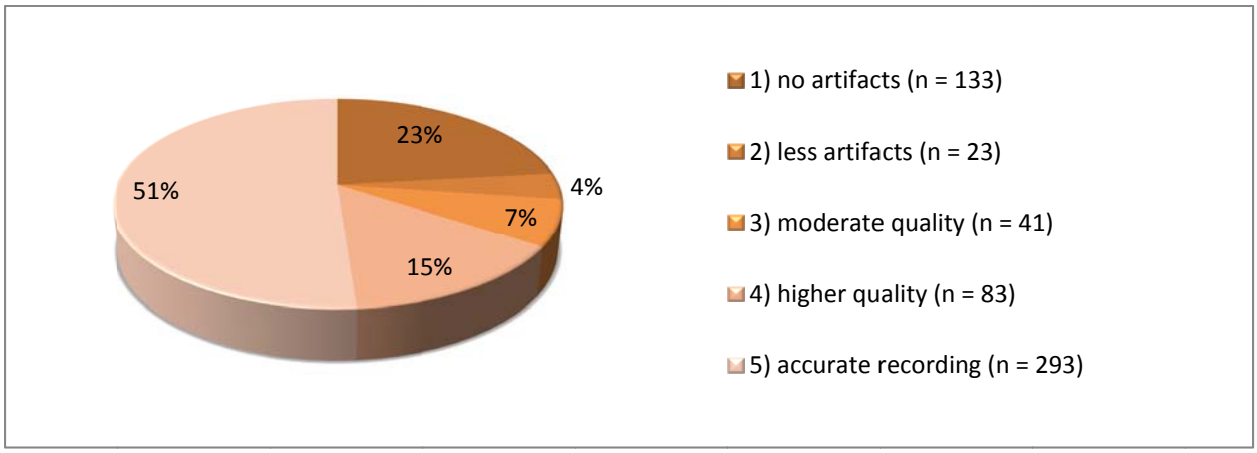


Figure 6 - Artifact data reflected in the CRGIS data base for Region 1.

PASS Form Types and Completeness

A total of 84% of PASS forms in the site sample from Region 1 are up to or greater than 75% complete (Figure 7). The remaining 16% of PASS forms in the site sample contain limited data. Almost all (95%) of the site sample for Region 1 is recorded on old version or middle version PASS forms, while only 5% are recorded on the newer version of the form that includes detailed artifact information (Figure 8). This suggests that for Region 1, the most reliable site information is likely to be locational rather than artifact data.

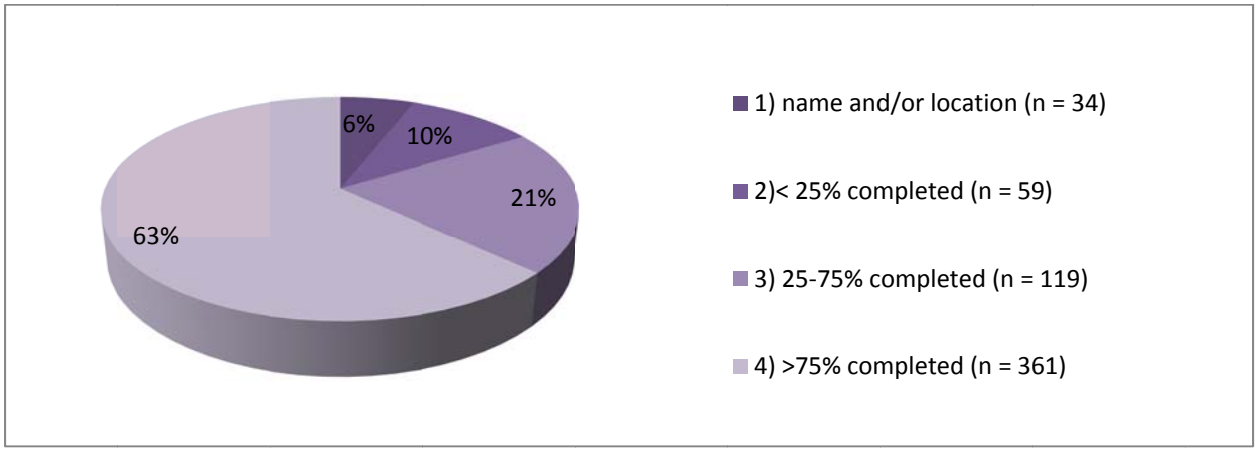


Figure 7 - Completeness of PASS form information in Region 1.

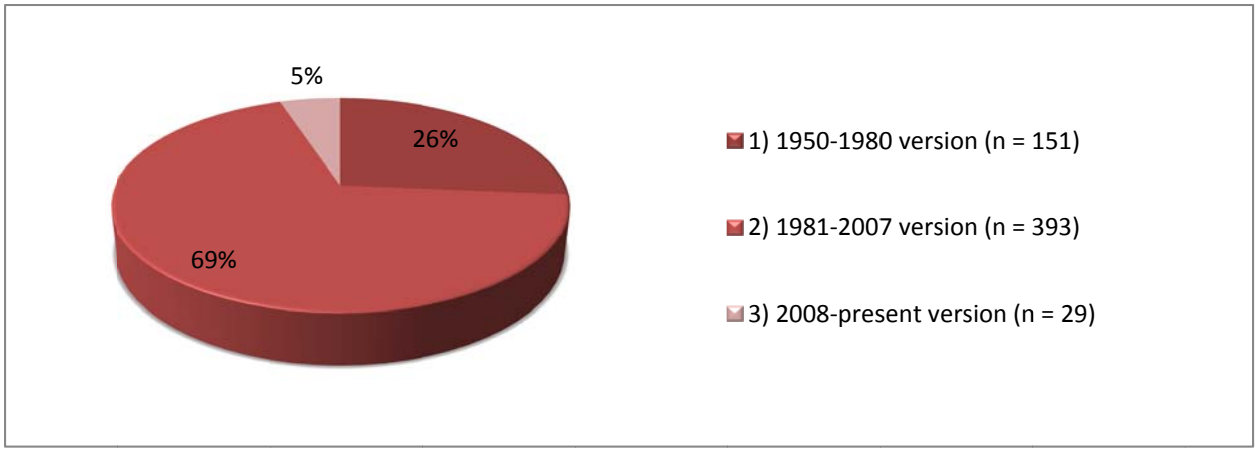


Figure 8 - Distribution of PASS form types in Region 1.

REGION 2

A total of 166 sites were included in the analysis for Region 2.

Location Accuracy

Of the 163 sites in the Region 2 sample, 69% are mapped on USGS maps or contain highly detailed maps on their PASS forms. The remaining 31% of forms contain no locational data, are only referenced by coordinates, or contain unreliable hand drawn maps (Figure 9). Within the CRGIS database, approximately 92% of the archaeological sites have been accurately mapped. The remaining 8% of sites in the CRGIS data base were mapped 250 m or more from the locations shown on the forms or have not yet been mapped in the database (Figure 10).

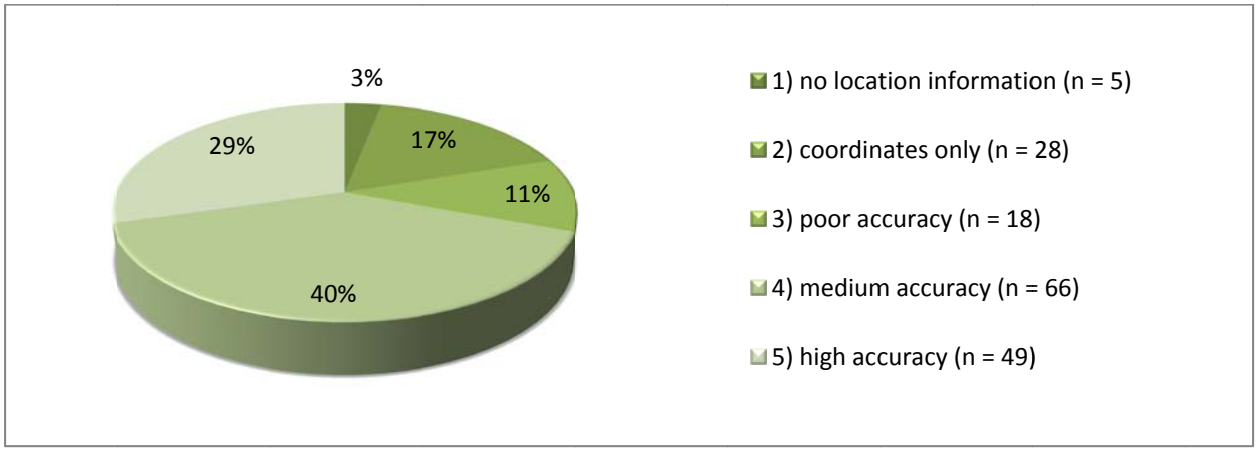


Figure 9 - Quality of location information on PASS forms within Region 2.

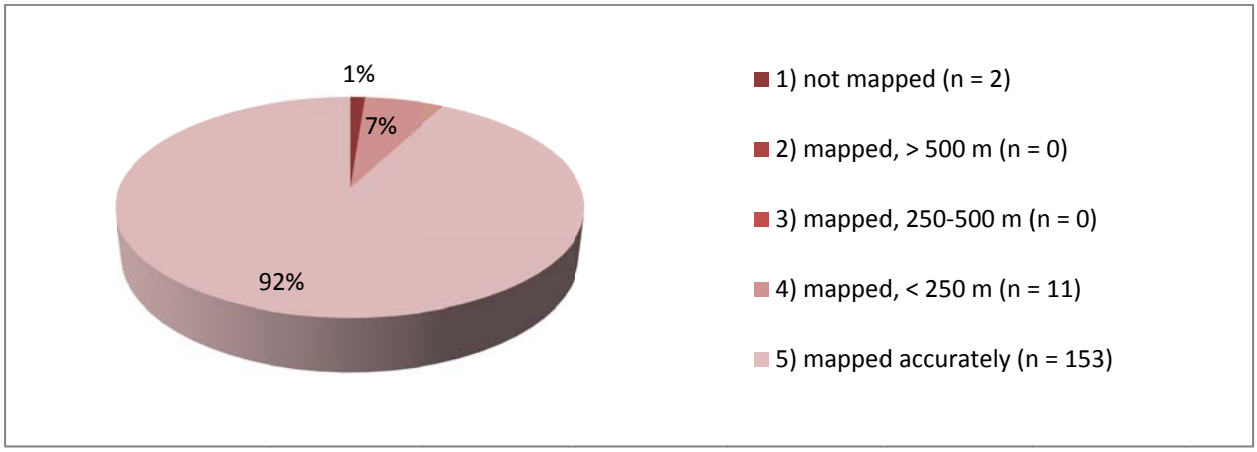


Figure 10 - Quality of location information reflected in CRGIS within Region 2.

Artifact Data

More than a third (37%) of the site sample in Region 2 has no artifact data recorded on the PASS form, while another 19% have poor quality artifact data (Figure 11). Less than half (44%) of the site sample has good or moderate artifact descriptions on the PASS forms. The CRGIS database performs somewhat better, with 56% of the sites represented by good quality artifact data, more artifacts listed than appear on the PASS forms, or data that match the PASS forms (Figure 12). Less than half (44%) of the site sample in the CRGIS data base has no artifacts listed, or fewer than appeared on the PASS form.

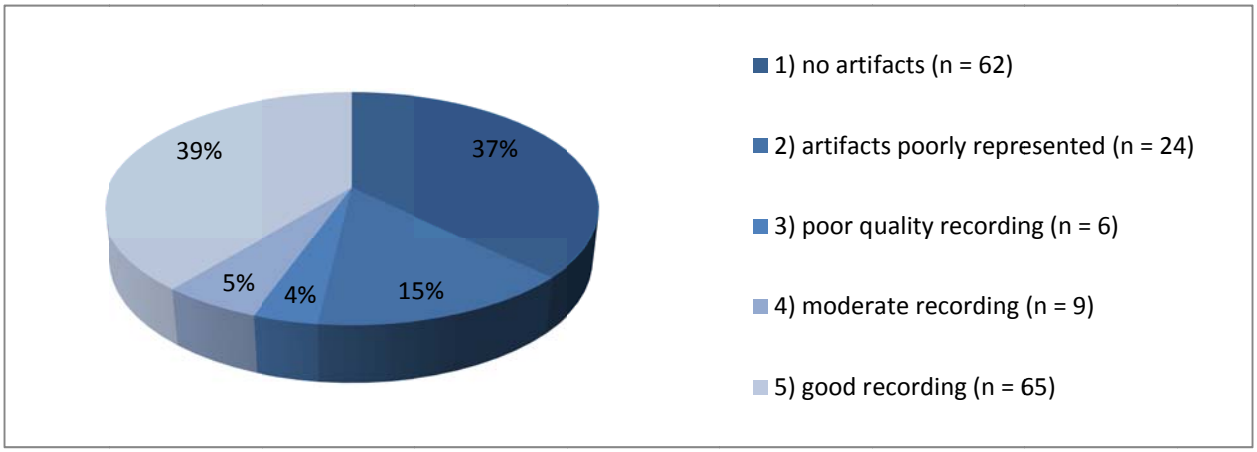


Figure 11 - Original artifact data recorded on PASS forms for Region 2.

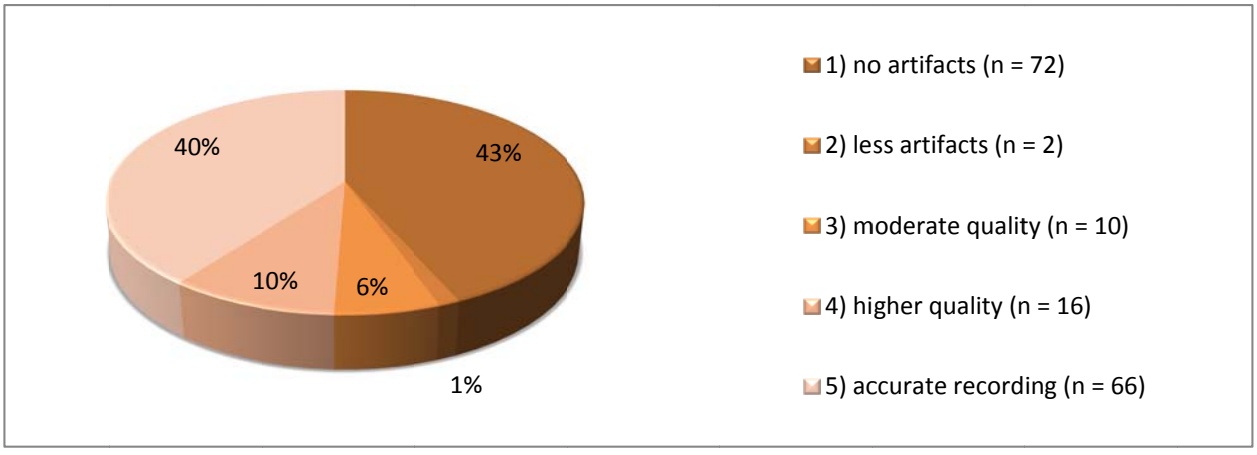


Figure 12 - Artifact data reflected in the CRGIS data base for Region 2.

PASS Form Types and Completeness

A total of 84% of the PASS forms in the Region 2 site sample are at least 75% complete. The remaining 16% of the forms contain limited data (Figure 13). The PASS form types for Region 2 are almost all either older or middle versions, with just 1% on new forms with detailed artifact data (Figure 14).

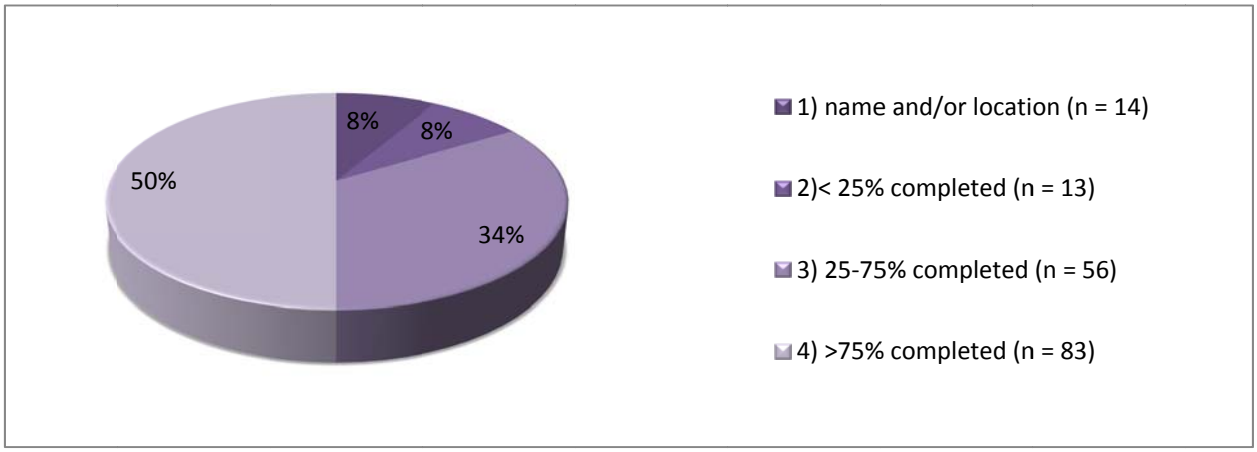


Figure 13 - Completeness of PASS form information in Region 2.

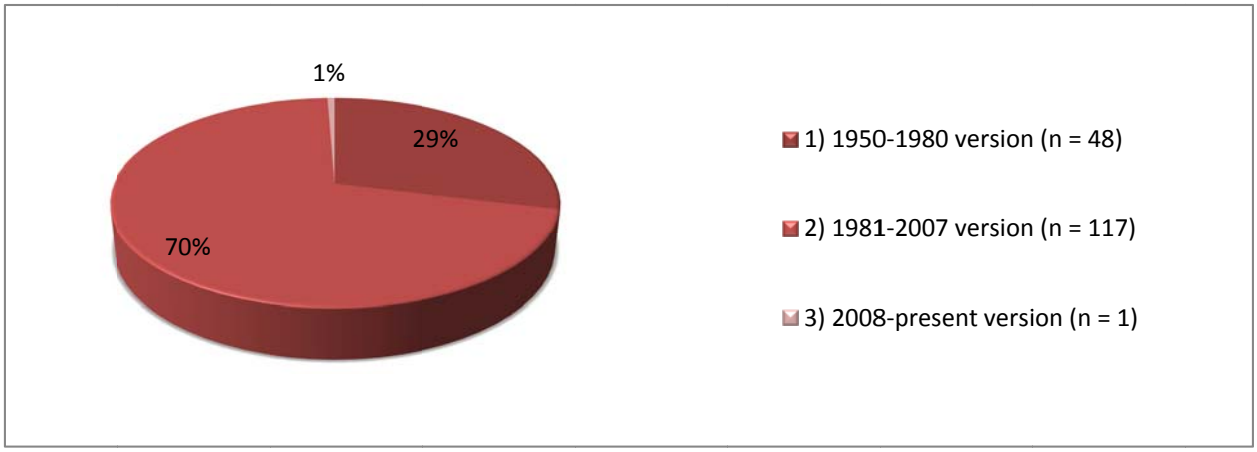


Figure 14 - Distribution of PASS form types in Region 2.

REGION 3

A total of 16 prehistoric archaeological sites were included in this analysis for Region 3. Region 3 is the smallest of the three regions and therefore contains the smallest sample size.

Location Accuracy

The accuracy of mapped locations for the site sample within Region 3 is similar to that for the other two regions. A total of 62.5% of the PASS form sites were mapped on highly detailed or USGS maps, while 37.5% were either not mapped or were referenced only by coordinates (Figure 15). Within the CRGIS data base, 94% percent of the sites are accurately mapped and only 6% are unmapped or mapped more than more than 250 m away from their location as indicated on the PASS forms (Figure 16).

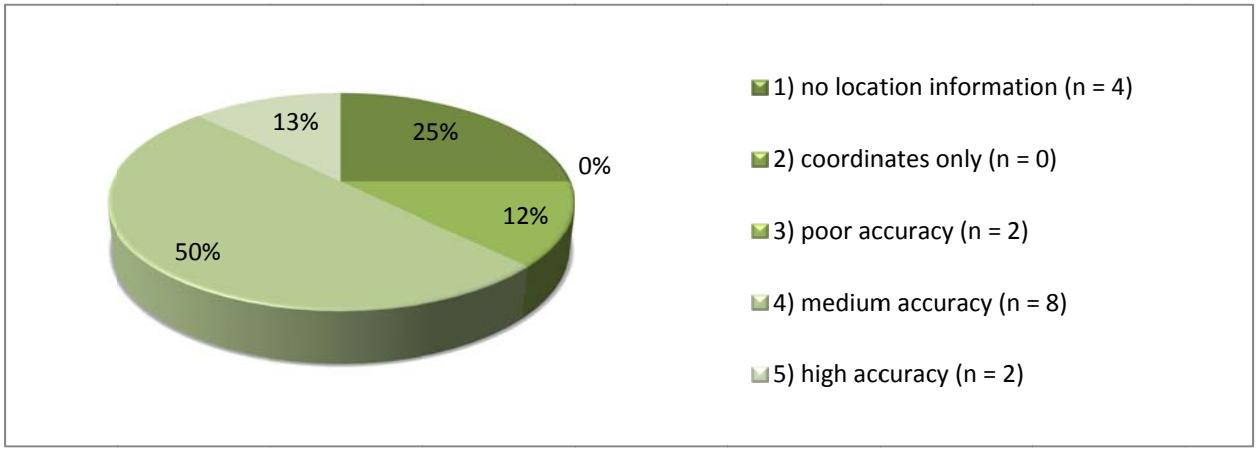


Figure 15 - Quality of location information on PASS forms within Region 3.

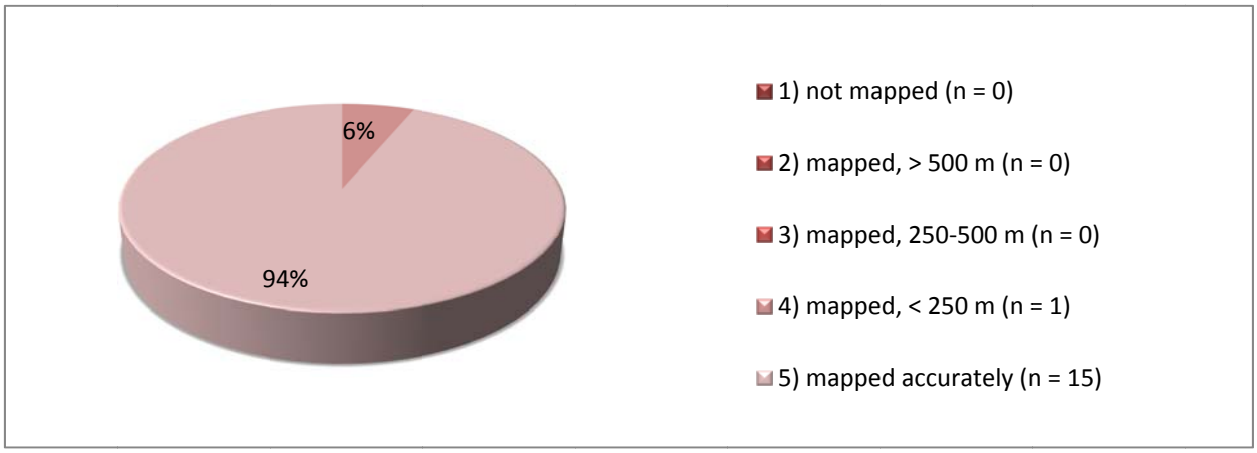


Figure 16 - Quality of location information reflected in CRGIS within Region 3.

Artifact Data

The PASS forms for three-quarters (75%) of the site sample for Region 3 did not contain artifact data or contained poor data (Figure 17). The remaining 25% of PASS forms contained artifact data considered to be moderate or good in quality. While only 12% of the sites within the CRGIS data base contained artifact data that matched the PASS forms, 44% of the sites have more artifact data than is found on the original PASS forms, suggesting that the CRGIS has corrected a large portion of the 69% of sites that had no artifact data on the PASS forms (Figure 18). Only 44% of the sites in the CRGIS data base have no artifact data.

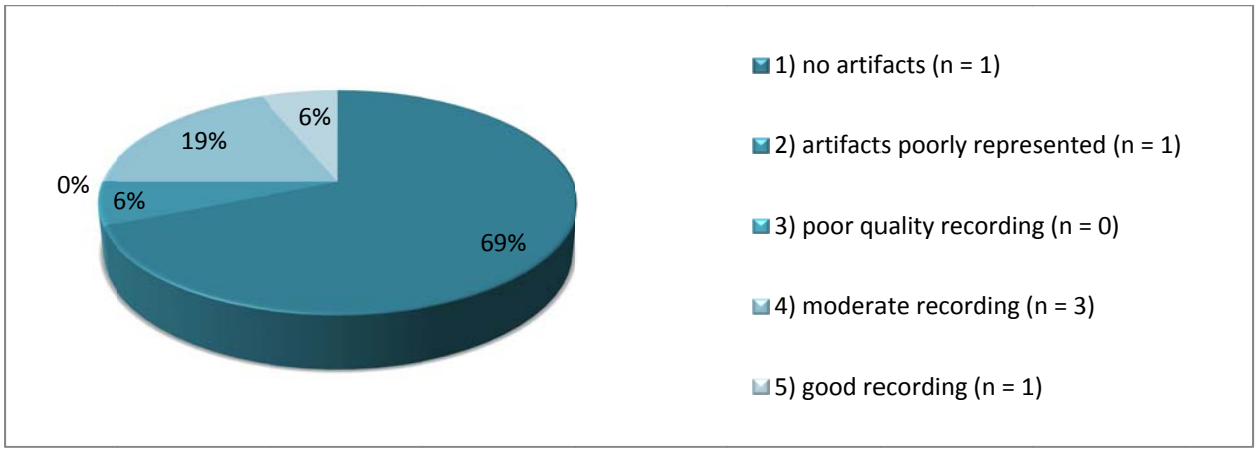


Figure 17 - Original artifact data recorded on PASS forms for Region 3.

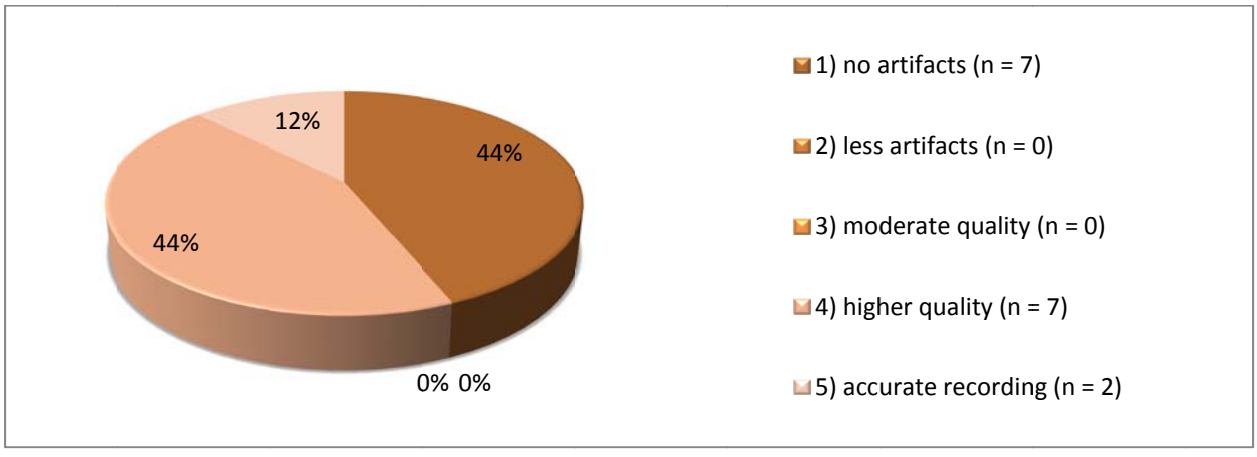


Figure 18 - Artifact data reflected in the CRGIS data base for Region 3.

PASS Form Types and Completeness

The Region 3 results are similar to the two previous regions: 75% of the PASS forms in the site sample are at least 75% complete (Figure 19). This is directly related to the fact that 75% of the PASS forms are middle version forms (Figure 20), which are often filled out completely or contain very little missing data. The remaining 25% of PASS forms within the site sample contain limited information. There were no new version PASS forms within the Region 3 site sample.

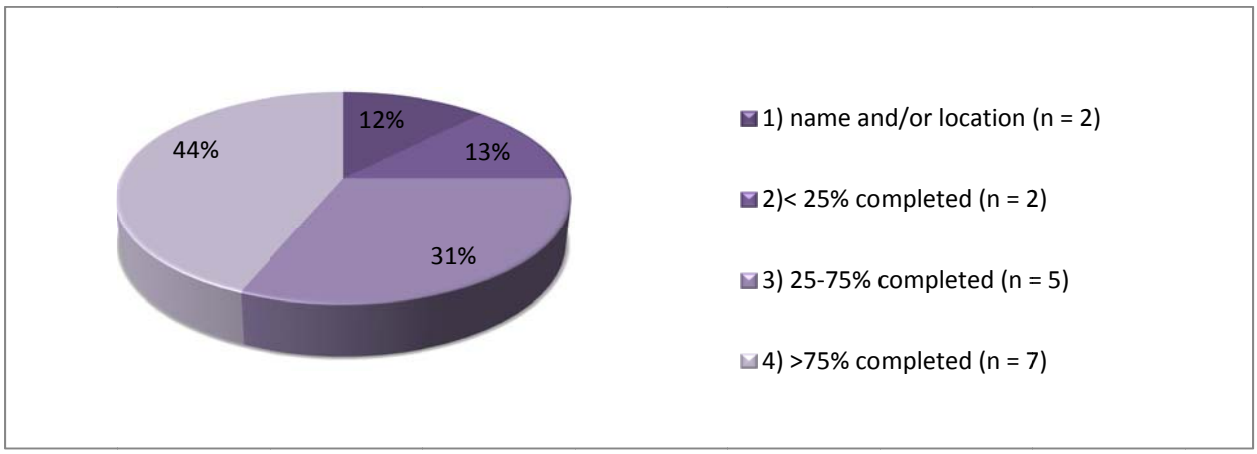


Figure 19 - Completeness of PASS form information in Region 3.

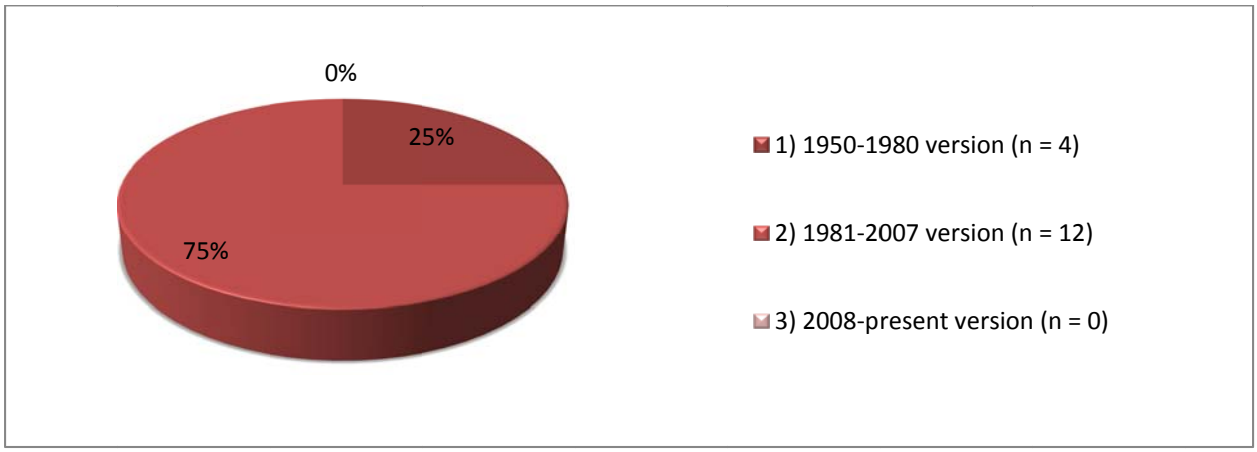


Figure 20 - Distribution of PASS form types in Region 3.

CONCLUSIONS

Overall, the analysis shows that the data derived from the CRGIS data base are at least as complete and accurate as the data included in the original PASS forms, and in some cases, more so. Of the 755 sites in the sample for Regions 1, 2, and 3, only 8 (1%) were missing locational information in CRGIS compared to 26 (3%) that had no locational information on the PASS forms. Errors and missing information on the PASS forms were addressed in the transition to CRGIS, and sites that had no mapping were located and plotted. In some cases, CRGIS staffers navigated to the site locations using non-map information provided on the PASS forms, such as landmarks, creeks, road names, or other locational references. Mapping locations in CRGIS diverged very little from locations provided on the PASS forms, reflecting the accurate transcription of data: of the 755 sites in the sample, only 2 sites (both in Region 1) were mapped 250 m or more from the locations shown on the PASS forms.

Of the 755 PASS forms examined for Regions 1, 2, and 3, 451 (59%) contain good artifact data, while 209 (27%) contain no artifact data, with both categories accounting for 87% of the total site sample. This suggests that most PASS form submitters are recording artifact data thoroughly or not at all. Most of the forms with no artifact data were of the older version that did not provide space for artifact descriptions. Artifact data that was provided on the PASS forms was, overall, accurately transferred into the CRGIS data base: artifact information in the CRGIS data base matched the information in the PASS form for 361 (48%) of the 755 sites. Further, the quality of artifact data was improved upon in the CRGIS data for 106 (14%) of the 755 sites. This reflects a successful effort by CRGIS staffers to track down missing artifact information.

PASS forms have changed over time, and the current version requires more thorough recordation of site locations and artifact data. Most of the sites considered for this analysis were recorded on the "middle" version of the PASS form (n = 522; 69%) and were at least 75% complete (n = 451; 60%). These forms do not include as much information as the newer version, and the data in the CRGIS data base are therefore limited.

4
MODEL METHODOLOGY – REGIONS 1, 2, AND 3

The general approach to modeling Regions 1, 2, and 3 followed the same process used for the pilot model area as described in the Task 3 report. Steps included the following tasks:

- delineation of study areas;
- preparation of PASS data;
- creation of environmental variables;
- extraction of variables for each known site and 500,000 background samples;
- statistical comparison of the variables at sites and various background samples;
- selection of variables that are able to discriminate sites from the background;
- parameterization, creation, and validation of statistical models (Logistic Regression, Adaptive Regression Splines, and Random Forest);
- application of the statistical models to create study area wide predictions;
- collection of predicted probability distributions from sites and the entire study area background;
- establishment of cut-off values to create high, moderate, and low classes; and
- mosaicking of the selected models into a final assessment of prehistoric site location sensitivity.

This process was described in detail within the Task 3 report and will not be repeated here. There are, however, a number of improvements to aspects of the model building process that were adopted for the Regions 1, 2, and 3 models. These adaptations led to a more streamlined process with better parameterization, more consistent sensitivity thresholds, and all together better models. Described below are the aspects of the modeling methodology that have changed from those described in the pilot model (Task 3) report.

ADAPTATIONS FROM PILOT MODEL METHODOLOGY

While the methodology used for the pilot study was very effective in creating successful models that assessed the sensitivity for prehistoric archaeological site locations as well or better than any previously published models, there were aspects that could be modified to improve organization, model processing speed, and model performance. These aspects included a new hierarchy for the delineation and naming of study areas; the creation of a wider array of environmental predictor variables and the inclusion of more variables within each model; the creation of models specific to certain site types in situations where they formed a large percentage of the site sample for a study area; and a new method for the creation of thresholds to distinguish high, moderate, and low

potential; and finally the introduction and discussion of the Cohen's Kappa statistic that is a compliment to the Kvamme Gain and will be used to assess model performance.

Study Region Delineation

Dividing the area of the Commonwealth into smaller sections is essential for two main reasons: 1) models that require the correlation of environmental variables to site locations are most effective in areas of similar environmental character; and 2) given the extremely dense data and computationally expensive statistical methods of this approach, it is necessary to partition the data into tractable chunks. As discussed in the pilot model report, these two requirements are handled in different ways, but are not mutually exclusive. The first constraint is more concerned with using natural boundary definitions and not too concerned with the overall size of each unit—only the variability of the environments within. The second constraint is entirely concerned with the overall size of an area, but most often uses completely arbitrary boundaries that maximize data efficiency. The purpose of the delineation for this project is to partition the landscape into study areas that are small enough to be computationally tractable, but non-arbitrarily bounded and large enough to contain adequate site samples and environmental variability. Each division multiplies the number of models and modeling steps, however. Therefore, choosing a set of boundaries that is efficient in regards to model creation, organization, tractability, and environmental variability is a very important part of this project. A hierarchy was developed to address these requirements that is based principally on physiographic sections and watershed boundaries). The terms used to describe the hierarchy are as follows from largest to smallest:

Region → Zone → Section → Subarea

Regions are the largest partition of the Commonwealth (Figure 21). There are ten total modeling regions, a number decided on an arbitrary basis for the overall organization of the Pennsylvania Model Set project. The regions used for the Task 4 report are the same as those described in the Task 2 report (Harris 2013b) and will continue to be the same throughout the modeling project. The boundaries for the 10 regions are based on grouping similar physiographic sections into regions of very roughly equal size. The exceptions to this are the very small regions numbered 3 and 10. The current report deals with the creation of models for Regions 1, 2, and 3; where Region 3 is merged with Region 2 to create an area comparable in area to Region 1. This is referred to as Region 2/3.

Each region is broken down into a small number of zones based on drainage basin boundaries within physiographic provinces (Table 26 and Figure 22). The use of zones is primarily for organizing the regions into more manageable sizes for the modeling effort. In this report, Region 1 is divided into an east, north, and west zone. Region 2/3 did not require subdivision into zones. From here, zones were further subdivided into units referred to as sections. Sections are defined based on watershed boundaries within physiographic sections. These are referred to as physio-sheds in previous reports for this project. Table 26 shows the number of sections within each region and zone. Note that the

final row indicates the information for Region 3, but, as described earlier, Region 3 is modeled as part of Region 2 for the sake of organization. It is split out in this table only to show that is contained within a different physiographic province.

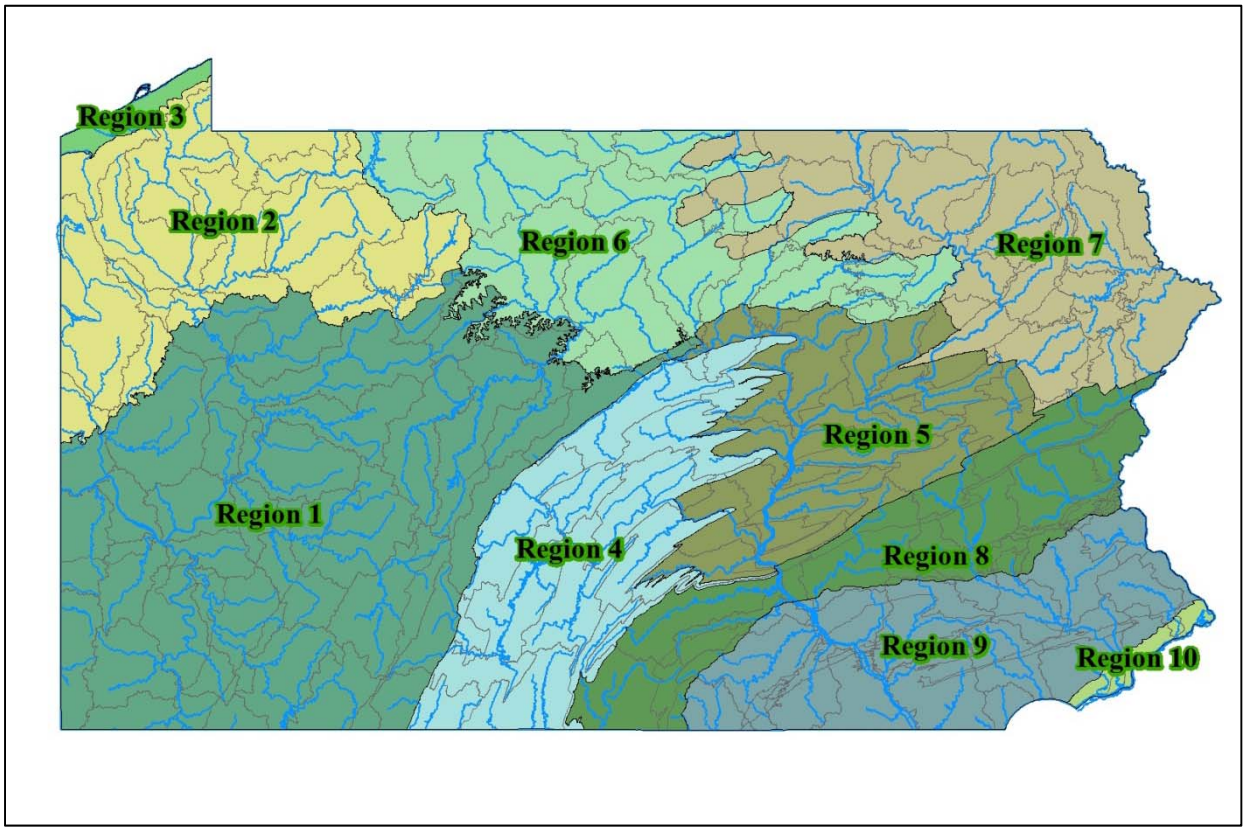


Figure 21 - Modeling regions for the Pennsylvania Model Set project.

The final column in this table shows the subarea. A subarea is simply a section divided into riverine and upland areas. Each subarea represents the study area for a single model, meaning that each subarea was run through the entire modeling process as an individual unit exclusive from the rest. Therefore, for Regions 1,2, and 3 there were a total of 30 subareas and 30 separate model building efforts. This large number of modeling efforts illustrates how rapidly the overall task grows as the areas are split into tractable and manageable chunks. Throughout this report, the subareas will be referred to when discussing individual models and study areas. The results of various statistical tests and model metrics will be displayed and categorized by the subareas since these are the unit of analysis. Subareas will be differentiated by including other elements of the hierarchy such that the expression "R1_east_riverine_section_1" will refer to the riverine subarea of section 1 of the east zone of Region 1. Table 26 includes the physiographic provinces and sections within the study area hierarchy for reference, but the physiographic information will not appear in the study area names or descriptions in the remainder of this report.

Table 26 - Relationship between Regions, Zones, Sections, Subareas, and Physiography

Physiographic Providence	Region	Zone	Physiographic Section	Section	Subarea
Appalachian Plateaus	1	east	Allegheny Front	1	riverine section 1
					upland section 1
			Allegheny Mountain	2	riverine section 2
					upland section 2
				3	riverine section 3
					upland section 3
		north	Pittsburgh Low Plateau	1	riverine section 1
					upland section 1
				2	riverine section 2
					upland section 2
		west	Waynesburg Hills	1	riverine section 1
					upland section 1
				2	riverine section 2
					upland section 2
			Pittsburgh Low Plateau	3	riverine section 3
					upland section 3
				4	riverine section 4
					upland section 4
				5	riverine section 5
					upland section 5
	2	all	Northwestern Glaciated Plateau	1	riverine section 1
					upland section 1
				2	riverine section 2
					upland section 2
			High Plateau	4	riverine section 4
					upland section 4
				5	riverine section 5
					upland section 5
Central Lowlands	3	all	Eastern Lake	3	riverine section 3
					upland section 3

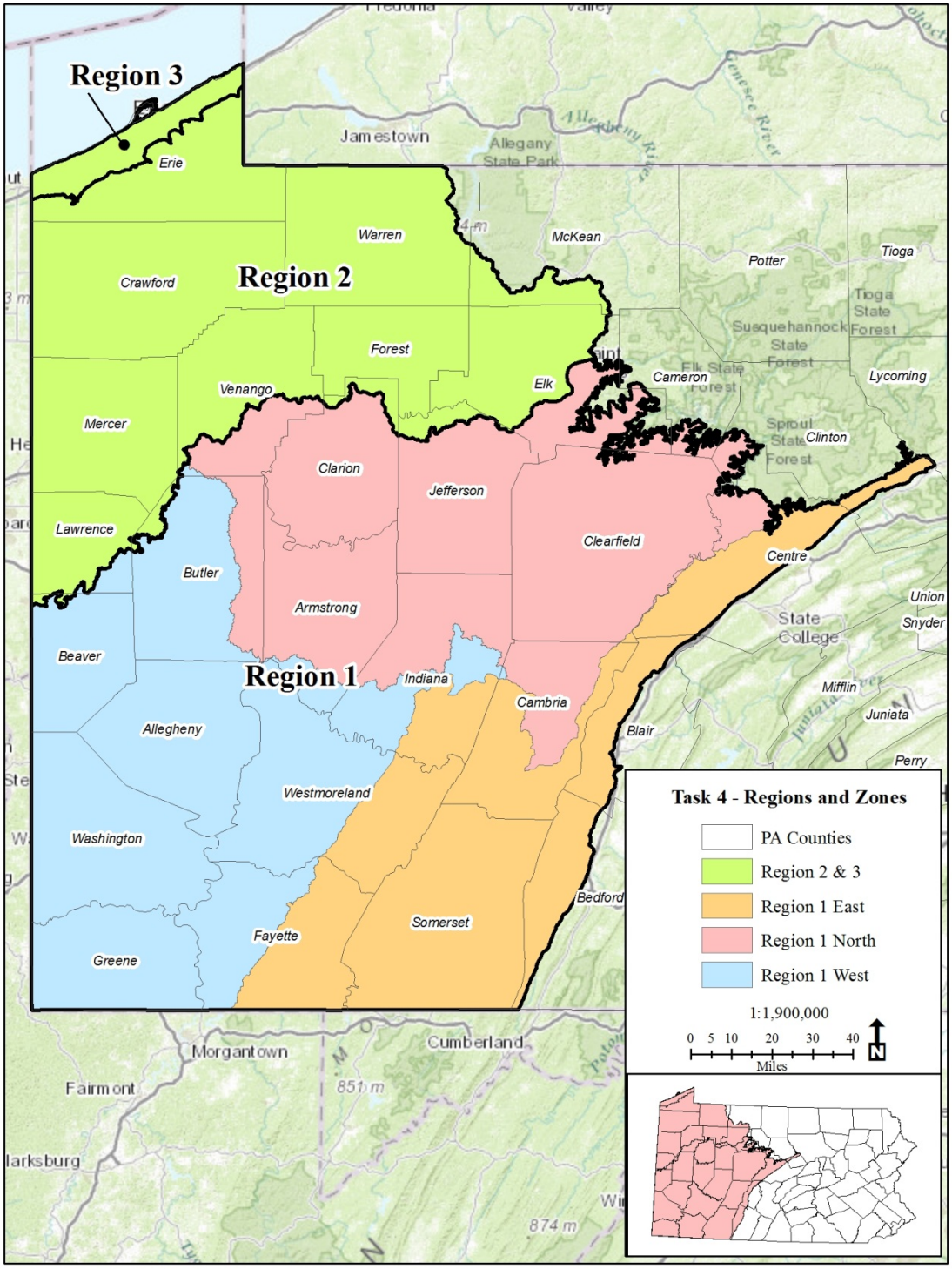


Figure 22 - Task 4 report Regions and Zones.

