PENNSYLVANIA DEPARTMENT OF TRANSPORTATION
ARCHAEOLOGICAL PREDICTIVE MODEL SET

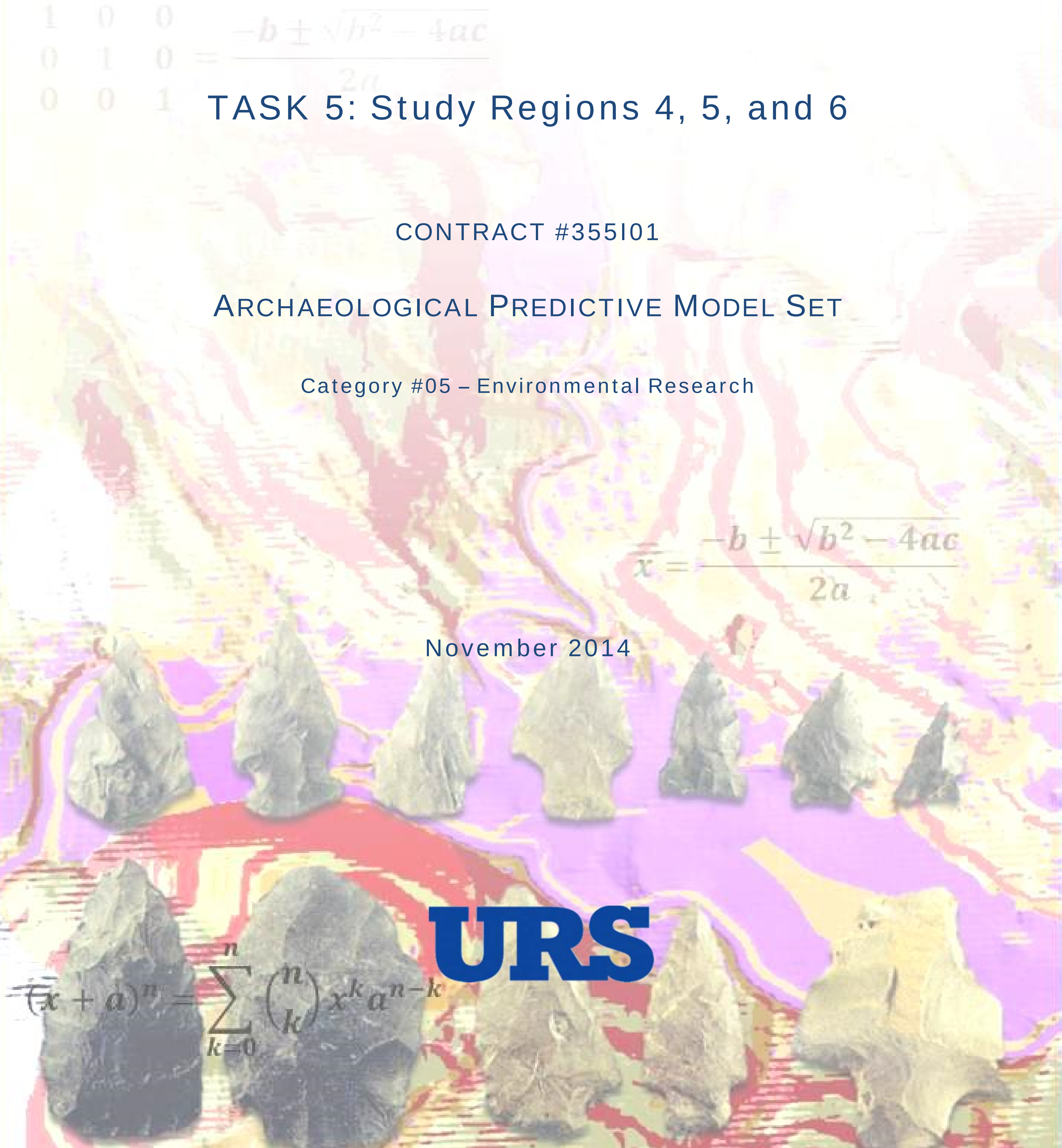
TASK 5: Study Regions 4, 5, and 6

CONTRACT #355I01

ARCHAEOLOGICAL PREDICTIVE MODEL SET

Category #05 – Environmental Research

November 2014


$\begin{matrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{matrix} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$
$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$
$(x + a)^n = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}$

URS

PENNSYLVANIA DEPARTMENT OF TRANSPORTATION ARCHAEOLOGICAL PREDICTIVE MODEL SET TASK 5: STUDY REGIONS 4, 5, AND 6

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

November 2014

ABSTRACT

This report is the documentation for Task 5 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 5, discussed in the following report, is to create a series of archaeological predictive models for Regions 4, 5, and 6. In total, this area covers 13,870.8 square miles, which is 30.1% of the state. These three regions cover much of central Pennsylvania, including the Ridge and Valley Province and part of the Appalachian Plateau Province. A total of 3,173 prehistoric archaeological components were incorporated into this modeling effort. One hundred and thirty-one individual candidate models were created to cover these three regions. The final ensemble is created from 36 models selected for their representation of the archaeological sensitivity of each of the subareas. This final model correctly classifies 95.2% of known site-present cells within 29.9% of the study area, for a Kg of 0.685 and an average hold-out sample prediction error of RMSE = 0.176.

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 (PASS) file database to produce a baseline model for the sensitivity of prehistoric site-presence throughout the entire Commonwealth using Archaeological Predictive Modeling (APM). The resulting assessments of archaeological sensitivity will be used by transportation, planning, and other Cultural Resources 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 from the previous tasks in this project—a review of APM literature (Harris 2013a), designation of study regions (Harris 2013b), the creation of a pilot model for central Pennsylvania (Harris 2014), and modeling three regions in western Pennsylvania (Harris et al. 2014), this report documents the second 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 4, 5, and 6 (Figure 1). These regions comprise a total of 13,870.8 square miles, which is 30.1% of the entire state. Covering almost the entirety of central Pennsylvania, this process involved creating 36 individual models from a dataset of 3,173 prehistoric archaeological sites or site components.

The process reported below consisted of the development of proportionally weighted models and three statistical models (Logistic Regression [LR], Multivariate Adaptive Regression Splines [MARS], and Random Forest [RF]) for each of the 36 subareas. Each of these three model types is discussed and detailed in the previous Task 3 report. The final model selected to represent the three regions is a composite of each of the four different model types: five proportionally weighted models, five LR models, 13 MARS models, and 13 RF models. The selection of a model for each subarea was based on the quality, quantity, and representativeness of the known data, the model metrics and error rates, and the distribution of site-present cells versus background cells summed up by the Kvamme Gain (Kg) statistic (Kvamme 1988). 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.

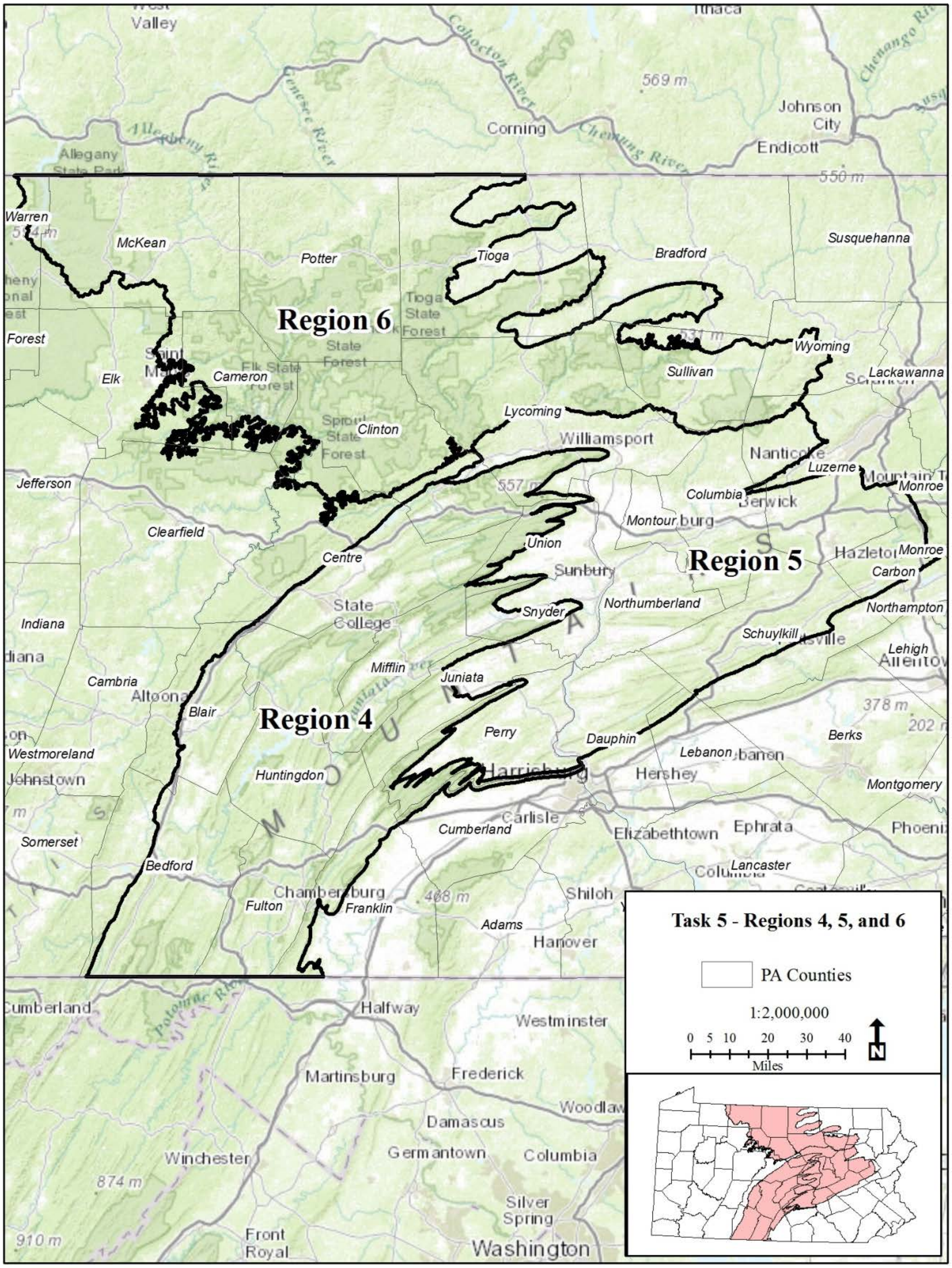


Figure 1 - Overview of Regions 4, 5, and 6

PREDICTIVE MODELING IN REGIONS 4, 5, AND 6

Numerous predictive model studies have been undertaken within Pennsylvania, many for compliance-related projects. Because of this association, the models often focused on an area determined by the location of the specific project, and were not generated to answer questions about settlement patterns. Only two predictive models were located for Regions 4, 5, and 6, and they did not attempt to predict anything beyond a general archaeological sensitivity for prehistoric resources (Duncan and Schilling 1999; Duncan et al. 1999). The dearth of predictive models in Regions 4, 5, and 6 is 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. Duncan and Schilling (1999:27) also observed that the reliability of information in the PASS database was sometimes questionable. They found that nearly all the sites in their model dataset represented surface collected material, and that most of the well-documented sites were multi-component sites with mixed plowzone contexts.

Duncan and Shilling (1999) used GIS to identify archaeological sensitivity areas for a road improvement project in Region 5. Their preliminary model was developed from data obtained from the study area consisting of the seven USGS quadrangle maps surrounding the road project area, and tested against a randomly selected group of archaeological sites within the study area. This model was based on correlating site locations with environmental factors, and then ranking the importance of the environmental factors based on archaeological theory. The study area was subdivided into lowland and upland settings, as it was felt that the environmental factors associated with lowlands and uplands contrasted to a degree that would have seriously affected the viability of a predictive model for the area. Site information was retrieved from the PASS database, as well as from interviews with local amateur archaeologists (Duncan and Schilling 1999:26). A total of 345 sites were used in the predictive model dataset. The environmental factors included elevation, cost distance to water, and soil types.

Within their study area, Duncan and Schilling (1999:28) observed a general trend of increasing numbers of sites during the Late and Terminal Archaic, and again during the Late Woodland period. They noted that nearly all the sites within their study area consisted of Open habitation, prehistoric, or unknown function types, and determined that site type would not be a useful variable for the predictive model because of this. They also noted that sites were predominately located in lowland settings, such as flood plains, terraces, and stream benches.

The predictive model was tested against a set of 101 sites in the study area that were excluded from the predictive model dataset, and the results of the testing found that the model was about 80%

accurate in predicting site locations (Duncan and Schilling 1999:51). The model did not attempt to predict anything beyond the level of probability a particular cell in the model had to possess a prehistoric archaeological site; it did not attempt to predict site types or temporal periods.

The other predictive model identified for Regions 4, 5, and 6 was one that was generated for a location just outside the northeastern corner of Region 6 (Duncan et al. 1999). The model was developed in the same way as was the Duncan and Schilling (1999) model. The model was about 80% accurate in predicting whether any given cell in the model area would contain an archaeological site. Within their study area, Duncan et al. (1999:62) noted that sites are primarily found in flood plain settings, with very few sites in the uplands. The upland sites appear to solely represent short-term logistical camps, rather than longer seasonal occupations.

2
STUDY AREA – REGIONS 4, 5, AND 6

PHYSICAL CHARACTER

Regions 4 and 5 are located within the Ridge and Valley physiographic province, which is characterized by long, even ridges punctuated by long valleys that run in a southwesterly to northeasterly direction through the central and eastern portions of the state. Two sections of the Ridge and Valley province fall within Region 5 (Susquehanna Lowland), while one section is within Region 4 (Appalachian Mountain). Region 6 is 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. Two sections of the Appalachian Plateaus fall within Region 6 (Deep Valley and Glaciated High Plateau). (Table 1; Figure 2).

Table 1 - Physiographic Provinces and Sections for Modeling Regions 4, 5, and 6

Modeling Region	Physiographic Province	Physiographic Section
4	Ridge and Valley	Appalachian Mountain
5	Ridge and Valley	Susquehanna Lowland
		Anthracite Upland
6	Appalachian Plateaus	Deep Valleys
		Glaciated High Plateau

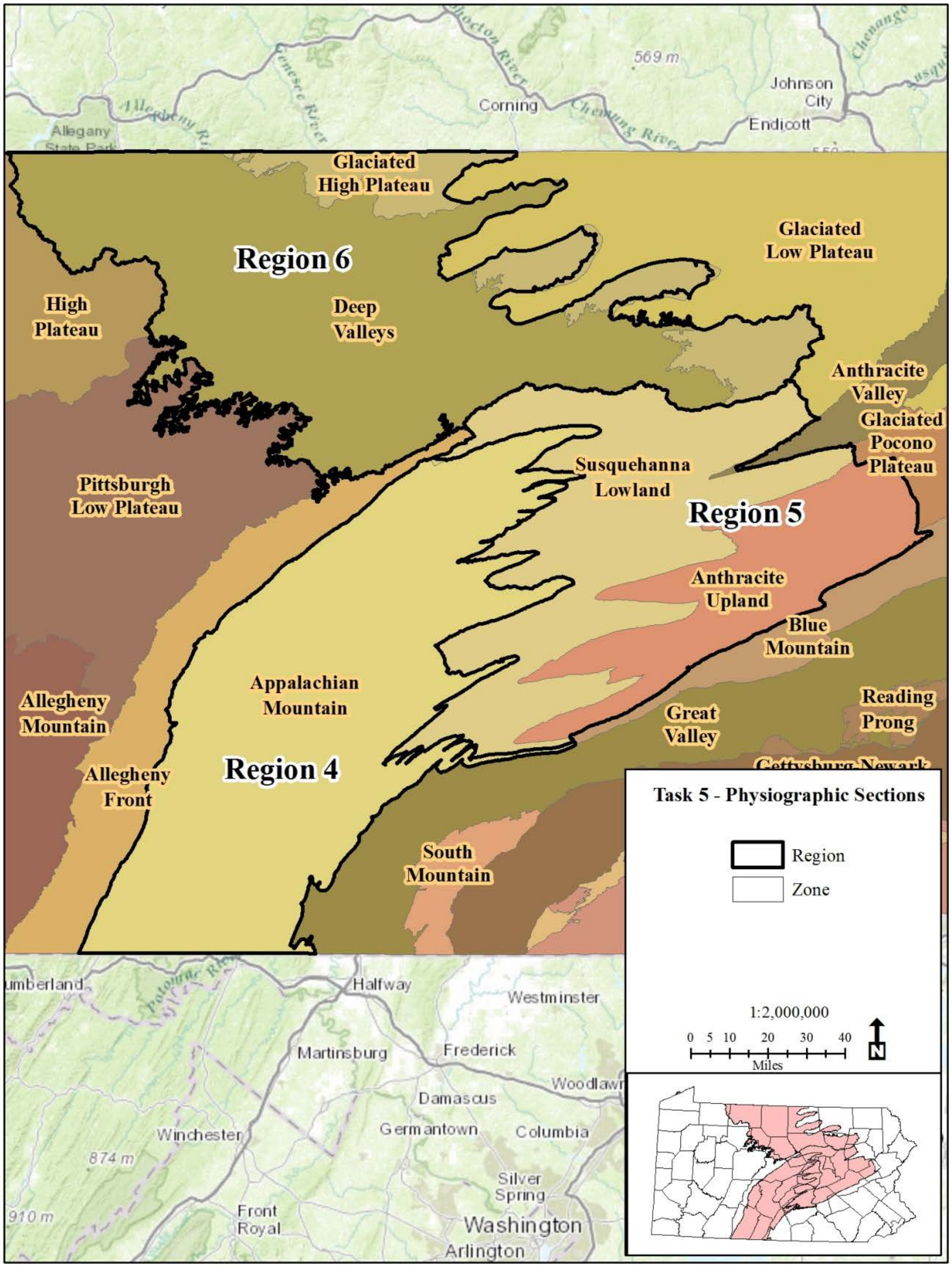


Figure 2 - Regions 4, 5, and 6 physiographic sections.

Appalachian Plateaus

Deep Valleys

The Deep Valleys section is located in the north-central part of the state. It is bordered to the north by New York State as well as by a portion of the Glaciated High Plains section, which is divided into four small non-contiguous sections along its eastern edge. Where the eastern edge of the Deep Valleys section does not abut a pocket of the Glaciated High Plateau section, it abuts fingers of the Glaciated Low Plateau section that are interspersed between the pockets of the Glaciated High Plateau section along the eastern flank of the Deep Valleys section. To the south, the Deep Valleys section abuts the Susquehanna Lowland section along its southeastern edge and the Allegheny Front section along its southwestern extreme. To the west the Deep Valleys section abuts the Pittsburgh Low Plateau section and High Plateau section. The boundaries of this section are based on physical attributes such as deep valley basins, drainage divides, or top of valley slopes. The dominant topographic features contained within this section are deep angular valleys cutting into a mix of both broad and narrow uplands. The local relief is defined as moderate to very high, with elevations ranging from 560 to 2,650 feet above mean sea level (amsl). The landforms within this section were initially created by a combination of fluvial erosion and periglacial mass wasting. The drainage pattern within the section is characterized by an angulate and rectangular network of streams and rivers. The underlying rock types that can be found in the section include sandstone, siltstone, shale, and conglomerate. The section’s geologic structure is composed of moderate amplitude and open folds. These open folds are responsible for the directionality and orientation of valleys contained within the section, and are therefore also responsible for the angulate and rectangular nature of the section’s drainage pattern.

Glaciated High Plateau section

The Glaciated High Plateau section is broken up into four smaller sections surrounded by the Deep Valleys section and the Glaciated Low Plateau section; it also abuts the Susquehanna Lowland section on its southernmost border. The boundaries are defined along the east by the base of escarpment or long steep slopes at the edge of plateaus. The remaining borders are defined by arbitrary margins of deep valleys. The dominant topographic forms that can be found in the section range from broad to narrow and rounded to flat elongated uplands and shallow valleys. The local relief is defined as low to high (101–1,000 feet). The elevation of the section ranges from a minimum of 620 feet to 2,560 feet amsl. The formations of the Glaciated High Plateau section were shaped by fluvial and glacial erosion. Glacial melting and deposition of sediment and minerals are other reasons for the sculpting of the landforms in the area. These processes of shaping the topography also contribute to the present drainage patterns found within the section, which are characterized as angulate and dendritic. The underlying rock types that can be found in this section include sandstone, siltstone, shale, and conglomerate stone. Coal is also present within the region. The section’s geologic structure is composed of moderate amplitude and open folds, defined as bends in stratified rock.

Ridge and Valley

Anthracite Upland Section

The Anthracite Upland section is located south of the Anthracite Valley section and stretches to the southwest. The section abuts the Glaciated Pocono Plateau section to the northeast, the Blue Mountain section along its southern border, and protrudes and recedes through the Susquehanna Lowland section. The section's boundaries are delineated by the separation of the coal and non-coal producing areas in the northeast. The remaining perimeter is defined by the outer base of the surrounding mountain. The area's underlying rock type consists of mainly sandstone, shale, conglomerate stone, and anthracite. The geologic structure of the section has many narrow folds with steep limbs and many faults. The origin or formation of the section and its landforms was executed mostly by fluvial erosion. Glacial erosion and periglacial mass wasting also contributed to the sculpting of the section. The natural dominant topographic landform is the upland surface with low, linear to rounded hills. The upland is surrounded by an escarpment, a valley, and a mountain rim. The most impactful cause of changes to the topography is anthropogenic in origin. The area is riddled with strip mines and waste piles left behind from commercially based industries. The way the landforms and topography were naturally formed and cut also explains the drainage pattern of the section. The waterway pattern reflects that of a trellis, with smaller tributaries pouring into larger rivers at 90-degree angles. The elevation of the section varies from a low of 320 feet amsl to a high of 2,094 feet amsl.

Susquehanna Lowland Section

The Susquehanna Lowland section is located in the central part of the state. It is bordered to the north by the Deep Valleys and Glaciated High Plateau sections. To the south it abuts the Anthracite Upland section and to the west it abuts the Appalachian Mountain section. Its eastern boundary is small, but abuts three separate sections: the Glaciated Low Plateau section, the Anthracite Valley section, and the Glaciated Pocono Plateau section (listed from north to south). The boundaries of the Susquehanna Lowland section are defined along the bases of the steep slopes that form the surrounding sections, but in valley areas the boundaries are defined arbitrarily. The dominant topographic features contained within this section are typified by low to moderately high linear ridges and valleys, as well as the large Susquehanna River valley from which the section takes its name. The local relief is defined as low to moderate, with elevations ranging from 260 to 1,715 feet amsl. The origin of the landforms within this section are principally the result of fluvial erosion, but there are also areas that were created by glacial erosion, and, in the northeastern part of the section, glacial deposition. The underlying rock types that can be found in the section include sandstone, siltstone, shale, conglomerate, limestone, and dolomite. The geologic structure of the section is characterized by a series of open and closed plunging folds with narrow hinges and planar limbs. These are indicative of compression faults and are known as kinks, where there is no real deflection in the limb of the fold. The resulting folded pattern looks like a zigzag and is very angular. These series of linear compression geologic folds within the section have created a drainage pattern characterized by trellis and angulate networks of streams and rivers in which streams and rivers flow in parallel valleys and join at right angles between ridges or breaks in the folds.

Appalachian Mountain Section

The Appalachian Mountain section is located in the central part of the state. It is bordered to the south by the state of Maryland, to the west by the Allegheny Front section, which, along with the Susquehanna Lowland section, also forms its northern boundary. The northeastern boundary is defined by the Susquehanna Lowland section and its southeastern border occurs at its interface with the Great Valley section. The dominant topographic features contained within this section are long narrow ridges separated by a mixture of broad to narrow valleys. This area is also known to include numerous examples of karst formations (a landscape feature formed by the dissolving of water-soluble rock, resulting in voids such as caves and sinkholes). The local relief is defined as low to moderate to very high, with elevations ranging from 440 to 2,775 feet amsl. The origin of the landforms within this section is predominately the result of fluvial erosion, but there are also areas contained within this section that were formed as a result of periglacial mass wasting and, in rare instances, the dissolution of carbonate rock (this is the process that results in the karst features described above). The underlying rock types that can be found in the section include sandstone, siltstone, shale, conglomerate, limestone, and dolomite. The geologic structure of the section is characterized by a series of open and closed plunging folds with narrow hinges and planar limbs. These are indicative of compression faults are known as kinks, where there is no real deflection in the limb of the fold. The resulting folded pattern looks like a zigzag and is very angular. This section also contains a number of faults of differing kinds. The drainage pattern within the section is characterized by trellis and angulate networks of streams and rivers that flow in parallel valleys and join at right angles between ridges or breaks in the geologic folds. The karst processes also play a part in drainage, creating a network of underground streams, lakes, etc.

Study Region Delineation

As described in the report for Regions 1, 2, and 3 (Harris et al. 2014), the state was divided into 10 modeling regions to ensure uniform modeling within similar landscapes and to help manage the large datasets (Figure 3). The boundaries for the 10 regions are based on grouping similar physiographic sections into regions of very roughly equal size (with the exception of Regions 3 and 10). The current report deals with the Regions 4, 5, and 6. Regions 4 and 5 were merged for data management purposes into Region 4/5. Note that while Regions 4 and 5 were combined, the subsequent splitting of Region 4/5 into smaller zones led to Region 4/5 West taking on the same boundaries as the original Region 4 and Region 4/5 East the original boundaries of Region 5. This is purely a coincidence of data organization and has no effect on the model’s outcome.

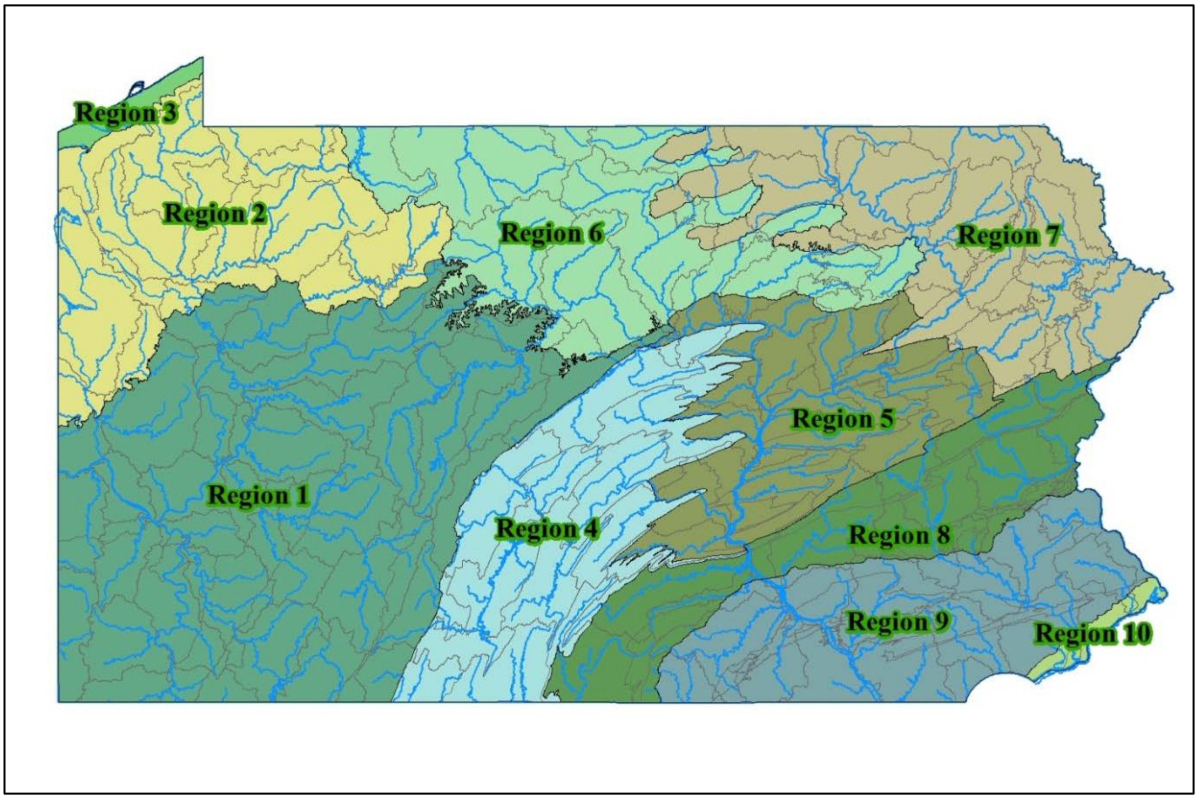


Figure 3 - Modeling regions for the Pennsylvania Model Set project.

Each region is broken down into a small number of zones based on drainage basin boundaries within physiographic province, largely for data management purposes (Table 2, Figure 4). Region 4/5 is divided into a west and east zone (equivalent to the original Regions 4 and 5, respectively), but Region 6 did not require division into zones. Zones are further subdivided into units referred to as sections, which are based on watershed boundaries within physiographic sections (sections were referred to as "physio-sheds" in earlier reports for this project). As shown in Table 2, Region 4/5 west contains six sections, Region 4/5 east has seven sections, and Region 6 has five sections.

Finally, each section was divided into upland and riverine subareas, shown in the final column in Table 2. 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. For Regions 4, 5, and 6 there are a total of 36 subareas and, therefore, 36 separate model building efforts. The rationale and methodology for dividing the sections into upland and riverine settings is discussed in detail in the Task 4 report (Harris et al. 2014). 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 "R4/5_east_riverine_section_1" will refer to the riverine subarea of section 1 of the east zone of Region 4/5. The modeled subareas are shown in Figure 5, Figure 6, and Figure 7.

Table 2 - Relationship between Regions, Zones, Sections, Subareas, and Physiography

Physiographic Province	Region	Zone	Physiographic Section	Section	Subarea
Ridge and Valley	4 (4/5)	west	Appalachian Mountain	1	riverine section 1
					upland section 1
				2	riverine section 2
					upland section 2
				3	riverine section 3
					upland section 3
				4	riverine section 4
					upland section 4
				5	riverine section 5
					upland section 5
				6	riverine section 6
					upland section 6
	5 (4/5)	east	Anthracite Upland	1	riverine section 1
					upland section 1
				2	riverine section 2
					upland section 2
				3	riverine section 3
					upland section 3
			Susquehanna Lowland	4	riverine section 4
					upland section 4
				5	riverine section 5
					upland section 5
				6	riverine section 6
					upland section 6
				7	riverine section 7
					upland section 7
Appalachian Plateaus	6	all	Glaciated High Plateau	1	riverine section 1
					upland section 1
			Deep Valleys	2	riverine section 2
					upland section 2
				3	riverine section 3
					upland section 3
				4	riverine section 4
					upland section 4
				5	riverine section 5
					upland section 5

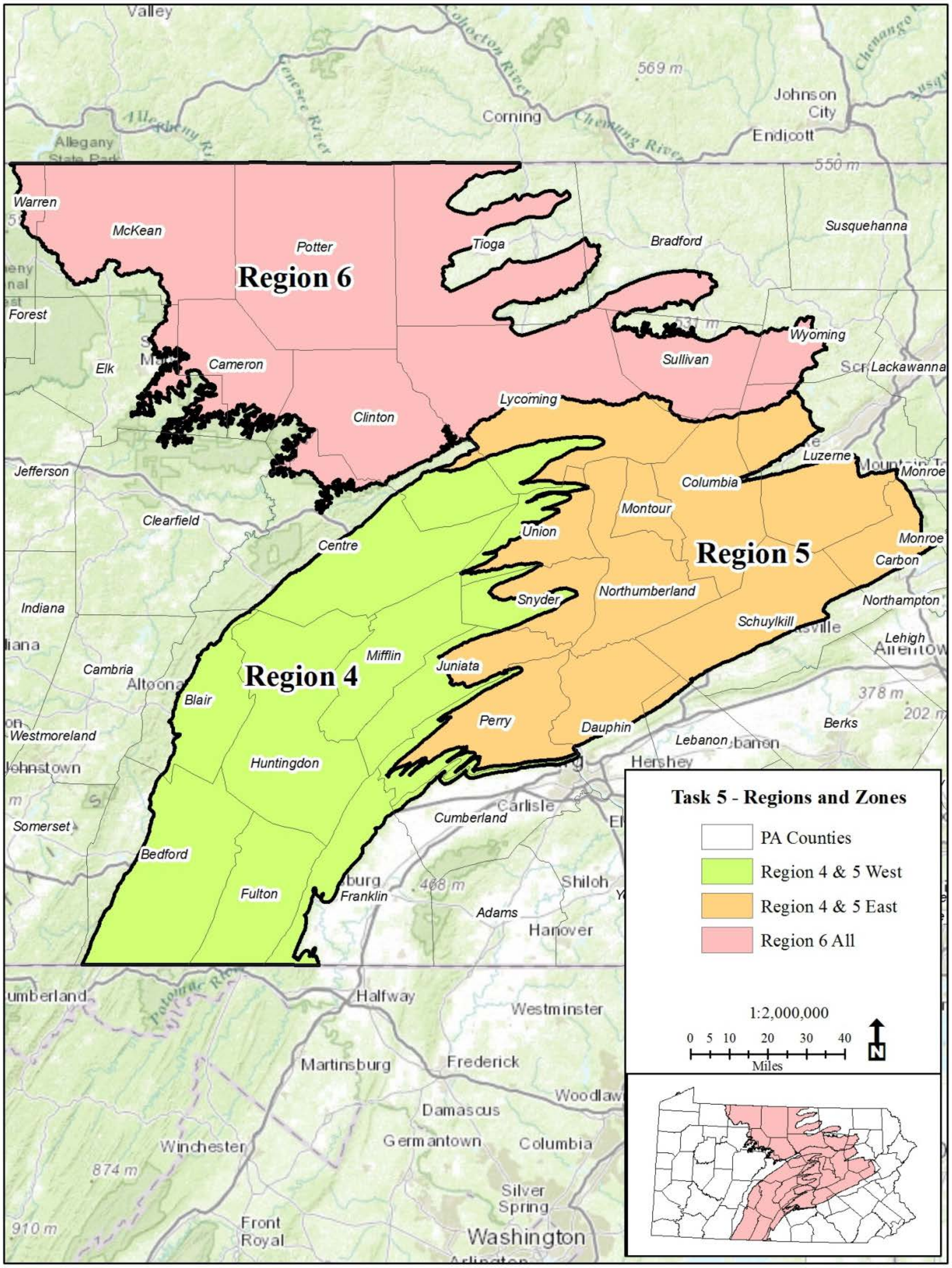


Figure 4 - Task 5 report regions and zones.

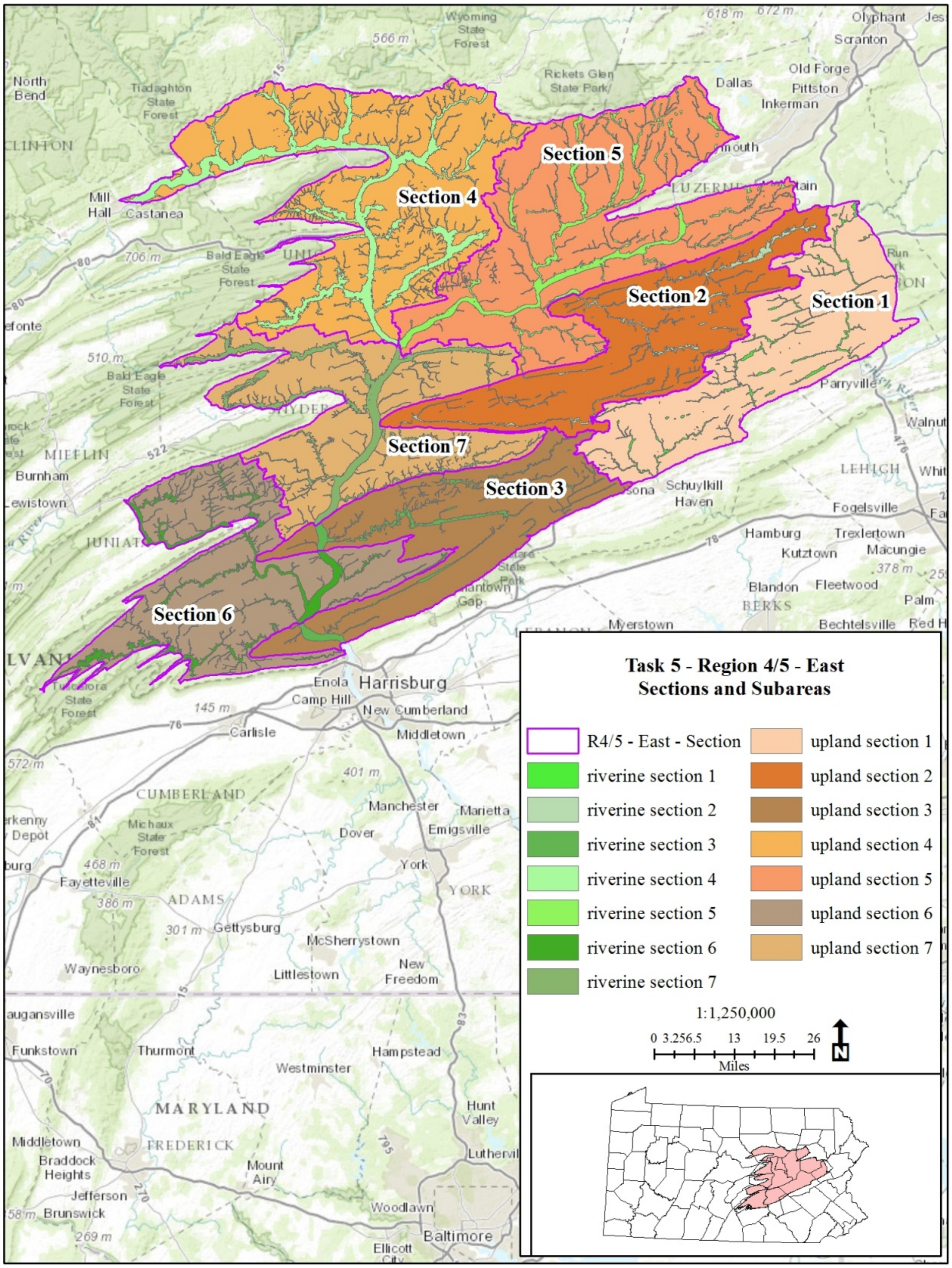


Figure 5 - Modeling subareas of Region 4/5 east.

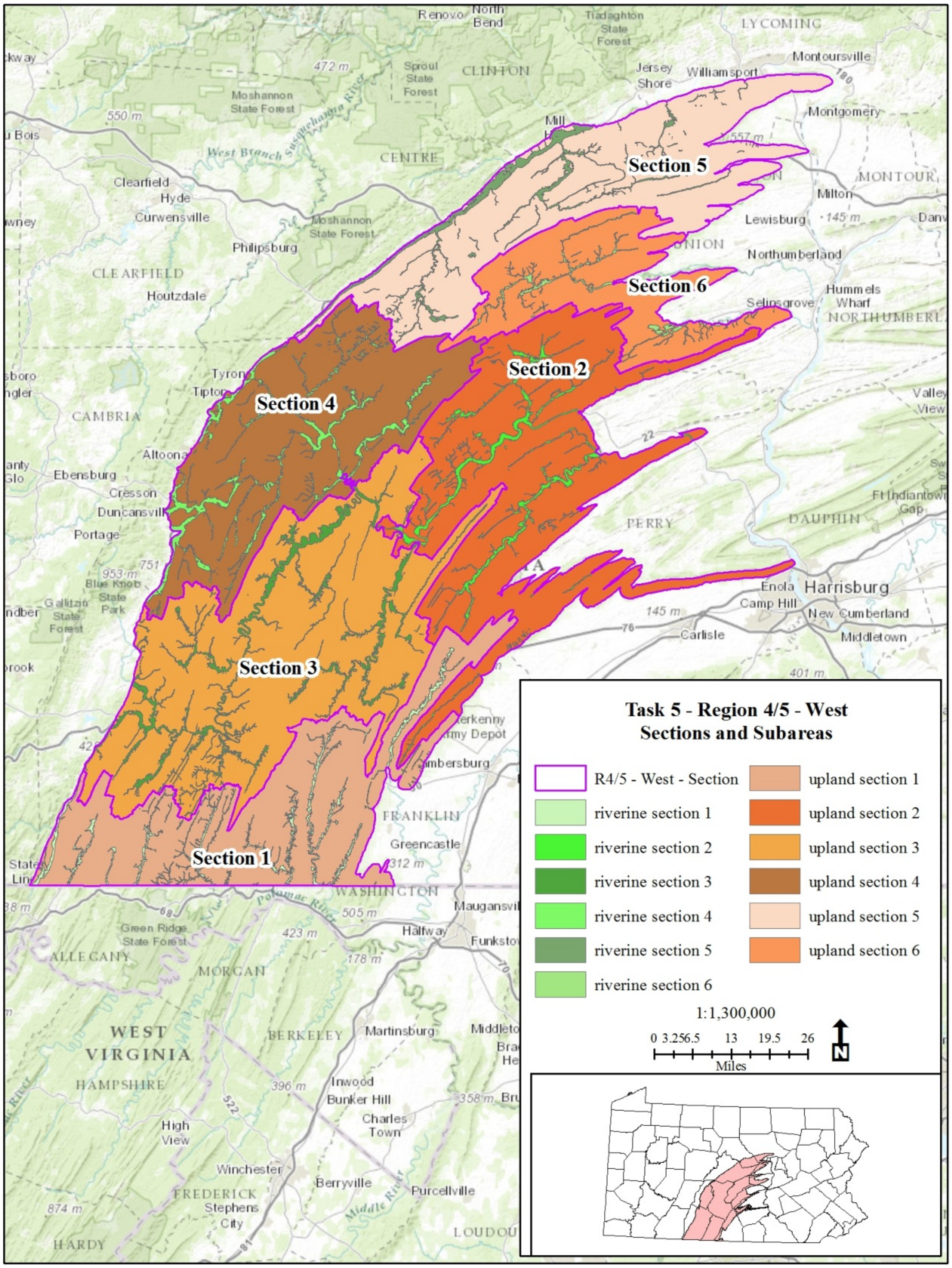


Figure 6 - Modeling subareas of Region 4/5 west.

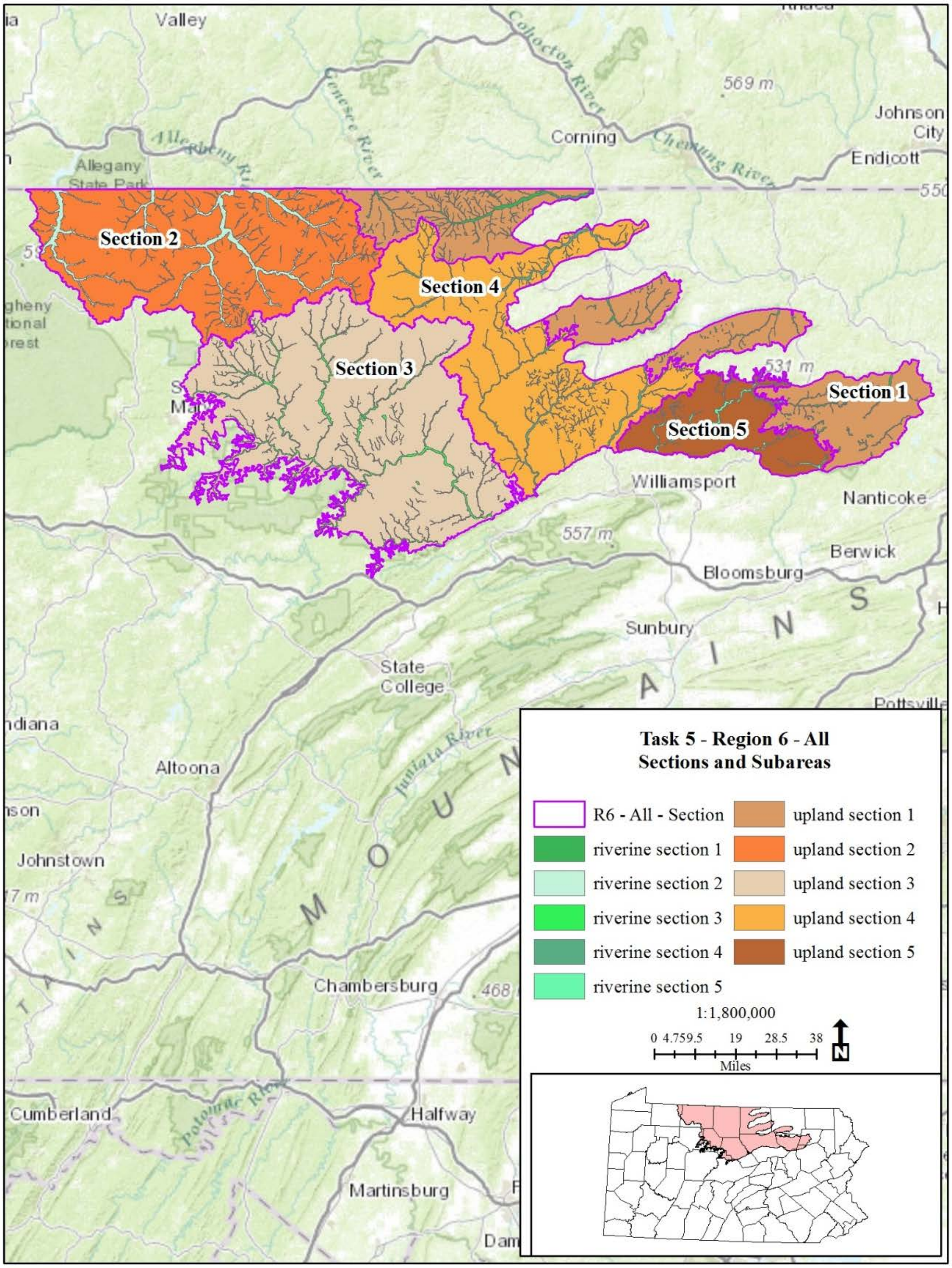


Figure 7 - Modeling subareas of Region 6.

PREHISTORIC BACKGROUND

The Peopling of the Americas and the Paleoindian Period

The first humans likely reached North America no earlier than about 30,000 years ago. The chronology of the Paleoindian period in Pennsylvania begins with a period known as Pre-Clovis, dating from about 14,000 to 9500 B.C. (Quinn et al. 1994). This date range is largely supported through extensive research performed at the Meadowcroft Rockshelter in southwest Pennsylvania, which has a minimum early date of 9300 B.C., though 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 (Quinn et al. 1994).

Most evidence of early human occupation in eastern North America is associated with the Clovis period (9500–8000 B.C.), which is characterized primarily by its distinctive lithic assemblage. Fluted projectile points, usually produced from high-quality chert, are generally considered to be the diagnostic marker of the time period, along with scrapers and spurred gravers. 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). One of the most well-known Paleoindian sites in the country, the Shoop Site, is located in Dauphin County in Region 5. The Shoop Site was excavated in the mid-twentieth century and yielded dozens of fluted points and hundreds of flake tools. The site covered an area of approximately 20 acres, representing a series of repeated visitations to the site (Custer 1996:118–122). Boyd et al. (2000:38) note that Paleoindians in the eastern United States likely employed a settlement pattern in which a small group would be highly mobile through part of the year, then practice a semi-sedentary lifestyle the rest of the year, in accordance with the specific seasonally available resources that were the focus of subsistence at any particular time. This pattern resulted in two basic types of Paleoindian sites: base camps and short-term resource procurement camps. The short-term camp categorization subsumes other specialized site types, such as hunting stations, quarries, and isolated point finds. Boyd et al. also use the same site types for the subsequent Early Archaic period (Boyd et al. 2000:43).

Carr and Adovasio (2002:36) provide data indicating that upland/interior locations in the Ridge and Valley province comprise only 18% of Paleoindian sites; 82% of all Paleoindian sites are located on flood plains and higher terraces of major streams. Carr and Adovasio (2002:41–42) suggest that Custer et al's (1983) cyclical, quarry-focused settlement pattern model best explains the high frequencies of New York Onondaga and Coxackie cherts on Paleoindian sites in the middle and upper Susquehanna drainage. This view of local Paleoindian settlement patterns and lithic raw material use may soon change with the presentation of data from site 36PE16, the first stratified Paleoindian site excavated in the Susquehanna drainage. Bibler and Miller's (2002) preliminary report on this site suggests that the majority of the component's lithic assemblage is composed of local cherts.

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 four subperiods: Early Archaic (8500–6000 B.C.), Middle Archaic (6000–4000 B.C.), Late Archaic (4000–1800 B.C.), and Terminal Archaic (1800–1000 B.C.), based on marked differences in subsistence and settlement patterns (Quinn et al. 1994).

The Early Archaic Period (8500–6000 B.C.)

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 used contemporaneously: included are Early Archaic Side-Notched, Charleston, Thebes, Kirk/Palmer, Kirk Stemmed, Large Bifurcate, and Small Bifurcate (Purtill 2009:569). Raber et al. (1998) note that Early Archaic sites in the Ridge and Valley province are often found close to sources of high-quality stone tool materials, such as jasper. MacDonald (2003) notes that in Region 4, Early Archaic sites are predominately open camps in lowland settings close to water.

The Early Archaic period is not well represented in the archaeological record for Regions 4, 5, and 6, although stratified Early Archaic components were present at Sheep Rock Shelter (36HU1), Huntingdon County (Michels and Smith 1967), and the West Water Street Site in Clinton County (Custer et al. 1994). Based on PASS file data, Carr (1998a:58–59) noted a drop in the use of riverine settings and a lack of patterned use of different topographic settings by Early Archaic peoples, in comparison to earlier Paleoindian groups and later bifurcate-using groups; Carr attributes the difference to rapid environmental change during the Early Holocene. Carr agrees with several authors (Custer 1989, 1996; Gardner 1989; Geier 1990; Stewart and Cavallo 1991) that greater organizational differences existed between Early Archaic groups and those of the Middle Archaic period.

The Middle Archaic Period (6000–4000 B.C.)

By the Middle Archaic, populations had shifted their movement strategies from high mobility to reduced mobility; the period saw a substantial increase in size of the regional population (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 are diverse and include MacCorkle, LeCroy, St. Albans, Kanawha, Neville, Otter Creek, and Stanly (Justice 1995; Carr

1998b:80). The bifurcated bases on these tools are typically seen as first occurring in the early Middle Archaic and are considered diagnostic of the period. 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). The use of rhyolite as a lithic raw material increased from the preceding periods.

Boyd et al. (2000:50) characterize Middle Archaic sites by the same two basic site types as the preceding periods (base camps and short-term camps), but note they display a tendency to exploit a wider range of physiographic 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. Middle Archaic groups likely practiced residential mobility of family units utilizing large base camps with satellite special-purpose camps, as previously. This possibility may explain an increase in Middle Archaic site visibility in PASS data, as a foraging system may result in a larger number of sites occupied on a temporary basis without a corresponding major increase in population. Site types may include small base camps on terraces, specialized resource procurement camps in the uplands, and lithic processing camps near quarry locations (MacDonald 2003:63).

The majority of Middle Archaic sites in the study area are located in the lowlands, and all sites are located close to a water source. Many sites are located near stream confluences. Coppock (2009:51) notes that large Middle Archaic base camps have not been found in the Juanita River Basin, corresponding with much of Region 4. Custer (1996:139–143) suggests that, at least in the lower Susquehanna Valley, several diagnostic projectile points typically assigned to the Late Archaic were first produced in the late Middle Archaic, resulting in the inflation of the number of Late Archaic sites. This is probably also the case in the Susquehanna’s West Branch Valley: for example, radiocarbon dates associated with Brewerton Series projectiles at the Memorial Park site indicate that the type was in use by 4720–3790 B.C. (Hart 1995:table 47).

The Late Archaic Period (4000–1800 B.C.)

Trends first seen in the Middle Archaic, such as the diversification of utilized plant resources, increased sedentism, and the establishment of cemeteries, continued into the Late Archaic period. Raber (2010) notes a general shift from early Middle Archaic residential mobility to a collecting strategy with base camps occupied for longer periods of time, possibly even for entire seasons, by the Terminal Archaic. The early Late Archaic in the Susquehanna drainage is best represented at the Memorial Park, East Bank, and Raker I sites (Hart 1995; East et al. 2002; Wyatt et al. 2005). The Memorial Park and East Bank sites, both on broad flood plains of the West Branch, produced numerous artifacts and fire-related features that ranged between 4000 and 2500 B.C.

The Late Archaic Period in central and eastern Pennsylvania is associated with the Laurentian and Piedmont traditions. The Laurentian Tradition has its roots to the north in New York and adjacent states. The Laurentian 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). Lithic material choices made by Late Archaic people show that they strongly favored jasper, chert, and rhyolite.

As the name implies, the Piedmont Tradition occupies the Appalachian Piedmont physiographic zone, extending from the Carolinas into New England. The tradition is characterized by long, narrow, lanceolate projectile points, frequently fashioned from non-local raw materials such as argillite (e.g., Kingsley et al. 1991). The occurrence of non-local lithic raw materials indicates trade and/or communication and interaction between Piedmont and Laurentian groups.

Some evidence from sites in the southeastern United States indicates that Late Archaic populations began to experiment with fired clay at this time (Sassaman 1993; Milanich 1994), though as yet no firm evidence has been found that Late Archaic groups in Regions 4, 5, and 6 were familiar with this technology before the very end of the period.

In general, site types include large base camps located on mid- to high-order streams occupied by all members of a band, with smaller bivouacs or short-term resource extraction camps that would have been occupied either as a single-night encampment by the whole group in transit, or by sub-groups or single members of the group focused on a singular activity (MacDonald 2003:77). Late Archaic base camps were strategically located to take advantage of resources that could be exploited with minimal expenditures of labor (Raber et al 1998:126). The number of sites located in upland settings as a portion of the whole population of sites increased from the preceding periods, although a primary focus on flood plain resources by Late Archaic groups is suggested by Duncan and Schilling (1999:16).

The Terminal Archaic Period (1800–1000 B.C.)

The Terminal Archaic, also known as the Transitional period, is thought to be linked with a climatic change that resulted in warmer and dryer conditions (Custer 1996:187). Diagnostic artifacts associated with the Terminal Archaic include the Broadspear type projectile points such as Lehigh, Susquehanna, 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). The increased use of jasper and rhyolite indicates expansion of trade networks during the Terminal Archaic (MacDonald 2003). Transitional Archaic sites are often characterized by high densities of fire-cracked rock, suggesting intensive cooking techniques. Carved steatite (soapstone) bowls first appear in this period; evidence of burning on many of these examples indicates use as cooking vessels. The earliest occurring pottery in Region 4 was found at the Sunny Side site, dated at ca. 1900 B.C., and was identified as Selden Island Cordmarked, characteristically tempered with steatite (Macdonald 2003:108).

Sites associated with the Terminal Archaic have yielded 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. A strong preference for terraces above the confluences of streams is noted for Terminal Archaic sites (MacDonald 2003:104). Sites with stratified, well-separated Terminal Archaic components in Regions 4, 5, and 6 include Gould Island, Jacobs (Weed and Wenstrom 1993), and Site 36CO17 (Jacoby et al. 1998) on the North Branch of the Susquehanna, and Site 36UN82 (Wall 1994, 2000), Memorial Park (Hart 1995), and East Bank (East et al. 2002) on the West Branch.

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.), Middle Woodland (100 B.C.–A.D. 1000), and Late Woodland Periods (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 evident continuity with preceding Archaic lifeways. There appears to be a significant decline in Early and Middle Woodland sites in the Susquehanna Basin (containing Regions 5 and 6), but it is unknown if this reflects an actual demographic change for the region or if there is a masking effect resulting from difficulties sorting out regional variants of Early and Middle Woodland points from similar Late Archaic styles (Wyatt 2003). Wyatt notes that in general, Early and Middle Woodland sites in the Susquehanna Basin differ from preceding Terminal Archaic sites through smaller size, lack of large thermal features, and preference for local lithic materials for tool manufacture.

<u>The Early Woodland Period (1000–100 B.C)</u>

The Early Woodland cultural complexes in the larger part of Pennsylvania that includes Region 4 consist of the Meadowood and Adena cultures. As noted, the number of identified Early Woodland sites in the region drops sharply from the preceding Archaic periods, a trend that possibly began in the Terminal Archaic (MacDonald 2003:116). There are several main interpretations of this trend: it may represent a wholesale movement of Early Woodland people out of the area, or possibly increased nucleation of Early Woodland groups, with populations staying steady but living in larger groups at fewer numbers of locations. Alternatively, the fewer numbers of sites might indicate an overall population decrease across the region at this time. While this supposition is attractive, a possible cause for such a demographic change has not been forthcoming. Another possibility is that archaeological surveys in the region have simply missed landforms preferred by Early Woodland groups. Early Woodland pottery is characteristically tempered with a variety of materials, including steatite, mussel shell, quartz, and other mixed grit, representing a technological change from the pottery associated with the late Transitional Archaic. Early Woodland pottery types include Vinette I, Marcey Creek, and Brodhead Net-marked (MacDonald 2003:117). Site types are mainly open camps and lithic-reduction sites in lowland settings, with a few upland sites on ridgetops and rock shelters, representing a continuation of the base camp and short-term resource procurement camp model of settlement. Diagnostic projectile points include Meadowood, Hellgrammite, Cresap Stemmed, Robbins, and Adena Stemmed types

In Pennsylvania, Early Woodland settlement patterns resembled those of the Late Archaic and Terminal Archaic periods, with seasonal base camps situated in flood plain settings with smaller upland resource procurement locations (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. In general, evidence for Early Woodland horticulture is rarely encountered in Regions 4, 5, and 6. The burial mounds associated with Early Woodland cultures to the west in the Ohio River Valley are largely absent from Regions 4, 5, and 6.

The Middle Woodland Period (100 B.C.–A.D. 1000)

Stewart (2003:17) notes there is a general lack of information about the habitation sites that may be associated with early Middle Woodland mound locations in the Ridge and Valley province, but trade goods suggest a moderate amount of trade with Ohio Valley Hopewell. In particular, in a section of the Juniata River Valley in Region 4, Hopewellian trade goods occur in unexpectedly high densities, which is especially notable due to the lack of such artifacts in adjacent watersheds. Similar to the Early Woodland, Middle Woodland sites occur in lower frequency than those of the preceding Archaic periods. Again, this trend can be attributed to a continuation of either a population movement away from the area or increased nucleation of resident populations into fewer and fewer sites, as part of a trend of increasing sedentism. However, the lack of large occupational sites within the study area would support the argument for population reduction over aggregation as an explanation for the low frequency of Middle Woodland sites. Small camp sites were the only site type identified within MacDonald's (2003:129) study area in the Upper Juniata Sub-Basin, with most Middle Woodland sites located in the lowlands. All of the sites in MacDonald's study were located within 150 m of a stream, and locations near confluences were widely preferred. The extravagant mound and earthwork-building practices of the Hopewell culture are not present in central Pennsylvania; rather, only a few burial mounds are associated with Middle Woodland cultural groups.

Middle Woodland cultural phases in Region 4, 5, and 6 include Fox Creek, Kipp Island, Clemson Island, and Jack's Reef (Wyatt 2003:41). Clemson Island is a late Middle Woodland culture appearing about 700 A.D. Clemson Island groups lived in hamlets and practiced horticulture. In the upper and middle Susquehanna, the dominant late Middle Woodland cultural expression is referred to as the Kipp Island phase after the type site in central New York (Ritchie 1994), dated between 500 and 850 A.D. (Funk 1993:206). Diagnostic artifacts of the Middle Woodland period in central Pennsylvania include Raccoon Notched, Rossville, Fox Creek, Levanna, and Jack's Reef projectile point types; Middle Woodland ceramic types include the Point Peninsula series for the Upper Susquehanna River Valley and the Fox Creek and subsequent Kipp Island series in the north-central part of the state, including Region 6.

The Late Woodland Period (A.D. 1000–1550)

The Late Woodland period in general is marked by a move toward nucleated, fortified settlements and the emergence of maize-based agricultural groups (Griffin 1967). 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; covered storage pits are frequently associated with the houses (Means 2008:8). By the end of the Late Woodland period, villages typically consisted of concentric circles of houses with a large central building.

The Late Woodland period in central Pennsylvania is characterized by an apparent population expansion or large-scale movement of people, with several times the number of sites identified than in the preceding period. MacDonald (2003:4) notes that the hunter-gatherer subsistence strategy persisted through the Late Woodland in the Upper Juniata River Sub-Basin. Only one site, the Sheep Rock Shelter, produced corn in substantial amounts. The Upper Juniata Sub-Basin may have been a peripheral area for one or more Late Woodland groups, with most sites showing the influence of Clemson Island and Shenks Ferry groups from the east, with a minor influence of the Monongahela culture to the west (MacDonald 2003:133). Clemson Island has its roots in the late Middle Woodland period, notable through its cordmarked, punctated ceramics. The later part of the Late Woodland in Regions 4 and 5 is associated with the Shenks Ferry culture with its highly decorated pottery, and to a lesser extent with the northern McFate-Quiggle cultures. In the northernmost parts of Region 6, the Point Peninsula Hunter’s Home phase represents the early Late Woodland period, followed by Clemson Island, Owasco, and Shenks Ferry cultural phases (Duncan et al. 1999).

Late Woodland site types include villages, burial mounds, agricultural hamlets, and special-purpose short-duration camps (MacDonald 2003:144). One Late Woodland village site identified in Region 4, Bedford Village, is possibly the easternmost Monongahela village site in Pennsylvania. Bedford Village featured a defensive stockade, and the material culture of the site shows influences from Monongahela, Clemson Island, and Potomac Valley groups located outside of the region. As with preceding periods, Late Woodland sites are all located near water, but in contrast, few sites show preferences for locations at stream confluences.

REGION 4 SITES

There are 1,156 archaeological sites with prehistoric components in Region 4 (Table 3 shows a breakdown of the Region 4 sites by site type and landform; individual tables for each of the time periods are included in Appendix B). A total of 603 sites in the PASS database did not possess diagnostic material and were not assigned to a temporal period. In addition, there are 101 Archaic-period sites that could not be assigned to one of the Archaic sub-periods, and 46 Woodland-period sites with a similar issue.

Table 3 - Region 4 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
Burial Mound	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2
Cemetery	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Earthwork	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Isolated Find	0	1	0	0	2	2	1	1	0	0	0	0	0	2	0	0	9
Lithic Reduction	0	11	1	0	6	6	3	0	1	2	3	1	0	1	0	1	36
Open Habitation, Prehistoric	0	283	6	1	195	91	49	29	2	5	1	0	3	25	1	8	699
Open Prehistoric Site, Unknown Function	0	61	7	0	28	43	4	1	2	6	10	2	6	11	0	5	186
Other Specialized Aboriginal Site	0	2	0	0	1	0	0	0	0	0	1	0	0	0	0	0	4
Petroglyph/ Pictograph	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Quarry	0	1	0	0	2	0	1	1	0	2	0	0	0	1	1	0	9
Rock shelter/cave	0	0	0	0	3	0	0	15	0	0	2	0	1	0	1	1	23
Unknown Function Open Site Greater than 20 m Radius	0	10	1	0	33	11	3	8	1	1	2	0	2	1	0	1	74
Unknown Function Surface Scatter Less than 20 m Radius	0	9	0	0	2	5	1	0	0	2	0	0	1	1	0	0	21
Village	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
(blank)	0	24	0	0	12	13	5	1	0	3	1	1	1	2	0	26	89
Total	0	405	15	1	284	172	68	56	6	22	20	4	14	44	3	42	1156

Site locations in Region 4 appear to show a strong trend for lowland settings, with 78.7% of all sites with landform information in the PASS database located in lowland settings (n = 877). The flood plain landform alone accounts for 36.3% of all site locations with landform data in Region 4 (n = 405). The only site type unique to lowland settings in Region 4 is the Village site type. The two most commonly occurring site types, Open habitation, prehistoric (n = 699) and Open prehistoric site, unknown function (n = 186), both predominately occur in lowland settings. The apparent trend toward site location in lowland settings in Region 4, however, may simply reflect survey bias. A visual examination of survey area locations for the counties that compose the majority of Region 4 suggests that much of the survey effort in this part of Pennsylvania has been directed at lowland settings. Thus, the lack of sites in upland settings may reflect survey bias rather than an actual prehistoric landform preference.

Paleoindian

Within Region 4, there have been 12 sites identified with Paleoindian components. Ten sites with Paleoindian components also contain one or more components dating to later time periods. Paleoindian sites in Region 4 have only been identified in lowland settings, primarily on flood plains. The two single-component Paleoindian sites are both isolated finds of fluted points; interestingly, these findspots occurred on a stream bench and a terrace and represent the only Paleoindian sites with landform data that do not occur in a flood plain setting.

Early Archaic

The PASS database records 45 sites with Early Archaic components in Region 4. Early Archaic sites in Region 4 are largely found in physiographic settings that are close to water sources, although the Early Archaic sites appear more spread out among the different lowland settings in comparison to the preceding Paleoindian period. The single-component Early Archaic sites in the PASS data that probably represent some form of resource extraction camp include Open habitation, prehistoric and Open prehistoric site, unknown function. Sites of this type are only found in lowland settings according to the PASS data.

Middle Archaic

The PASS database includes 77 sites with Middle Archaic components in Region 4. As with the preceding periods, Middle Archaic sites in Region 4 are mostly located in lowland physiographic settings. When single-component Middle Archaic site types are considered, however, sites are more evenly distributed, with slightly more sites (n = 6) in lowland settings than in upland settings (n = 4). 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 Open habitation, prehistoric site type occurs in equal numbers in lowland and upland settings, so it could represent either a base camp or a resource exploitation camp.

Late Archaic

The PASS database includes 293 sites with Late Archaic components in Region 4, an increase in site numbers by a factor of 3.8 from the preceding Middle Archaic. The increase in the number of recorded sites may indicate a population expansion within existing groups in the area, or a migration of outside groups to Region 4 during the Late Archaic period. Late Archaic sites in Region 4 show less of a focus toward lowland physiographic settings than during the preceding Early Archaic and Middle Archaic periods, although the emphasis on lowland settings is still strong for Late Archaic site distribution, with 84.7% of all sites with landform data found in lowland settings (n = 282). Flood plain settings alone account for 40.1% of all sites with landform information in the PASS database. 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 include 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 also occurs in greater numbers in lowland settings than in upland settings. The Unknown function open site, greater than 20 m radius site type occurs in nearly equal numbers between upland and lowland settings, and may represent repeatedly occupied short-term resource extraction camps.

Terminal Archaic

The PASS database includes only 120 sites with Terminal Archaic components in Region 4, perhaps indicating a continuity of Late Archaic cultural tendencies among central Pennsylvania populations at the end of the Archaic period. Alternatively, the drop in frequency of Terminal Archaic sites could represent the start of a regional depopulation trend that carried over into the Woodland period. There are 89 Terminal Archaic multi-component sites possessing components from either or both the Late Archaic and Early Woodland periods, representing 74.1% of the total population of Terminal Archaic sites. The fact that Terminal Archaic site components are strongly associated with preceding Late Archaic and subsequent Early Woodland components suggests group continuity within Region 4 between the Late Archaic and Early Woodland periods.

Terminal Archaic sites in Region 4 show a similar focus toward lowland physiographic settings as with the preceding Late Archaic period, with 83.5% of all Terminal Archaic sites with landform information in the PASS database located in lowland settings (n = 96, out of 115 sites with landform data). 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; Open prehistoric site, unknown function; and Unknown function open site, greater than 20 m radius. Nearly all sites of these types are found in lowland settings. Multi-component sites with Terminal Archaic components occur in a greater number of different upland settings in comparison to the single-component sites.

Early Woodland

The PASS database includes 77 sites with Early Woodland components in Region 4, a drop in site frequency from the Terminal Archaic by 35.8%. There are 71 Early Woodland multi-component sites possessing Terminal Archaic and Middle Woodland components (either one or both), representing 92.2% of the total population of Early Woodland sites, suggesting strong group continuity within Region 4 between the Terminal Archaic and Middle Woodland periods.

There are only six single-component Early Woodland sites in Region 4, none of which represent ceremonial sites such as earthworks or burial mounds. There is one single-component Early Woodland rock shelter site, 36HU60, possibly representing an upland resource exploitation camp. The other single-component site types include Open habitation, prehistoric; Open prehistoric site, unknown function; and Unknown function surface scatter, less than 20 m radius. These site types likely represent a mixture of base camps and temporary resource exploitation camps, similar to the preceding Archaic periods.

Middle Woodland

The PASS database includes 92 sites with Middle Woodland components in Region 4. There are 68 Middle Woodland multi-component sites possessing either or both Early and Late Woodland components, representing 73.9% of the total population of Late 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 4 between the three Woodland periods. Middle Woodland sites in Region 4 show a marked focus toward lowland physiographic settings, with 85.1% of all Middle Woodland sites with landform information located in lowlands. There does not appear to be a particular lowland setting preferred by Middle Woodland groups, with similar numbers of sites appearing on flood plains, stream benches, and terraces. No Middle Woodland ceremonial site types are recorded for Region 4. There are only five single-component Middle Woodland site types, and they likely represent seasonal occupation sites such as base camps and short-term resource extraction camps rather than year-round occupations such as hamlets or villages.

Late Woodland

The PASS data for Region 4 includes 193 sites with Late Woodland components. There are 52 Late Woodland multi-component sites possessing Middle Woodland components, representing 26.9% of the total population of Late Woodland sites. A lesser degree of group continuity within Region 4 between the Middle Woodland and Late Woodland periods is possibly indicated by the smaller percentage of Late Woodland sites with Middle Woodland components, suggesting migration into the region during the Late Woodland. Conversely, a population increase could also result in site locations being selected that were previously unoccupied during the Middle Woodland.

Late Woodland sites in Region 4 show a general focus toward lowland physiographic settings, with some exceptions. Village sites are perhaps the defining site type for the Late Woodland, but only two such sites appear in Region 4 and represent a reoccupation of the same landform: 36CN0210, the Piper Airport 1 site, and 36CN0211, the Piper Airport 2 site, both located in Lock Haven, Clinton County. Site 36CN0211 represents an early Late Woodland village dating to ca. 100 A.D., with a subsequent reoccupation of the landform by site 36CN0210, ca. 1300. This later occupation was originally unfortified, but a stockade was added ca. 1450 A.D. (MacDonald 2003:104). 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/caves; and Unknown function open site greater than 20 m radius. Open habitation, prehistoric sites, which likely include a number of base camp sites, are primarily found in lowland settings (93.5%). The data for the Open prehistoric site, unknown function site type, which may represent short-term resource extraction camps, shows 66.6% of this site type occurring in lowland settings (although only three such sites are present in Region 4). Rock shelters or caves also may have served as short-term resource extraction camps or base camps during the Late Woodland; all single-component Late Woodland Rock shelter/cave sites are found in upland settings in Region 4.

REGION 5 SITES

There are 1,280 archaeological sites with prehistoric components in Region 5 (Table 4 shows a breakdown of the Region 5 sites by site type and landform; individual tables for each of the time periods are included in Appendix B). A total of 563 sites in the PASS database did not possess diagnostic material and were not assigned to a temporal period. In addition, there are 102 Archaic sites that could not be assigned to one of the Archaic sub-periods, and 106 Woodland sites with a similar issue.

Sites in Region 5 overwhelmingly are found in lowland settings, with 86.4% of all sites (n = 1,106) located there. The flood plain landform alone accounts for 46.6% of all site locations in Region 5 (n = 597). Three site types are only found in lowland settings in Region 5, including the Burial mound, Cemetery, and Unknown function surface scatter less than 20 m radius site types; no site types are exclusive to upland settings. The most commonly occurring site type is Open habitation, prehistoric (n = 876), which occurs almost entirely in lowland settings. The apparent trend toward site location in lowland settings in Region 5 may, however, simply reflect survey bias. As with Region 4, visual analysis of survey areas in the PASS system suggests survey effort in Region 5 has been largely directed at lowland settings. Thus, the lack of sites in upland settings may reflect survey bias rather than an actual prehistoric landform preference.

Paleoindian

Within Region 5, there have been 41sites identified with Paleoindian components, according to the PASS database. Thirty-five sites with Paleoindian components also contain one or more components dating to later time periods. Paleoindian sites in Region 5 are nearly all found in lowland settings, with only three sites in upland settings. The six single-component Paleoindian sites in the PASS database for Region 5 include two isolated findspots, three Open habitation, prehistoric sites, and one site that lacked a site type description. Both isolated point findspots were in lowland settings.

Table 4 - Region 5 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
Burial Mound	0	1	0	0	2	1	0	0	0	0	0	0	0	0	0	0	4
Cemetery	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Earthwork	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Isolated Find	0	1	0	0	1	2	0	0	0	0	0	0	0	0	0	0	4
Lithic Reduction	0	3	3	0	6	15	0	1	0	3	1	0	0	1	1	0	34
Open Habitation, Prehistoric	0	447	2	8	187	170	21	3	8	3	2	0	0	14	5	6	876
Open Prehistoric Site, Unknown Function	0	41	2	2	8	32	3	6	1	7	4	3	0	6	4	4	123
Other Specialized Aboriginal Site	1	3	0	0	1	3	2	4	0	0	0	0	0	0	0	0	14
Petroglyph/ Pictograph	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Quarry	0	0	0	0	0	1	2	3	0	1	1	0	0	0	0	0	8
Rock shelter/cave	0	3	0	1	1	4	0	5	0	2	1	0	0	0	0	1	18
Unknown Function Open Site Greater than 20 m Radius	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	2
Unknown Function Surface Scatter Less than 20 m Radius	0	2	0	0	1	2	0	0	0	0	0	0	0	0	0	0	5
Village	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	1	4
(blank)	0	93	0	1	15	34	1	0	5	8	0	0	1	6	0	22	186
Total	1	597	7	12	222	267	29	22	14	24	10	3	1	27	10	34	1280

Early Archaic

The PASS database records 59 sites with Early Archaic components in Region 5. Similar to the preceding Paleoindian Period, Early Archaic sites in Region 5 are predominately found in lowland physiographic settings. Nearly all of the Early Archaic sites in the PASS database were part of a multi-component site. The single-component Early Archaic sites include four Open habitation, prehistoric sites and one site without a site type entered in the database. One of the Open habitation, prehistoric sites was located on a flood plain, two were on a terrace, and the other Open habitation, prehistoric site did not have a landform type entered in the PASS database.

Middle Archaic

The PASS database includes 125 sites with Middle Archaic components in Region 5. As noted previously, the Middle Archaic period in the Susquehanna River Valley was a time of apparent dramatic population increase, with over twice the number of sites with Middle Archaic components in comparison to the preceding Early Archaic period. Continuing an apparent trend, almost all of the Middle Archaic sites in Region 5 are found in lowland settings, with 54.4% of all Middle Archaic sites occurring in flood plain settings alone (n = 68). There are only 11 single-component Middle Archaic sites types recorded in the PASS database: eight Open habitation, prehistoric sites and three Open prehistoric site, unknown function sites. These two types represent likely candidates for occupation sites. All of the single-component sites occur in lowland settings. 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; thus, the single-component sites may represent Middle Archaic base camps.

Late Archaic

The PASS database includes 341 sites with Late Archaic components in Region 5. There are 169 Late Archaic multi-component sites that also possess a Terminal Archaic component (49.6% of all Late Archaic sites), indicating a continuity of occupational use of these sites during these periods by local Archaic peoples.

Late Archaic sites in Region 5 appear to strongly focus toward lowland physiographic settings. There are 79 single-component Late Archaic sites. 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 (n = 53) and Open prehistoric site, unknown function (n = 10). The other single-component site types include Lithic reduction sites (n = 7) and sites without identified site types in the PASS database (n = 9). All of these site types are predominately found in lowland settings.

Terminal Archaic

The PASS database includes 272 sites with Terminal Archaic components in Region 5, a decrease from the Late Archaic period. There are 181 Terminal Archaic multi-component sites possessing Late Archaic and Early Woodland components, representing 66.5% of the total population of Late Woodland sites. The fact that Terminal Archaic site components are strongly associated with preceding Late Archaic and subsequent Early Woodland components suggests group continuity within Region 5 between the Late Archaic and Early Woodland periods.

Terminal Archaic sites in Region 5 show a very strong focus toward lowland physiographic settings, with 92.3% of all Terminal Archaic sites found in that setting (n = 251). Flood plain settings alone account for 65.1% of all Terminal Archaic site locations (n = 177). 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 exclusively found in lowland settings. The Open prehistoric site, unknown function site type, which may represent short-term resource extraction camps, are evenly divided between upland and lowland landforms, although there are only two such sites in the PASS data, so this distribution cannot be said to represent a pattern or trend.

Early Woodland

The PASS database includes 112 sites with Early Woodland components in Region 5. There are 95 Early Woodland multi-component sites possessing Terminal Archaic and Middle Woodland components, representing 84.8% of the total population of Early Woodland sites. Early Woodland site components are strongly associated with preceding Terminal Archaic and subsequent Middle Woodland components, suggesting group continuity within Region 5 between the Terminal Archaic and Middle Woodland periods, even though there is a decline in total site numbers by nearly 60% from the preceding Terminal Archaic period. This decline may represent a population decrease, such as through out-migration; alternatively, the decline in site numbers could reflect a coalescence of small Archaic bands into larger Woodland groups as a more sedentary lifestyle was adopted. In this scenario, the overall population has not significantly decreased, but is instead concentrated at fewer sites with higher occupational densities. A third possibility is that the decline in site frequency represents both scenarios, as the coalescence of Archaic groups into Woodland settlements outside of Region 5 resulting in the migration of people to locations outside of the region.

Early Woodland sites in Region 5 show a marked focus toward lowland physiographic settings, with 101 Early Woodland sites identified in lowlands (83.5% of all Early Woodland sites). 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. There are very few examples of either site type for the Early Woodland period in Region 5, with four Open habitation, prehistoric sites and two Open prehistoric site, unknown function sites (all found in lowland settings). No single-component ceremonial sites (burial mounds, earthworks) or sites indicative of a more sedentary lifestyle (such as villages) are present in the PASS data.

Middle Woodland

The PASS database includes only 85 sites with Middle Woodland components in Region 5. There are 74 Middle Woodland multi-component sites possessing both Early and Late Woodland components, representing 88.1% of the total population of Middle Woodland sites. The fact that Middle Woodland site components are very strongly associated with preceding Early Woodland and subsequent Late Woodland components suggests group continuity within Region 5 between the three Woodland periods. The decline in site frequency first noted for the Early Woodland period in Region 5 continued in the Middle Woodland period, with 24.1% fewer Middle Woodland sites than in the Early Woodland. The reason for the decline in site frequency could be a continuation of populations aggregating at fewer numbers of sites, but the site types identified for both single-component and multi-component Middle Woodland sites do not suggest that hamlets or villages existed in Region 5 during this time period. Additionally, no ceremonial sites attributable to the Middle Woodland are present in Region 5. One possibility that could explain the decline in site frequency during both the Early and Middle Woodland periods is that the local Woodland groups moved out of the region to be closer to ceremonial centers elsewhere in the state, while still returning to Region 5 on resource acquisition forays.

Middle Woodland sites in Region 5 show a marked focus toward lowland physiographic settings, with 88.2% of all Middle Woodland sites located in lowlands. There are only five single-component Middle Woodland sites in the PASS database: four Open habitation, prehistoric sites and one site without an identified site type. The Open habitation, prehistoric site type may represent the likeliest candidate for seasonal occupation sites, such as base camps and short-term resource extraction camps; the four such sites in Region 5 are all found on flood plains.

Late Woodland

The PASS data for Region 5 includes 293 sites with Late Woodland components in Region 5. There are 56 Late Woodland multi-component sites possessing Middle Woodland components, representing 19.1% of the total population of Late Woodland sites, a reflection of the huge increase in site frequency between the Middle and Late Woodland periods. This increase in site frequency indicates either a large population explosion in residential groups, or an influx of Late Woodland groups expanding into the region from elsewhere; the latter explanation seems the likeliest hypothesis for the dramatic increase in site numbers from the Middle Woodland to the Late Woodland.

Late Woodland sites in Region 5 show a very strong focus toward lowland physiographic settings, with 93.8% of all Late Woodland sites occurring in lowland settings. Village sites are perhaps the defining site type for the Late Woodland. There are four Late Woodland villages in the PASS data, with three in lowland settings and one without landform data. Two Late Woodland Burial Mounds are present: one on a flood plain and the other on a terrace. A cemetery is also present on a flood plain. 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; and Rock shelter/cave. Open Habitation, prehistoric sites, which likely include a number of base camp sites, are nearly exclusively found in lowland settings (98.1%). The Open prehistoric site, unknown function and Rock shelter/cave site types, which may represent short-term resource extraction camps, occur predominantly in lowland settings (87.5%).

REGION 6 SITES

There are 376 archaeological sites with prehistoric components in Region 6 (Table 5 shows a breakdown of the Region 6 sites by site type and landform; individual tables for each of the time periods are included in Appendix B). A total of 198 sites in the PASS database did not possess diagnostic material and were not assigned to a temporal period. In addition, there are 26 Archaic-period sites that could not be assigned to one of the Archaic sub-periods, and 32 Woodland-period sites with a similar issue.

Sites in Region 6 are primarily found in lowland settings, with 72.6% of all sites (n = 273) located in lowland settings. The flood plain landform alone accounts for 53.7% of all site locations in Region 6 (n = 202). Two site types are only found in lowland settings in Region 5, including the Isolated find and Other specialized aboriginal site; no site types are exclusive to upland settings. The Burial mound site type may also be restricted to lowland settings, as all such sites with landform data in the PASS database are in lowlands (one burial mound did not have landform data). The most commonly occurring site type is Open habitation, prehistoric (n = 221), which occurs almost entirely in lowland settings. The apparent trend toward site location in lowland settings in Region 6, however, may simply reflect survey bias. As with Regions 4 and 5, visual analysis of survey areas in the PASS system suggests survey effort in Region 6 has been largely directed at lowland settings. Thus, the lack of sites in upland settings may reflect survey bias rather than an actual prehistoric landform preference. Indeed, survey coverage may also be the best explanation for the apparent low frequency of sites occurring across all time periods in Region 6 when compared to Regions 4 and 5.

Paleoindian

Within Region 6, there have been 13 sites identified with Paleoindian components. Seven sites with Paleoindian components also contain one or more components dating to later time periods.

Paleoindian sites in Region 6 with landform data included in the PASS database are only found in lowland physiographic settings, with 53.8% of all Paleoindian sites located on flood plains. Of the single-component Paleoindian sites, the Open habitation, prehistoric site type may represent camp locations.

Table 5 - Region 6 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
Burial Mound	0	7	1	0	0	1	0	0	0	0	0	0	0	0	0	1	10
Cemetery	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2
Earthwork	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Isolated Find	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
Lithic Reduction	0	6	1	0	0	0	0	0	2	1	0	0	0	0	0	1	11
Open Habitation, Prehistoric	0	148	7	2	4	29	4	4	0	1	7	0	1	1	2	11	221
Open Prehistoric Site, Unknown Function	0	10	0	0	0	8	0	0	0	0	0	0	0	0	1	4	23
Other Specialized Aboriginal Site	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Rock shelter/cave	0	1	0	0	1	1	4	18	1	0	8	0	2	1	7	1	45
Unknown Function Open Site Greater than 20 m Radius	0	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	4
Unknown Function Surface Scatter Less than 20 m Radius	0	2	0	0	0	2	0	0	0	0	0	2	1	1	0	3	11
Village	0	5	0	0	1	3	0	0	1	1	0	0	0	0	0	1	12
(blank)	0	13	3	0	4	3	0	0	2	0	0	2	0	0	1	3	31
Total	0	202	12	2	10	47	8	22	6	3	15	5	4	4	11	25	376

Early Archaic

The PASS database records only four sites with Early Archaic components in Region 6. Such a small number of sites does not lend itself to identifying trends in site types or locations in relation to physiographic settings. All of the Early Archaic sites in Region 6 co-occur with other prehistoric components; there are no single-component Early Archaic sites present in the PASS database. The Region 6 Early Archaic sites occur mainly in lowland settings.

Middle Archaic

The PASS database includes 13 sites with Middle Archaic components in Region 6. Middle Archaic sites in Region 6 are nearly all found in lowland physiographic settings; one site did not have landform information in the PASS database, while all the other sites were in the lowlands. Flood plain settings account for 61.5% of all Middle Archaic site locations in Region 6 (n = 8). There are five Open habitation, prehistoric sites that may represent resource exploitation camps. The other eight Middle Archaic sites are all multi-component sites. As with the Early Archaic, the low number of Middle Archaic sites makes identification of possible trends in site types or locations unfeasible.

Late Archaic

The PASS database includes 48 sites with Late Archaic components in Region 6, an increase in site frequency by a factor of 3.7, indicating either a population increase or the expansion into Region 6 of Late Archaic groups from outside the region. Late Archaic sites in Region 6 show a focus toward lowland physiographic settings, with 70.8% of all Late Archaic sites occurring in lowlands (n = 34). For the first time in Region 6, however, prehistoric groups appear to have moved into the uplands.

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 Rock shelter/cave. The Open habitation, prehistoric sites, which likely include a number of base camps, are evenly split between upland and lowland settings (one site did not possess landform information). No landform data was available for the two Open prehistoric site, unknown function sites. One upland single-component Late Archaic rock shelter is present in Region 6. A burial mound (36WA82) listed with a Late Archaic affiliation is present in the PASS database for Region 6; it is unclear from the PASS data if this is actually a Woodland mound with some Late Archaic diagnostics present in the same location. The mound has not been professionally evaluated.

Terminal Archaic

The PASS database includes 41 sites with Terminal Archaic components in Region 6. There are 22 Terminal Archaic multi-component sites also possessing either or both Late Archaic and Early Woodland components, representing 53.6% of the total population of Terminal Archaic sites. The fact that Terminal Archaic site components are commonly associated with preceding Late Archaic and subsequent Early Woodland components suggests a certain degree of group continuity within Region 6 between the Late Archaic and Early Woodland periods.

Terminal Archaic sites in Region 6 show a marked focus toward lowlands, with 80.5% of all Terminal Archaic sites located in that physiographic setting. 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; Open prehistoric site, unknown function; and Rock shelter/cave. All the open sites are exclusively found in lowland settings, although any apparent landform preferences in the data could be a result of survey bias due to the low number of sites. The one rock shelter site is located in an upland setting.

Early Woodland

The PASS database includes 32 sites with Early Woodland components in Region 6, showing a decline in frequency from the preceding Late and Terminal Archaic periods. There are 22 Early Woodland multi-component sites possessing Terminal Archaic and Middle Woodland components, representing 68.7% of the total population of Early Woodland sites. Early Woodland site components are strongly associated with preceding Terminal Archaic and subsequent Middle Woodland components, suggesting group continuity within Region 6 between the Terminal Archaic and Middle Woodland periods.

Early Woodland sites in Region 6 show a marked focus toward lowlands, with 78.1& of all Early Woodland sites occurring in that physiographic setting. One Early Woodland earthwork (36WA305) is present in Region 6. 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 Rock shelter/cave, represented by two sites in each site type category in Region 6. The two Open habitation, prehistoric sites, possibly representing base camps, are both located in lowland settings, while the two rock shelters, possibly representing short-term resource extraction camps, are both located in the uplands. The earthwork is located on a flood plain.

Middle Woodland

The PASS database includes 29 sites with Middle Woodland components in Region 6. There are 24 Middle Woodland multi-component sites possessing either or both Early and Late Woodland components, representing 82.6% 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 6 between the three Woodland periods.

Middle Woodland sites in Region 6 show a marked focus toward lowlands, with 75.9% of Middle Woodland sites located in that physiographic setting. Two burial mounds located in the lowlands represent the only single-component Middle Woodland ceremonial site type. One isolated find and a site without site type information in the PASS database are the other two single-component sites. There are no single-component Middle Woodland site types that may represent the candidates for seasonal occupation sites in Region 6, such as base camps and short-term resource extraction camps.

Late Woodland

The PASS data for Region 6 includes 73 sites with Late Woodland components. There are only 17 Late Woodland multi-component sites possessing Middle Woodland components, representing 23.3% of the total population of Late Woodland sites. The doubling in frequency of occurrence from the Middle to the Late Woodland may obscure the relationship between Middle and Late Woodland groups. A population increase could lead to more sites being occupied, although a movement into Region 6 by outside Late Woodland groups would also explain the increase in site frequency.

Late Woodland sites in Region 6 show a general focus toward lowlands, with 76.7% of all Late Woodland sites occurring in that physiographic setting. Flood plain settings alone account for 58.9% of all Late Woodland sites. Village sites are perhaps the defining site type for the Late Woodland. There are six Late Woodland villages in the PASS database, with five occurring in lowland settings and a single village located on a hilltop. Single-component Late Woodland site types that may represent likely 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 Rock shelter/cave. Open habitation, prehistoric sites, which likely include a number of base camp sites, are only found in lowland settings. The Open prehistoric site, unknown function site type, which may represent short-term resource extraction camps, shows 66.6% of this site type occurring in lowland settings. There are four rock shelter/caves, three of which are in upland settings; these sites likely represent seasonal camps.

3
DATA QUALITY – REGIONS 4, 5, AND 6

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 database. 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 4, 5, and 6.

METHODS

Within Regions 4, 5, and 6, PASS forms and CRGIS data were examined for 283 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 6 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 database. Discrepancies between the two data sets were categorized using the ranking outlined in Table 6

Table 6 - Rating Criteria for Site Data

Rating	Criterion
	<u>Location Accuracy, PASS Form</u>
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.
	<u>How Well Location is Reflected in CRGIS</u>
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.
	<u>Artifact Data Quality, PASS Form</u>
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.
	<u>How Well Artifacts are Reflected in CRGIS</u>
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.
	<u>PASS Form Completeness</u>
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.
	<u>PASS Form Type</u>
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 4

PASS forms and CRGIS data were examined for a total of 164 sites within Region 4.

Location Accuracy

Examination of the PASS forms in Region 4 indicates that 48% of the sites sampled are mapped with medium to high accuracy (that is, on detailed maps or USGS topographic quadrangles), while 52% of the sites are poorly mapped or provide little location information (Figure 8). By comparison, 96% of the same site sample has accurately mapped locations in the CRGIS database, and another 3% are mapped within 250 m from the location indicated on the PASS forms (Figure 9). Just 1% of sites in the sample remained unmapped, suggested an increase in mapping accuracy in CRGIS as compared to the PASS forms.

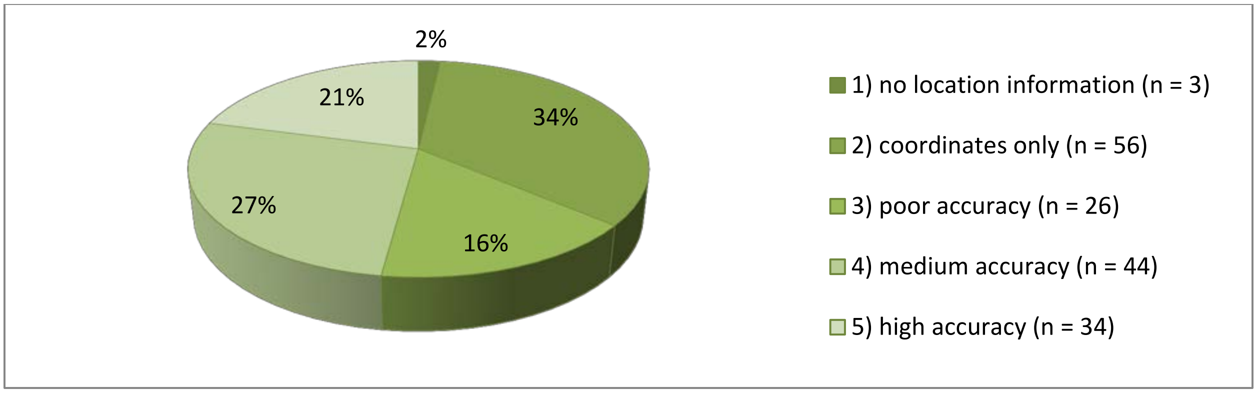


Figure 8 - Quality of location information on PASS forms within Region 4.

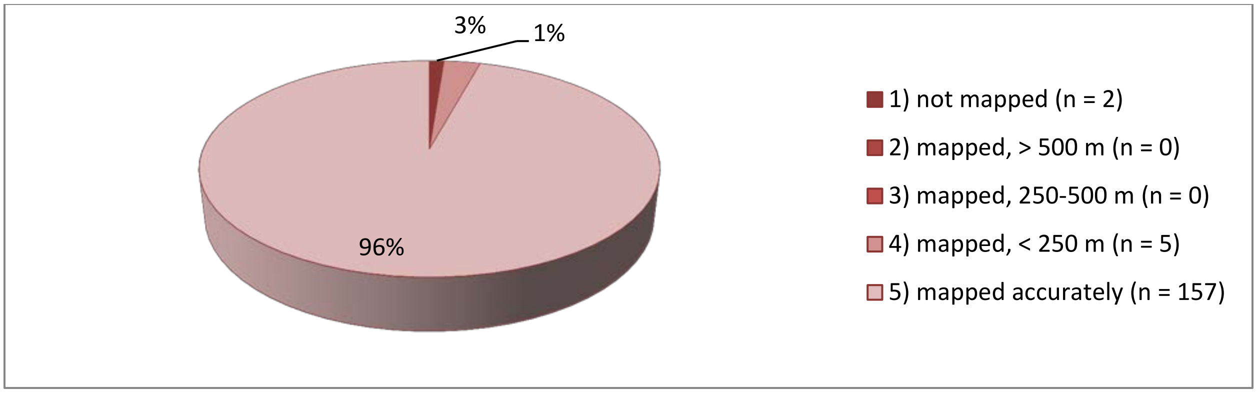


Figure 9 - Quality of location information reflected in CRGIS within Region 4.

Artifact Data

More than half (56%) of the archaeological site sample of PASS forms within Region 4 contain artifact descriptions that can be described as good or moderate. The remaining 44% of sites contain poor descriptions or none at all (Figure 10). Just over one-third (37%) of the sites in the CRGIS database contain artifact information that accurately matches the information found in their PASS forms, while almost as many (32%) contain more artifact data than the original PASS form, and another 9% have good quality artifact data (Figure 11). The percentage of sites with no artifact data in the CRGIS database (18%) is greatly reduced from the percentage of sites with no artifact data in the PASS forms (34%).

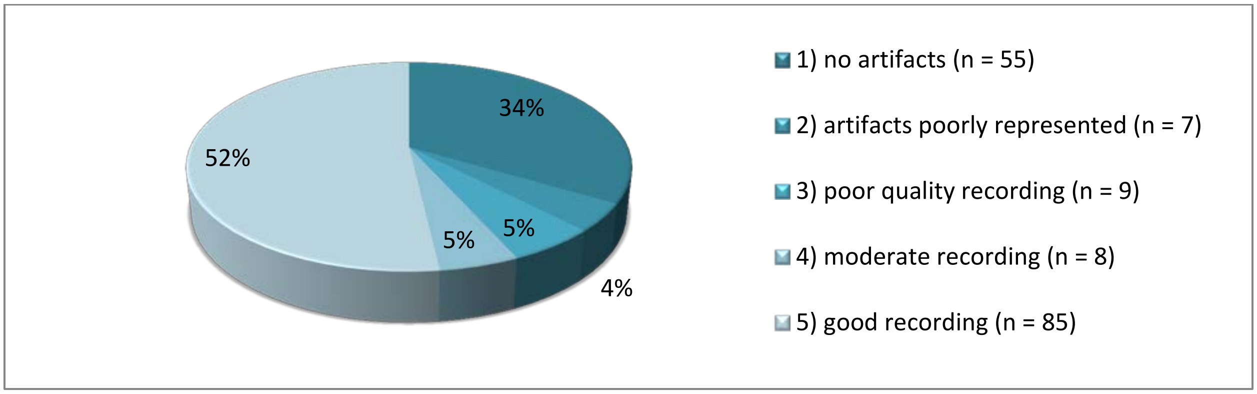


Figure 10 - Original artifact data recorded on PASS forms for Region 4.

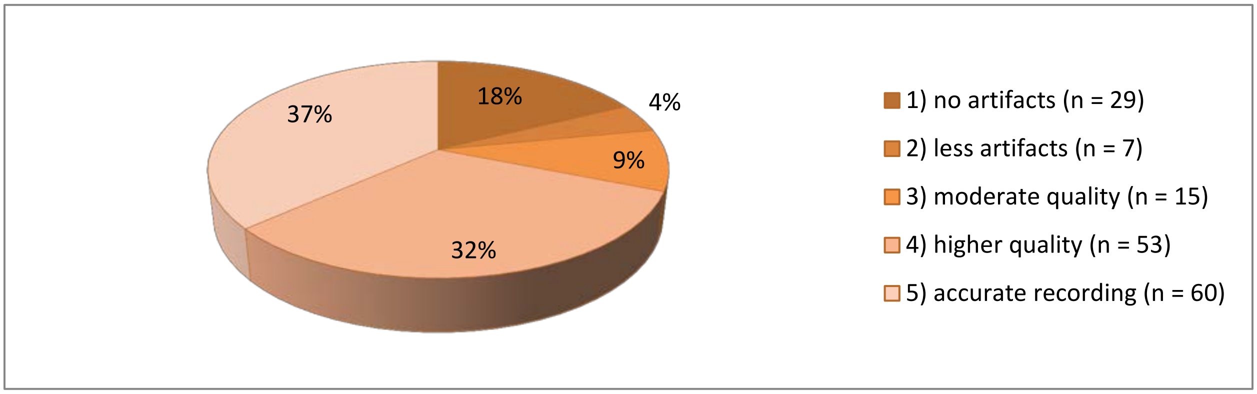


Figure 11 - Artifact data reflected in the CRGIS database for Region 4.

PASS Form Types and Completeness

A little over half (57%) of the PASS forms in the site sample from Region 4 are up to or greater than 75% complete (Figure 12). The remaining 43% of PASS forms in the site sample contain limited data. Almost all (96%) of the site sample for Region 4 is recorded on old version or middle version PASS forms, while only 4% are recorded on the newer version of the form that includes detailed artifact information (Figure 13). This suggests that for Region 4, the most reliable site information is likely to be locational rather than artifact data.

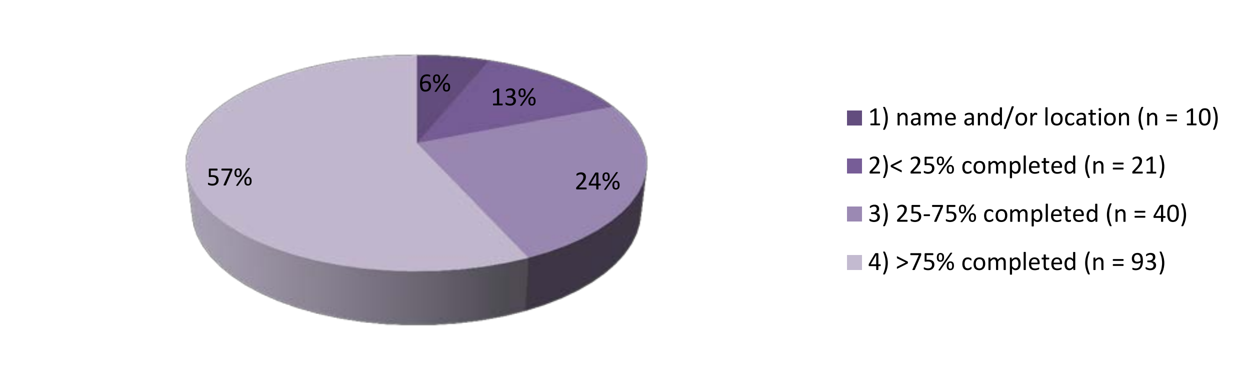


Figure 12 - Completeness of PASS form information in Region 4.

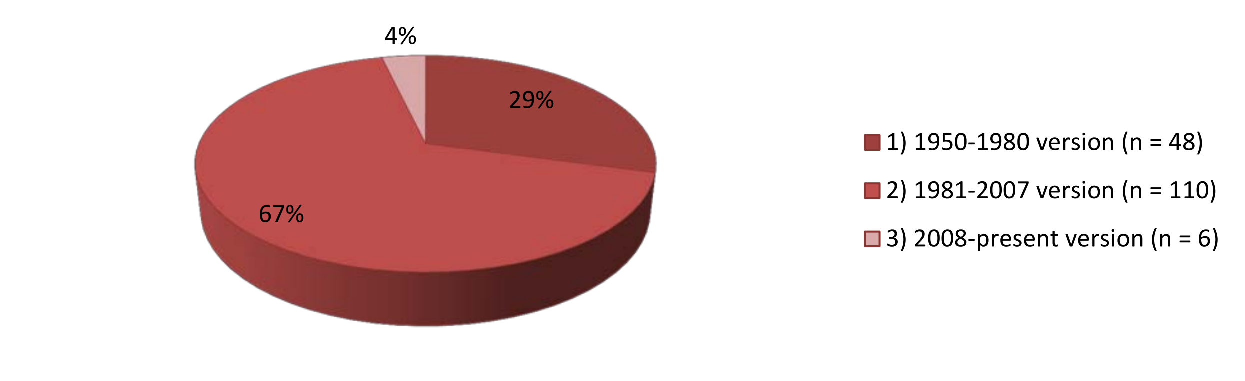


Figure 13 - Distribution of PASS form types in Region 4..

REGION 5

Within Region 5, PASS forms and CRGIS data were examined for 83 sites.

Location Accuracy

Of the 83 sites in the Region 5 sample, 39% are mapped on USGS maps or contain highly detailed maps on the PASS forms. The remaining 61% of forms contain no location data, are only referenced by coordinates, or contain unreliable hand drawn maps (Figure 14). Within the CRGIS database, almost all (96%) of the site locations match the mapping in the PASS forms. Two sites (3%) were mapped within 250 m of the locations indicated on the PASS forms, and just one site (1%) was not mapped. (Figure 15).

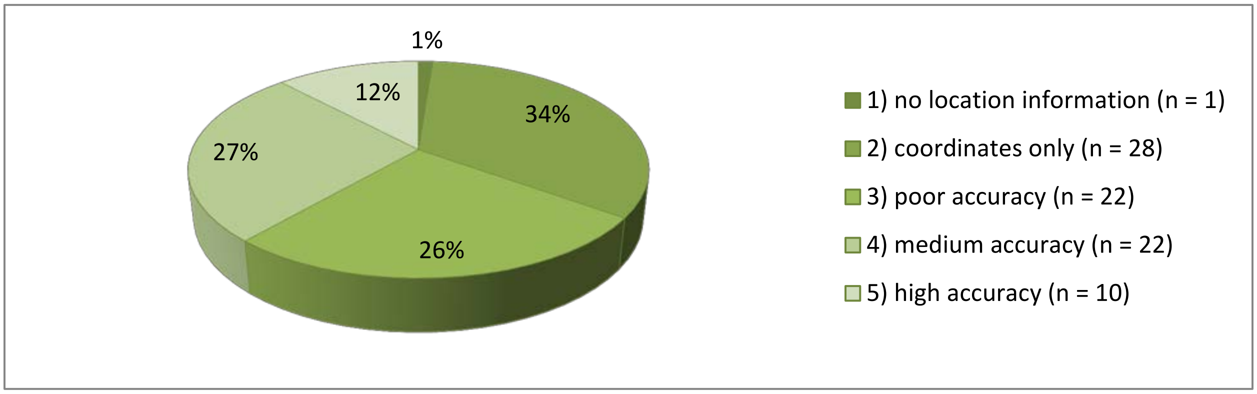


Figure 14 - Quality of location information on PASS forms within Region 5.

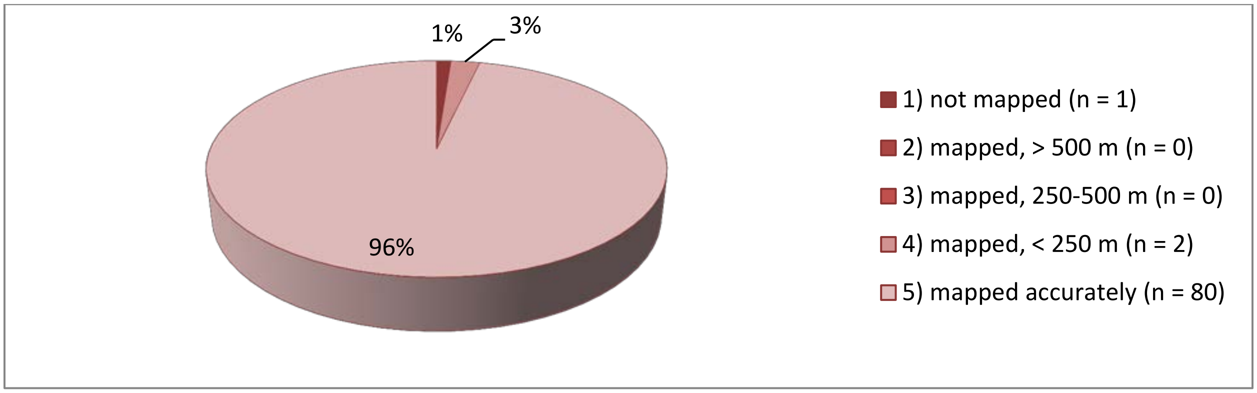


Figure 15 - Quality of location information reflected in CRGIS within Region 5.

Artifact Data

More than half (63%) of the site sample in Region 5 has good or moderate artifact description on the PASS forms, while just 8% have poor quality data and 29% have no artifact data at all (Figure 16). By comparison, a full 75% of the sites in the Region 5 site sample have moderate to high quality artifact data, while only 2% have poor quality artifact data and 23% have no data (Figure 17), suggesting that data quality was improved in the transition from PASS forms to CRGIS.

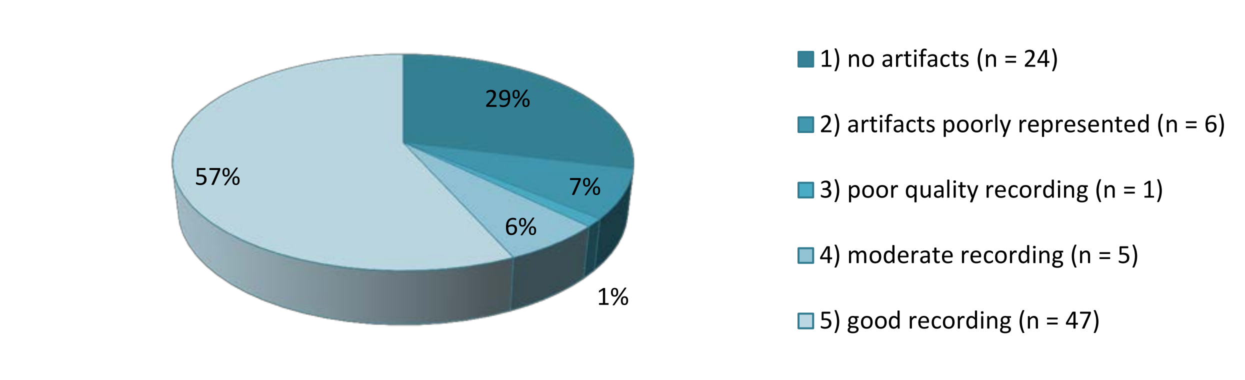


Figure 16 - Original artifact data recorded on PASS forms for Region 5.

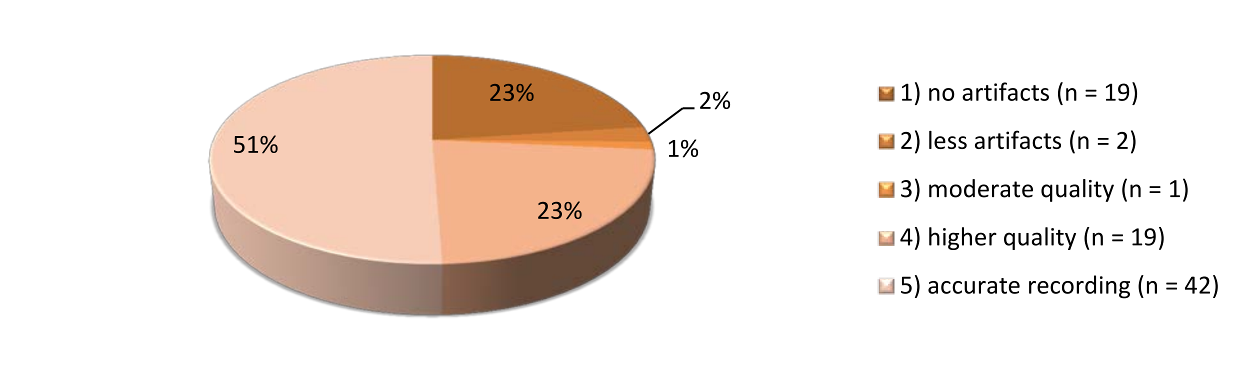


Figure 17 - Artifact data reflected in the CRGIS database for Region 5.