The division between riverine and upland areas, leading to the distinction of subareas from sections, was done to organize the landscape into two distinct settings. While a physiographic section or a watershed can contain a relatively homogenous environment, the difference between riverine areas and uplands can be quite dramatic, especially in regions of moderate to high relief. Additionally, throughout time riverine areas have offered a variety of different plant and animal resources that are not found in upland areas. Finally, in general the riverine areas of the Commonwealth have a greater density of known recorded sites as compared to upland areas. This may be due in large part to any number of survey biases, but it is nonetheless the case. For these reasons, the model building process needs to consider riverine and upland settings separately.

Riverine settings for these models consider not only modern floodplains, but also a river's terrace system and other near river landforms such as benches. For this project, an automated process of landscape division combines a simple landform model with mapped floodplains to delineate the riverine areas. Everything that is not riverine is upland. Figure 23 depicts the process by which the study area is divided into riverine and upland areas. The first step uses the variables of Euclidian distance from third order and higher streams, elevation above third order and higher streams, and topographic slope to find areas that are near larger streams, relatively flat, and not too high above the stream. The cut points of 600 m from streams, 8% slope, and 12 m in elevation are chosen arbitrarily, but with an understanding of the variation in the region and a conservative approach that does not make the riverine area too narrow, thereby missing important landforms. As shown in Figure 23, the *intersection* of these three variables is taken, resulting in an area that only contains the overlap of these three variables. The result of the intersection is then *unioned* with a layer of floodplains that was created by the Office of Remote Sensing for Earth Resources at Penn State University (PSU) to assist in the permitting of stream encroachment permits within Pennsylvania (PSU 1996). Unlike the intersection operation that took only the overlap of the three variables, the union operation takes both the results of the intersection and the floodplains and combines all areas of both layers. What remains is the portion of the study area that is near larger streams *and* relatively flat *and* not too high above the stream *or* mapped as a floodplain by PSU. In the majority of the cases, the PSU floodplains were contained within the landform intersection. This process can be formalized as:

$$X_R \in F \cup (D \cap E \cap S)$$
$$X_U \notin X_R$$

Where X is the universe of every raster cell in the study area, X_R is the riverine cells and X_U are the upland cells. The notation shows that the raster cells of X_R are members of the set defined by floodplains (F) unioned with the intersection of distance to water (D), elevation above streams (E), and slope (S). Further, the notation shows that the upland raster cells of X_U are all of those cells not in the set of riverine raster cells X_R. Figure 24 through Figure 27 illustrate the results.

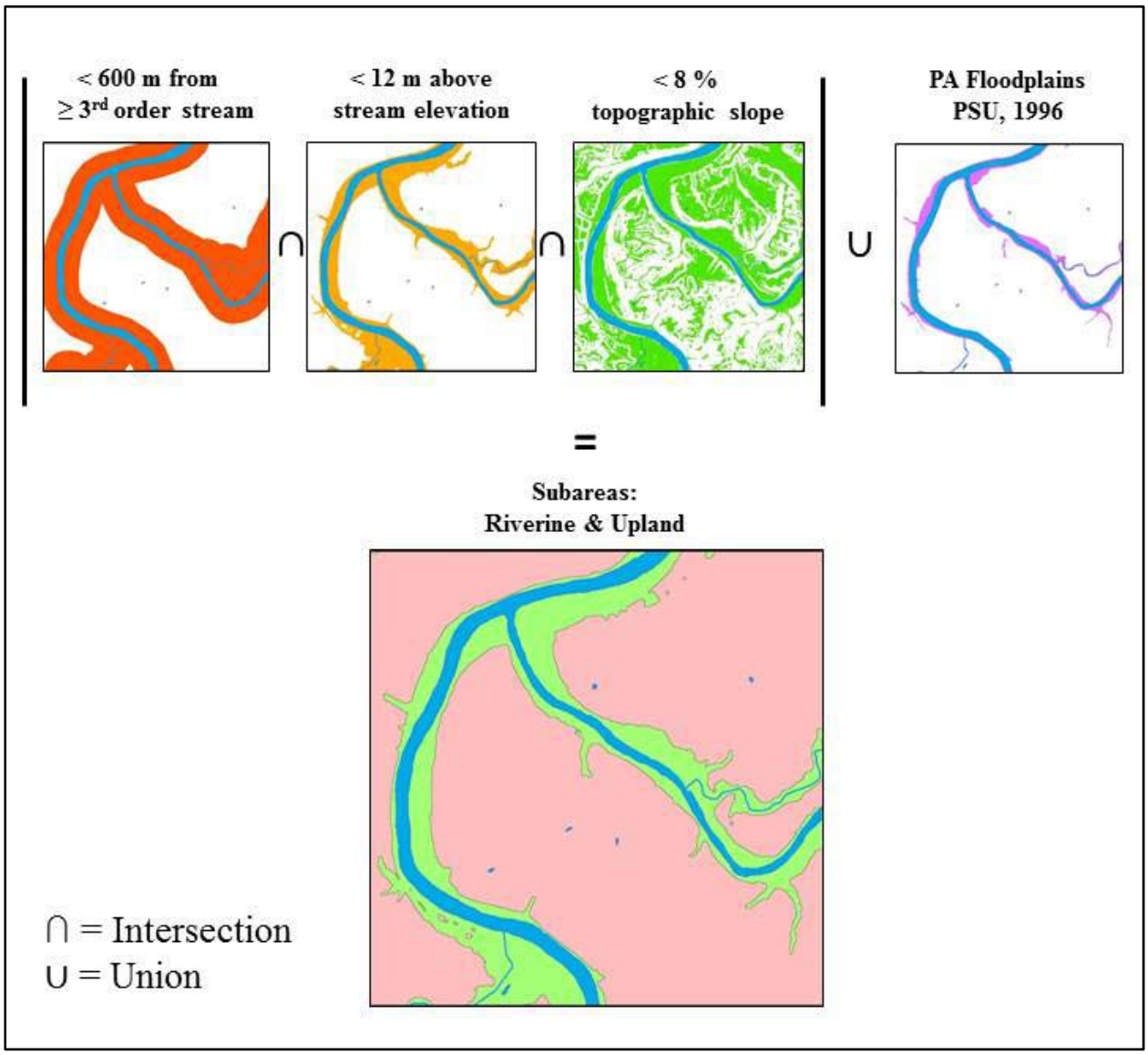


Figure 23 - Schematic process of delineating riverine and upland subareas.

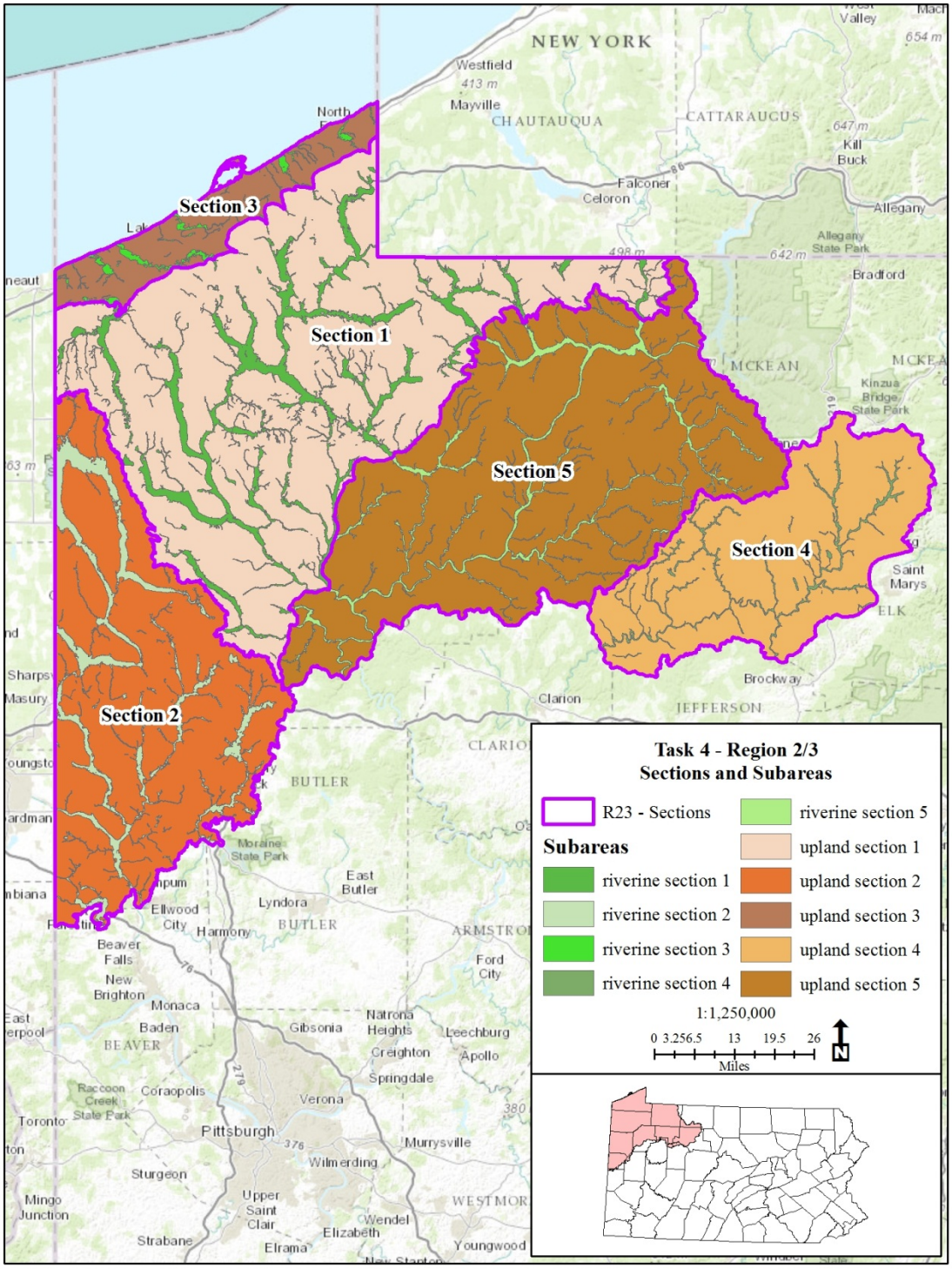


Figure 24 - Modeling subareas of Region 2/3.

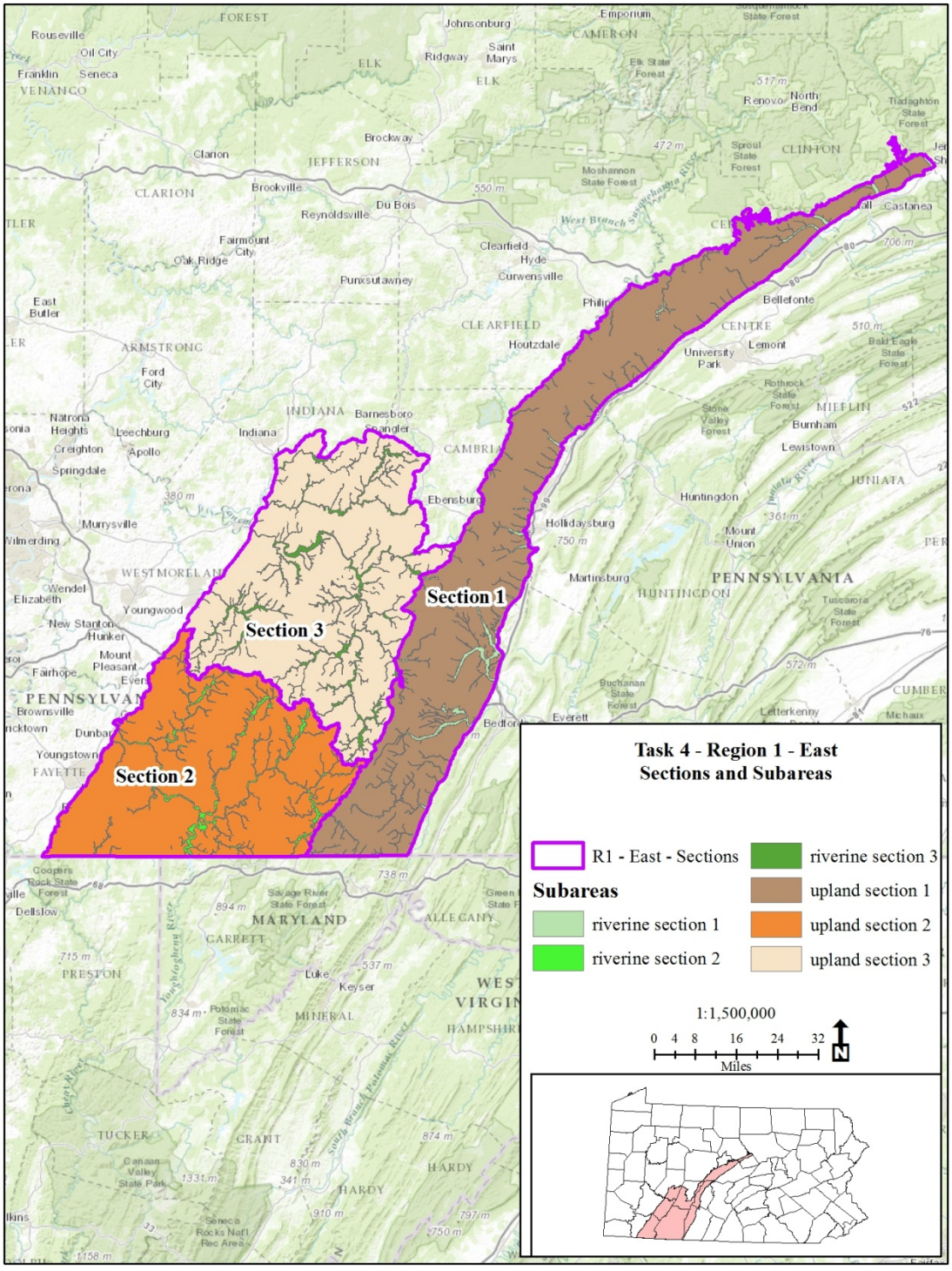


Figure 25 - Modeling subareas of Region 1 East.

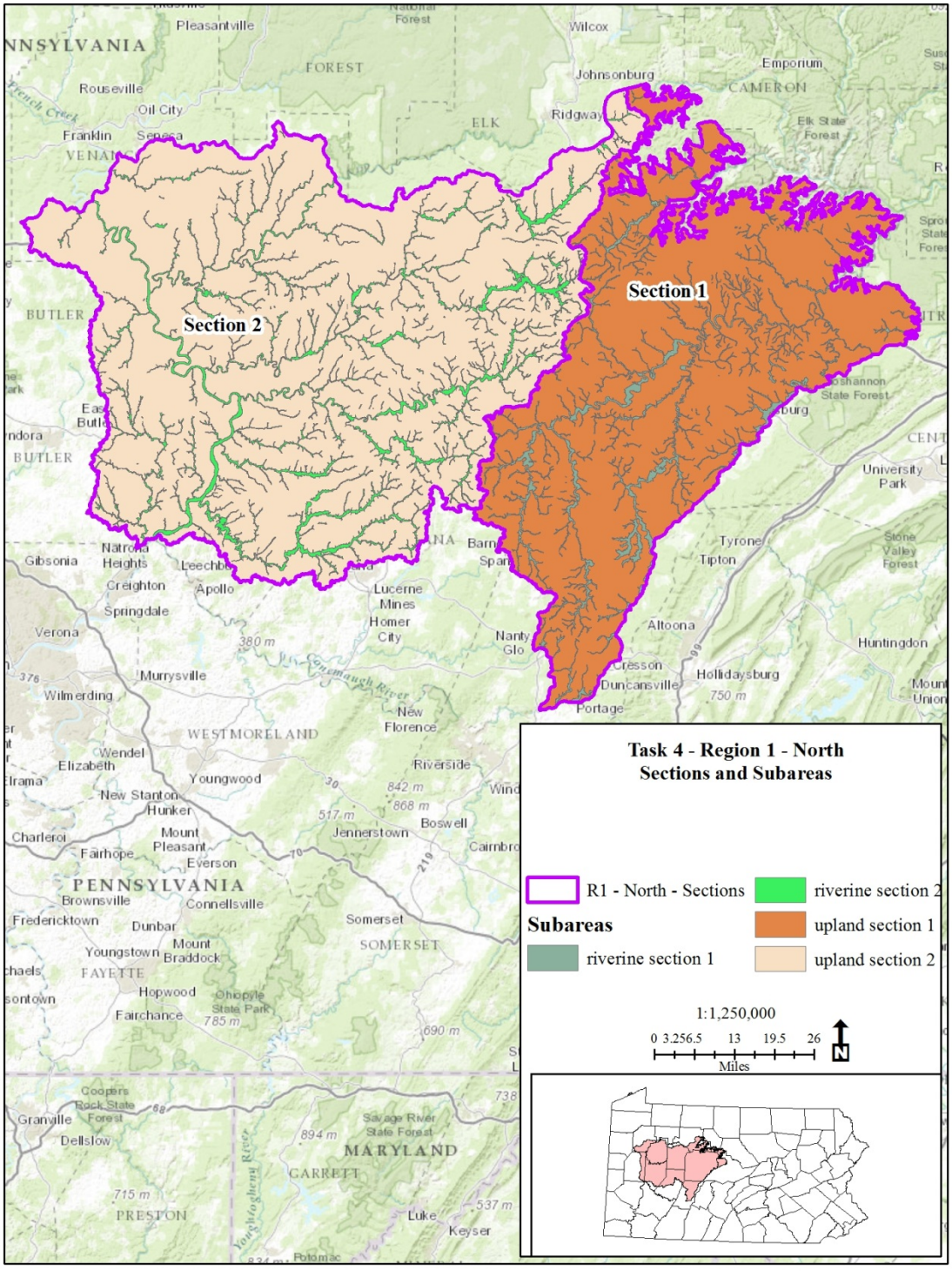


Figure 26 - Modeling subareas of Region 1 North.

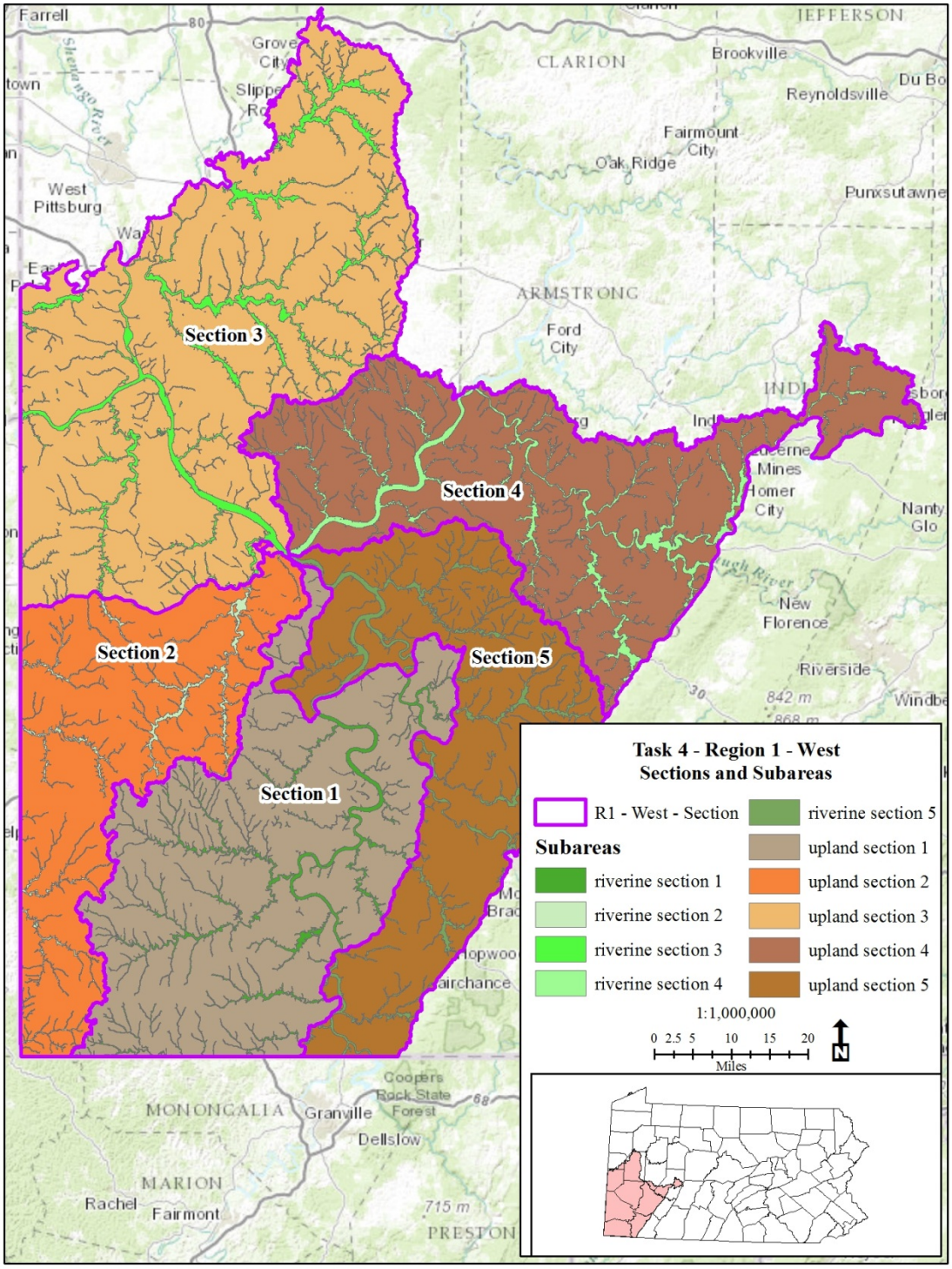


Figure 27 - Modeling subareas of Region 1 West.

Environmental Variables

The pilot model processed a total of 72 environmental variables. From these, a small number of variables (typically four or five) were chosen based on their ability to discriminate site locations from the background. The variables selected for each model were the four or five most highly discriminant. The modeling process documented in this report takes a broader approach to the creation and selection of environmental variables. For this report, a total of 89 variables were created (Appendix B). The wider range of variables includes both new variables (e.g., Euclidian distance to Indian Paths [Wallace 1965]) and a wider range of neighborhoods for previously used variables (e.g., Topographic Position Index for 5, 10, 50, 100, and 250 cell neighborhoods instead of only 5, 10, and 50 cell neighborhoods in the pilot model).

Once all of these variables were tested for their ability to discriminate, approximately half were dropped and the other half were given a second evaluation to identify those that were highly correlated with others. Some variables were highly correlated because they simply measure the same aspect of the environment in two different ways (e.g., standard deviation of slope and range of slope, or distance from third order streams and distance from the historic streams data set). After removing these redundant variables, there remained on average 14 or so variables to use in the creation of each model. The greater number of variables allows the models to find more intricate relationships among the environmental parameters and the presence of site locations. The addition of this larger set of variables was made possible through streamlining of the modeling process and improvements in computational efficiency.

Site Type Models

In the pilot model, all site locations with a prehistoric archaeological component were used within the model building process without any differentiation of site type. As the results demonstrated, this did not have a major impact on the quality of the resulting models given a sufficient site sample size. However, it is intuitive that sites of different types will occupy different portions of the landscape and in essence make sub-samples that have different patterns. Technically, this is an important distinction because the statistical models seek to define a pattern of site location relative to the predictor variables and to project that pattern across the landscape. If there are multiple distinctive patterns, then the model may have more difficulty deciding which patterns to project. While it is true that there are likely variations in the settlement pattern between all site types (e.g., open, village, isolated find, etc…) the largest discrepancy is between rock shelters and other open sites.

Rock shelters by their nature generally occur on steep slopes and are created by specific combinations of bedrock geology, glacial history, hydrology, and vegetation that act on a very local level. While some bedrock geologic units have characteristics that foster the creation of rock shelters, caves, cliff faces, and overhangs (all considered as rock shelters), and this can lead to clusters of rock shelters in a given area, they do not occur with a predictable regularity of landforms such as terraces

and hill tops. Further, the places that might make good rock shelters rarely appeal as good open air site locations. Differentiating site types between rock shelters and all other "open" sites may be very important in subareas that contain a high percentage of rock shelters.

Prior to modeling each of the subareas in this report, the percentage of rock shelter sites was quantified. For any subarea where rock shelters composed greater than 35% of the total number of sites in the subarea, separate models were created, one for only rock shelters and another for all sites excluding rockshelters. Each of these separate models covered the entire surface of the subarea. In two cases, R2/3 Upland Section 4 and R2/3 Upland Section 5, rock shelters composed 64% and 71% of the site sample, respectively. Following the creation of separate rock shelter and non-rock shelter models, a final model was created by combining them into a single raster layer that covers the entire subarea. The final model is created by overlaying the two separate models and taking the maximum predicted value for each cell of the subarea (the combined models are referenced with the suffix of "c" or "combined" in the text and tables below). In this way, each cell in the subarea is evaluated as to whether it is modeled as likely to contain a rock shelter or non-rock shelter site and classified accordingly. In both cases, the final composite model performed better than the first model that considered all sites as a single sample. For future models, the same methodology will be applied if a site type group is found to represent a significant portion (e.g., > 35%) of the site sample within a subarea.

Model Threshold Selection

For each statistical model created for this project, the output is a continuous distribution of probabilities ranging from zero to one. These represent the relative probability that a specific raster cell on a map is a sensitive location for the presence of an archaeological site. For the purposes of assessing model performance and for the final classification into high, moderate, and low sensitivity, the continuous probability distribution needs to be divided at specific thresholds (e.g., 0.0–0.25 = low sensitivity, 0.25–0.65 = moderate sensitivity, and 0.65–1.0 = high sensitivity). The selection thresholds to represent a model's site-present versus site-absent regions is a very important, somewhat difficult, and unfortunately often poorly considered decision. This process is by no means unique to APM, but requires consideration in any field that uses predictive modeling or forecasting (Metz 1978).

Threshold values are necessary for both assessing the ability of competing models to classify and for the final classification into presence versus absence (in this project site presence is considered the combined high and moderate sensitivity areas and site absence is the low sensitivity areas). There is a wide range of methods by which to choose a threshold value for both of these needs. Three obvious ways to choose a threshold are by the model characteristics, by the objectives and needs of the project for which the model was built, or by subjective selection (Liu et al. 2005). The following description of methods applies mainly to using a threshold to partition the final models into site-presence versus site-absence. Choosing a threshold to compare competing models can be done

relatively simply by selecting a static threshold, say 0.5, and comparing characteristics. This is not the only way to address competing models, but the issue of final model thresholds deserves more attention.

Most, if not all, of the previous APM models reviewed in the Task 1 report of this project (Harris 2013a) used the subjective selection method to choose thresholds. By this method, the model author simply chooses a threshold for the sensitivity classes based on personal judgment. This judgment can be based on any number of things such as personal preference, the shape of the probability distribution, or more nefarious reasons such as creating the appearance of model success when, in reality, success is lacking. The latter situation can be uncovered if the author provides a detailed analysis of the full probability distribution without thresholds. This method of threshold selection is *ad hoc* and often difficult to justify or repeat. While it is best to avoid this method of threshold determination, it may have its place depending on the model, results, and initial problem. Regardless, the full probability distribution must also be provided and the choice of thresholds must be thoroughly justified.

The second group of methods for choosing the appropriate thresholds for model selection include project needs, predetermined thresholds, and standards for a particular field of study. In this group, the thresholds may be derived arbitrarily or quantitatively, but either way they are broadly agreed upon by the project team, institution, or across an academic discipline. For example, if a project is designed to correctly predict the outcome of a medical treatment the project team or funding agency may declare that the final model must have an 85% success rate. This number may be arbitrary or may be linked to a previous clinical study, but unlike the subjective threshold above, it is a previously determined and justified goal. Similarly, within a given field of study, it may be documented through peer reviewed journals that the appropriate false-positive rate for any predictive model is less than or equal to five. In this case, a model's authors will threshold the model to achieve the field's standards or use it as a performance benchmark. Note that the threshold selected by the project team or institution may be derived from the quantitative methods discussed below. This group of methods allow for reproducible and comparable models of a similar subject matter or purpose. While the thresholds may still be subjective in some cases, they are generally well documented.

The final group of threshold selection methods is based on quantitative characteristics of a statistical model. The selection of a threshold in this manner often relies on finding the optimal probability value break-point based on the balance of certain criteria. The optimal threshold is often the maximum or minimum of some model metric that seeks to balance the model results for a specific objective. There are many metrics available for this task, many of which are derived from the confusion matrix or ROC plot. Recent studies by Liu et al. (2005) and Freeman and Moisen (2008) use 11 and 12 different model metrics, respectively, to assess how each affects the final model and its implementation. The metrics used in these two studies overlap, but Freeman and Moisen are more focused on ROC based methods, while Liu et al. look at ROC based methods as well as other model metrics. The results of these studies were generally agreeable on many points. Both studies agreed

that thresholding methods that incorporated a measure of prevalence and ROC characteristics were preferable to using a fixed threshold (e.g., 0.5). For models that are created for specific management objectives, Freeman and Moisen argue that thresholds that maximize for a required specificity or sensitivity are preferable. A threshold that requires a *sensitivity* of 0.85 will minimize the site-likely area so that no less that 85% of the known sites are contained within it. Conversely, a threshold that requires a *specificity* of 0.67 will maximize the number of known sites it can fit in a site-likely area of no less than 33% of the study area. Perhaps the more insightful conclusion from Freeman and Moisen (2008:57) is that since the selected threshold is so critical in determining the implementation of the model, it is best to provide the end-user with the complete predicted probability surface in addition to the recommended thresholds. This is the approach that was taken in the Task 3 pilot model (Harris 2014) of this study and will continue on this and future tasks.

In general, all of the methods looked at by Freeman and Moisen and most of the methods used by Liu et al. require the use of the entire probability distribution to calculate the selected metric at many different points along the distribution. The resulting threshold is chosen by selecting the point on the probability distribution where the metric is at its optimum. Some examples of metrics include finding the balance between sensitivity and specificity, maximizing the Kappa statistic, Youden's J statistic, or minimizing the distance from the ROC line to the upper-left hand corner of the ROC plot. The Kavamme Gain (Kg) statistic, the most common model metric used in archaeology, can be used as a continuous metric by which to choose a threshold. Additional methods of this group include using the prevalence measure, using the median of the probability distribution, or selecting a static metric as opposed to an optimized metric. Further, many of these measures can be adjusted to the particulars of the dataset being modeled by including weights for the cost of false-positive and false-negative predictions, as well as for the prevalence of the positive case (e.g., site-present).

The archaeological literature on predictive modeling does not frequently discuss threshold selection or come to any kind of a consensus. The importance of threshold selection was discussed by Kvamme (1988:389–417), however. Following the establishment of the Kg statistic as a means to compare model performance, Kvamme introduced the cross-over graph as a means to visualize and select an appropriate model threshold. The cross-over graph has an x-axis with all of the predicted model probabilities ranging from 0 to 1. The y-axis ranges from 0 to 100 and measures the percent site-present cells within a given predicted probability and 100 minus the percent of background cells present in the same probability (Figure 28). Essentially, this is a measure of the balance between site-present versus background for the full range of predicted probabilities. The importance of the cross-over graph is that it allows for the visualization of the point at which site-present versus background percentages are optimized—the point at which the two lines cross. Given a sufficiently large background sample, this also approximates the point at which sensitivity and specificity are equal. This point can be found quantitatively as:

$$P_x = \min_{0 \le P \le 1} \left(\left| 1 - \left(S_p + B_p \right) \right| \right)$$

Where P_x is the probability value at the cross-over threshold (in Figure 28, P_x = 0.62), S_p is the cumulative proportion of site cells at a given probability threshold, and B_p is the cumulative proportion of background cells at the same threshold. The value for P_x is the probability threshold value that corresponds to minimum value of p for all values 0 to 1. Calculating this for pairs of cumulative site and background cell proportions at all probability threshold values in our example model (the red line in Figure 28), we find the minimum value (min = 0.09) at a probability threshold of 0.62, the same location at which the lines graphically cross-over. This corresponds to S_p= 0.9697 and B_p = 0.0294, representing a model that at the *optimal* threshold correctly predicts a total of about 97% of the site-present cells in a site-likely area of about 3% of the total modeled study area: a Kg of 0.97.

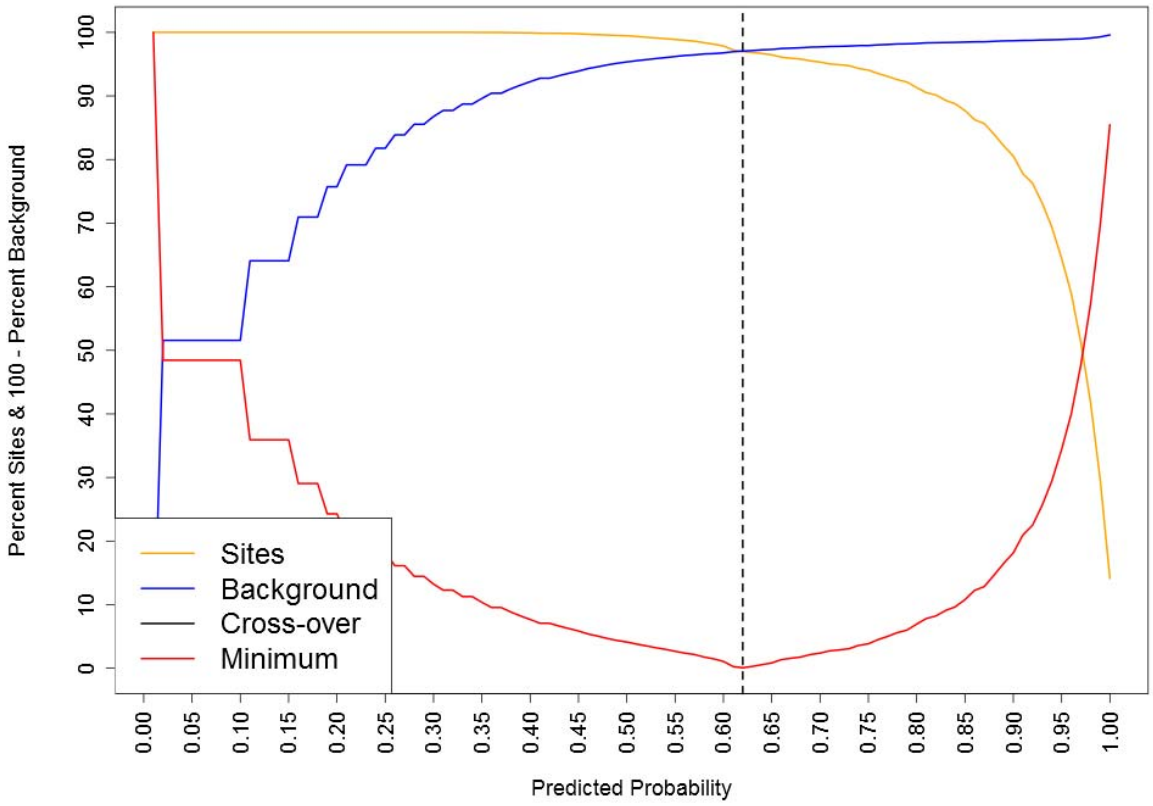


Figure 28 - Example of cross-over graph from a representative model output.

Emphasis is added to the term *optimal* in the previous sentence because the cross-over point is optimal in the sense that it is the point at which wasteful errors (Type I) and gross errors (Type II) are balanced. The problem with utilizing this threshold is that by balancing the Type I and Type II errors,

it assumes that they are equally weighted. It assumes that the cost (in a general sense) associated with predicting a location that truly does not contain a site as site-likely (Type I wasteful error) is equal to the cost of predicting an actual site location as site-unlikely (Type II gross error). A key assumption of this project and many other APM projects is that these two errors are not equal and that inadvertently finding a site during construction is more costly than a negative Phase I survey. While this last point is more within the domain of project implementation and policy than the technical model, it is important to recognize that any threshold decision, even when *optimal*, incurs costs and requires deep consideration.

Within the field of archaeology, there are no standards for the balance of errors, model metrics, or widely accepted sensitivity/specificity targets. The Kg has been used most consistently to discuss model ability, and being a statistic that incorporates both S_p and B_p, it can be calculated across the range of probability thresholds. The Kg is, however, a poor choice for establishing thresholds for two reasons: 1) as a standardized measure it fails because it can have the same value at many different combinations of S_p and B_p (taking two hypothetical model examples; Model A [S_p = 0.45, B_p = 0.20] and Model B [S_p = 0.90, B_p = 0.40] both equal a Kg of 0.56); and 2) as a metric, selecting a threshold that maximizes Kg often ends up in a model biased toward a high gross error over wasteful areas. This means that the highest Kg is achieved when the site-likely area is minimized and known sites are excluded from site-likely areas in order to achieve the smallest site-likely area. The Kg statistic is more effectively used to describe a model at a threshold selected by a different method and to compare similar models at a given threshold.

The Minnesota statewide archaeological predictive model (Mn model) offers the most relevant example of a standardized threshold selection method. The initial goal of the Phase I Mn model was that "these models be accurate enough to predict 85% of known archaeological sites without designating more than 33% of the state's area as high and medium site probability" (Mn/Model n.d.). The Phase I and Phase II Mn models achieved this goal, and the Phase III model, completed in 2007, bettered this goal by predicting 85.5% of the known sites in 23% of the state's area. In a presentation at the 1997 ESRI User Conference, Hobbs (1997) stated that, "[o]ur goal is to have the high and medium probability areas (red and orange on the model maps) occupy as little of the landscape as possible, while still containing approximately 85% of known archaeological sites." This statement shows that the requirement of finding 85% of the sites remained, but the need to minimize the area below 33% was no longer needed. This same goal was reiterated by Oehlert and Shea (2007) in their analysis of the Mn model's statistical methods. Oehlert and Shea (2007:13) state that "[b]ecause the costs associated with accidentally discovering an archaeological site are considered large compared to the costs of avoiding some locations unnecessarily, we recommend fixing a minimum sensitivity, as was done in Mn/Model Phase 3. Among the prediction rules attaining that sensitivity level, choose the rule that is most specific. Here we recommend choosing rules that maximize specificity for 85% sensitivity." Here again, a minimum of 85% of the known sites were required to be contained within the smallest site-likely area possible. In the summary from Oehlert and Shea, they characterize the

85% known site true-positive rate as the 85% sensitivity. This rule is the same as the required sensitivity rule advocated by Freeman and Moisen (2008) in the study discussed above. The threshold generated by this rule simply states that the division between site-likely and site-unlikely areas will be at the probability that accounts for 85% of the known sites in the smallest area possible. Intuitively, this is a pretty reasonable rule given that the goal of most APM has been to make the site-likely area as small as possible while maintaining a reasonable false-negative error rate: 15% in this case.

This rule for selecting thresholds would be applicable here if not for the fact that the accuracy of statistical predictive methods has drastically improved in the years since those recommendations were made. In most of the subareas modeled in Regions 1, 2, and 3, we were able to correctly classify 85% of the sites in a site-likely area of 1% of the subarea, on average. Clearly, there is a significant difference between models trying to minimize the area to less than 33% to models only considering 1% of the study area as archaeologically sensitive. In reality, given the relatively small size and rarity of archaeological sites, the actual prevalence of prehistoric archaeological remains as a percent of the Commonwealth's total area is much closer to 1% than 33%. Currently, the 18,323 identified archaeological sites with a prehistoric component within the PASS files make up 0.2% of the area of Pennsylvania. However, to address the issues of preservation and management, selecting 33% of the landscape to survey for sites is much more satisfactory. Surveying only 1% of the study area leaves too many opportunities to be wrong and miss sites.

Greater model efficiency conveys greater importance on model threshold selection methods. Rules such as seeking 85% sensitivity cannot be universally applied with the same expected outcome; deeper issues of model weights for errors and site prevalence need to be considered. Essentially, the modern statistical models require a rethinking of the goals of APM and a reconsideration of how much survey is enough. The old goal of maximizing known sites while minimizing the site-likely area is no longer valid. The new goal requires a consideration of what we know of site prevalence and relative weights of errors to maximize the known sites while generalizing the site-likely area enough to achieve the objectives of a given project. For example, if a project's goal is to identify the most ideal locations to find a site in a given area or to find the most similar landforms to known sites in that area, then using a higher threshold that minimizes the site-likely area would be appropriate. On the other hand, if the goal is to find all landforms with similar characteristics to the known sites in an area and make sure that few if any known sites are excluded, then using a lower threshold that maximizes known sites and generalizes the site-likely area is preferable. Following Freeman and Moisen's (2008) recommendation and providing the complete predicted probability surface to those decision makers initiating the projects will best ensure that the appropriate thresholds are chosen.

As with Task 3, the models created for this and following tasks will provide the full predicted probability surface. This and future tasks will also provide layers and maps with the probability surface divided into high, moderate, and low thresholds based on thresholds derived from required

specificity and prevalence based approaches. Chapter 6 of this report provides additional discussion on threshold selection methods and appropriateness.

Cohen's Kappa Statistic

As mentioned above, the Kg statistic is a threshold-dependent composite statistic that attempts to synthesize both sides of the balance between Type I and Type II errors (also termed as a balance between sensitivity and specificity, or accuracy and precision depending on the different fields of study). This report will introduce an additional measure that was not used in the previous pilot model report. This measure is called Cohen's Kappa Coefficient, or simply Kappa (Cohen 1960). Kappa is a measure of the observed accuracy of correct predictions versus the expected by-chance accuracy of correct predictions (Kuhn and Johnson 2013). Or stated another way, it is a calculation of total accuracy while accounting for random accuracy. This is actually quite close to what is attempted by the Kg statistic. The Kg uses the percent of the area of the site-likely classification as by-chance percent of finding a site. As such, the Kg evaluates the percent of known sites found against the by-chance percent of finding a site, using the percent of the site-likely area as a proxy. With the Kg, finding 80% of the sites in a model that considers 80% of the study area as site-likely results in Kg = 0. This is because it assumes that the chance of finding a site is equally distributed across the study area; if you predict 80% of the study area as site-likely, then there is an 80% chance of randomly finding a site in that area. The kappa statistic uses a more sophisticated method of calculating the by-chance prediction percent as shown in the equations below.

$$Kappa = \frac{totalAccuracy - randomAccuracy}{1 - randomAccuracy}$$

$$totalAccuracy = \frac{Total\ Correct}{Total}$$

$$randomAccuracy = \frac{Actual\ False * Predcited\ False + Actual\ True * Predicted\ True}{Total * Total}$$

The random accuracy (also referred to as expected accuracy) is generated from the confusion matrix by calculating the by-chance rate using the marginal sums of the confusion matrix and dividing by the total observations. Kappa ranges from -1 to 1, with a value of zero indicating no agreement between prediction and observation. A value of k = 1 indicates perfect agreement, and a negative Kappa value indicates agreement, but in the wrong direction. There is no set range of Kappa to interpret as a *good* versus *bad* model because the meaning of the statistic depends on the problem, field, and data. Landis and Koch (1977) give some guideline by stating that k < 0 indicates no agreement, k = 0 to k = 0.20 are in slight agreement, k = 0.21 to k = 0.40 are a fair agreement, k =

0.41 to k = 0.60 are in moderate agreement, k = 0.61 to k = 0.80 are a substantial agreement, and k = 0.81 to k = 1 are in almost perfect agreement. This is likely good general guidance, but it cannot be applied universally without consideration.