PASS Form Types and Completeness

Of the 83 total sites sampled within Region 5, nearly two-thirds (64%) are at least 75% complete. The remaining 36% of the forms contain limited data (Figure 18). The PASS form types for Region 5 are almost all (94%) either older or middle version, with just 6% on new forms with detailed artifact data (Figure 19).

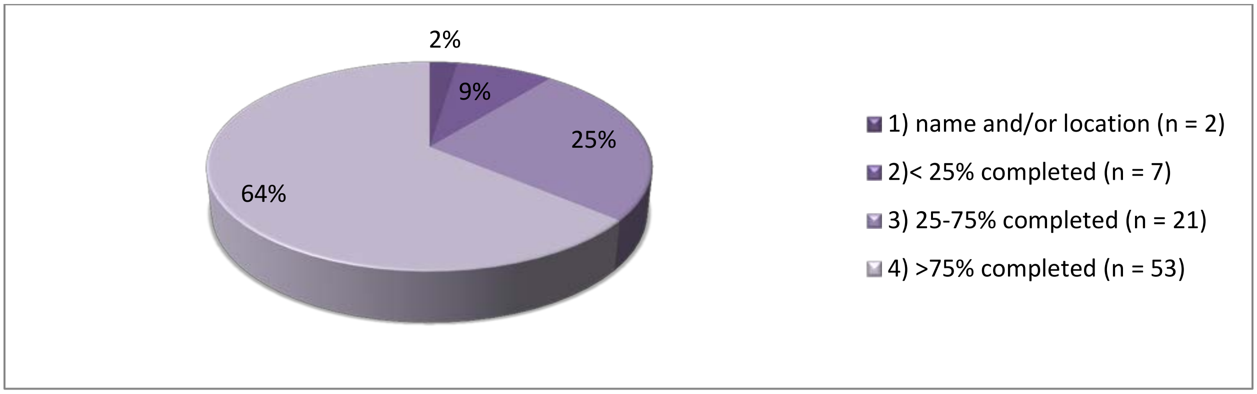


Figure 18 - Completeness of PASS form information in Region 5.

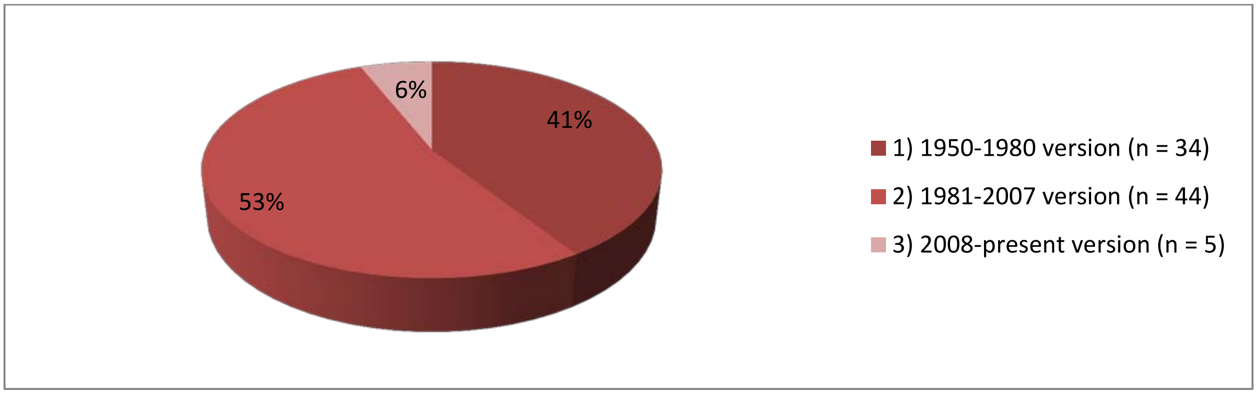


Figure 19 - Distribution of PASS form types in Region 5.

REGION 6

A total of 39 prehistoric archaeological sites were included in this analysis for Region 6.

Location Accuracy

The accuracy of mapped locations for the site sample within Region 6 is more evenly distributed among the categories than in the other two regions (Figure 20). A little over half (54%) of the sites were mapped on USGS maps or other highly detailed maps, while the remaining half (46%) were either not mapped (8%) or were poorly mapped (38%). Within the CRGIS database, a full 90% of sites are mapped accurately or within 250 m of the location indicated in the PASS database (Figure 21). Four sites (10%) are not mapped in CRGIS, which is one more than in the PASS forms.

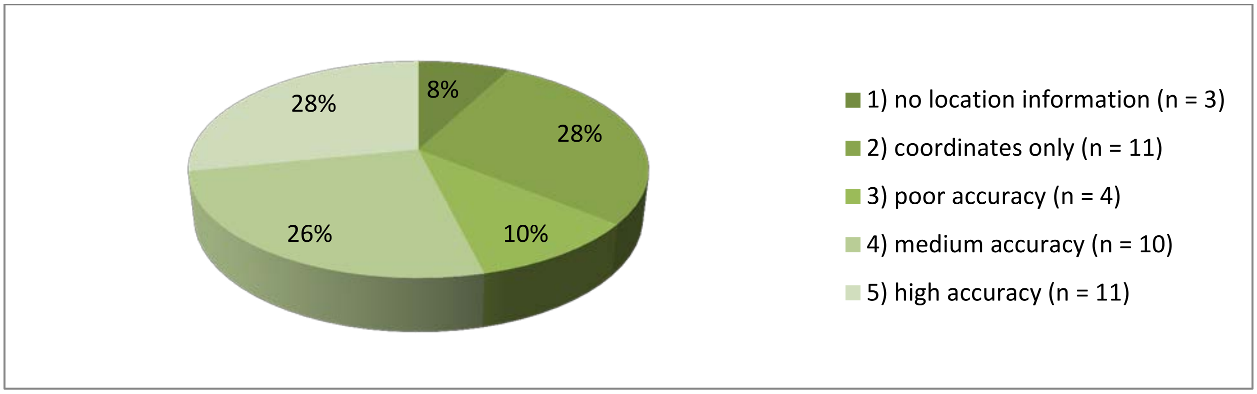


Figure 20 - Quality of location information on PASS forms within Region 6.

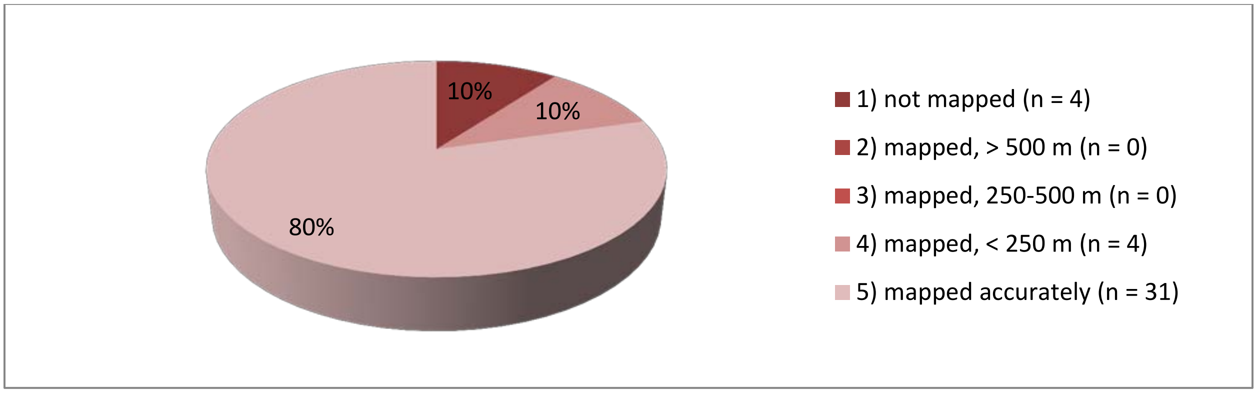


Figure 21 - Quality of location information reflected in CRGIS within Region 6.

Artifact Data

Equal numbers of sites in the Region 6 sample (44%) have good artifact data and no artifact data on the PASS forms (Figure 22). In between those two extremes is just 2% of sites with moderate artifact data quality and 10% with poor artifact data quality. The transition to CRGIS appears to have improved the artifact data quality to some extent, with 57% of sites having high quality or accurate artifact data, while 12% have moderate artifact data quality and 31% have no artifact data (Figure 23).

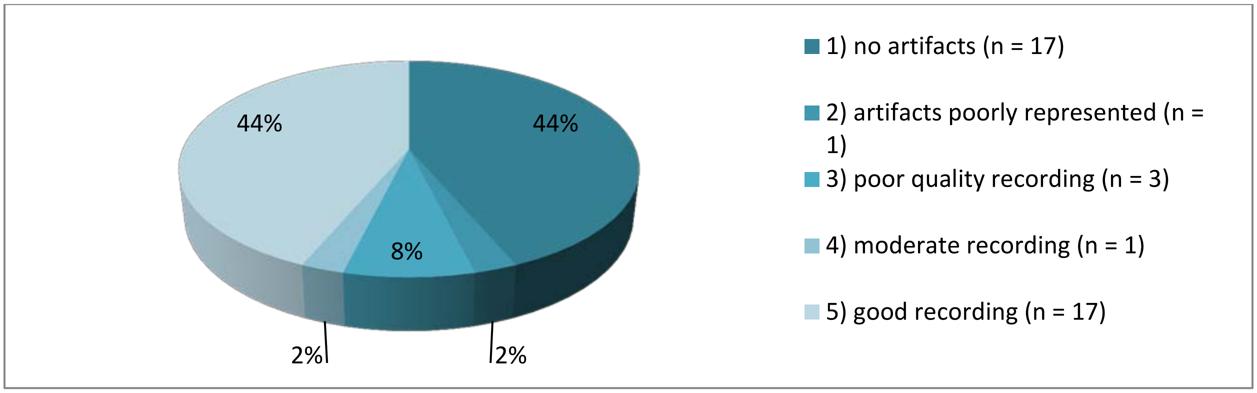


Figure 22 - Original artifact data recorded on PASS forms for Region 6.

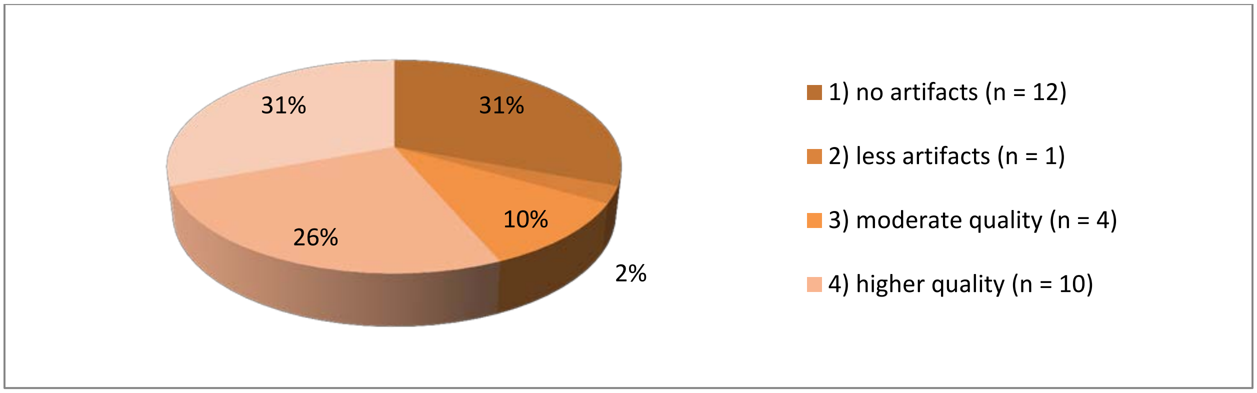


Figure 23 - Artifact data reflected in the CRGIS database for Region 6.

PASS Form Types and Completeness

The Region 6 results are similar to the two previous regions: more than half (54%) of the PASS forms in the site sample are at least 75% complete and a quarter (26%) are only minimally completed (Figure 24). The majority of the forms (90%) are the early or middle version (Figure 25). The large number of middle version forms, which are often filled out completely or contain very little missing data, probably accounts for the overall completeness of the site sample.

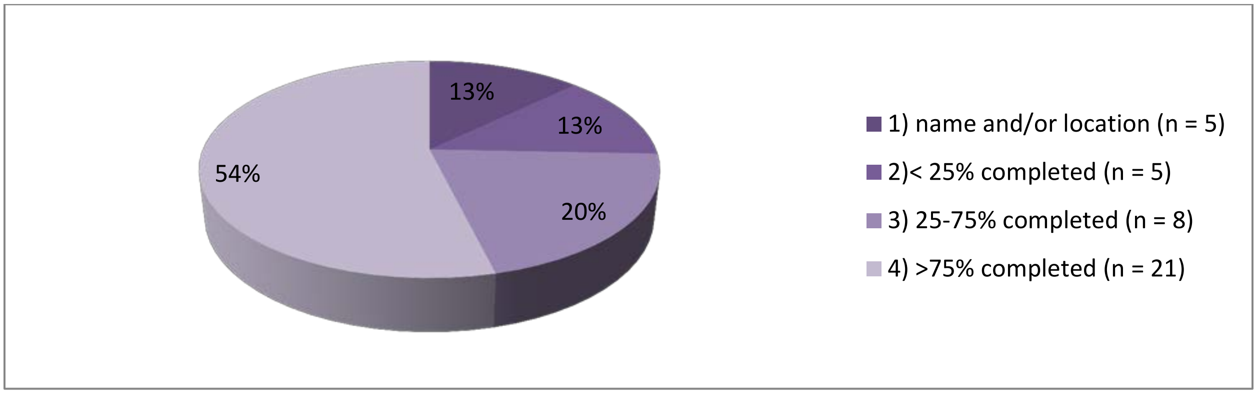


Figure 24 - Completeness of PASS form information in Region 6.

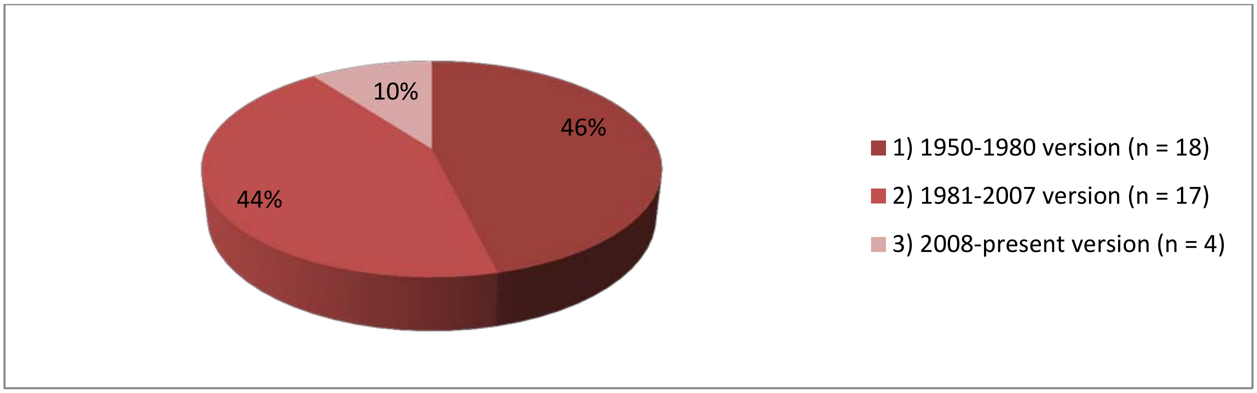


Figure 25 - Distribution of PASS form types in Region 6.

CONCLUSIONS

Overall, the analysis shows that the data derived from the CRGIS database are at least as complete and accurate as the data included in the original PASS forms, and in some cases, more so. Although the sample of 286 sites in Regions 4, 5, and 6 includes the same number (n = 7) of unmapped sites in the PASS forms and the CRGIS database, errors and missing information on the PASS forms were addressed in the transition to CRGIS. Mapping locations in CRGIS diverged very little from locations provided on the PASS forms, reflecting the accurate transcription of data: of the 286 sites in the sample, 4% (n = 11) sites were mapped 250 m or more from the locations shown on the PASS forms, which may in fact represent an improvement in mapping accuracy.

Of the 286 PASS forms examined for Regions 4, 5 and 6, 149 (52%) contain good artifact data, while 96 (34%) contain no artifact data, with both categories accounting for 86% 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 database: artifact information in the CRGIS database matched the information in the PASS form for 114 of the 286 sites (40%). Further, the quality of artifact data was improved upon in the CRGIS data for 82 (29%) of the 286 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 provides for more thorough recordation of site locations and artifact data. Most of the sites considered for this analysis (60%; n = 171) were recorded on the “middle” version of the PASS form and 58% (n = 167) were considered at least 75% complete. These forms do not include as much information as the newer version, and the data in the CRGIS is therefore limited.

4
MODEL METHODOLOGY – REGIONS 4, 5, AND 6

The general approach to modeling Regions 4, 5, and 6 followed the same process used for Regions 1, 2, and 3. The methodology is documented in detail in the Task 3 report (Harris 2014), with adaptations documented in the Task 4 report (Harris et al. 2014). Broadly, the steps leading to the final sensitivity model are as follows:

- 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, Multivariate 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.

A single yet significant change was applied to the Task 5 models that warrants further discussion. This change is the inclusion of soils data as environmental variables in the modeling process. Described below are the aspects of the modeling methodology that have changed as a result of including these data.

ADAPTATION 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. As discussed in the Task 4 report (Harris et al. 2014), changes to the process 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; 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 (Cohen 1960), which is a compliment to the Kvamme Gain and will be used to assess model performance. These improvements were carried over into the modeling for Regions 4, 5, and 6 without any significant changes. However, the Task 5 methods incorporated a new advancement that was not present in the previous Task 4 or Pilot Model studies: the inclusion of soils data as a predictive variable. The text below describes how the models for Regions 4, 5, and 6 incorporate data derived from soil surveys.

Soils Data

Within Regions 4, 5, and 6, four environmental factors derived from soils data were incorporated into the modeling process. Two factors represent soil drainage types under dry ("drcdry") and wet ("drcwet") conditions. One factor represents the available water capacity at a depth of 50 cm below the surface ("aws050"). The fourth attribute represents the agricultural capability of a soil under non-irrigated conditions ("niccdcd"). These data were derived from the United States Department of Agriculture's (USDA) National Cooperative Soil Survey as hosted through the USDA Web Soil Survey (WSS) portal.

The inclusion of attributes derived from soils survey data as environmental variables into the modeling process is not uncommon in the APM literature and is very intuitive to archaeologists given the we undertake our work almost entirely within the soil. Generally, it is understood that the variation in soil drainage, depth, texture, composition, and other factors correlate to the distribution of archaeological sites, or at least suggest patterning. This assumption holds even though the soils of today may or may not accurately reflect the soil conditions of the archaeological sites within them. However, as with many of the variables incorporated into empirical models such as these, the causation is secondary to correlation. The methods for employing soils data into these models do not require an *a priori* assumption of soil preference. For example, many archaeologists will not quibble if you say that sites are frequently located on well-drained soils. In fact, according to the locations of 18,000 prehistoric sites recorded in the PASS database, sites are found with a higher than expected frequency in poorly and very poorly drained soils and a lower than expected frequency in well-drained soils state-wide. Within the modeling process undertaken here, the relative frequency of sites versus background locations determines the importance, or lack thereof, of drainage classes or other soil attributes. However, the inclusion of these data into the statistical modeling process requires the consideration of a few issues, namely: 1) the aggregation of soils data into map areas; and 2) the inclusion of the nominal data type (e.g., categorical data) in the form of factors. The text below will discuss these issues.

Through the WSS, soils can be mapped and downloaded for counties or smaller, arbitrary study areas. The downloaded data comes in the form of a spatial data layer and a Microsoft Access database of tabular data. The spatial data, represented as polygons, and tabular data have a single identifier in common: "mukey" (short for map unit key). The database contains numerous tables that describe many attributes of soil units such as engineering, agricultural, industrial, and so on. In the database, the data are organized by a different identifier than in the spatial data: the component key.

The component key describes spatial areas smaller than those mapped in the spatial data under mukey. The database attributes that correspond to any one mapped soil unit (mukey) may be divided among numerous unmapped components (component key), and the attributes of the component may vary widely. Therefore, you can have one mapped soil unit with multiple database entries for drainage class that range from poorly to well drained. The reason for this many-to-one relationship is that the soil scientists who create these data understand that there is a great deal of variation within a soil unit and that variations in attributes such as drainage class vary with the other attributes. However, this level of variation is not suitable for mapping on the scale of spatial polygons—polygons that are very much a legacy of the previous county soil survey books. In order to assign a single value, such as drainage class, to a single mapped soil unit polygon, the numerous components must be aggregated to remove the variation. In order to flatten the many-to-one relationship between components and map units, a weighted average of the components is calculated so that the attributes of the component that contributes the most to the map unit is selected to represent that map unit. For each of the four attributes utilized in the modeling process, these data are collected into a new table, "muaggatt" (short for map unit aggregated), and joined to the spatial map units. This process was repeated for each of the 67 counties in Pennsylvania for each of the four soils attributes. Differences in attribute coding between counties and missing values were standardized and cleaned up, and each county was converted into a raster layer. These rasters were then mosaicked to form a state-wide coverage of each of these variables.

The inclusion of these four variables in the modeling process requires very careful consideration due to the way the statistical models and methods address nominal data. Previous to the inclusion of soils data, the environmental variables used were either ratio or interval data types. This means that the measurements, such as distance to wetlands, have a natural zero point (e.g., within a wetland), and the interval between units is the same magnitude for the same interval along the length of the scale (e.g., the interval between 5 and 6 m is the same as the interval between 1,005 and 1,006 m). These quantitative data types offer properties that are very convenient for mathematical operations such as those used in the models employed here. However, most of the soils attributes are not on the interval or ratio scale, but instead on the nominal scale, often referred to as categorical (the variable of "aws050" is a quantitative variable and is excluded from this treatment). With this type of qualitative data, there is no set zero point to reference, the interval between units is not standardized, and the order of the categories is essentially arbitrary. Categorical data do not have the same convenient arithmetic properties as quantitative data, but can still be utilized in the model process with some adjustments.

Briefly, the primary technical hurdle in incorporating categorical variables lies in addressing factors, factor levels, and model formula. Factors are the mechanism by which the R statistical language stores categorical data such as the soils attributes. In a factor, the data are stored as arbitrary integers and assigned a factor level that is that actual category. Therefore, the full value of each category label, such as "very well drained," does not need to be stored for the many thousands of times it occurs in a given region. Instead, a representative number is assigned to this category and tied to the label "very well drained." However, an issue that frequently arises in dealing with data sampling is that the levels must match from sample to sample and between training data and testing data, otherwise the models fail. This is due to the fact that the model cannot know how to predict for a level such as "excessively well drained" when the training data did not contain any data points with that label. One option is to assign a Null value to any new observation that contains an unknown label, but this potentially ignores valuable data. The other option is to ensure that a fully representative training sample is used or to insure that the training sample recognizes all of the labels that are in the testing data, regardless of whether they exist in the training sample or not. This project chose to harmonize the levels throughout the process so that the final model predictions are never surprised by levels that were not present before.

The other issue in dealing with categorical data is the reworking of the formula interface for each model. From the technical perspective, this change required a reworking of the way in which the statistical models interpret the relationship between the predictor variables and the outcome (e.g., site presence or absence). The previous method involved feeding the model a simple matrix of variables and the outcome, whereas the new method involves creating a formula to express the relationship between variables. This results in a formula such as y = x1 + x2 + … + xn, where y is the response, x1 through xn are the variables, and the plus sign is how the variables interact; in this case they are always additive. On the more general level, the change to a formula reflects how the categorical data are addressed within the statistical models. The use of categorical data requires the use of what are called dummy variables. Essentially, each level of a categorical variable is split into its own dummy variable so that every time that label is present it is coded as a 1 and when it is absent, coded as a zero. For example, for the seven levels of the "drcdry" variable, ranging from "excessively well drained" to "very poorly drained," an entirely new variable was created, leading to seven new model variables (one for each level). To continue the example, if in the original "drcdry" variable, there were 500 observations and 100 of them were coded as "excessively well drained," then the dummy variable created for the "excessively well drained" label would contain a 1 for each of the 100 observations that were "excessively well drained" and a zero for the rest of the observations that were some other drainage class. The next dummy variable would be created for the next level, "well drained," and ones and zeros applied to where they are present or absent, and so on until there is a dummy variable for each of the original levels. These dummy variables are then entered into the formula as essentially presence/absence variables for each category. The models then use the dummy variables to fit and predict. One drawback of this—again from the technical side—is that each factor level for each categorical variable is turned into a dummy variable thereby dramatically increasing the number of variables for any given model, in turn increasing the computation of fitting and predicting. However, the models created for this report indicate that the variations on these soil attributes are useful in distinguishing the pattern inherent in our known archaeological site locations.

5
MODEL VALIDATION – REGIONS 4, 5, AND 6

The total number of known archaeological sites within each of the 36 subareas range from as few as 3 sites to as many as 473 sites. The density, measured as the number of sites per square mile, ranges from a low of 0.009 to a high of 6.074, with riverine areas having a higher site density on average (1.578) than upland areas (0.095). With this high variability in the density of known site locations, both the suite of statistical models, Logistic Regression (LR), Multivariate Adaptive Regression Splines (MARS), and randomForest (RF) and the proportionally weighted model (Model 2) were used to try to find the best model to capture the available data. The proportionally weighted models were not used in Regions 1, 2, and 3 due to adequate numbers of sites, but with eight subareas containing 20 known sites or fewer in Regions 4, 5, and 6, it seemed likely that not all of the site samples would be adequate for the statistical models. Proportionally weighted models (Model 2) were created for each subarea that contained 20 or fewer known prehistoric PASS sites. The judgmentally weighted model (Model 1) was not created for any subareas within Regions 4, 5, or 6. The theoretical basis and technical components of these models are covered in detail in the Task 3 and Task 4 reports (Harris 2014; Harris et al. 2014).

This model validation section is organized by model type. For each of the 36 subareas for which models were created, a single model was selected as being the best balance between model fit, predictive ability, and the distribution of sensitivity values. As for Regions 1, 2, and 3, the metrics used to assess the most representative model include the Root Mean Squared Error (RMSE), Area Under the Curve (AUC), Kvamme Gain (KG) and Kappa (K) at a 0.5 threshold, with the thresholds calculated empirically from final sensitivity raster layers. Each of these metrics was presented and discussed in the previous Task 4 report (Harris et al. 2014). Table 7 lists the model type chosen to best represent each subarea. The text that follows will be organized by these model types, beginning with Model 2, followed by LR, MARS, and finally RF.

Table 7 - Selected Model Type for Each Subarea

Region	Zone	Subarea	Model Type
6	all	riverine section 1	LR
		riverine section 2	MARS
		riverine section 3	LR
		riverine section 4	LR
		riverine section 5	MARS
		upland section 1	LR
		upland section 2	MARS*
		upland section 3	Model 2
		upland section 4	Model 2
		upland section 5	Model 2
4/5	east	riverine section 1	RF
		riverine section 2	Model 2
		riverine section 3	MARS
		riverine section 4	RF
		riverine section 5	MARS
		riverine section 6	MARS
		riverine section 7	MARS
		upland section 1	LR
		upland section 2	Model 2
		upland section 3	MARS
		upland section 4	RF
		upland section 5	RF
		upland section 6	MARS
		upland section 7	RF
	west	riverine section 1	RF
		riverine section 2	MARS
		riverine section 3	RF
		riverine section 4	RF
		riverine section 5	MARS
		riverine section 6	MARS
		upland section 1	MARS
		upland section 2	RF
		upland section 3	RF
		upland section 4	RF
		upland section 5	RF
		upland section 6	RF

*two models: rock shelter and non-rock shelter sites

PREDICTOR VARIABLES

As with the previous models in Task 4, a large number of environmental variables was created and then pared down. The final selection was based on a variable's 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 (MW) 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 93 variables (including a purely random noise variable) was tested against 100 random samples of 50,000 background values (the variables tested are listed in Appendix C). The results were tabulated and the test statistics and p-values were compared to identify those variables that were most discriminant, as well as detect indications of how site location patterns were expressed within the variable pool. From the list of all variables, those with a K-S D statistic that is 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. Finally, variables that were very highly correlated were removed, resulting in the final selection of predictors, which averaged 19 per subarea.

The inclusion of the soils variables as factors required the models to consider many additional dummy variables. A described in Chapter 4, for each factor variable included in these models, a series of presence/absence variables, referred to as dummy variables, had to be created for each level of the factor. A variable of soil drainage requires the creation of a new dummy variable for each category (e.g., well-drained, moderately well-drained, poorly-drained, etc…). If the drainage variable contains seven different levels (categories) it will be represented within the model as seven separate dummy variables instead of just one. Because of this, if a model includes one of the three soil variables, the total number of soil drainage variables used within each model will include the dummy variables and therefore will be greater than the number of selected variables. As shown in Table 8, Table 9, and Table 10 an additional field is added to show the total number of variables after the inclusion of the dummy variables. The tables included in Appendix D show the variables that were selected to represent each subarea, the K-S D statistic, the MW U statistic, with associated p-values, and the statistics for the variable that represents random noise, for a basis of comparison. These tables provide information on the parameterization of each of the three statistical models for each subarea. While the final model selection for a subarea may be a proportionally weighted model, as opposed to one of the statistical models, the information is presented here.

Each of the variables tabulated in Table 8, Table 9, and Table 10 and detailed in the tables in Appendix D 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 LR, MARS, and RF statistical methods have means of dealing with correlated variables and variables that do not contribute to the success of the prediction. For LR, a backwards stepwise routine removes noncontributing variables based on their reduction of the Akaike Information Criterion (AIC) metric. For the MARS algorithm, the backwards elimination routine minimizes the effects of variables that do little to reduce the generalized cross-validation (GCV) metric. Additionally, the *nprune* parameter of the MARS algorithm controls the maximum number of terms within the model. This parameter is optimized to reduce misclassification through 10-fold Cross-Validation (CV). Finally, the RF algorithm reduces the effects of those variables that contribute little to the classification success through repeating predictions for each variable with random data. If the success of the model's classification is changed little by randomizing a given variable, then that variable likely contributes little to the overall success and its effect is minimized. Additionally, RF uses the *mtry* parameter to randomly select a set of variables to try at each node in a tree; the variable that leads to the most successful classification is retained. This serves to reduce the influence of ineffective variables and reduce the influence of variable correlation. Like the *nprune* parameter, *mtry* is also optimized through the use of 10-fold CV as was done and described in the Region 1, 2, and 3 models. These mechanisms are discussed in greater detail in the Task 3 report (Harris 2014) and for RF in Chapter 5 of the Task 4 report (Harris et al. 2014).

Table 8 - Optimized Number of Variables for Region 4/5 East Models

Subarea	Total Variables	Total w/ Dummy Variables	LR Selected Variables	LR AIC	nprune	mtry
riverine_section_1	20	32	25	791.16	36	9
riverine_section_2	17	26	20	1583.38	22	8
riverine_section_3	18	33	28	17788.16	20	9
riverine_section_4	19	37	34	284141.82	23	19
riverine_section_5	19	24	22	24162.52	32	13
riverine_section_6	19	24	23	73485.03	33	13
riverine_section_7	18	34	31	218122.08	24	18
upland_section_1	20	38	25	1469.84	38	11
upland_section_2	20	31	25	3074.24	16	9
upland_section_3	22	39	30	4595.55	8	11
upland_section_4	21	33	27	37532.92	23	9
upland_section_5	19	23	17	3459.27	29	7
upland_section_6	22	31	29	23274.17	17	9
upland_section_7	21	32	30	61911.48	12	9

Table 9 - Optimized Number of Variables for Region 4/5 West Models

Subarea	Total Variables	Total w/ Dummy Varibles	LR Selected Variables	LR AIC	nprune	mtry
riverine_section_1	18	23	21	18267.92	30	7
riverine_section_2	18	31	28	20120.64	35	16
riverine_section_3	17	26	22	38874.66	31	14
riverine_section_4	17	32	25	22563.27	21	17
riverine_section_5	21	34	33	68207.86	32	18
riverine_section_6	15	25	20	4957.95	29	7
upland_section_1	20	31	23	4189.14	22	9
upland_section_2	20	30	24	5745.42	27	9
upland_section_3	21	31	23	7901.64	35	16
upland_section_4	19	24	19	6501.91	15	7
upland_section_5	20	31	28	24506.50	24	9
upland_section_6	21	32	25	4560.16	26	9

Table 10 - Optimized Number of Variables for Region 6 Models

Subarea	Total Variables	Total w/ Dummy Variables	LR Selected Variables	LR AIC	nprune	mtry
riverine_section_1	17	29	24	26687.24	17	8
riverine_section_2	23	35	25	5735.21	41	26
riverine_section_3	20	25	22	53516.42	23	13
riverine_section_4	18	29	23	10483.48	32	15
riverine_section_5	19	29	19	2865.67	25	8
upland_section_1	20	32	23	5428.39	5	17
upland_section_2_RS	20	44	42	55657.31	45	12
upland_section_2_nonRS	18	18	16	1109.79	25	6
upland_section_3	23	35	27	4457.30	10	10
upland_section_4	19	31	26	1501.75	27	9
upland_section_5	23	35	7	16.00	42	2

MODEL 2 – PROPORTIONALLY WEIGHTED

The theory and development of the proportionally weighted Model 2 methodology are covered in some detail in the Task 3 report (Harris 2014). Briefly, this type of model is designed for areas with few recorded sites and is intended to limit the effects that a small and potentially unrepresentative sample may have on the statistical models.

Model 2 is created by first assessing which variables have the ability to discriminate site-present from background cells using the K-S and MW statistics, in the same way that it is done for the statistical models. While the statistical models use a set of around 20 variables selected from those that have discriminatory ability, Model 2 uses a smaller set of 5 variables selected to represent each of the broader classes of variable types (e.g., Euclidian or cost distance, hydrology types, and measures of slope or variation in topography). The known site locations are then compared to the selected variables to see where sites are located relative to each variable. For example, if distance to streams is one of the variables, the distance from a stream is broken down into numerous distance bands, and the portion of site-present cells within each band is calculated (e.g., 35% of site-present cells within 0–100 m of a stream, 25% within 100–200 m of a stream, and so on). The proportions of site-likely cells are rescaled to weights ranging from 1 to 20. These weights are then assigned to each distance band of the variable so that the bands with the highest proportion of sites now have the highest weight (e.g., 20); the distance band with the second highest proportion of site-likely cells receives the second highest weight (e.g., something less than 20 depending on the proportion), and so on until the bands with no site-likely cells are assigned the lowest weights. This is repeated for all five variables. The variables are then added together for a final model with weights ranging from 100 to 0, with a weight of 100 being a location that is at the intersection of the highest weight of 20 for all five variables. Following this, the model raster is divided by 100 to bring the weights into a scale of 0–1 to match the results of the other statistical models. This final model represents sensitivity based on the cumulative total of weights derived from the locations of known sites using variables that are demonstrated to distinguish site locations from the environmental background. Further, the intersection of weights helps to not only find those locations that are known to have sites, but can combine to indicate areas of similar yet varying landforms that may also contain archaeological sites.

Being that these models are not constructed from the statistical regression or classification approaches used for LR, MARS, or RF, there are no internal metrics to demonstrate the efficacy of a model—only the soundness and transparency of the method itself. The primary validation of Model 2 is through the Kg statistic once the model is applied to the full subarea. However, the establishment of model thresholds and the construction of confusion matrices is accomplished in essentially the same way and therefore will be discussed in the following chapter. Finally, although this modeling approach was applied to every subarea with fewer than 20 known sites, the LR, MARS, and RF models were also created for these subareas. The results of all four model types were compared to derive the final model selection. In some instances, the statistical models were selected over the proportional model if the outcome did not appear to be adversely effected by the small site sample. In these cases, it was common that while the sample was relatively small, the density of sites per area was relatively high.

MODEL 3 – SELECTED MODEL TEST SET AND CV ERROR RATES

The final LR, MARS, and RF models were fit on the complete dataset using the selected variables and *nprune* and *mtry* parameter values listed in the tables above. The models were run through 10-fold CV to derive error estimates and the 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 or the entire background sample if there were less than 500,000 cells. The final models were fit using the complete set of data and then calculated for the full population of raster cells within each subarea.

Table 11, Table 12, and Table 13 detail the error estimates and AUC values for each of the selected statistical model types for each subarea. The second column in these tables contains the Root Mean Square Error (RMSE) for the model prediction on a 25% hold-out sample of site-present cells that were not used in fitting the prediction. The third column contains the RMSE (LR model) or Accuracy (MARS and RF models) value for each model calculated as the average error/accuracy from each of the 10 CV out-of-fold 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. Similarly, Accuracy (for the MARS and RF models) measures the percentage of observations that were correctly classified as either site-present or site-absent. The fourth column is the Coefficient of Variation (CoV) for the error/accuracy expressed as a percentage. The MARS and RF models report Accuracy for the internal CV out-of-fold testing, as opposed to RMSE for the regression based LR model, because these models perform a classification that is measured by how often each observation is correctly classified. The column for AUC presents a single metric that describes the ability of the model to discriminate site-present from site-absent out-of-fold samples averaged across the 10 CV repetitions. This metric was described in detail in the Task 3 report (Harris 2014). Finally, the column for data samples contains the total number of site-present cells for the hold-out and training samples combined.

Table 11 - LR Model Prediction Errors from Test Set and 10-Fold CV

Subarea	Test RMSE	CV RMSE	CV RMSECoV	AUC	Data Samples
R4/5 East					
upland_section_1	0.248	0.242	7.240	0.968	1137
R6					
riverine_section_1	0.169	0.167	2.279	0.988	22831
riverine_section_3	0.314	0.309	0.635	0.925	24337
riverine_section_4	0.268	0.269	2.466	0.953	5744
upland_section_1	0.072	0.071	3.690	0.999	17293

Table 12 - MARS Model Prediction Errors and Accuracy from Test Set and 10-Fold CV

Subarea	Test RMSE	CV Accuracy	CV AccuracyCoV	AUC	Data Samples
R4/5 East					
riverine_section_3	0.216	0.939	0.486	0.976	10900
riverine_section_5	0.261	0.909	0.421	0.963	14185
riverine_section_6	0.253	0.911	0.334	0.962	30498
riverine_section_7	0.341	0.824	0.278	0.881	70629
upland_section_3	0.116	0.984	0.347	0.996	7274
upland_section_6	0.147	0.972	0.207	0.993	23645
R4/5 West					
riverine_section_2	0.325	0.861	0.684	0.902	7462
riverine_section_5	0.369	0.802	0.583	0.835	20100
riverine_section_6	0.246	0.913	0.784	0.958	2109
upland_section_1	0.177	0.963	0.471	0.988	2782
R6					
riverine_section_2	0.284	0.876	0.971	0.939	2749
riverine_section_5	0.182	0.965	0.533	0.988	3844
upland_section_2_RS	0.195	0.967	1.513	0.985	828
upland_section_2_nonRS	0.203	0.949	1.459	0.982	498

Table 13 - RF Model Prediction Errors and Accuracy from test set and 10-fold CV

Subarea	Test RMSE	CV Accuracy	CV AccuracyCoV	AUC	Data Samples
R4/5 East					
riverine_section_1	0.107	0.992	0.691	0.998	555
riverine_section_4	0.169	0.967	0.093	0.988	97780
upland_section_4	0.057	0.997	0.064	1.000	19377
upland_section_5	0.075	0.995	0.234	1.000	1780
upland_section_7	0.070	0.994	0.050	0.999	32071
R4/5 West					
riverine_section_1	0.120	0.986	0.216	0.921	6198
riverine_section_3	0.122	0.985	0.208	0.868	12581
riverine_section_4	0.113	0.985	0.236	0.901	8489
upland_section_2	0.089	0.994	0.220	0.961	2937
upland_section_3	0.077	0.995	0.133	0.979	4949
upland_section_4	0.070	0.996	0.204	0.957	2658
upland_section_5	0.080	0.994	0.106	0.966	14166
upland_section_6	0.081	0.994	0.281	0.955	2528

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; 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 hold-out test sample RMSE numbers for each of the selected 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 some degree. For example, upland subareas have a lower RMSE on average than do the riverine subareas (0.213 vs. 0.298 RMSE for all LR held-out samples; 0.180 vs. 0.266 for all MARS held-out samples; and 0.06 vs. 0.12 for all RF held-out samples).

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 than is considered sensitive to archaeological sites. Therefore, there are more cells that are observed to be background (a value of zero) than 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 and accuracy CoV show the percent change in the error/accuracy within the 10 out-of-fold samples for each CV repetition. The largest RMSE CoV value, which shows a larger magnitude of variation between the error/accuracy rates, is 7.2%. While this shows notable swings in the RMSE of the out-of-fold 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/Accuracy CoV on average than the riverine RMSE/Accuracy CoV sample mean (11.33 vs. 0.62 RMSE CoV for all LR out-of-fold samples; 0.63 vs. 0.58 Accuracy CoV for all MARS out-of-fold samples; however a reversal of 0.13 vs. 0.24 Accuracy CoV for all RF out-of-fold samples). While not a significant trend, the difference in CoV between riverine and upland areas is derived from the same biases of prevalence and area noted above.

The tables and discussion above show the steps for variable selection, parameterization, and error rates based on a 25% hold-out sample and 10-fold CV. The error rates resulting from the 10-fold CV, expressed as average RMSE, Accuracy, and the CoV of each show that the LR, MARS, and RF algorithms are variably successful in identifying the pattern of predictor variables that define the location of known sites within all selected 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 selected subareas. Based on these findings, all of the selected 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 4, 5, AND 6

In the previous chapter, the subarea models for LR, MARS, and RF were validated using a hold-out sample, 10-fold CV to produce prediction error estimates (RMSE) and percent accuracy, prediction error stability across hold-out samples (CoV), and a measure of a model's ability to discriminate site-present and background cells across the range of predicted probabilities (AUC). From these values, the LR, MARS, and RF models selected 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 in Chapter 5, along with RMSE and accuracy, give impressions of the models' overall ability to predict site-present cells. However, as elaborated in the beginning of this report, models that seek to define presence and absence are best evaluated at a given threshold. 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 14, Table 15, Table 16, and Table 17 compare each of the models at an arbitrary predicted probability threshold of p = 0.5. This threshold choice is essentially arbitrary, but choosing a threshold halfway between the extremes of the predicted probability distribution (p = 0 and p = 1) offers the most balanced point to compare results. The point of choosing this arbitrary threshold is to compare model results without the assumptions derived from implicitly selected thresholds as described in the section following this.

These tables present a series of metrics that allow the models to be directly compared with one another. As discussed in Chapter 4 of the Task 4 report, 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 these reasons, 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 downsamples 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 guard 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.

Table 14 - Comparing Kg and Kappa at a Threshold of 0.5, Selected Model 2 Subareas

Subarea	Back-ground %	Site-Present %	Kg @ 0.5	3:1 Balanced Mean Kappa	Upper 95%	Lower 95%
R4/5 East						
riverine section 2	14.47	78.37	0.82	0.60	0.63	0.57
upland section 2	4.57	88.60	0.95	0.83	0.84	0.83
R6						
upland section 3	1.86	41.60	0.96	0.44	0.47	0.42
upland section 4	2.09	44.77	0.95	0.51	0.54	0.48
upland section 5	4.49	92.59	0.95	0.86	0.92	0.79

Table 15 - Comparing Kg and Kappa at a Threshold of 0.5, Selected LR Models

Subarea	Back-ground %	Site-Present %	Kg @ 0.5	3:1 Balanced Mean Kappa	Upper 95%	Lower 95%
R4/5 East						
uplandsection1	12.53	96.13	0.87	0.74	0.77	0.72
R6						
riverine section 1	5.31	78.55	0.93	0.74	0.75	0.73
riverine section 3	20.31	92.13	0.78	0.62	0.62	0.61
riverine section 4	14.95	95.61	0.84	0.72	0.73	0.71
upland section 1	0.82	78.61	0.99	0.82	0.85	0.79

Table 16 - Comparing Kg and Kappa at a Threshold of 0.5, Selected MARS Models

Subarea	Back-ground %	Site-Present %	Kg @ 0.5	3:1 Balanced Mean Kappa	Upper 95%	Lower 95%
R4/5 East						
riverine section 3	7.23	83.58	0.91	0.75	0.76	0.74
riverine section 5	11.97	90.80	0.87	0.73	0.73	0.72
riverine section 6	11.86	81.63	0.85	0.66	0.67	0.65
riverine section 7	26.77	84.65	0.68	0.48	0.49	0.48
upland section 3	2.04	93.74	0.98	0.92	0.93	0.91
upland section 6	3.49	49.32	0.93	0.53	0.54	0.51
R4/5 West						
riverine section 2	20.54	83.85	0.76	0.56	0.57	0.55
riverine section 5	30.02	82.87	0.64	0.44	0.45	0.43
riverine section 6	12.48	92.49	0.87	0.74	0.75	0.72
upland section 1	5.13	87.56	0.94	0.82	0.84	0.80
R6						
riverine section 2	14.59	91.30	0.84	0.69	0.71	0.67
riverine section 5	5.29	95.21	0.94	0.87	0.88	0.86
upland section 2*	11.56	97.06	0.88	0.78	0.80	0.76

* combined rock shelter and non-rock shelter specific models

Table 17 - Comparing Kg and Kappa at a Threshold of 0.5, Selected RF Models

Subarea	Back-ground %	Site-Present %	Kg @ 0.5	3:1 Balanced Mean Kappa	Upper 95%	Lower 95%
R4/5 East						
riverine section 1	3.69	94.91	0.96	0.90	0.92	0.87
riverine section 4	5.01	84.05	0.94	0.80	0.80	0.79
upland section 4	0.75	70.74	0.99	0.77	0.78	0.76
upland section 5	1.34	87.19	0.98	0.89	0.90	0.87
upland section 7	1.34	89.89	0.99	0.90	0.91	0.90
R4/5 West						
riverine section 1	2.57	99.96	0.97	0.95	0.96	0.95
riverine section 3	2.99	99.96	0.97	0.95	0.95	0.94
riverine section 4	2.51	99.91	0.97	0.95	0.96	0.95
upland section 2	2.00	99.95	0.98	0.97	0.97	0.96
upland section 3	1.53	99.94	0.98	0.97	0.98	0.97
upland section 4	1.30	100.00	0.99	0.98	0.98	0.97
upland section 5	1.46	99.95	0.99	0.97	0.98	0.97
upland section 6	2.00	100.00	0.98	0.96	0.97	0.96

The above tables show that the models as applied to the full subarea study area are generally very good at identifying site-present locations relative to a random chance of finding a site. Between the models, the Kappa results show a relatively consistent trend within the different model types. As illustrated in Figure 26, across all model types the mean Kappa statistics range from a low of *k* = 0.44 to a high of *k* = 0.98; most with relatively narrow 95% confidence intervals. Unsurprisingly, the average Kappa for all models of a particular model type are lowest with Model 2 (*k* = 0.65) and highest with the RF models (*k* = 0.92), with LR (*k* = 0.73) and MARS (*k* = 0.69) in between. The most notable trend in Figure 26, is the majority of upland subareas scoring a higher Kappa (average *k* = 0.89) than the majority of riverine subareas (average *k* = 0.69). 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 successful at capturing the known site pattern within a small portion of the model.

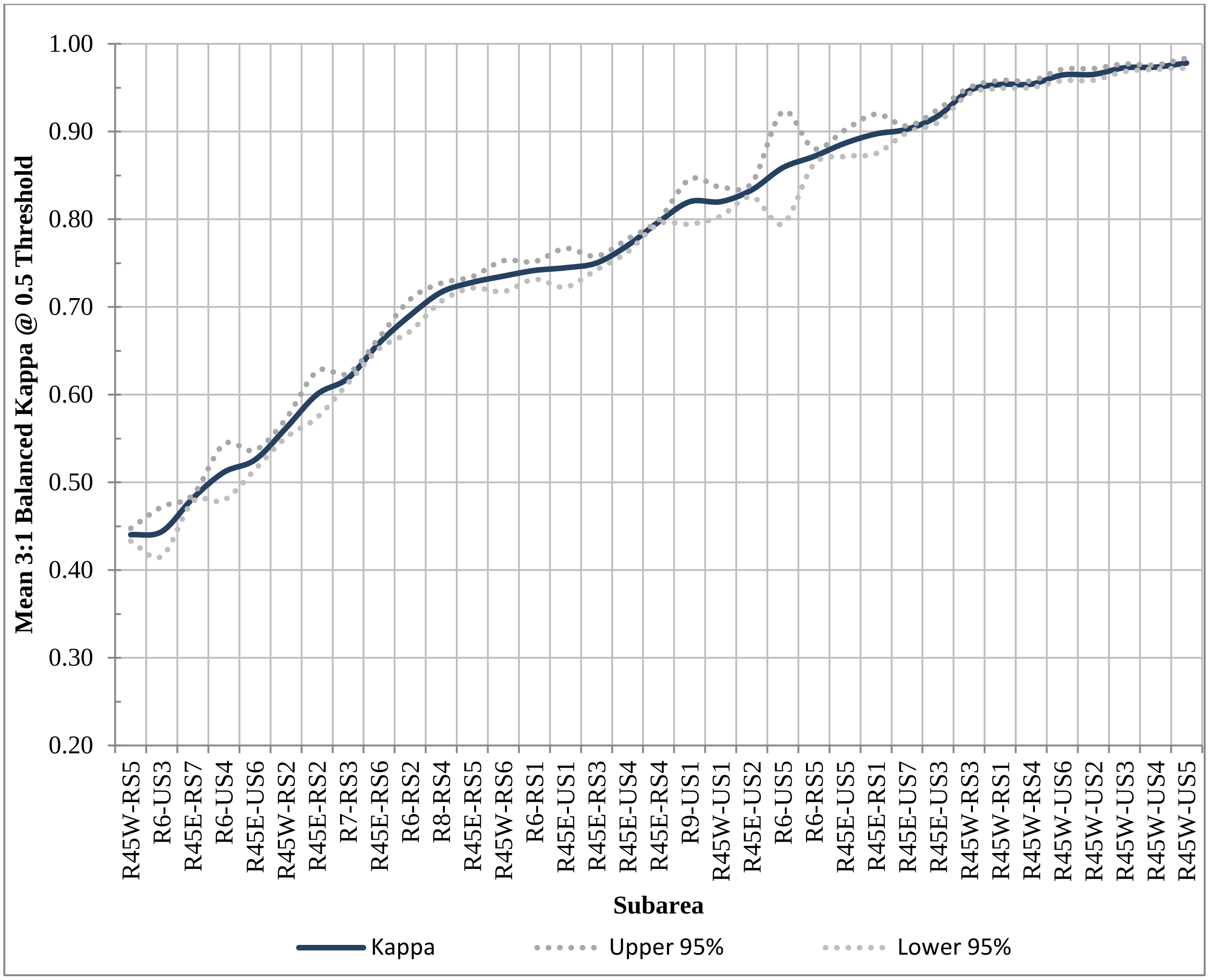


Figure 26 - 3:1 balance mean Kappa and 95-percent confidence intervals for all subarea models.

ESTABLISHING MODEL THRESHOLDS

As discussed in detail in the Task 4 report and repeated here for clarity, the discriminatory ability of the models created in this project 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 in 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 has 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.

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. 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 MARS model, RF model, and other innovations in statistical modeling, achieving very well fit—and at times overfit—models is not as great 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 the Task 4 report, 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. To illustrate the points above, the Task 4 report provided a series of different thresholds appropriate for different model objectives. Although only two thresholds were chosen to partition the final models, the full variety of thresholds is also presented here. This is for the purpose of comparison between the models of Task 4 and Task 5, but also to provide these thresholds in the event that these models are to be repartitioned for a different purpose.

On the other hand, the proportionally weighted models are much more akin to traditional models that sought to primarily maximize the correct site prediction while secondarily trying to limit the growth of the site-likely area. The use of discriminatory variables and proportional weighting definitely lift these models above the common judgmental APM, but not to the level of the statistical models. This is not a bad thing; it is, however, an inescapable reality of the method used in areas of low site counts. The proportional models suffer the same fate as the statistical models in being subject to the need for clearly defined and justified thresholds. For that reason, the proportionally weighted models were put through the same threshold creation routine as the statistical models and will be presented along with them for the remainder of the report. It may be helpful to repeat that the output sensitivity of the proportionally weighted models are on the same zero to 1 scale as the statistical models so the thresholds, Kg, and Kappa are also scaled appropriately.

Table 18, Table 19, Table 20, and Table 21present 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 F. 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 18 - Optimal Thresholds for Various Selection Methods; Selected Model 2 Subareas

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
R4/5 East								
riverine section 2	0.89	0.92	0.48	0.50	0.47	0.34	0.54	0.43
upland section 2	0.74	0.80	0.45	0.48	0.53	0.25	0.41	0.31
R6								
upland section 3	0.92	0.86	0.24	0.26	0.24	0.06	0.25	0.08
upland section 4	0.73	0.76	0.19	0.20	0.18	0.11	0.28	0.14
upland section 5	0.95	0.96	0.44	0.46	0.67	0.23	0.40	0.27

Table 19 - Optimal Thresholds for Various Selection Methods; Selected LR Models

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
R4/5 East								
upland section 1	0.96	0.98	0.59	0.60	0.68	0.10	0.56	0.27
R6								
riverine section 1	0.93	1.00	0.32	0.34	0.41	0.05	0.33	0.14
riverine section 3	0.83	1.00	0.57	0.60	0.57	0.30	0.70	0.50
riverine section 4	0.93	0.96	0.63	0.64	0.68	0.17	0.63	0.36
upland section 1	0.99	1.00	0.04	0.06	0.28	0.01	0.04	0.02

Table 20 - Optimal Thresholds for Various Selection Methods; Selected MARS Models

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
R4/5 East								
riverine section 3	0.92	0.96	0.21	0.24	0.38	0.11	0.33	0.16
riverine section 5	0.93	1.00	0.53	0.54	0.61	0.16	0.57	0.29
riverine section 6	0.87	1.00	0.43	0.44	0.43	0.19	0.56	0.33
riverine section 7	0.76	0.92	0.54	0.56	0.47	0.41	0.70	0.58
upland section 3	0.95	0.98	0.33	0.34	0.89	0.02	0.13	0.04
upland section 6	0.97	1.00	0.09	0.10	0.06	0.06	0.18	0.09
R4/5 West								
riverine section 2	0.91	1.00	0.51	0.52	0.46	0.37	0.64	0.49
riverine section 5	0.78	0.94	0.53	0.56	0.46	0.46	0.69	0.57
riverine section 6	0.88	0.98	0.58	0.60	0.72	0.17	0.57	0.33
upland section 1	0.99	1.00	0.29	0.30	0.62	0.08	0.28	0.14
R6								
riverine section 2	0.90	0.92	0.57	0.58	0.61	0.17	0.65	0.35
riverine section 5	0.95	1.00	0.52	0.54	0.88	0.09	0.31	0.15
upland section 2*	0.99	1.00	0.71	0.74	0.83	0.13	0.55	0.25

* combined rock shelter and non-rock shelter specific models

Table 21 - Optimal Thresholds for Various Selection Methods; Selected RF Models

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
R4/5 East								
riverine section 1	0.99	1.00	0.42	0.44	0.93	0.13	0.26	0.17
riverine section 4	0.81	1.00	0.28	0.30	0.40	0.11	0.36	0.21
upland section 4	0.97	1.00	0.10	0.12	0.12	0.01	0.14	0.08
upland section 5	0.99	1.00	0.23	0.24	0.81	0.07	0.20	0.12
upland section 7	0.95	1.00	0.23	0.24	0.80	0.01	0.14	0.08
R4/5 West								
riverine section 1	0.93	1.00	0.72	0.74	0.91	0.13	0.27	0.18
riverine section 3	0.93	1.00	0.71	0.74	0.92	0.13	0.29	0.20
riverine section 4	0.94	1.00	0.72	0.74	0.93	0.13	0.25	0.17
upland section 2	0.98	1.00	0.68	0.70	0.93	0.07	0.27	0.16
upland section 3	0.99	1.00	0.72	0.74	0.96	0.07	0.20	0.12
upland section 4	0.97	1.00	0.68	0.70	0.93	0.07	0.22	0.14
upland section 5	0.98	1.00	0.68	0.70	0.95	0.07	0.20	0.12
upland section 6	0.98	1.00	0.73	0.76	0.95	0.07	0.21	0.12

The full description and technical details of each of these thresholds is presented in the Task 4 report; a summary of each is provided here. 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 (maximizing the proportion of correctly classified sites while accounting for random agreement) and maximizing Kg (maximizing the proportion of correctly classified sites while accounting for the area of the classification). 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 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. 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; 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. 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. 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.

Throughout Regions 4, 5, and 6, the overall prevalence of known archaeological sites with a prehistoric component is 0.0011. Riverine subareas have an overall prevalence of 0.0083 and upland subareas have an overall prevalence of 0.0003. Figure 27 shows the prevalence of all subareas within Regions 4, 5, and 6. The lowest prevalence is within Region 6 Upland Section 5 at 0.00001 and the highest is within Region 4/5 East Riverine Section 7 at 0.0302. 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.

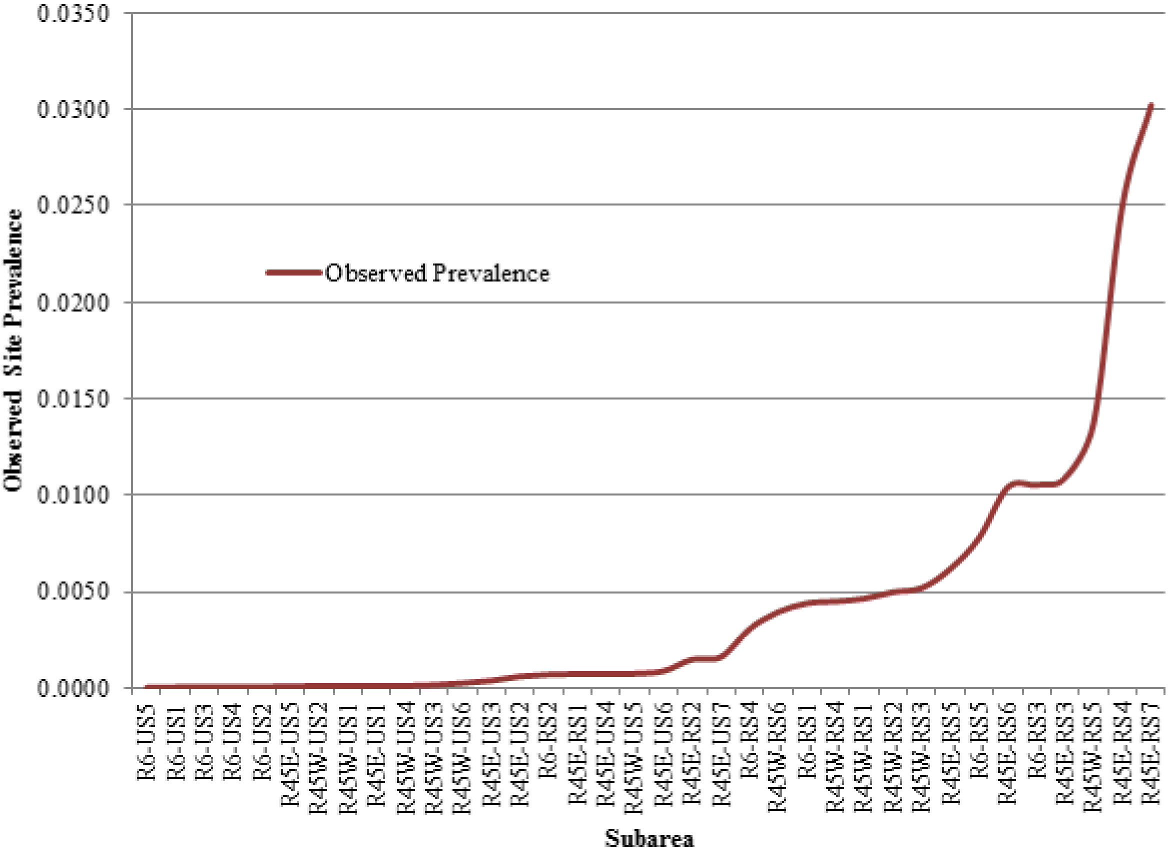


Figure 27 - Average prevalence of prehistoric sites by subarea.

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 thresholds 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 concluded 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 22, Table 23, Table 24, and Table 25, the combined site-likely area of high and moderate probability includes from 85% to 100% of the known site-present cells in a site-likely area from 25% to 37% of the study area, for Kg statistics ranging from 0.603 to 0.740: an average Kg of 0.678. The boxplots in Figure 28 show the variation on Kg statistics for the 36 selected models across the four model types. As anticipated, the mean Kg increases and the variation in Kg decreases as the models become more powerful. 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 G. The overall confusion matrix representing the site-likely classification for the entirety of Regions 4, 5, and 6 is presented in Table 26. Figure 29 depicts an overview of high, moderate, and low sensitivity for the entirety of Regions 4, 5, and 6. These data will be provided as ESRI raster grids for detailed viewing and analysis.

Table 22 - Kg and Cell Percentages at Suggested Final Thresholds, Selected Model 2 Subareas

	Pred=Obs @ 0.1, High Sensitivity				Specificity @ 0.67, Moderate Sensitivity			
Subarea	Threshold	% Background	% Sites	Kg	Threshold	% Background	% Sites	Kg
R4/5 East								
riverine section 2	0.54	11%	75%	0.86	0.34	36%	95%	0.621
upland section 2	0.41	11%	96%	0.89	0.25	35%	99%	0.644
R6								
upland section 3	0.25	14%	85%	0.84	0.06	37%	92%	0.603
upland section 4	0.28	11%	62%	0.82	0.11	35%	98%	0.637
upland section 5	0.40	11%	100%	0.89	0.23	33%	100%	0.666

Table 23 - Kg and Cell Percentages at Suggested Final Thresholds, Selected LR Models

	Pred=Obs @ 0.1, High Sensitivity				Specificity @ 0.67, Moderate Sensitivity			
Subarea	Threshold	% Background	% Sites	Kg	Threshold	% Background	% Sites	Kg
R4/5 East								
upland section 1	0.56	10%	92%	0.892	0.1	33%	97%	0.665
R6								
riverine section 1	0.33	10%	90%	0.893	0.05	31%	97%	0.683
riverine section 3	0.70	9%	69%	0.865	0.30	33%	98%	0.666
riverine section 4	0.63	10%	90%	0.893	0.17	33%	99%	0.668
upland section 1	0.04	10%	91%	0.895	0.01	25%	95%	0.741

Table 24 - Kg and Cell Percentages at Suggested Final Thresholds, Selected MARS Models

	Pred=Obs @ 0.1, High Sensitivity				Specificity @ 0.67, Moderate Sensitivity			
Subarea	Threshold	% Background	% Sites	Kg	Threshold	% Background	% Sites	Kg
R4/5 East								
riverine section 3	0.33	9%	86%	0.893	0.11	31%	97%	0.682
riverine section 5	0.57	10%	87%	0.890	0.16	33%	99%	0.668
riverine section 6	0.56	9%	75%	0.877	0.19	32%	97%	0.670
riverine section 7	0.7	9%	45%	0.811	0.41	32%	90%	0.641
upland section 3	0.13	10%	98%	0.899	0.02	26%	100%	0.740
upland section 6	0.18	10%	61%	0.840	0.06	32%	86%	0.627
R4/5 West								
riverine section 2	0.64	10%	67%	0.859	0.37	32%	93%	0.653
riverine section 5	0.69	9%	50%	0.818	0.46	33%	85%	0.615
riverine section 6	0.57	10%	91%	0.892	0.17	33%	97%	0.659
upland section 1	0.28	10%	90%	0.889	0.08	30%	99%	0.691
R6								
riverine section 2	0.65	10%	83%	0.878	0.17	33%	99%	0.669
riverine section 5	0.28	9%	97%	0.905	0.31	30%	99%	0.694
upland section 2*	0.55	10%	97%	0.896	0.13	31%	99%	0.684

* combined rock shelter and non-rock shelter specific models

Table 25 - Kg and Cell Percentages at Suggested Final Thresholds, Selected RF Models

	Pred=Obs @ 0.1, High Sensitivity				Specificity @ 0.67, Moderate Sensitivity			
Subarea	Threshold	% Background	% Sites	Kg	Threshold	% Background	% Sites	Kg
R4/5 East								
riverine section 1	0.26	10%	100%	0.897	0.13	29%	100%	0.715
riverine section 4	0.36	8%	86%	0.907	0.11	33%	96%	0.658
upland section 4	0.14	10%	83%	0.885	0.01	30%	94%	0.678
upland section 5	0.2	10%	100%	0.902	0.07	30%	100%	0.704
upland section 7	0.14	10%	97%	0.893	0.01	32%	98%	0.677
R4/5 West								
riverine section 1	0.27	9%	100%	0.908	0.13	30%	100%	0.697
riverine section 3	0.29	9%	100%	0.908	0.13	32%	100%	0.676
riverine section 4	0.25	10%	100%	0.901	0.13	30%	100%	0.699
upland section 2	0.27	10%	100%	0.901	0.07	33%	100%	0.671
upland section 3	0.2	11%	100%	0.895	0.07	28%	100%	0.719
upland section 4	0.22	10%	100%	0.899	0.07	33%	100%	0.672
upland section 5	0.2	10%	100%	0.903	0.07	29%	100%	0.714
upland section 6	0.21	10%	100%	0.904	0.07	29%	100%	0.707

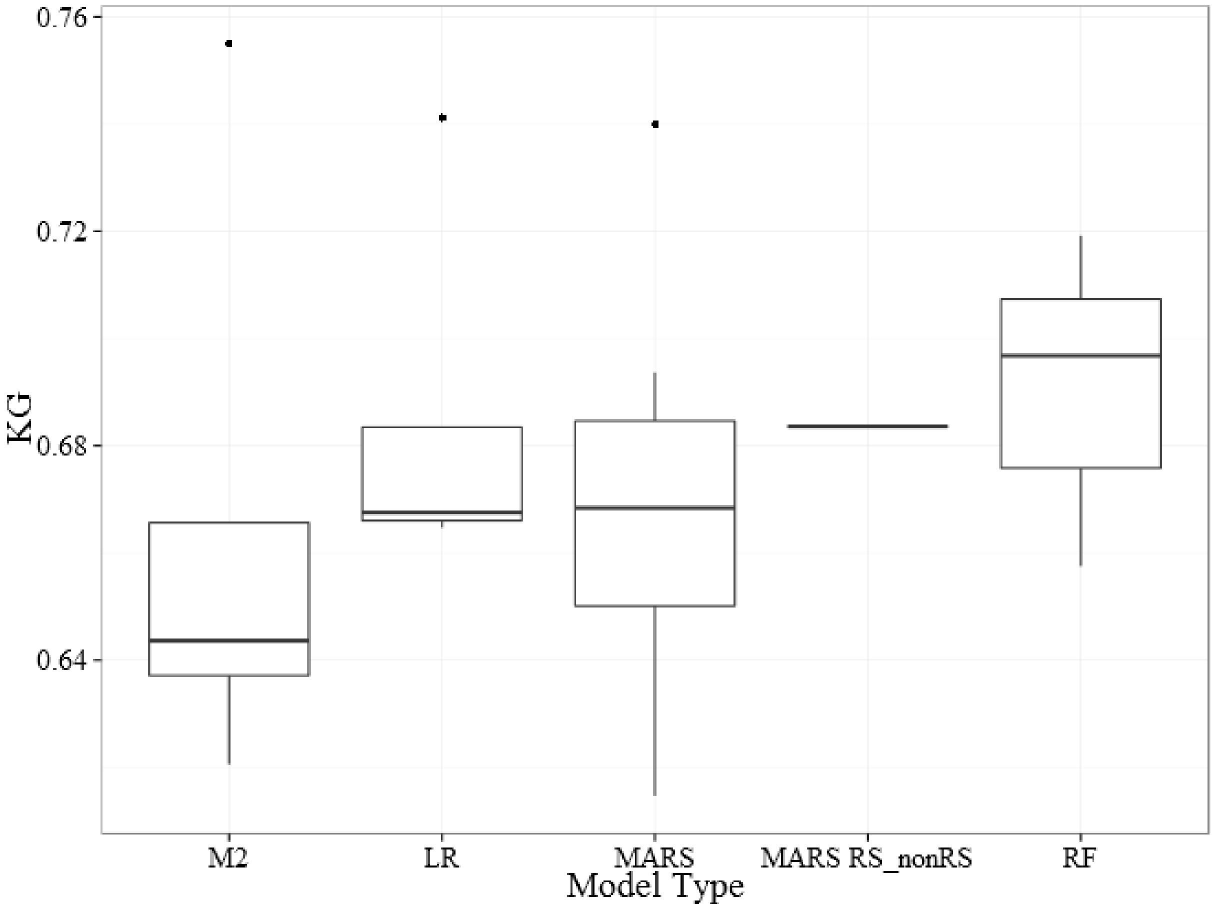


Figure 28 - Distribution of Kg statistics for each of the four model types.

Table 26 - Confusion Matrix for Site-Likely Area of Complete Regions 4, 5, and 6 Selected Models

		Known Sites		
		Present	Absent	
Model Prediction	Present	343773	97632926	97976699
	Absent	17149	228997869	229015018
		360922	326630795	326991717

Sensitivity / TPR =	0.952
Specificity / TNR =	0.701
Prevalence =	0.0011
Kvamme Gain (Kg) =	0.685
Accuracy =	0.701
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) =	3.179
Negative Prediction Gain (NPG) =	0.068
False Negative Rate (FNR) =	0.048
Detection Prevalence =	0.300

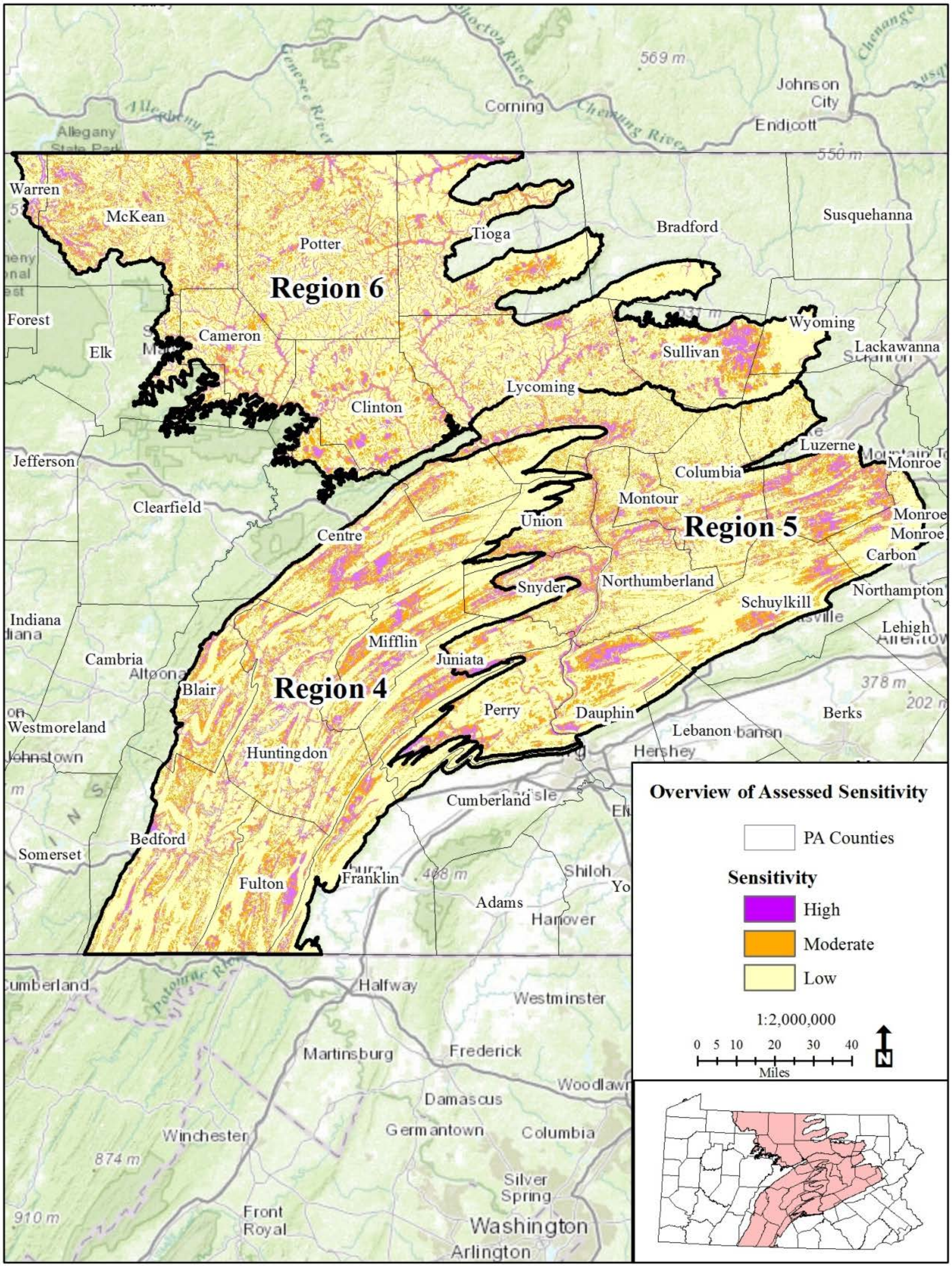


Figure 29 - Overview of assessed prehistoric sensitivity for Regions 4, 5, and 6.

7
CONCLUSIONS AND RECOMMENDATIONS

Over the course of modeling archaeological sensitivity in Regions 4, 5, and 6, 131 individual models were created for the 36 subareas. These included LR, MARS, RF, and proportionally weighted (Model 2) models for non-rock shelter sites and, in some subareas, for rock shelter sites as well. The total area covered by these models is 13,871 square miles, constituting much of central Pennsylvania. The methodology used to create these models involved the preparation of PASS site data, the development of 93 individual environmental variables, and the division of the regions into 36 separate subareas. Through the testing of each of the variables against the environmental background of each subarea, the parameterization and validation of statistical models, creation of additional models where there are few known sites or high proportions of rock shelters, and the final model selection based on error estimate results, Kg, and other metrics, a total of 36 models was selected from the candidates. The establishment of numerous potential thresholds based on variable criteria, and, finally, the application of selected thresholds and mosaicking of 36 separate subarea models into the final model for each of the regions completed the task. The end result is a model of all three regions that correctly classifies 95.2% of known site-present cells within 29.9% of the study area, for a Kg of 0.685. In actuality, the model is capable of correctly predicting the location of all archaeological sites and minimizing the site-likely area to a much smaller percent of the study area, but the selection of a low end threshold for the site-likely area was intentionally set to approximately 33% of the study area. Compared to a random survey, the chances of finding a site in the combined high and moderate sensitivity area are 3.179 times greater.

The 36 subarea models created for Regions 4, 5, and 6 are derived from a variety of model types, including proportionally weighted (Model 2) and LR, MARS, and RF statistical models. Each of these models has their own strengths, weaknesses, and assumptions, as well as ability to address the bias-variance tradeoff that is amplified when using correlated environmental variables and often sparse site location data. However, each model type has been shown to be effective at identifying the patterns within known site locations and extrapolating that pattern to landforms that share similar characteristics. Further, each type of model has different abilities in addressing variations in data quality and sample size issues. With the exception of the proportionally weighted models, each of the statistical models is capable of providing internal metrics that offer information on the model's prediction errors and qualities of fit.

The results of the internal prediction error rate tests on the 10-fold CV samples (average RMSE = 0.212 for the LR models and an accuracy of 95.2% for MARS and RF models) and an average RMSE of 0.176 for all models on the held-out sample 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 can 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 representation 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 analyses.

Three of the four recommendations from the Task 4 report were implemented within the modeling of Regions 4, 5, and 6: 1) the inclusion of relevant aggregated soils data as predictive variables; 2) continuing to refine and integrate the modeling of rock shelter site-types; and 3) increasing model-building efficiency. The methodology for the inclusion of soils data is discussed in Chapter 4 of this report. The need and methods for integrating rock shelter specific models in subareas where they account for 30% or more of the site total was discussed in the Task 4 report and implemented in Regions 1, 2, and 3. For this report, the integration of the rock shelter/non-rock shelter modeling methods was streamlined through modifications to the code that runs the models and to the sequence by which the code is executed. Finally, the third recommendation was achieved by continued streamlining of the modeling code, designing functions to utilize multiple processing cores, using memory more efficiently, unifying separate routines into a single unified code base, and, most importantly, implementing the very time consuming model fitting routine onto an internet based “cloud” server. This final step allows for very effective scaling of computer resources, access to very powerful servers, and the ability to have many servers working simultaneously. Further, through this method, the modeling process can be monitored, interacted with, and reported on at any hour of the day. This method is not yet practical for the implementation of the model prediction and raster creation routine due to the extremely large file size of the raster prediction layers. However, the implementation of this method into the “cloud,” likely the first archaeological predictive model to utilize such technology, has led to much greater efficiency in the modeling process.

The single recommendation from Task 4 that was not fully implemented in this report is the use of class weights and thresholds within the RF model to attempt to reduce model variance and increase generalizability of the results. A number of experiments were undertaken, but no conclusive method by which to employ this strategy was found. It was decided to continue with the methods used in Task 4 for the sake of the schedule and consistency. However, this recommendation will remain and further experiments will be conducted.

The methods for creating models are streamlined and well-implemented at this point in the overall project. There are a number of minor variations and cleaning up of the code that can be undertaken to save time and provide better feedback, but no major advancement or inclusions are suggested at this time. The following recommendations include the restated recommendation from the Task 4 report as well as a few minor recommendations that are designed to lead up to the final report in this project, which will outline the outcome of the Pennsylvania Model program and the next steps forward.

1. 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 MARS 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.
2. Create proportionally weighted models for all subareas, not only those with few sites.
    - To date, all subareas have been modeled with the three statistical models, but only those with 20 sites or fewer were modeled with the proportionally weighted model. The creation of this model for all subareas will lead to complete state-wide coverage of this model in the interest of consistency and future use. While not as powerful as the statistical models, this method has proven to be valid and should be part of the final deliverable.
3. Testing additional model types for comparison with existing model results.
    - The landscape of statistical models is vast and there are many powerful methods available today. Through this project, a framework was constructed to not only handle and process large amounts of data from site locations to prediction rasters, but also to include interchangeable parts such as model types and analysis. With the existing framework, additional modeling methods can be dropped in and tested on existing data. This process will help to inform the future direction of this project.

8
REFERENCES CITED

Akaike, Hirotugu

1974 A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* 19(6):716–723.

Bibler, David, and Patricia Miller

2002 A Report on the First Pennsylvanians: A Stratified Paleo-Indian Site in Liverpool, Perry County. Paper presented at the Third Annual Byways to the Past Conference, 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.

Breiman, Leo

2001 Random Forests. Machine Learning 45(1):5–32.

Carr, Kurt W.

1998a The Early Archaic Period in Pennsylvania. *Pennsylvania Archaeologist* 68(2):42-69.

1998b 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.

Cohen, Jacob

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

Conover, W, J.

1999 *Practical Nonparametric Statistics*. 3rd ed. Wiley, New York, NY.

Coppock, Gary F.

2009 The Martin Site: A Bifurcate Phase Hunting Camp in Central Pennsylvania. *Journal of Middle Atlantic Archaeology* 25:31–58.

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.

Custer, Jay F.

1989 P*rehistoric Cultures of the Delmarva Peninsula: An Archaeological Study*. Associated University Presses, Inc., Cranbury, New Jersey.

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

Custer, Jay F., John A. Cavallo, and R. Michael Stewart

1983 Lithic Procurement and Paleo-Indian Settlement Patterns on the Middle Atlantic Coastal Plain. *North American Archaeologist* 4(4):263-275.

Custer, Jay F., Scott Watson, and Daniel Bailey

1994 Data Recovery Investigations at the West Water Street Site 36CN175, Lock Haven, Clinton County, Pennsylvania. E.R. # 1981-0405-035. Prepared for Kise, Frank, and Straw Historic Preservation Group for the United States Army Corps of Engineers, Baltimore District.

Dragoo, Don W.

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

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 Brian F. Schilling

1999 U.S. Route 15 Improvement Project, Tioga County, Pennsylvania. S.R. 6015, Sections G20 and G22, Steuben County, New York. E.R. 1997-2018-117-H. Prepared for the Pennsylvania Department of Transportation and New York State Department of Transportation, Harrisburg, Pennsylvania. Skelly and Loy Inc., Monroeville, Pennsylvania.

East, Thomas C., Frank J. Vento, Christopher T. Espenshade, Margaret G. Sams, and Brian C. Henderson

2002 Phase I/II/III Archaeological Investigations, Northumberland and Union Counties, S.R. 0080, Section 52D, Bridge Expansion and Highway Improvement Project. E.R. # 1999-8000-042. Report prepared by Skelly and Loy, Inc., Monroeville, Pennsylvania and submitted to the Pennsylvania Department of Transportation, Engineering District 3-0, Montoursville, Pennsylvania.

Efron, B., and R. Tibshirani

1997 Improvements on Cross-Validation: The .632 + Bootstrap Method. *Journal of the American Statistical Association* 92(438):548–560.

Fawcett, Tom

2004 ROC Graphs: Notes and Practical Considerations for Researchers. *Pattern Recognition Letters* 27(8):882–891.

2006 An Introduction to ROC Analysis. *Pattern Recognition Letters* 27(2006):861–874.

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.

Friedman, J. H.

1991 Multivariate Adaptive Regression Splines. *The Annals of Statistics* 19:1.

Funk, Robert E.

1993 *Archaeological Investigations in the Upper Susquehanna Valley, New York State*, Volume 1. Persimmon Press Monographs in Archaeology, Buffalo, New York.

Gardner, William M.

1989 An Examination of Cultural Change in the Late Pleistocene and Early Holocene (circa 9200-6800 BC). In *Paleoindian Research in Virginia: A Synthesis*, edited by J.M. Wittkofski and T.R. Reinhart, pp. 5–51. The Deitz Press, Richmond, Virginia.

Geier, Clarence

1990 The Early and Middle Archaic Periods: Material Culture and Technology. In *Early and Middle Archaic Research in Virginia: A Synthesis*, edited by Theodore R. Reinhart and Mary Ellen N. Hodges, pp. 81–98. The Dietz Press, Richmond, Virginia.

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.

Harris, Matthew D., Susan Landis, and Andrew R. Sewell

2014 Pennsylvania Department of Transportation Archaeological Predictive Model Set, Task 4: Study Regions 1, 2, and 3. Prepared for Pennsylvania Department of Transportation, Bureau of Planning and Research, Harrisburg. URS Corporation, Burlington, New Jersey.

Hart, John P. (editor)

1995 Archaeological Investigations at the Memorial Park Site (36CN164), Clinton County, Pennsylvania. E.R. # 1981-0405-035. Report prepared GAI Consultants, Inc., Pittsburgh for the Baltimore District of the U.S. Army Corps of Engineers.

Jacoby, Robert, Robert Wall, Rhea Rodgers, and John Killeen

1998 Phase III Archaeological Investigation of Site 36Co17 and 36Co18 for the Mifflinville Bridge Replacement. E.R. # 1988-0959-037. Report prepared by the Cultural Resource Group, Louis Berger & Associates, Inc., East Orange, New Jersey for the Pennsylvania Department of Transportation, Engineering District 3-0, Montoursville, Pennsylvania.

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.

Justice, Noel D.

1987 *Stone Age Spear and Arrow Points of the Midcontinental and Eastern United States: A Modern Survey and Reference*. Bloomington, Indiana: Indiana University Press.

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

Kingsley, Robert G., Joseph Schuldenrein, James A. Robertson, and Daniel R. Hayes

1991 The Archeology of the Lower Black's Eddy Site, Bucks County, Pennsylvania: Final Report. Report submitted to the Bucks County Commissioners. John Milner Associates, Inc., West Chester, PA.

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.

Lehmann, Erich L

1986 *Testing Statistical Hypothesis*. 2nd ed. Wiley, New York, NY.

Liaw, Andy, and Matthew Wiener

2002 Classification and Regression by randomForest. *R News* 2(3):18–22.

MacDonald, Douglas H.

2003 Pennsylvania Archaeological Data Synthesis: The Upper Juniata River Sub-Basin (Watersheds A-D). E.R. # 00-2888-013. Walter Industrial Park: Mitigation of Adverse Effects, U.S. Department of Commerce, Economic Development Administration, Greenfield Township, Blair County, Pennsylvania. Reported by GAI Consultants, Inc., Monroeville, Pennsylvania, to Keller Engineers, Inc., Hollidaysburg, Pennsylvania.

Means, Bernard K.

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

Michels, Joseph W., and Ira F. Smith (editors)

1967 *A Preliminary Report of Archaeological Investigations of the Sheep Rock Shelter, Huntingdon County, Pennsylvania*, Volume 2. The Pennsylvania State University, Department of Anthropology, University Park, Pennsylvania.

Milanich, Jerald T.

1994 Archaeology of Precolumbian Florida. University Press of Florida, Gainesville.

Milborrow, Stephen

2014 Notes on the Earth Package. Electronic document: http://cran.r-project.org/web/packages/earth/vignettes/earth-notes.pdf.

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.

Pampel, F. C. (editor)

2000 *Logistic Regression: A Primer.* Vol. 132. Sage, Thousand Oaks, CA.

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.,

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.

2010 Chert Use and Local Settlement in Central Pennsylvania: Investigations at 36Ce523. *Journal of Middle Atlantic Archaeology* 26:115–140.

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.

Ritchie, William A.

1994 The Archaeology of New York State. Revised Edition. Purple Mountain Press, Fleischmanns, New York.

Salkind, Neil J. (editor)

2007 *Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks, CA.

Sassaman, Kenneth E.

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

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.

Stewart, R. Michael, and John Cavallo

1991 Delaware Valley Middle Archaic. *Journal of Middle Atlantic Archaeology* 7: 19-42.

Viera, Anthony J., and Joanne M. Garrett

2005 Understanding Interobserver Agreement: The Kappa Statistic. Family Medicine 37(5):360-363.

Wall, Robert D.

1994 Phase III Archaeological Investigations (36Un82), Allenwood Federal Correctional Complex Sanitary Water Treatment Facility, Union County, Pennsylvania. E.R. # 1989-1630-119. Report prepared by Louis Berger and Associates, Washington, D.C. for the United States Department of Justice, Federal Bureau of Prisons, Washington, D.C.

2000 A Buried Lamoka Occupation in Stratified Contexts West Branch Valley of the Susquehanna River. *Pennsylvania. Pennsylvania Archaeologist* 70 (1):1-44).

Weed, Carol, and William Wenstrom

1993 Cultural Resources Investigations of 36Lu90 (Jacobs Site) and 36Lu105 (Gould Island Site), Luzerne County, Pennsylvania. E.R. # 1982-0648-042. Report prepared by Emanco, Inc. for the Transcontinental Gas Pipeline Corporation, Houston, Texas.

Wheatley, David, and Mark Gillings

2002 *Spatial Technology and Archaeology. The Archaeological Application of GIS*. Taylor & Francis, London, UK.

Wyatt, Andrew

2003 Early and Middle Woodland Settlement Data for the Susquehanna Basin. In *Foragers and Farmers of Early and Middle Woodland Periods in Pennsylvania*, edited by Paul A. Raber and Verna L. Cowin, pp. 35–48. Pennsylvania Historical and Museum Commission, Recent Research in Pennsylvania Archaeology No. 3. Harrisburg.

Wyatt, Andrew, Robert H. Eiswert, Richard C. Petyk, and Richard T. Baublitz

2005 Phase III Archaeological Investigations at the Raker I Site (36Nb58), Route 147 Climbing Lane Project, Upper Augusta Township, Northumberland County, Pennsylvania. E.R. # 2000-6173-097. Prepared by McCormick Taylor, Inc., Harrisburg for the Pennsylvania Department of Transportation, Engineering District 3-0, Montoursville.

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

AIC	Akaike Information Criterion
APM	Archaeological Predictive Modeling
AUC	Area Under Curve
CoV	Coefficient of Variation
CRGIS	Cultural Resources Geographic Information System
CV	Cross-Validation
GCV	Generalized 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
PASS	Pennsylvania Archaeological Site Survey
PPG	Positive Predictive Gain
PPV	Positive Prediction Value
RF	Random Forests/randomForest
RMSE	Root Mean Square Error
TNR	True-Negative Rate
TPR	True-Positive Rate
UDR	Unexpected Discovery Rate

TERMS

page in report text
(first used)

Accuracy (in error estimates for MARS and RF models) 59

The measurement of accuracy is used in many classification methods. This measure is simply the percent of observations (site-present or site-absent) that are correctly classified by the algorithm. As used in this report, the accuracy is the percentage of observations from the out-of-bag sample that were correctly classified by the model. This is an internal metric that assess the model's ability to correctly predict data that were not used in the fitting of the model.

Adaptive Regression Splines (see Multivariate Adaptive Regression Splines) 1

Akaike Information Criterion (AIC)..56

A measure of relative model quality that balances goodness of fit and model complexity. This measure is used in model selection to choose the model that has the best fit relative to complexity for a given data set. Within a series of nested candidate models, the one with the lowest AIC will likely represent the model with the best goodness of fit without being over-fit or over-parameterized (see Akaike 1974).

Archaeological Predictive Modeling (APM)..1

The field of study concerning the use of existing archaeological data or theory to predict the sensitivity of locations for the presence of archaeological material.

Area Under Curve (AUC) (see also Receiver Operating Characteristics) 53

Also referred to as Area Under Receiver Operating Characteristics Curve (AUROC), AUC is a measure of the balance between a model's Sensitivity and Specificity across the full range of cut-off points. The AUC is a single measure that captures a model's ability to balance True Positive Rate and False Positive Rate across the full range of the model's output. The higher the AUC, the higher the Sensitivity and Specificity across the full range of the model, and the more likely the model is to correctly classify a randomly chosen positive instance. AUC is used in model selection to assess a model's ability to correctly classify observations (see Fawcett 2006).

Confusion Matrix 73

A classification table in the form of a 2-cell × 2-cell contingency table that shows how many sites were correctly predicted as sites and how much of the non-site area was correctly predicted as such. This method is frequently used as a means to assess the ability of a model to classify observations (see Fawcett 2006).

Coefficient of Variation (CoV) 59

The CoV is a statistic that measures the normalized dispersion within a frequency distribution. The acronym CoV is used in this study to avoid confusion with the acronym 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)...49

Cross-Validation (CV) (see Generalized Cross Validation and K-folds Cross-Validation)..........56

Cultural Resources Geographic Information System (CRGIS) 37

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.)

Factor & Factor Level 51

A factor is the data type used by the R statistical language to code data that are categorical (nominal), as opposed to quantitative data such as continuous integers. The factor data type is composed of the qualitative categories represented as levels (e.g., "high," "moderate," "low") and a string of integers to represent the categories (e.g., 1, 2, 3). The categorical data are actually stored as a string of representative integers, but referenced back to the levels so that the data can be converted to its original category when needed. Among other reasons, this allows the program to work very efficiently with integers as opposed to storing and computing a long list of category labels.

False Negative Rate (FNR)..76

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).

Generalized Cross-Validation (GCV)..56

GCV is a statistical method that estimates performance or prediction error from within a model based on weight assigned to model complexity. GCV approximates the measure of performance that would be derived through leave-one-out Cross-Validation. In this project, the GCV relates to the internal performance measure derived from the Multivariate Adaptive Regression Splines model (see Milborrow 2014).

Geographic Information Systems (GIS) 3

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

K-folds Cross-Validation 56

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 49

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 55

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)..1

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 55

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).

Model Formula 51

As used in this project, the model formula is a symbolic representation of the *a priori* relationship between the model predictors (*x1*, *x2*, *x3*,… *xn*) and the outcome (*y*). Typically, the tilde symbol (~) is used to specify that the response is a function of one or more predictors. For example, the formula (*y* ~ *x1*) specifies to the statistical model that *y* as the response variable is a function of the linear predictor *x1*. Further, the formula symbols specify the relationship between the predictor variables. For example, the formula (*y* ~ *x1* + *x2* + *x3*) specifies that *y* is an additive function of the linear predictors *x1*, *x2*, and *x3*. Additional symbols can be used in the formula to represent interactions between predictors, non-additive relationship, and polynomials. However, this project uses only linear and additive formulae.

mtry 56

This is the name of a key parameter in the RF model. One of the key features of RF is the random selection of a subset of the predictor variables to test at each node in the tree building process. The number of randomly selected variables to try is called "*mtry*.. By default, *mtry* is set to $\sqrt{p}$ for classification problems and $p/3$ in regression problems. In this project, *mtry* is optimized through cross-validation to the lowest error rate of the out-of-fold sample.

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) 76

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) 76

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).

nprune 56

This is the name of a key parameter in the MARS model. This algorithm includes a backwards pass that prunes the model down to reduce variance and eliminate unneeded model terms. The *nprune* parameter is used to set the *maximum* number of terms that are allowed to remain in the model; the fewer terms, the more simple the model. Through this parameter, models can be trimmed for the purpose of model size, complexity, or generality of the fit. By default, *nprune* is set to NULL so that the model is unrestrained in the number of terms. For this project, the *nprune* parameter is set through cross-validation to the lowest error rate of the out-of-fold sample.

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)..76

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)..76

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).

Prevalence

Prevalence is the proportion of a population found to have a particular condition. In this case, the population is the total number of ~10 × 10-m raster cells that make up each subarea and the condition is that a cell be within a known archaeological site. Determining prevalence is important in these models because the low number of cells within known archaeological sites is very small compared to the overall area being predicted, leading to highly imbalanced data in terms of site-presence versus site-absence. 61

Random Forests ....n/a

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).

Root Mean Square Error (RMSE)...53

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) 68

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)..68

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) 76

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) 76

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)...76

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

SITE TYPES AND LANDFORMS RECORDED IN THE PASS DATABASE, BY TIME PERIOD

Chart 1 - Region 4 Site Types by Landform, Paleoindian Period

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	0	0	0	0	0	0	0	0	2
Part of Multi-Component Site	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	1	10
Total	0	9	0	0	1	1	0	0	0	0	0	0	0	0	0	1	12

Chart 2 - Region 4 Site Types by Landform, Early Archaic Period

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	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Open Prehistoric Site, Unknown Function	0	1	0	0	1	1	0	0	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	0	17	1	0	8	4	3	3	1	0	0	0	1	0	0	1	39
Total	0	18	1	0	10	7	3	3	1	0	0	0	1	0	0	1	45

Chart 3 - Region 4 Site Types by Landform, Middle Archaic Period

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	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Open Habitation, Prehistoric	0	2	0	0	1	0	1	0	0	0	0	0	0	2	0	0	6
Open Prehistoric Site, Unknown Function	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Unknown Function Open Site Greater than 20 m Radius	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
(blank)	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Part of Multi-Component Site	0	22	2	0	14	16	2	3	1	0	0	0	2	1	0	4	67
Total	0	25	3	0	15	17	3	4	1	0	0	0	2	3	0	4	77

Chart 4 - Region 4 Site Types by Landform, Late Archaic Period

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	0	1	0	0	0	0	0	0	0	0	0	1
Open Habitation, Prehistoric	0	28	2	0	20	6	6	1	0	0	0	0	2	0	0	2	67
Open Prehistoric Site, Unknown Function	0	5	0	0	5	1	0	0	0	0	1	0	0	2	0	0	14
Other Specialized Aboriginal Site	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Unknown Function Open Site Greater than 20 m Radius	0	2	0	0	2	2	2	2	0	0	2	0	1	0	0	1	14
Unknown Function Surface Scatter Less than 20 m Radius	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
(blank)	0	2	0	0	1	1	1	0	0	0	1	0	0	0	0	3	9
Part of Multi-Component Site	0	76	1	1	45	38	6	5	1	1	3	0	2	2	0	5	186
Total	0	113	3	1	73	49	16	8	1	1	8	0	5	4	0	11	293

Chart 5 - Region 4 Site Types by Landform, Terminal Archaic Period

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	0	0	1	0	0	0	0	0	1	0	0	3
Lithic Reduction	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3
Open Habitation, Prehistoric	0	6	0	0	1	3	1	1	1	0	0	0	0	0	0	0	13
Open Prehistoric Site, Unknown Function	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	2	4
Unknown Function Open Site Greater than 20 m Radius	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
(blank)	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Part of Multi-Component Site	0	40	0	1	19	19	3	4	1	2	1	0	1	1	0	3	95
Total	0	49	0	1	23	23	5	6	2	2	1	0	1	2	0	5	120

Chart 6 - Region 4 Site Types by Landform, Early Woodland Period

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	1	1	0	0	0	0	0	0	0	0	0	0	2
Open Prehistoric Site, Unknown Function	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	1	0	0	0	0	0	0	0	0	1
Unknown Function Surface Scatter Less than 20 m Radius	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
(blank)	0	0	0	0	0	1	0	0		0	0	0	0	0	0	0	1
Part of Multi-Component Site	0	15	0	0	20	18	2	7	1	1	1	0	2	0	0	4	71
Total	0	16	0	0	21	20	2	8	1	2	1	0	2	0	0	4	77

Chart 7 - Region 4 Site Types by Landform, Middle Woodland Period

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	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Open Habitation, Prehistoric	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Open Prehistoric Site, Unknown Function	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	3
Part of Multi-Component Site	0	28	0	0	26	16	1	9	1	1	0	0	1	0	0	4	87
Total	0	29	0	0	27	18	1	9	1	2	0	0	1	0	0	4	92

Chart 8 - Region 4 Site Types by Landform, Late Woodland Period

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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Lithic Reduction	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Open Habitation, Prehistoric	0	17	1	0	7	4	1	0	1	0	0	0	0	0	0	0	31
Open Prehistoric Site, Unknown Function	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	3
Rock shelter/cave	0	0	0	0	0	0	0	3	0	0	1	0	0	0	0	0	4
Unknown Function Open Site Greater than 20 m Radius	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Village	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
(blank)	0	4	0	0	1	0	0	1	0	0	0	0	0	0	0	1	7
Part of Multi-Component Site	0	66	2	1	23	24	7	8	1	3	1	0	2	1	1	3	143
Total	0	91	3	1	32	30	8	12	2	4	2	0	2	1	1	4	193

Chart 9 - Region 5 Site Types by Landform, Paleoindian Period

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	1	0	0	0	0	0	0	0	0	0	0	2
Open Habitation, Prehistoric	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	3
(blank)	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Part of Multi-Component Site	0	21	0	1	2	8	1	0	1	0	1	0	0	0	0	0	35
Total	0	24	0	1	3	10	1	0	1	0	1	0	0	0	0	0	41

Chart 10 - Region 5 Site Types by Landform, Early Archaic Period

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	1	0	0	0	2	0	0	0	0	0	0	0	0	0	1	4
(blank)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Part of Multi-Component Site	0	31	3	0	5	11	0	0	0	1	0	0	0	2	0	1	54
Total	0	32	3	0	5	13	0	0	1	1	0	0	0	2	0	2	59

Chart 11 - Region 5 Site Types by Landform, Middle Archaic Period

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	3	0	0	2	3	0	0	0	0	0	0	0	0	0	0	8
Open Prehistoric Site, Unknown Function	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
Part of Multi-Component Site	0	62	4	1	16	21	1	0	0	1	1	0	0	2	3	2	114
Total	0	68	4	1	18	24	1	0	0	1	1	0	0	2	3	2	125

Chart 12 - Region 5 Site Types by Landform, Late Archaic Period

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	2	0	1	1	0	0	0	1	0	0	1	0	0	0	7
Open Habitation, Prehistoric	0	29	0	0	12	7	3	1	0	0	0	0	0	1	0	0	53
Open Prehistoric Site, Unknown Function	0	4	0	0	0	3	0	0	0	1	0	1	0	1	0	0	10
(blank)	0	3	0	0	2	2	0	0	1	0	0	0	0	0	0	1	9
Part of Multi-Component Site	0	171	1	0	26	45	4	2	1	1	1	0	0	3	2	5	262
Total	0	208	3	0	41	58	7	3	2	3	1	1	1	5	2	6	341

Chart 13 - Region 5 Site Types by Landform, Terminal Archaic Period

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	26	0	0	5	4	0	0	0	0	0	0	0	0	0	0	35
Open Prehistoric Site, Unknown Function	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	2
Other Specialized Aboriginal Site	0	1	0	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	0	0	0	0	2	2
Part of Multi-Component Site	0	150	3	4	20	37	6	1	2	2	1	0	0	1	2	3	232
Total	0	177	3	4	26	41	6	1	2	2	1	1	0	1	2	5	272

Chart 14 - Region 5 Site Types by Landform, Early Woodland Period

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	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	4
Open Prehistoric Site, Unknown Function	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	2
Part of Multi-Component Site	0	67	3	1	8	16	1	0	1	1	0	0	0	1	1	6	106
Total	0	70	3	1	9	18	1	0	1	1	0	0	0	1	1	6	112

Chart 15 - Region 5 Site Types by Landform, Middle Woodland Period

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	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
(blank)	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Part of Multi-Component Site	0	48	1	1	5	15	2	0	0	0	1	0	0	0	2	5	80
Total	0	53	1	1	5	15	2	0	0	0	1	0	0	0	2	5	85

Chart 16 - Region 5 Site Types by Landform, Late Woodland Period

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	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
Cemetery	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Open Habitation, Prehistoric	0	42	0	1	3	6	1	0	0	0	0	0	0	0	0	0	53
Open Prehistoric Site, Unknown Function	0	3	0	1	1	1	0	0	1	0	0	0	0	0	0	0	7
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	1	0	0	0	0	0	0	0	0	0	0	1
Village	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	1	4
(blank)	0	9	0	0	0	2	0	0	0	1	0	0	0	1	0	1	14
Part of Multi-Component Site	0	134	4	6	21	33	3	1	1	2	1	0	0	0	1	3	210
Total	0	192	4	8	25	46	4	1	2	3	1	0	0	1	1	5	293

Chart 17 - Region 6 Site Types by Landform, Paleoindian Period

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	0	0	0	0	0	0	1
Open Habitation, Prehistoric	0	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	4
(blank)	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Part of Multi-Component Site	0	3	1	0	0	2	0	0	0	0	0	0	0	0	0	1	7
Total	0	7	1	0	0	4	0	0	0	0	0	0	0	0	0	1	13

Chart 18 - Region 6 Site Types by Landform, Early Archaic Period

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	1	0	0	0	2	0	0	0	0	1	0	0	0	0	0	4
Total	0	1	0	0	0	2	0	0	0	0	1	0	0	0	0	0	4

Chart 19 - Region 6 Site Types by Landform, Middle Archaic Period

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	3	0	0	0	1	0	0	0	0	0	0	0	0	0	1	5
Part of Multi-Component Site	0	5	2	0	1	0	0	0	0	0	0	0	0	0	0	0	8
Total	0	8	2	0	1	1	0	0	0	0	0	0	0	0	0	1	13

Chart 20 - Region 6 Site Types by Landform, Late Archaic Period

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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Lithic Reduction	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Open Habitation, Prehistoric	0	2	0	0	0	2	0	0	0	1	3	0	0	0	0	1	9
Open Prehistoric Site, Unknown Function	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
Rock shelter/cave	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Part of Multi-Component Site	0	18	2	1	1	6	0	2	0	0	0	0	1	0	1	2	34
Total	0	22	2	1	1	8	0	2	0	1	4	0	1	0	1	5	48

Chart 21 - Region 6 Site Types by Landform, Terminal Archaic Period

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	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Open Habitation, Prehistoric	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
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	0	0	0	1	0	0	0	0	0	0	0	0	1
(blank)	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Part of Multi-Component Site	0	22	1	1	1	3	1	2	0	0	0	0	1	0	1	1	34
Total	0	26	1	1	1	4	1	3	1	0	0	0	1	0	1	1	41

Chart 22 - Region 6 Site Types by Landform, Early Woodland Period

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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Open Habitation, Prehistoric	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
Rock shelter/cave	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	2
Part of Multi-Component Site	0	15	0	1	1	5	1	0	0	0	0	0	0	0	2	2	27
Total	0	17	0	1	1	6	2	1	0	0	0	0	0	0	2	2	32

Chart 23 - Region 6 Site Types by Landform, Middle Woodland Period

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	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Isolated Find	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
(blank)	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Part of Multi-Component Site	0	12	1	0	1	4	1	1	0	1	0	0	0	0	1	3	25
Total	0	15	2	0	1	4	1	1	0	1	0	0	0	0	1	3	29

Chart 24 - Region 6 Site Types by Landform, Late Woodland Period

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	0	0	0	0	0	0	1
Open Habitation, Prehistoric	0	13	0	1	0	1	0	0	0	0	0	0	0	0	0	0	15
Open Prehistoric Site, Unknown Function	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	3
Rock shelter/cave	0	0	0	0	1	0	0	0	0	0	1	0	0	0	2	0	4
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
Village	0	4	0	0	0	1	0	0	1	0	0	0	0	0	0	0	6
(blank)	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Part of Multi-Component Site	0	22	2	1	1	4	2	2	0	1	0	0	0	0	2	5	42
Total	0	43	2	2	2	7	2	2	1	1	1	0	0	0	5	5	73

APPENDIX C

VARIABLES CONSIDERED

WITHIN REGIONS 4, 5, AND 6

Predictor	Family	Measure	Neighborhood Sizes	Description
aspect	Topography	bearing	n/a	Orientation of slope relative to north
aws050	Soils - aggregate	water storage - integer	n/a	water that is available to plants in the top 50cm of soil. AWS is expressed as centimeters of water, reported as the average of all components in the map unit.
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
drcdry	Soils - aggregate	classification, nominal	n/a	drainage class (dominant condition) - The NRCS describes natural soil drainage classes that represent the moisture condition of the soil in its natural condition throughout the year
drcwet	Soils - aggregate	classification, nominal	n/a	drainage class (wet conditions) - The NRCS describes natural soil drainage classes that represent the moisture condition of the wettest soil component in its natural condition throughout the year

Predictor	Family	Measure	Neighborhood Sizes	Description
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)
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

Predictor	Family	Measure	Neighborhood Sizes	Description
niccdcd	Soils - aggregate	classification, nominal	n/a	The broadest category in the land capability classification system for soils; the dominant capability class, under nonirrigated conditions, for the map unit based on composition percentage of all components in the map unit.
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
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 D

VARIABLES SELECTED FOR EACH OF 36 MODELS WITHIN REGIONS 4, 5, AND 6

Region 4/5 East - Riverine Section 1				
Predictor	**Mean D**	**Mean KS p**	Mean U	**Mean MW p**
aspect	0.274	p < 0.001	10417313	p < 0.001
c_hyd_min	0.246	p < 0.001	13364786	p < 0.001
c_trail_dist	0.443	p < 0.001	8021388	p < 0.001
cd_conf	0.303	p < 0.001	10949137	p < 0.001
cd_drnh	0.369	p < 0.001	11000542	p < 0.001
cd_h1	0.505	p < 0.001	6221029	p < 0.001
drcwet	0.319	p < 0.001	22053527	p < 0.001
ed_h4	0.267	p < 0.001	12069778	p < 0.001
ed_h5	0.388	p < 0.001	10079123	p < 0.001
ed_h6	0.509	p < 0.001	7854123	p < 0.001
eldrop10c	0.290	p < 0.001	13155832	p < 0.001
elev_2_strm	0.480	p < 0.001	7184140	p < 0.001
niccdcd	0.301	p < 0.001	21148688	p < 0.001
rel_32c	0.314	p < 0.001	11932534	p < 0.001
rng_32c	0.257	p < 0.001	12742755	p < 0.001
slope_pct	0.273	p < 0.001	13582805	p < 0.001
tpi_50c	0.260	p < 0.001	19523029	p < 0.001
tpi_sd50c	0.261	p < 0.001	19526853	p < 0.001
twi32c	0.267	p < 0.001	19189904	p < 0.001
vrf_8c	0.261	p < 0.001	11502208	p < 0.001
random	0.038	p = 0.303	16587405	p = 0.664