One serious issue that affects the Kappa value is class imbalance; this is also a serious issue that affects prediction using archaeological datasets. This observation was summed up by Viera and Garrett (2005:362): "Kappa is affected by prevalence of the finding under consideration… For rare findings, very low values of kappa may not necessarily reflect low rates of overall agreement." Jeni et al. (2013) used simulated data to model the effects of class imbalance (having many more negative observations than positive observation or vice versa) on the Kappa statistic and found that it can have a significant effect on the results. Their results showed that the Kappa statistic calculated for predictors with the same accuracy but varying degrees of imbalance dropped quickly as the imbalanced was increased. Further, they found that this trend was more drastic as the accuracy of the prediction decreased. As an example, for a classifier with an accuracy of 95%, Kappa was approximately k = 0.80 for a balanced dataset, but dropped to k = 0.6 with a 20:1 imbalance, and k = 0.4 with a 50:1 imbalance (Jeni et al. 2013:5). These results show that a model with the same classification accuracy can result in a wide range of Kappa statistics based on the prevalence of positive observations to negative observations in the prediction.

Class imbalances such as this are a pervasive problem in modeling archaeological site locations. The models and tests presented within this report are calculated at typically a 3:1 imbalance of background cell to site-present cells. For all of the PASS sites with a prehistoric component within the Commonwealth, however, the true imbalance of background cells to site-present cells is 469:1. Because of this massive imbalance, the Kappa statistic becomes very hard to interpret if it is used as a measure of model quality for the final models applied to each subarea. If all of the subareas had roughly equal prevalence of site-present cells, then the Kappa could be used to compare these models to each other, but the scale at which a model is considered *good* would be much lower than the guidelines suggested by Landis and Koch (1977). Between subareas, and especially between the upland and lowland subareas, however, there is large variation on site-present prevalence, so the Kappa statistics would be difficult to compare. An alternative approach is taken in this project to attempt to make a comparable Kappa statistic for the final models as applied to full site-present and background cell populations of each subarea. Presented in Chapter 6, the Kappa statistic is calculated for each subarea model based on the average of all Kappa statistics computed for 1,000 sample background cells and site-present cells at a balance of 3:1. Additionally, the 95% confidence interval is calculated and tabulated. For each of the 1,000 samples for which the Kappa is calculated, a new background cell dataset is drawn (with replacement) from the total population of background cells in that subarea. This is in effect a method of down-sampling the background data to achieve a bootstrapped estimate and confidence interval of the Kappa statistic for a lightly imbalanced data set. While this 3:1 imbalance does not match the much higher imbalance of the real-world site-present dataset, it does give a basis on which to compare each model's relative ability to predict site-present locations better than chance alone.

5
MODEL VALIDATION – REGIONS 1, 2, AND 3

Given the large number of sites located within the subareas of Regions 1, 2, and 3, only statistical models were created for these regions. The judgmentally and proportionally weighted models, referred to in the pilot model Task 3 report as Model 1 and Model 2, were not created for these subareas. The purpose of those models was to provide a base model with few assumptions and clear weights to be used in locations where the statistical models were unlikely to have enough sites to be effective. The group of statistical models, referred to in the pilot model report as Model 3, was able to accurately define the environmental pattern of site locations and extrapolate to unsurveyed areas in Regions 1, 2, and 3. The same three classes of statistical models were used for Regions 1, 2, and 3 as were used in the pilot model study: Logistic Regression (LR), Multivariate Adaptive Regression Splines (MARS), and randomForest (RF). The final models chosen to represent these regions were all of the RF type. These models were more consistent, had high accuracy, greater stability, and had better variable selection as compared to the LR and MARS models. The main body of the report from this point on will only discuss the results of the RF models.

In the Task 3 report, it was concluded that the RF model results were too specific for our purposes. New techniques of parameterizing the RF models were employed for Regions 1, 2, and 3 and a greater number of variables was used for each model, resulting in models with less specificity, but equal sensitivity (accuracy). Essentially, these models were considerably more *well behaved*. Further, when the data quality is adequately high, using the same type of statistical model for all subareas will create a much more consistent result with fewer variations in interpretation. While the LR and MARS models worked well for these subareas, the results of the RF models were preferable because of their ability to discriminate similar landforms and deal with a high level of noise in the data. The RF models will serve as the basis for the final layer divided into low, moderate, and high sensitivity.

PREDICTOR VARIABLES

For Regions 1, 2, and 3 a large number of environmental variables was created and then pared down based on ability to discriminate site locations from background locations. The ability to discriminate was judged based on the Kolmogorov-Smirnov (K-S) test and Mann-Whitney (M-W) U test statistics. Both are non-parametric tests that measure the dissimilarity of two distributions, in this case environmental variables measured at known site locations and those randomly picked from the background. There are specific differences in each test that contribute information valuable to understanding the way in which the two samples are different. Within each region modeled, each of the 89 variables (including a purely random noise variable) was tested against 100 random samples of 50,000 background values. The results are tabulated and the test statistics and p-values are compared to identify those variables that are most discriminant, as well as indications of how site location patterns are expressed within the variable pool. From the list of all variables, those with a K-

S D statistic that was higher than the median were selected; typically this was about 35 variables. From this group, the variables that measured the same aspect of the landscape, but on a different scale (e.g., range in elevation within 10 cells or 16 cells) were pared down so that only the scale with the highest D statistic was left. This resulted in an average of 17 variables for each subarea. Finally, variables with that were very highly correlated were removed resulting in the final selection of predictors, which averaged 14 per subarea. The tables included in Appendix C show the variables that were selected for each subarea, the K-S D statistic, its p-value, the M-W U statistic with its p-value, and the statistics for the random variable for a basis of comparison.

Each of the variables in the Appendix C tables was selected to represent the most discriminant version of the particular part of the landscape that it measures. It is understood that many of these variables will be correlated naturally or by the design of what they measure. The previously discussed steps were taken to eliminate highly correlated or redundant variables, but it cannot be assumed that the remaining variables are truly independent. These are simply the facts of dealing with environmentally based variables. However, the RF statistical method has means of dealing with correlated variables and variables that do not contribute to the success of the prediction.

First, the mechanism by which the RF algorithm tests only a certain number of randomly selected variables at each node of each statistical tree helps to split up potentially correlated variables. As described in the more technical discussion of the RF methods in the Task 3 report, the model parameter of *mtry* sets the number of randomly selected variables that the algorithm tests at each split in the tree. By default, for classification problems, the *mtry* is set to $\sqrt{p}$, where p is the total number of predictor variables. From these randomly selected variables, the split is made based on the variable that leads to the most improvement in results. Therefore, if a fraction of the available variables are selected at each node, then the potentially correlated variables are often separated. This aspect of RF is helpful in dealing with correlated variables, but RF has a more direct way of establishing variable importance.

The RF algorithm is able to measure variable importance in two different ways. The first is by calculating the Gini Index for the purity of each node following a split in the tree. The second is to calculate the decrease in model accuracy with the exclusion of each variable. Since the latter method is what we used for Regions 1, 2, and 3, the former method of the Gini Index will not be elaborated on. Suffice to say that while the Gini Index is not used here to illustrate variable importance, it is used as an internal component of the RF algorithm. The method of illustrating variable importance here is based on a clever method of utilizing the Out-Of-Bag (OOB) sample from the tree building process. To recap the discussion of the RF algorithm in the Task 3 report, the RF method created numerous (in this case 500) individual branching models called trees. In the building of each tree, the training data set is split into two groups, about two-thirds for testing and one-third for training that specific tree. For each tree, the data are reshuffled and split again. At the completion of building each tree with two-thirds of the data set, the remaining one-third of the data set, known as the OOB

sample, is used to test the accuracy of the tree's prediction. The RF algorithm is able to determine variable importance by using each variable and making a prediction with randomized data. The accuracy of the prediction for randomized data is compared to the accuracy of the OOB data to see which predictors have the largest influence on the accuracy of the results. Using random data essentially removes that variable from contention, so if the overall accuracy is not significantly affected by this, then that variable is likely not very influential. On the other hand, if the overall accuracy is greatly reduced by randomizing a particular variable, then it stands to reason that the variable has a lot of influence on the overall accuracy. This process is repeated for each variable on each of the trees and then calculated across all of the trees to determine the relative importance of each variable in achieving high accuracy for the RF model. In addition to providing a way in which to illustrate variable importance, the RF algorithm uses this information to determine if a split will make a significant improvement to the overall accuracy of the final model. The charts in Appendix D illustrate the relative importance of each variable within each of the 30 subareas, as well as for the rock shelter specific models of Region 2/3 Upland Section 4 and Region 2/3 Upland Section 5.

RF MODEL PARAMETERIZATION

From these variables, RF models were built for each of the 30 subareas, as well as for the 2 rock shelter specific models of Region 2/3. The construction of these models followed a similar framework as in the pilot model, with a few variations. Many of the differences between the workflow of the pilot model and this study were increases in efficiency through parallelization of the computer code and more efficient utilization of in-memory models. These changes do not impact the outcome or accuracy of the models, but make the modeling process much more efficient and decrease the opportunity for error. The largest change that does affect the outcome of the models is use of repeated Cross-Validation (CV) to parameterize the RF model. In this case the parameter that is optimized through CV is *mtry*, which is the number of variables that are randomly selected to test at each node in the RF tree.

This method of parameterization uses CV to find the value of *mtry* that leads to the highest prediction accuracy. This optimization is done through repeating a 10-folds CV for the RF model for each of 10 different values of *mtry*. For example, if the total number of variables for a given subarea is 10, then the CV would be repeated for *mtry* values of 1, 2, 3 … 10. Starting with *mtry* = 1, the RF model is created 10 times, one for each of the 10 CV folds, and the accuracy of the model is tested with the data in the hold-out sample (the data that were left out). This process is repeated for each of the 10 values of *mtry* and the error rates of the prediction on the hold-out sample are compared. The *mtry* that leads to the highest prediction accuracy is what is used in the final RF model. Table 27, Table 28, Table 29, and Table 30 display the total predictor variables and the number of *mtry* variables optimized to increase predictive accuracy for the 32 models within each of the 4 modeling zones.

Table 27 - Optimized Number of Variables for RF Parameter *mtry* in Region 1 East Models

Region 1 - East		
Subarea	Total Variables	*mtry*
riverine_section_1	15	10
riverine_section_2	14	6
riverine_section_3	16	11
upland_section_1	17	7
upland_section_2	14	6
upland_section_3	17	7

Table 28 - Optimized Number of Variables for RF Parameter *mtry* in Region 1 North Models

Region 1 - North		
Subarea	Total Variables	*mtry*
riverine_section_1	14	10
riverine_section_2	13	9
upland_section_1	14	6
upland_section_2	13	5

Table 29 - Optimized Number of Variables for RF Parameter *mtry* in Region 1 West Models

Region 1 - West		
Subarea	Total Variables	*mtry*
riverine_section_1	14	6
riverine_section_2	12	12
riverine_section_3	15	10
riverine_section_4	10	7
riverine_section_5	12	8
upland_section_1	10	10
upland_section_2	10	10
upland_section_3	14	6
upland_section_4	15	6
upland_section_5	10	7

Table 30 - Optimized Number of Variables for RF Parameter *mtry* in Region 2/3 Models

Region 2/3 - All		
Subarea	Total Variables	*mtry*
riverine_section_1	11	5
riverine_section_2	9	6
riverine_section_3	14	2
riverine_section_4	10	4
riverine_section_5	12	5
upland_section_1	11	5
upland_section_2	13	5
upland_section_3	17	7
upland_section_4_RS	14	4
upland_section_5_RS	14	6
upland_section_4_nonRS	15	4
upland_section_5_nonRS	16	4

RF MODEL CV ERROR RATES

The final RF models were run on the complete dataset using the *mtry* parameter values listed above and the *ntree* parameter set to 250 for all models. The models were run through 10-fold CV to derive error estimates and the Receiver Operator Characteristics (ROC) Area Under the Curve (AUC) value. The balance between background and site-present data points for model creation was set at a ratio of 3:1, with the background values randomly selected from a pool of 500,000 background values. The final models are fit using the complete set of data and then calculated for the full population of raster cells within the subarea.

Table 31, Table 32, Table 33, and Table 34 detail the error estimates and AUC values for each of the subareas modeled. The first column in these tables contains the Root Mean Square Error (RMSE) value for each model calculated as the average RMSE from each of the 10 hold-out samples. As detailed in the Task 3 report, the RMSE is an error estimate that measures the variation and magnitude of errors between the predicted value and the actual value (e.g., site present vs. site absent); simply put, it is the square root of the average of all squared errors.

Table 31 - RF Model Prediction Errors from 10-fold CV; Region 1 East

Region 1 - East				
Subarea	RMSE	RMSECoV	AUC	Data Sample
riverine_section_1	0.154	2.428	0.990	79896
riverine_section_2	0.136	3.222	0.993	48564
riverine_section_3	0.127	4.694	0.996	37276
upland_section_1	0.068	8.613	1.000	21856
upland_section_2	0.073	4.825	0.999	77952
upland_section_3	0.081	3.804	1.000	51108

Table 32 - RF Model Prediction Errors from 10-fold CV; Region 1 North

Region 1 - North				
Subarea	RMSE	RMSECoV	AUC	Data Sample
riverine_section_1	0.143	4.074	0.995	41624
riverine_section_2	0.197	1.000	0.990	167788
upland_section_1	0.075	4.919	1.000	66596
upland_section_2	0.086	2.571	0.999	151620

Table 33 - RF Model Prediction Errors from 10-fold CV; Region 1 West

Region 1 - West				
Subarea	RMSE	RMSECoV	AUC	Data Sample
riverine_section_1	0.159	4.429	0.995	36380
riverine_section_2	0.210	2.338	0.984	67308
riverine_section_3	0.149	2.459	0.991	157364
riverine_section_4	0.184	1.905	0.989	110456
riverine_section_5	0.156	3.520	0.995	42376
upland_section_1	0.179	1.086	0.994	164212
upland_section_2	0.200	0.881	0.991	254372
upland_section_3	0.093	3.080	0.999	139256
upland_section_4	0.111	2.187	0.998	117136
upland_section_5	0.140	2.872	0.998	112700

Table 34 - RF Model Prediction Errors from 10-fold CV; Region 23

Region 23 - All				
Subarea	RMSE	RMSECoV	AUC	Data Sample
riverine_section_1	0.127	2.087	0.994	646984
riverine_section_2	0.151	2.878	0.991	472312
riverine_section_3	0.088	11.389	0.999	14812
riverine_section_4	0.129	14.209	0.997	13576
riverine_section_5	0.133	2.573	0.993	316408
upland_section_1	0.072	4.676	0.999	365712
upland_section_2	0.080	2.460	0.999	489904
upland_section_3	0.102	4.094	0.998	193920
upland_section_4_RS	0.106	8.381	1.000	47994
upland_section_5_RS	0.111	3.555	0.999	124602
upland_section_4_nonRS	0.069	18.242	1.000	30786
upland_section_5_nonRS	0.080	6.606	1.000	82464

The RMSE estimate ranges from 0 to infinity and is negatively oriented, so the lower the value, the lower the prediction error. In APM, which has a binary response variable (site present = 1 and background = 0), the RMSE is scaled such that 1 is a completely incorrect prediction, 0 is a perfect prediction, and 0.5 is an essentially random prediction. This allows the RMSE numbers for each of the 32 models to be compared relative to each other, but there are factors such as site prevalence and sample size that can influence the RMSE to a small degree (Figure 29). For example, upland subareas have a lower RMSE on average than do the riverine subareas (0.10 vs. 0.15 RMSE, K-S D = 0.765, two-tailed $p < 0.001$).

This is the result of a lower prevalence of site-present locations and an often more restricted choice of site locations in reference to the predictor variables in the upland subareas. The RMSE statistic is very sensitive to large magnitude errors, of which there are more in the riverine areas. This is because there is a higher prevalence of sites and more area that is considered sensitive to archaeological sites. Therefore, there are more cells that are observed to be background (a value of zero) that are predicted to be likely site locations (a value close to one). There are more of these high magnitude differences in the riverine areas, which tend to raise the RMSE; the opposite effect is true for the uplands. However, even with bias derived from known site prevalence and the overall size of the subareas, the RMSE values are all quite low and show models with a high degree of discrimination and the ability to correctly predict known site-present cells from the hold-out samples. The RMSE Coefficient of Variation (CoV) shows the percent change in the RMSE within the 10 RMSE values—one from each of the hold-out samples. The largest RMSE CoV values, which show a larger magnitude of variation between the error rates of the 10 hold-out samples, are from 14.2% to 18.2%. While these show

notable swings in the RMSE of the hold-out samples, the fact that they are percentages of very small RMSE values leads to low error rates even at the upper end of the variation. In general, upland subareas have a slightly higher RMSE CoV on average than the riverine RMSE CoV sample mean (4.87 vs. 4.21, K-S D = 0.235, two-tailed p = 0.734). This insignificant trend is derived from the same biases of prevalence and area noted above.

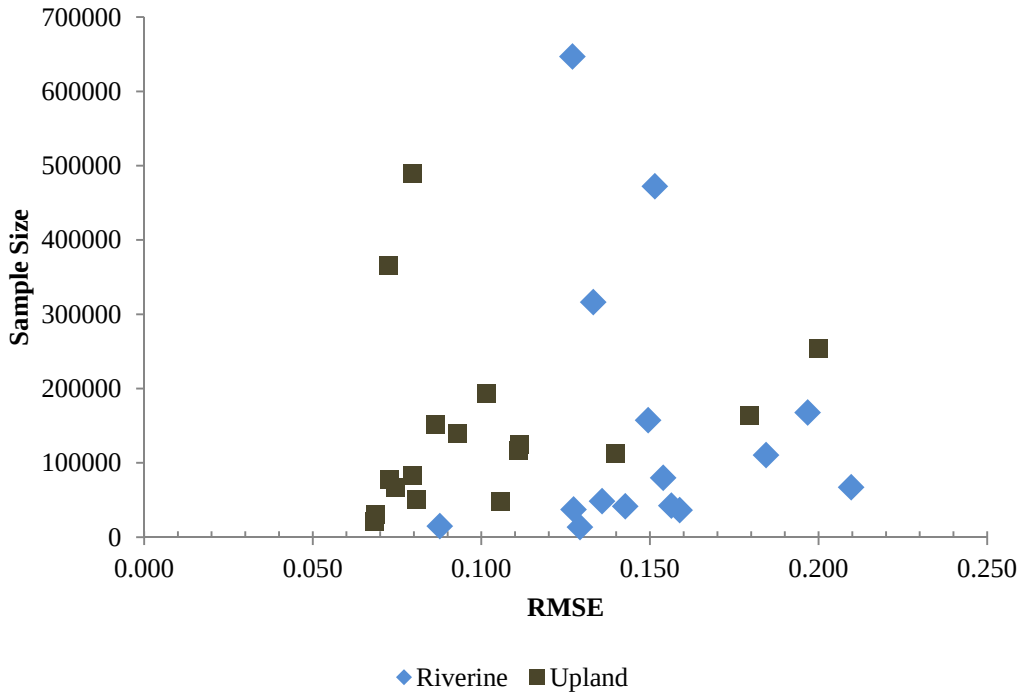


Figure 29 - Comparison of average RMSE values for all upland versus all riverine subareas.

The tables and discussion above show the steps for variable selection, parameterization, and error rates based on 10-fold CV. The error rates resulting from the 10-fold CV, expressed as average RMSE and RMSE CoV, show that the RF algorithm was very successful in identifying the pattern of predictors that define the location of known sites within all subareas. Additionally, the AUC values (a single number that is designed to show the quality of a model across all thresholds) show that the models are very accurate for each of the subareas. Based on these findings, all of the RF models appear to be capable of detecting the known sites as well as predicting the location of site-present cells that were held-out from the model building. There are no red-flags that would indicate that any one subarea has an inadequate or poorly performing model. The findings in the next chapter will demonstrate how these models are applied to each subarea and how the thresholds for sensitivity strata are determined.

6
THRESHOLD SELECTION AND FINALIZATION – REGIONS 1, 2, AND 3

In the previous chapter, the subarea models were validated using 10-fold CV to produce prediction error estimates (RMSE), prediction error stability across hold-out samples (RMSE CoV), and a measure of a model's ability to classify site-present and background cells across the range of predicted probabilities (AUC). From these values, the RF models for each subarea appear to accurately classify known site locations and do so with a relatively low variation in prediction accuracy. Whereas the previous chapter detailed the model building and validation process using random samples of sites and background from each subarea, the data presented in this chapter will show the results of the models applied to the full population of data for each subarea, as well as how choosing different thresholds affects the final evaluation of sensitivity.

COMPARING MODELS AT 0.5 PREDICTED PROBABILITY

The AUC statistic presented in the tables above, along with RMSE, give impressions of the models' accuracy overall. However, as elaborated in the beginning of this report, models that seek to define presence and absence are best evaluated at a given threshold. As discussed, there are many different methods and issues for finding optimal and useful thresholds, but the best method is specific to a single model problem or field of study. For these reasons, a model's applicability and usefulness for a certain purpose is directly related to the threshold that is selected to represent presence and absence. Further along in this chapter, each model will be evaluated at a selected threshold, but this creates an uneven field from which to compare models. In order to better compare the results of models on more level terms, it is best to pick a common threshold and calculate model metrics uniformly. Table 35, Table 36, Table 37, and Table 38 compare each of the models at an arbitrary predicted probability threshold of 0.5.

These tables present a series of metrics that allow the models to be directly compared with one another. As discussed in Chapter 4, the Kappa statistic can be greatly affected by the balance of positive and negative observation; in the case of these models that is effectively controlled by the prevalence of known archaeological sites. For the reasons discussed in Chapter 4, the tables below present a mean from a sample of Kappa statistics drawn from the site-present prediction compared to 1,000 bootstrapped background cell samples, at a ratio of three background cells to one site-present cell. Using the 3:1 ratio down-samples the background cell data set and removes the drastic imbalance created by modeling large areas with low known site prevalence. Further, the 1,000 bootstrapped samples of background cells guards against drawing an unrepresentative sample to represent the environmental background. Even with these safeguards in place, the prevalence of known sites still has some influence on the Kappa, as can be seen in the trend of higher Kappa statistics for upland subareas. Since the Kappa compares the model against an estimate of the chances of randomly finding a site, and known sites are generally dispersed in upland areas, the by-

chance occurrence of sites is lower and therefore the Kappa will be a bit higher for a successful model. However, despite this small bias, the mean Kappa statistics presented in the tables below offer a way to compare the models outright and against each other. The 95% confidence intervals of Kappa sample are also listed. Finally, the tables below present the percent-sites, percent-background, and Kg at the 0.5 threshold. Interestingly, the Kg for the full dataset and 3:1 balanced mean Kappa share a nearly 1 to 1 relationship ($r^2 = 0.9994$) across all subareas.

Table 35 - Comparing Kg and Kappa at a Threshold of 0.5, Region 1 East

Region 1 - East						
Subarea	background %	site-present %	Kg	3:1 Balanced Mean Kappa	Upper 95%	Lower 95%
riverine_section_1	3.37	100.00	0.966	0.935	0.932	0.938
riverine_section_2	2.95	100.00	0.971	0.943	0.940	0.947
riverine_section_3	2.91	100.00	0.971	0.944	0.940	0.948
upland_section_1	0.93	100.00	0.991	0.982	0.979	0.985
upland_section_2	0.87	100.00	0.991	0.983	0.981	0.984
upland_section_3	1.00	100.00	0.990	0.980	0.979	0.982

Table 36 - Comparing Kg and Kappa at a Threshold of 0.5, Region 1 North

Region 1 - North						
Subarea	background %	site-present %	Kg	3:1 Balanced Mean Kappa	Upper 95%	Lower 95%
riverine_section_1	3.19	100.00	0.968	0.939	0.935	0.942
riverine_section_2	7.62	100.00	0.924	0.860	0.857	0.863
upland_section_1	0.94	99.83	0.991	0.980	0.979	0.982
upland_section_2	1.32	100.00	0.987	0.974	0.973	0.976

Table 37 - Comparing Kg and Kappa at a Threshold of 0.5, Region 1 West

Region 1 - West						
Subarea	background %	site-present %	Kg	3:1 Balanced Mean Kappa	Upper 95%	Lower 95%
riverine_section_1	4.66	100.00	0.953	0.912	0.907	0.917
riverine_section_2	7.57	100.00	0.924	0.861	0.856	0.865
riverine_section_3	3.18	100.00	0.968	0.939	0.937	0.941
riverine_section_4	6.06	100.00	0.939	0.887	0.884	0.890
riverine_section_5	3.88	100.00	0.961	0.926	0.922	0.931
upland_section_1	6.56	100.00	0.934	0.878	0.875	0.881
upland_section_2	8.44	100.00	0.916	0.846	0.844	0.848
upland_section_3	1.43	100.00	0.986	0.972	0.971	0.974
upland_section_4	1.89	100.00	0.981	0.963	0.962	0.965
upland_section_5	2.82	100.00	0.972	0.946	0.944	0.948

Table 38 - Comparing Kg and Kappa at a Threshold of 0.5, Region 2/3

Region 2/3						
Subarea	background %	site-present %	Kg	3:1 Balanced Mean Kappa	Upper 95%	Lower 95%
riverine_section_1	2.39	99.94	0.976	0.953	0.952	0.955
riverine_section_2	3.73	99.92	0.963	0.928	0.926	0.930
riverine_section_3	1.63	100.00	0.984	0.968	0.960	0.976
riverine_section_4	3.95	100.00	0.961	0.925	0.912	0.938
riverine_section_5	2.89	99.96	0.971	0.944	0.942	0.946
upland_section_1	1.07	99.98	0.989	0.979	0.978	0.980
upland_section_2	1.00	100.00	0.990	0.980	0.979	0.982
upland_section_3	1.84	99.94	0.982	0.964	0.961	0.966
upland_section_4c*	3.59	100.00	0.964	0.932	0.926	0.938
upland_section_5c*	4.42	99.96	0.956	0.916	0.912	0.921

* combined rock shelter and non-rock shelter specific models

The above tables show that the models as applied to the full subarea study area are very good at identifying the location of site-present locations relative to a random chance of finding a site. Between the models, the Kappa results show a relatively consistent trend. As illustrated in Figure 30, the mean Kappa statistics range from a low of $k = 0.846$ to a high of $k = 0.983$: a rather narrow range. Furthermore, the 95% confidence interval is generally quite small. The most notable trend in this figure is the majority of upland subareas scoring a higher Kappa than the majority of riverine

subareas. This trend is most likely attributable to the lower prevalence of known sites in the uplands and the lower chance of randomly findings a site there. The Kg statistic and site/background percentages show that the models are very successful at capturing the known site pattern within a very small portion of the model.

As explained in Chapter 4, the discriminatory ability of these models is at a level not yet seen in APM and raises a new host of questions regarding the purpose and intention of these models. The low background percentages of these models relative to the site-present percentages are drastically smaller than most previous APM, but in fact reflect the reality of a low prevalence phenomenon such as archaeological sites. While the models and methodology employed here have been adjusted to account for low prevalence and unequal weights between false-positives (low weight) and false-negatives (high weight) the reality that archaeological site occurrence only comprises a very finite portion of the total landscape is inescapable. The means of dealing with this reality have now been shifted from using the lower discriminant, less accurate, and obfuscated models of the past to using more thoughtful interpretation, problem specific model applications, and a better understanding of the model's abilities and limitations. A large part of this reckoning is the better understanding and application of model thresholds (discussed below).

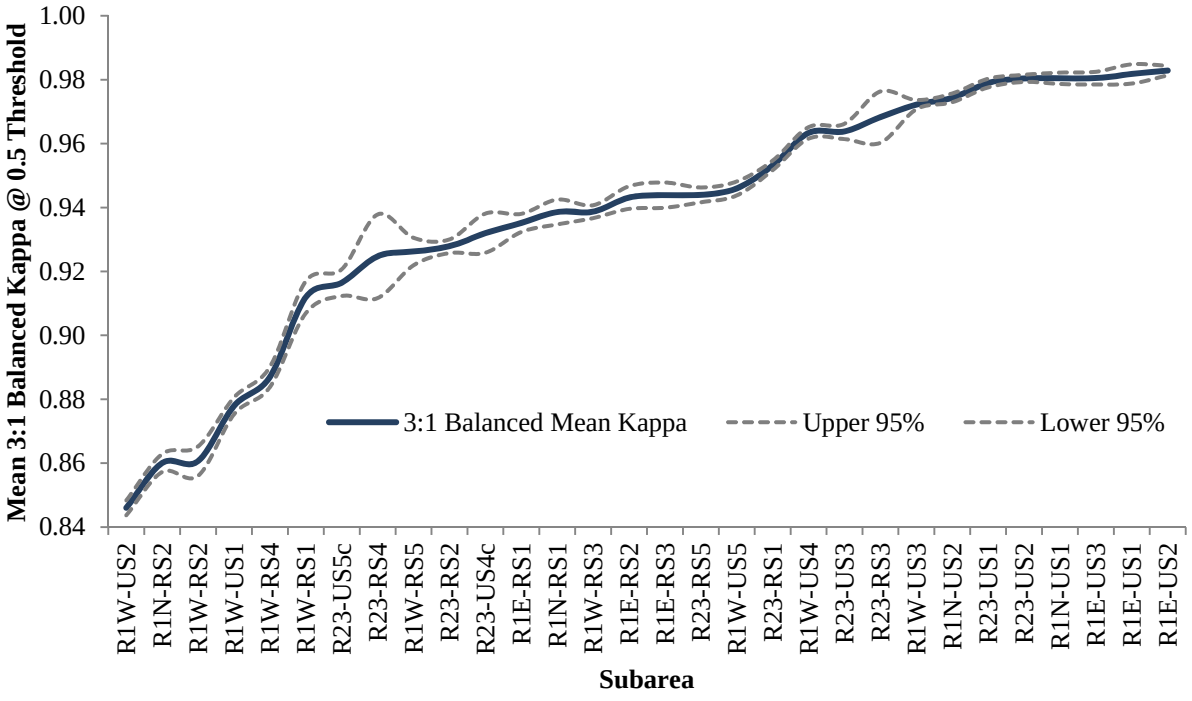


Figure 30 - 3:1 balance mean Kappa and 95-percent confidence intervals for all subarea models.

ESTABLISHING MODEL THRESHOLDS

The issues and opportunities associated with model threshold selection are discussed in some detail in Chapter 4. Essentially, due to the ability of modern statistical models to identify patterns and discriminate site locations much more effectively than in the past, the onus of portioning site-present from site-absent areas has shifted. In the past, many model-building efforts had the simple goal of maximizing the site-present percent and minimizing the site-likely area to as small as the model would allow. This was the primary challenge of the modeling effort, and the thresholds that determined site-likely areas were often an afterthought or predicted on the low performance of the model. With the RF models and other innovations in statistical modeling, achieving very well fit, and at times overfit, models is not as great of a challenge. No longer is the goal of simply reducing the area within which a majority of the sites are contained sufficient. The models presented here are capable of minimizing that area to a small portion of the landscape that is closer to the true prevalence of known sites and more sensitive to previous survey bias. The new goal given these advances is to accurately model the site pattern with a low error rate and then select model thresholds that best achieve the goals of the project. If the project aims to minimize the site-likely area, then a higher threshold is useful. To generalize the site-likely area, a lower threshold is useful. As discussed in Chapter 4, the selection of an appropriate threshold can be based on a number of factors, including arbitrary decisions, field or project specific standards and goals, or optimization based on quantitative model metrics. For Regions 1, 2, and 3, a number of potential thresholds and the models' predictive ability at those thresholds were examined.

Table 39, Table 40, Table 41, and Table 42 present eight different potential thresholds based on optimized model metrics and previous research in APM. These values are graphically represented in a chart for each subarea, included as Appendix E. Much of the description to follow is drawn from Freeman and Moisen (2008); this reference can be used as a starting point to explore the more technical details of these threshold measures. The thresholds presented here are termed as:

- MaxKappa: the threshold that maximizes the Kappa statistic
- Max Kg: the threshold that maximizes the Kg statistic
- Sens=Spec: the threshold at which sensitivity and specificity are equal
- X-Over: the threshold at which site-present and background lines cross in the cross-over graph
- Sens @ 0.85: the threshold that is optimized for a sensitivity of 0.85
- Spec @ 0.67: the threshold that is optimized for a specificity of 0.67
- Pred=Obs: the threshold at which the predicted site prevalence equals the observed or assigned site prevalence (calculated at two different assigned values)

Table 39 - Optimal Thresholds for Various Selection Methods, Region 1 East

Region 1 - East								
Threshold Type	Maximize		Balanced		Domain Specific		Prevalence Based	
Subarea	MaxKappa	MaxKG	Sens=Spec	X-Over	Sens @ 0.85	Spec @ 0.67	Pred=Obs @ 0.1	Pred=Obs @ 0.2
riverine_section_1	1	1	0.770	0.770	0.950	0.100	0.28	0.160
riverine_section_2	1	1	0.790	0.788	0.950	0.090	0.23	0.140
riverine_section_3	1	1	0.790	0.790	0.960	0.110	0.26	0.160
upland_section_1	1	1	0.800	0.810	0.980	0.080	0.18	0.110
upland_section_2	1	1	0.810	0.810	0.990	0.080	0.15	0.100
upland_section_3	1	1	0.810	0.810	0.980	0.130	0.22	0.160

Table 40 - Optimal Thresholds for Various Selection Methods, Region 1 North

Region 1 - North								
Threshold Type	Maximize		Balanced		Domain Specific		Prevalence Based	
Subarea	MaxKappa	MaxKG	Sens=Spec	X-Over	Sens @ 0.85	Spec @ 0.67	Pred=Obs @ 0.1	Pred=Obs @ 0.2
riverine_section_1	1	1	0.790	0.794	0.950	0.110	0.29	0.19
riverine_section_2	1	0.998	0.770	0.770	0.910	0.220	0.47	0.33
upland_section_1	1	1	0.760	0.762	0.980	0.080	0.19	0.13
upland_section_2	1	1	0.800	0.804	0.980	0.130	0.23	0.16

Table 41 - Optimal Thresholds for Various Selection Methods, Region 1 West

Region 1 - West								
Threshold Type	Maximize		Balanced		Domain Specific		Prevalence Based	
Subarea	MaxKappa	MaxKG	Sens=Spec	X-Over	Sens @ 0.85	Spec @ 0.67	Pred=Obs @ 0.1	Pred=Obs @ 0.2
riverine_section_1	1	1	0.780	0.786	0.940	0.170	0.37	0.25
riverine_section_2	1	1	0.760	0.760	0.900	0.200	0.47	0.32
riverine_section_3	1	1	0.770	0.772	0.950	0.090	0.27	0.15
riverine_section_4	1	1	0.770	0.770	0.920	0.180	0.43	0.28
riverine_section_5	1	1	0.780	0.780	0.940	0.180	0.35	0.25
upland_section_1	1	1	0.810	0.810	0.930	0.160	0.41	0.26
upland_section_2	1	0.994	0.780	0.782	0.920	0.240	0.48	0.34
upland_section_3	1	1	0.800	0.798	0.970	0.110	0.22	0.15
upland_section_4	1	1	0.790	0.792	0.960	0.150	0.27	0.20
upland_section_5	1	1	0.790	0.794	0.950	0.190	0.34	0.25

Table 42 - Optimal Thresholds for Various Selection Methods, Region 2/3

Region 2/3								
Threshold Type	Maximize		Balanced		Domain Specific		Prevalence Based	
Subarea	MaxKappa	MaxKG	Sens=Spec	X-Over	Sens @ 0.85	Spec @ 0.67	Pred=Obs @ 0.1	Pred=Obs @ 0.2
riverine_section_1	1	1	0.750	0.756	0.950	0.120	0.24	0.16
riverine_section_2	1	1	0.760	0.764	0.930	0.100	0.28	0.16
riverine_section_3	1	1	0.800	0.802	0.960	0.120	0.22	0.16
riverine_section_4	1	0.998	0.750	0.748	0.950	0.150	0.34	0.22
riverine_section_5	1	1	0.750	0.756	0.940	0.100	0.27	0.16
upland_section_1	1	1	0.770	0.776	0.980	0.080	0.17	0.11
upland_section_2	1	1	0.770	0.774	0.970	0.130	0.21	0.16
upland_section_3	1	1	0.760	0.764	0.970	0.090	0.19	0.12
upland_section_4c*	0.98	1	0.760	0.764	0.960	0.240	0.38	0.29
upland_section_5c*	0.98	1	0.740	0.742	0.960	0.230	0.39	0.30

* combined rock shelter and non-rock shelter specific models

The first two thresholds, MaxKappa and MaxKg, are means of maximizing a particular metric to find a threshold. In this case it is maximizing Kappa, or maximizing the proportion of correctly classified sites while accounting for random agreement, and maximizing Kg, or maximizing the proportion of correctly classified sites while accounting for the area of the classification. As previously discussed, these two measures are related and highly correlated. Finding a threshold that maximizes for either of these two is akin to achieving the goal of finding the largest percent of sites in the smallest percent of the site-likely area, a goal that is no longer congruent with highly discriminant models such as RF. This can be easily observed in the fact that nearly all of the MaxKappa and MaxKG thresholds are a value of 1. This is a threshold that would identify a large percent of the sites, but include only a tiny fraction of the subarea. This would only be a preferable threshold if the intention of the model was to find the most highly sensitive areas or where one has the best chance of finding an unrecorded site to survey. These are not the understood goals of this project, but the MaxKappa and MaxKg thresholds are included as a reference point.

The second two threshold metrics Sens=Spec and X-Over are ways to find where the model balances false-positive and false-negative errors. This is the point where the model's prediction is just as likely to be right about correctly predicting a site as it is correctly predicting a background cell. The metric of Sens=Spec is calculated from the ROC curve to find the threshold at which those type measures are about equal. The X-Over is included here because it has been traditionally cited in APM literature as the optimal location to define a threshold (Kvamme 1988). The discussion of thresholds in Chapter 4 explains how the cross-over point, traditionally derived from a graphic, can be calculated quantitatively. It is easy to see from the tables below that these two metrics are nearly the same for

all of the models. This is because they are indeed measuring the same qualities of the predicted probability distribution. The cross-over statistic and discussion are included in this report because the idea of a point where two lines cross being an optimal balance is an intuitive notion and because it is a method used in previous APM studies. For future studies, it is suggested that the idea of sensitivity equaling specificity, common throughout other fields of study, replace the graphically oriented cross-over threshold.

Semantics aside, the error-balancing threshold approach is a logical and interpretable threshold selection method, but it has a drawback that is important to the problems that APM addresses. The point at which false-negative and false-positive errors are balanced assumes that these two types of errors are equally weighted (Fielding and Bell 1997). In archaeology, it is generally held that a false-negative is more costly, in a general sense, than a false-positive. This is because finding a site where not expected (construction discovery) is more costly than not finding a site where expected (negative Phase I survey). Furthermore, aside from that, an archaeological model requires a sufficiently large number of false-positive cells/regions in order to predict where undiscovered sites may be located. Essentially, all of the area that is in high or moderate sensitivity zones and that does *not* contain a known archaeological site is a false-positive. Without false-positive, we would only be predicting where known sites are. Therefore, the weight of a false-positive is negligible compared to the cost of a false-negative. This is not to say that the Sens=Spec threshold is always a poor choice; false-positives are still included within these predictions. For an easy to interpret and justifiable threshold, it may be a good choice. However, the next two sets of threshold types can help address the associated costs of misclassification.

The third group of threshold selection methods presented here, Sens @ 0.85 and Spec @ 0.67, are labeled as "Domain Specific" thresholds because these allow for the specification of sensitivity or specificity based on an arbitrary value established for a specific purpose. As stated by Freeman and Moisen (2008:57), "[f]or particular management applications the special cases of user specified required accuracy [threshold methods] (ReqSens and ReqSpec) may be most appropriate." Freeman and Moisen follow this with examples that can be easily translated into archaeological examples. If the goal of a project is to find which areas are most sensitive to site locations and therefore the most threatened, the model should avoid over-predicting true-absence as predicted presence. In this case the user required specificity would be relatively high, such as 0.95 to signify that no more than 5% of the true-negative observations (background cells) should be misclassified as site-likely. Alternately, the specificity could be used to set a lower end to the site-likely area, such as the case with the Spec @ 0.67 threshold presented here. In this case a specificity of 0.67 assures that no more than 33% of the true-negative observations (background cells) are classified as site-likely. This specificity was used here because of the Mn model's goals of maximizing the percent-sites within 33% of the study area. This is as close as a domain specific value for specificity as has been proposed for APM.

On the other hand, a threshold at an established specificity can be established. Translating Freeman and Moisen's (2008) example, if the goal of a model is to narrow the search area for a pre-stratified

survey (e.g., high, moderate, and low) then the threshold should be set to make sure to look at the full range of landforms that may contain sites. In this case, the user-required sensitivity would be relatively high, such as 0.95 to signify that no more than 5% of the true-positive observations (site-present cells) will be misclassified as site-unlikely. Note that this is the mirror image of the previous example of requiring a certain level of specificity where the threshold is based on managing true-negative misclassification, as opposed to here where sensitivity manages true-positive misclassification. In the tables below, the threshold for required sensitivity is set to 0.85. This assures that the site-likely area misclassifies no more than 15% of the known site-present cells. A sensitivity of 0.85 is also used in response to the Mn model example. Oehlert and Shea (2007:13) summarize their findings as follows: "[h]ere we recommend choosing rules that maximize specificity for 85% sensitivity." And that is what this threshold does: minimizes the area needed to correctly predict 85% of the known site cells. However, like the maximizing thresholds, Oehlert and Shea's recommendation still seeks to minimize area and maximize percent site-present, but at least puts a number—85%—to the equation. What is missing from this equation is a number for how far the site-likely area should be minimized.

The final two thresholds, Pred=Obs @ 0.1 and Pred=Obs @ 0.2, are labeled as "Prevalence Based" because they account for the prevalence of positive observations (sites) to adjust the threshold values. As previously stated, the low prevalence of archaeological sites across the landscape poses an obstacle to the modeling effort. This is because the data being modeled are heavily imbalanced toward the negative observation (site not-present cells), and most models will favor predictions for the larger of the two classes. A number of methods are used throughout the modeling process to combat this bias, and threshold selection can be another avenue. The Pred=Obs threshold selection methods set the threshold so that the predicted site-likely cell prevalence will equal the observed prevalence. The observed prevalence can either be calculated from the data set or assigned arbitrarily. For comparative purposes, two arbitrary prevalence values are used to adjust the threshold: 0.1 and 0.2. Throughout Regions 1, 2, and 3, the overall prevalence of known archaeological sites with a prehistoric component is 0.0016. Riverine subareas have an overall prevalence of 0.0083 and upland subareas have an overall prevalence of 0.0010. Figure 31 shows the prevalence of all subareas within Regions 1, 2, and 3. The lowest prevalence is within Region 1 – East Upland Section 1 at 0.00017 and the highest is within Region 1 West Riverine Section 2 at 0.0147. By setting the threshold for the site-likely area at 0.1, the threshold is compensating for survey and detection bias. Clearly, the density of archaeological sites varies widely throughout the state, but it is also clear that this is to some degree a function of survey bias. Establishing a baseline prevalence for site-likely predictions creates a basis for interpretation and consistency, much like Sens @ 0.85 and Spec @ 0.67.

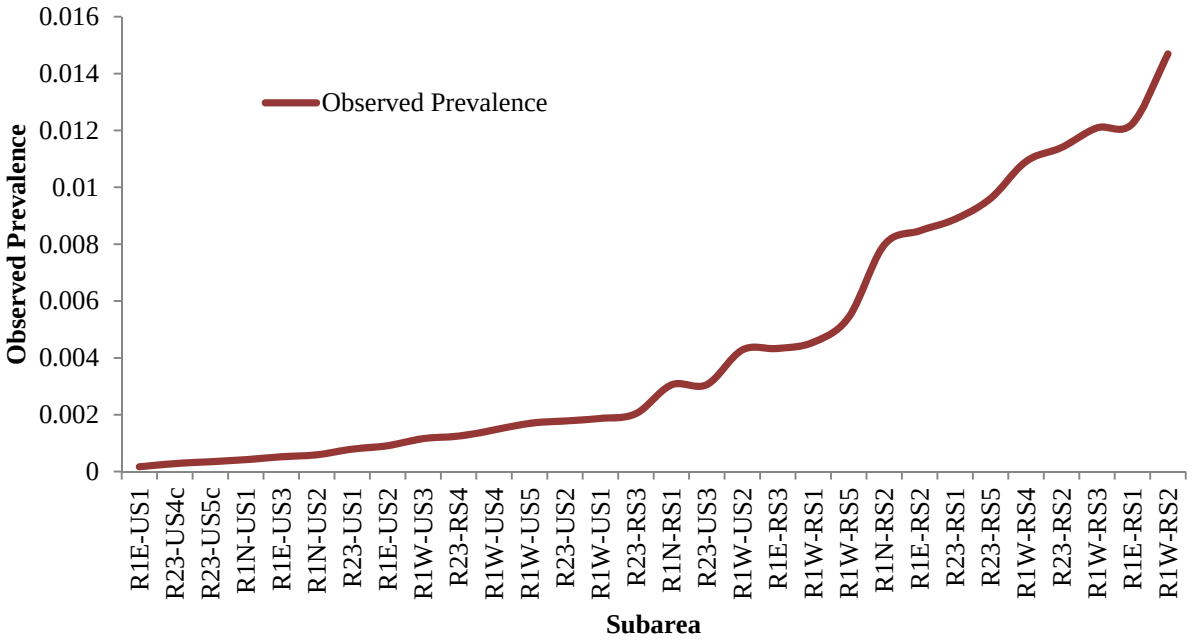


Figure 31 - Average prevalence of prehistoric sites by subarea.