Region 4/5 East - Riverine Section 2				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
c_trail_dist	0.755	p < 0.001	7148148	p < 0.001
cd_conf	0.364	p < 0.001	13281934	p < 0.001
cd_drnh	0.561	p < 0.001	9105591	p < 0.001
cd_h5	0.355	p < 0.001	15415809	p < 0.001
drcwet	0.364	p < 0.001	50751271	p < 0.001
ed_h1	0.609	p < 0.001	11496869	p < 0.001
ed_h4	0.426	p < 0.001	15976564	p < 0.001
ed_h6	0.481	p < 0.001	29066275	p < 0.001
elev_2_drainh	0.432	p < 0.001	46574198	p < 0.001
elev_2_strm	0.454	p < 0.001	21444191	p < 0.001
rng_16c	0.351	p < 0.001	18974770	p < 0.001
slpvr_32c	0.672	p < 0.001	14615458	p < 0.001
std_32c	0.402	p < 0.001	17685853	p < 0.001
tpi_10c	0.448	p < 0.001	51821033	p < 0.001
tpi_cls250c	0.422	p < 0.001	22075537	p < 0.001
tpi_sd10c	0.448	p < 0.001	51811250	p < 0.001
tri_32c	0.673	p < 0.001	14601590	p < 0.001
random	0.051	p = 0.001	33629107	p = 0.004

Region 4/5 East - Riverine Section 3				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.582	p < 0.001	698745314	p < 0.001
c_trail_dist	0.522	p < 0.001	231347984	p < 0.001
cd_drnh	0.560	p < 0.001	201349194	p < 0.001
cd_h4	0.372	p < 0.001	229755698	p < 0.001
drcdry	0.319	p < 0.001	501224516	p < 0.001
e_hyd_min	0.399	p < 0.001	614502170	p < 0.001
ed_h2	0.529	p < 0.001	577738478	p < 0.001
ed_h5	0.463	p < 0.001	238058450	p < 0.001
ed_h6	0.731	p < 0.001	79077682	p < 0.001
elev_2_strm	0.318	p < 0.001	359525372	p < 0.001
niccdcd	0.326	p < 0.001	326365480	p < 0.001
rel_16c	0.364	p < 0.001	606176204	p < 0.001
rng_16c	0.321	p < 0.001	294298883	p < 0.001
std_16c	0.301	p < 0.001	308457568	p < 0.001
tpi_50c	0.436	p < 0.001	621683178	p < 0.001
tpi_cls50c	0.397	p < 0.001	598023190	p < 0.001
tpi_sd50c	0.436	p < 0.001	621485004	p < 0.001
vrf_32c	0.340	p < 0.001	288562000	p < 0.001
random	0.013	p = 0.056	415961256	p = 0.489

Region 4/5 East - Riverine Section 4				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.185	p < 0.001	3148829414	p < 0.001
c_trail_dist	0.270	p < 0.001	2105153512	p < 0.001
drcdry	0.280	p < 0.001	3586998278	p < 0.001
e_hyd_min_wt	0.229	p < 0.001	3358938134	p < 0.001
ed_h2	0.273	p < 0.001	2997880767	p < 0.001
ed_h5	0.268	p < 0.001	1772307944	p < 0.001
ed_h6	0.382	p < 0.001	1438021292	p < 0.001
eldrop32c	0.199	p < 0.001	3421441269	p < 0.001
elev_2_conf	0.179	p < 0.001	3277510744	p < 0.001
elev_2_drainh	0.142	p < 0.001	3214796901	p < 0.001
niccdcd	0.221	p < 0.001	2155940275	p < 0.001
rel_16c	0.319	p < 0.001	3882202126	p < 0.001
rng_32c	0.183	p < 0.001	2269487092	p < 0.001
std_32c	0.145	p < 0.001	2351494094	p < 0.001
tpi_10c	0.309	p < 0.001	3939203143	p < 0.001
tpi_cls10c	0.145	p < 0.001	3226054108	p < 0.001
tpi_sd10c	0.309	p < 0.001	3939064244	p < 0.001
twi32c	0.150	p < 0.001	2189745129	p < 0.001
vrf_32c	0.148	p < 0.001	2418970621	p < 0.001
random	0.005	p = 0.517	2781558678	p = 0.547

Region 4/5 East - Riverine Section 5				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.407	p < 0.001	508405391	p < 0.001
c_trail_dist	0.571	p < 0.001	150894783	p < 0.001
cd_conf	0.288	p < 0.001	251645087	p < 0.001
cd_drnh	0.465	p < 0.001	157797931	p < 0.001
cd_h4	0.300	p < 0.001	244404565	p < 0.001
e_hyd_min_wt	0.262	p < 0.001	478571855	p < 0.001
ed_h2	0.309	p < 0.001	431437235	p < 0.001
ed_h5	0.398	p < 0.001	216804407	p < 0.001
ed_h6	0.490	p < 0.001	180744177	p < 0.001
elev_2_strm	0.293	p < 0.001	308146529	p < 0.001
rel_16c	0.398	p < 0.001	553617170	p < 0.001
rng_32c	0.419	p < 0.001	195598476	p < 0.001
slpvr_32c	0.403	p < 0.001	203473719	p < 0.001
std_32c	0.386	p < 0.001	197906785	p < 0.001
tpi_10c	0.404	p < 0.001	590590415	p < 0.001
tpi_cls10c	0.350	p < 0.001	511721859	p < 0.001
tpi_sd10c	0.404	p < 0.001	590613604	p < 0.001
tri_32c	0.404	p < 0.001	203109521	p < 0.001
vrf_32c	0.313	p < 0.001	259698782	p < 0.001
random	0.007	p = 0.661	377294394	p = 0.547

Region 4/5 East - Riverine Section 6				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.418	$p < 0.001$	1852664464	$p < 0.001$
c_trail_dist	0.498	$p < 0.001$	582191041	$p < 0.001$
cd_conf	0.282	$p < 0.001$	1229839986	$p < 0.001$
cd_drnh	0.387	$p < 0.001$	747737335	$p < 0.001$
e_hyd_min_wt	0.381	$p < 0.001$	1788616814	$p < 0.001$
ed_h2	0.311	$p < 0.001$	1319904836	$p < 0.001$
ed_h5	0.343	$p < 0.001$	836452141	$p < 0.001$
ed_h6	0.461	$p < 0.001$	606685678	$p < 0.001$
eldrop32c	0.439	$p < 0.001$	1848611965	$p < 0.001$
elev_2_conf	0.282	$p < 0.001$	1533045754	$p < 0.001$
elev_2_strm	0.284	$p < 0.001$	1397219277	$p < 0.001$
rel_32c	0.554	$p < 0.001$	2035509020	$p < 0.001$
rng_32c	0.480	$p < 0.001$	722275831	$p < 0.001$
slpvr_32c	0.426	$p < 0.001$	790310141	$p < 0.001$
std_32c	0.407	$p < 0.001$	868649450	$p < 0.001$
tpi_10c	0.460	$p < 0.001$	1963643860	$p < 0.001$
tpi_cls10c	0.387	$p < 0.001$	1704210672	$p < 0.001$
tpi_sd10c	0.460	$p < 0.001$	1963467073	$p < 0.001$
tri_32c	0.427	$p < 0.001$	788440254	$p < 0.001$
random	0.010	$p = 0.050$	1221462631	$p = 0.229$

Region 4/5 East - Riverine Section 7				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.294	p < 0.001	3250980077	p < 0.001
c_trail_dist	0.288	p < 0.001	1841399994	p < 0.001
cd_conf	0.170	p < 0.001	2354927337	p < 0.001
cd_drnh	0.178	p < 0.001	1777327084	p < 0.001
cd_h2	0.227	p < 0.001	2625042095	p < 0.001
drcwet	0.270	p < 0.001	3157851033	p < 0.001
e_hyd_min_wt	0.344	p < 0.001	3367628962	p < 0.001
ed_h5	0.181	p < 0.001	2076573303	p < 0.001
ed_h6	0.260	p < 0.001	1595244664	p < 0.001
eldrop32c	0.179	p < 0.001	2941148473	p < 0.001
elev_2_conf	0.258	p < 0.001	3028025339	p < 0.001
elev_2_strm	0.162	p < 0.001	2662061704	p < 0.001
niccdcd	0.184	p < 0.001	2158071217	p < 0.001
rel_32c	0.273	p < 0.001	3214951828	p < 0.001
tpi_50c	0.254	p < 0.001	3150379176	p < 0.001
tpi_cls50c	0.246	p < 0.001	2990454441	p < 0.001
tpi_sd50c	0.254	p < 0.001	3149953093	p < 0.001
twi32c	0.171	p < 0.001	1830278705	p < 0.001
random	0.005	p = 0.512	2375537967	p = 0.527

Region 4/5 East - Upland Section 1				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aspect	0.356	p < 0.001	19855992	p < 0.001
aws050	0.468	p < 0.001	41196089	p < 0.001
c_hyd_min	0.332	p < 0.001	23231140	p < 0.001
c_trail_dist	0.619	p < 0.001	14507662	p < 0.001
cd_conf	0.582	p < 0.001	14610116	p < 0.001
cd_drnh	0.387	p < 0.001	20900131	p < 0.001
cd_h2	0.404	p < 0.001	21153955	p < 0.001
cd_h5	0.466	p < 0.001	17864496	p < 0.001
cd_h7	0.731	p < 0.001	9085528	p < 0.001
ed_h4	0.419	p < 0.001	37592954	p < 0.001
elev_2_conf	0.490	p < 0.001	17954712	p < 0.001
elev_2_drainh	0.482	p < 0.001	17580672	p < 0.001
elev_2_strm	0.700	p < 0.001	7959548	p < 0.001
flowdir	0.312	p < 0.001	20858347	p < 0.001
niccdcd	0.525	p < 0.001	18698872	p < 0.001
rng_32c	0.394	p < 0.001	23693939	p < 0.001
std_32c	0.318	p < 0.001	27082524	p < 0.001
tpi_250c	0.541	p < 0.001	12376052	p < 0.001
tpi_cls250c	0.534	p < 0.001	11711535	p < 0.001
tpi_sd250c	0.541	p < 0.001	12387611	p < 0.001
random	0.015	p = 0.949	31004045	p = 0.847

Region 4/5 East - Upland Section 2				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
c_hyd_min	0.381	p < 0.001	125692968	p < 0.001
c_trail_dist	0.908	p < 0.001	9263463	p < 0.001
cd_conf	0.503	p < 0.001	21031529	p < 0.001
cd_drnh	0.497	p < 0.001	13686581	p < 0.001
cd_h2	0.480	p < 0.001	18946503	p < 0.001
cd_h4	0.531	p < 0.001	14419007	p < 0.001
cd_h5	0.418	p < 0.001	22172725	p < 0.001
cd_h7	0.707	p < 0.001	11758884	p < 0.001
eldrop10c	0.391	p < 0.001	94110482	p < 0.001
elev_2_conf	0.489	p < 0.001	106693119	p < 0.001
elev_2_strm	0.646	p < 0.001	67365142	p < 0.001
niccdcd	0.496	p < 0.001	94103083	p < 0.001
rng_32c	0.608	p < 0.001	56868032	p < 0.001
slope_deg	0.434	p < 0.001	80638251	p < 0.001
slpvr_32c	0.726	p < 0.001	35758344	p < 0.001
std_32c	0.580	p < 0.001	62716455	p < 0.001
tpi_250c	0.378	p < 0.001	174200245	p = 0.325
tpi_cls250c	0.371	p < 0.001	152038364	p < 0.001
tpi_sd250c	0.378	p < 0.001	174160134	p = 0.338
tri_32c	0.737	p < 0.001	35659068	p < 0.001
random	0.011	p = 0.498	173779476	p = 0.475

Region 4/5 East - Upland Section 3				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
aws050	0.687	p < 0.001	584451732	p < 0.001
c_hyd_min	0.623	p < 0.001	70080826	p < 0.001
c_trail_dist	0.886	p < 0.001	33845091	p < 0.001
cd_conf	0.731	p < 0.001	30656264	p < 0.001
cd_drnh	0.678	p < 0.001	54602173	p < 0.001
cd_h2	0.669	p < 0.001	57657642	p < 0.001
cd_h4	0.737	p < 0.001	35658464	p < 0.001
cd_h5	0.660	p < 0.001	50165750	p < 0.001
cd_h6	0.943	p < 0.001	10262607	p < 0.001
drcwet	0.437	p < 0.001	183563897	p < 0.001
eldrop32c	0.545	p < 0.001	99268450	p < 0.001
elev_2_conf	0.733	p < 0.001	46954934	p < 0.001
elev_2_drainh	0.585	p < 0.001	137881328	p < 0.001
elev_2_strm	0.855	p < 0.001	19827644	p < 0.001
niccdcd	0.720	p < 0.001	71450447	p < 0.001
rel_32c	0.475	p < 0.001	169060105	p < 0.001
rng_32c	0.630	p < 0.001	89278135	p < 0.001
std_32c	0.585	p < 0.001	105011467	p < 0.001
tpi_250c	0.689	p < 0.001	108813465	p < 0.001
tpi_cls250c	0.601	p < 0.001	116951120	p < 0.001
tpi_sd250c	0.689	p < 0.001	108831675	p < 0.001
vrf_32c	0.455	p < 0.001	136237634	p < 0.001
random	0.010	p = 0.306	325140918	p = 0.710

Region 4/5 East - Upland Section 4

Predictor	Mean D	Mean KS p	Mean U	Mean MW p
aws050	0.351	p < 0.001	1157666266	p < 0.001
c_hyd_min_wt	0.510	p < 0.001	318101739	p < 0.001
c_trail_dist	0.634	p < 0.001	226440147	p < 0.001
cd_conf	0.645	p < 0.001	172280115	p < 0.001
cd_drnh	0.436	p < 0.001	349073907	p < 0.001
cd_h2	0.495	p < 0.001	254992463	p < 0.001
cd_h4	0.666	p < 0.001	195439677	p < 0.001
cd_h5	0.515	p < 0.001	261083039	p < 0.001
cd_h6	0.806	p < 0.001	69355069	p < 0.001
eldrop32c	0.418	p < 0.001	400534374	p < 0.001
elev_2_conf	0.564	p < 0.001	308765590	p < 0.001
elev_2_drainh	0.311	p < 0.001	541611849	p < 0.001
elev_2_strm	0.715	p < 0.001	133342509	p < 0.001
niccdcd	0.413	p < 0.001	405964517	p < 0.001
rng_8c	0.420	p < 0.001	395431789	p < 0.001
slope_pct	0.398	p < 0.001	392879181	p < 0.001
std_10c	0.397	p < 0.001	413180520	p < 0.001
tpi_250c	0.522	p < 0.001	344792171	p < 0.001
tpi_cls250c	0.448	p < 0.001	426767561	p < 0.001
tpi_sd250c	0.522	p < 0.001	344697331	p < 0.001
vrf_32c	0.339	p < 0.001	488748810	p < 0.001
random	0.006	p = 0.421	818764579	p = 0.539

Region 4/5 East - Upland Section 5				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
c_hyd_min	0.623	p < 0.001	16955451	p < 0.001
c_trail_dist	0.729	p < 0.001	13664295	p < 0.001
cd_conf	0.741	p < 0.001	9426214	p < 0.001
cd_drnh	0.672	p < 0.001	14538042	p < 0.001
cd_h2	0.655	p < 0.001	13909262	p < 0.001
cd_h4	0.614	p < 0.001	19351987	p < 0.001
cd_h5	0.664	p < 0.001	12971505	p < 0.001
cd_h6	0.438	p < 0.001	38003968	p < 0.001
eldrop32c	0.520	p < 0.001	21406212	p < 0.001
elev_2_conf	0.636	p < 0.001	16434568	p < 0.001
elev_2_drainh	0.428	p < 0.001	44011998	p < 0.001
elev_2_strm	0.416	p < 0.001	47179246	p < 0.001
rel_32c	0.366	p < 0.001	34456550	p < 0.001
rng_32c	0.551	p < 0.001	26652639	p < 0.001
std_32c	0.474	p < 0.001	29704861	p < 0.001
tpi_100c	0.468	p < 0.001	31023403	p < 0.001
tpi_cls250c	0.379	p < 0.001	48129306	p < 0.001
tpi_sd100c	0.468	p < 0.001	31014661	p < 0.001
vrf_32c	0.392	p < 0.001	36469979	p < 0.001
random	0.018	p = 0.437	64625377	p = 0.384

Region 4/5 East - Upland Section 6				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.715	p < 0.001	2085397273	p < 0.001
c_hyd_min	0.597	p < 0.001	334235081	p < 0.001
c_trail_dist	0.768	p < 0.001	327109683	p < 0.001
cd_conf	0.794	p < 0.001	115059610	p < 0.001
cd_drnh	0.649	p < 0.001	238843948	p < 0.001
cd_h2	0.674	p < 0.001	197433559	p < 0.001
cd_h4	0.791	p < 0.001	129110421	p < 0.001
cd_h5	0.773	p < 0.001	116938497	p < 0.001
cd_h6	0.873	p < 0.001	34091321	p < 0.001
eldrop32c	0.496	p < 0.001	504036014	p < 0.001
elev_2_conf	0.665	p < 0.001	315694169	p < 0.001
elev_2_drainh	0.568	p < 0.001	479272209	p < 0.001
elev_2_strm	0.799	p < 0.001	182634380	p < 0.001
niccdcd	0.544	p < 0.001	399447824	p < 0.001
rng_32c	0.672	p < 0.001	250534754	p < 0.001
slope_pct	0.451	p < 0.001	476206690	p < 0.001
slpvr_16c	0.445	p < 0.001	607637745	p < 0.001
std_32c	0.632	p < 0.001	371993089	p < 0.001
tpi_250c	0.654	p < 0.001	270819866	p < 0.001
tpi_cls250c	0.618	p < 0.001	439190850	p < 0.001
tpi_sd250c	0.654	p < 0.001	271381770	p < 0.001
tri_16c	0.479	p < 0.001	560554926	p < 0.001
random	0.011	p = 0.028	1150641381	p = 0.205

Region 4/5 East - Upland Section 7				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
aws050	0.482	p < 0.001	2246960906	p < 0.001
c_hyd_min	0.485	p < 0.001	585856656	p < 0.001
c_trail_dist	0.675	p < 0.001	290072880	p < 0.001
cd_conf	0.653	p < 0.001	249869247	p < 0.001
cd_drnh	0.465	p < 0.001	430319586	p < 0.001
cd_h2	0.486	p < 0.001	462393295	p < 0.001
cd_h4	0.666	p < 0.001	242363689	p < 0.001
cd_h5	0.594	p < 0.001	295356996	p < 0.001
cd_h6	0.779	p < 0.001	96184057	p < 0.001
eldrop32c	0.437	p < 0.001	594957004	p < 0.001
elev_2_conf	0.592	p < 0.001	428770476	p < 0.001
elev_2_drainh	0.348	p < 0.001	756713770	p < 0.001
elev_2_strm	0.716	p < 0.001	193049234	p < 0.001
niccdcd	0.343	p < 0.001	751448768	p < 0.001
rel_32c	0.390	p < 0.001	644867820	p < 0.001
rng_16c	0.364	p < 0.001	771676560	p < 0.001
slope_pct	0.368	p < 0.001	724394348	p < 0.001
std_8c	0.353	p < 0.001	718395156	p < 0.001
tpi_250c	0.613	p < 0.001	428863524	p < 0.001
tpi_cls250c	0.610	p < 0.001	533519235	p < 0.001
tpi_sd250c	0.613	p < 0.001	428951344	p < 0.001
random	0.005	p = 0.599	1410315183	p = 0.582

Region 4/5 West - Riverine Section 1				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.258	p < 0.001	250881448	p < 0.001
c_trail_dist	0.251	p < 0.001	185568230	p < 0.001
cd_drnh	0.297	p < 0.001	128290112	p < 0.001
cd_h4	0.286	p < 0.001	155183200	p < 0.001
cd_h5	0.256	p < 0.001	154431280	p < 0.001
cd_h6	0.325	p < 0.001	134257528	p < 0.001
e_hyd_min	0.346	p < 0.001	283896717	p < 0.001
ed_h2	0.337	p < 0.001	281223176	p < 0.001
elev_2_drainh	0.267	p < 0.001	245333015	p < 0.001
rel_10c	0.322	p < 0.001	285745859	p < 0.001
rng_16c	0.438	p < 0.001	96558540	p < 0.001
slpvr_8c	0.360	p < 0.001	128964704	p < 0.001
std_16c	0.432	p < 0.001	104213494	p < 0.001
tpi_10c	0.445	p < 0.001	313222412	p < 0.001
tpi_cls10c	0.417	p < 0.001	293893436	p < 0.001
tpi_sd10c	0.444	p < 0.001	313278823	p < 0.001
tri_8c	0.362	p < 0.001	126665781	p < 0.001
vrf_8c	0.222	p < 0.001	145587235	p < 0.001
random	0.009	p = 0.604	202233993	p = 0.789

Region 4/5 West - Riverine Section 2				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
drcdry	0.256	p < 0.001	278877529	p < 0.001
e_hyd_min	0.284	p < 0.001	290294089	p < 0.001
ed_conf	0.214	p < 0.001	176006618	p < 0.001
ed_drnh	0.242	p < 0.001	286691905	p < 0.001
ed_h2	0.292	p < 0.001	297082236	p < 0.001
ed_h5	0.368	p < 0.001	158105345	p < 0.001
ed_h6	0.320	p < 0.001	173188178	p < 0.001
niccdcd	0.275	p < 0.001	162342584	p < 0.001
rel_16c	0.375	p < 0.001	327836740	p < 0.001
rng_32c	0.358	p < 0.001	127486611	p < 0.001
slpvr_32c	0.215	p < 0.001	175823470	p < 0.001
std_32c	0.357	p < 0.001	128568609	p < 0.001
tpi_10c	0.364	p < 0.001	329290114	p < 0.001
tpi_cls50c	0.333	p < 0.001	294465899	p < 0.001
tpi_sd10c	0.364	p < 0.001	329338770	p < 0.001
tri_32c	0.219	p < 0.001	174933533	p < 0.001
random	0.010	p = 0.486	219514213	p = 0.385

Region 4/5 West - Riverine Section 3				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
cd_h5	0.231	$p < 0.001$	347628615	$p < 0.001$
cd_h6	0.270	$p < 0.001$	313890366	$p < 0.001$
drcwet	0.190	$p < 0.001$	449660611	$p < 0.001$
e_hyd_min	0.262	$p < 0.001$	497719104	$p < 0.001$
e_trail_dist	0.259	$p < 0.001$	281254919	$p < 0.001$
ed_h2	0.249	$p < 0.001$	487822243	$p < 0.001$
ed_h4	0.374	$p < 0.001$	234841080	$p < 0.001$
elev_2_conf	0.199	$p < 0.001$	429691885	$p < 0.001$
rel_10c	0.301	$p < 0.001$	499038718	$p < 0.001$
rng_10c	0.285	$p < 0.001$	240971890	$p < 0.001$
slpvr_10c	0.269	$p < 0.001$	273019736	$p < 0.001$
std_10c	0.292	$p < 0.001$	245956314	$p < 0.001$
tpi_10c	0.336	$p < 0.001$	525722462	$p < 0.001$
tpi_cls10c	0.290	$p < 0.001$	481712829	$p < 0.001$
tpi_sd10c	0.336	$p < 0.001$	525670995	$p < 0.001$
tri_10c	0.274	$p < 0.001$	269874172	$p < 0.001$
vrf_32c	0.211	$p < 0.001$	294836436	$p < 0.001$
random	0.008	$p = 0.460$	378965073	$p = 0.561$

Region 4/5 West - Riverine Section 4				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.226	p < 0.001	291513211	p < 0.001
c_trail_dist	0.371	p < 0.001	167582077	p < 0.001
cd_drnh	0.307	p < 0.001	168579453	p < 0.001
cd_h4	0.244	p < 0.001	204269971	p < 0.001
drcdry	0.241	p < 0.001	331363392	p < 0.001
ed_h1	0.191	p < 0.001	221810811	p < 0.001
ed_h5	0.352	p < 0.001	155846701	p < 0.001
ed_h6	0.348	p < 0.001	166859948	p < 0.001
elev_2_drainh	0.305	p < 0.001	359433065	p < 0.001
niccdcd	0.304	p < 0.001	169080613	p < 0.001
rel_8c	0.282	p < 0.001	353482700	p < 0.001
rng_10c	0.210	p < 0.001	204380455	p < 0.001
std_10c	0.204	p < 0.001	209378798	p < 0.001
tpi_10c	0.276	p < 0.001	354283604	p < 0.001
tpi_cls100c	0.228	p < 0.001	222205603	p < 0.001
tpi_sd10c	0.276	p < 0.001	354261422	p < 0.001
vrf_32c	0.222	p < 0.001	198012048	p < 0.001
random	0.007	p = 0.767	263539656	p = 0.756

Region 4/5 West - Riverine Section 5				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.265	p < 0.001	755584059	p < 0.001
cd_conf	0.235	p < 0.001	495458661	p < 0.001
cd_drnh	0.129	p < 0.001	579310909	p < 0.001
cd_h4	0.213	p < 0.001	505853195	p < 0.001
cd_h5	0.167	p < 0.001	511554909	p < 0.001
drcwet	0.147	p < 0.001	662351287	p < 0.001
e_hyd_min	0.142	p < 0.001	688992347	p < 0.001
e_trail_dist	0.134	p < 0.001	580698344	p < 0.001
ed_h2	0.152	p < 0.001	676820983	p < 0.001
ed_h7	0.213	p < 0.001	493525944	p < 0.001
elev_2_conf	0.174	p < 0.001	530387822	p < 0.001
elev_2_strm	0.255	p < 0.001	462341322	p < 0.001
niccdcd	0.239	p < 0.001	472285176	p < 0.001
rel_10c	0.119	p < 0.001	656235205	p < 0.001
rng_10c	0.144	p < 0.001	516864621	p < 0.001
slope_pct	0.134	p < 0.001	529303098	p < 0.001
slpvr_32c	0.169	p < 0.001	694729848	p < 0.001
std_8c	0.149	p < 0.001	515850467	p < 0.001
tpi_10c	0.139	p < 0.001	723825528	p < 0.001
tpi_sd10c	0.139	p < 0.001	723940233	p < 0.001
tri_32c	0.167	p < 0.001	693106309	p < 0.001
random	0.010	p = 0.094	608294279	p = 0.047

Region 4/5 West - Riverine Section 6				
Predictor	Mean D	Mean KS p	Mean U	Mean MW p
c_trail_dist	0.466	p < 0.001	39857482	p < 0.001
cd_drnh	0.405	p < 0.001	35379951	p < 0.001
ed_h1	0.299	p < 0.001	55155742	p < 0.001
ed_h5	0.417	p < 0.001	42591257	p < 0.001
ed_h6	0.438	p < 0.001	39654503	p < 0.001
elev_2_drainh	0.524	p < 0.001	113344173	p < 0.001
niccdcd	0.320	p < 0.001	46607463	p < 0.001
rel_32c	0.338	p < 0.001	89821770	p < 0.001
rng_16c	0.427	p < 0.001	42573272	p < 0.001
slpvr_16c	0.326	p < 0.001	50248750	p < 0.001
std_32c	0.468	p < 0.001	34612159	p < 0.001
tpi_100c	0.509	p < 0.001	103627841	p < 0.001
tpi_cls50c	0.465	p < 0.001	106070621	p < 0.001
tpi_sd100c	0.509	p < 0.001	103628429	p < 0.001
tri_16c	0.332	p < 0.001	49810214	p < 0.001
random	0.028	p = 0.035	72875440	p = 0.089

Region 4/5 West - Upland Section 1				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.720	p < 0.001	197044392	p < 0.001
c_hyd_min_wt	0.612	p < 0.001	34925420	p < 0.001
c_trail_dist	0.487	p < 0.001	63753074	p < 0.001
cd_conf	0.786	p < 0.001	13900070	p < 0.001
cd_drnh	0.543	p < 0.001	43050130	p < 0.001
cd_h2	0.623	p < 0.001	34420170	p < 0.001
cd_h4	0.859	p < 0.001	9268964	p < 0.001
cd_h5	0.706	p < 0.001	22402087	p < 0.001
cd_h7	0.776	p < 0.001	11827276	p < 0.001
eldrop32c	0.612	p < 0.001	28707446	p < 0.001
elev_2_conf	0.744	p < 0.001	23733878	p < 0.001
elev_2_strm	0.790	p < 0.001	11370423	p < 0.001
niccdcd	0.574	p < 0.001	41384917	p < 0.001
rng_10c	0.662	p < 0.001	27449807	p < 0.001
slope_pct	0.495	p < 0.001	43055618	p < 0.001
std_16c	0.608	p < 0.001	30702622	p < 0.001
tpi_250c	0.661	p < 0.001	28534853	p < 0.001
tpi_cls250c	0.655	p < 0.001	40239005	p < 0.001
tpi_sd250c	0.661	p < 0.001	28513138	p < 0.001
tri_8c	0.427	p < 0.001	66210436	p < 0.001
random	0.014	p = 0.345	116549202	p = 0.699

Region 4/5 West - Upland Section 2				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.625	p < 0.001	181280184	p < 0.001
c_hyd_min	0.607	p < 0.001	35797665	p < 0.001
c_trail_dist	0.477	p < 0.001	46393745	p < 0.001
cd_conf	0.680	p < 0.001	19203446	p < 0.001
cd_drnh	0.514	p < 0.001	45531902	p < 0.001
cd_h2	0.605	p < 0.001	35711719	p < 0.001
cd_h4	0.618	p < 0.001	32661165	p < 0.001
cd_h5	0.618	p < 0.001	22388801	p < 0.001
cd_h7	0.708	p < 0.001	12494523	p < 0.001
eldrop32c	0.560	p < 0.001	33518033	p < 0.001
elev_2_conf	0.652	p < 0.001	27056422	p < 0.001
elev_2_drainh	0.488	p < 0.001	53253914	p < 0.001
elev_2_strm	0.737	p < 0.001	12584356	p < 0.001
niccdcd	0.606	p < 0.001	32739790	p < 0.001
rng_32c	0.674	p < 0.001	23037043	p < 0.001
slope_pct	0.478	p < 0.001	42558501	p < 0.001
std_32c	0.684	p < 0.001	24559525	p < 0.001
tpi_250c	0.628	p < 0.001	30601345	p < 0.001
tpi_cls250c	0.623	p < 0.001	37780910	p < 0.001
tpi_sd250c	0.628	p < 0.001	30585323	p < 0.001
random	0.027	p = 0.010	103636103	p = 0.001

Region 4/5 West - Upland Section 3				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.313	$p < 0.001$	213049995	$p < 0.001$
c_hyd_min	0.492	$p < 0.001$	62963830	$p < 0.001$
c_trail_dist	0.586	$p < 0.001$	47341470	$p < 0.001$
cd_conf	0.681	$p < 0.001$	30915817	$p < 0.001$
cd_h2	0.490	$p < 0.001$	65489922	$p < 0.001$
cd_h4	0.713	$p < 0.001$	30982341	$p < 0.001$
cd_h5	0.595	$p < 0.001$	44591450	$p < 0.001$
cd_h7	0.664	$p < 0.001$	33986263	$p < 0.001$
ed_drnh	0.426	$p < 0.001$	254020892	$p < 0.001$
eldrop32c	0.418	$p < 0.001$	73881712	$p < 0.001$
elev_2_conf	0.632	$p < 0.001$	45224309	$p < 0.001$
elev_2_drainh	0.407	$p < 0.001$	93557143	$p < 0.001$
elev_2_strm	0.600	$p < 0.001$	63456181	$p < 0.001$
niccdcd	0.349	$p < 0.001$	99693736	$p < 0.001$
rel_32c	0.371	$p < 0.001$	108177368	$p < 0.001$
rng_16c	0.492	$p < 0.001$	75705617	$p < 0.001$
slope_pct	0.354	$p < 0.001$	95643953	$p < 0.001$
std_32c	0.490	$p < 0.001$	75493182	$p < 0.001$
tpi_250c	0.415	$p < 0.001$	95210283	$p < 0.001$
tpi_cls250c	0.412	$p < 0.001$	99584588	$p < 0.001$
tpi_sd250c	0.415	$p < 0.001$	95174423	$p < 0.001$
random	0.013	$p = 0.308$	174934969	$p = 0.278$

Region 4/5 West - Upland Section 4				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.379	p < 0.001	150091666	p < 0.001
c_hyd_min	0.506	p < 0.001	38530363	p < 0.001
c_trail_dist	0.508	p < 0.001	45864224	p < 0.001
cd_conf	0.595	p < 0.001	24810817	p < 0.001
cd_drnh	0.486	p < 0.001	47415147	p < 0.001
cd_h2	0.536	p < 0.001	35551306	p < 0.001
cd_h4	0.549	p < 0.001	35103325	p < 0.001
cd_h5	0.609	p < 0.001	28522748	p < 0.001
cd_h7	0.650	p < 0.001	20336691	p < 0.001
eldrop32c	0.460	p < 0.001	43920441	p < 0.001
elev_2_conf	0.543	p < 0.001	30582921	p < 0.001
elev_2_drainh	0.362	p < 0.001	66592331	p < 0.001
elev_2_strm	0.633	p < 0.001	20485048	p < 0.001
rel_32c	0.498	p < 0.001	40066039	p < 0.001
rng_32c	0.393	p < 0.001	67288671	p < 0.001
std_32c	0.376	p < 0.001	64663108	p < 0.001
tpi_100c	0.634	p < 0.001	35476537	p < 0.001
tpi_cls100c	0.604	p < 0.001	39208209	p < 0.001
tpi_sd100c	0.635	p < 0.001	35386220	p < 0.001
random	0.018	p = 0.206	103933240	p = 0.218

Region 4/5 West - Upland Section 5				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.604	p < 0.001	781074094	p < 0.001
c_hyd_min	0.492	p < 0.001	198431454	p < 0.001
c_trail_dist	0.564	p < 0.001	138381393	p < 0.001
cd_conf	0.577	p < 0.001	149214260	p < 0.001
cd_drnh	0.479	p < 0.001	189691764	p < 0.001
cd_h2	0.512	p < 0.001	187917220	p < 0.001
cd_h4	0.521	p < 0.001	184095387	p < 0.001
cd_h5	0.627	p < 0.001	125979367	p < 0.001
cd_h7	0.607	p < 0.001	133554216	p < 0.001
eldrop32c	0.562	p < 0.001	123755240	p < 0.001
elev_2_conf	0.512	p < 0.001	189522433	p < 0.001
elev_2_strm	0.708	p < 0.001	94116429	p < 0.001
niccdcd	0.577	p < 0.001	167872954	p < 0.001
rel_32c	0.416	p < 0.001	219549118	p < 0.001
rng_32c	0.501	p < 0.001	164181845	p < 0.001
slope_deg	0.437	p < 0.001	187412469	p < 0.001
std_32c	0.508	p < 0.001	164294347	p < 0.001
tpi_250c	0.512	p < 0.001	218159901	p < 0.001
tpi_cls250c	0.371	p < 0.001	246457251	p < 0.001
tpi_sd250c	0.512	p < 0.001	218336542	p < 0.001
random	0.008	p = 0.449	464352295	p = 0.343

Region 4/5 West - Upland Section 6				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.475	p < 0.001	124863067	p < 0.001
c_hyd_min	0.549	p < 0.001	27245273	p < 0.001
c_trail_dist	0.729	p < 0.001	15274269	p < 0.001
cd_conf	0.676	p < 0.001	17707145	p < 0.001
cd_drnh	0.546	p < 0.001	26571363	p < 0.001
cd_h2	0.488	p < 0.001	32227474	p < 0.001
cd_h4	0.516	p < 0.001	35310089	p < 0.001
cd_h5	0.747	p < 0.001	12153896	p < 0.001
cd_h7	0.660	p < 0.001	21089037	p < 0.001
eldrop32c	0.554	p < 0.001	32599757	p < 0.001
elev_2_conf	0.625	p < 0.001	27908217	p < 0.001
elev_2_strm	0.653	p < 0.001	19014377	p < 0.001
niccdcd	0.574	p < 0.001	31048836	p < 0.001
rng_32c	0.657	p < 0.001	19612738	p < 0.001
slope_pct	0.439	p < 0.001	37747183	p < 0.001
slpvr_32c	0.566	p < 0.001	37653607	p < 0.001
std_32c	0.669	p < 0.001	21034661	p < 0.001
tpi_250c	0.585	p < 0.001	35237596	p < 0.001
tpi_cls250c	0.433	p < 0.001	39613717	p < 0.001
tpi_sd250c	0.584	p < 0.001	35338142	p < 0.001
tri_32c	0.574	p < 0.001	35837632	p < 0.001
random	0.018	p = 0.292	82812786	p = 0.802

Region 6 All - Riverine Section 1				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.475	$p < 0.001$	1377574058	$p < 0.001$
c_trail_dist	0.811	$p < 0.001$	148990363	$p < 0.001$
e_hyd_min	0.557	$p < 0.001$	1662980478	$p < 0.001$
ed_drnh	0.765	$p < 0.001$	1740504845	$p < 0.001$
ed_h2	0.631	$p < 0.001$	1765492293	$p < 0.001$
ed_h4	0.522	$p < 0.001$	577570748	$p < 0.001$
ed_h5	0.544	$p < 0.001$	532264524	$p < 0.001$
ed_h6	0.698	$p < 0.001$	352747533	$p < 0.001$
elev_2_drainh	0.719	$p < 0.001$	1690393656	$p < 0.001$
elev_2_strm	0.483	$p < 0.001$	595997820	$p < 0.001$
niccdcd	0.543	$p < 0.001$	510135531	$p < 0.001$
rng_16c	0.614	$p < 0.001$	432345830	$p < 0.001$
std_32c	0.633	$p < 0.001$	305132448	$p < 0.001$
tpi_10c	0.576	$p < 0.001$	1662133134	$p < 0.001$
tpi_cls10c	0.485	$p < 0.001$	1537654587	$p < 0.001$
tpi_sd10c	0.577	$p < 0.001$	1662107943	$p < 0.001$
vrf_32c	0.428	$p < 0.001$	524654464	$p < 0.001$
random	0.018	$p < 0.005$	985534525	$p = 0.039$

Region 6 All - Riverine Section 2				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aspect	0.490	$p < 0.001$	34806396	$p < 0.001$
aws050	0.505	$p < 0.001$	30834095	$p < 0.001$
c_trail_dist	0.646	$p < 0.001$	24200105	$p < 0.001$
cd_conf	0.514	$p < 0.001$	30681161	$p < 0.001$
cd_drnh	0.387	$p < 0.001$	55621122	$p < 0.001$
cd_h2	0.445	$p < 0.001$	39785945	$p < 0.001$
cd_h4	0.389	$p < 0.001$	39531883	$p < 0.001$
cd_h6	0.512	$p < 0.001$	26062271	$p < 0.001$
drcdry	0.496	$p < 0.001$	46085511	$p < 0.001$
e_hyd_min	0.446	$p < 0.001$	43232687	$p < 0.001$
ed_h5	0.521	$p < 0.001$	28810345	$p < 0.001$
eldrop32c	0.509	$p < 0.001$	32156705	$p < 0.001$
elev_2_drainh	0.425	$p < 0.001$	44500739	$p < 0.001$
elev_2_strm	0.425	$p < 0.001$	39188745	$p < 0.001$
niccdcd	0.495	$p < 0.001$	35570779	$p < 0.001$
rel_32c	0.382	$p < 0.001$	33094220	$p < 0.001$
rng_8c	0.501	$p < 0.001$	32751306	$p < 0.001$
slope_deg	0.494	$p < 0.001$	31196661	$p < 0.001$
slpvr_8c	0.510	$p < 0.001$	36881151	$p < 0.001$
std_8c	0.508	$p < 0.001$	33275813	$p < 0.001$
tri_8c	0.510	$p < 0.001$	36577453	$p < 0.001$
twi32c	0.467	$p < 0.001$	105963178	$p < 0.001$
vrf_10c	0.551	$p < 0.001$	26761535	$p < 0.001$
random	0.011	$p = 0.900$	71567391	$p = 0.811$

Region 6 All - Riverine Section 3				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aspect	0.344	p < 0.001	513171275	p < 0.001
aws050	0.359	p < 0.001	993019942	p < 0.001
c_trail_dist	0.441	p < 0.001	466010303	p < 0.001
cd_h4	0.433	p < 0.001	331206423	p < 0.001
cd_h5	0.460	p < 0.001	332758468	p < 0.001
cd_h6	0.501	p < 0.001	293800372	p < 0.001
e_hyd_min	0.426	p < 0.001	1171586006	p < 0.001
ed_h2	0.498	p < 0.001	1217134137	p < 0.001
eldrop32c	0.341	p < 0.001	481489412	p < 0.001
elev_2_strm	0.348	p < 0.001	481841767	p < 0.001
rel_32c	0.342	p < 0.001	477553624	p < 0.001
rng_8c	0.542	p < 0.001	257733803	p < 0.001
slope_pct	0.371	p < 0.001	355335225	p < 0.001
slpvr_8c	0.501	p < 0.001	290128455	p < 0.001
std_10c	0.537	p < 0.001	254645471	p < 0.001
tpi_5c	0.547	p < 0.001	1202587121	p < 0.001
tpi_cls10c	0.529	p < 0.001	1222273961	p < 0.001
tpi_sd5c	0.547	p < 0.001	1202751175	p < 0.001
tri_8c	0.508	p < 0.001	283946076	p < 0.001
vrf_32c	0.473	p < 0.001	282375971	p < 0.001
random	0.006	p = 0.467	751967526	p = 0.568

Region 6 All - Riverine Section 4				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.329	p < 0.001	249724463	p < 0.001
cd_h4	0.392	p < 0.001	118403993	p < 0.001
e_hyd_min_wt	0.385	p < 0.001	272672380	p < 0.001
e_trail_dist	0.674	p < 0.001	60404117	p < 0.001
ed_h2	0.369	p < 0.001	264144688	p < 0.001
ed_h5	0.459	p < 0.001	112126099	p < 0.001
ed_h6	0.564	p < 0.001	92464789	p < 0.001
elev_2_drainh	0.390	p < 0.001	119262554	p < 0.001
elev_2_strm	0.361	p < 0.001	124503318	p < 0.001
niccdcd	0.374	p < 0.001	102188083	p < 0.001
rng_8c	0.405	p < 0.001	102910631	p < 0.001
slpvr_8c	0.342	p < 0.001	110747003	p < 0.001
std_10c	0.404	p < 0.001	101196493	p < 0.001
tpi_250c	0.579	p < 0.001	77364442	p < 0.001
tpi_cls250c	0.503	p < 0.001	76725893	p < 0.001
tpi_sd250c	0.580	p < 0.001	77330088	p < 0.001
tri_8c	0.355	p < 0.001	108014280	p < 0.001
vrf_32c	0.309	p < 0.001	123601782	p < 0.001
random	0.013	p = 0.246	180954259	p = 0.570

Region 6 All - Riverine Section 5				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.405	p < 0.001	129218649	p < 0.001
c_trail_dist	0.609	p < 0.001	44308243	p < 0.001
cd_drnh	0.716	p < 0.001	16841611	p < 0.001
cd_h4	0.440	p < 0.001	67465781	p < 0.001
cd_h5	0.542	p < 0.001	51171796	p < 0.001
e_hyd_min_wt	0.549	p < 0.001	160132889	p < 0.001
ed_h2	0.504	p < 0.001	153425066	p < 0.001
elev_2_conf	0.435	p < 0.001	43435088	p < 0.001
elev_2_drainh	0.587	p < 0.001	152960416	p < 0.001
niccdcd	0.451	p < 0.001	65163258	p < 0.001
rng_8c	0.574	p < 0.001	29161275	p < 0.001
slope_deg	0.484	p < 0.001	39880485	p < 0.001
slpvr_8c	0.598	p < 0.001	34491014	p < 0.001
std_10c	0.577	p < 0.001	29555074	p < 0.001
tpi_50c	0.684	p < 0.001	157689459	p < 0.001
tpi_cls50c	0.674	p < 0.001	157978905	p < 0.001
tpi_sd50c	0.684	p < 0.001	157720076	p < 0.001
tri_8c	0.602	p < 0.001	33945306	p < 0.001
vrf_32c	0.623	p < 0.001	28329048	p < 0.001
random	0.015	p = 0.423	95347557	p = 0.446

Region 6 All - Upland Section 1				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aws050	0.759	p < 0.001	1655596919	p < 0.001
c_hyd_min	0.672	p < 0.001	201528833	p < 0.001
c_trail_dist	0.919	p < 0.001	45384229	p < 0.001
cd_conf	0.907	p < 0.001	42348608	p < 0.001
cd_h2	0.683	p < 0.001	237141457	p < 0.001
cd_h4	0.755	p < 0.001	158257447	p < 0.001
cd_h5	0.871	p < 0.001	56666171	p < 0.001
cd_h6	0.984	p < 0.001	3321599	p < 0.001
ed_drnh	0.904	p < 0.001	1834028096	p < 0.001
eldrop32c	0.658	p < 0.001	172057333	p < 0.001
elev_2_conf	0.848	p < 0.001	120218156	p < 0.001
elev_2_strm	0.968	p < 0.001	13060515	p < 0.001
niccdcd	0.778	p < 0.001	171951709	p < 0.001
rel_32c	0.666	p < 0.001	201658886	p < 0.001
rng_16c	0.739	p < 0.001	198151593	p < 0.001
std_16c	0.711	p < 0.001	210174923	p < 0.001
tpi_250c	0.950	p < 0.001	56628368	p < 0.001
tpi_cls250c	0.821	p < 0.001	82376462	p < 0.001
tpi_sd250c	0.949	p < 0.001	56750139	p < 0.001
vrf_32c	0.521	p < 0.001	337957802	p < 0.001
random	0.018	p < 0.005	946554781	p = 0.035

Region 6 All - Upland Section 2				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aspect	0.147	p < 0.001	38724379	p < 0.010
aws050	0.286	p < 0.001	29642387	p < 0.001
c_trail_dist	0.215	p < 0.001	39727396	p < 0.001
cd_conf	0.213	p < 0.001	45265237	p < 0.001
cd_drnh	0.159	p < 0.001	31431590	p < 0.001
cd_h1	0.278	p < 0.001	43805158	p < 0.001
cd_h5	0.383	p < 0.001	49997563	p < 0.001
drcdry	0.171	p < 0.001	43898852	p < 0.001
ed_h4	0.220	p < 0.001	43651094	p < 0.001
ed_h6	0.226	p < 0.001	30187528	p < 0.001
elev_2_conf	0.262	p < 0.001	45436394	p < 0.001
elev_2_strm	0.248	p < 0.001	45008633	p < 0.001
flowdir	0.236	p < 0.001	29761969	p < 0.001
niccdcd	0.217	p < 0.001	40387294	p < 0.001
rel_32c	0.254	p < 0.001	44906961	p < 0.001
rng_16c	0.165	p < 0.001	31586338	p < 0.001
slope_deg	0.150	p < 0.001	32177623	p < 0.001
std_8c	0.157	p < 0.001	32319867	p < 0.001
tpi_50c	0.220	p < 0.001	46697641	p < 0.001
tpi_cls250c	0.202	p < 0.001	44582129	p < 0.001
tpi_sd50c	0.220	p < 0.001	46705103	p < 0.001
random	0.045	p = 0.007	36154844	p = 0.068

Region 6 All - Upland Section 3				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
aspect	0.734	p < 0.001	103544808	p < 0.001
aws050	0.634	p < 0.001	477969461	p < 0.001
c_hyd_min	0.754	p < 0.001	28074739	p < 0.001
c_trail_dist	0.876	p < 0.001	40226882	p < 0.001
cd_conf	0.884	p < 0.001	15101508	p < 0.001
cd_h2	0.759	p < 0.001	27104769	p < 0.001
cd_h4	0.877	p < 0.001	15231134	p < 0.001
cd_h5	0.905	p < 0.001	9215764	p < 0.001
cd_h6	0.938	p < 0.001	11986956	p < 0.001
eldrop32c	0.769	p < 0.001	26155219	p < 0.001
elev_2_conf	0.838	p < 0.001	31650117	p < 0.001
elev_2_drainh	0.862	p < 0.001	42557112	p < 0.001
elev_2_strm	0.926	p < 0.001	13967741	p < 0.001
niccdcd	0.680	p < 0.001	96849036	p < 0.001
rel_32c	0.838	p < 0.001	16862066	p < 0.001
rng_8c	0.648	p < 0.001	64106985	p < 0.001
slope_pct	0.734	p < 0.001	63588436	p < 0.001
std_8c	0.660	p < 0.001	66371965	p < 0.001
tpi_250c	0.932	p < 0.001	19537417	p < 0.001
tpi_cls250c	0.907	p < 0.001	20296395	p < 0.001
tpi_sd250c	0.932	p < 0.001	19570721	p < 0.001
twi32c	0.744	p < 0.001	552942136	p < 0.001
vrf_32c	0.747	p < 0.001	36302194	p < 0.001
random	0.013	p = 0.113	299471717	p = 0.415

Region 6 All - Upland Section 4				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
c_hyd_min	0.568	p < 0.001	25835380	p < 0.001
c_trail_dist	0.824	p < 0.001	12507420	p < 0.001
cd_conf	0.739	p < 0.001	16972435	p < 0.001
cd_h2	0.572	p < 0.001	25303439	p < 0.001
cd_h4	0.640	p < 0.001	20654692	p < 0.001
cd_h6	0.854	p < 0.001	14384385	p < 0.001
ed_drnh	0.489	p < 0.001	132897544	p < 0.001
ed_h5	0.821	p < 0.001	9821069	p < 0.001
eldrop32c	0.512	p < 0.001	29027792	p < 0.001
elev_2_conf	0.685	p < 0.001	20914028	p < 0.001
elev_2_drainh	0.671	p < 0.001	26582691	p < 0.001
elev_2_strm	0.830	p < 0.001	8675308	p < 0.001
niccdcd	0.516	p < 0.001	45733336	p < 0.001
rel_32c	0.820	p < 0.001	9725431	p < 0.001
slpvr_32c	0.750	p < 0.001	165156011	p < 0.001
tpi_250c	0.826	p < 0.001	16160715	p < 0.001
tpi_cls250c	0.802	p < 0.001	17381119	p < 0.001
tpi_sd250c	0.826	p < 0.001	16138911	p < 0.001
tri_32c	0.747	p < 0.001	163783262	p < 0.001
random	0.019	p = 0.158	90755765	p = 0.749

Region 6 All - Upland Section 5				
Predictor	**Mean D**	**Mean KS p**	**Mean U**	**Mean MW p**
c_hyd_min	0.756	p < 0.001	555258	p < 0.001
cd_drnh	0.584	p < 0.001	681092	p < 0.001
cd_h2	0.761	p < 0.001	516741	p < 0.001
cd_h4	0.533	p < 0.001	1428897	p < 0.001
eldrop16c	0.518	p < 0.001	1056390	p < 0.001
rng_32c	0.596	p < 0.001	991614	p < 0.001
std_32c	0.615	p < 0.001	761980	p < 0.001
tri_10c	0.490	p < 0.001	1206220	p < 0.001
aws050	0.471	p < 0.001	2639611	p < 0.001
ed_conf	0.456	p < 0.001	2980950	p < 0.001
aspect	0.448	p < 0.001	1148023	p < 0.001
elev_2_drainh	0.447	p < 0.001	2617773	p < 0.001
slpvr_10c	0.447	p < 0.001	1446204	p < 0.005
drcdry	0.432	p < 0.001	1129496	p < 0.001
flowdir	0.428	p < 0.001	1084858	p < 0.001
ed_h6	0.421	p < 0.001	1618916	p < 0.005
ed_h5	0.419	p < 0.001	1612704	p < 0.005
vrf_32c	0.414	p < 0.001	1649425	p < 0.010
elev_2_strm	0.411	p < 0.001	2371781	p < 0.005
tpi_sd250c	0.396	p < 0.001	2675840	p < 0.001
tpi_250c	0.396	p < 0.001	2674878	p < 0.001
slope_deg	0.389	p < 0.001	1211041	p < 0.001
elev_2_conf	0.369	p < 0.001	1828777	p < 0.200
random	0.094	p = 0.477	2159892	p = 0.219

APPENDIX E

VARIABLE IMPORTANCE

FOR SELECTED RF MODELS

WITHIN REGIONS 4, 5, AND 6

Chart 1. Region 4/5 East - Riverine Section 1

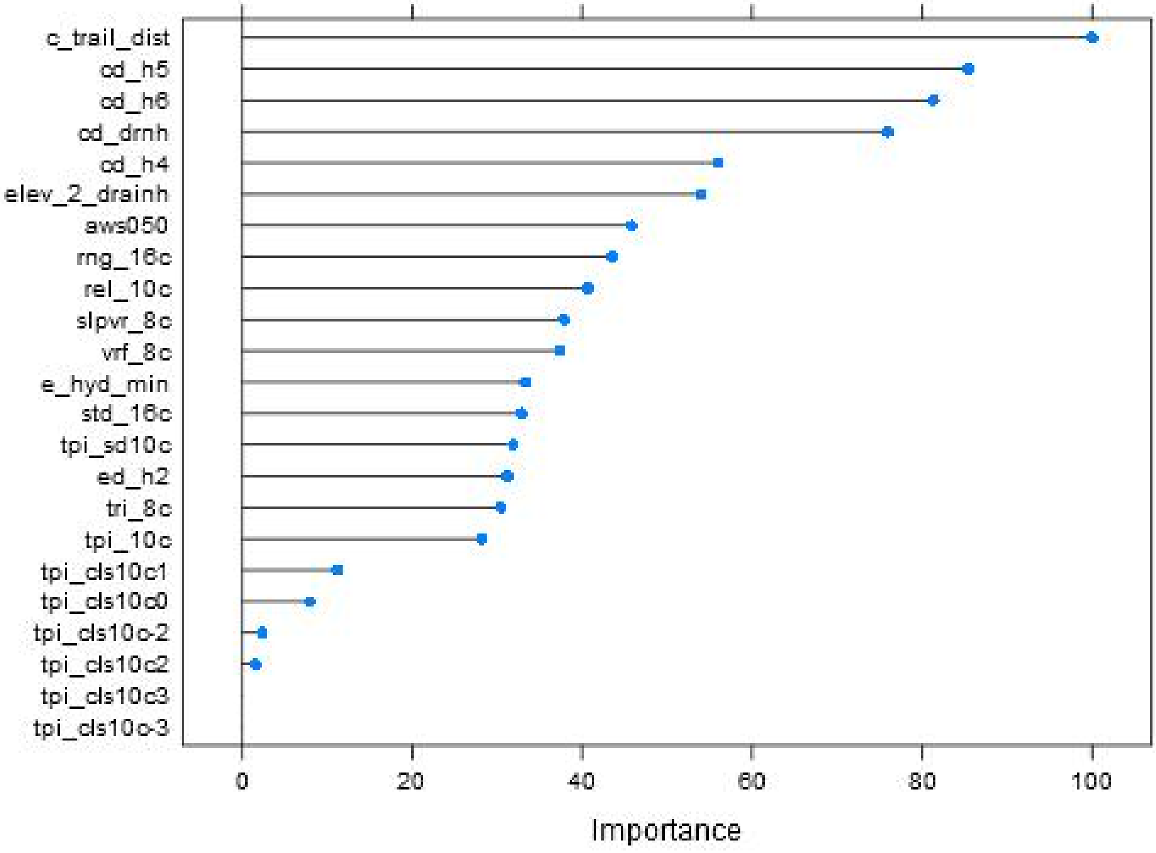


Chart 2. Region 4/5 East - Riverine Section 4

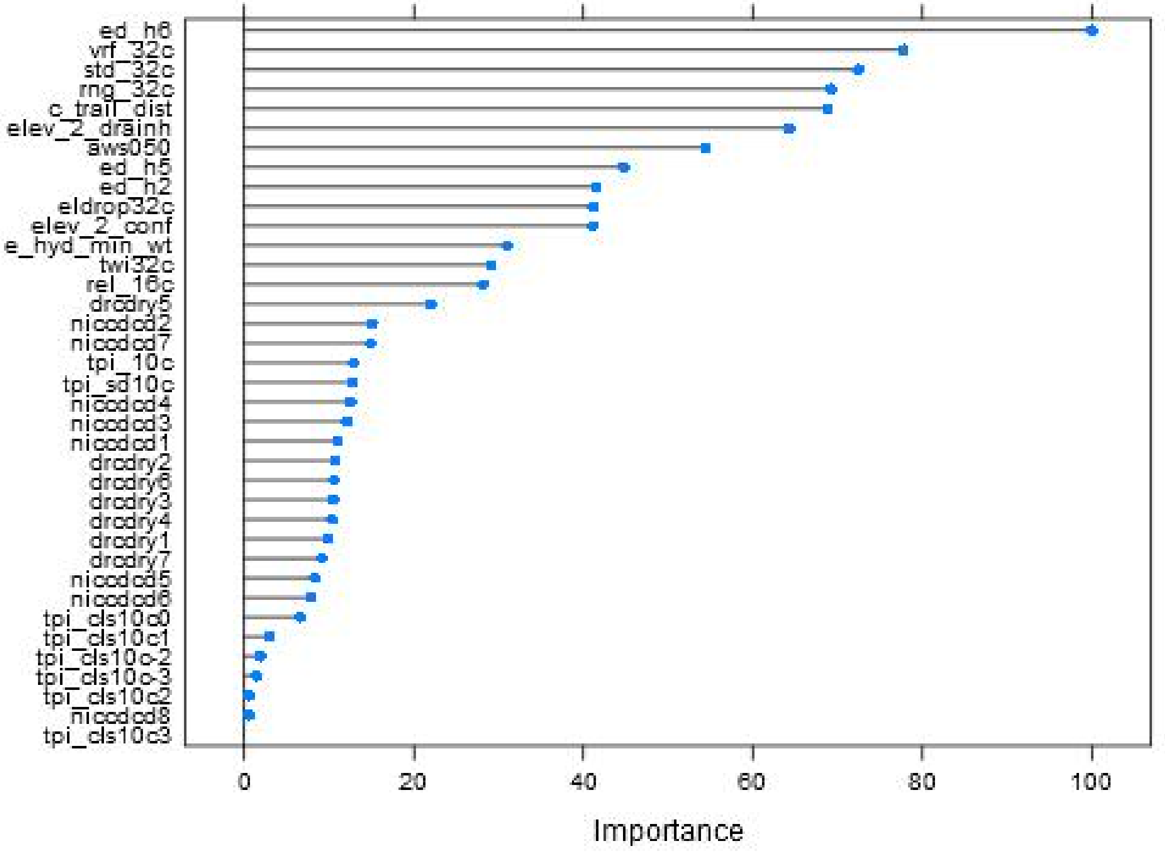


Chart 3. Region 4/5 East - Upland Section 4

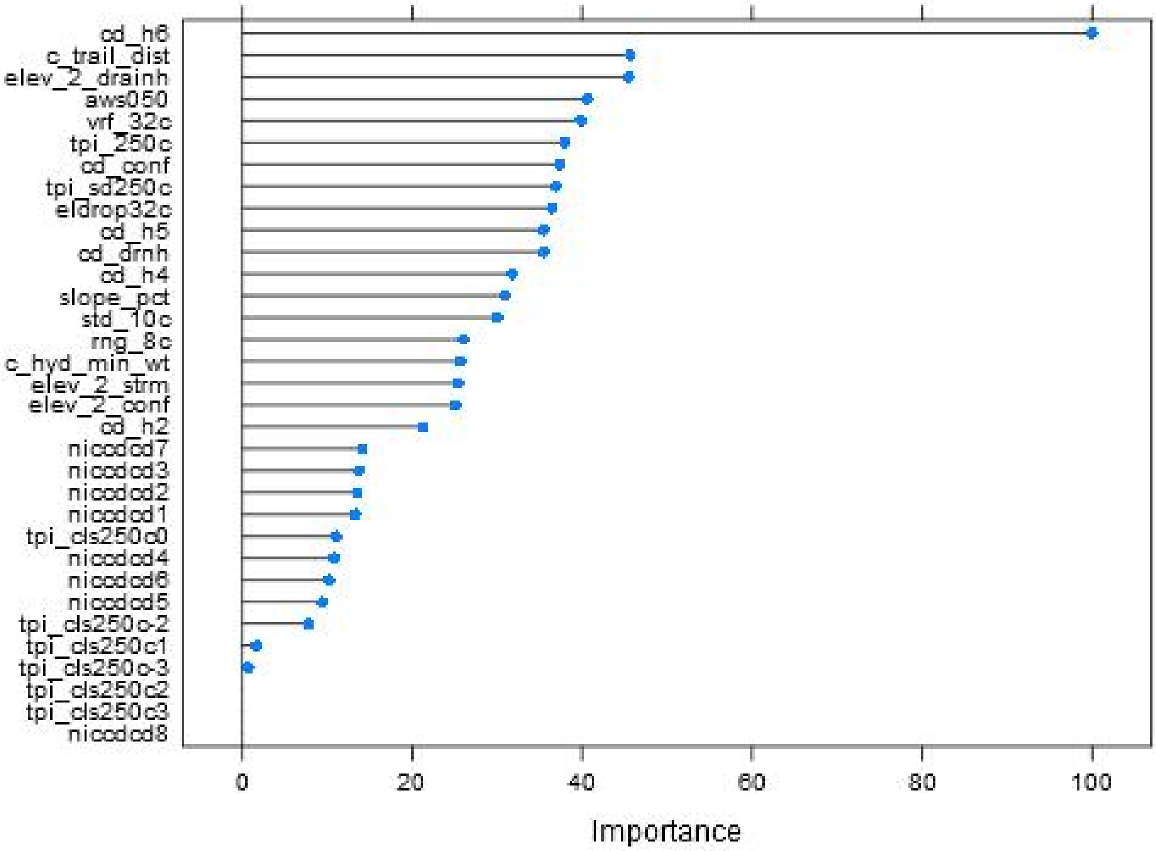


Chart 4. Region 4/5 East - Upland Section 5

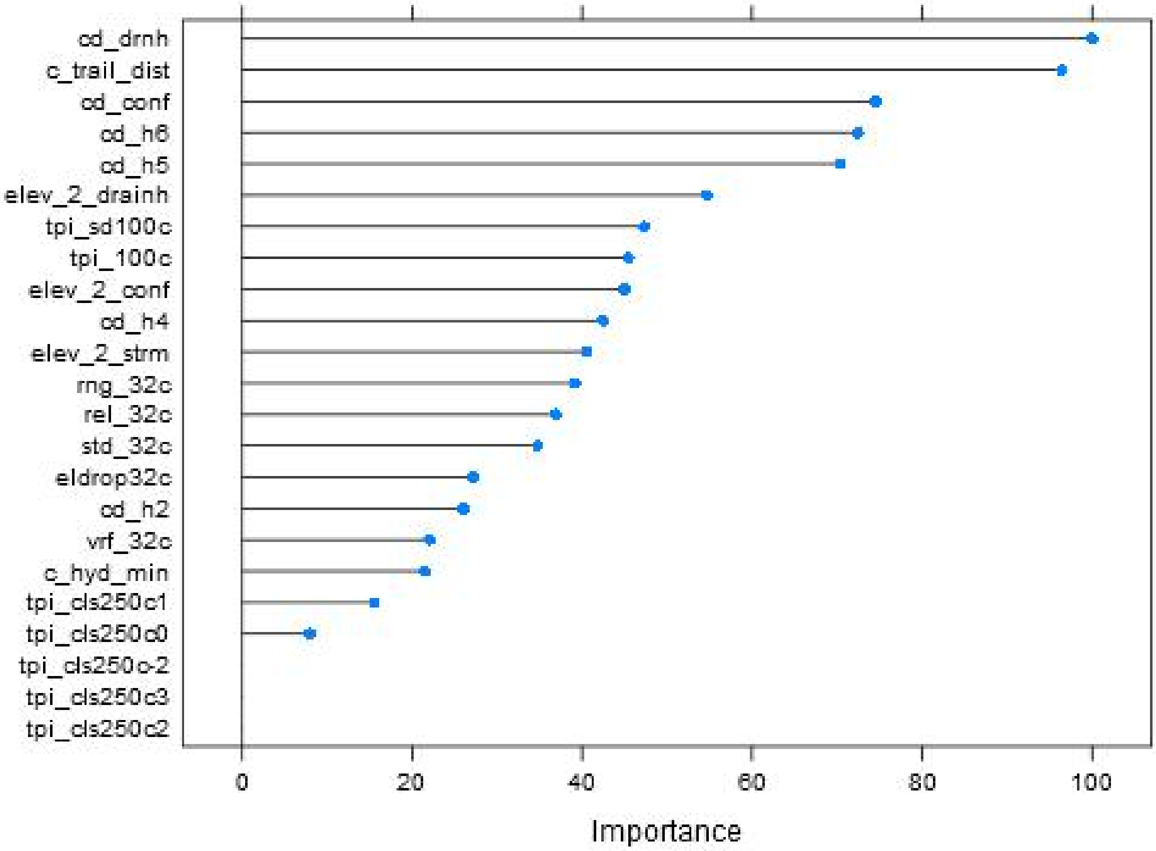


Chart 5. Region 4/5 East - Upland Section 7

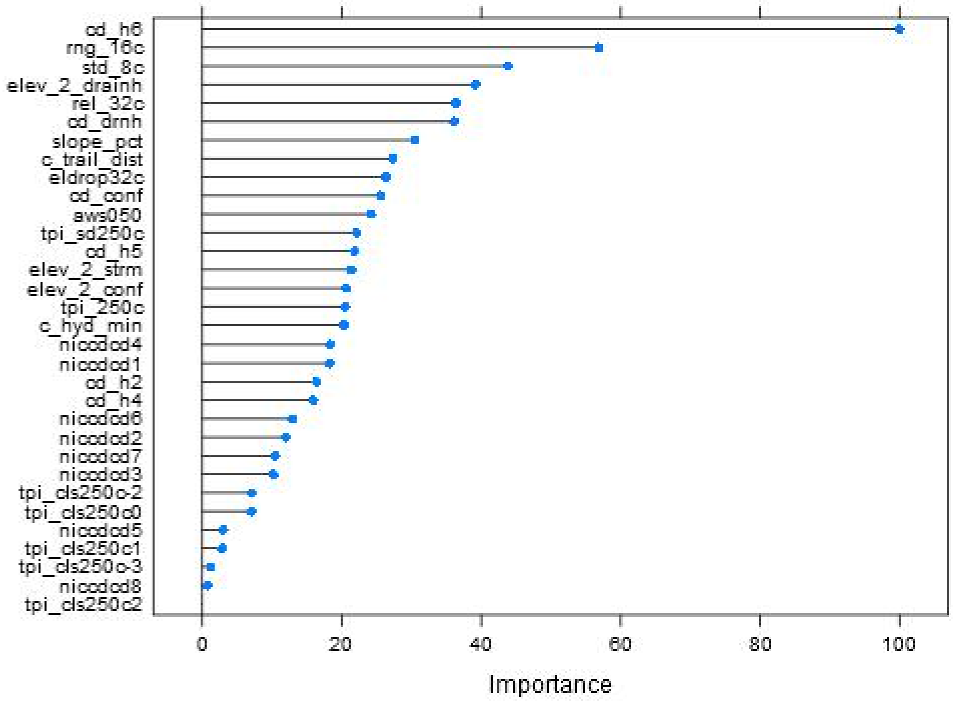


Chart 6. Region 4/5 West - Riverine Section 1

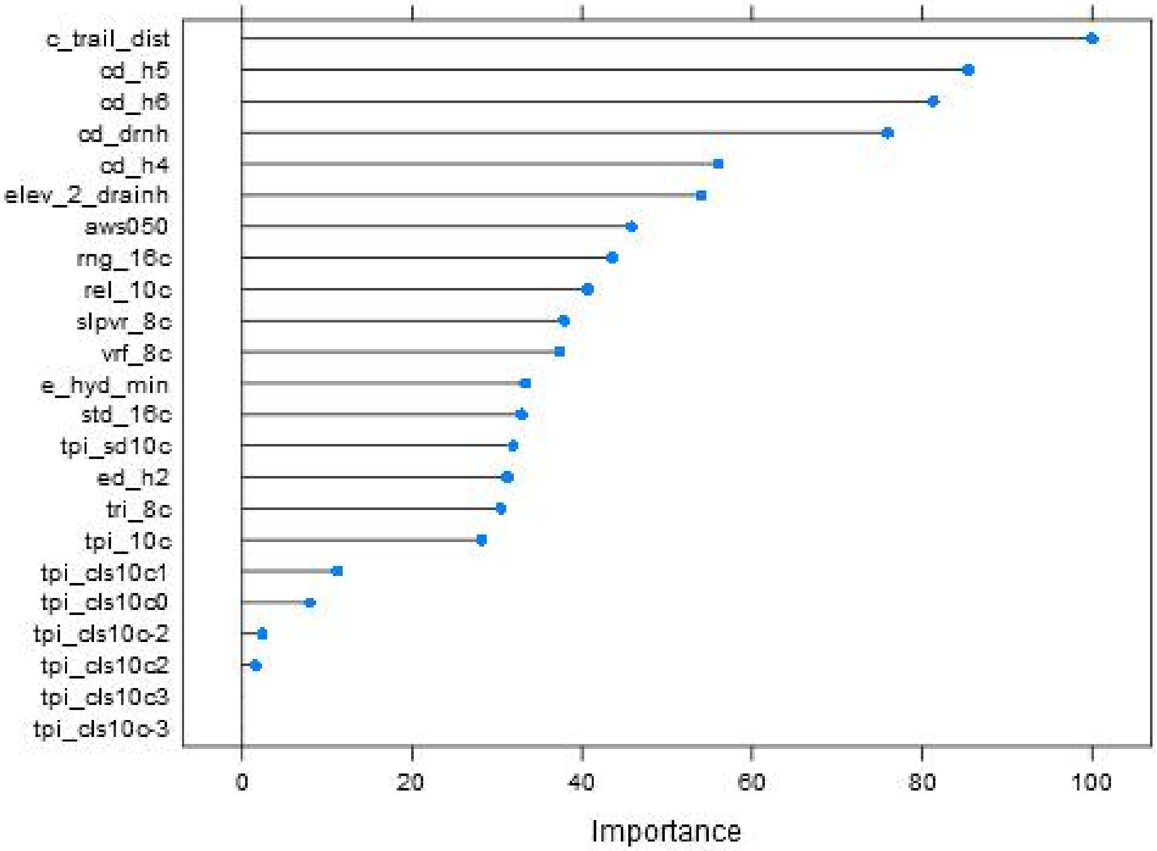


Chart 7. Region 4/5 West - Riverine Section 3

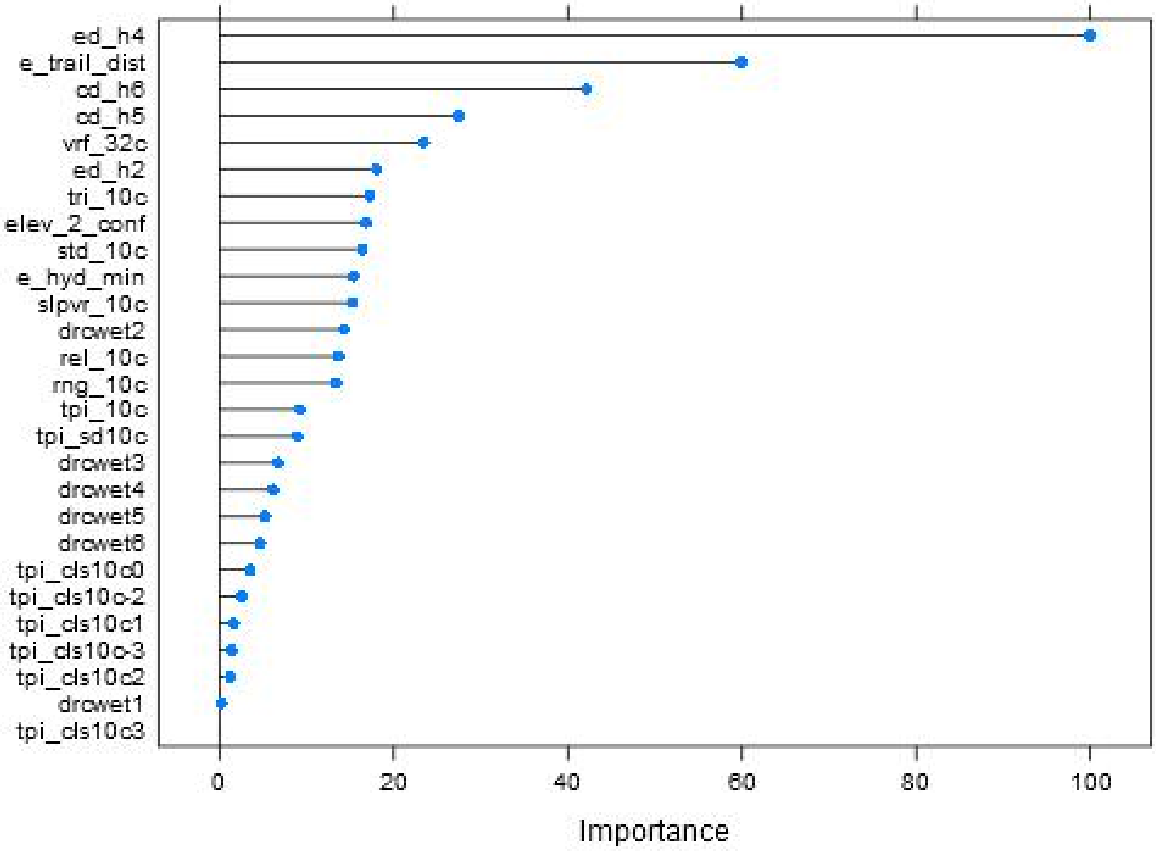


Chart 8. Region 4/5 West - Riverine Section 4

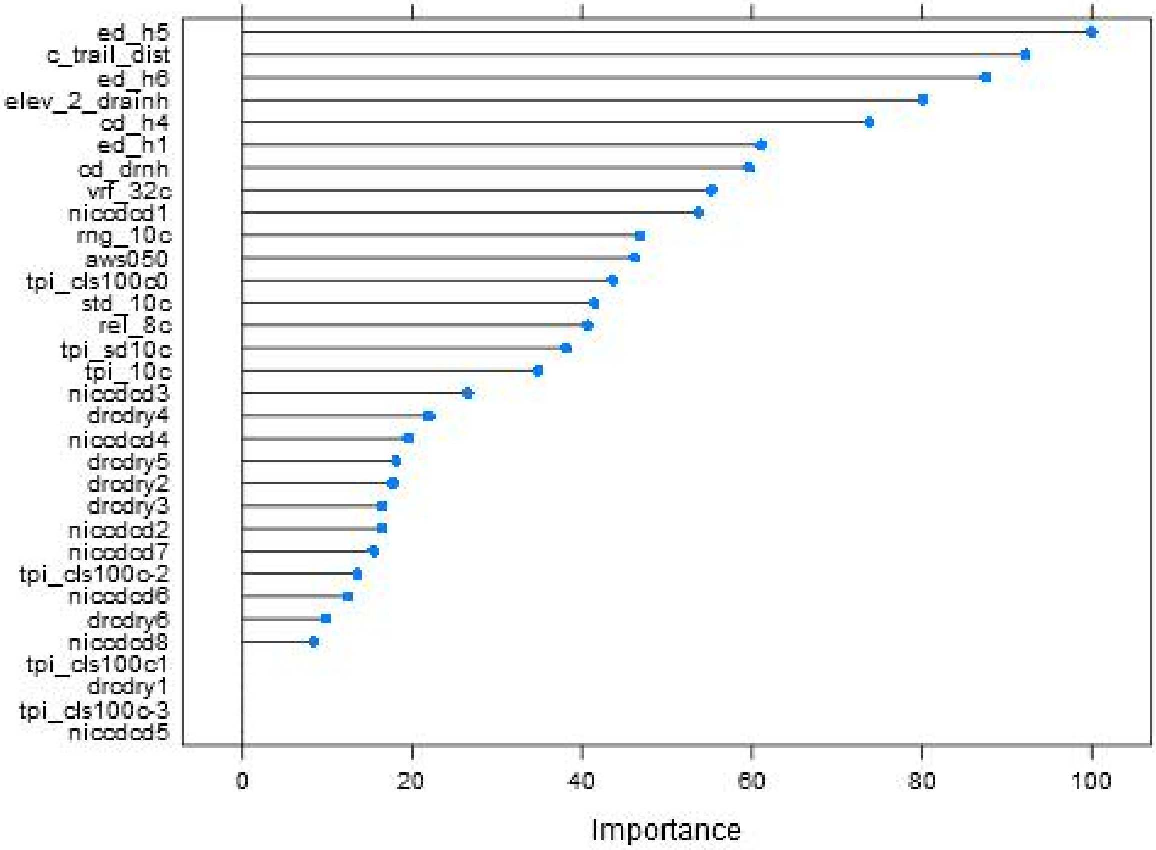


Chart 9. Region 4/5 West - Upland Section 2

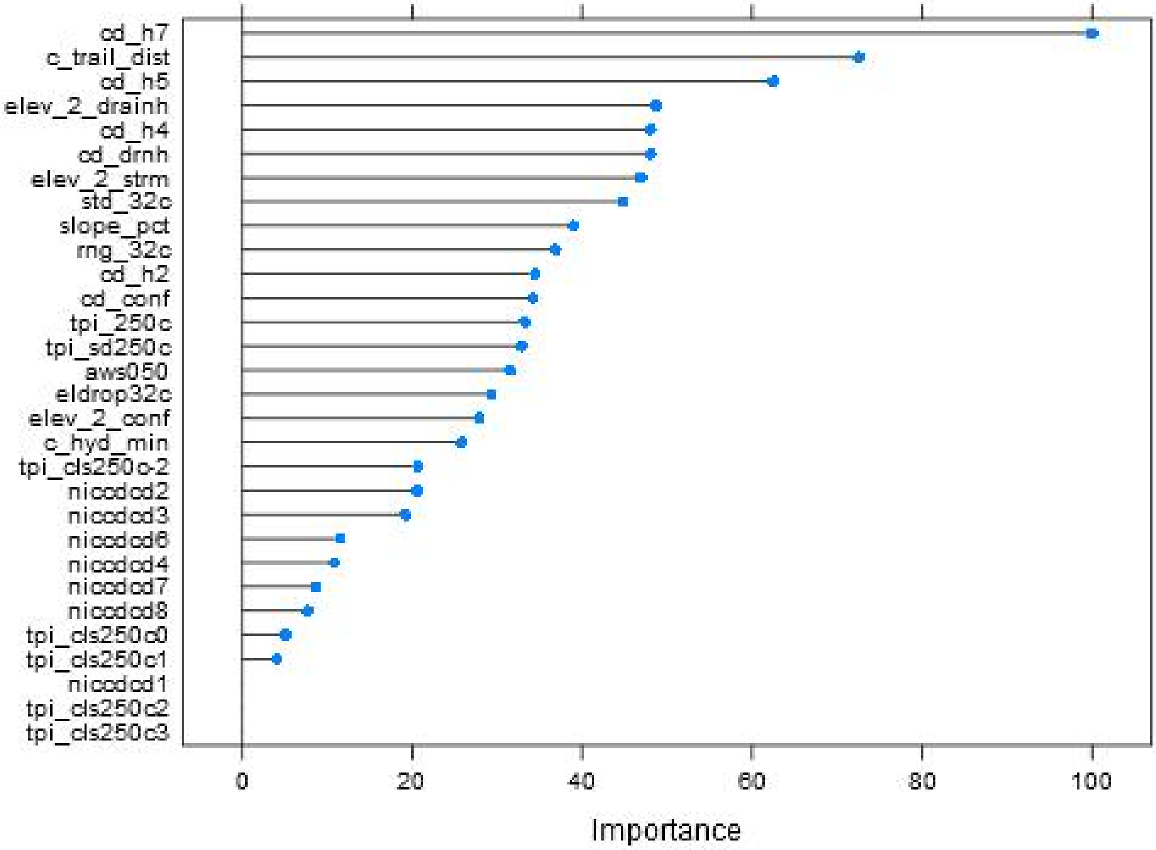


Chart 10. Region 4/5 West - Upland Section 3

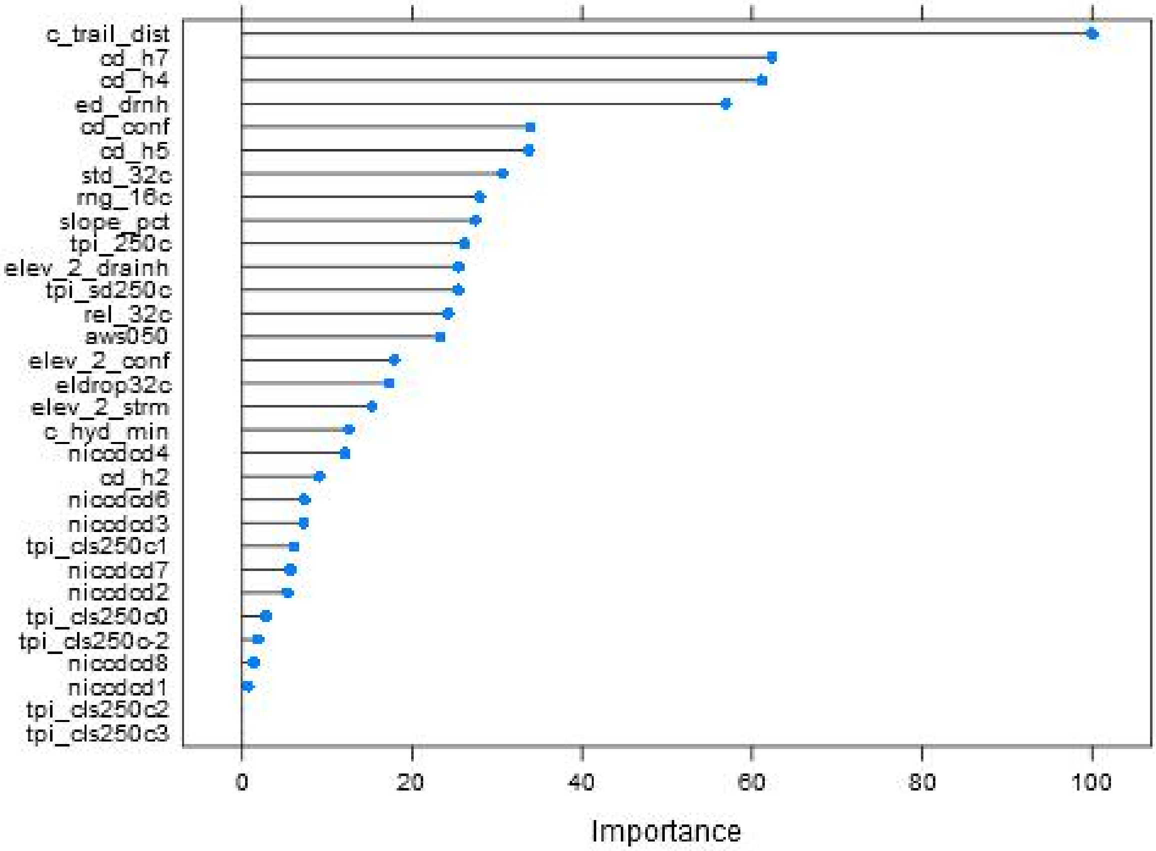


Chart 11. Region 4/5 West - Upland Section 3

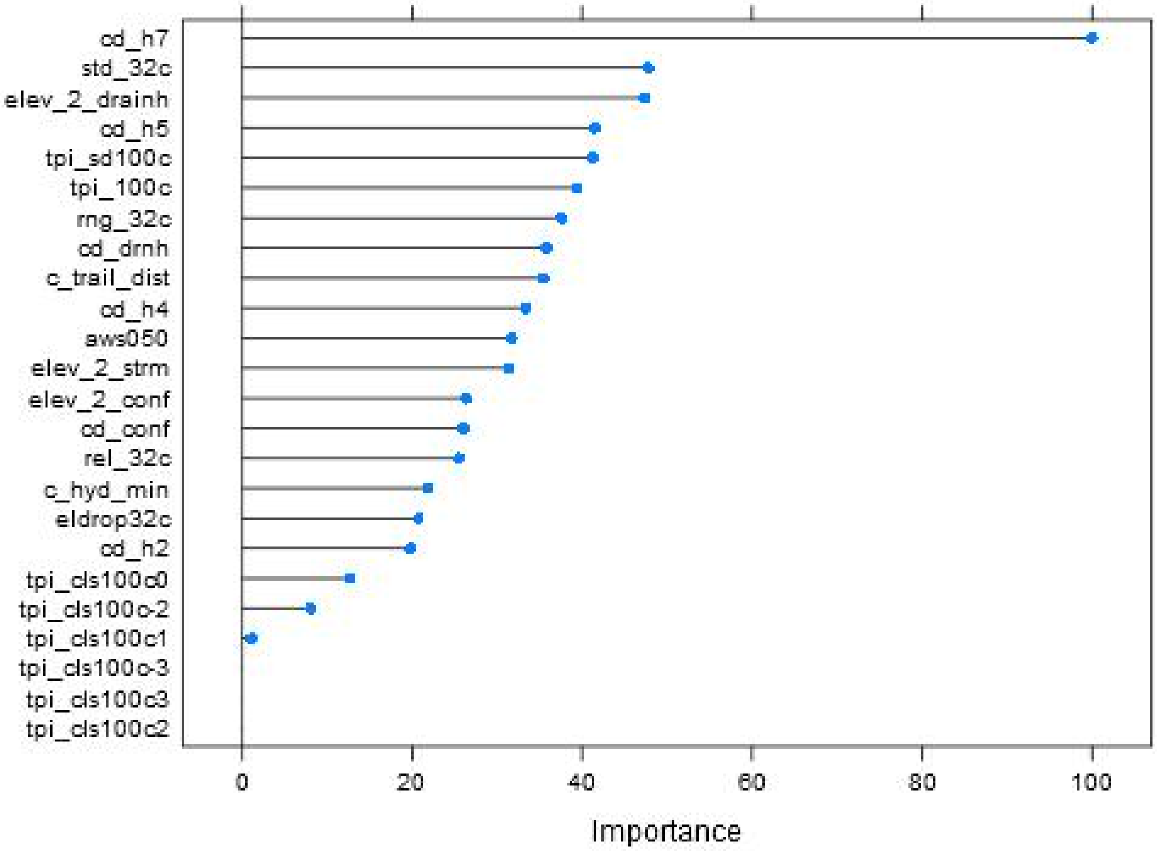


Chart 12. Region 4/5 West - Upland Section 5

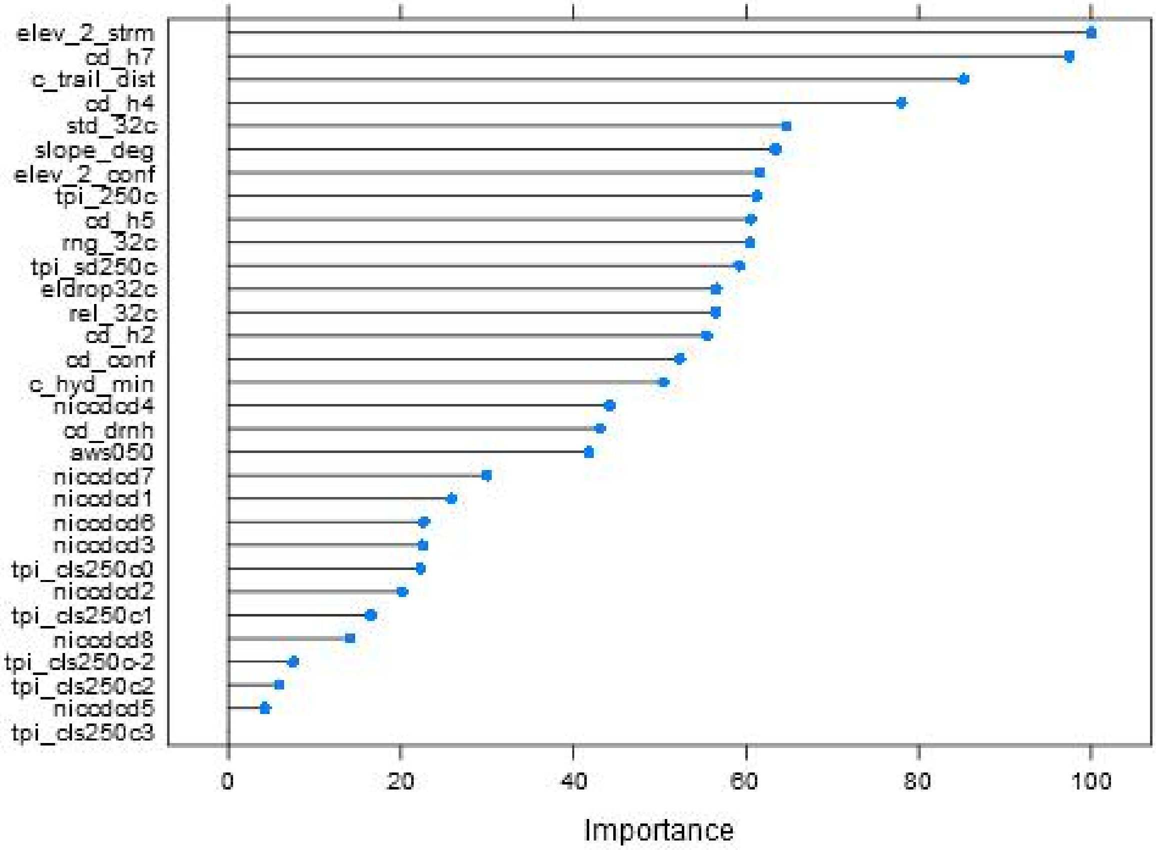
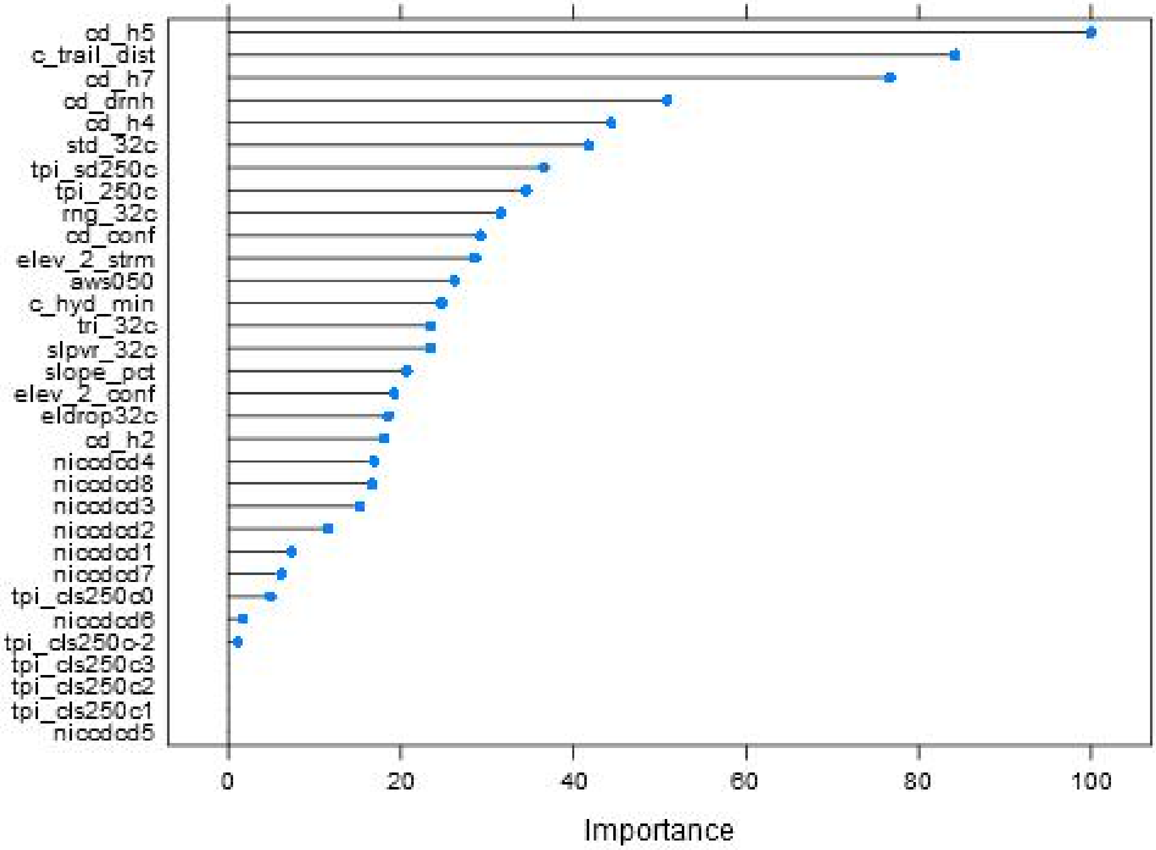