The choice of appropriate thresholds for model prediction is driven by project needs and management goals. The threshold selection methods and thresholds discussed above are all appropriate for these models, depending on how they are to be used; ranging maximized thresholds are the most conservative, the cross-over thresholds are the most balanced, and the prevalence thresholds are the most liberal. Any one of these approaches could be effective given the problem at hand, but approaches such as the requirements of sensitivity or specificity, and prevalence-based threshold are likely the most applicable to APM. Freeman and Moisen (2008:57) came to the same conclusion based on studies in ecological modeling, which shares many of the same obstacles and goals as APM. Additionally, Freeman and Moisen conclude that no one set of thresholds or the resulting map can fulfill all of the objectives for which a model could be used, and that essentially the model should be viewed as a tool that needs to be adapted to a specific task through the use of thresholds. They state that, "[u]ltimately, maps will typically have multiple and sometimes conflicting management applications and thus providing users with a continuous probability surface may be the most versatile method … allowing threshold choice to be matched up with map use" Freeman and Moisen (2008:57).

SELECTED MODEL THRESHOLDS

This project supports Freeman and Moisen's conclusion and will provide the continuous probability distribution maps as a part of the final deliverable. However, this project also recognizes that with the insight gained through this analysis, a recommended set of thresholds should be provided and maps based on these thresholds should be created.

The thresholds selected for this project are based on both the required specificity and prevalence methods. The threshold for high sensitivity sets the predicted site-likely prevalence to 0.1. This threshold assumes that there is a large portion of the archaeological record that has not yet been discovered in each subarea. The true prevalence of archaeological sites in a region would be very difficult to estimate, especially in a region where very few sites are easily detected from surface survey (as opposed to arid desert regions with many sites on the surface). However, a prevalence target of 0.1 is well higher than the highest observed prevalence and incorporates approximately 9%–11% of the subarea for each model.

The threshold for the low end of moderate probability, and therefore the low end of the site-likely area, is set at a specificity target of 0.67. This assures that no more than 33% of the true-negative observations (background cells) are classified as site-likely. In essence, this sets the site-likely area at close to 33% of the total subarea. This threshold is used in response to the Mn model goal of maximizing site-present locations within 33% of the study area (Mn/Model n.d.). As discussed earlier, the recommendation by Oehlert and Shea (2007) of requiring a sensitivity of 0.85 and minimizing specificity is not very useful here because it does not set a lower bound on specificity. The implementation of the specificity at a 0.67 threshold used here establishes a lower bound (at 0.67) and takes a more conservative approach than suggested by Oehlert and Shea.

On balance, the use of these two threshold measures creates a standardized set of high, moderate, and low classifications across the three regions. As evident in Table 43, Table 44, Table 45, and Table 46, the combined site-likely area of high and moderate probability always includes 100% of the known sites in an average area of approximately 31% of the study area, with an average Kg of 0.69. The confusion matrices for each of the models, classified as site-likely (high and moderate sensitivity) and site-unlikely (low sensitivity), are presented in Appendix F. The overall confusion matrix representing the site-likely classification for the entirety of Regions 1, 2, and 3 is presented in Table 47. Figure 32 depicts an overview of high, moderate, and low sensitivity for the entirety of Regions 1, 2, and 3. These data will be provided as ESRI raster grids for detailed viewing and analysis.

Table 43 - Kg and Cell Percentages at Suggested Final Thresholds, Region 1 East

Region 1 - East								
Classification	Pred=Obs @ 0.1, High Sensitivity				Specificity @ 0.67, Moderate Sensitivity			
Subarea	Threshold	% background	% sites	Kg	Threshold	% background	% sites	Kg
riverine_section_1	0.28	9.00	100.00	0.910	0.10	33.55	100.00	0.665
riverine_section_2	0.23	9.45	100.00	0.906	0.09	32.58	100.00	0.674
riverine_section_3	0.26	10.00	100.00	0.900	0.11	29.05	100.00	0.710
upland_section_1	0.18	10.15	100.00	0.899	0.08	26.84	100.00	0.732
upland_section_2	0.15	9.59	100.00	0.904	0.08	24.64	100.00	0.754
upland_section_3	0.22	9.85	100.00	0.902	0.13	27.94	100.00	0.721

Table 44 - Kg and Cell Percentages at Suggested Final Thresholds, Region 1 North

Region 1 - North								
Classification	Pred=Obs @ 0.1, High Sensitivity				Specificity @ 0.67, Moderate Sensitivity			
Subarea	Threshold	% background	% sites	Kg	Threshold	% background	% sites	Kg
riverine_section_1	0.29	10.12	100.00	0.899	0.11	33.27	100.00	0.667
riverine_section_2	0.47	9.14	100.00	0.909	0.22	32.92	100.00	0.671
upland_section_1	0.19	10.44	100.00	0.896	0.08	32.24	100.00	0.678
upland_section_2	0.23	9.84	100.00	0.902	0.13	28.63	100.00	0.714

Table 45 - Kg and Cell Percentages at Suggested Final Thresholds, Region 1 West

Region 1 - West								
Classification	Pred=Obs @ 0.1, High Sensitivity				Specificity @ 0.67, Moderate Sensitivity			
Subarea	Threshold	% background	% sites	Kg	Threshold	% background	% sites	Kg
riverine_section_1	0.37	9.79	100.00	0.902	0.17	32.86	100.00	0.671
riverine_section_2	0.47	8.90	100.00	0.911	0.20	33.07	100.00	0.669
riverine_section_3	0.27	8.76	100.00	0.912	0.09	31.60	100.00	0.684
riverine_section_4	0.43	8.97	100.00	0.910	0.18	31.77	100.00	0.682
riverine_section_5	0.35	9.70	100.00	0.903	0.18	32.40	100.00	0.676
upland_section_1	0.41	9.82	100.00	0.902	0.16	33.73	100.00	0.663
upland_section_2	0.48	9.51	100.00	0.905	0.24	31.87	100.00	0.681
upland_section_3	0.22	9.56	100.00	0.904	0.11	33.40	100.00	0.666
upland_section_4	0.27	10.39	100.00	0.896	0.15	32.02	100.00	0.680
upland_section_5	0.34	9.84	100.00	0.902	0.19	31.80	100.00	0.682

Table 46 - Kg and Cell Percentages at Suggested Final Thresholds, Region 2/3

Region 2/3								
Classification	Pred=Obs @ 0.1, High Sensitivity				Specificity @ 0.67, Moderate Sensitivity			
Subarea	Threshold	% background	% sites	Kg	Threshold	% background	% sites	Kg
riverine_section_1	0.24	9.54	100.00	0.905	0.12	31.18	100.00	0.688
riverine_section_2	0.28	8.91	100.00	0.911	0.10	31.00	100.00	0.690
riverine_section_3	0.22	9.81	100.00	0.902	0.12	31.98	100.00	0.680
riverine_section_4	0.34	9.72	100.00	0.903	0.15	31.43	100.00	0.686
riverine_section_5	0.27	9.20	100.00	0.908	0.10	32.73	100.00	0.673
upland_section_1	0.17	10.59	100.00	0.894	0.08	30.04	100.00	0.700
upland_section_2	0.21	10.45	100.00	0.896	0.13	30.37	100.00	0.696
upland_section_3	0.19	10.30	100.00	0.897	0.09	30.50	100.00	0.695
upland_section_4c*	0.38	10.55	100.00	0.895	0.24	31.16	100.00	0.688
upland_section_5c*	0.39	10.67	100.00	0.893	0.23	34.31	100.00	0.657

* combined rock shelter and non-rock shelter specific models

Table 47 - Confusion Matrix for Site-Likely area of Complete Regions 1, 2, and 3 Model

Complete Model

		Known Sites		
		Present	Absent	
Model Prediction	Present	678154	127533633	128211787
	Absent	0	288725095	288725095
		678154	416258728	416936882

Sensitivity / TPR =	1.000
Specificity / TNR =	0.694
Prevalence =	0.0016
Kvamme Gain (Kg) =	0.692
Accuracy =	0.694
Positive Prediction Value (PPV) =	0.005
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.002
Positive Prediction Gain (PPG) =	3.252
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.308

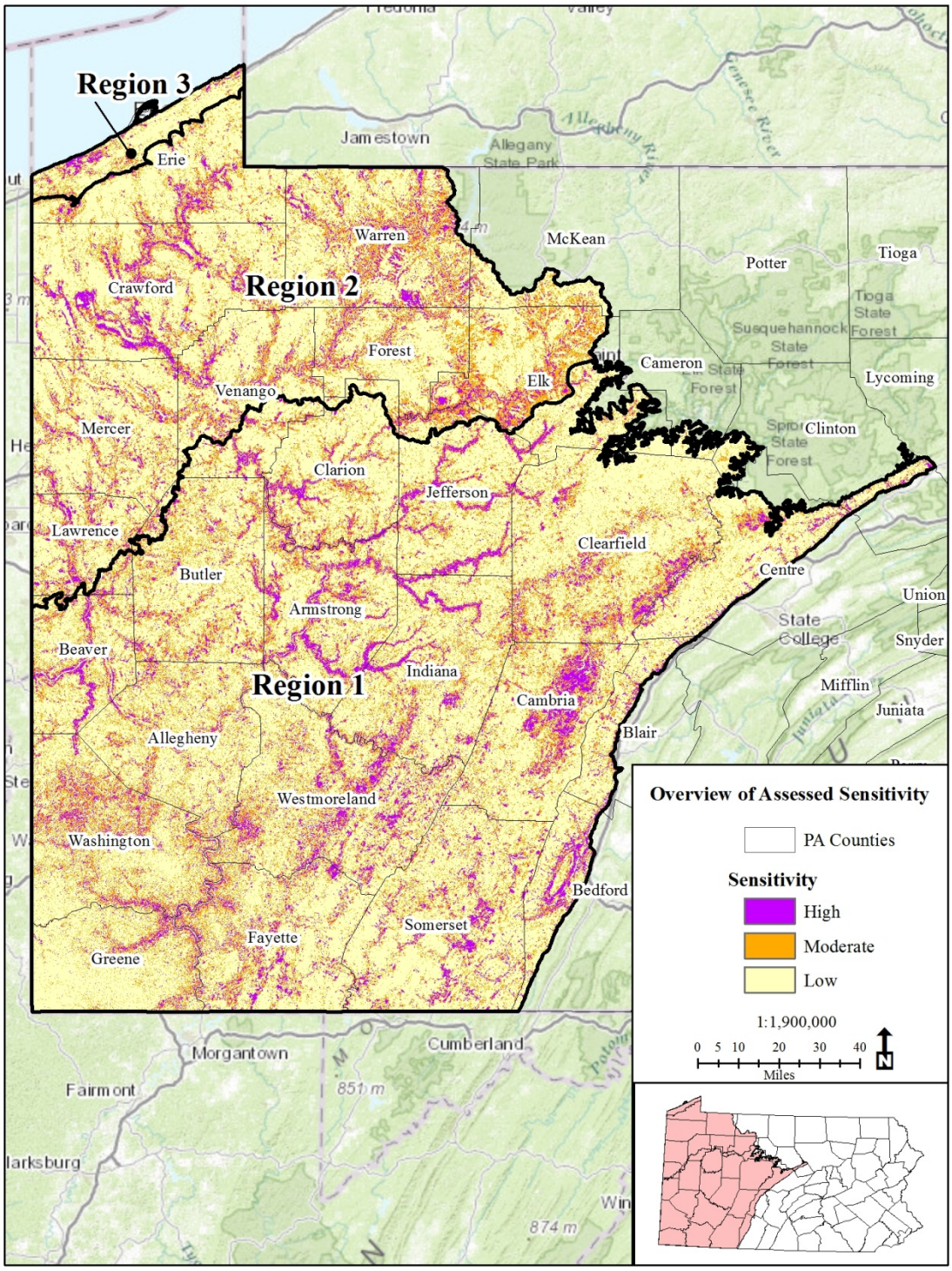


Figure 32 - Overview of assessed prehistoric sensitivity for Regions 1, 2, and 3.

7
CONCLUSIONS AND RECOMMENDATIONS

The previous chapters describe the creation of 30 individual archaeological predictive models for Regions 1, 2, and 3, comprising much of western Pennsylvania. The methodology used to create these models involved the preparation of PASS site data, the development of 89 individual environmental variables, the division of the regions into 30 separate subareas, the testing of each of these variables against the environmental background of each subarea, the parameterization and validation of a logistic regression, adaptive regression splines, and random forest models for each subarea, final model selection based on error estimate results, the establishment of numerous potential thresholds based on variable criteria, and, finally, the application of selected thresholds and mosaicking of 30 separate subarea models into the final model for each region. The end result is a model of all regions that correctly classifies all archaeological sites within 30.8 percent of the study area, for a Kg of 0.692. In actuality, the model is capable of correctly predicting the location of all archaeological sites and minimizing the site-likely area to on average 5 percent of the study area, but the selection of a low end threshold for the site-likely area was intentionally set to approximately 33 percent of the study area. Compared to a random survey, the chances of finding a site in the combined high and moderate sensitivity area are 3.252 times greater.

The 30 subarea models created for Regions 1, 2, and 3 are all derived from the random forest statistical model and have a high degree of accuracy in discriminating known site locations from the background. The results of the prediction error rate tests (average RMSE = 0.124) on the 10-fold CV samples demonstrate that these models are capable of accurately predicting site-present cells that were not part of the model-building sample. This adds confidence that these models are not only able to identify landforms that the test sites are found on, but also extrapolate this pattern to site locations outside of the test set. The suite of validation and testing statistics presented in the previous chapters all agree that these models are a good presentation of the site sample from previously identified prehistoric archaeological sites. Further, these models better approximate a more realistic prevalence of prehistoric sites than previous and more generalized models. With the choice of classification thresholds that are appropriate for the particular management or research objective, these models should be valid and accurate tools to assist in project planning and sensitivity analysis.

Most of the recommendations from the Task 3 report were successfully incorporated into this study, including: a greater focus on statistical models over judgmental models, a larger and standardized body of predictor variables, revision of the resampling methods to better account for low prevalence, and incorporation of more and more varied predictor variables for consideration within statistical models. Addressing these recommendations and developing new means to handle large data was critical to the success of implementing the pilot model methodology onto an area nearly five times as large. To continually increase the rate of efficiency, model accuracy, metric analysis, interpretation, and overall model applicability, additional recommendations are provided here.

1. Inclusion of aggregated soil characteristic data as predictor variables.
    - The use of USDA soils data is a relatively common and intuitive predictive variable used within many archaeological predictive models, because the vast majority of archaeological material in the Commonwealth is found within soils and archaeologists are trained to understand and interpret soils. However, the incorporation of soils data into the modeling process has a number of methodological considerations, as well as theoretical considerations. In an effort to utilize this potentially important data source, the incorporation of USDA soils data should be tested and possibly utilized in the regional models to follow.
2. Continue to develop rock shelter and open-site specific models, where appropriate.
    - In the pilot model, rockshelter sites were included within the site sample, but steep slopes were excluded from the final model sensitivity layers. Following the pilot model, and through group discussions, it was decided to attempt to create final sensitivity models that can predict for rockshelters type sites on steep slopes. This discussion resulted in the combined rockshelter and open-site sensitivity models of Region 2/3 upland sections 4 and 5.
3. Continue to increase model-building efficiency and ability to assess validity.
    - The ability to efficiently and accurately process and model the large amounts of data presented in this report is paramount to achieving this project. The iteration of the model-building process utilized in this task is much more efficient than the version built for the pilot model, based in a large part on the recommendations made in the Task 3 report. The next iteration of the modeling methodology should continue to incorporate new and faster modeling techniques, as well as new routines to effectively use available site data to validate predictions and provide insight into model behavior. Creating faster models can be achieved through removing redundancies in the code base, efficient storage of large volumes of data, and increased use of parallel processing. More effective use of the available site data can be achieved through better partitioning of training, testing, and validation data sets, background sample bootstrapping, and variable selection through cross-validation. These elements will be tested in the upcoming modeling tasks.
4. Test and incorporate use of class weighting and cost thresholds in future models.
    - The use of weighting false-positive versus false-negative error rates was discussed throughout this report in the context of threshold selection methods. The use of class weighting can also be incorporated into the RF and adaptive regression spline statistical models. The next set of models should continue to develop the use of relative costs in threshold selections, as well as explore the adaptation of class weights as a means to manage both class imbalance and relative error weights.

8
REFERENCES CITED

Adovasio, J.M., R. Fryman, A.G. Quinn, and D.R. Pedlar
2003 The Appearance of Cultigens and the Early and Middle Woodland Periods in Southwestern Pennsylvania. In *Foragers and Farmers of the Early and Middle Woodland Periods in Pennsylvania*, edited by P. Raber and V. Cowin, pp. 67-84. Pennsylvania Historical and Museum Commission, Recent Research in Pennsylvania Archaeology No. 3. Harrisburg.

Asch, David, and Nancy Asch
1985 Prehistoric Plant Cultivation in West-Central Illinois. In *Prehistoric Food Production in North America*, edited by R. Ford, pp. 149–203. University of Michigan, Museum of Anthropology Anthropological Papers No. 75, Ann Arbor.

Baublitz, Richard T., Charles A. Richmond, and Barbara J. Shaffer
2003 Archaeological Predictive Model, S.R. 0830, Section 590, DuBois-Jefferson County Airport Access Project, Jefferson and Clearfield Counties, Pennsylvania.ER# 93-0231-065. Reported by McCormick, Taylor, and Associates, Inc., Pittsburgh, Pennsylvania, to Pennsylvania Department of Transportation, Distric 10-0, Indiana, Pennsylvania.

Boyd, Varna G., Gary F. Coppock, Kathleen A. Ferguson, Benjamin R. Fischler, Bernard K. Means, and Frank Vento
2000 Prehistoric Archaeological Synthesis: The U.S. 219 Meyersdale Bypass Project. U.S. 219 Meyersdale Bypass Project, S.R. 6219, Section B08, Somerset County, Pennsylvania. Reported by Greenhorne & O'Mara, Inc., Mechanicsburg, Pennsylvania, to Pennsylvania Department of Transportation, Engineering District 9-0, Hollidaysburg.

Brose, David S.
2000 Late Prehistoric Societies of Northeastern Ohio and Adjacent Portions of the South Shore Of Lake Erie: A Review. In *Cultures Before Contact: The Late Prehistory of Ohio and Surrounding Regions*, edited by Robert A. Genheimer, pp. 96–123, The Ohio Archaeological Council, Columbus.

Caldwell, Joseph R.
1964 Interaction Spheres in Prehistory. In *Hopewellian Studies*, edited by J. R. Caldwell and R. L. Hall, pp. 133–144. Illinois State Museum Scientific Papers Volume 12. Illinois State Museum, Springfield.

Carr, Kurt W.
1998 Archaeological Site Distribution and Patterns of Lithic Utilization During the Middle Archaic in Pennsylvania. In *The Archaic Period in Pennsylvania: Hunter-Gatherers of the Early and Middle Holocene Period,* edited by Paul A. Raber, Patricia E. Miller, and Sarah M. Neusius, pp. 77–90. Pennsylvania Historical and Museum Commission, Harrisburg.

Carr, Kurt W., and James M. Adovasio
2002 Paleoindians in Pennsylvania. In Ice Age Peoples of Pennsylvania, Kurt Carr and James Adovasio, editors, pp. 1–50. Pennsylvania Historical and Museum Commission, Recent Research in Pennsylvania Archaeology No. 2. Harrisburg.

Chiarulli, Beverly A.
2001 Prehistoric Settlements Patterns in Upland Settings: An Analysis of Site Data in Watershed D (Conemaugh River – Blacklick Creek). Prepared for the Pennsylvania Historical and Museum Commission under a Historic Preservation Grant awarded to the Pennsylvania Archaeological Council. http://home.earthlink.net/~pacweb/spframeset.html. Accessed 10 January 2014.

Cohen, Jacob
1960 A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20(1):37–46.

Coppock, Gary F., and Scott D. Heberling
2001 Interim Report, Predictive Model for Archaeological Resources, US 219 Improvements Project, SR0219, Seciton 020, Somerset County, Pennsylvania. ER# 01-8012-111. Reported by Heberling and Associates, Inc., Huntingdon, Pennsylvania, to Pennsylvania Department of Transportation, Engineering District 9-0, Hollidaysburg, Pennsylvania, and Federal Highway Administration, Washington, D.C.

Coppock, Gary F., Scott D. Heberling, David A. Krilov, and Ronan A. Carthy
2003 Predictive Model for Archaeological Resources and Phase I Archaeology Work Plan, U.S. 219 Improvements Project, Meyersdale to I-68, Somerset County, Pennsylvania, and Garrett County, Maryland. ER# 2002-0842-111. Reported by Heberling and Associates, Inc., Huntingdon, Pennsylvania, in association with McCormick, Taylor, and Associates, Inc, Philadelphia, to Pennsylvania Department of Transportation, Engineering District 9-0, Hollidaysburg, Pennsylvania, Maryland State Highway Administration, Baltimore, and Federal Highway Administration, Washington, D.C.

Cowin, Verna L.
1981 1980-1981 Archaeological Survey in Region VII, West Central Pennsylvania. E.R. # 1981-R003-042. Reported by Carnegie Museum of Natural History, Pittsburgh, Pennsylvania, to Pennsylvania Historical and Museum Commission, Harrisburg, Pennsylvania.

Custer, Jay F.
1996 *Prehistoric Cultures of Eastern Pennsylvania*. Pennsylvania Historical and Museum Commission, Anthropological Series No. 7. Harrisburg.

Davis, Christine E., Curtis L. Biondich, and David A. Roth
2004 Phase I Archaeology Survey and Phase II Archaeology Surveys for the Latimer Farm Site (36LR60) and the Abraham Number 6 Site (36LR266), Millennium Park, Neshannock Township, Lawrence County, Pennsylvania. E.R. # 2003-1727-073-N. Reported by Christine Davis Consultants, Inc., Verona, Pennsylvania, to Pennsylvania Historical and Museum Commission, Bureau of Historic Preservation, Harrisburg.

Dragoo, Don W.
1976 Some Aspects of Eastern North American Prehistory: A Review, 1975. *American Antiquity* 41(1):3–27.

1989 [1963] *Mounds for the Dead: An Analysis of the Adena Culture*. The Carnegie Museum of Natural History, Pittsburgh, Pennsylvania.

Duncan, Richard B.
2002 Centre and Clearfield Counties, Pennsylvania, S.R. 0322, Section B02, Corridor O Project, Phase IA Archaeological Investigations and Archaeological Predictive Model, Executive Summary.ER# 1999-2755-033. Reported by Skelly and Loy, Inc., Monroeville, Pennsylvania, to Pennsylvania Department of Transportation, District 2-0, Clearfield, Pennsylvania.

Duncan, Richard B., and Brian F. Schilling
1999 Fayette and Washington Counties, Mon/Fayette Expressway Project, Uniontown to Brownsville, Archaeological Predictive Model Development. ER# 87-1002-042. Reported by Skelly and Loy, Inc., Monroeville, Pennsylvania, to Pennsylvania Turnpike Commission, Harrisburg.

Duncan, Richard B., Thomas C. East, and Kristen A. Beckman, Ph.D
1996 Allegheny and Washington Counties, Mon/Fayette Transportation Project, Interstate 70 to Route 51, Evaluation of the Crooked Creek Predictive Model. ER# 87-1002-024-A02 and A03. Reported by Skelly and Loy, Inc., Monroeville, Pennsylvania, to Pennsylvania Turnpike Commission, Harrisburg.

Fielding, Alan H. and John F. Bell
1997 A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*. 42(1):38-49.

Freeman, Elizabeth A. and Gretchen G. Moisen
2008 A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modeling*. 217:48-58.

Glenn, Jonathon
2010 Archaeological Overview and Sensitivity Models Erie National Wildlife Refuge Crawford County, Pennsylvania. E.R. 2012-1218-042-A. Prepared for U.S. Fish and Wildlife Service, Region 5 Hadley, Massachusetts. GAI Consultants, Inc., Homestead, Pennsylvania.

Griffin, James B.
1967 Eastern North American Archaeology: A Summary. *Science* 156(3772):175–191.

Harris, Matthew D.
2013a Pennsylvania Department of Transportation Archaeological Predictive Model Set, Task 1: Literature Review. Prepared for Pennsylvania Department of Transportation, Bureau of Planning and Research, Harrisburg. URS Corporation, Burlington, New Jersey.

2013b Pennsylvania Department of Transportation Archaeological Predictive Model Set, Task 2: Designating Modeling Regions. Prepared for Pennsylvania Department of Transportation, Bureau of Planning and Research, Harrisburg. URS Corporation, Burlington, New Jersey.

2014 Pennsylvania Department of Transportation Archaeological Predictive Model Set, Task 3: Pilot Model Study. Prepared for Pennsylvania Department of Transportation, Bureau of Planning and Research, Harrisburg. URS Corporation, Burlington, New Jersey.

Hart, John
1994 Development of Predictive Models of Prehistoric Archaeological Site Location for the Lake Erie Plain and Glacial Escarpment in the Erie East Side Access Project Area, Erie County, Pennsylvania. E.R. 1992-0858-049-E. Prepared for the Pennsylvania Department of Transportation, Harrisburg, Pennsylvania. GAI Consultants, Monroeville, Pennsylvania.

Herbstritt, James T.
1981 Prehistoric Archaeological Site Survey in Pennsylvania Region II, Southwestern Pennsylvania, 1989. E.R. # 1981-R002-51. Reported by Center for Prehistoric and Historic Site Archaeology, California State College, California, Pennyslyvania, to Pennsylvania Historical and Museum Commission, Harrisburg.

Hobbs, Elizabeth
1997 Mn/Model: An Archaeological Predictive Model for Minnesota. Proceedings, ESRI User Conference, San Diego.

Jefferies, Richard W.
1990 Archaic Period. In *The Archaeology of Kentucky: Past Accomplishments and Future Directions*, edited by D. Pollack, pp. 143–246. Kentucky Heritage Council, Frankfort.

1996 Hunters and Gatherers after the Ice Age. In *Kentucky Archaeology*, edited by R. Barry Lewis, pp. 39–78. University of Kentucky Press, Lexington.

Jeni, Laszlo, Jeffrey F. Cohn, and Fernando De la Torre
2013 Facing Imbalanced Data Recommendations for the use of Performance Metrics. Proceedings of the International Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland.

Johnson, William C.
1981 Archaeological Review Activities in Survey Region IV, Northwestern Pennsylvania: Year End Report of the Regional Archaeologist for the Period September 1980 through August 1981. Report prepared for Pennsylvania Historical and Museum Commission. Cultural Resources Management Program, University of Pittsburgh.

Justice, Noel D.
1995 *Stone Age Spear and Projectile Points of the Midcontinental and Eastern United States*. Indiana University Press, Bloomington.

Katz, Gregory M., Joh P. Branigan, Paul W. Schopp, and Steven J. Blondo
2002 Reconnaissance Survey/Predictive Model Report, S.R. 0228, Section 290, Cranberry, Adams, and Middlesex Townships, Butler County, Pennsylvania, and Marshall, Pine, and Richland Townships, Allegheny County, Pennsylvania, Volume I. Reported by A. D. Marble and Company, Conshohocken, Pennsylvania, to Pennsylvania Department of Transportation, Engineering District 10-0, Indiana, Pennsylvania.

Koetje, Todd
1998 The Archaic in Northwestern Pennsylvania: A View from 36ME105, Mercer County. In *The Archaic Period in Pennsylvania: Hunter-Gatherers of the Early and Middle Holocene Period*, edited by Paul A. Raber, Patricia E. Miller, and Sarah M. Neusius, pp. 29–44. Pennsylvania Historical and Museum Commission, Harrisburg.

Kuhn, Max and Kjell Johnson
2013 *Applied Predictive Modeling*. Springer, New York.

Kvamme, Kenneth L.
1988 Development and Testing of Quantitative Models. In *Quantifying the Present and Predicting the Past*, edited by W. Judge and L. Sebastian, pp. 325-428. U.S. Government Printing Office, Washington, D.C.

Landis, J.R., and G.G. Koch
1977 The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1):159–174.

Liu, Canran, Berry, Pam M., Dawson, Terence P. and Richard G. Pearson
2005 Selecting Thresholds of Occurrence in the Prediction of Species Distributions. *Ecography*, 28(3): 385-393.

Means, Bernard K.
2008 Resurrecting a Forgotten Monongahela Tradition Village: The Phillips (36FA22) Site . *Journal of Middle Atlantic Archaeology* 24:1–12.

Means, Bernard K., Benjamin R. Fischler, Elizabeth Moore, Susannah Dean Leslie Raymer, Richard Fuss, Antonio V. Segovia, and Daniel P. Wagner
1998 Volume VII: Interpretations and Special Analyses Prepared for Phase I and Phase II Archaeological Investigations of the U.S. 219 Meyersdale Bypass Project. ER# 92-0237-0111. Reported by Greenhorne & O'Mara, Inc., Mechanicsburg, Pennsylvania, to Pennsylvania Department of Transportation, District 9-0, Hollidaysburg, Pennsylvania.

Metz, Charles E.
1978 Basic Principles of ROC analysis. *Seminars in Nuclear Medicine*. Vol. III, No. 4.

Meyers, Andrew J., and Malinda Moses Meyers
2014 Paleoindian Research in Western Pennsylvania: A Preliminary Report on the Paleoindian Assemblage From Indian Camp Run No. 1 (36FO65). Allegheny Archaeological Research, Brockway, Pennsylvania. http://www.orgsites.com/pa/alleghenyarchaeology/_pgg10.php3

Michael, Ronald L., and James T. Herbstritt
1980 Archaeological Reconnaissance of North Sewickley Township, Beaver County, South Project (with Eastvale Borough) Sewage Project. E.R. # 1981-0739-007-A. Reported by New Consultants, Inc.. Uniontown, Pennsylvania, to Duncan, Lagnese, and Associates, Inc., Pittsburgh, Pennsylvania.

Milanich, Jerald T.
1994 *Archaeology of Precolumbian Florida*. University Press of Florida, Gainesville.

Mn/Model
n.d. Project Background.Electronic document: http://www.dot.state.mn.us/mnmodel/about/history.html, Accessed May 8, 2014.

Oehlert, Gary W., and Brian Shea
2007 Statistical Methods for Mn/Model Phase 4. Research Services Section of Minnesota Department of Transportation, St. Paul.

Prufer, Olaf H., and Dana A. Long
1986 *The Archaic of Northeastern Ohio*. Kent State University Press, Kent, Ohio.

Purtill, Matthew P.
2009 The Ohio Archaic: A Review. In Archaic Societies: Diversity and Complexity Across the Midcontinent, edited by Thomas E. Emerson, Dale L. McElrath, and Andrew C. Fortier, pp. 565–606. State University of New York Press, Albany, New York.

Quinn, Allen G. , with contributions by Judith E. Thomas and David C. Hyland
1994 Phase I Archaeological reconnaissance of the Shades Beach Park Study Area: A Report of the Pennsylvania Department of Environmental Resources to the National Oceanic and Atmospheric Administration Pursuant to NOAA Award No. NA370Z0351. Report to Harborcreek Township, Harborcreek, Pennsylvania, from Mercyhurst Archaeological Institute, Erie, PA. U.S. Government Printing Office <http://www.gpo.gov/fdsys/pkg/CZIC-qh76-5-h3-q56-1994/html/CZIC-qh76-5-h3-q56-1994.htm>. Accessed 3 January 2014.

Raber, Paul A., Patricia E. Miller, and Sarah M. Neusius
1998 The Archaic Period in Pennsylvania: Current Models and Future Directions. In The Archaic Period in Pennsylvania: Hunter-Gatherers of the Early and Middle Holocene Period, edited by Paul A. Raber, Patricia E. Miller, and Sarah M. Neusius, pp. 121–137, Pennsylvania Historical and Museum Commission, Harrisburg.

Sassaman, Kenneth E.
1993 *Early Pottery in the Southeast: Tradition and Innovation in Cooking Technology*. The University of Alabama Press, Tuscaloosa.

Raber, Paul A.,
2003 Problems and Prospects in the Study of the Early and Middle Woodland Periods. In *Foragers and Farmers of Early and Middle Woodland Periods in Pennsylvania*, edited by Paul A. Raber and Verna L. Cowin, pp. v–vii. Pennsylvania Historical and Museum Commission, Recent Research in Pennsylvania Archaeology No. 3. Harrisburg.

Stafford, C. Russell
1994 Structural Changes in Archaic Landscape Use in the Dissected Uplands of Southwestern Indiana. *American Antiquity* 59(2):219–237.

Stewart, R. Michael
2003 A Regional Perspective on Early and Middle Woodland Prehistory in Pennsylvania. In *Foragers and Farmers of Early and Middle Woodland Periods in Pennsylvania*, edited by Paul A. Raber and Verna L. Cowin, pp. 1–33. Pennsylvania Historical and Museum Commission, Recent Research in Pennsylvania Archaeology No. 3. Harrisburg.

Viera, Anthony J., and Joanne M. Garrett
2005 Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine* 37(5):360-363.

Wallace, Paul A.W.
1965 *Indian Paths of Pennsylvania*. Pennsylvania Historical and Museum Commission, Harrisburg.

Weed, Carol S.
2004 Prehistoric Context Study (Chapter Three) in Support of Data Recovery at Site 36AL480, Leetsdale, Allegheny County, Pennsylvania (Contract No. DACW-69-98-D-0027, Task Order DV01). Report to U.S. Army Corps of Engineers, Pittsburgh District, PA, from David Miller & Associates, Vienna, Virginia.

Yerkes, Richard W.
1988 The Woodland and Mississippian Traditions in the Prehistory of Midwestern North America. *Journal of World Prehistory* 2(3):307–358.

APPENDIX A

ACRONYMS AND GLOSSARY OF TERMS

ACRONYMS

APM	Archaeological Predictive Modeling
AUC	Area Under Curve
CoV	Coefficient of Variation
CRGIS	Cultural Resources Geographic Information System
CV	Cross-Validation
GIS	Geographic Information Systems
Kg	Kvamme Gain
K-S	Kolmogorov–Smirnov
LR	Logistic Regression
MARS	Multivariate Adaptive Regression Splines
MW	Mann-Whitney
NPG	Negative Prediction Gain
NPV	Negative Prediction Value
OOB	Out-of-bag Sample
PASS	Pennsylvania Archaeological Site Survey
PPG	Positive Predictive Gain
PPV	Positive Prediction Value
RF	Random Forests/randomForest
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristics
TNR	True-Negative Rate
TPR	True-Positive Rate
UDR	Unexpected Discovery Rate

TERMS

page in report text
(first used)

[1] Bolded text indicates a term that is defined elsewhere in the glossary.

used for **Cross-Validation** (CV). The CoV is calculated as the ratio of the standard deviation to the mean and is also referred to as Relative Standard Deviation (RSD). The CoV represents the percentage of standard deviation from the sample mean (see Lehmann 1986).

Cohen's Kappa Coefficient (see Kappa).....61

Cross-Validation (CV) (see **Generalized Cross Validation** and **K-folds Cross-Validation**) 82

Cultural Resources Geographic Information System (CRGIS) 48
Computerized database and mapping tool for the visualization and analysis of cultural resources data within the Commonwealth of Pennsylvania. This tool is developed and administered through a join agreement between the Pennsylvania Historical and Museum Commission and the Pennsylvania Department of Transportation. (This tool is available at: www.portal.state.pa.us/portal/server.pt/community/crgis/3802.)

Earth (see also **Multivariate Adaptive Regression Splines**) 1
Earth is an implementation of the **Multivariate Adaptive Regression Splines** algorithm written in the **R Statistical Language** (see Milborrow 2011).

Euclidian Distance 65
The simple, or straight-line, distance between two points, colloquially described as "as the crow flies."

False Negative Rate (FNR).....101
The fraction of the positive observation (site locations) that are incorrectly classified as a negative observation (site not-likely). The FNR is derived from the **Confusion Matrix** and calculated by dividing the number of false negatives by total number of observed positive observations. This number is also interpreted as the Type-II error rate, or beta (β).

False Positive Rate (FPR) 101
The fraction of the negative observations (background locations) that are incorrectly classified as a positive observations (site likely). The FPR is derived from the **Confusion Matrix** and calculated by dividing the number of false positives by total number of observed negative observations. This number is also interpreted as the Type-I error rate.

Geographic Information Systems (GIS) 4

A GIS is a computer application that stores, manages, displays, and manipulates information with a spatial component (see Wheatley and Gillings 2002).

Gini Importance criterion or Gini impurity 81

The Gini Importance criterion is a metric used within the **random forest** algorithm for both branch splitting and variable importance. For the former, the Gini criterion is used to measure the purity, or how well segregated the representation of sites versus background values, of the node following its split on one of *p* variables. The split is made using the variable the leads to the largest increase in node purity from the parent node to the two descendent nodes. For the latter, the Gini Importance criterion is used to assess the value of each variable's contribution to the model. For each instance that a variable is chosen to split a node, the decrease in Gini is added up and compared to the other variables. Those variables that contributed to a greater decrease in the Gini are considered to be more important to the model's ability to correctly classify (see Breiman 2001).

K-folds Cross-Validation 82

Cross-Validation is the method by which a sample of observations is split into a number of different but equal-sized classes. The number of classes is referred to as K and the classes themselves are referred to as folds, hence "K-folds Cross-Validation." This is a method by which models can be validated on test sets that were not part of the training set, while at the same time, using the entire data set for modeling (see Efron and Tibshirani 1997).

Kappa coefficient 61

The Kappa coefficient, or **Cohen's Kappa coefficient**, is a statistical measure of a predictions agreement with real observations after accounting for chance agreement. In this project, the Kappa is used in a similar fashion as the **Kvamme Gain statistic**. However, the Kappa's calculation of by-chance observation is more inclusive that the **Kvamme Gain**. The Kappa statistic is derived from the confusion matrix and is used to compare model results of similar prevalence (see Viera and Garrett 2005).

Kolmogorov–Smirnov (K-S) Test 80

A non-parametric statistical test that measures the equality of continuous unpaired probability distributions to each other (two-sample test) or a reference distribution (one-sample test). In this study, the K-S test is used to test whether the distribution of an environmental variable is significantly different between known site locations and the overall environmental background (see Conover 1999).

Kvamme Gain (Kg).....74

The Kg is a metric used to assess the ability of a model to correctly classify positive observations (site present) given the area in which positive observations are predicted to occur (site-likely area). The higher the gain, the greater the ratio of percent sites present to percent of the modeled area considered site-likely. This measure does not take into account model precision or **True Positive Rate** (**Sensitivity**), meaning that an equivalent Kg statistic can be reached by correctly predicting 16% of known sites in 5% of the area or 95% of known sites in 30% of the area (see Kvamme 1988).

Logistic Regression (LR).....1

Logistic Regression is a statistical model used to predict for a binary response (0 or 1) or to classify a categorical response ("dead" or "alive") based on one or more predictors. This method uses a S-shaped logistic transformation to model the binary response probability as the log odds of the linear function of the predictor variables. Simply, the model fits the linear model to the S-shaped curve so that the prediction is kept between 0 and 1 (see Pampel 2000).

Mann-Whitney (MW) U Test 80

The Mann-Whitney U Test is a non-parametric statistical test that evaluates the dissimilarity of unpaired distributions by ranking the observations and comparing the mean ranks. This test is similar in concept to the **Kolmogorov–Smirnov Test**, but uses a ranked approach as opposed to a distance approach. The MW U Test is more sensitive to changes in the median of two distributions (see Lehman 1975).

Multivariate Adaptive Regression Splines (MARS).....1

A statistical model that is an extension of the **Generalized Linear Model**. This method approximates a non-linear model by fitting piecewise linear segments that are connected at nodes referred to as hinge functions. The hinge functions provide the point at which the two straight lines join. A sequence of lines and hinges approximates a non-linear **Spline**. The MARS model uses a forward pass to find the best fit that minimizes the **Sum of Squared Error**. This first pass is referred to as "greedy" because it seeks the best fit regardless of how many terms, or line and hinge segments, it creates. To avoid over-fitting, the MARS method has a second pass that prunes the terms created in the first path to assess which can be removed without having large negative effects on the model's performance; this lowers the model's complexity and variance. The MARS method uses **Generalized Cross-Validation** to assess how pruning affects performance. This method was introduced by Friedman (1991).

Negative Prediction Gain (NPG) 101

The NPG is a statistic that is derived from the confusion matrix to assess a model's ability to correctly classify site-unlikely areas. The NPG quantifies how much less likely a site discovery is at a location labeled site-unlikely using the model than if surveying at random. Ideally, a model would have a low NPG and a high **Positive Predictive Gain** (see Oehlert and Shea 2007).

Negative Prediction Value (NPV) 101

The NPV is a measure that is derived from the confusion matrix. This measures the probability that a non-site cell is correctly labeled as a background cell (see Oehlert and Shea 2007).

Pennsylvania Archaeological Site Survey (PASS).....1

The PASS files are a collection of paper forms, maps, reports, and photographs that document the location and attributes of known archaeological sites within the Commonwealth of Pennsylvania. These files have been digitized and can be accessed through the **Cultural Resources Geographic Information System**.

Positive Predictive Gain (PPG).....101

The PPG is a statistic that is derived from the **Confusion Matrix** to assess a model's ability to correctly classify site-likely areas. The PPG quantifies how much more likely a site discovery is at a location labeled site-likely using the model than if surveying at random. Ideally, a model would have a high PPG and a low **Negative Prediction Value** (see Oehlert and Shea 2007).

Positive Prediction Value (PPV).....101

The PPV is a measure that is derived from the **Confusion Matrix**. This measures the probability that a site cell is correctly labeled as a site-likely cell (see Oehlert and Shea 2007).

Random Forests 14

Random Forests is trademarked statistical classification algorithm created by Leo Breiman and Adele Cutler. Random Forests is a tree based ensemble method that builds off the ideas of **Classification and Regression Trees** and **Bagging**. The primary features of Random Forests include internal testing through **Bootstrap Aggregating** and variable importance via random subset selection (see Breiman 2001).

randomForest (RF) (see also **Random Forests**) 1

RF is an implementation of the **Random Forests** classification algorithm written in the **R Statistical Language** (see Liaw and Wiener 2002).

Receiver Operating Characteristics (ROC).....73

The ROC is a graphical representation of statistical classification model results. The ROC graph typically takes on a curved shape and is therefore often referred to as the ROC curve. The x-axis of the ROC graph is a model's **False Positive Rate** and the y-axis is the **True Positive Rate**; both are scaled from 0 to 1. The quantities on the x- and y-axes are also referred to as 1 – Specificity and **Sensitivity**, respectively. The actual curve in the graphic is generated by calculating the **True Positive Rate** and **False Positive Rate** for each cut-point of the model's prediction. The graphic also contains a line (often dashed) that originates at point 0,0 and goes at a 45-degree angle to point 1,1. This line represents a model that has no predictive power. The closer the ROC curve is to the upper left corner of the graph (which is point x = 0, y = 1), the greater the predictive power. Put another way, the best classification has the largest area under the curve. A line of this description will have a high **True Positive Rate** for the entire range of **False Positive Rates**. The ROC curve can be used to estimate the total predictive power of the model, often enumerated as the **Area Under Curve**, to compare similar models across all cut-points, or select an optimal cut-point to use for classification, resulting in a **Confusion Matrix** (see Fawcett 2004).

Root Mean Square Error (RMSE)..... i

The RMSE is a statistic, or loss function, used to quantify the difference between an estimate and a true value. The RMSE is calculated as the square root of the **Mean Squared Error**. When calculated on **Out-of-Sample** predictions, such as in this project, the RMSE represents the sample standard deviation of the prediction errors. The formula below is how RMSE is calculated, where n = the number of data values, y_j is the observed j^{th} value and $\hat{y}_j$ is the predicted j^{th} value for all j values from 1 to n. Therefore the RMSE is the square root of the average of all squared errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

A benefit of RMSE over **Mean Squared Error** is that it is scaled to the dependent variable and is therefore directly interpretable. With a binary dependent variable (0 to 1), the RMSE is taken as the distance on average between the predicted probability and the true value (see Salkind 2007).

Sensitivity (see also **True Positive Rate**) 73

Sensitivity is a term used for a classification's **True Positive Rate**; this value is also referred to as Recall. Sensitivity is the total fraction of sites that are classified by the model to be in the site-likely area. This measure is akin to the concept of precision and Type II errors. Sensitivity is calculated for a cut-point within a classification model as the number of correctly predicted positive observations (correctly classified sites) divided by the total number of actual positive observations (known sites) (see Oehlert and Shea 2007).

Specificity (see also **True Negative Rate**).....73

Specificity is a termed used for a classification's **True Negative Rate**. Specificity is the fraction of background that is classified as site-unlikely by the model. This measure is akin to the concept of accuracy and Type I errors. Specificity is calculated for a cut-point within a classification model as the number of correctly predicted negative observations (correctly classified non-sites) divided by the total number of actual negative observations (background cells) (see Oehlert and Shea 2007).

True Negative Rate (TNR) (see also **Specificity**).....101

The TNR is a measure of a model's classification at a given cut-point. Often referred to as a model's Specificity, the TNR is calculated as the percent of negative observations correctly classified as such. In this project, this would be the rate at which background cells are correctly classified as site un-likely cells (see Oehlert and Shea 2007).

True Positive Rate (TPR) (see also **Sensitivity**).....101

The TPR is a measure of a model's classification at a given cut-point. Often referred to as a models Sensitivity, the TPR is calculated as the percent of positive observations correctly classified as such. In this project, this would be the rate at which known site-present cells are correctly classified as site-likely cells (see Oehlert and Shea 2007).