Chart 13. Region 4/5 West - Upland Section 6



APPENDIX F

POTENTIAL THRESHOLDS

FOR EACH OF 36 MODELS

WITHIN REGIONS 4, 5, AND 6

Chart 1. Region 4/5 East - Riverine Section 1

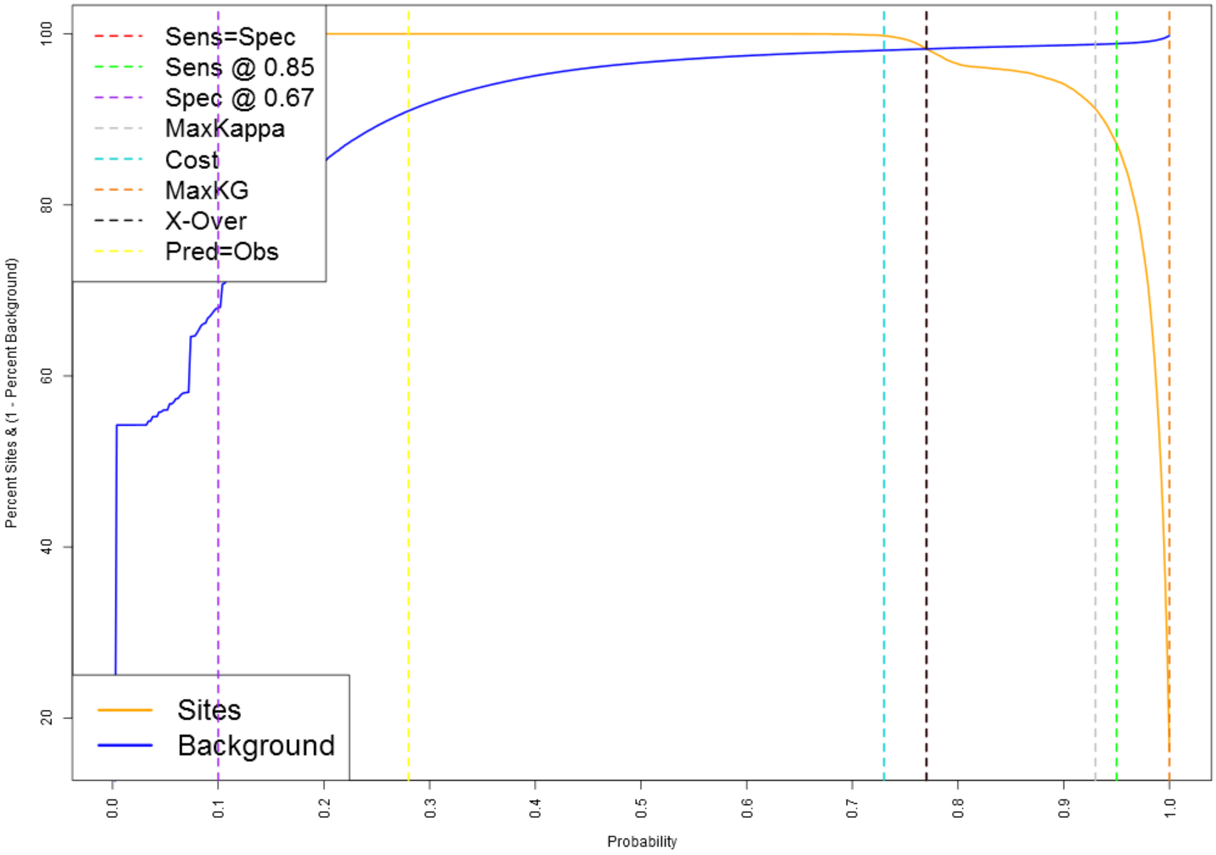


Chart 2. Region 4/5 East -Riverine Section 2

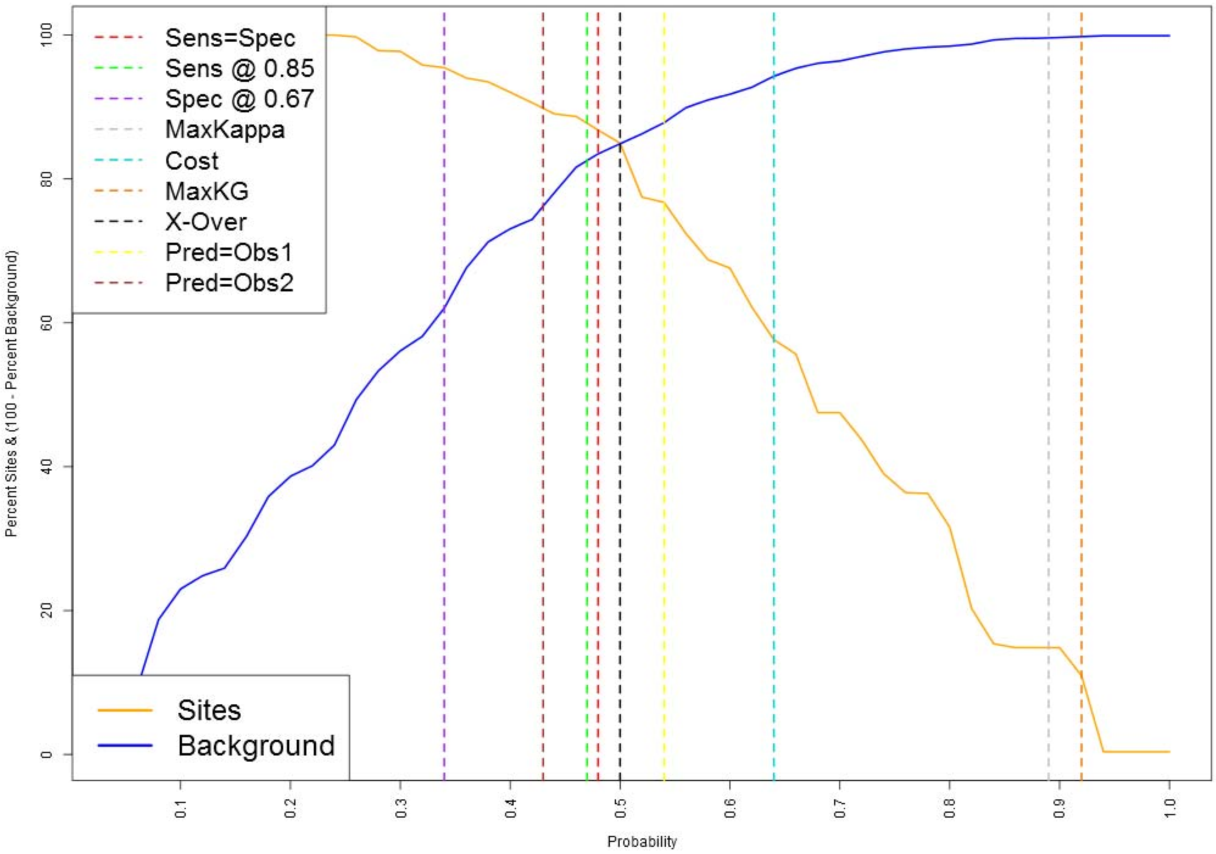


Chart 3. Region 4/5 East -Riverine Section 3

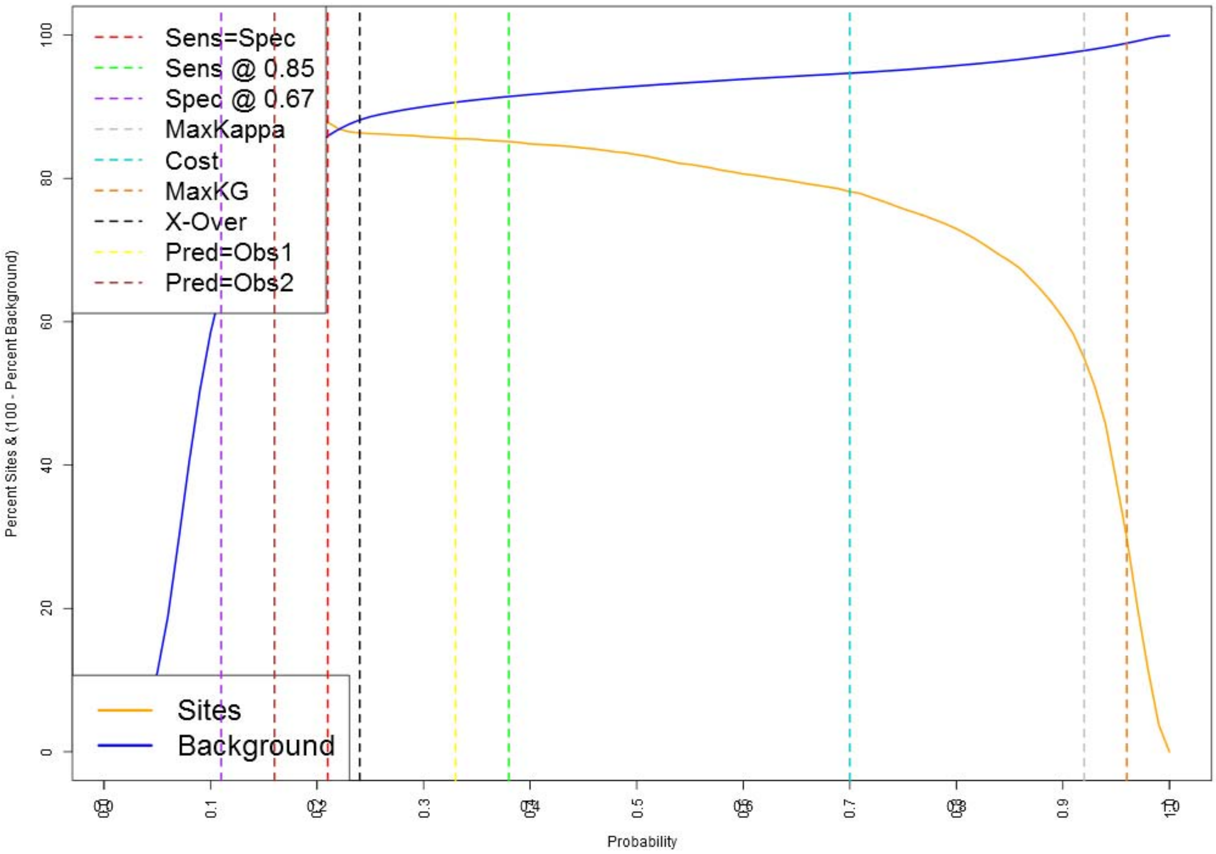


Chart 4. Region 4/5 East -Riverine Section 4

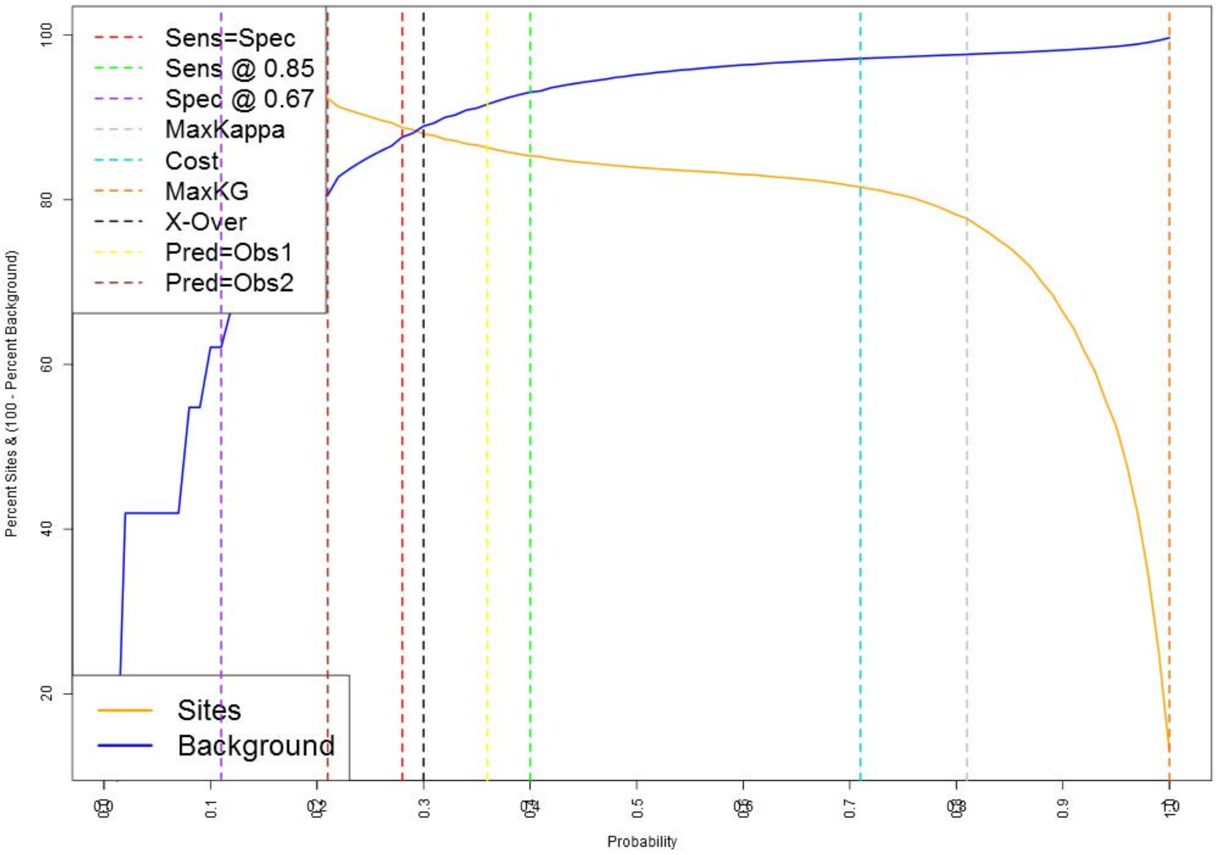


Chart 5. Region 4/5 East -Riverine Section 5

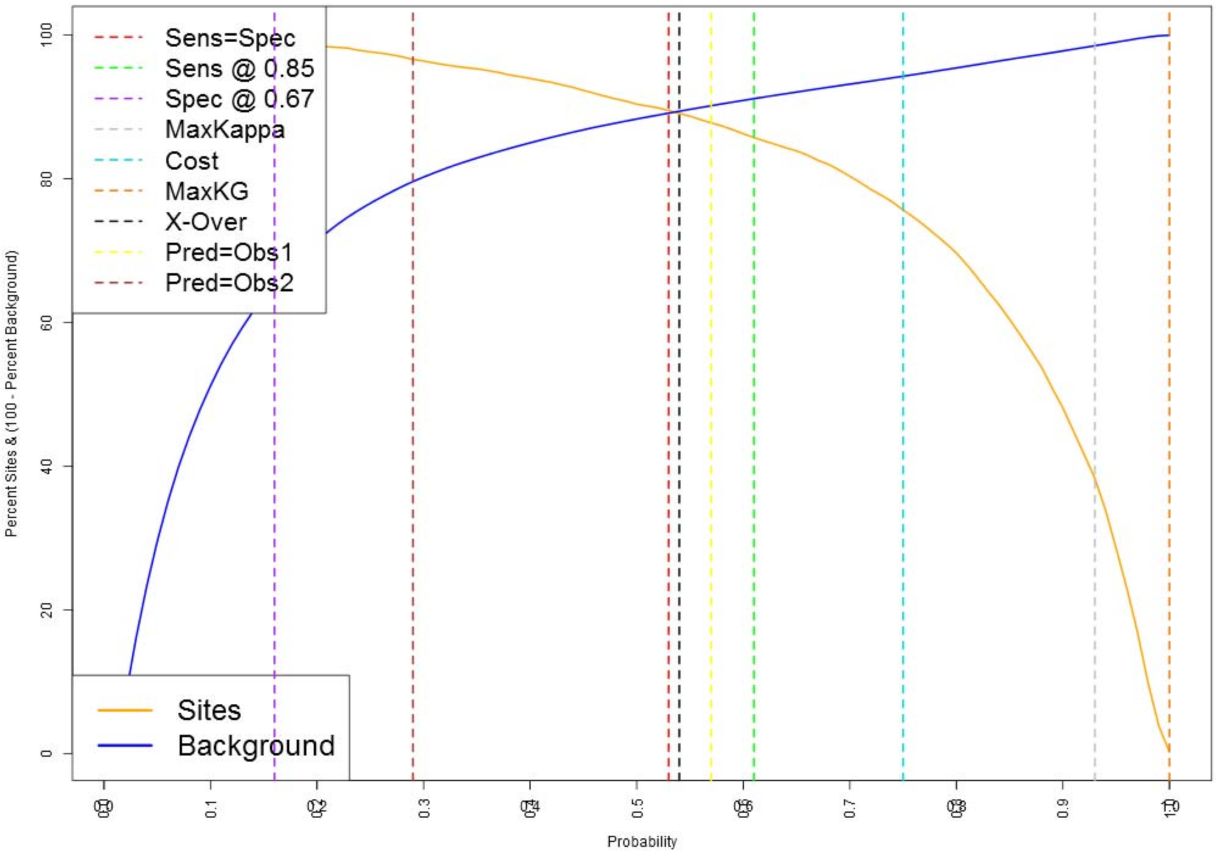


Chart 6. Region 4/5 East -Riverine Section 6

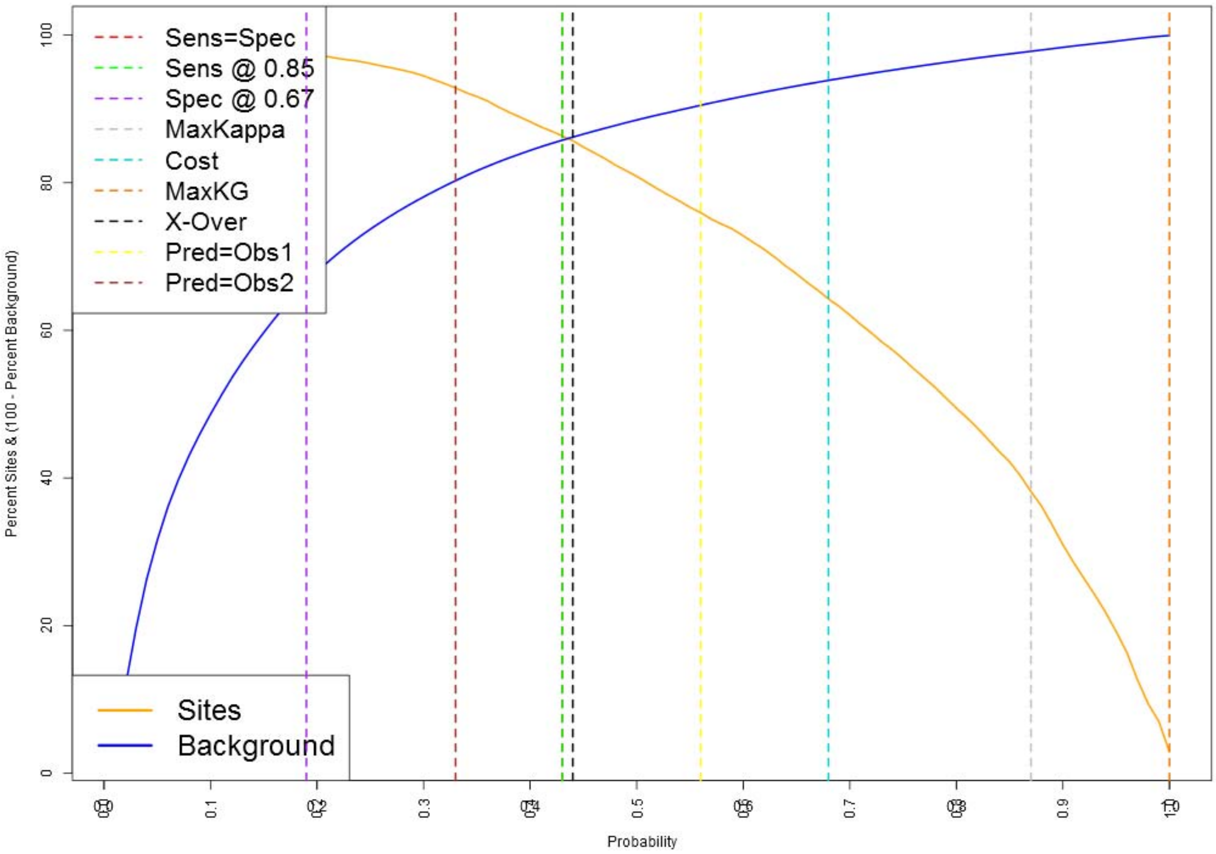


Chart 7. Region 4/5 East -Riverine Section 7

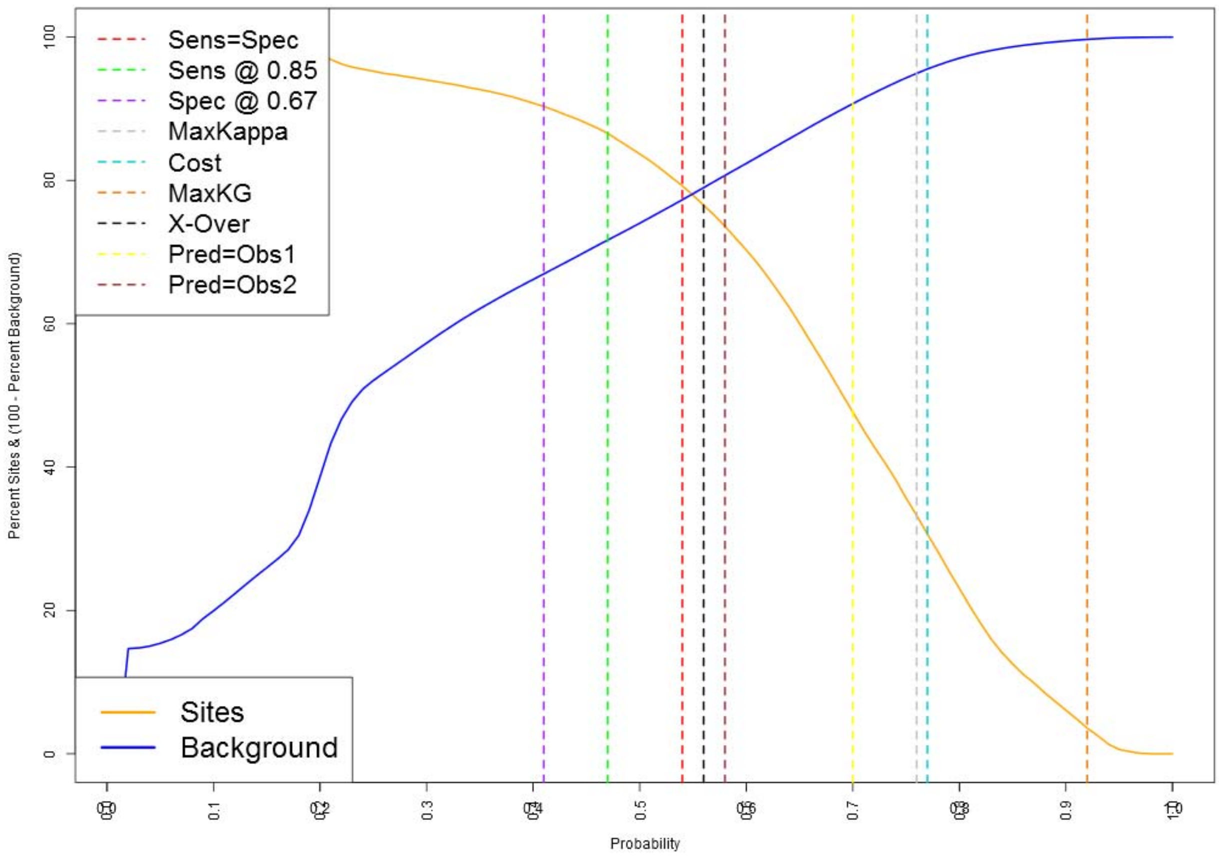


Chart 8. Region 4/5 East –Upland Section 1

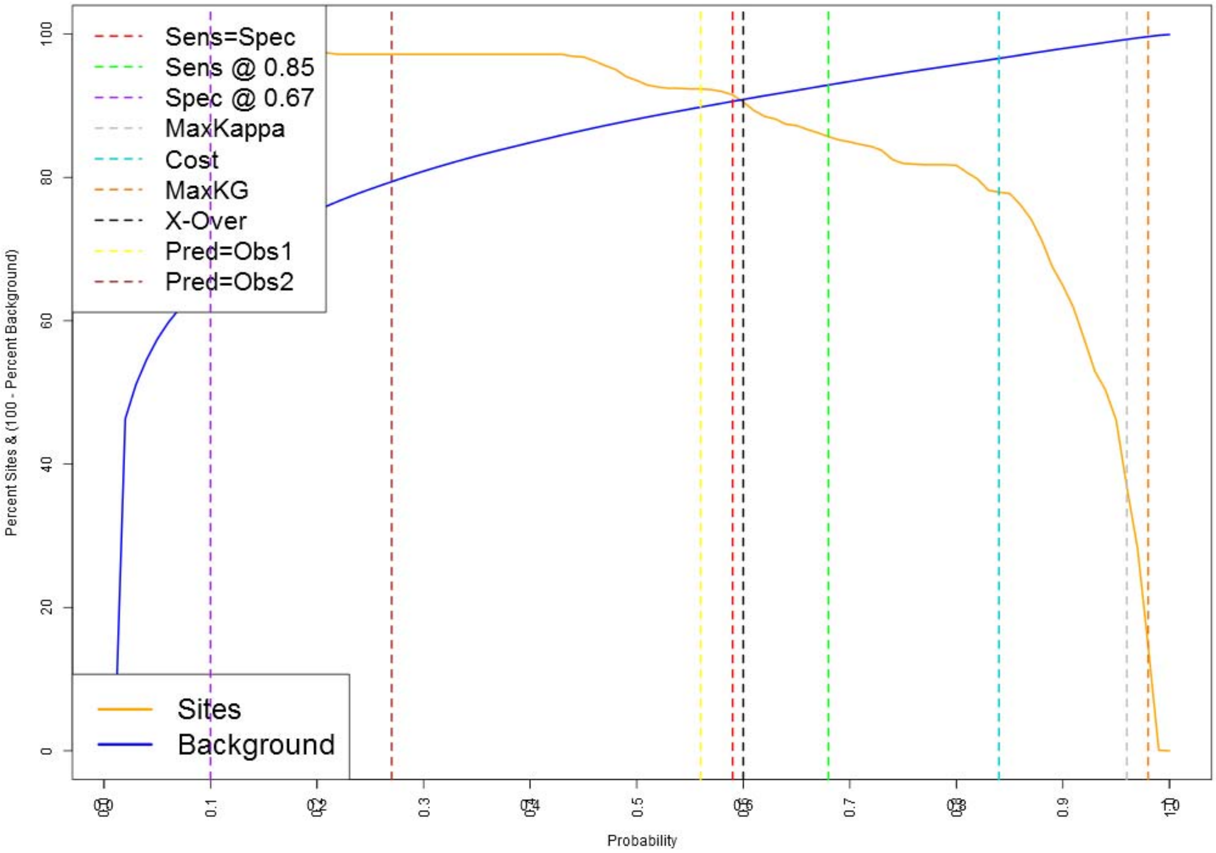


Chart 9. Region 4/5 East –Upland Section 2

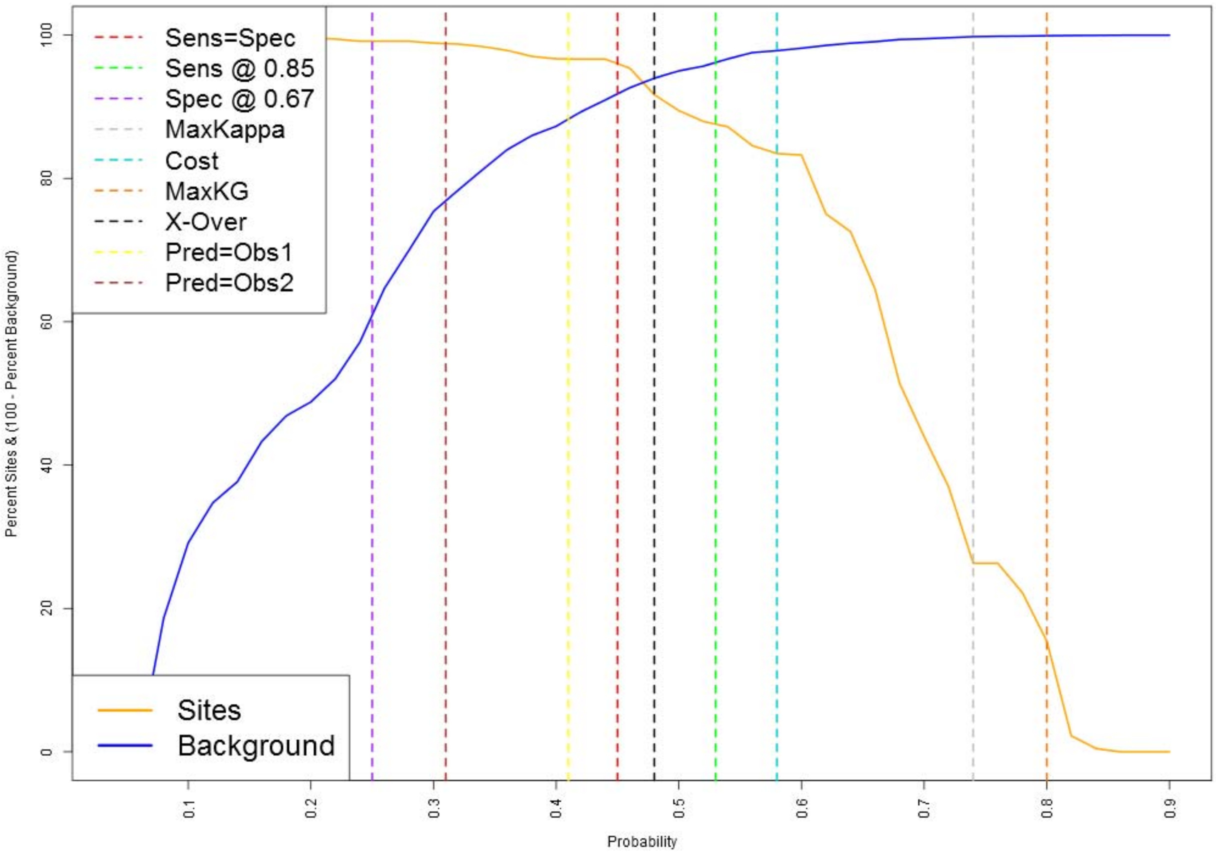


Chart 10. Region 4/5 East –Upland Section 3

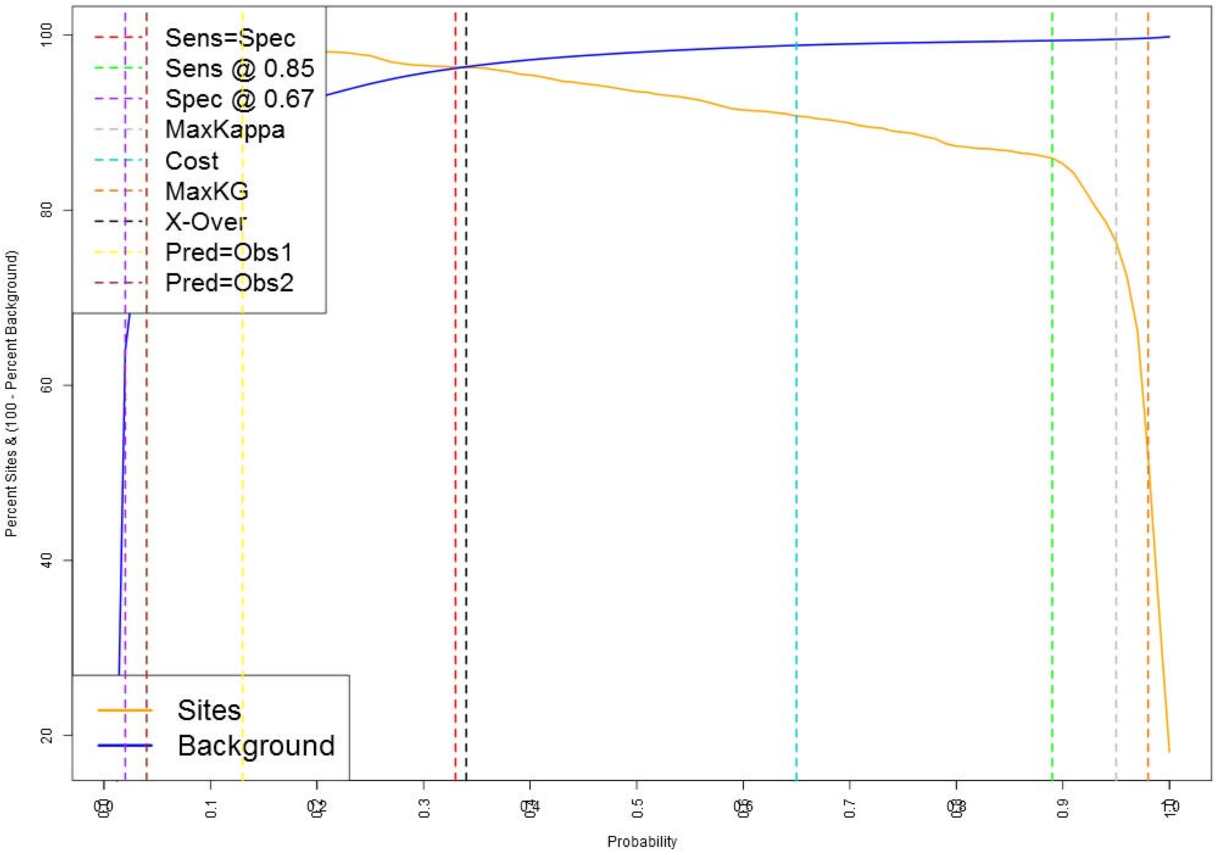


Chart 11. Region 4/5 East –Upland Section 4

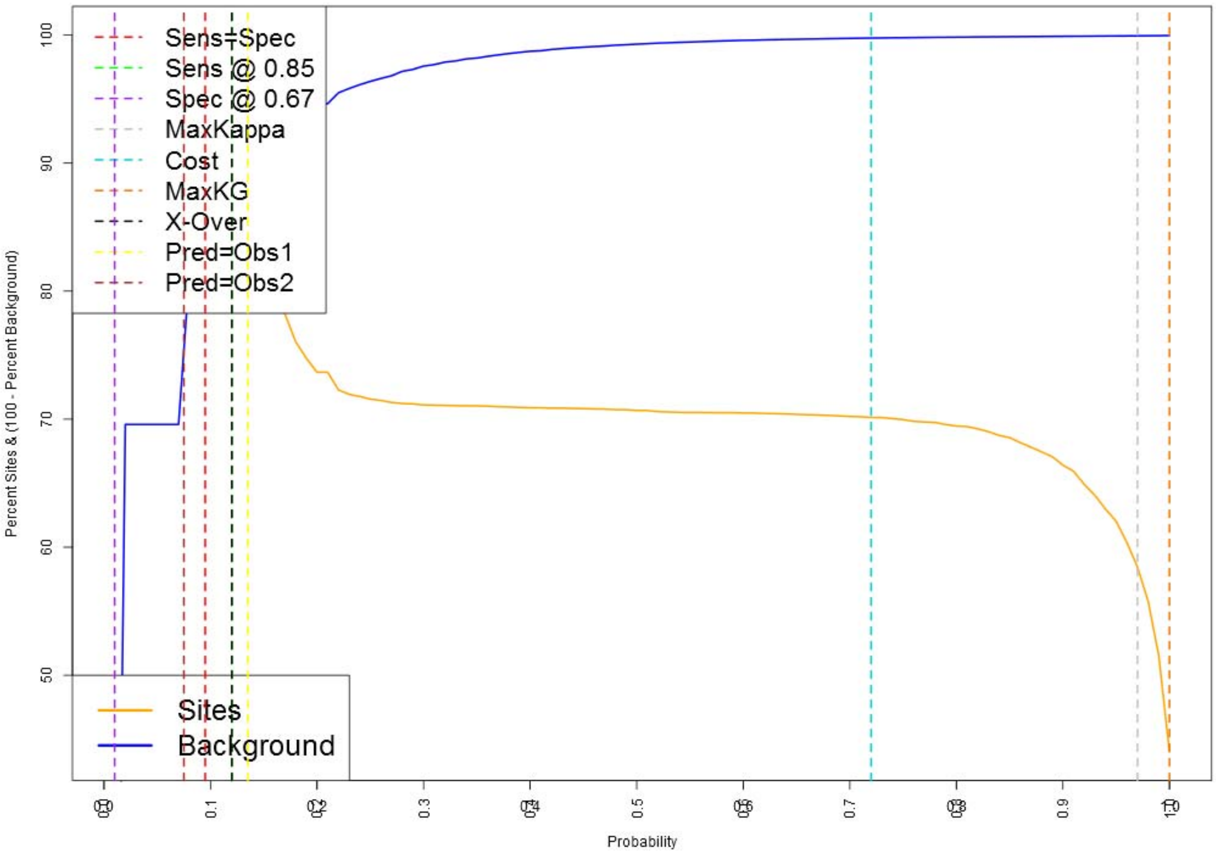


Chart 12. Region 4/5 East –Upland Section 5

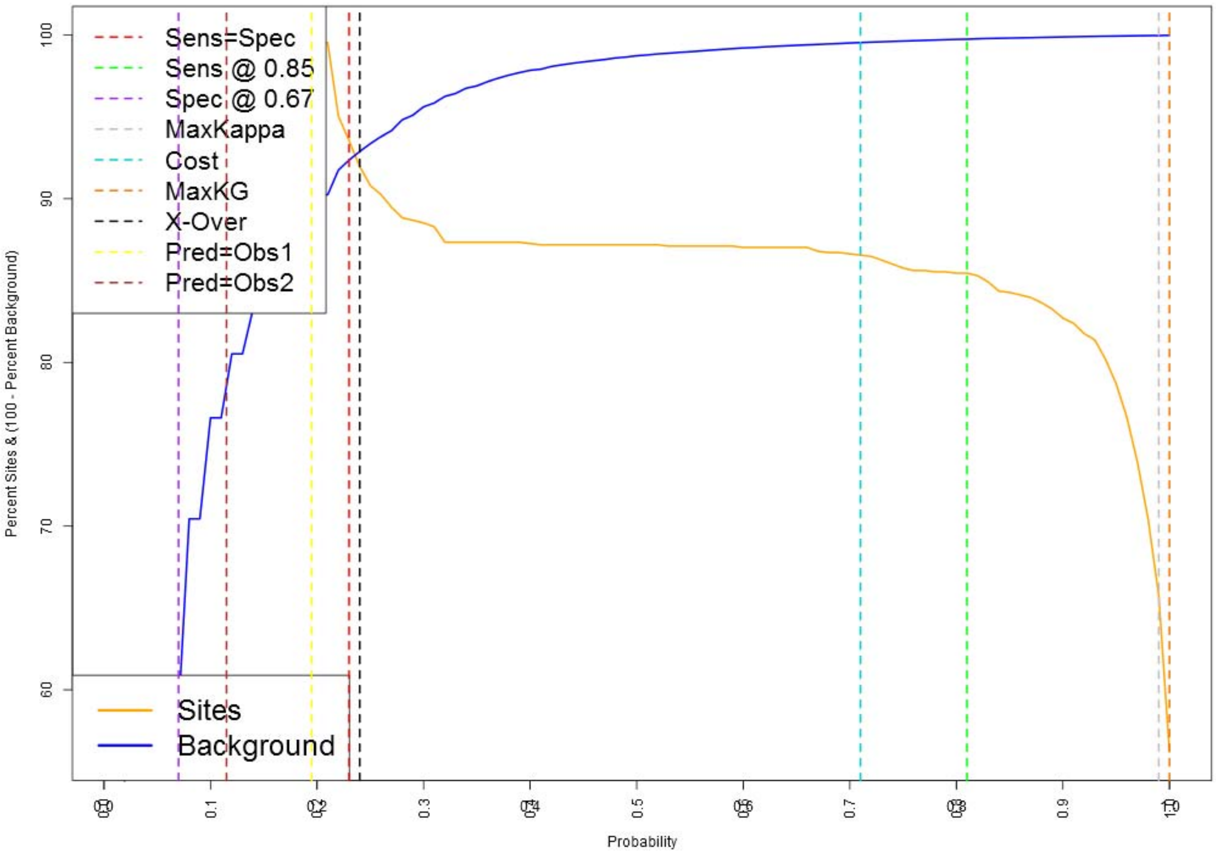


Chart 13. Region 4/5 East –Upland Section 6

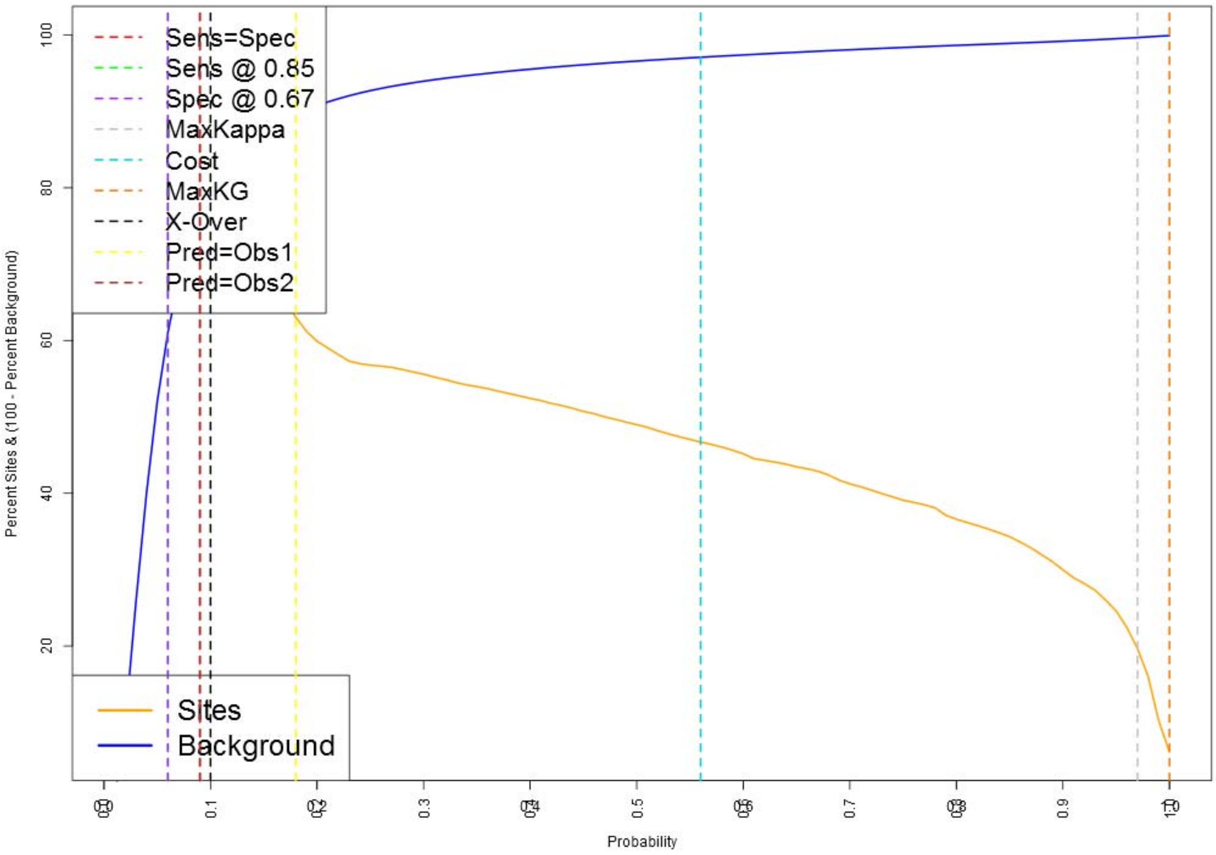


Chart 14. Region 4/5 East –Upland Section 7

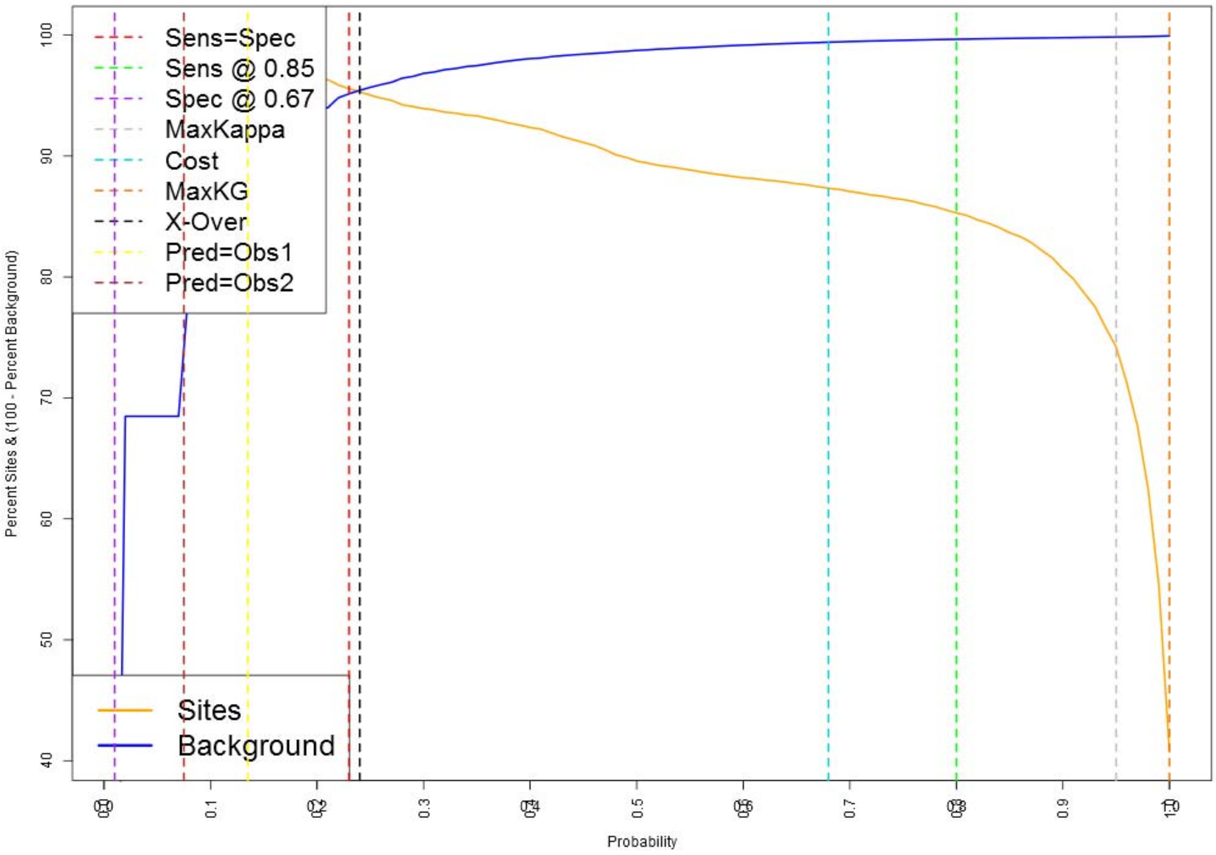


Chart 15. Region 4/5 West – Riverine Section 1

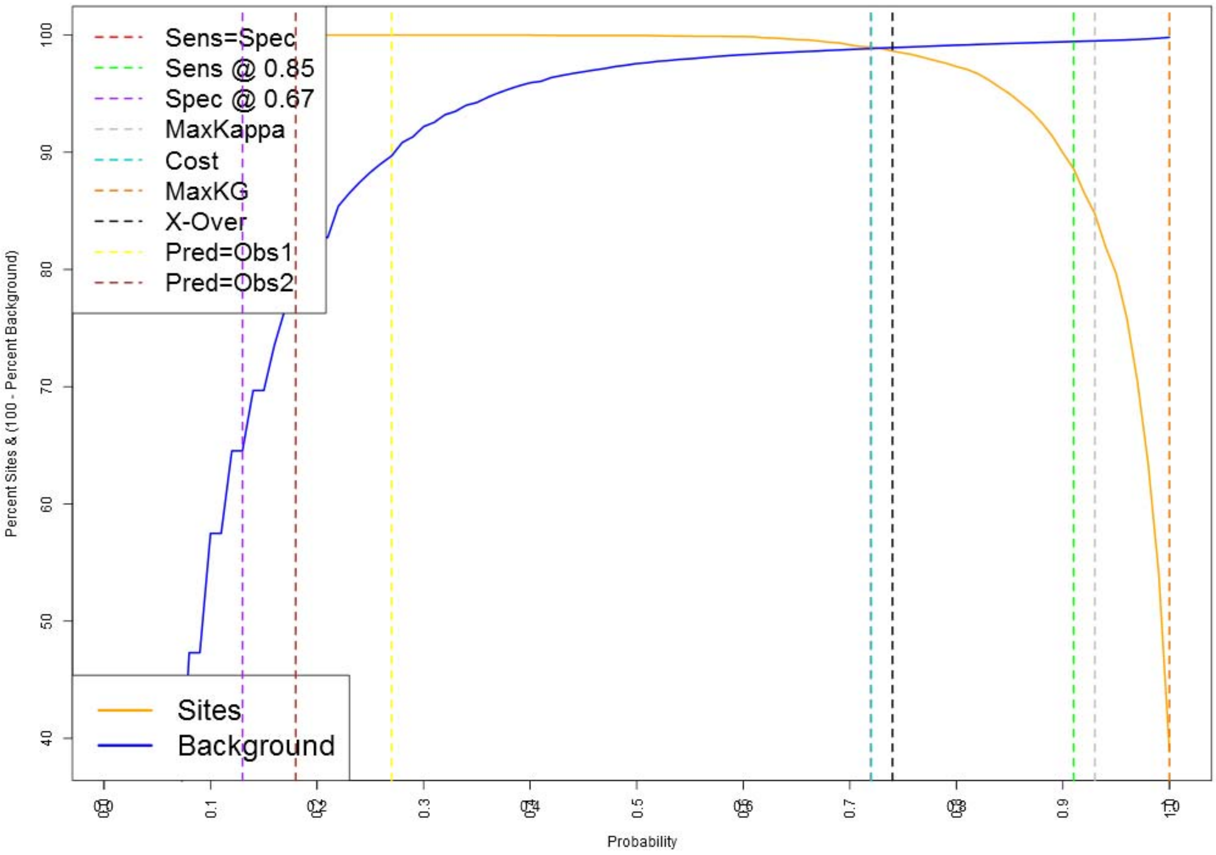


Chart 16. Region 4/5 West – Riverine Section 2

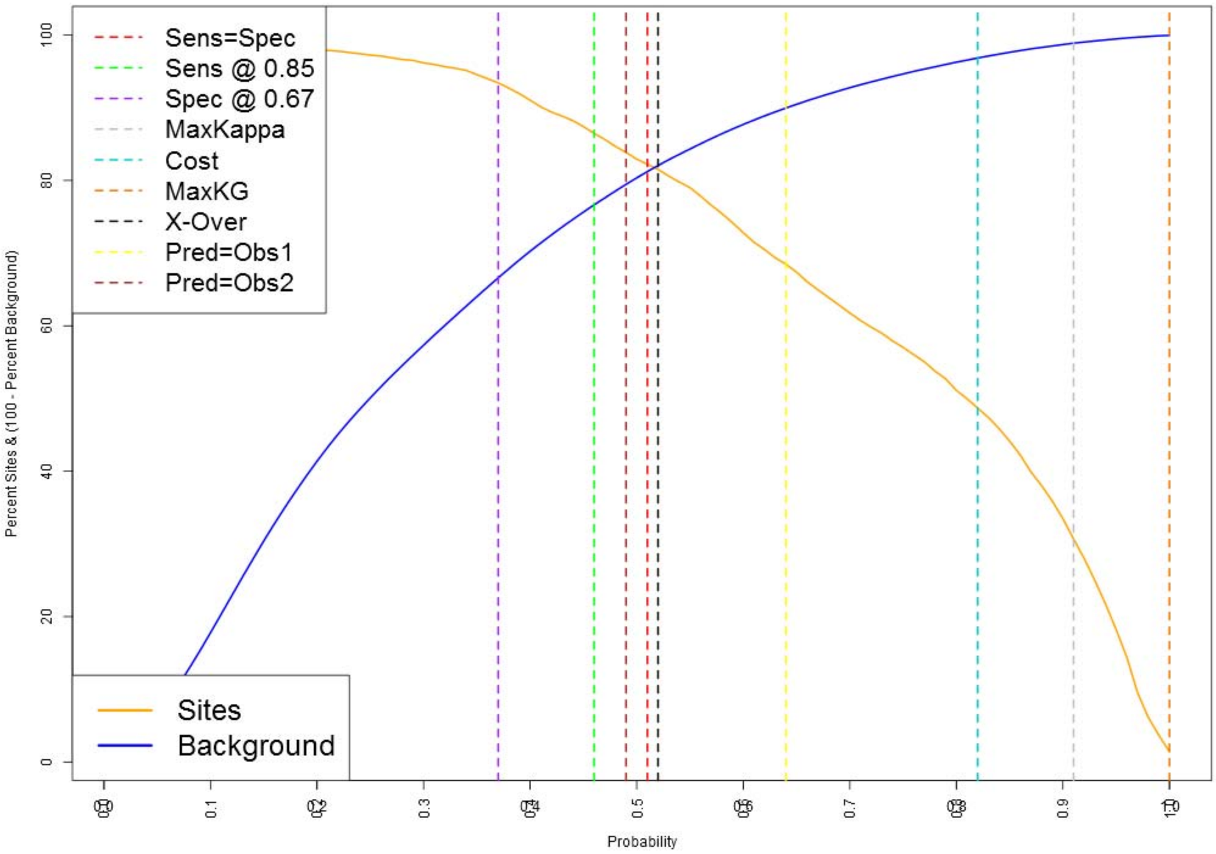


Chart 17. Region 4/5 West – Riverine Section 3

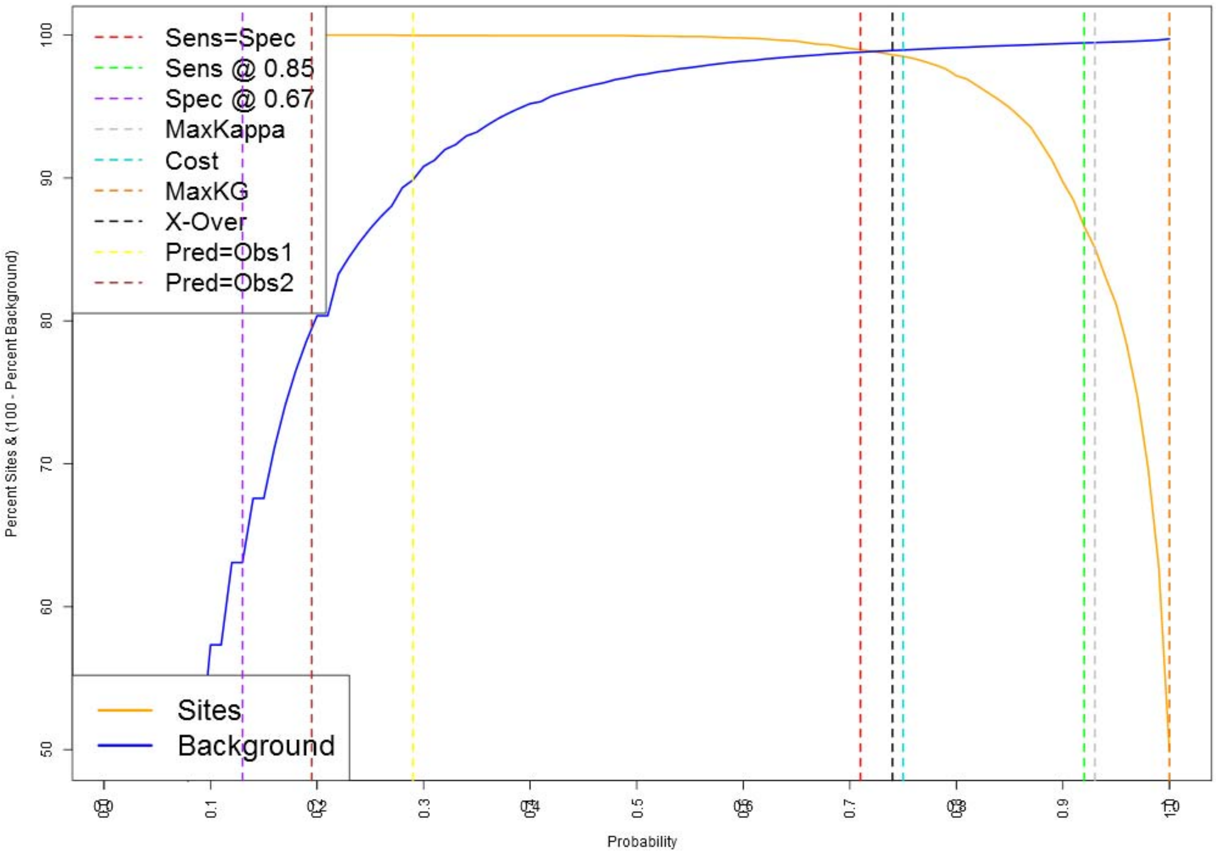


Chart 18. Region 4/5 West – Riverine Section 4

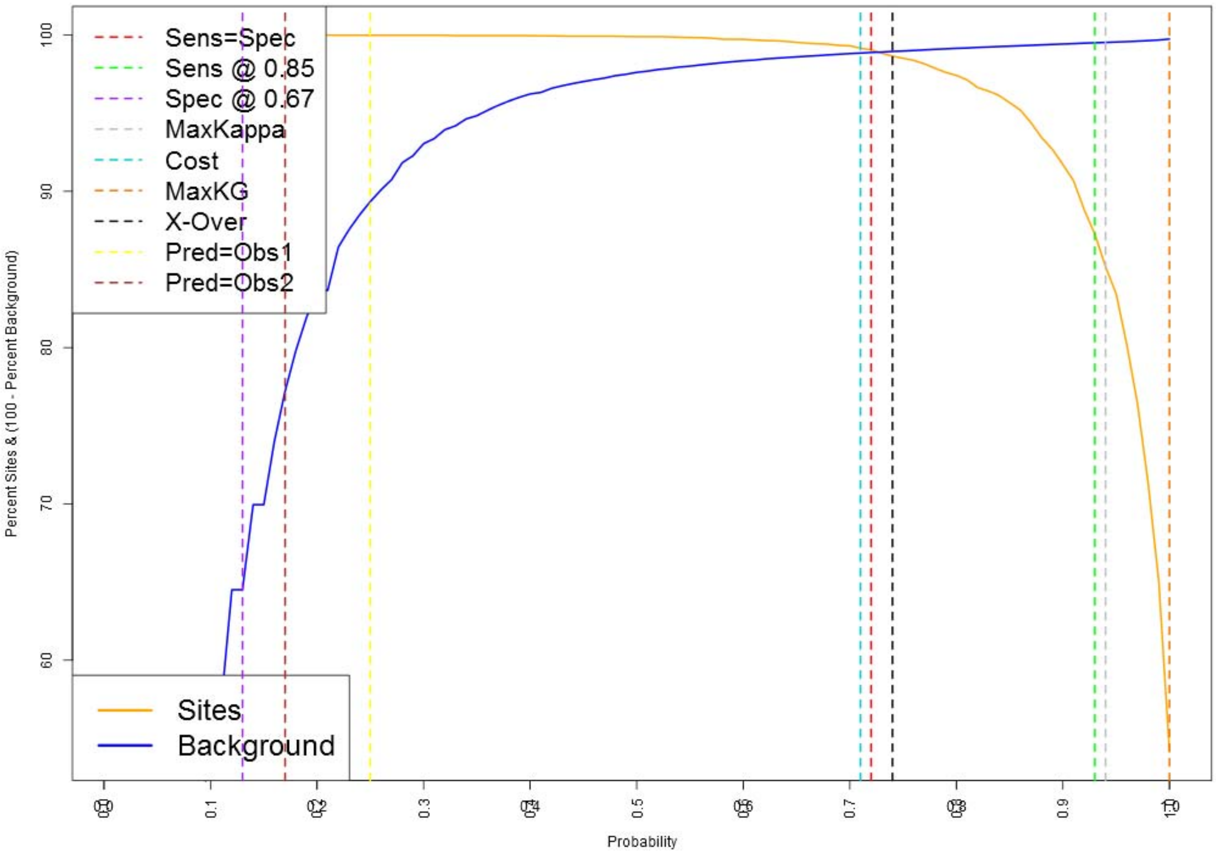


Chart 19. Region 4/5 West – Riverine Section 5

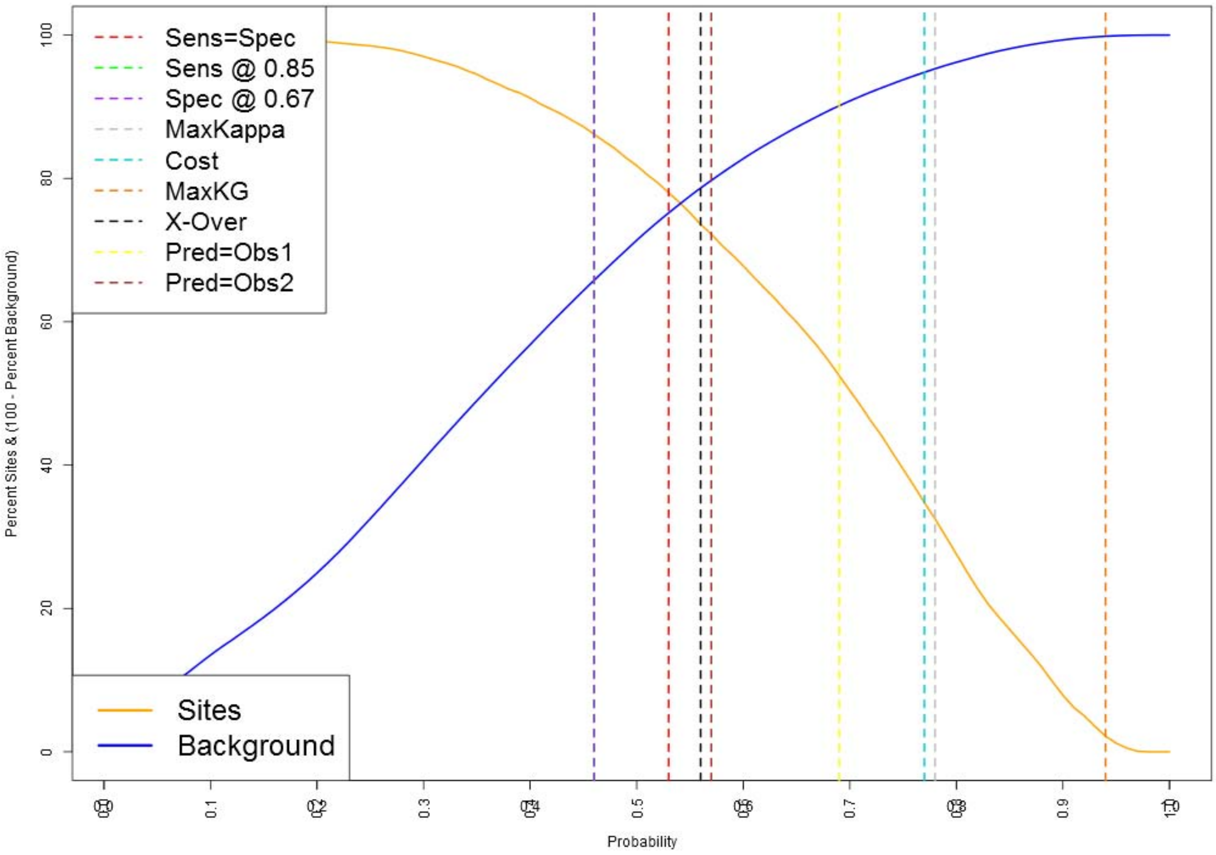


Chart 20. Region 4/5 West – Riverine Section 6

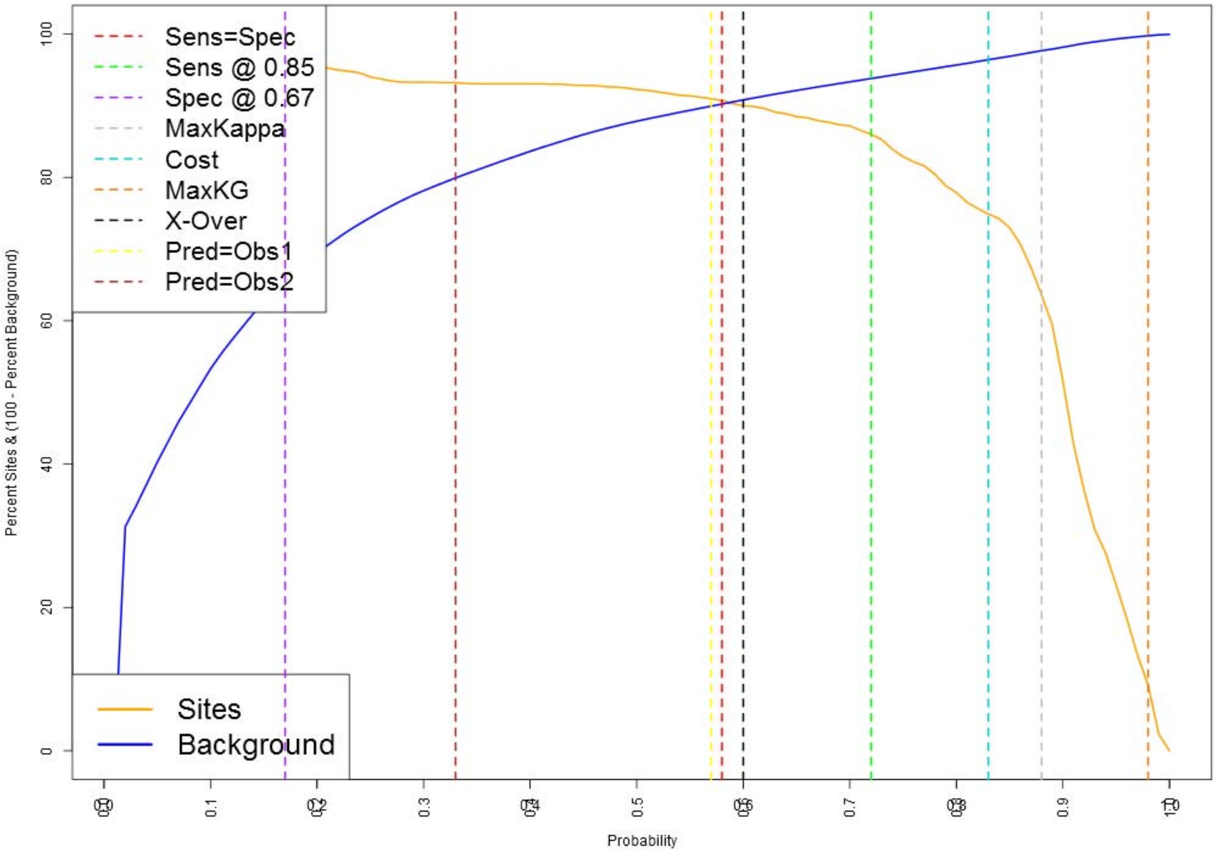


Chart 21. Region 4/5 West – Upland Section 1

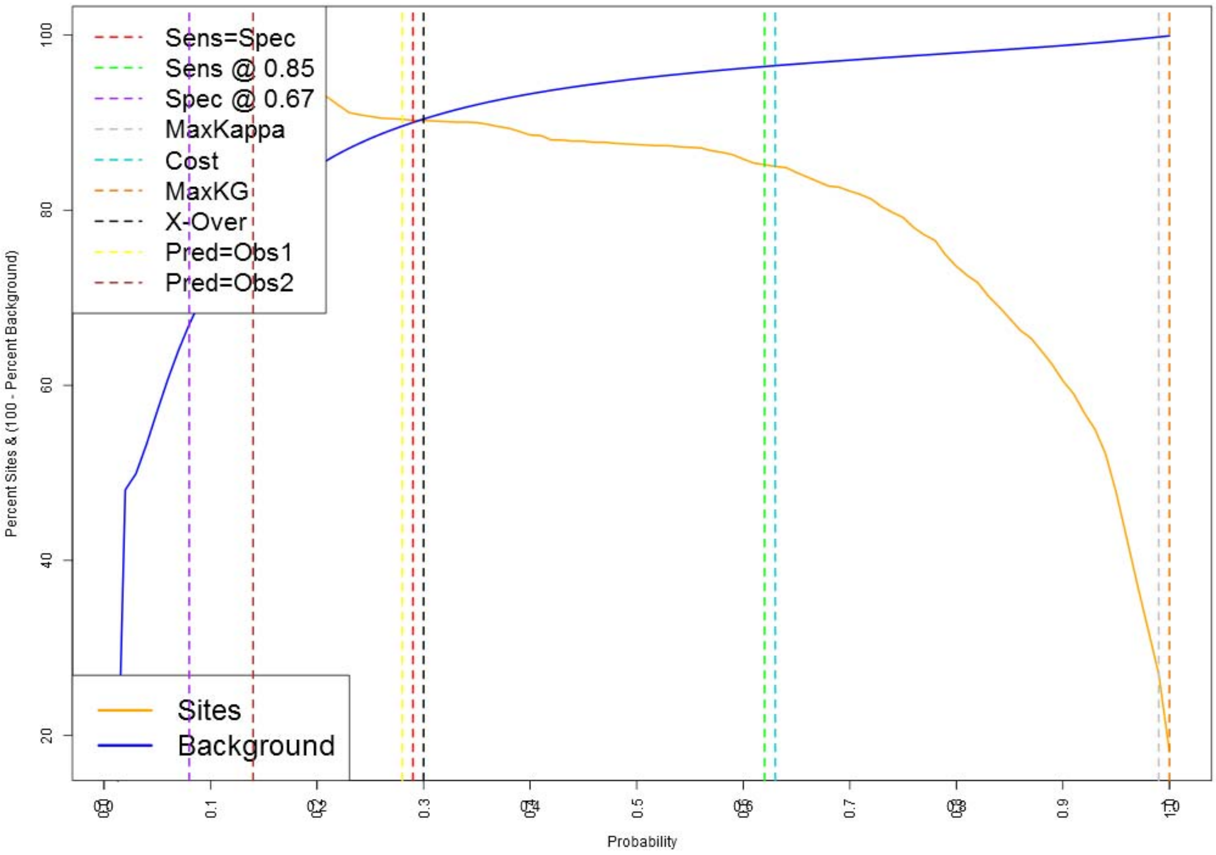


Chart 22. Region 4/5 West – Upland Section 2

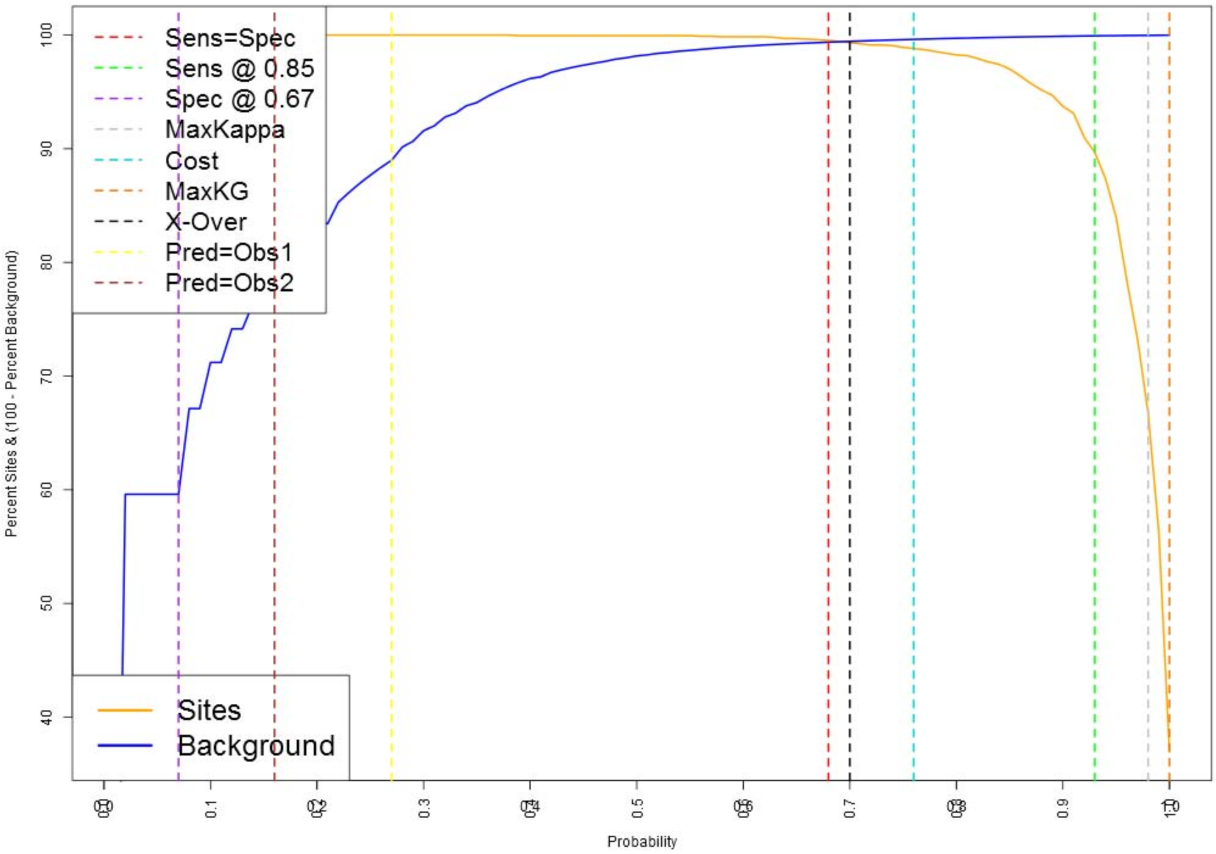


Chart 23. Region 4/5 West – Upland Section 3

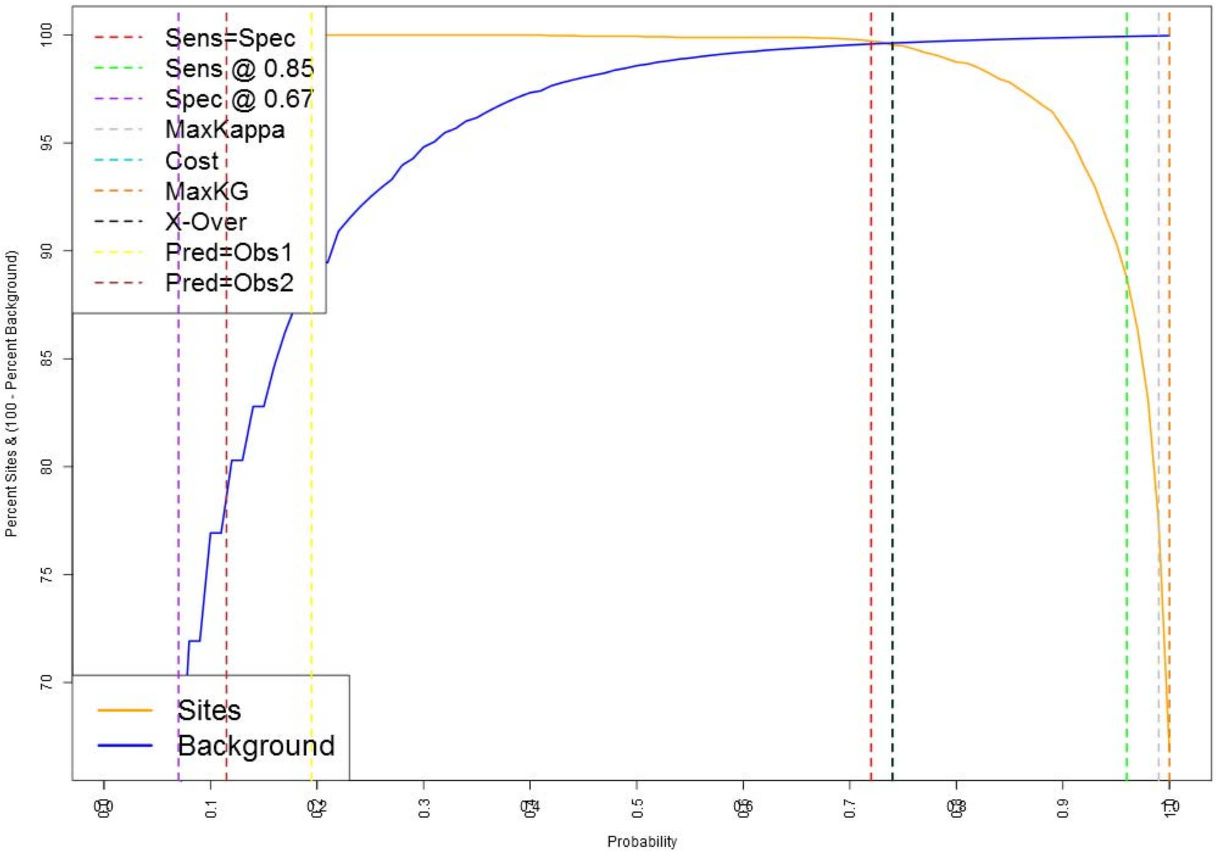


Chart 24. Region 4/5 West – Upland Section 4

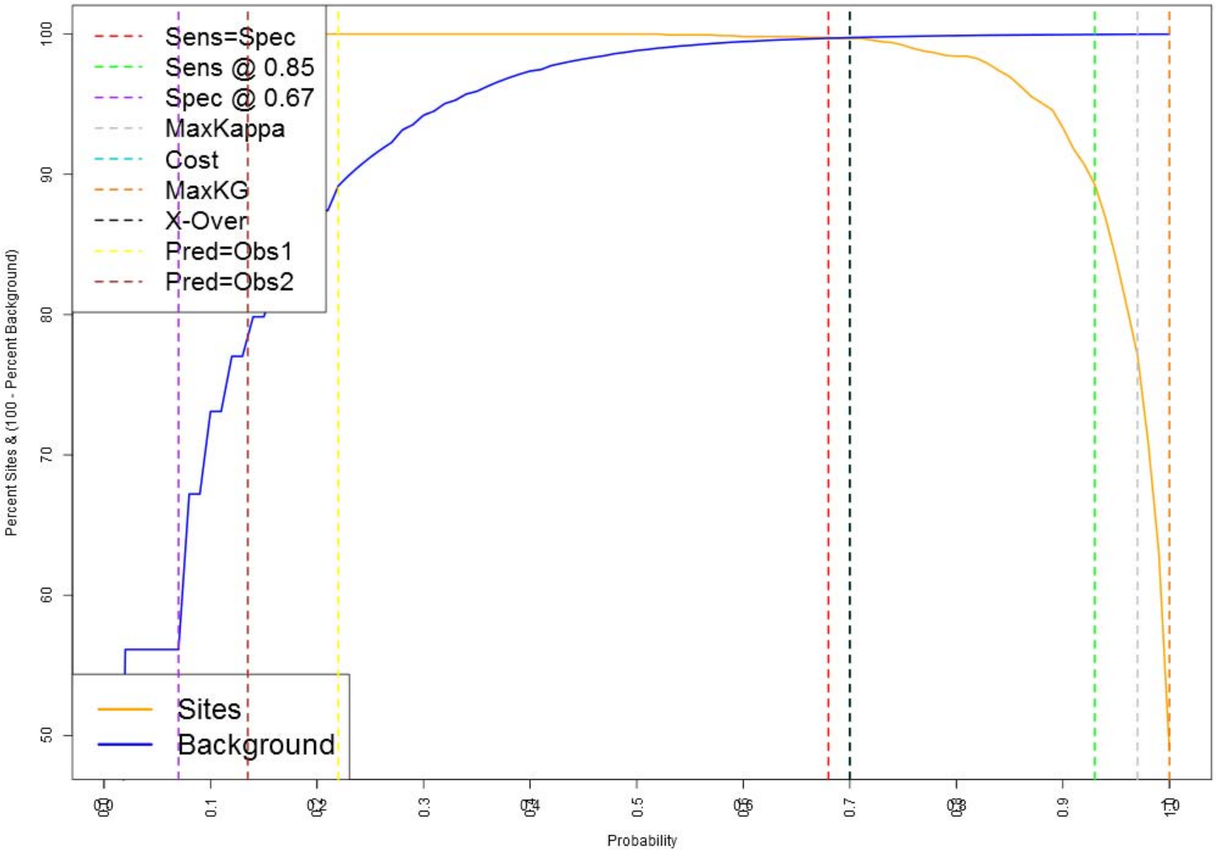


Chart 25. Region 4/5 West – Upland Section 5

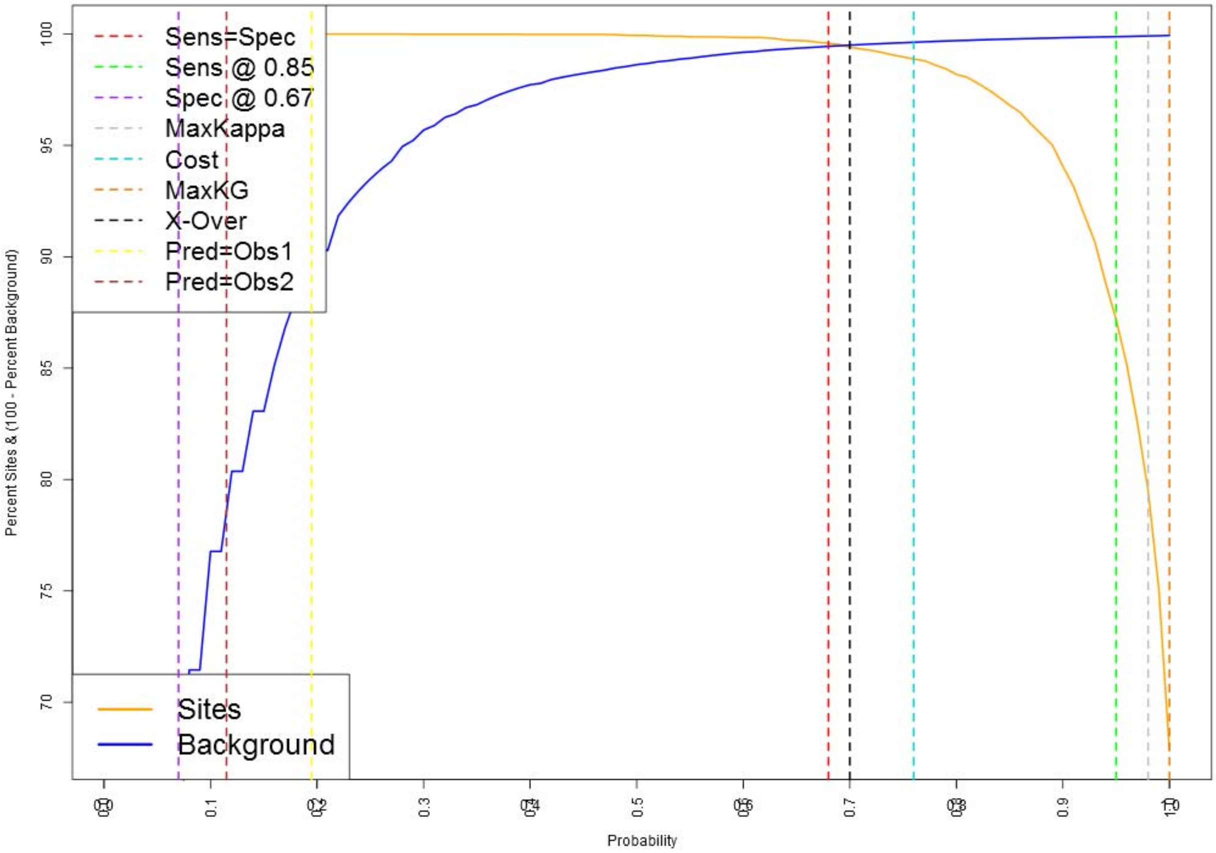


Chart 26. Region 4/5 West – Upland Section 6

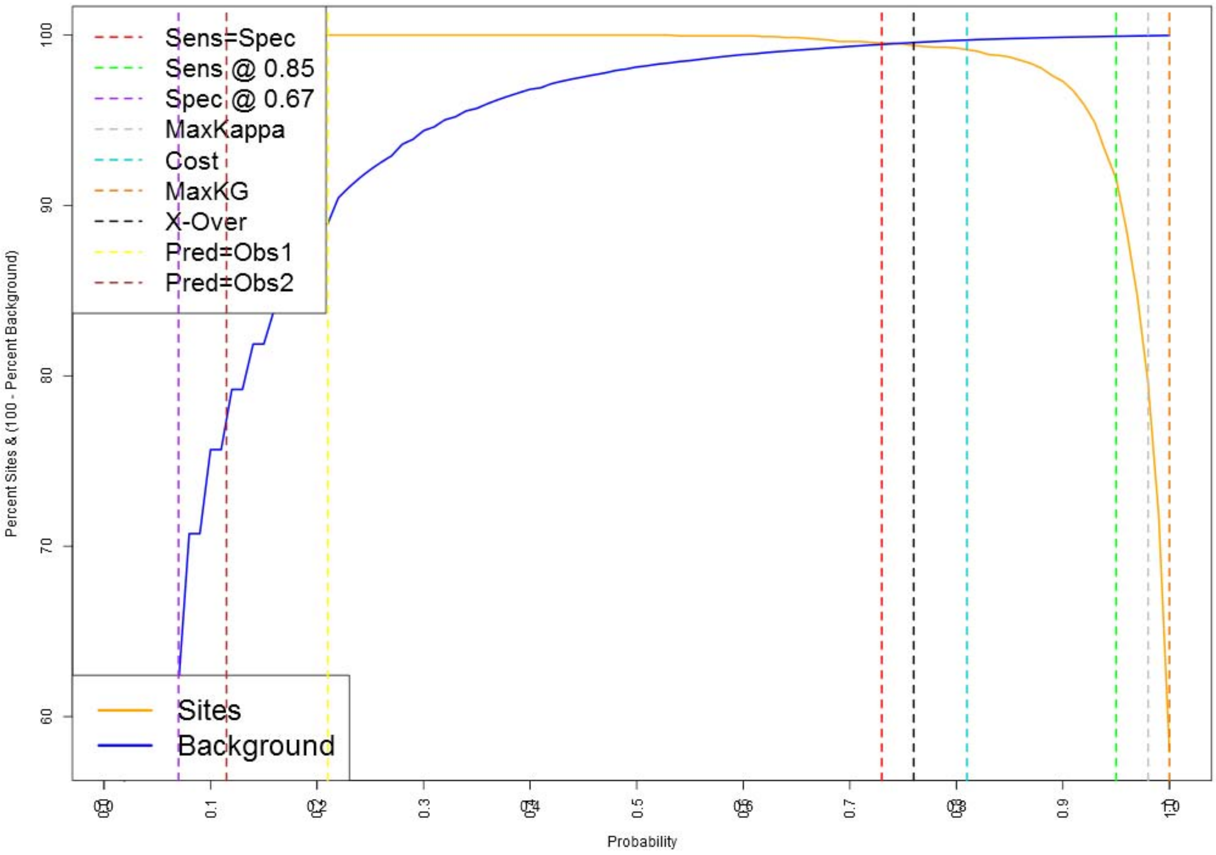


Chart 27. Region 6 All – Riverine Section 1

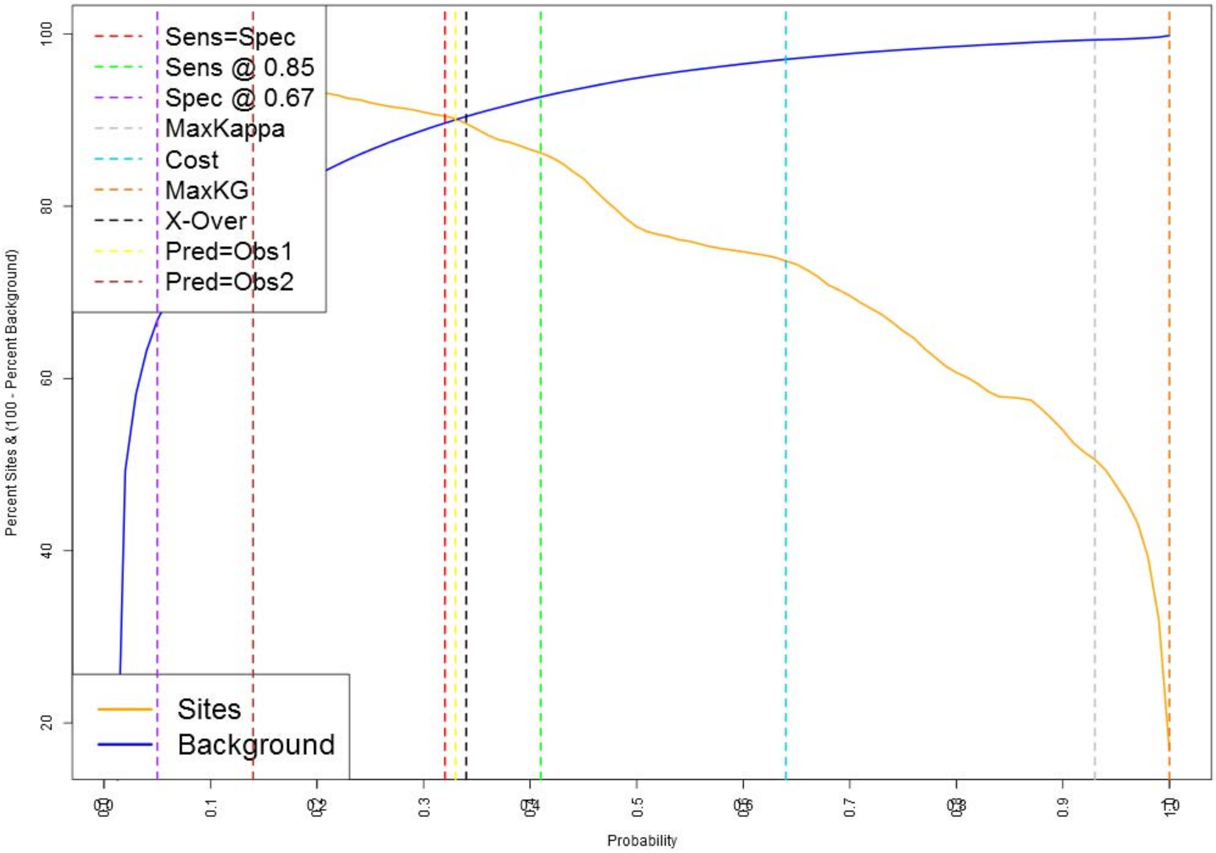