Unexpected Discovery Rate (UDR).....101

The UDR is a measurement of a model's classification ability at a given cut-point. The UDR is defined as the probability of a cell containing a site given that the model predicted it as site-unlikely. That can be thought of as the rate of unintentional discovery, or "oops" rate (see Oehlert and Shea 2007).

APPENDIX B

VARIABLES CONSIDERED WITHIN REGIONS 1, 2, AND 3

Predictor	Family	Measure	Neighborhood Sizes	Description
aspect	Topography	bearing	n/a	Orientation of slope relative to north
c_hyd_min	Hydrology	cost-distance	n/a	Minimum distance to stream or water body
c_hyd_min_wt	Hydrology	cost-distance	n/a	Minimum distance to stream, water body, or wetland
c_trail_dist	Topography - Cultural	cost-distance	n/a	Cost-distance to historically documented Native American trails (Wallace 1965).
cd_conf	Hydrology	cost-distance	n/a	Cost-Distance to stream confluence (NHD flow lines)
cd_drnh	Hydrology	cost-distance	n/a	Cost-Distance to stream heads (NHD flow lines)
cd_h1	Hydrology	cost-distance	n/a	Cost-distance to historic streams
cd_h2	Hydrology	cost-distance	n/a	Cost-distance to NHD flow lines
cd_h3	Hydrology	cost-distance	n/a	Cost-distance to NHD water bodies
cd_h4	Hydrology	cost-distance	n/a	Cost-distance to NWI wetlands
cd_h5	Hydrology	cost-distance	n/a	Cost-distance to NWI water bodies
cd_h6	Hydrology	cost-distance	n/a	Cost-distance to 4th order and higher streams
cd_h7	Hydrology	cost-distance	n/a	Cost-distance to 3rd order and higher streams
dem_fll	Topography	elevation, meters (float)	n/a	1/3rd Arc-second digital elevation model as float, with sinks filled
e_hyd_min	Hydrology	Euclidian-distance, meters	n/a	Minimum distance to stream or water body
e_hyd_min_wt	Hydrology	Euclidian-distance, meters	n/a	Minimum distance to stream, water body, or wetland
e_trail_dist	Topography - Cultural	Euclidian-distance, meters	n/a	Euclidian-Distance to historically documented Native American trails (Wallace 1965).
ed_conflu	Hydrology	Euclidian-distance, meters	n/a	Euclidian-Distance to stream confluence (NHD flow lines)
ed_drnh	Hydrology	Euclidian-distance, meters	n/a	Euclidian-Distance to stream heads (NHD flow lines)

Predictor	Family	Measure	Neighborhood Sizes	Description
ed_h1	Hydrology	Euclidian-distance, meters	n/a	Euclidian-distance to historic streams
ed_h2	Hydrology	Euclidian-distance, meters	n/a	Euclidian-distance to NHD flow lines
ed_h3	Hydrology	Euclidian-distance, meters	n/a	Euclidian-distance to NHD water bodies
ed_h4	Hydrology	Euclidian-distance, meters	n/a	Euclidian-distance to NWI wetlands
ed_h5	Hydrology	Euclidian-distance, meters	n/a	Euclidian-distance to NWI water bodies
ed_h6	Hydrology	Euclidian-distance, meters	n/a	Euclidian-distance to 4th order and higher streams
ed_h7	Hydrology	Euclidian-distance, meters	n/a	Euclidian-distance to 3rd order and higher streams
eldrop#c	Topography	elevation, meters	1,8,10,16,32 cells	Drop in elevation over # cell neighborhood
elev_2_conf	Topography - Hydrology	vertical-distance, meters	na	Elevation to stream confluence (NHD flow lines)
elev_2_drainh	Topography - Hydrology	vertical-distance, meters	na	Elevation to stream head (NHD flow lines)
elev_2_strm	Topography - Hydrology	vertical-distance, meters	na	Elevation to stream (NHD flow lines)
flowdir	Hydrology	direction, bearing	na	Flow direction based on DEM
flw_acum	Hydrology	accumulation, cells	na	Flow accumulation based on DEM
random	Random	random float (0 to 1)	na	Randomly selected number between 1 and 0
rel_#c	Topography	index, 0 to 1	1,8,10,16,32 cells	Relative topographic position
rng_#c	Topography	elevation range, integer	1,8,10,16,32 cells	Range of elevation in # cell neighborhood
slope_deg	Topography	slope, degrees	n/a	Topographic slope measured in degrees
slope_pct	Topography	slope, percent	n/a	Topographic slope measured in percent rise over run
slpvr_#c	Topography	slope range, integer	1,8,10,16,32 cells	Slope variability within # cell neighborhood
std_#c	Topography	standard deviation	1,8,10,16,32 cells	Standard deviation of elevation range within # cell neighborhood

Predictor	Family	Measure	Neighborhood Sizes	Description
tpi_#c	Topography	index, integer	5,10,50,100,250 cells	Topographic Position Index. Position of cell relative to surrounding landscape within # cell neighborhood
tpi_cls#c	Topography	classification, nominal	5,10,50,100,250 cells	TPI standardized and classified into 1 standard deviation groups within # cell neighborhood
tpi_sd#c	Topography	standard deviation	5,10,50,100,250 cells	Standard deviation of TPI within # cell neighborhood
tri_#c	Topography	index, integer	1,8,10,16,32 cells	Topographic Ruggedness Index. Measure of terrain roughness within # cell neighborhood
twi#c	Topography - Hydrology	index, integer	1,8,10,16,32 cells	Topographic Wetness Index. Measure of upslope accumulation within # cell neighborhood
vrf_#c	Topography	index, integer	1,8,10,16,32 cells	Vector Roughness Factor. Measure of three-dimensional variation in slope within # cell neighborhood

APPENDIX C

VARIABLES SELECTED

FOR EACH OF 32 MODELS

WITHIN REGIONS 1, 2, AND 3

Region 1 East - Riverine Section 1				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
vrf_32c	0.525	p < 0.001	194616666	p < 0.001
cd_h1	0.522	p < 0.001	229921531	p < 0.001
rng_10c	0.518	p < 0.001	191987492	p < 0.001
std_16c	0.512	p < 0.001	196704585	p < 0.001
tpi_sd10c	0.508	p < 0.001	799210654	p < 0.001
elev_2_strm	0.501	p < 0.001	244428342	p < 0.001
ed_h1	0.477	p < 0.001	264579102	p < 0.001
tri_1c	0.460	p < 0.001	221756677	p < 0.001
e_hyd_min	0.458	p < 0.001	754127844	p < 0.001
splvr_8c	0.442	p < 0.001	224322584	p < 0.001
slope_pct	0.439	p < 0.001	239626694	p < 0.001
cd_conf	0.351	p < 0.001	326274606	p < 0.001
cd_h4	0.344	p < 0.001	301641618	p < 0.001
cd_drnh	0.339	p < 0.001	319128556	p < 0.001
ed_drnh	0.302	p < 0.001	677671039	p < 0.001
random	0.006	p = 0.656	500383285	p = 0.622

Region 1 East - Riverine Section 2				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
tpi_sd10c	0.433	p < 0.001	466957293	p < 0.001
slpvr_8c	0.390	p < 0.001	147094368	p < 0.001
rng_16c	0.380	p < 0.001	150414837	p < 0.001
std_16c	0.377	p < 0.001	145463729	p < 0.001
cd_h4	0.348	p < 0.001	170052438	p < 0.001
elev_2_drainh	0.345	p < 0.001	247605322	p < 0.001
ed_h5	0.327	p < 0.001	203535705	p < 0.001
vrf_10c	0.314	p < 0.001	182037047	p < 0.001
ed_drnh	0.299	p < 0.001	409911547	p < 0.001
cd_drnh	0.290	p < 0.001	359834248	p < 0.001
ed_h6	0.282	p < 0.001	271282347	p < 0.001
cd_h6	0.279	p < 0.001	193518573	p < 0.001
slope_pct	0.274	p < 0.001	188551715	p < 0.001
ed_h2	0.272	p < 0.001	409752227	p < 0.001
random	0.009	p = 0.412	301701370	p = 0.354

Region 1 East - Riverine Section 3				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
cd_h6	0.438	p < 0.001	106361476	p < 0.001
cd_drnh	0.413	p < 0.001	120235698	p < 0.001
std_8c	0.403	p < 0.001	111883730	p < 0.001
cd_h4	0.396	p < 0.001	116694469	p < 0.001
ed_h4	0.393	p < 0.001	127689435	p < 0.001
rng_8c	0.392	p < 0.001	111330214	p < 0.001
ed_h6	0.386	p < 0.001	146423504	p < 0.001
vrf_32c	0.374	p < 0.001	122114765	p < 0.001
tri_8c	0.365	p < 0.001	120901629	p < 0.001
tpi_sd10c	0.363	p < 0.001	344279147	p < 0.001
slpvr_8c	0.361	p < 0.001	122905390	p < 0.001
cd_h1	0.350	p < 0.001	133128794	p < 0.001
elev_2_strm	0.341	p < 0.001	147582383	p < 0.001
slope_pct	0.326	p < 0.001	132362379	p < 0.001
cd_conf	0.287	p < 0.001	150391604	p < 0.001
ed_h1	0.286	p < 0.001	187030587	p < 0.001
random	0.008	p = 0.683	232755913	p = 0.776

Region 1 East - Upland Section 1				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
cd_h7	0.645	p < 0.001	35984188	p < 0.001
elev_2_strm	0.631	p < 0.001	39530838	p < 0.001
cd_h3	0.606	p < 0.001	42025442	p < 0.001
cd_conf	0.554	p < 0.001	54556139	p < 0.001
cd_drnh	0.526	p < 0.001	50344440	p < 0.001
ed_h7	0.517	p < 0.001	62282001	p < 0.001
elev_2_conf	0.483	p < 0.001	60972522	p < 0.001
tpi_sd250c	0.461	p < 0.001	82614717	p < 0.001
c_hyd_min	0.455	p < 0.001	54579589	p < 0.001
rng_32c	0.439	p < 0.001	59655543	p < 0.001
eldrop32c	0.414	p < 0.001	56870148	p < 0.001
ed_h4	0.373	p < 0.001	74845285	p < 0.001
std_32c	0.358	p < 0.001	63959901	p < 0.001
tri_8c	0.354	p < 0.001	75947730	p < 0.001
slope_pct	0.352	p < 0.001	69008686	p < 0.001
tri_10c	0.346	p < 0.001	78946962	p < 0.001
slpvr_10c	0.343	p < 0.001	86309945	p < 0.001
random	0.009	p = 0.773	136900487	p = 0.731

Region 1 East - Upland Section 2				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
ed_h6	0.457	p < 0.001	688311177	p < 0.001
ed_h1	0.282	p < 0.001	583624746	p < 0.001
cd_h3	0.237	p < 0.001	393113591	p < 0.001
tpi_sd250c	0.227	p < 0.001	423641506	p < 0.001
cd_conf	0.204	p < 0.001	414057704	p < 0.001
elev_2_strm	0.202	p < 0.001	379201630	p < 0.001
ed_h3	0.194	p < 0.001	397002632	p < 0.001
cd_h7	0.191	p < 0.001	386266152	p < 0.001
ed_conf	0.183	p < 0.001	430897913	p < 0.001
rng_8c	0.145	p < 0.001	406440011	p < 0.001
elev_2_conf	0.141	p < 0.001	441571334	p < 0.001
std_1c	0.138	p < 0.001	402209094	p < 0.001
slope_pct	0.136	p < 0.001	403922650	p < 0.001
c_hyd_min	0.134	p < 0.001	438773708	p < 0.001
random	0.007	p = 0.593	486767756	p = 0.643

Region 1 East - Upland Section 3				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
elev_2_strm	0.338	p < 0.001	212596563	p < 0.001
cd_conf	0.322	p < 0.001	209066613	p < 0.001
cd_h6	0.316	p < 0.001	202595767	p < 0.001
cd_h1	0.302	p < 0.001	208264944	p < 0.001
elev_2_conf	0.300	p < 0.001	226158347	p < 0.001
c_hyd_min_wt	0.285	p < 0.001	212658603	p < 0.001
std_16c	0.281	p < 0.001	203491341	p < 0.001
cd_h3	0.279	p < 0.001	221719810	p < 0.001
rng_8c	0.275	p < 0.001	203481463	p < 0.001
slope_pct	0.257	p < 0.001	208220083	p < 0.001
cd_drnh	0.247	p < 0.001	213878353	p < 0.001
tri_1c	0.231	p < 0.001	218871409	p < 0.001
ed_h4	0.230	p < 0.001	225961927	p < 0.001
eldrop32c	0.227	p < 0.001	222682651	p < 0.001
tpi_sd100c	0.192	p < 0.001	264819497	p < 0.001
ed_h6	0.180	p < 0.001	274904382	p < 0.001
ed_conf	0.170	p < 0.001	273801779	p < 0.001
random	0.010	p = 0.360	317442442	p = 0.338

Region 1 North - Riverine Section 1				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
elev_2_strm	0.338	p < 0.001	212596563	p < 0.001
cd_conf	0.322	p < 0.001	209066613	p < 0.001
cd_h7	0.301	p < 0.001	208301343	p < 0.001
cd_h3	0.279	p < 0.001	221719810	p < 0.001
rng_10c	0.275	p < 0.001	204708334	p < 0.001
std_10c	0.272	p < 0.001	202637442	p < 0.001
slope_pct	0.257	p < 0.001	208220083	p < 0.001
cd_drnh	0.247	p < 0.001	213878353	p < 0.001
tri_8c	0.200	p < 0.001	237535878	p < 0.001
slpvr_8c	0.157	p < 0.001	259385451	p < 0.001
vrf_32c	0.149	p < 0.001	265498897	p < 0.001
ed_h7	0.124	p < 0.001	292223365	p < 0.001
ed_h5	0.062	p < 0.001	334445402	p < 0.001
tpi_sd10c	0.053	p < 0.001	318390177	p < 0.001
random	0.010	p = 0.360	317442442	p = 0.338

Region 1 North - Riverine Section 2				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
ed_h2	0.373	p < 0.001	1527697518	p < 0.001
ed_h7	0.332	p < 0.001	862912657	p < 0.001
cd_h1	0.316	p < 0.001	823902937	p < 0.001
e_hyd_min_wt	0.314	p < 0.001	1475328107	p < 0.001
tpi_sd5c	0.297	p < 0.001	1457996473	p < 0.001
elev_2_strm	0.265	p < 0.001	866965421	p < 0.001
rng_10c	0.234	p < 0.001	762915929	p < 0.001
vrf_32c	0.230	p < 0.001	791297802	p < 0.001
std_10c	0.217	p < 0.001	778860060	p < 0.001
c_hyd_min_wt	0.201	p < 0.001	1254572090	p < 0.001
elev_2_conf	0.200	p < 0.001	1232173734	p < 0.001
slpvr_32c	0.177	p < 0.001	1241133541	p < 0.001
tri_8c	0.176	p < 0.001	843961123	p < 0.001
random	0.006	p = 0.387	1043797068	p = 0.320

Region 1 North - Upland Section 1				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
elev_2_strm	0.510	p < 0.001	155284553	p < 0.001
cd_h1	0.501	p < 0.001	176790076	p < 0.001
tri_32c	0.341	p < 0.001	264281614	p < 0.001
ed_h4	0.333	p < 0.001	244329614	p < 0.001
cd_h5	0.317	p < 0.001	274653038	p < 0.001
rng_16c	0.316	p < 0.001	279727274	p < 0.001
std_16c	0.300	p < 0.001	284163668	p < 0.001
cd_conf	0.281	p < 0.001	285729128	p < 0.001
ed_conf	0.275	p < 0.001	308051711	p < 0.001
cd_drnh	0.271	p < 0.001	299848452	p < 0.001
slope_pct	0.229	p < 0.001	304399923	p < 0.001
tpi_sd10c	0.214	p < 0.001	518584324	p < 0.001
ed_drnh	0.202	p < 0.001	318853214	p < 0.001
elev_2_conf	0.197	p < 0.001	336114607	p < 0.001
random	0.007	p = 0.582	416107660	p = 0.708

Region 1 North - Upland Section 2				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
ed_h7	0.391	p < 0.001	503792264	p < 0.001
tpi_sd250c	0.380	p < 0.001	543013386	p < 0.001
cd_h6	0.363	p < 0.001	565052270	p < 0.001
cd_h1	0.362	p < 0.001	535885766	p < 0.001
elev_2_strm	0.335	p < 0.001	580859711	p < 0.001
elev_2_drainh	0.261	p < 0.001	669736239	p < 0.001
ed_drnh	0.255	p < 0.001	1253650815	p < 0.001
tri_32c	0.245	p < 0.001	1255048764	p < 0.001
ed_conf	0.185	p < 0.001	745407111	p < 0.001
cd_h4	0.178	p < 0.001	759938750	p < 0.001
cd_conf	0.174	p < 0.001	747969039	p < 0.001
elev_2_conf	0.168	p < 0.001	776269365	p < 0.001
cd_drnh	0.158	p < 0.001	1114504139	p < 0.001
random	0.005	p = 0.729	946838351	p = 0.594

Region 1 West - Riverine Section 1				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
tpi_sd250c	0.275	p < 0.001	181707569	p < 0.001
ed_h6	0.246	p < 0.001	167944992	p < 0.001
cd_h1	0.238	p < 0.001	194553961	p < 0.001
std_10c	0.224	p < 0.001	167362254	p < 0.001
rng_8c	0.222	p < 0.001	169590585	p < 0.001
c_hyd_min_wt	0.206	p < 0.001	271390636	p < 0.001
elev_2_strm	0.203	p < 0.001	199195913	p < 0.001
slope_pct	0.173	p < 0.001	199973471	p < 0.001
eldrop32c	0.152	p < 0.001	261400553	p < 0.001
elev_2_conf	0.140	p < 0.001	243705997	p < 0.001
slpvr_10c	0.139	p < 0.001	190120165	p < 0.001
cd_conf	0.130	p < 0.001	226070290	p < 0.005
cd_h4	0.125	p < 0.001	209922772	p < 0.001
cd_drnh	0.120	p < 0.001	242952383	p < 0.001
random	0.009	p = 0.550	227080319	p = 0.744

Region 1 West - Riverine Section 2				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
tpi_sd10c	0.349	p < 0.001	617769516	p < 0.001
ed_h2	0.342	p < 0.001	593893906	p < 0.001
std_10c	0.339	p < 0.001	236037223	p < 0.001
rng_10c	0.338	p < 0.001	235841795	p < 0.001
e_hyd_min	0.320	p < 0.001	578902796	p < 0.001
ed_h7	0.284	p < 0.001	323179460	p < 0.001
slpvr_8c	0.281	p < 0.001	264498043	p < 0.001
cd_h1	0.242	p < 0.001	303696628	p < 0.001
cd_h3	0.223	p < 0.001	317992732	p < 0.001
ed_h5	0.216	p < 0.001	349302638	p < 0.001
slope_pct	0.199	p < 0.001	321541264	p < 0.001
elev_2_strm	0.183	p < 0.001	331428356	p < 0.001
random	0.006	p = 0.694	421419064	p = 0.663

Region 1 West - Riverine Section 3				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
ed_h6	0.455	p < 0.001	613025889	p < 0.001
tpi_sd250c	0.424	p < 0.001	485885477	p < 0.001
cd_h6	0.374	p < 0.001	649003708	p < 0.001
ed_h2	0.346	p < 0.001	1404383655	p < 0.001
elev_2_drainh	0.320	p < 0.001	636615275	p < 0.001
e_hyd_min	0.297	p < 0.001	1325302997	p < 0.001
ed_h5	0.296	p < 0.001	795569066	p < 0.001
cd_h2	0.227	p < 0.001	1239955685	p < 0.001
elev_2_conf	0.222	p < 0.001	1186601881	p < 0.001
tri_32c	0.219	p < 0.001	1152382861	p < 0.001
cd_h5	0.208	p < 0.001	833274211	p < 0.001
ed_drnh	0.204	p < 0.001	1231569989	p < 0.001
eldrop32c	0.203	p < 0.001	1199109209	p < 0.001
rng_16c	0.199	p < 0.001	784869210	p < 0.001
c_hyd_min	0.194	p < 0.001	1178484753	p < 0.001
random	0.005	p = 0.677	982886641	p = 0.631

Region 1 West - Riverine Section 4				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
tpi_sd10c	0.394	p < 0.001	1037566798	p < 0.001
std_16c	0.354	p < 0.001	396313624	p < 0.001
rng_16c	0.352	p < 0.001	398394180	p < 0.001
tri_8c	0.300	p < 0.001	459871357	p < 0.001
ed_h7	0.299	p < 0.001	564044609	p < 0.001
tri_10c	0.297	p < 0.001	453742849	p < 0.001
cd_h1	0.270	p < 0.001	547471406	p < 0.001
e_hyd_min	0.262	p < 0.001	906797672	p < 0.001
cd_h5	0.251	p < 0.001	575072976	p < 0.005
elev_2_strm	0.222	p < 0.001	579731467	p < 0.001
random	0.008	p = 0.320	685774789	p = 0.182

Region 1 West - Riverine Section 5				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
tpi_sd10c	0.378	p < 0.001	402449583	p < 0.001
ed_h2	0.327	p < 0.001	364888146	p < 0.001
std_16c	0.319	p < 0.001	165457408	p < 0.001
cd_drnh	0.318	p < 0.001	189139147	p < 0.001
ed_h5	0.309	p < 0.001	198099741	p < 0.001
cd_h5	0.290	p < 0.001	203937045	p < 0.001
rng_10c	0.281	p < 0.001	182070073	p < 0.001
e_hyd_min_wt	0.268	p < 0.001	346201187	p < 0.001
tri_10c	0.262	p < 0.001	191088015	p < 0.001
cd_h2	0.261	p < 0.001	323606505	p < 0.001
eldrop32c	0.221	p < 0.001	316350780	p < 0.001
ed_h7	0.209	p < 0.001	275632301	p < 0.001
random	0.008	p = 0.613	263507821	p = 0.456

Region 1 West - Upland Section 1				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
std_8c	0.372	p < 0.001	516473407	p < 0.001
rng_8c	0.364	p < 0.001	523554253	p < 0.001
slope_pct	0.360	p < 0.001	537438926	p < 0.001
ed_h6	0.296	p < 0.001	676242476	p < 0.001
ed_h1	0.294	p < 0.001	661493964	p < 0.001
tri_1c	0.280	p < 0.001	632009661	p < 0.001
tpi_sd10c	0.248	p < 0.001	1351001865	p < 0.001
cd_h6	0.242	p < 0.001	702859058	p < 0.001
twi1c	0.185	p < 0.001	795505688	p < 0.001
flw_acum	0.183	p < 0.001	797754567	p < 0.001
random	0.006	p = 0.453	1024654569	p = 0.553

Region 1 West - Upland Section 2				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
rng_8c	0.413	p < 0.001	715738653	p < 0.001
std_8c	0.405	p < 0.001	723733881	p < 0.001
slope_pct	0.366	p < 0.001	809643069	p < 0.001
tri_1c	0.297	p < 0.001	953713071	p < 0.001
cd_h4	0.225	p < 0.001	1102695900	p < 0.001
slpvr_16c	0.217	p < 0.001	1175237596	p < 0.001
cd_drnh	0.202	p < 0.001	1210664315	p < 0.001
tpi_sd10c	0.185	p < 0.001	1916659116	p < 0.001
cd_h6	0.174	p < 0.001	1234897016	p < 0.001
eldrop8c	0.165	p < 0.001	1265774467	p < 0.001
random	0.005	p = 0.465	1595218889	p = 0.390

Region 1 West - Upland Section 3				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
ed_h6	0.426	p < 0.001	499468171	p < 0.001
ed_h1	0.405	p < 0.001	481084722	p < 0.001
tpi_sd250c	0.344	p < 0.001	570502661	p < 0.001
cd_h1	0.323	p < 0.001	554943990	p < 0.001
tri_32c	0.311	p < 0.001	1158951958	p < 0.001
elev_2_strm	0.296	p < 0.001	643256545	p < 0.001
ed_drnh	0.293	p < 0.001	1132741416	p < 0.001
elev_2_drainh	0.288	p < 0.001	581355345	p < 0.001
cd_conf	0.274	p < 0.001	635303467	p < 0.001
cd_h4	0.248	p < 0.001	670450018	p < 0.001
ed_h5	0.227	p < 0.001	625405208	p < 0.001
slope_pct	0.194	p < 0.001	660737843	p < 0.001
rng_1c	0.194	p < 0.001	659911533	p < 0.001
elev_2_conf	0.179	p < 0.001	704368947	p < 0.001
random	0.005	p = 0.690	871784432	p = 0.598

Region 1 West - Upland Section 4				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
rng_8c	0.376	p < 0.001	399521975	p < 0.001
std_10c	0.368	p < 0.001	406011831	p < 0.001
cd_h4	0.335	p < 0.001	449003166	p < 0.001
tri_10c	0.331	p < 0.001	464159157	p < 0.001
slope_pct	0.326	p < 0.001	441853439	p < 0.001
cd_h1	0.284	p < 0.001	484618971	p < 0.001
tpi_sd250c	0.271	p < 0.001	489232605	p < 0.001
cd_h2	0.270	p < 0.001	492719355	p < 0.001
elev_2_strm	0.269	p < 0.001	475962239	p < 0.001
cd_drnh	0.264	p < 0.001	484815356	p < 0.001
c_hyd_min_wt	0.249	p < 0.001	501671781	p < 0.001
cd_conf	0.241	p < 0.001	493869714	p < 0.001
eldrop32c	0.217	p < 0.001	538303307	p < 0.001
elev_2_conf	0.213	p < 0.001	535113681	p < 0.001
ed_h6	0.203	p < 0.001	583712044	p < 0.001
random	0.006	p = 0.551	732684240	p = 0.650

Region 1 West - Upland Section 5				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
rng_8c	0.398	p < 0.001	364009431	p < 0.001
std_10c	0.396	p < 0.001	365880280	p < 0.001
slope_pct	0.329	p < 0.001	414163035	p < 0.001
tri_1c	0.291	p < 0.001	439514148	p < 0.001
cd_h4	0.283	p < 0.001	481470451	p < 0.001
slpvr_32c	0.251	p < 0.001	495829578	p < 0.001
ed_h4	0.233	p < 0.001	502495090	p < 0.001
cd_drnh	0.228	p < 0.001	506930977	p < 0.001
eldrop8c	0.177	p < 0.001	539737048	p < 0.001
cd_h1	0.165	p < 0.001	533469246	p < 0.001
random	0.007	p = 0.378	705492758	p = 0.608

Region 2/3 - Riverine Section 1				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
tpi_sd250c	0.455	p < 0.001	673534520	p < 0.001
ed_h6	0.397	p < 0.001	797757266	p < 0.001
e_hyd_min_wt	0.332	p < 0.001	1879633020	p < 0.001
ed_h2	0.254	p < 0.001	1685543804	p < 0.001
cd_drnh	0.232	p < 0.001	1055770862	p < 0.001
ed_h4	0.228	p < 0.001	1671907421	p < 0.001
cd_h1	0.213	p < 0.001	1128228756	p < 0.001
vrf_32c	0.181	p < 0.001	1128458227	p < 0.001
cd_conf	0.170	p < 0.001	1491025510	p < 0.001
elev_2_strm	0.169	p < 0.001	1119225794	p < 0.001
tri_32c	0.163	p < 0.001	1619082668	p < 0.001
random	0.005	p = 0.561	1344232300	p = 0.457

Region 2/3 - Riverine Section 2				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
ed_h6	0.398	p < 0.001	584615052	p < 0.001
tpi_sd250c	0.394	p < 0.001	494861778	p < 0.001
cd_drnh	0.263	p < 0.001	1191352812	p < 0.001
ed_h2	0.251	p < 0.001	1197189674	p < 0.001
elev_2_conf	0.233	p < 0.001	1189994021	p < 0.001
ed_h5	0.232	p < 0.001	760588265	p < 0.001
eldrop32c	0.202	p < 0.001	1162101139	p < 0.001
e_hyd_min	0.198	p < 0.001	1171840177	p < 0.001
tri_32c	0.181	p < 0.001	1134269481	p < 0.001
random	0.006	p = 0.543	986368998	p = 0.512

Region 2/3 - Riverine Section 3				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
cd_drnh	0.596	p < 0.001	11296286	p < 0.001
cd_h6	0.569	p < 0.001	13593335	p < 0.001
tpi_250c	0.475	p < 0.001	48952243	p < 0.001
vrf_32c	0.462	p < 0.001	16991601	p < 0.001
cd_h4	0.436	p < 0.001	20743458	p < 0.001
cd_h1	0.412	p < 0.001	18389466	p < 0.001
cd_h5	0.404	p < 0.001	20789981	p < 0.001
cd_conf	0.386	p < 0.001	18226437	p < 0.001
std_32c	0.374	p < 0.001	19103640	p < 0.001
tri_32c	0.360	p < 0.001	22534158	p < 0.001
rng_8c	0.333	p < 0.001	21979809	p < 0.001
tri_16c	0.313	p < 0.001	24430276	p < 0.001
tpi_10c	0.289	p < 0.001	42536477	p < 0.001
ed_h3	0.279	p < 0.001	33552147	p < 0.001
random	0.019	p = 0.773	30547332	p = 0.592

Region 2/3 - Riverine Section 4				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
tri_32c	0.502	p < 0.001	42809740	p < 0.001
cd_h6	0.443	p < 0.001	17209558	p < 0.001
tpi_250c	0.416	p < 0.001	15802361	p < 0.001
ed_h1	0.348	p < 0.001	18058916	p < 0.001
cd_drnh	0.300	p < 0.001	26137639	p < 0.001
cd_h5	0.297	p < 0.001	22147114	p < 0.001
tri_16c	0.286	p < 0.001	37877221	p < 0.001
elev_2_strm	0.284	p < 0.001	19993765	p < 0.001
rel_32c	0.248	p < 0.001	22384092	p < 0.001
cd_conf	0.212	p < 0.001	24429166	p < 0.001
random	0.020	p = 0.739	28057415	p = 0.698

Region 2/3 - Riverine Section 5				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
ed_h5	0.459	p < 0.001	435076089	p < 0.001
ed_h6	0.456	p < 0.001	426033002	p < 0.001
tpi_sd250c	0.450	p < 0.001	335017954	p < 0.001
std_32c	0.360	p < 0.001	365118451	p < 0.001
rng_16c	0.341	p < 0.001	370641379	p < 0.001
std_16c	0.328	p < 0.001	368867654	p < 0.001
tri_16c	0.315	p < 0.001	419428700	p < 0.001
ed_h2	0.289	p < 0.001	877222211	p < 0.001
slpvr_8c	0.284	p < 0.001	431793200	p < 0.001
cd_drnh	0.265	p < 0.001	528134586	p < 0.001
e_hyd_min_wt	0.257	p < 0.001	865345825	p < 0.001
vrf_32c	0.244	p < 0.001	456872709	p < 0.001
random	0.006	p = 0.580	660132539	p = 0.596

Region 2/3 - Upland Section 1				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
cd_h6	0.323	p < 0.001	538534455	p < 0.001
tpi_250c	0.311	p < 0.001	500669489	p < 0.001
elev_2_strm	0.298	p < 0.001	524614341	p < 0.001
cd_h1	0.295	p < 0.001	483110043	p < 0.001
cd_h5	0.207	p < 0.001	584620553	p < 0.001
cd_conf	0.204	p < 0.001	623626014	p < 0.001
ed_h1	0.178	p < 0.001	586786184	p < 0.001
elev_2_conf	0.164	p < 0.001	674686946	p < 0.001
ed_h4	0.155	p < 0.001	647162075	p < 0.001
rel_32c	0.152	p < 0.001	644724579	p < 0.001
e_hyd_min	0.115	p < 0.001	674783626	p < 0.001
random	0.008	p = 0.202	756044013	p = 0.12

Region 2/3 - Upland Section 2				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
tpi_sd250c	0.224	p < 0.001	729889280	p < 0.001
tri_32c	0.202	p < 0.001	1237615884	p < 0.001
elev_2_strm	0.200	p < 0.001	795692104	p < 0.001
ed_drnh	0.198	p < 0.001	1259272503	p < 0.001
cd_h7	0.196	p < 0.001	778297954	p < 0.001
elev_2_drainh	0.171	p < 0.001	882873729	p < 0.001
std_32c	0.148	p < 0.001	1174696504	p < 0.001
ed_conf	0.139	p < 0.001	838159244	p < 0.001
eldrop32c	0.138	p < 0.001	1191811567	p < 0.001
rng_32c	0.120	p < 0.001	1164355813	p < 0.001
cd_h5	0.112	p < 0.001	1165898493	p < 0.001
vrf_32c	0.106	p < 0.001	1159029834	p < 0.001
c_hyd_min_wt	0.095	p < 0.001	1123026886	p < 0.001
random	0.006	p = 0.51	1017134480	p = 0.43

Region 2/3 - Upland Section 3				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
cd_h6	0.481	p < 0.001	139883684	p < 0.001
cd_h1	0.409	p < 0.001	185256348	p < 0.001
tpi_250c	0.374	p < 0.001	576753937	p < 0.001
cd_h5	0.365	p < 0.001	268455048	p < 0.001
cd_drnh	0.297	p < 0.001	285054433	p < 0.001
cd_conf	0.281	p < 0.001	274361693	p < 0.001
tri_1c	0.240	p < 0.001	284649072	p < 0.001
elev_2_conf	0.239	p < 0.001	344912098	p < 0.001
std_1c	0.233	p < 0.001	286659327	p < 0.001
rng_1c	0.233	p < 0.001	286809228	p < 0.001
vrf_32c	0.229	p < 0.001	298300841	p < 0.001
rel_32c	0.228	p < 0.001	503441442	p < 0.001
cd_h4	0.228	p < 0.001	331635270	p < 0.001
e_hyd_min_wt	0.225	p < 0.001	522243581	p < 0.001
slope_deg	0.219	p < 0.001	295395652	p < 0.001
slpvr_1c	0.218	p < 0.001	298734624	p < 0.001
elev_2_strm	0.172	p < 0.001	392147300	p < 0.001
random	0.008	p = 0.461	401627635	p = 0.314

Region 2/3 - Upland Section 4 without Rock Shelters				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
slpvr_16c	0.486	p < 0.001	99383963	p < 0.001
rng_16c	0.435	p < 0.001	92933066	p < 0.001
std_16c	0.423	p < 0.001	93320572	p < 0.001
eldrop32c	0.329	p < 0.001	92105964	p < 0.001
ed_h1	0.322	p < 0.001	49698562	p < 0.001
ed_h7	0.322	p < 0.001	49716007	p < 0.001
cd_drnh	0.306	p < 0.001	83636933	p < 0.001
vrf_32c	0.298	p < 0.001	89946697	p < 0.001
c_hyd_min_wt	0.286	p < 0.001	91314317	p < 0.001
eldrop16c	0.286	p < 0.001	89154358	p < 0.001
elev_2_conf	0.284	p < 0.001	85795584	p < 0.001
slope_deg	0.280	p < 0.001	88412233	p < 0.001
ed_h5	0.261	p < 0.001	52003819	p < 0.001
cd_conf	0.240	p < 0.001	83613553	p < 0.001
random	0.015	p = 0.647	66751737	p = 0.834

Region 2/3 - Upland Section 4 Rock Shelters				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
elev_2_strm	0.474	p < 0.001	61578033	p < 0.001
tpi_250c	0.445	p < 0.001	62975134	p < 0.001
elev_2_drainh	0.381	p < 0.001	61606940	p < 0.001
ed_conf	0.368	p < 0.001	57827833	p < 0.001
ed_conf	0.368	p < 0.001	57827833	p < 0.001
ed_h6	0.332	p < 0.001	48798428	p < 0.001
cd_drnh	0.331	p < 0.001	30724302	p < 0.001
std_32c	0.322	p < 0.001	29716384	p < 0.001
rng_16c	0.302	p < 0.001	31909277	p < 0.001
elev_2_conf	0.298	p < 0.001	51531047	p < 0.001
cd_h4	0.287	p < 0.001	31985707	p < 0.001
tri_32c	0.283	p < 0.001	34257797	p < 0.001
slpvr_32c	0.280	p < 0.001	34255241	p < 0.001
vrf_32c	0.260	p < 0.001	34046615	p < 0.001
cd_h1	0.254	p < 0.001	52818950	p < 0.001
e_hyd_min_wt	0.229	p < 0.001	51595354	p < 0.001
random	0.019	p = 0.597	42827114	p = 0.839

Region 2/3 - Upland Section 5 Rock Shelters				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
cd_h7	0.439	p < 0.001	53707997	p < 0.001
slpvr_32c	0.401	p < 0.001	160497862	p < 0.001
tri_32c	0.391	p < 0.001	160014166	p < 0.001
elev_2_strm	0.390	p < 0.001	59385670	p < 0.001
tpi_sd250c	0.386	p < 0.001	60576516	p < 0.001
elev_2_drainh	0.353	p < 0.001	70373326	p < 0.001
cd_conf	0.308	p < 0.001	74425081	p < 0.001
rel_32c	0.281	p < 0.001	89805536	p < 0.005
e_hyd_min	0.263	p < 0.001	75945888	p < 0.005
cd_h4	0.262	p < 0.001	80035583	p < 0.005
ed_drnh	0.259	p < 0.001	137506621	p < 0.005
std_32c	0.213	p < 0.001	135977633	p < 0.005
rng_32c	0.210	p < 0.001	136359107	p < 0.005
elev_2_conf	0.205	p < 0.001	86702438	p < 0.005
c_hyd_min	0.197	p < 0.001	86545537	p < 0.005
c_hyd_min_wt	0.195	p < 0.001	86747458	p < 0.005
random	0.011	p = 0.704	115194423	p = 0.509

Region 2/3 - Upland Section 5 without Rock Shelters				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
slpvr_16c	0.438	p < 0.001	267430669	p < 0.001
cd_conf	0.329	p < 0.001	238010392	p < 0.001
elev_2_strm	0.326	p < 0.001	245385724	p < 0.001
std_16c	0.310	p < 0.001	232236458	p < 0.001
elev_2_conf	0.307	p < 0.001	233630595	p < 0.001
eldrop32c	0.301	p < 0.001	239492413	p < 0.001
cd_h7	0.299	p < 0.001	226370080	p < 0.001
rng_10c	0.299	p < 0.001	230249835	p < 0.001
ed_h3	0.297	p < 0.001	235006030	p < 0.001
rel_32c	0.274	p < 0.001	226161544	p < 0.001
c_hyd_min_wt	0.272	p < 0.001	234577662	p < 0.001
cd_h4	0.268	p < 0.001	217489916	p < 0.001
tpi_cls10c	0.259	p < 0.001	222945606	p < 0.001
slope_deg	0.209	p < 0.001	215897959	p < 0.005
random	0.016	p = 0.123	171679425	p = 0.318

APPENDIX D

VARIABLE IMPORTANCE

FOR EACH OF 32 MODELS

WITHIN REGIONS 1, 2, AND 3

Chart 1. Region 1 East - Riverine Section 1

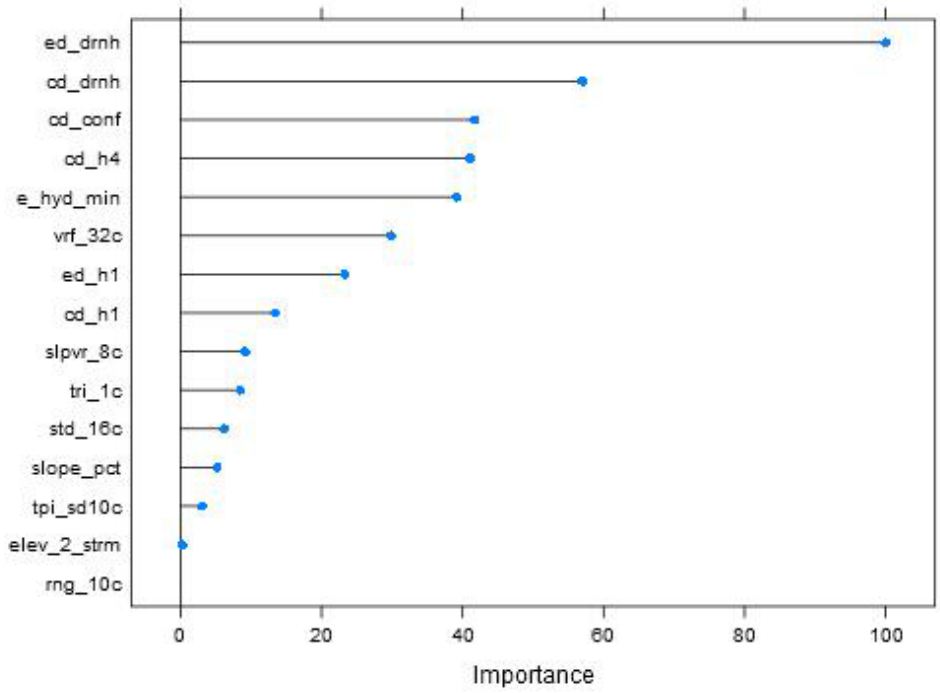


Chart 2. Region 1 East - Riverine Section 2

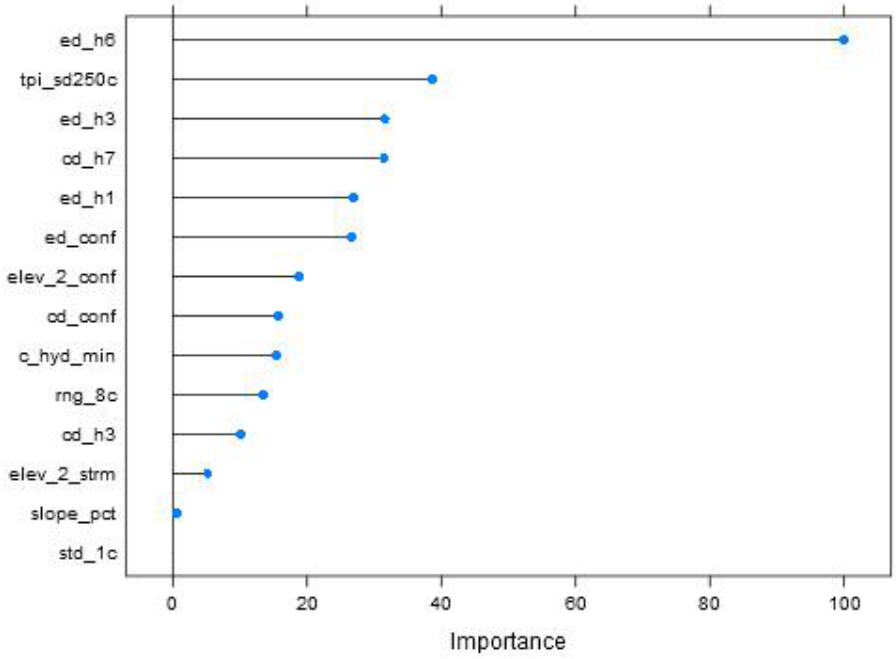


Chart 3. Region 1 East - Riverine Section 3

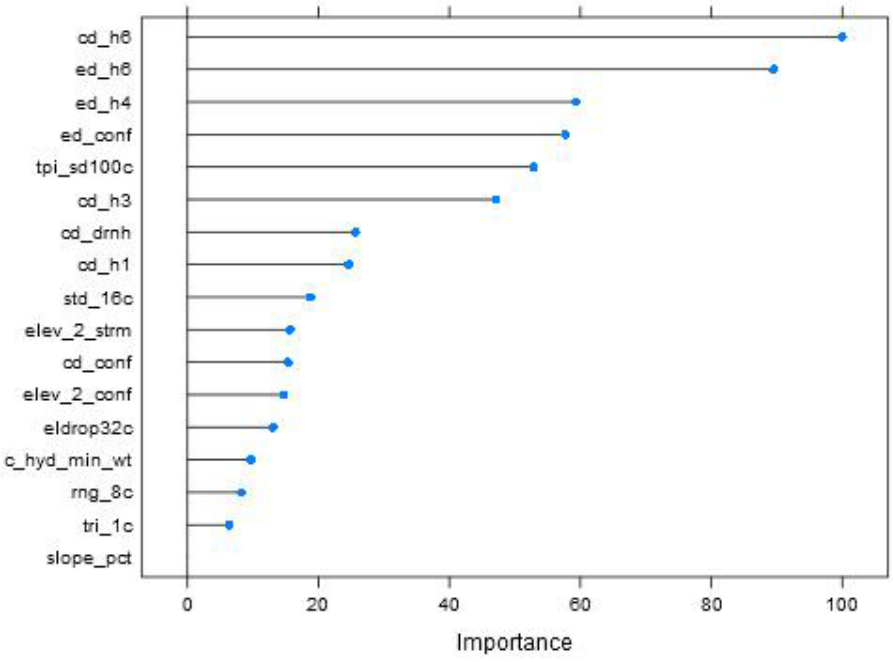


Chart 4. Region 1 East - Upland Section 1

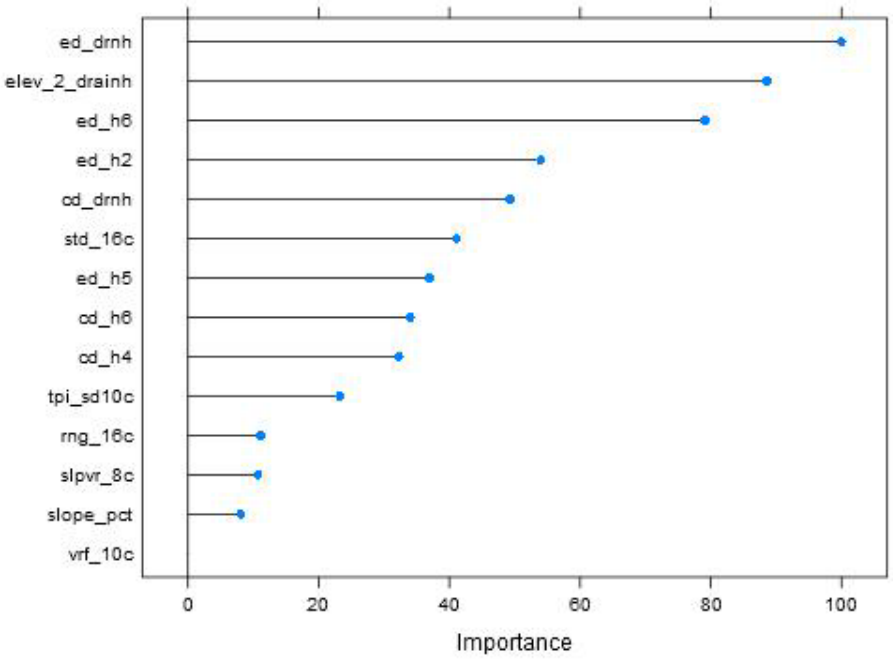


Chart 5. Region 1 East - Upland Section 2

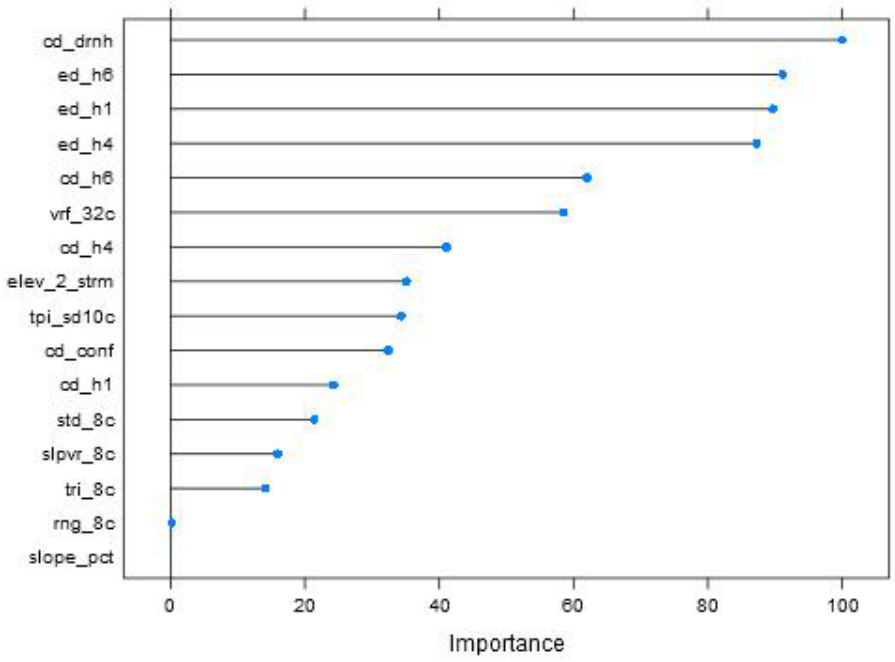


Chart 6. Region 1 East - Upland Section 3

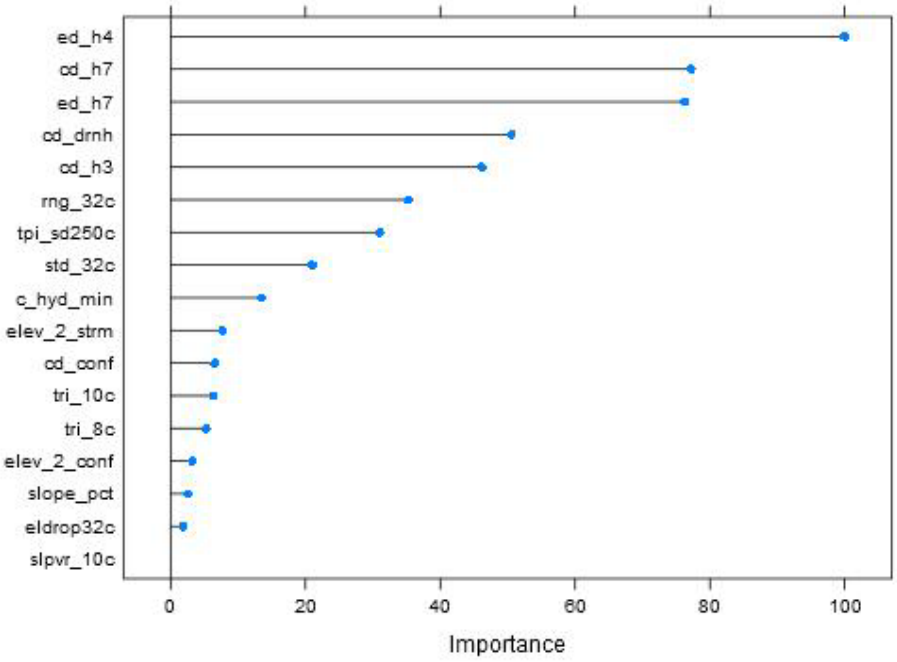


Chart 7. Region 1 North - Riverine Section 1

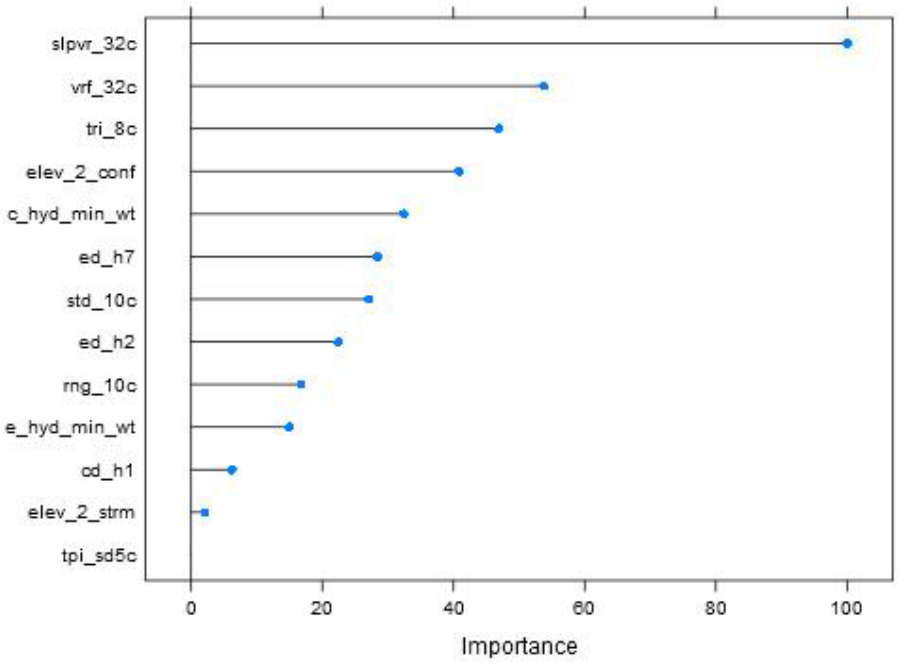


Chart 8. Region 1 North - Riverine Section 2

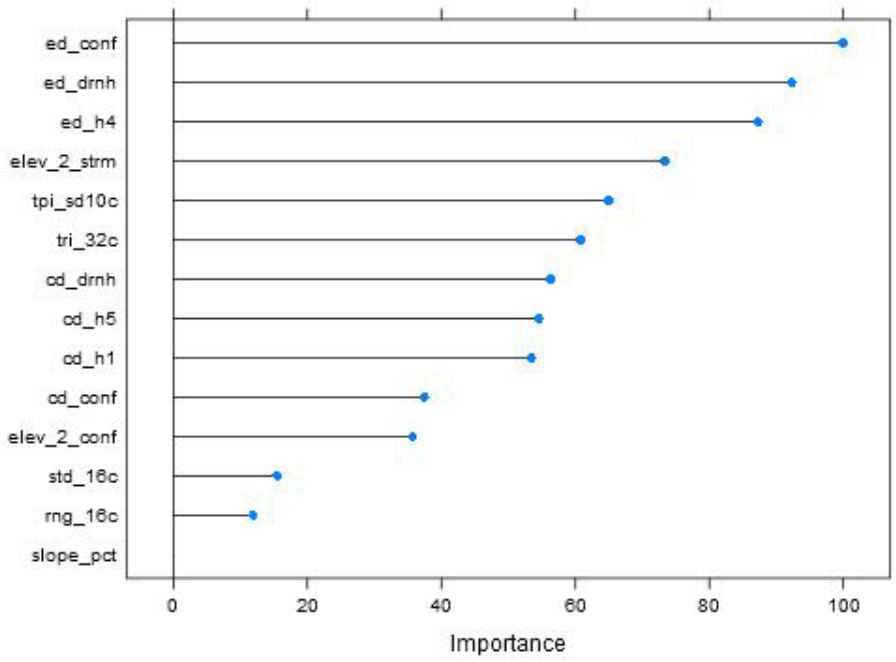


Chart 9. Region 1 North - Upland Section 1

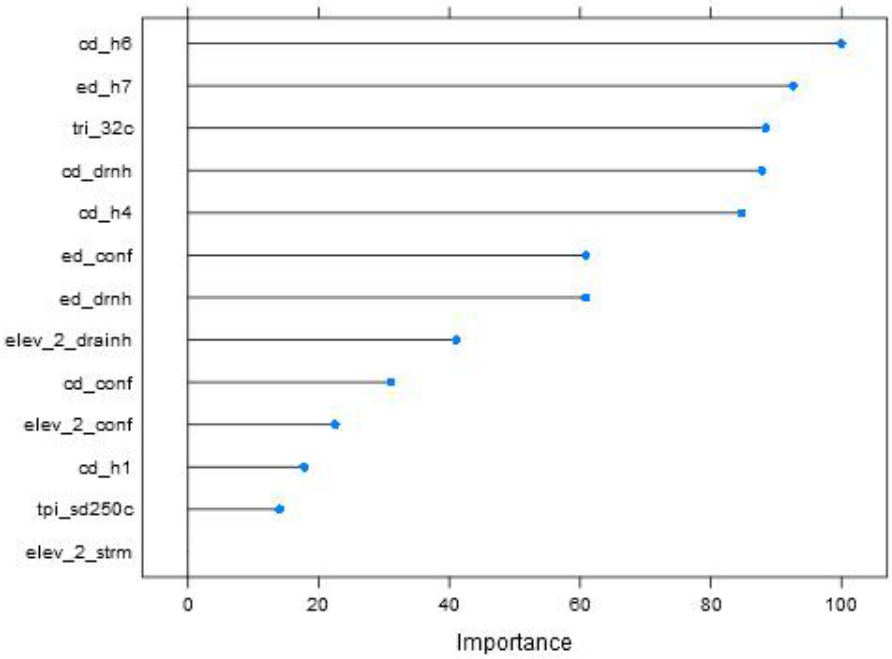


Chart 10. Region 1 North - Upland Section 2

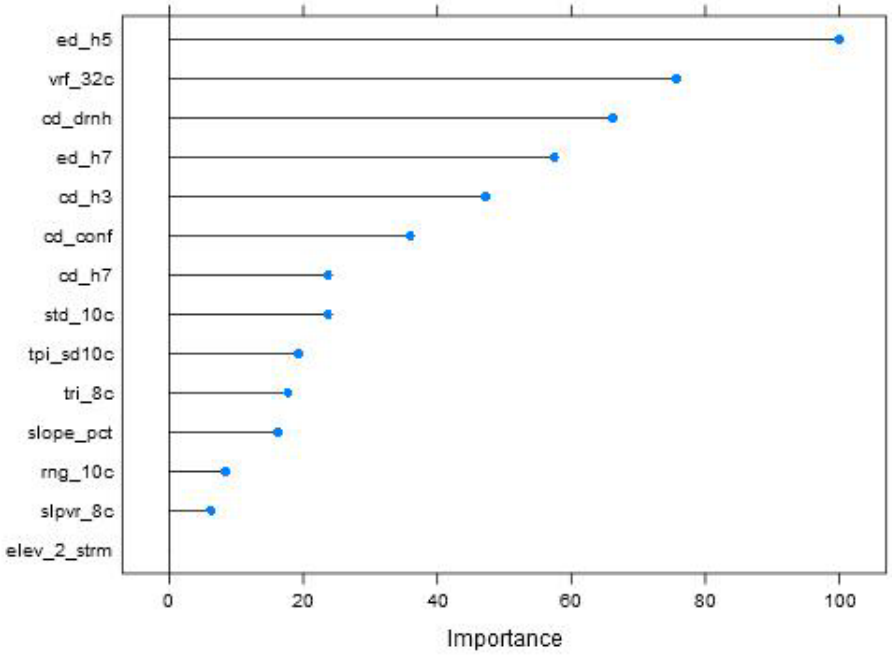


Chart 11. Region 1 West - Riverine Section 1

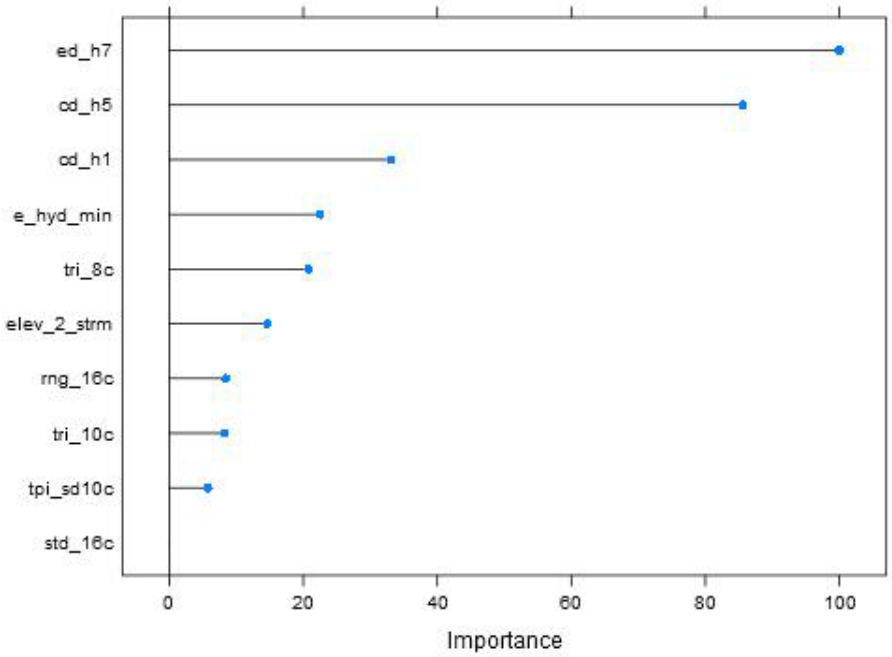


Chart 12. Region 1 West - Riverine Section 2

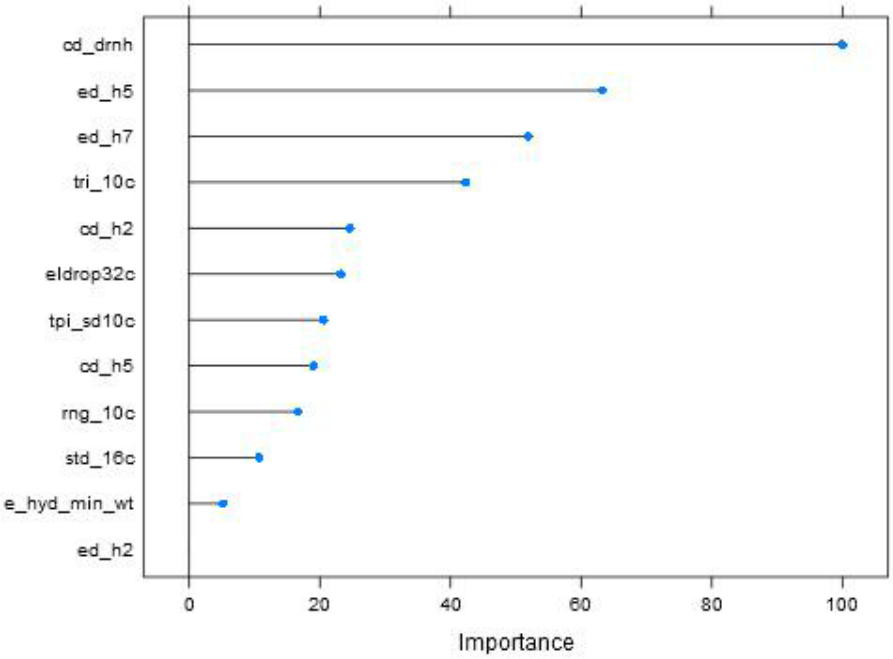


Chart 13. Region 1 West - Riverine Section 3

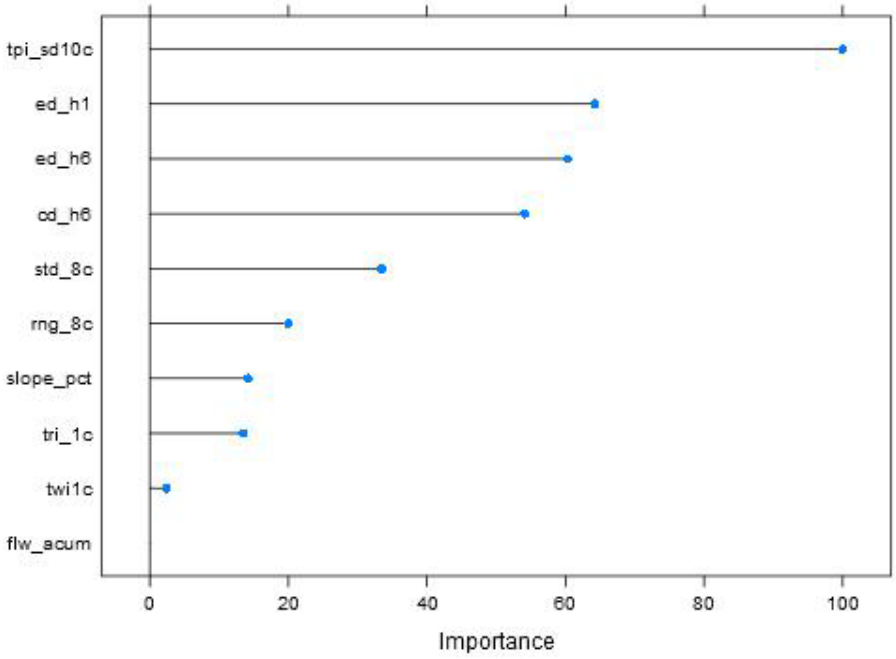


Chart 14. Region 1 West - Riverine Section 4

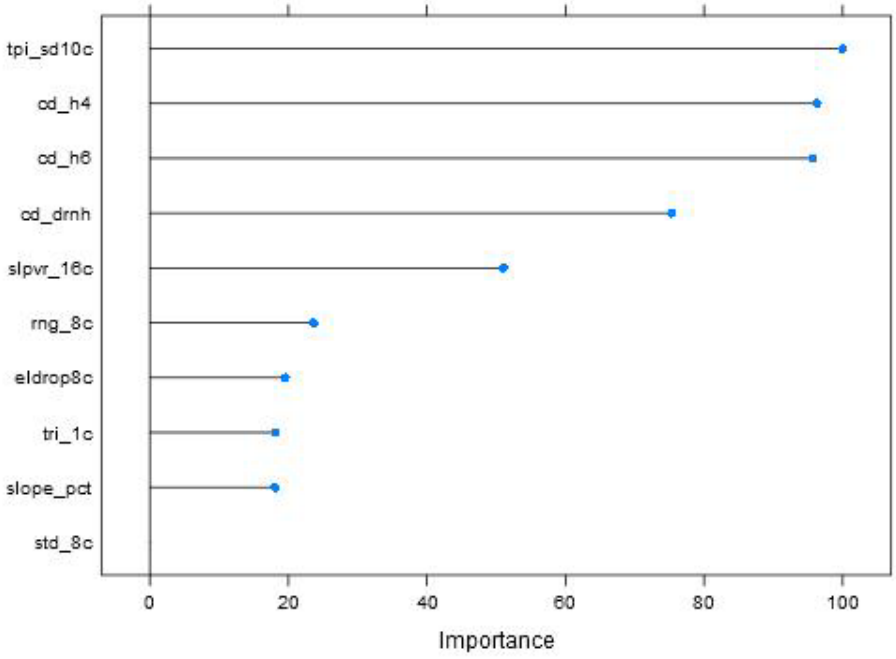


Chart 15. Region 1 West - Riverine Section 5

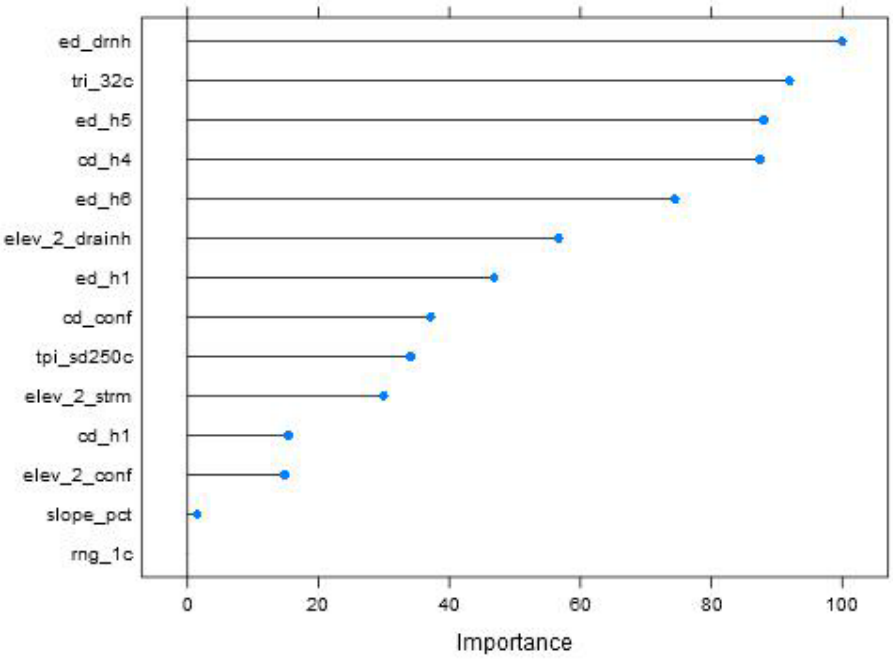


Chart 16. Region 1 West - Upland Section 1

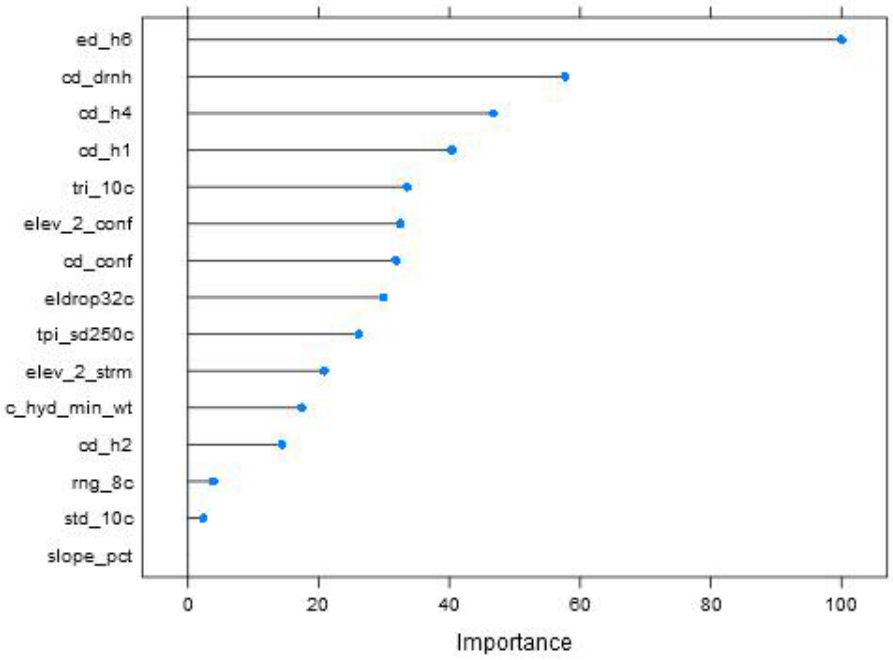


Chart 17. Region 1 West - Upland Section 2

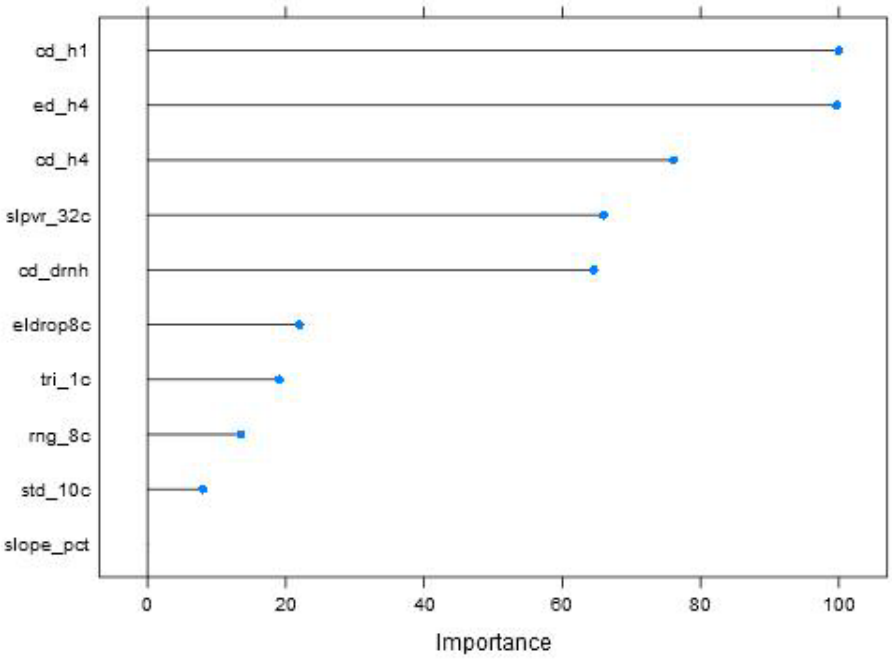


Chart 18. Region 1 West - Upland Section 3

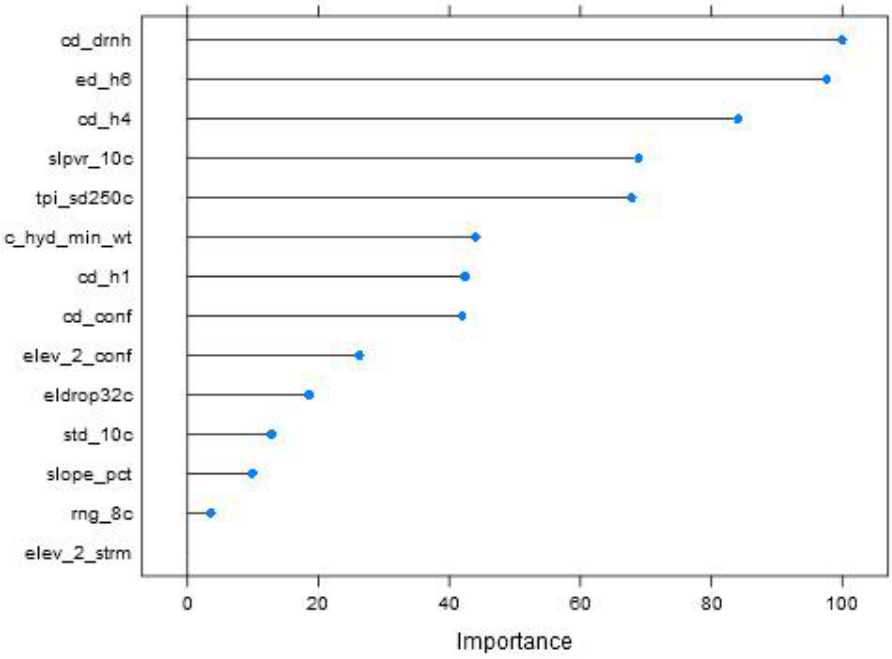


Chart 19. Region 1 West - Upland Section 4

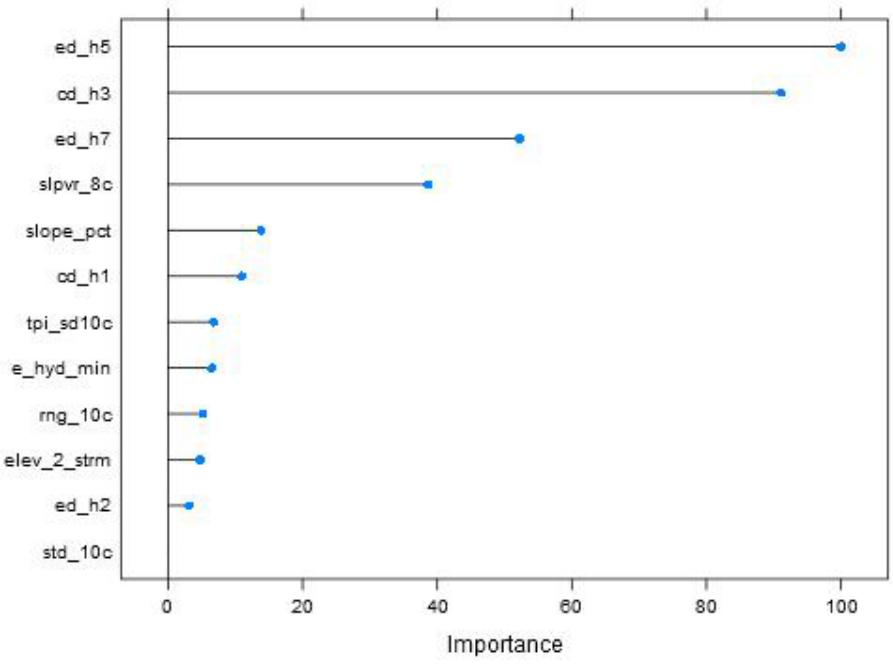


Chart 20. Region 1 West - Upland Section 5

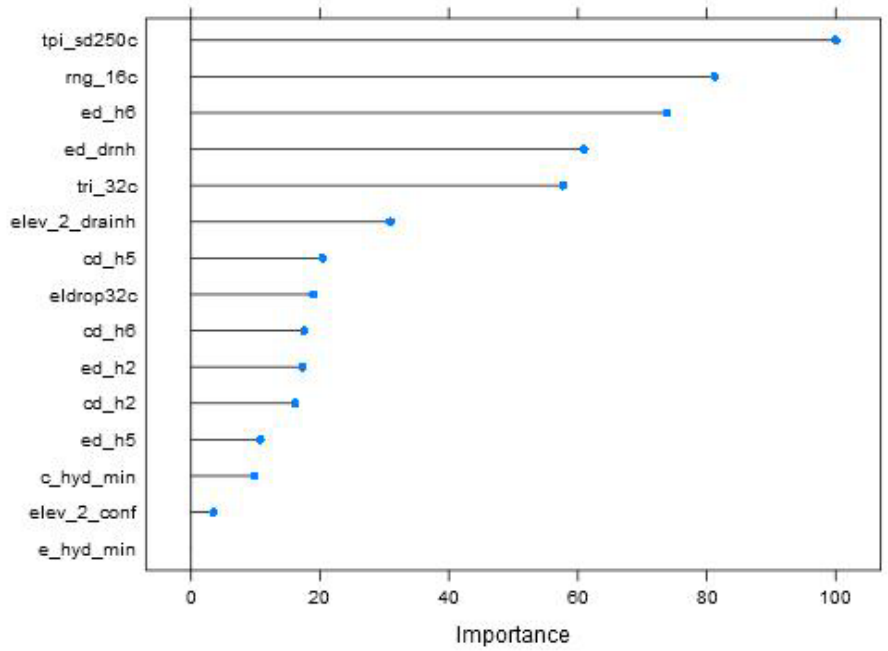


Chart 21. Region 2/3 - Riverine Section 1

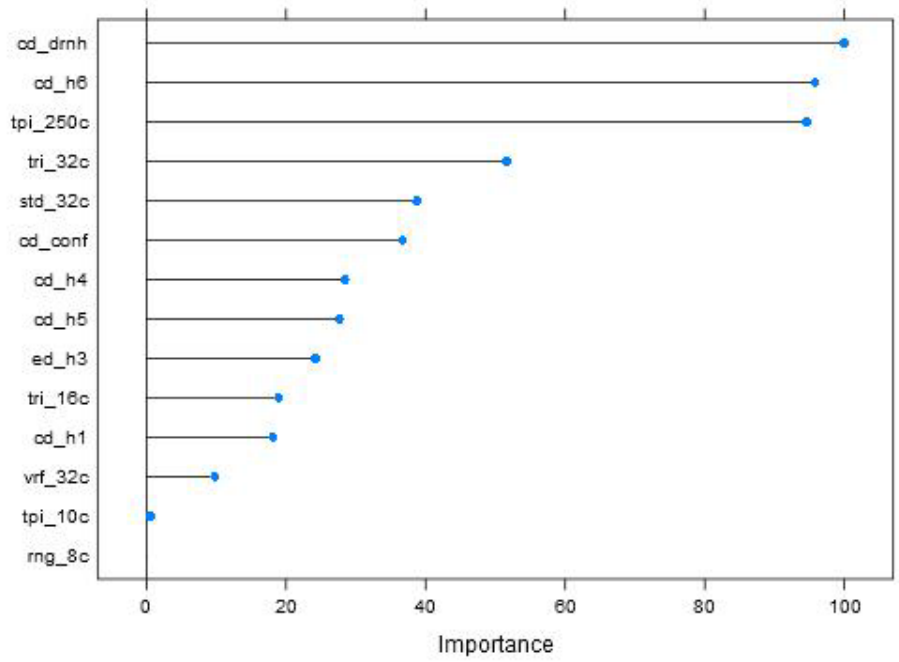


Chart 22. Region 2/3 - Riverine Section 2

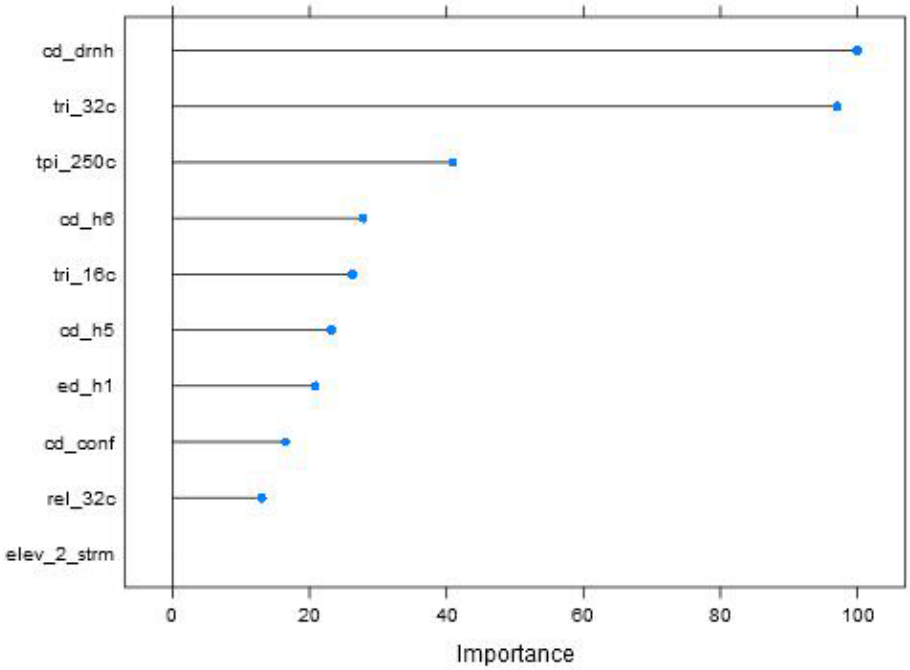


Chart 23. Region 2/3 - Riverine Section 3

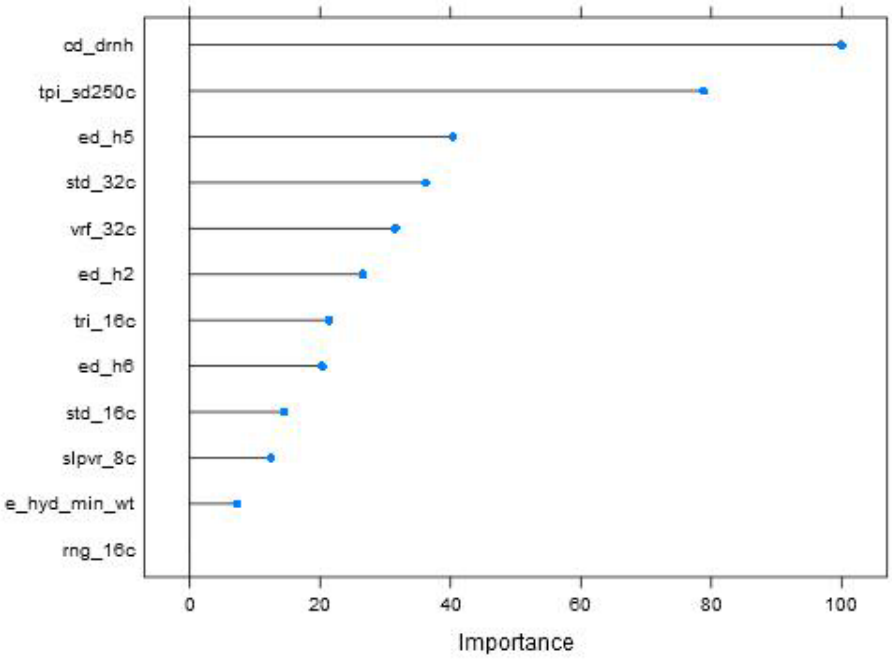


Chart 24. Region 2/3 - Riverine Section 4

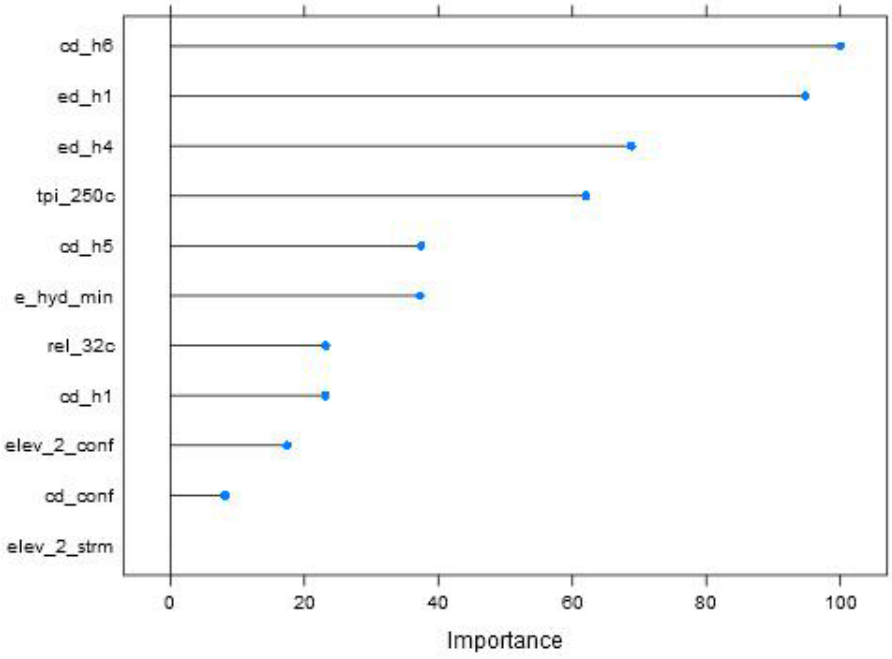


Chart 25. Region 2/3 - Riverine Section 5

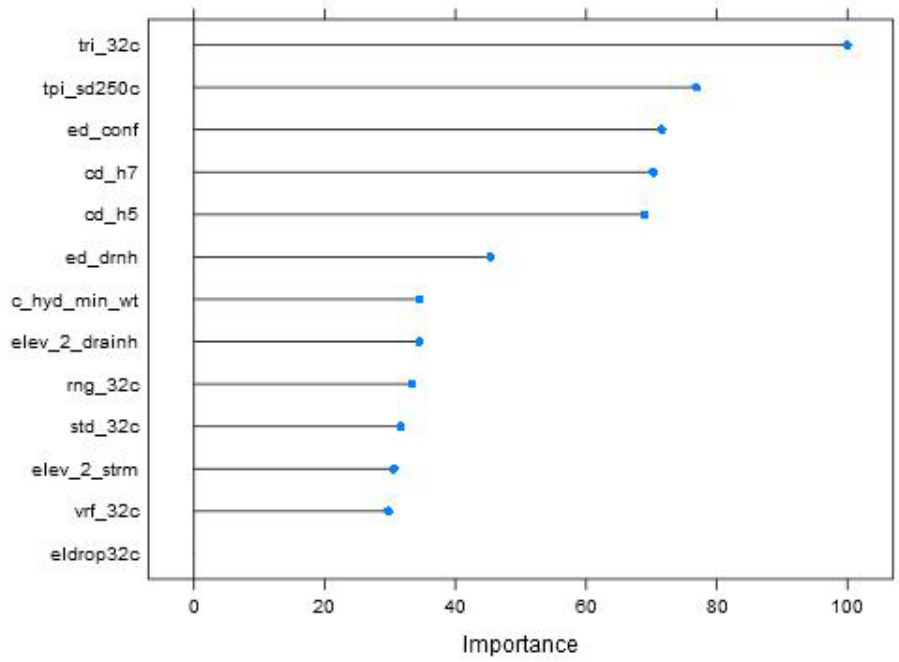


Chart 26. Region 2/3 - Upland Section 1

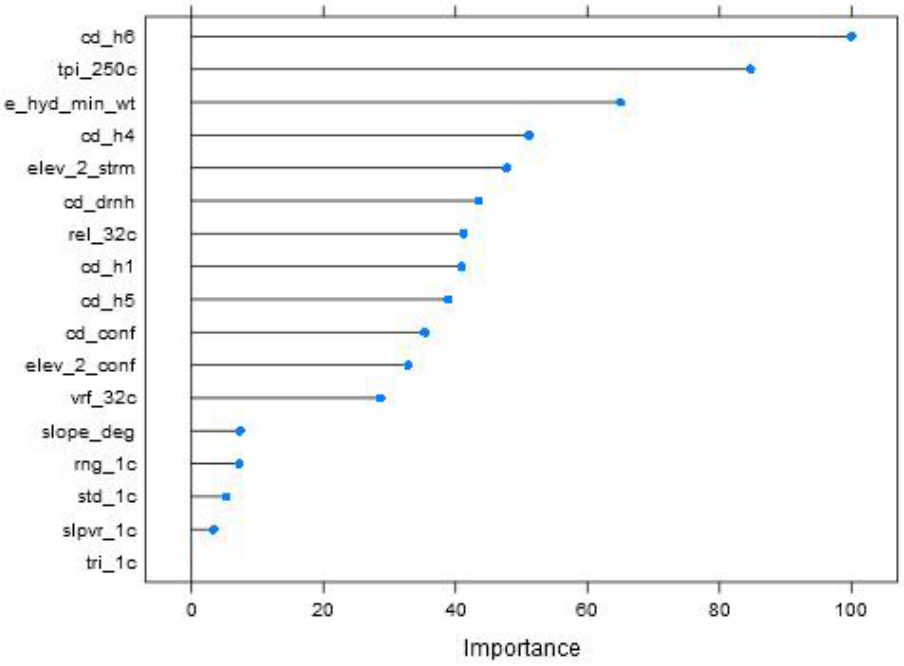


Chart 27. Region 2/3 - Upland Section 2

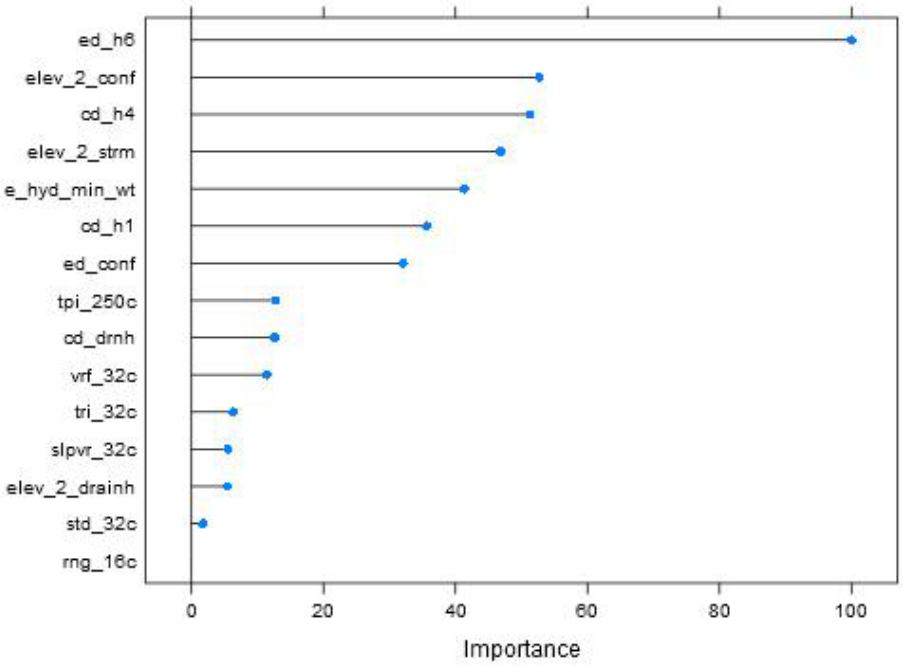


Chart 28. Region 2/3 - Upland Section 3

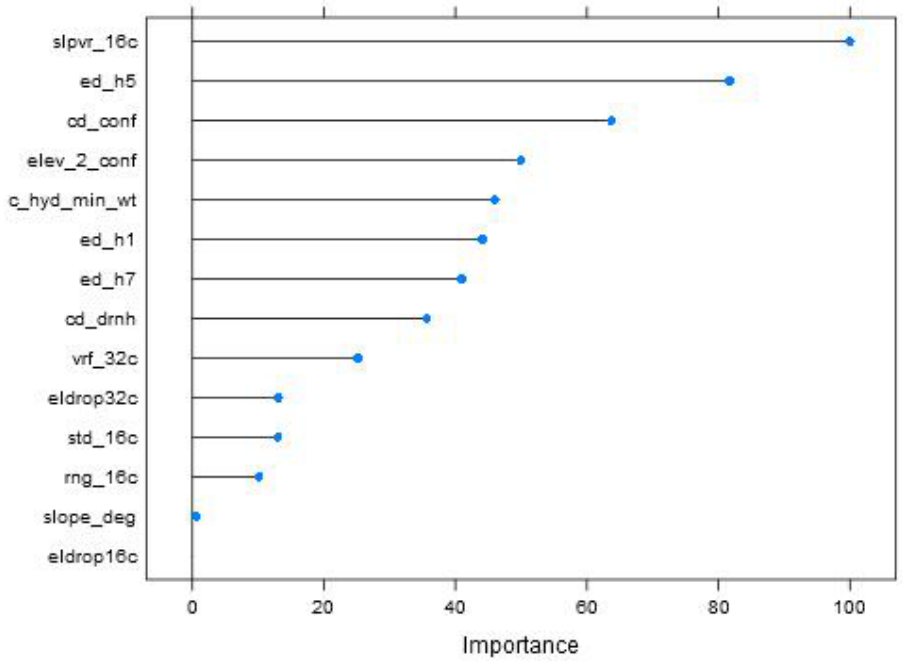


Chart 29. Region 2/3 - Upland Section 4 without Rock Shelters

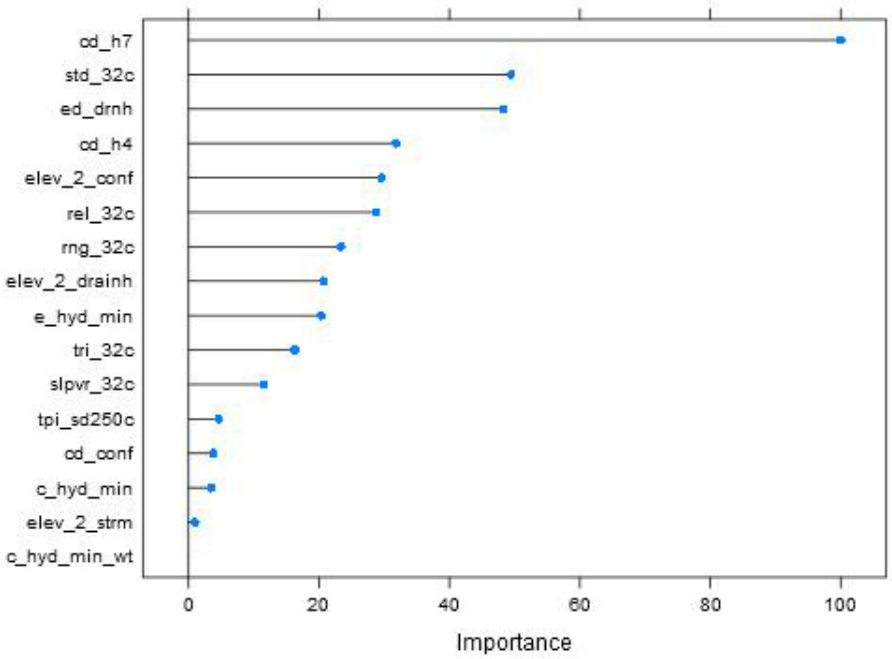


Chart 30. Region 2/3 - Upland Section 4 Rock Shelters

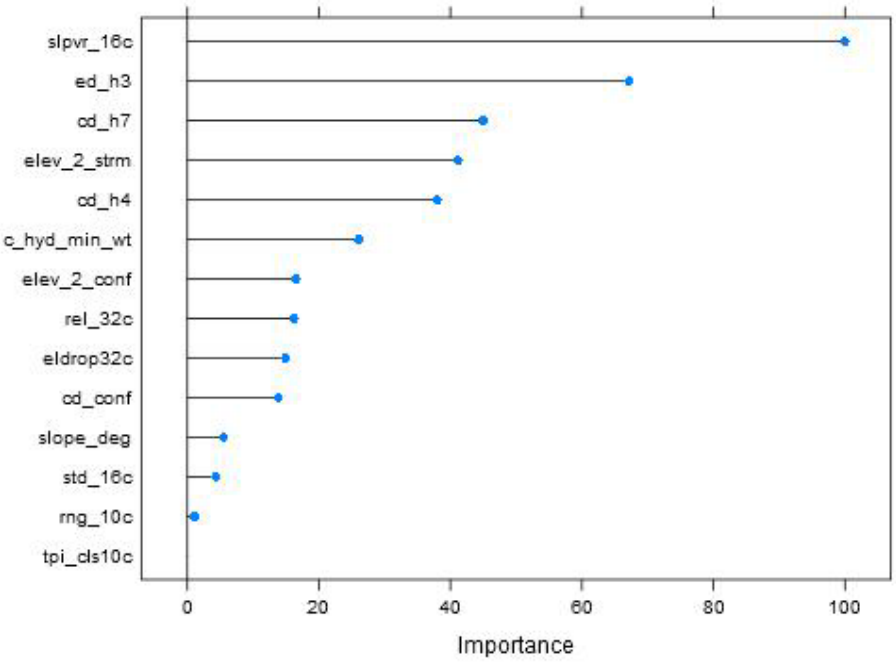


Chart 31. Region 2/3 - Upland Section 5 without Rock Shelters

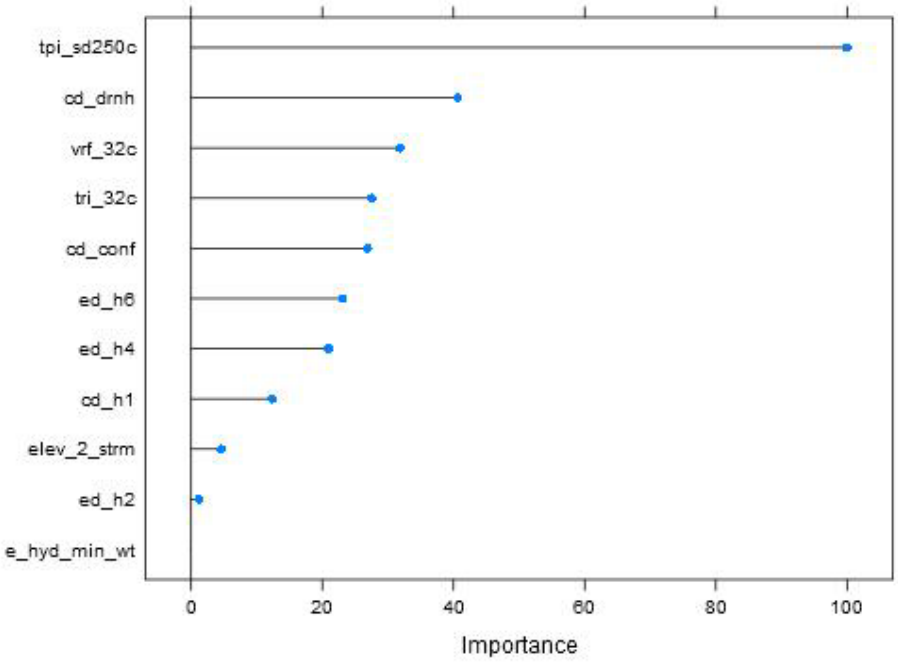
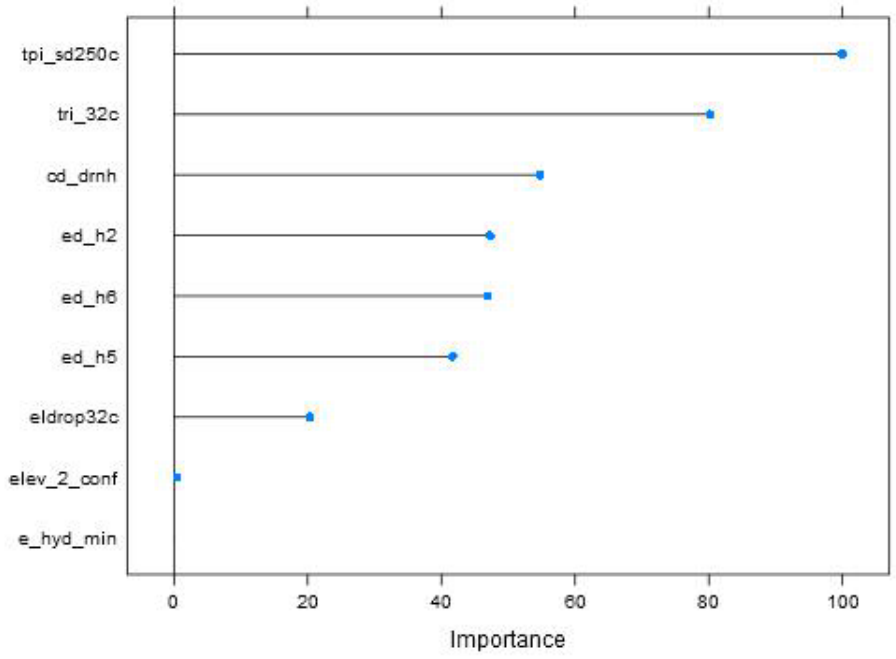


Chart 32. Region 2/3 - Upland Section 5 Rock Shelters



APPENDIX E

POTENTIAL THRESHOLDS

FOR EACH OF 30 MODELS

WITHIN REGIONS 1, 2, AND 3

Chart 1. Region 1 East - Riverine Section 1

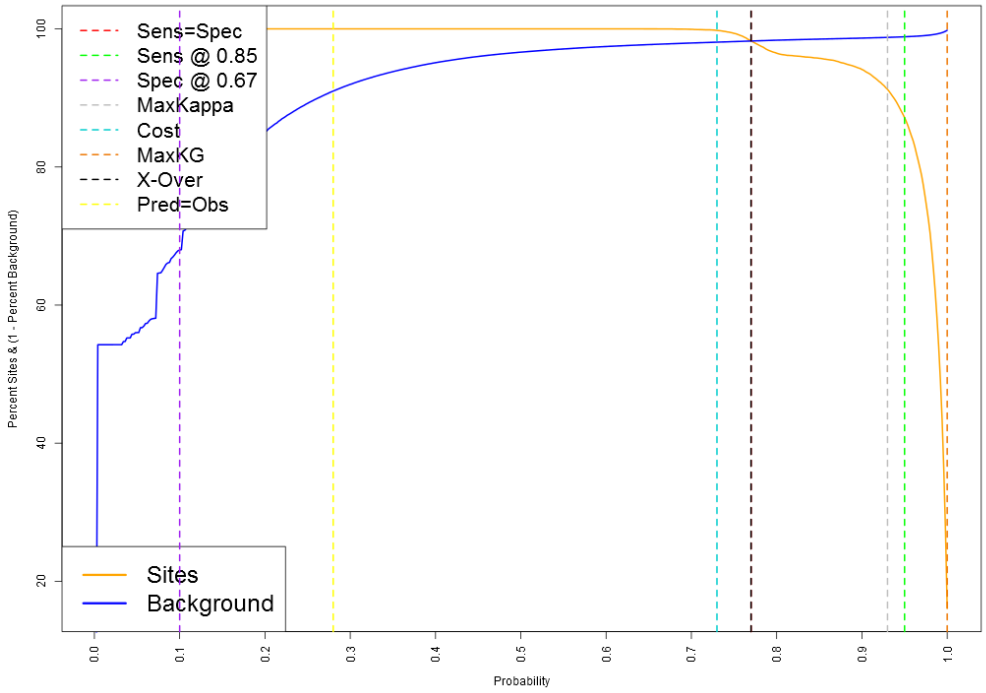


Chart 2. Region 1 East - Riverine Section 2

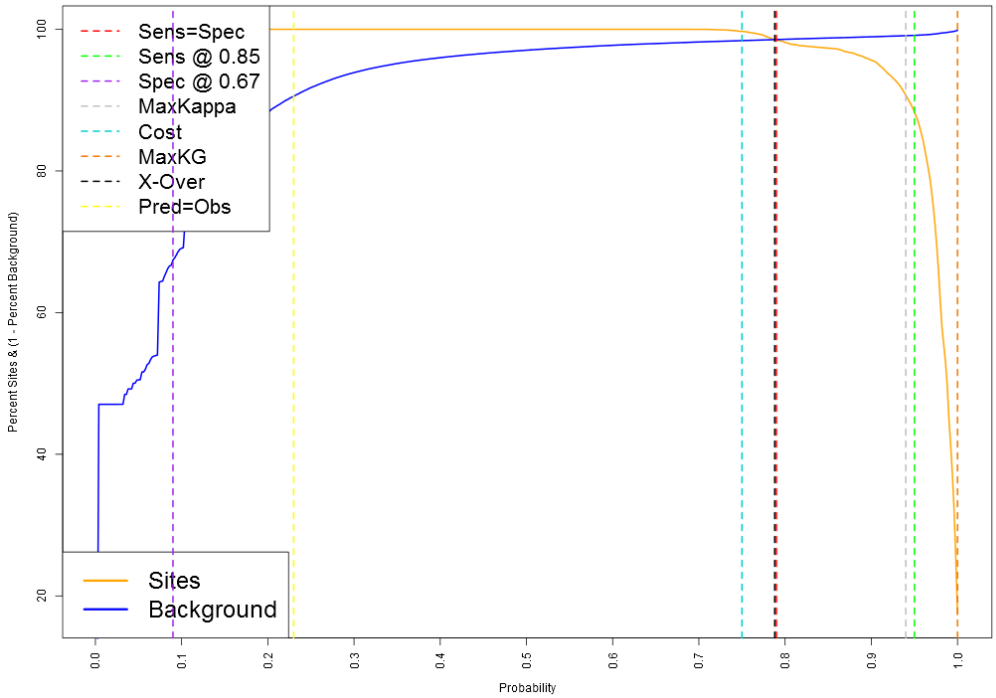


Chart 3. Region 1 East - Riverine Section 3

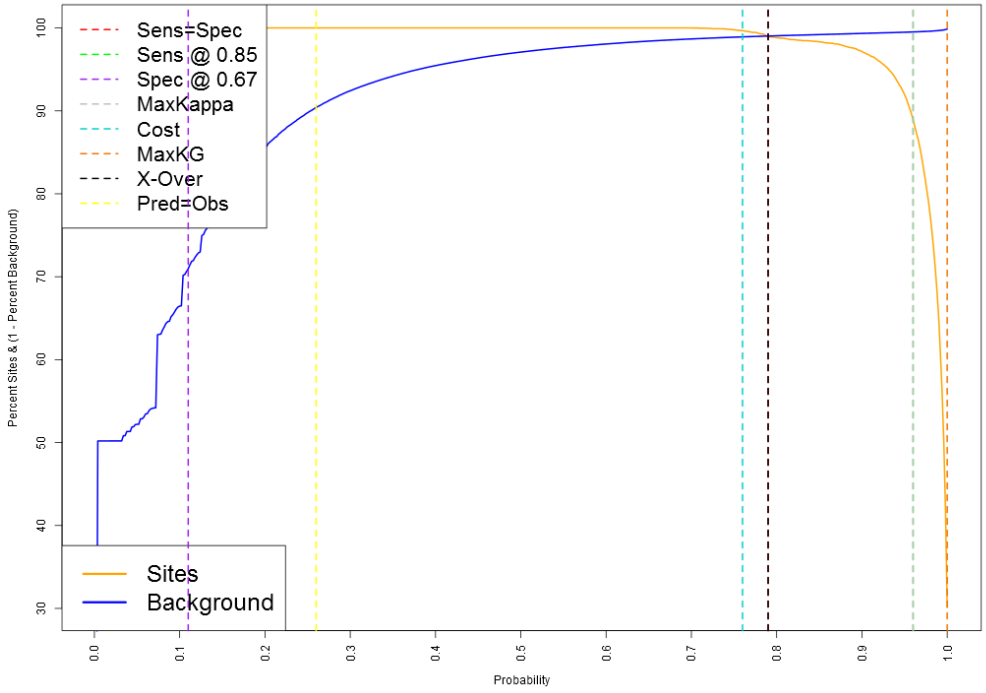


Chart 4. Region 1 East - Upland Section 1

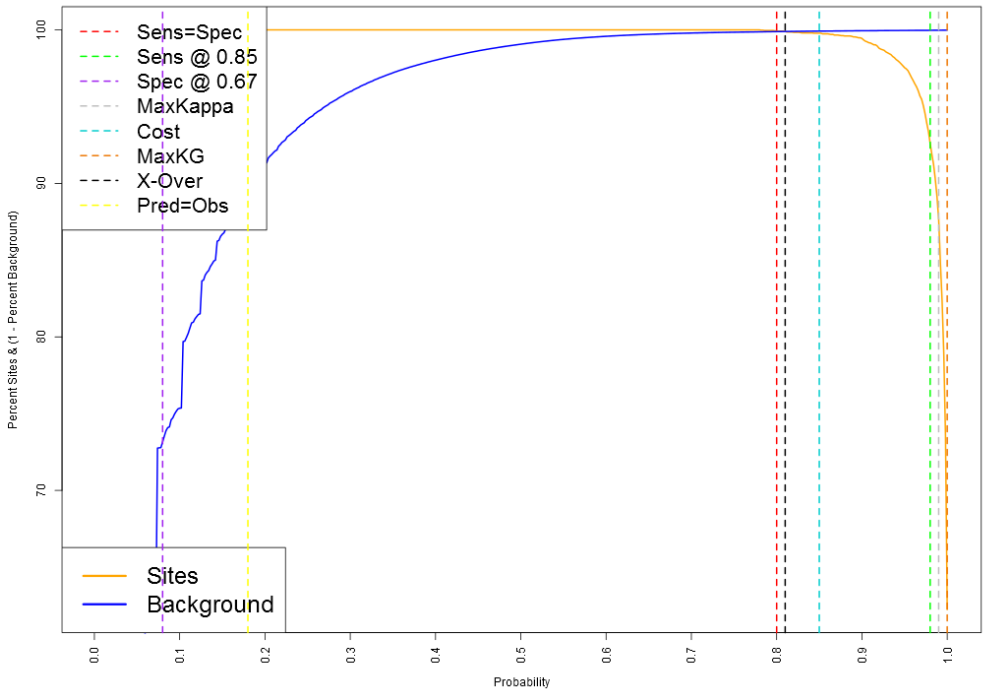


Chart 5. Region 1 East - Upland Section 2

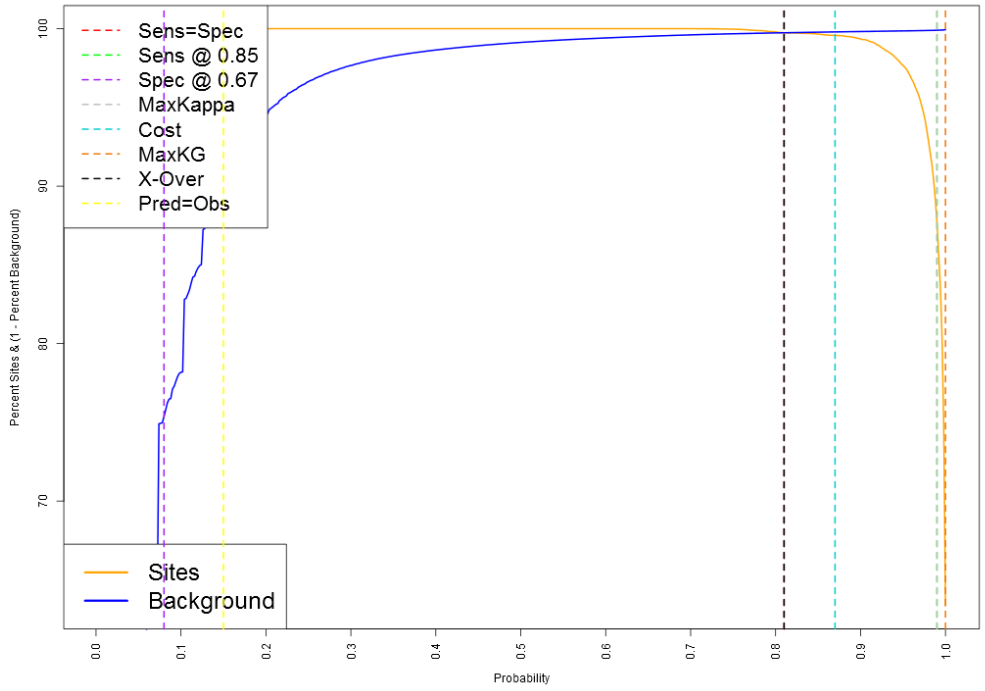


Chart 6. Region 1 East - Upland Section 3

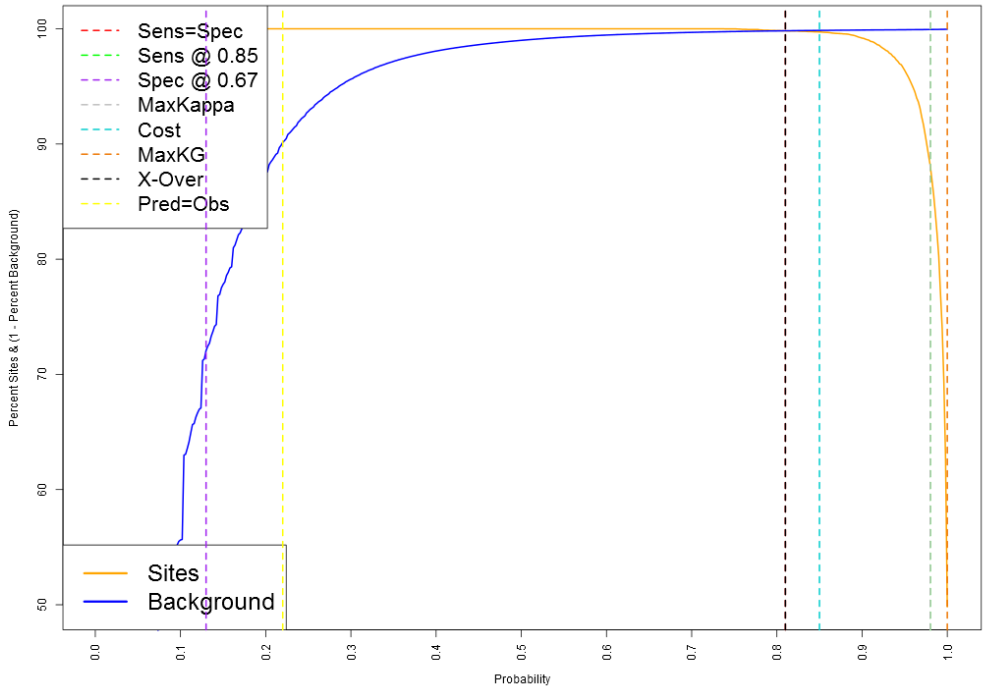


Chart 7. Region 1 North - Riverine Section 1

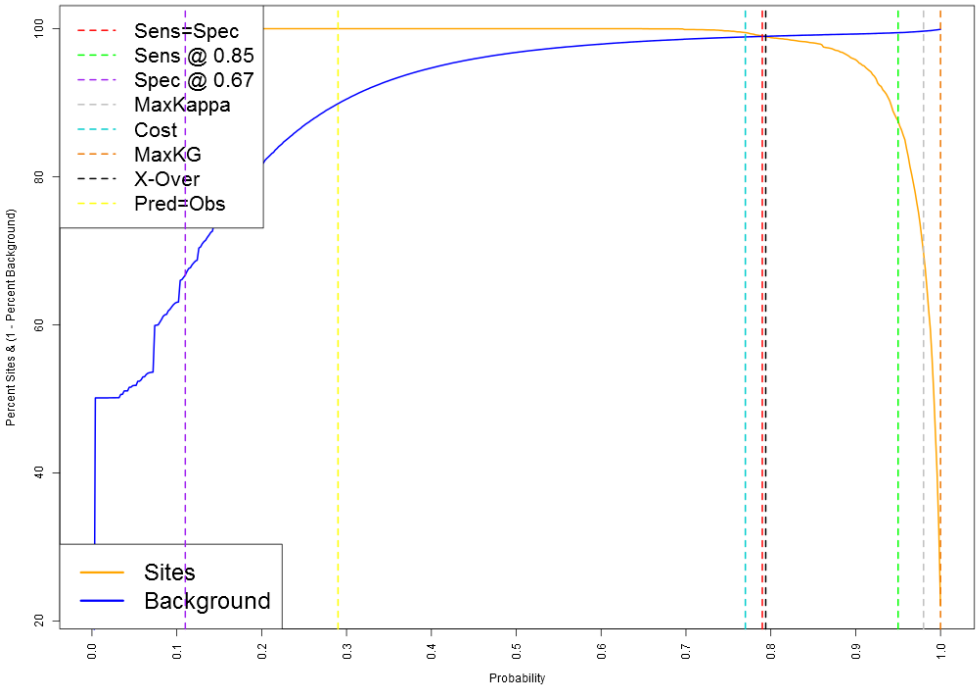


Chart 8. Region 1 North - Riverine Section 2

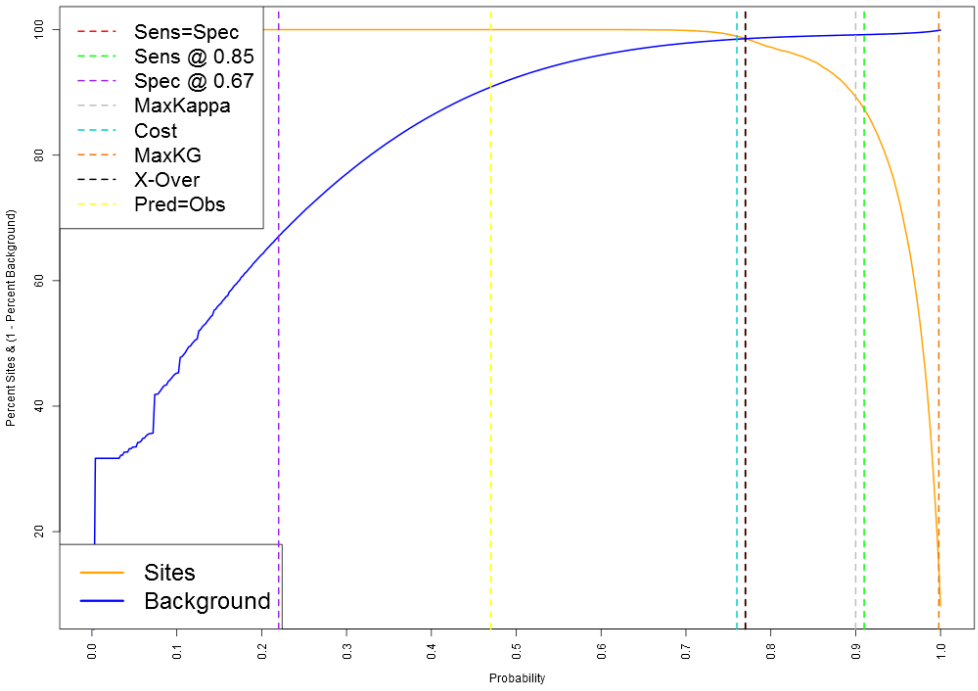


Chart 9. Region 1 North - Upland Section 1

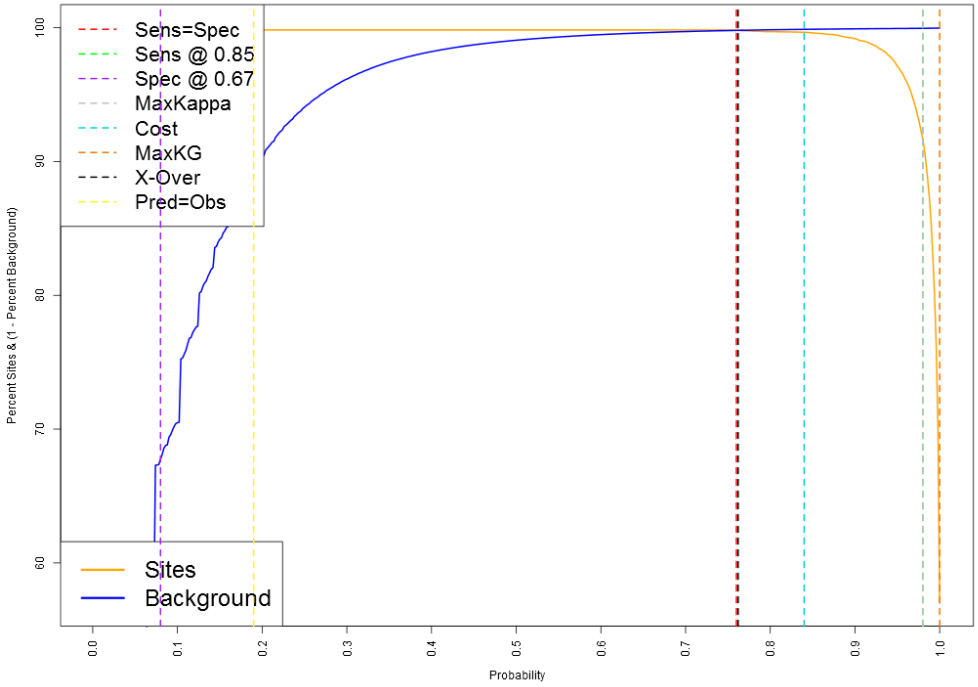


Chart 10. Region 1 North - Upland Section 2

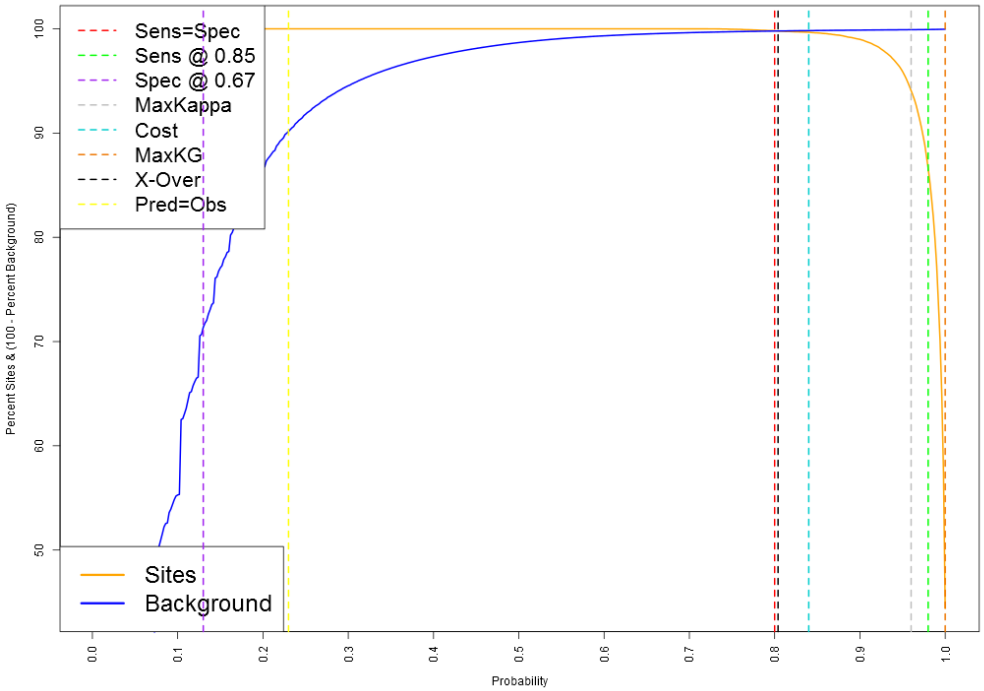


Chart 11. Region 1 West - Riverine Section 1

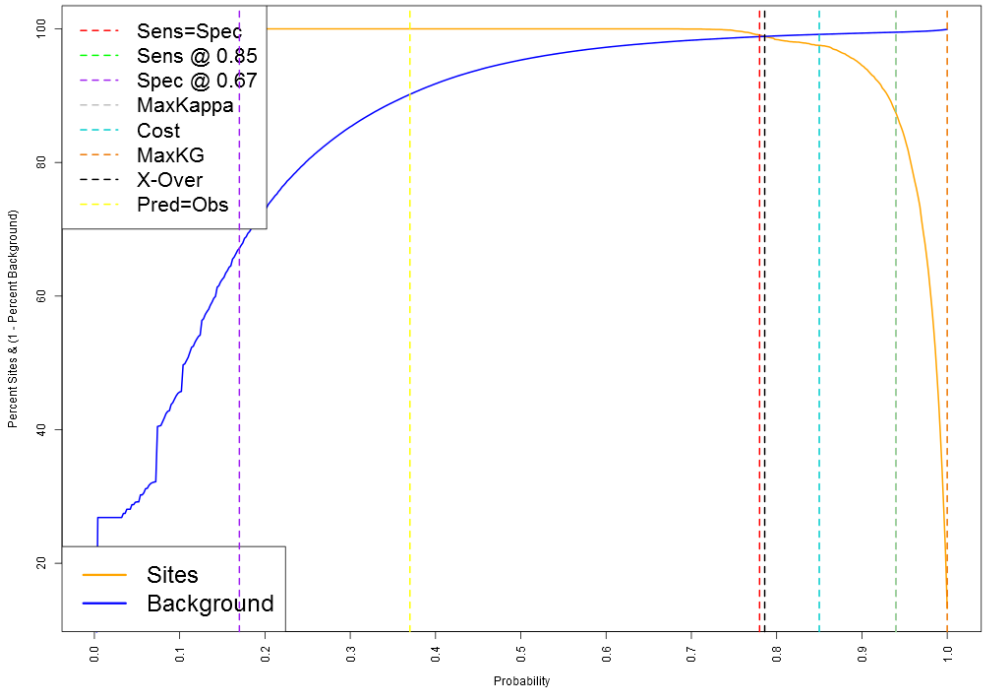


Chart 12. Region 1 West - Riverine Section 2

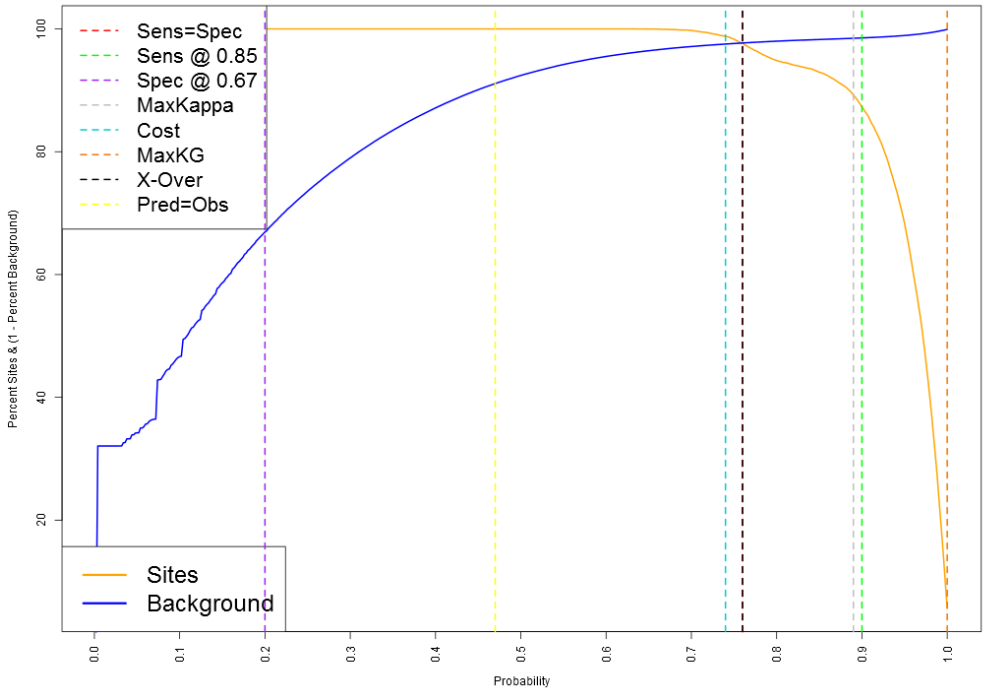


Chart 13. Region 1 West - Riverine Section 3

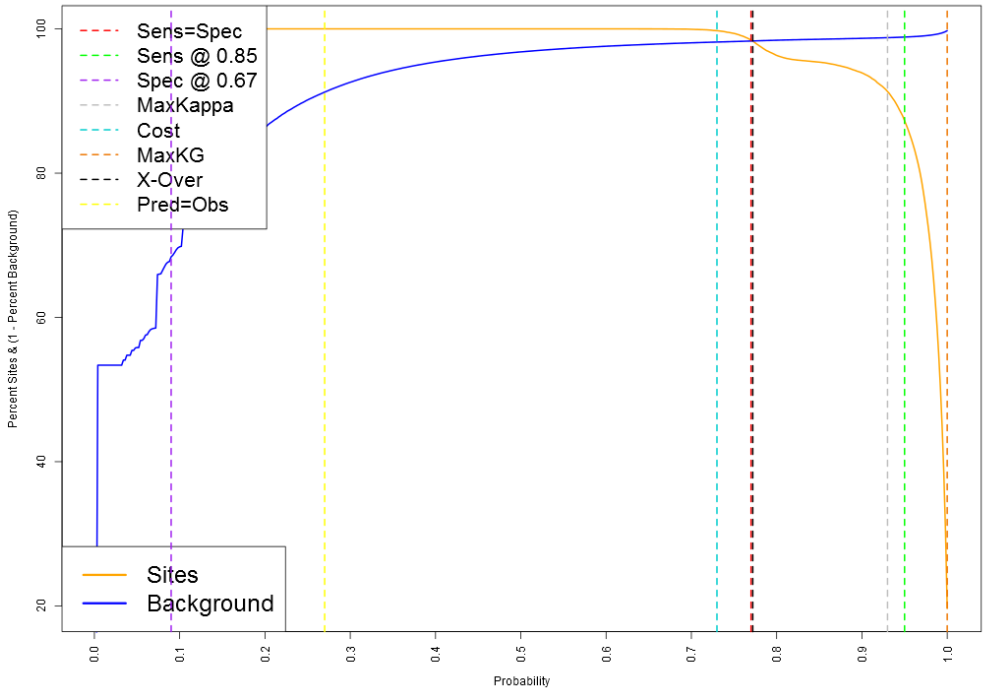


Chart 14. Region 1 West - Riverine Section 4

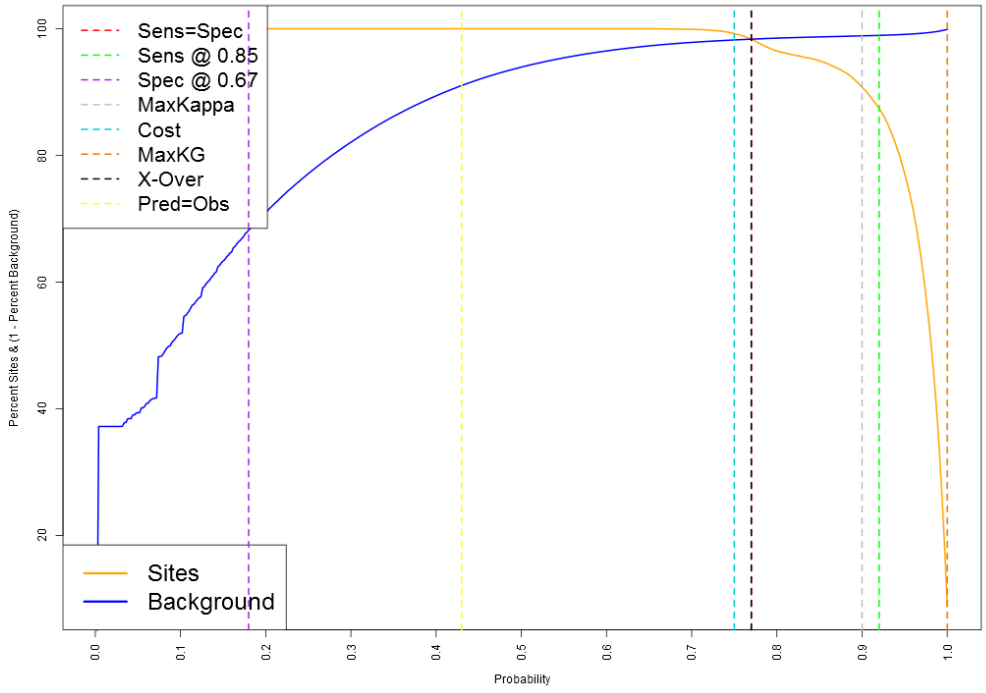


Chart 15. Region 1 West - Riverine Section 5

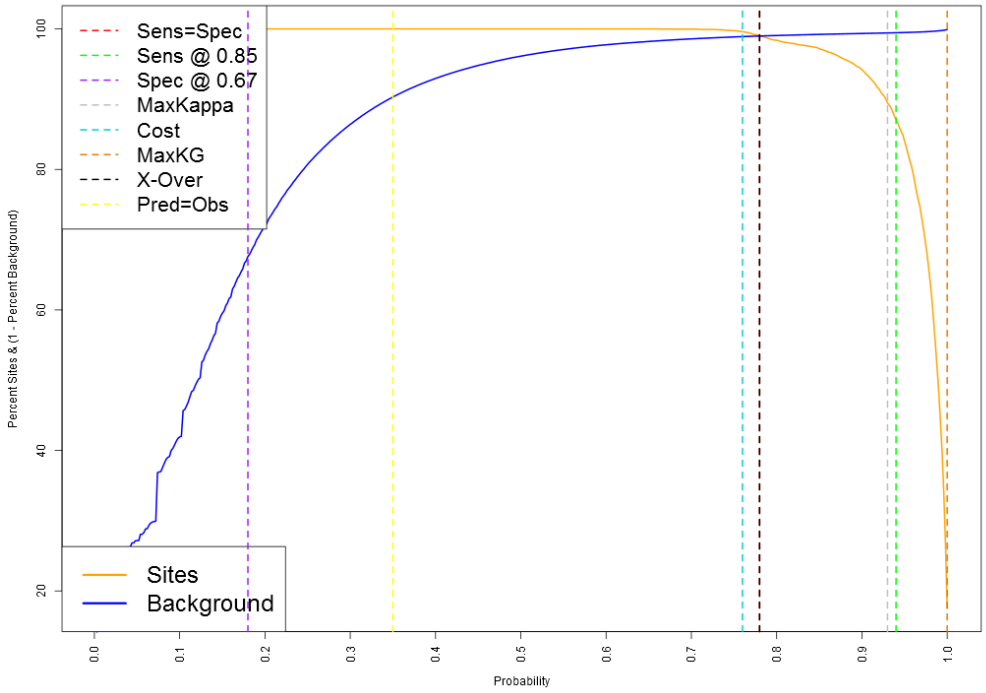


Chart 16. Region 1 West - Upland Section 1

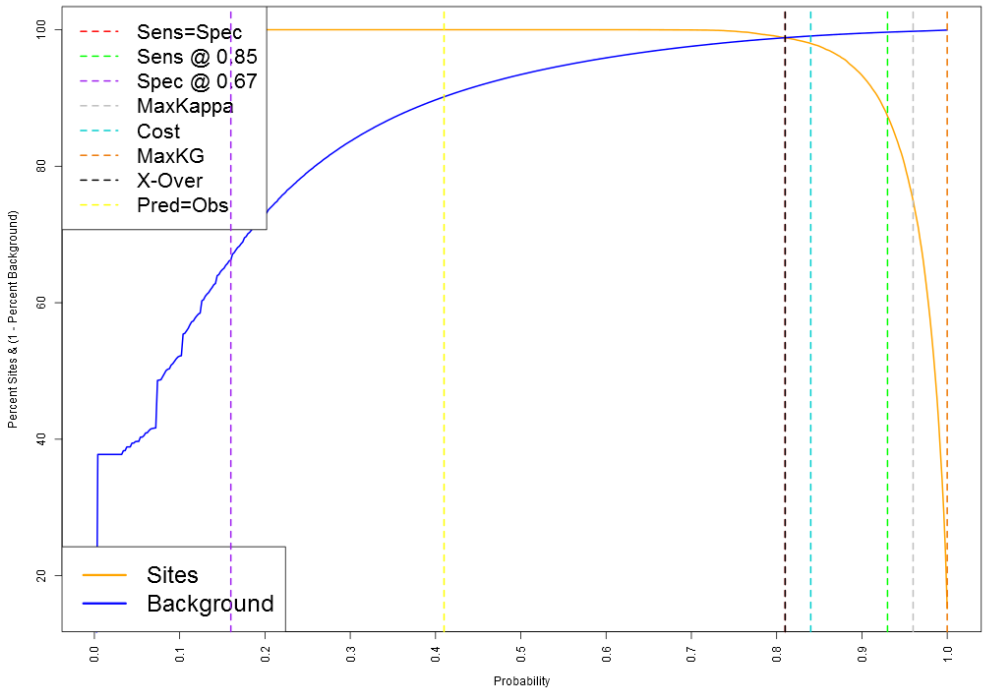


Chart 17. Region 1 West - Upland Section 2

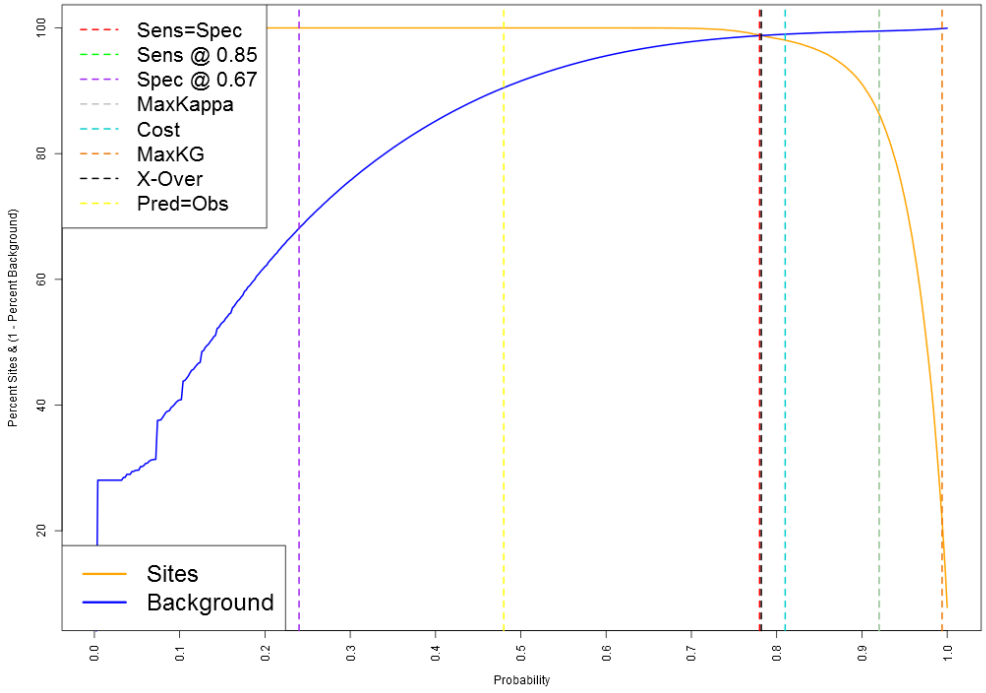


Chart 18. Region 1 West - Upland Section 3

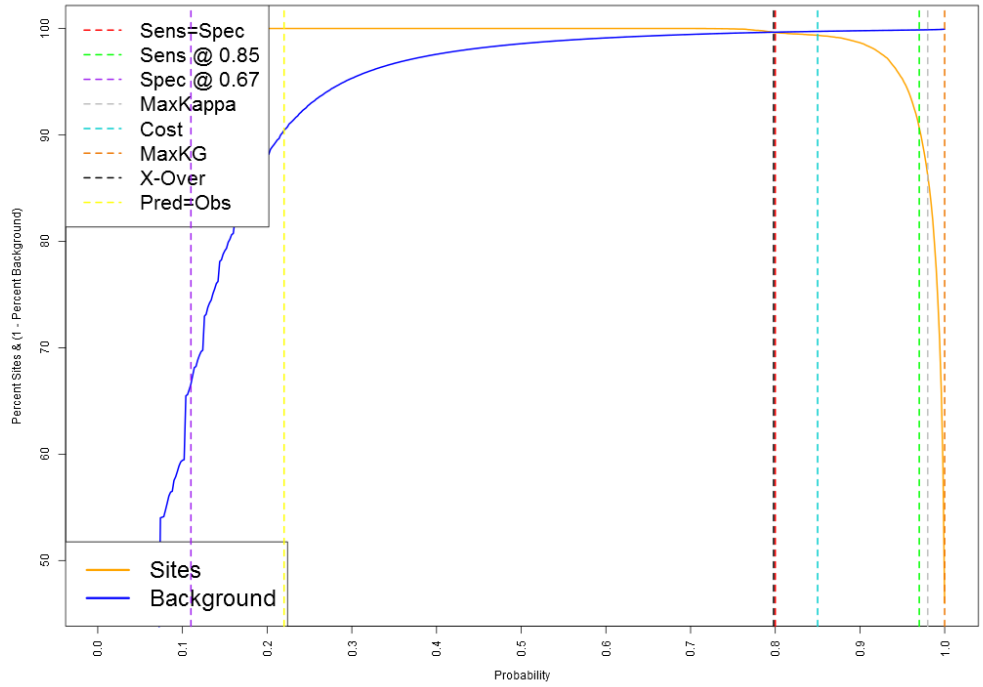


Chart 19. Region 1 West - Upland Section 4

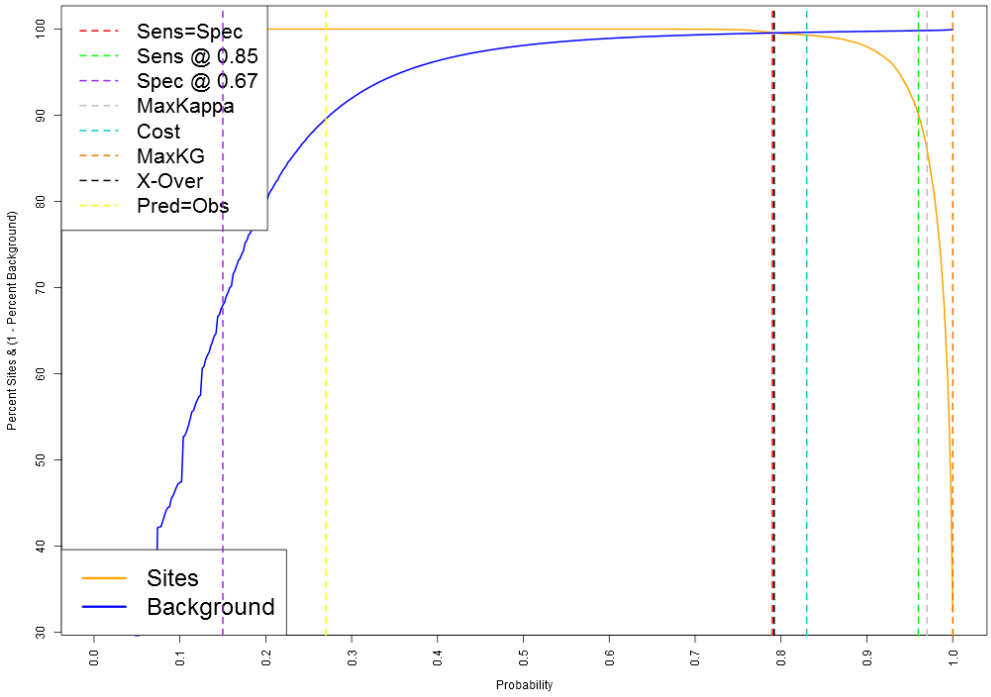


Chart 20. Region 1 West - Upland Section 5

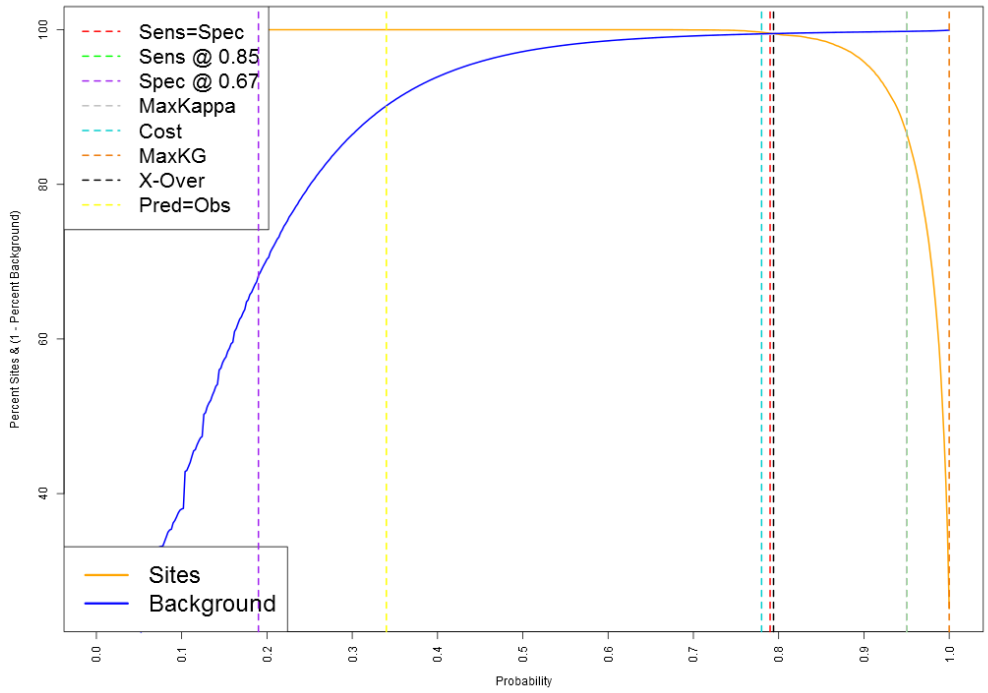


Chart 21. Region 2/3 - Riverine Section 1

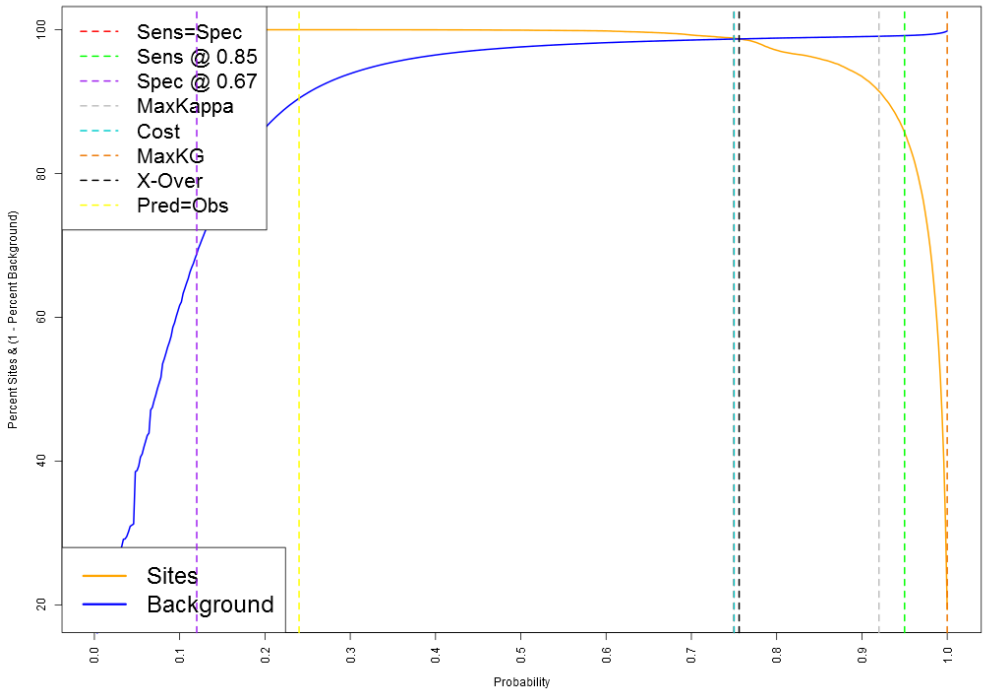


Chart 22. Region 2/3 - Riverine Section 2

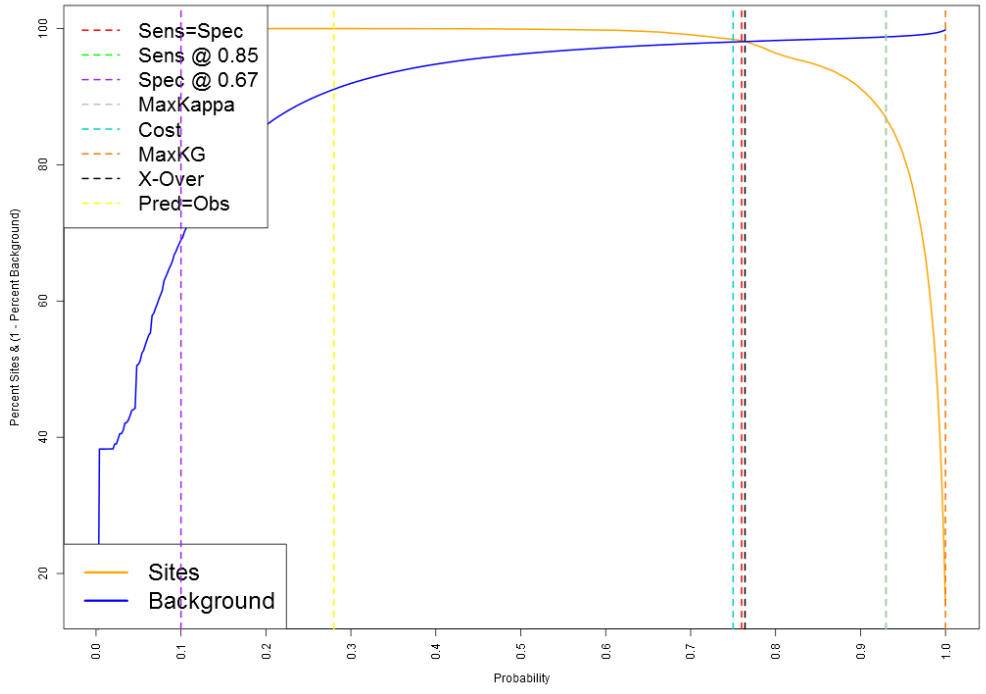


Chart 23. Region 2/3 - Riverine Section 3

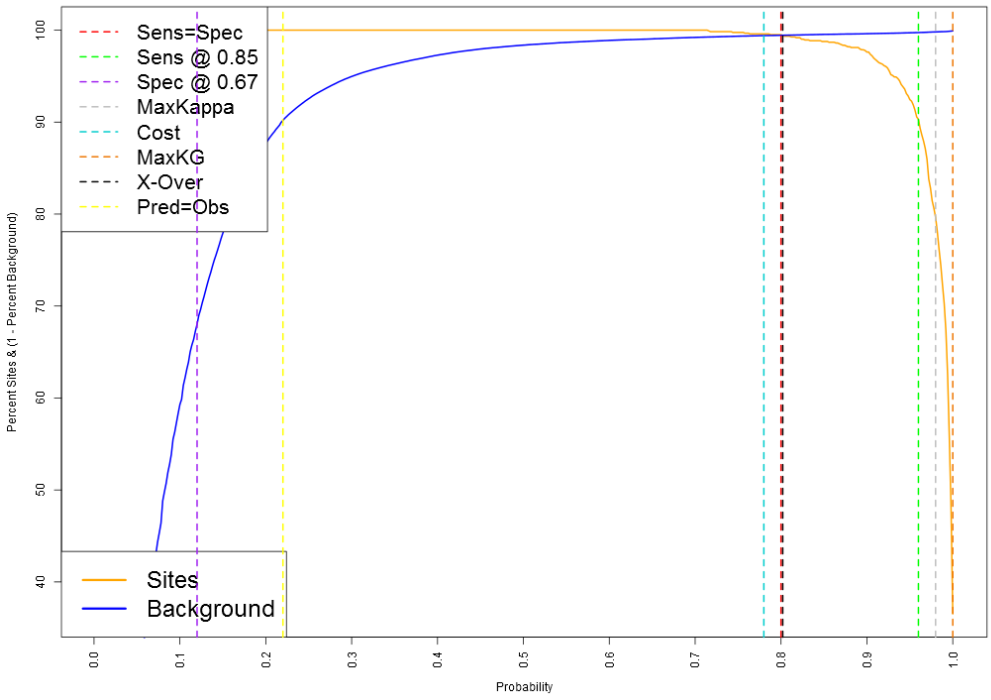


Chart 24. Region 2/3 - Riverine Section 4

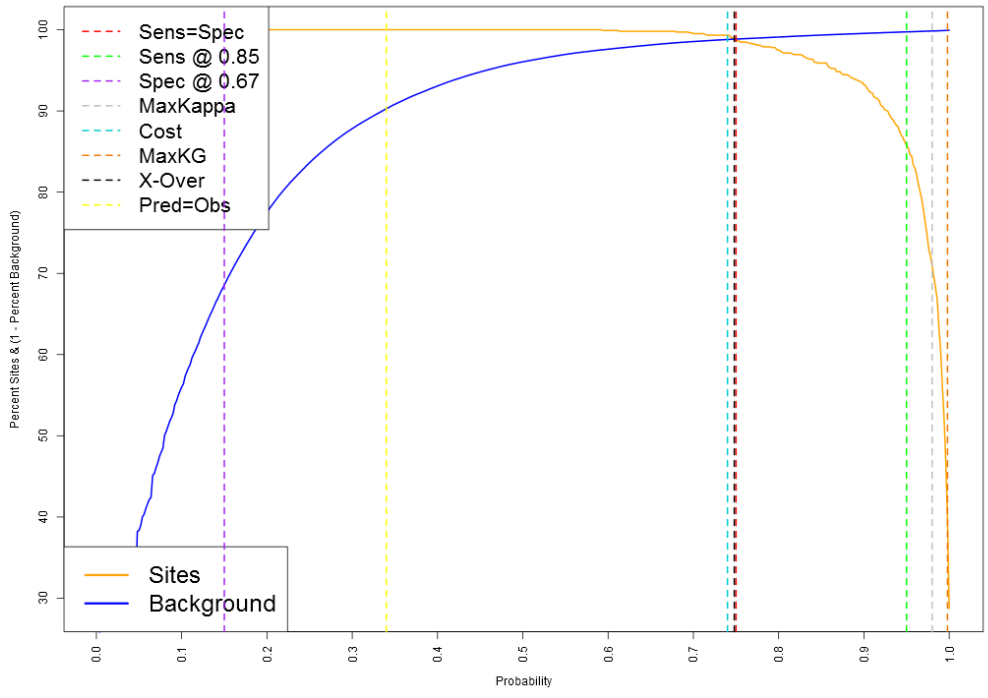


Chart 25. Region 2/3 - Riverine Section 5

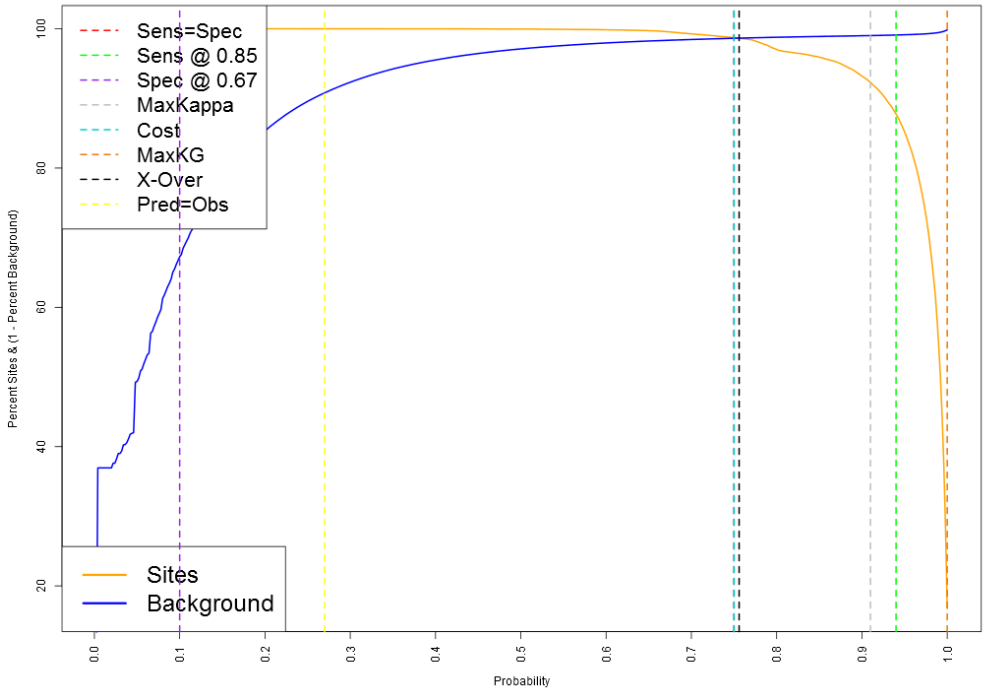


Chart 26. Region 2/3 - Upland Section 1

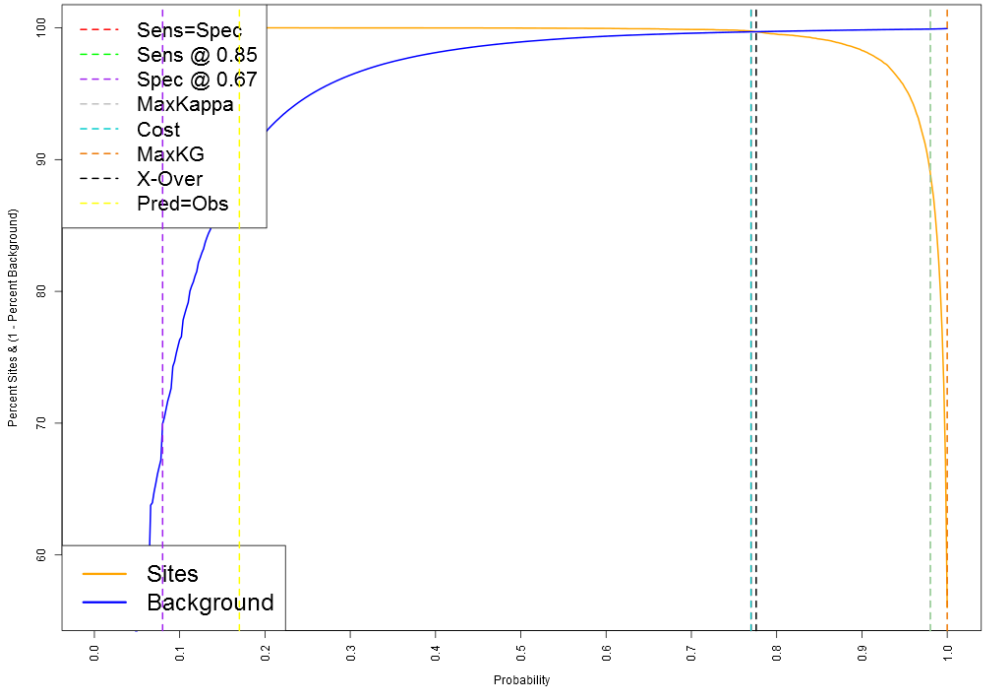


Chart 27. Region 2/3 - Upland Section 2

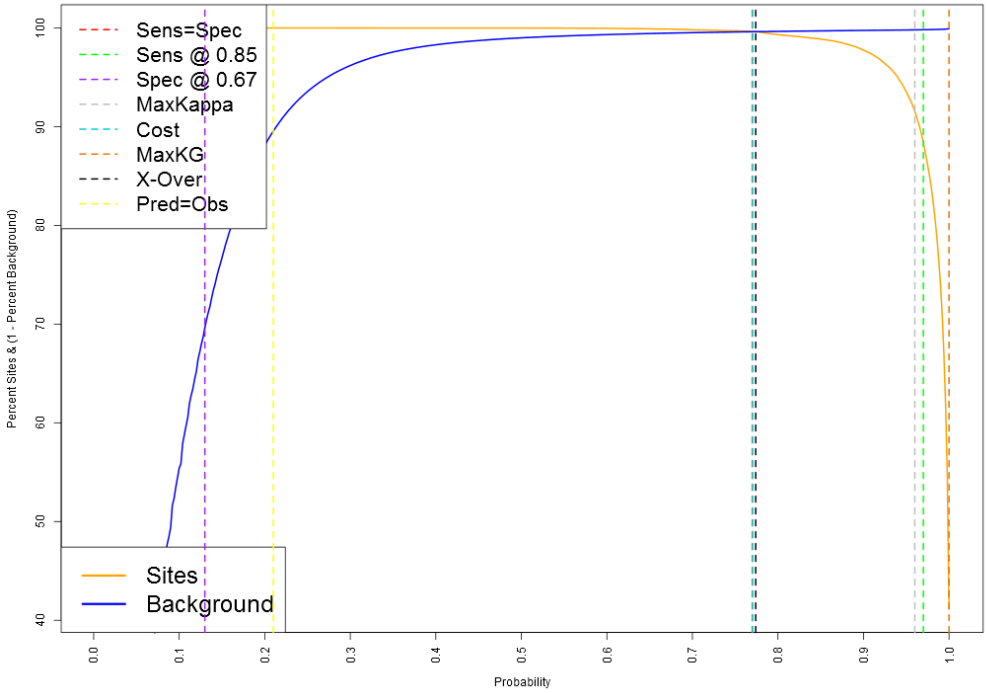


Chart 28. Region 2/3 - Upland Section 3

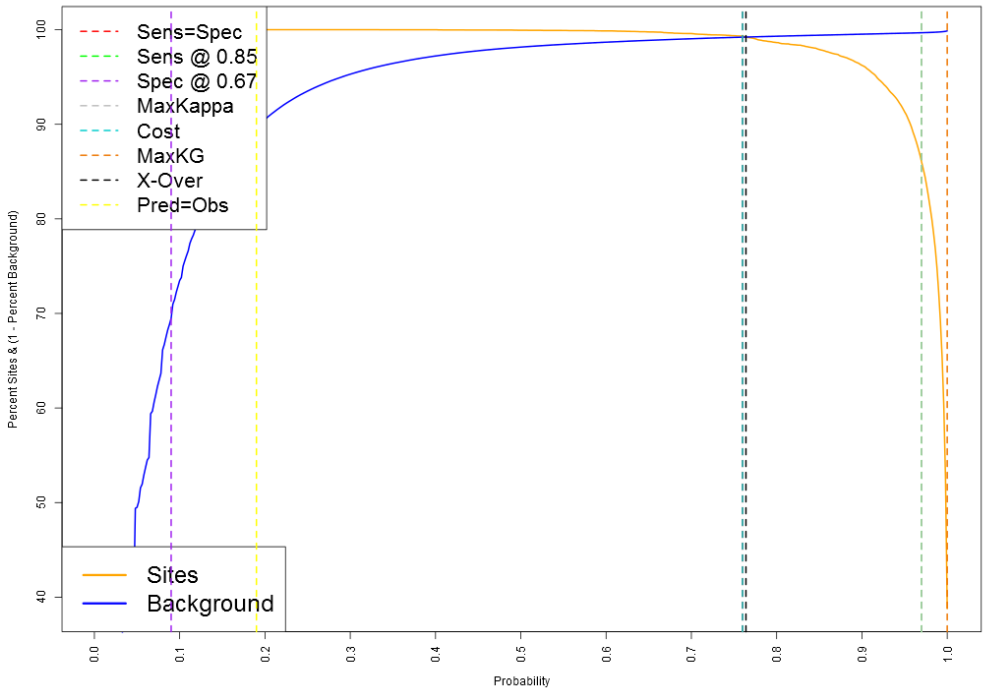


Chart 29. Region 2/3 - Upland Section 4
(combined rock shelter and non-rock shelter models)

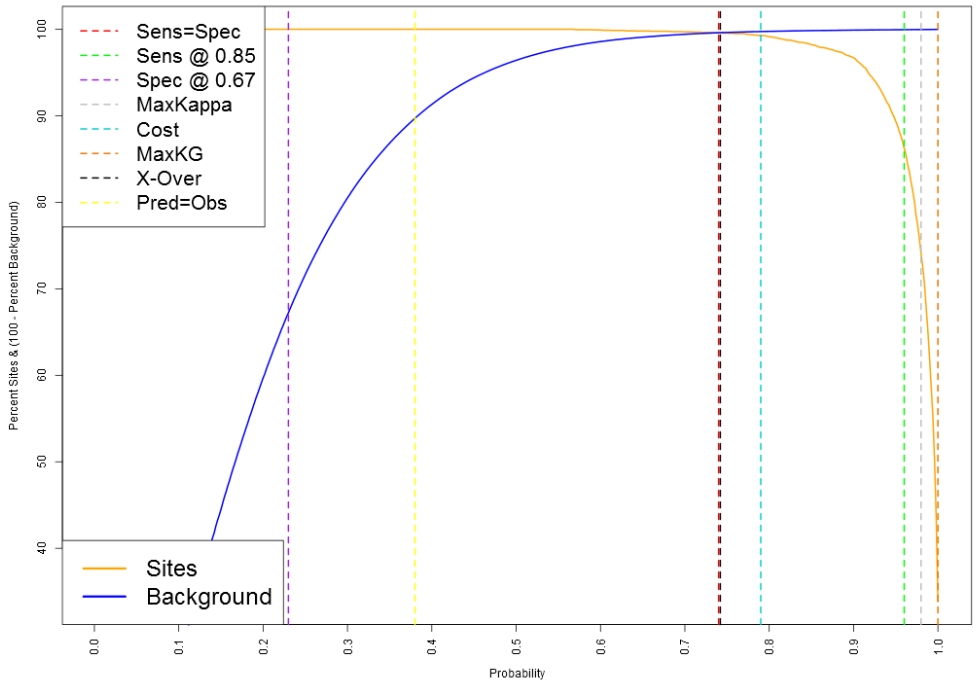
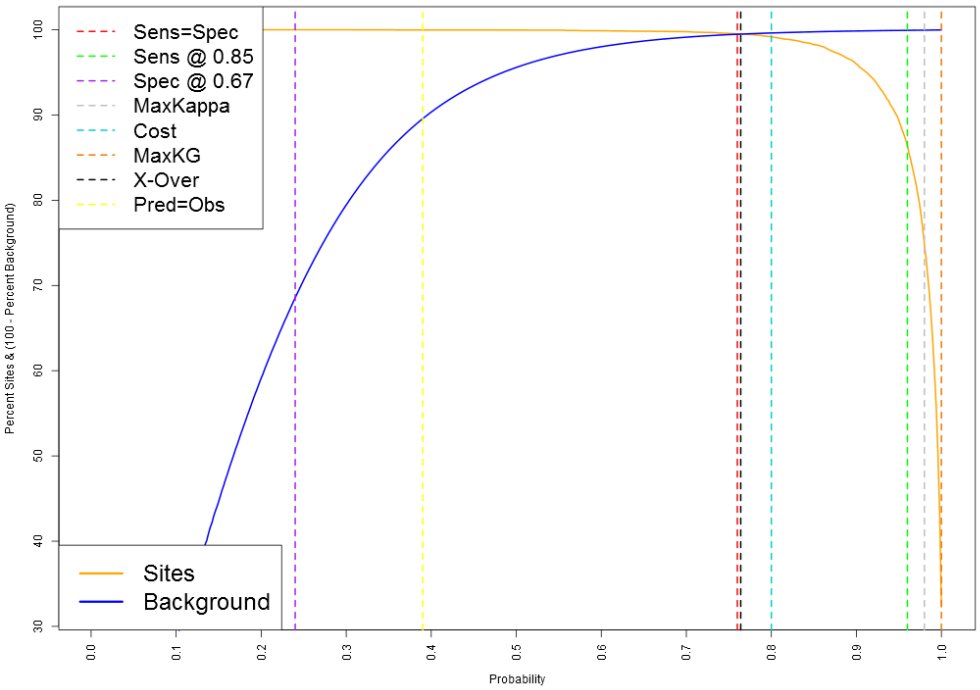


Chart 30. Region 2/3 - Upland Section 5
(combined rock shelter and non-rock shelter models)



APPENDIX F

CONFUSION MATRICES

FOR EACH OF 30 MODELS

WITHIN REGIONS 1, 2, AND 3

Region 1 East - Riverine Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	19374	500635	520009
	Absent	0	1062742	1062742
		19374	1563377	1582751

Sensitivity / TPR =	1.000
Specificity / TNR =	0.680
Prevalence =	0.0122
Kvamme Gain (Kg) =	0.671
Accuracy =	0.684
Positive Prediction Value (PPV) =	0.037
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.012
Positive Prediction Gain (PPG) =	3.044
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.329

Region 1 East - Riverine Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	11090	423157	434247
	Absent	0	875736	875736
		11090	1298893	1309983

Sensitivity / TPR =	1.000
Specificity / TNR =	0.674
Prevalence =	0.0085
Kvamme Gain (Kg) =	0.669
Accuracy =	0.677
Positive Prediction Value (PPV) =	0.026
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.008
Positive Prediction Gain (PPG) =	3.017
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.331

Region 1 East - Riverine Section 3

		Known Sites		
		Present	Absent	
Model Prediction	Present	8799	587112	595911
	Absent	0	1434275	1434275
		8799	2021387	2030186

Sensitivity / TPR =	1.000
Specificity / TNR =	0.710
Prevalence =	0.0043
Kvamme Gain (Kg) =	0.706
Accuracy =	0.711
Positive Prediction Value (PPV) =	0.015
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.004
Positive Prediction Gain (PPG) =	3.407
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.294

Region 1 East - Upland Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	5043	7758273	7763316
	Absent	0	21146115	21146115
		5043	28904388	28909431

Sensitivity / TPR =	1.000
Specificity / TNR =	0.732
Prevalence =	0.0002
Kvamme Gain (Kg) =	0.731
Accuracy =	0.732
Positive Prediction Value (PPV) =	0.001
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	3.724
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.269

Region 1 East - Upland Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	19142	5156870	5176012
	Absent	0	15771445	15771445
		19142	20928315	20947457

Sensitivity / TPR =	1.000
Specificity / TNR =	0.754
Prevalence =	0.0009
Kvamme Gain (Kg) =	0.753
Accuracy =	0.754
Positive Prediction Value (PPV) =	0.004
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.001
Positive Prediction Gain (PPG) =	4.047
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.247

Region 1 East - Upland Section 3

		Known Sites		
		Present	Absent	
Model Prediction	Present	12402	6641529	6653931
	Absent	0	17130478	17130478
		12402	23772007	23784409

Sensitivity / TPR =	1.000
Specificity / TNR =	0.721
Prevalence =	0.0005
Kvamme Gain (Kg) =	0.720
Accuracy =	0.721
Positive Prediction Value (PPV) =	0.002
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.001
Positive Prediction Gain (PPG) =	3.574
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.280

Region 1 North - Riverine Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	9843	1069965	1079808
	Absent	0	2145888	2145888
		9843	3215853	3225696

Sensitivity / TPR =	1.000
Specificity / TNR =	0.667
Prevalence =	0.0031
Kvamme Gain (Kg) =	0.665
Accuracy =	0.668
Positive Prediction Value (PPV) =	0.009
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.003
Positive Prediction Gain (PPG) =	2.987
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.335

Region 1 North - Riverine Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	39829	1632182	1672011
	Absent	0	3325612	3325612
		39829	4957794	4997623

Sensitivity / TPR =	1.000
Specificity / TNR =	0.671
Prevalence =	0.0080
Kvamme Gain (Kg) =	0.665
Accuracy =	0.673
Positive Prediction Value (PPV) =	0.024
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.008
Positive Prediction Gain (PPG) =	2.989
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.335

Region 1 North - Upland Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	15587	11927823	11943410
	Absent	0	25064877	25064877
		15587	36992700	37008287

Sensitivity / TPR =	1.000
Specificity / TNR =	0.678
Prevalence =	0.0004
Kvamme Gain (Kg) =	0.677
Accuracy =	0.678
Positive Prediction Value (PPV) =	0.001
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	3.099
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.323

Region 1 North - Upland Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	34100	16464934	16499034
	Absent	0	41041485	41041485
		34100	57506419	57540519

Sensitivity / TPR =	1.000
Specificity / TNR =	0.714
Prevalence =	0.0006
Kvamme Gain (Kg) =	0.713
Accuracy =	0.714
Positive Prediction Value (PPV) =	0.002
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.001
Positive Prediction Gain (PPG) =	3.488
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.287

Region 1 West - Riverine Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	8263	593506	601769
	Absent	0	1212398	1212398
		8263	1805904	1814167

Sensitivity / TPR =	1.000
Specificity / TNR =	0.671
Prevalence =	0.0046
Kvamme Gain (Kg) =	0.668
Accuracy =	0.673
Positive Prediction Value (PPV) =	0.014
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.005
Positive Prediction Gain (PPG) =	3.015
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.332

Region 1 West - Riverine Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	14801	328328	343129
	Absent	0	664528	664528
		14801	992856	1007657

Sensitivity / TPR =	1.000
Specificity / TNR =	0.669
Prevalence =	0.0147
Kvamme Gain (Kg) =	0.659
Accuracy =	0.674
Positive Prediction Value (PPV) =	0.043
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.015
Positive Prediction Gain (PPG) =	2.937
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.341

Region 1 West - Riverine Section 3

		Known Sites		
		Present	Absent	
Model Prediction	Present	36409	940334	976743
	Absent	0	2035586	2035586
		36409	2975920	3012329

Sensitivity / TPR =	1.000
Specificity / TNR =	0.684
Prevalence =	0.0121
Kvamme Gain (Kg) =	0.676
Accuracy =	0.688
Positive Prediction Value (PPV) =	0.037
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.012
Positive Prediction Gain (PPG) =	3.084
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.324

Region 1 West - Riverine Section 4

		Known Sites		
		Present	Absent	
Model Prediction	Present	26163	753811	779974
	Absent	0	1618788	1618788
		26163	2372599	2398762

Sensitivity / TPR =	1.000
Specificity / TNR =	0.682
Prevalence =	0.0109
Kvamme Gain (Kg) =	0.675
Accuracy =	0.686
Positive Prediction Value (PPV) =	0.034
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.011
Positive Prediction Gain (PPG) =	3.075
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.325

Region 1 West - Riverine Section 5

		Known Sites		
		Present	Absent	
Model Prediction	Present	9590	569669	579259
	Absent	0	1188475	1188475
		9590	1758144	1767734

Sensitivity / TPR =	1.000
Specificity / TNR =	0.676
Prevalence =	0.0054
Kvamme Gain (Kg) =	0.672
Accuracy =	0.678
Positive Prediction Value (PPV) =	0.017
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.005
Positive Prediction Gain (PPG) =	3.052
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.328

Region 1 West - Upland Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	40076	7213509	7253585
	Absent	0	14173320	14173320
		40076	21386829	21426905

Sensitivity / TPR =	1.000
Specificity / TNR =	0.663
Prevalence =	0.0019
Kvamme Gain (Kg) =	0.661
Accuracy =	0.663
Positive Prediction Value (PPV) =	0.006
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.002
Positive Prediction Gain (PPG) =	2.954
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.339

Region 1 West - Upland Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	62120	4606572	4668692
	Absent	0	9846449	9846449
		62120	14453021	14515141

Sensitivity / TPR =	1.000
Specificity / TNR =	0.681
Prevalence =	0.0043
Kvamme Gain (Kg) =	0.678
Accuracy =	0.683
Positive Prediction Value (PPV) =	0.013
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.004
Positive Prediction Gain (PPG) =	3.109
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.322

Region 1 West - Upland Section 3

		Known Sites		
		Present	Absent	
Model Prediction	Present	32660	9431094	9463754
	Absent	0	18798973	18798973
		32660	28230067	28262727

Sensitivity / TPR =	1.000
Specificity / TNR =	0.666
Prevalence =	0.0012
Kvamme Gain (Kg) =	0.665
Accuracy =	0.666
Positive Prediction Value (PPV) =	0.003
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.001
Positive Prediction Gain (PPG) =	2.986
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.335

Region 1 West - Upland Section 4

		Known Sites		
		Present	Absent	
Model Prediction	Present	28015	6124865	6152880
	Absent	0	12999020	12999020
		28015	19123885	19151900

Sensitivity / TPR =	1.000
Specificity / TNR =	0.680
Prevalence =	0.0015
Kvamme Gain (Kg) =	0.679
Accuracy =	0.680
Positive Prediction Value (PPV) =	0.005
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.001
Positive Prediction Gain (PPG) =	3.113
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.321

Region 1 West - Upland Section 5

		Known Sites		
		Present	Absent	
Model Prediction	Present	27629	5162530	5190159
	Absent	0	11072057	11072057
		27629	16234587	16262216

Sensitivity / TPR =	1.000
Specificity / TNR =	0.682
Prevalence =	0.0017
Kvamme Gain (Kg) =	0.681
Accuracy =	0.683
Positive Prediction Value (PPV) =	0.005
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.002
Positive Prediction Gain (PPG) =	3.133
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.319

Region 2/3 - Riverine Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	51223	1783493	1834716
	Absent	0	3935939	3935939
		51223	5719432	5770655

Sensitivity / TPR =	1.000
Specificity / TNR =	0.688
Prevalence =	0.0089
Kvamme Gain (Kg) =	0.682
Accuracy =	0.691
Positive Prediction Value (PPV) =	0.028
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.009
Positive Prediction Gain (PPG) =	3.145
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.318

Region 2/3 - Riverine Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	37994	1022001	1059995
	Absent	0	2273724	2273724
		37994	3295725	3333719

Sensitivity / TPR =	1.000
Specificity / TNR =	0.690
Prevalence =	0.0114
Kvamme Gain (Kg) =	0.682
Accuracy =	0.693
Positive Prediction Value (PPV) =	0.036
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.011
Positive Prediction Gain (PPG) =	3.145
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.318

Region 2/3 - Riverine Section 3

		Known Sites		
		Present	Absent	
Model Prediction	Present	1218	190151	191369
	Absent	0	404421	404421
		1218	594572	595790

Sensitivity / TPR =	1.000
Specificity / TNR =	0.680
Prevalence =	0.0020
Kvamme Gain (Kg) =	0.679
Accuracy =	0.681
Positive Prediction Value (PPV) =	0.006
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.002
Positive Prediction Gain (PPG) =	3.113
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000

Region 2/3 - Riverine Section 4

		Known Sites		
		Present	Absent	
Model Prediction	Present	1048	262933	263981
	Absent	0	573576	573576
		1048	836509	837557

Sensitivity / TPR =	1.000
Specificity / TNR =	0.686
Prevalence =	0.0013
Kvamme Gain (Kg) =	0.685
Accuracy =	0.686
Positive Prediction Value (PPV) =	0.004
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.001
Positive Prediction Gain (PPG) =	3.173
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.315

Region 2/3 - Riverine Section 5

		Known Sites		
		Present	Absent	
Model Prediction	Present	25307	853207	878514
	Absent	0	1753598	1753598
		25307	2606805	2632112

Sensitivity / TPR =	1.000
Specificity / TNR =	0.673
Prevalence =	0.0096
Kvamme Gain (Kg) =	0.666
Accuracy =	0.676
Positive Prediction Value (PPV) =	0.029
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.010
Positive Prediction Gain (PPG) =	2.996
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.334

Region 2/3 - Upland Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	29177	11049172	11078349
	Absent	0	25729079	25729079
		29177	36778251	36807428

Sensitivity / TPR =	1.000
Specificity / TNR =	0.700
Prevalence =	0.0008
Kvamme Gain (Kg) =	0.699
Accuracy =	0.700
Positive Prediction Value (PPV) =	0.003
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.001
Positive Prediction Gain (PPG) =	3.322
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.301

Region 2/3 - Upland Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	40024	6816763	6856787
	Absent	0	15626333	15626333
		40024	22443096	22483120

Sensitivity / TPR =	1.000
Specificity / TNR =	0.696
Prevalence =	0.0018
Kvamme Gain (Kg) =	0.695
Accuracy =	0.697
Positive Prediction Value (PPV) =	0.006
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.002
Positive Prediction Gain (PPG) =	3.279
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.305

Region 2/3 - Upland Section 3

		Known Sites		
		Present	Absent	
Model Prediction	Present	15520	1540922	1556442
	Absent	0	3510879	3510879
		15520	5051801	5067321

Sensitivity / TPR =	1.000
Specificity / TNR =	0.695
Prevalence =	0.0031
Kvamme Gain (Kg) =	0.693
Accuracy =	0.696
Positive Prediction Value (PPV) =	0.010
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.003
Positive Prediction Gain (PPG) =	3.256
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.307

Region 2/3 - Upland Section 4c*

		Known Sites		
		Present	Absent	
Model Prediction	Present	4359	4881128	4885487
	Absent	0	10784264	10784264
		4359	15665392	15669751

Sensitivity / TPR =	1.000
Specificity / TNR =	0.688
Prevalence =	0.0003
Kvamme Gain (Kg) =	0.688
Accuracy =	0.688
Positive Prediction Value (PPV) =	0.001
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	3.207
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.312

* combined rockshelter and non-rockshelter specific models

Region 2/3 - Upland Section 5c*

		Known Sites		
		Present	Absent	
Model Prediction	Present	11349	11247165	11258514
	Absent	0	21525035	21525035
		11349	32772200	32783549

Sensitivity / TPR =	1.000
Specificity / TNR =	0.657
Prevalence =	0.0003
Kvamme Gain (Kg) =	0.657
Accuracy =	0.657
Positive Prediction Value (PPV) =	0.001
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	2.912
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.343

* combined rockshelter and non-rockshelter specific models

Complete Model

		Known Sites		
		Present	Absent	
Model Prediction	Present	678154	127533633	128211787
	Absent	0	288725095	288725095
		678154	416258728	416936882

Sensitivity / TPR =	1.000
Specificity / TNR =	0.694
Prevalence =	0.0016
Kvamme Gain (Kg) =	0.692
Accuracy =	0.694
Positive Prediction Value (PPV) =	0.005
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.002
Positive Prediction Gain (PPG) =	3.252
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.308