Chart 28. Region 6 All – Riverine Section 2

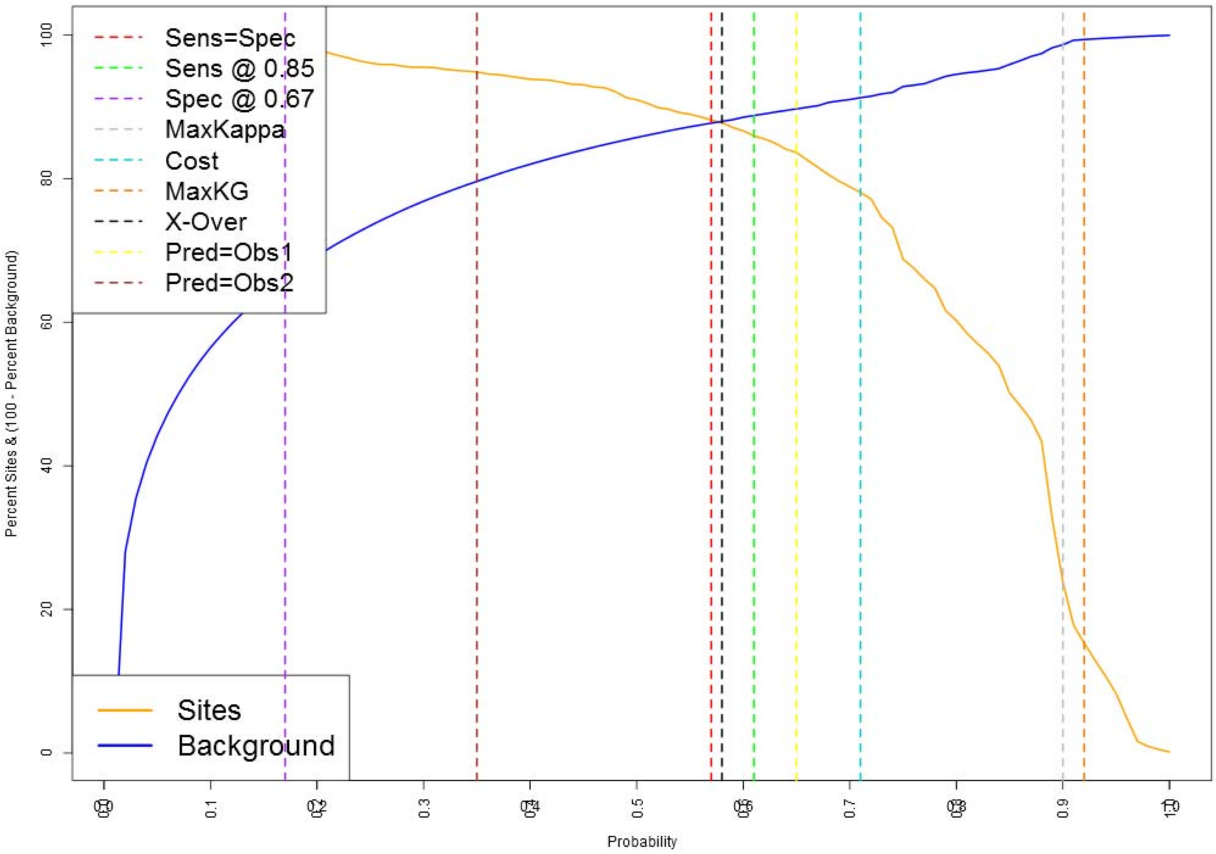


Chart 29. Region 6 All – Riverine Section 3

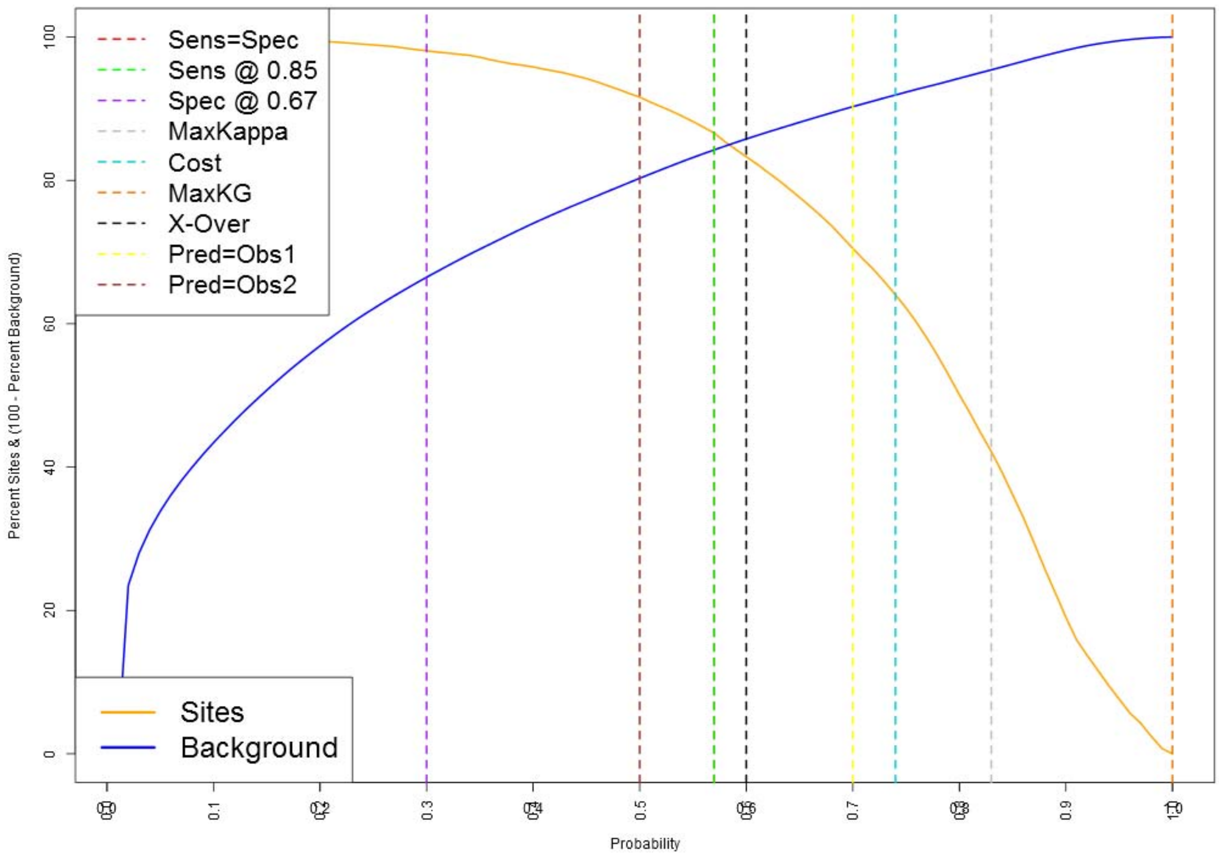


Chart 30. Region 6 All – Riverine Section 4

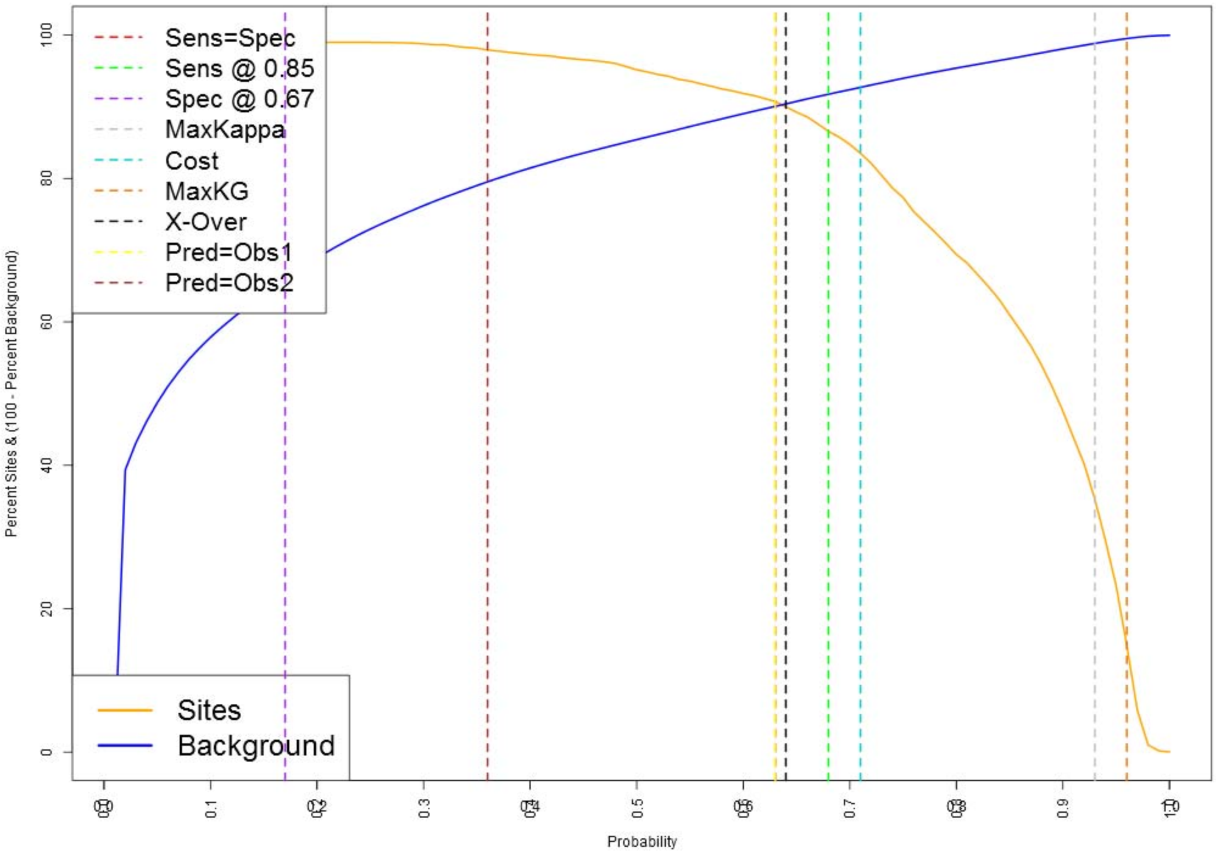


Chart 31. Region 6 All – Riverine Section 5

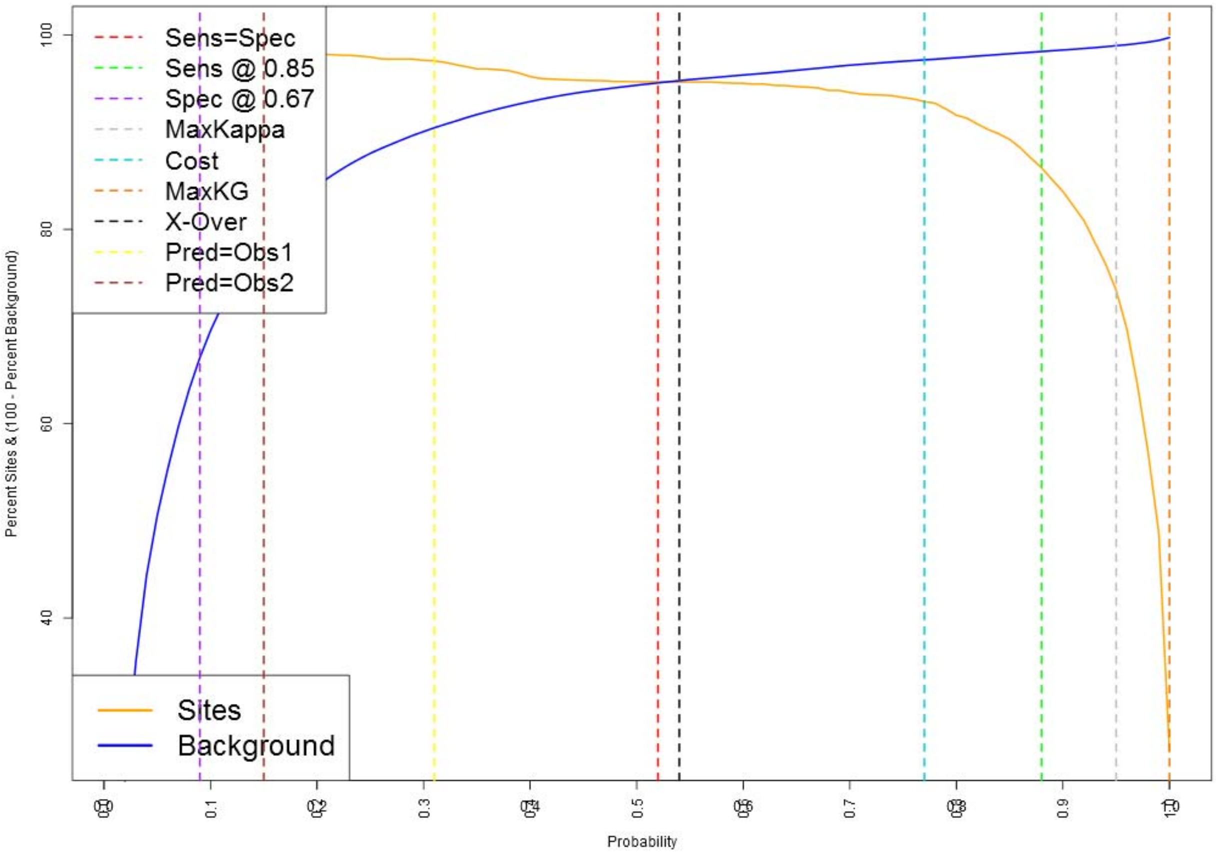


Chart 32. Region 6 All – Upland Section 1

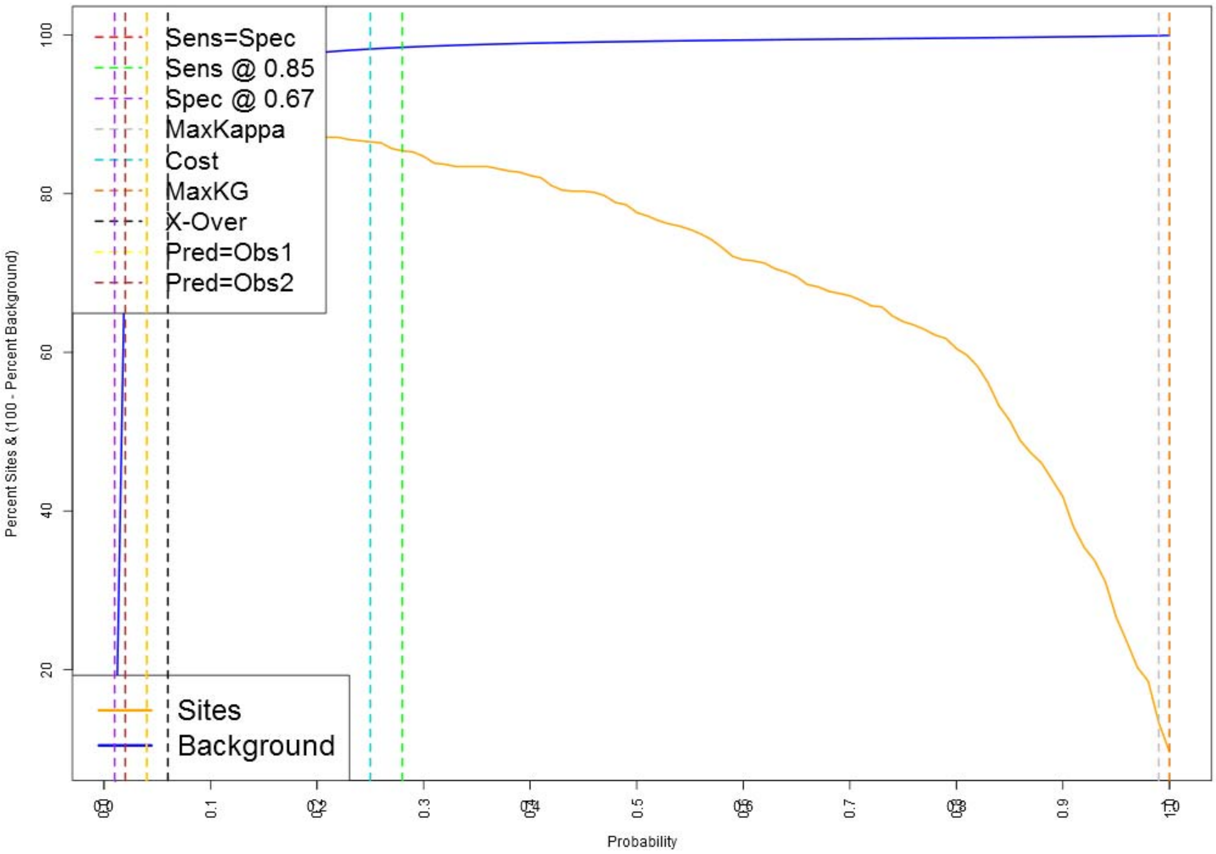


Chart 33. Region 6 All – Upland Section 2

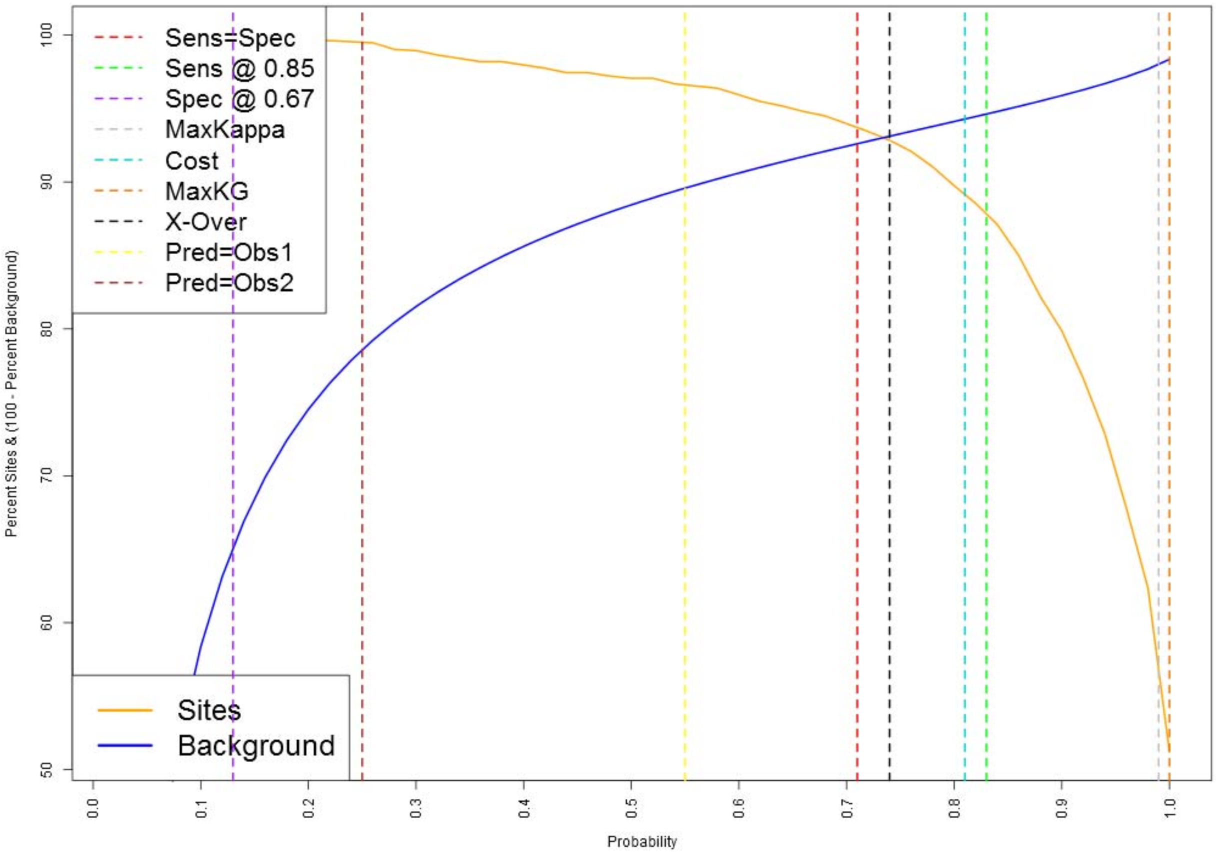


Chart 34. Region 6 All – Upland Section 3

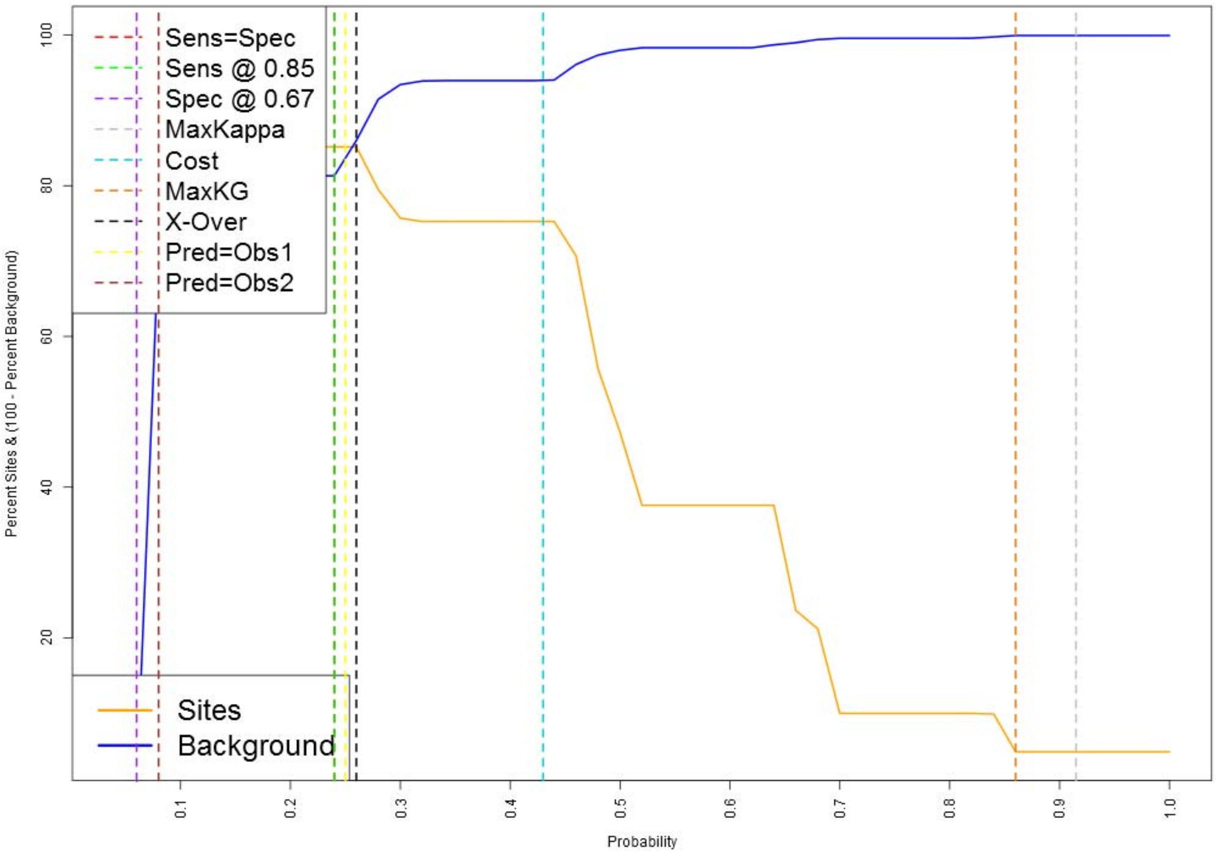


Chart 35. Region 6 All – Upland Section 4

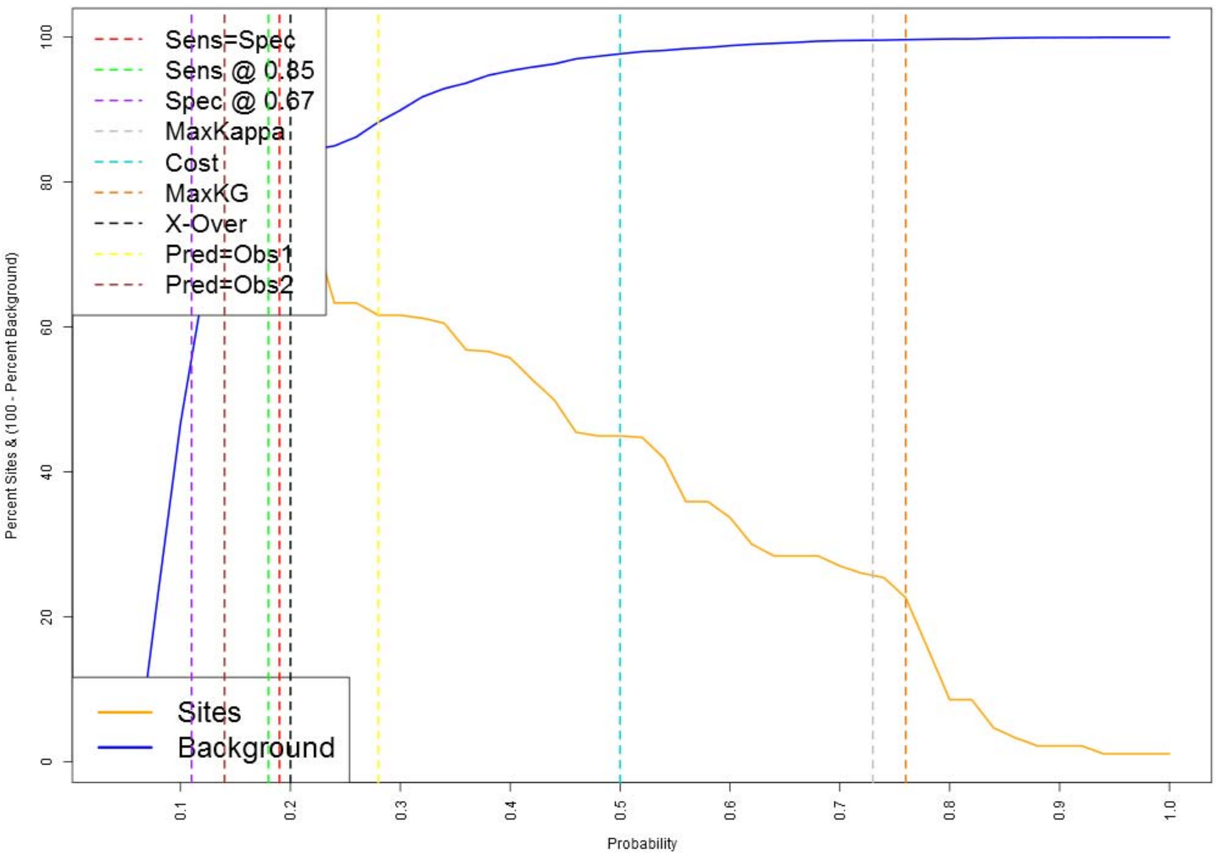
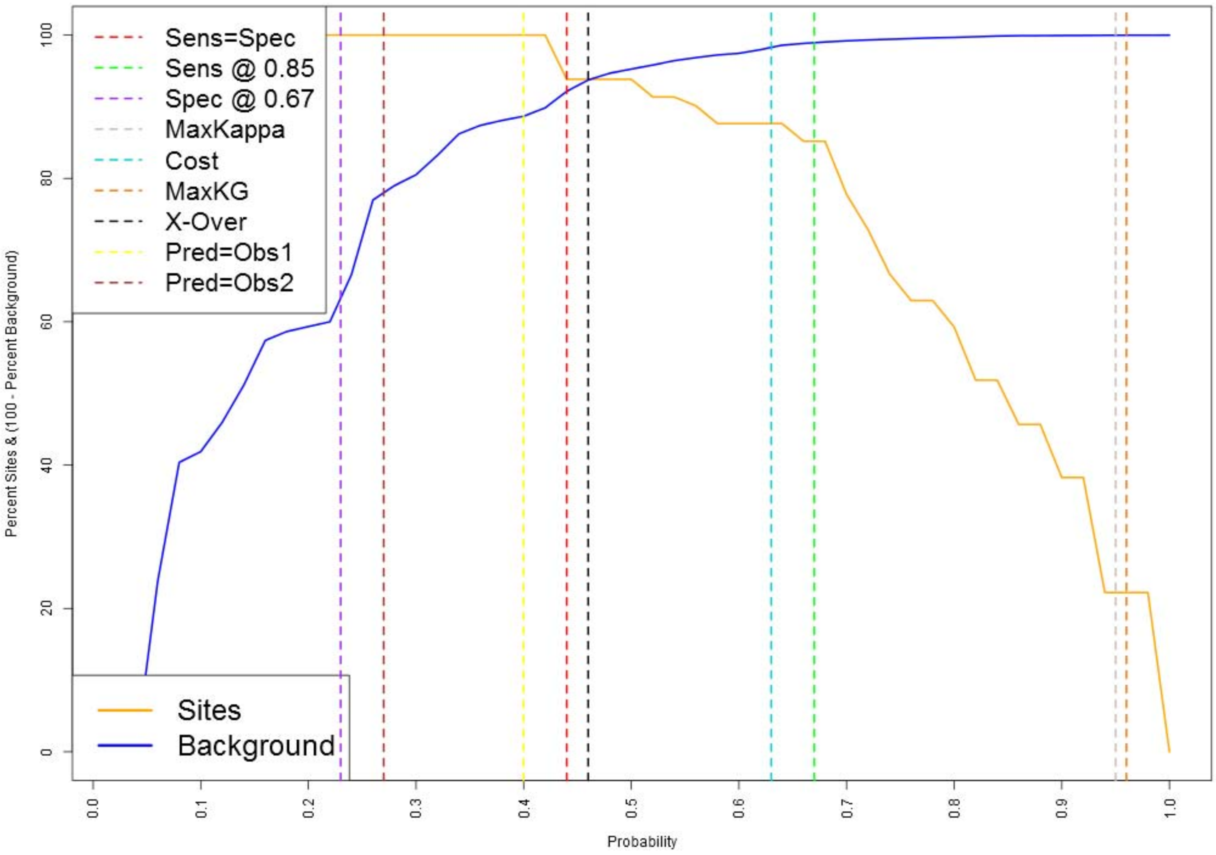


Chart 36. Region 6 All – Upland Section 5



APPENDIX G

CONFUSION MATRICES

FOR EACH OF 36 MODELS

WITHIN REGIONS 4, 5, AND 6

Region 4/5 East - Riverine Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	491	193437	193928
	Absent	0	485035	485035
		491	678472	678963

Sensitivity / TPR =	1.000
Specificity / TNR =	0.715
Prevalence =	0.0007
Kvamme Gain (Kg) =	0.714
Accuracy =	0.715
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) =	3.501
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.286

Region 4/5 East - Riverine Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	1046	267787	268833
	Absent	59	477724	477783
		1105	745511	746616

Sensitivity / TPR =	0.947
Specificity / TNR =	0.641
Prevalence =	0.0015
Kvamme Gain (Kg) =	0.620
Accuracy =	0.641
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) =	2.629
Negative Prediction Gain (NPG) =	0.083
False Negative Rate (FNR) =	0.053
Detection Prevalence =	0.360

Region 4/5 East - Riverine Section 3

		Known Sites		
		Present	Absent	
Model Prediction	Present	9213	265399	274612
	Absent	297	597369	597666
		9510	862768	872278

Sensitivity / TPR =	0.969
Specificity / TNR =	0.692
Prevalence =	0.0109
Kvamme Gain (Kg) =	0.675
Accuracy =	0.695
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) =	3.077
Negative Prediction Gain (NPG) =	0.046
False Negative Rate (FNR) =	0.031
Detection Prevalence =	0.315

Region 4/5 East - Riverine Section 4

		Known Sites		
		Present	Absent	
Model Prediction	Present	88333	1178545	1266878
	Absent	3438	2396968	2400406
		91771	3575513	3667284

Sensitivity / TPR =	0.963
Specificity / TNR =	0.670
Prevalence =	0.0250
Kvamme Gain (Kg) =	0.641
Accuracy =	0.678
Positive Prediction Value (PPV) =	0.070
Negative Prediction Value (NPV) =	0.999
Unexpected Discovery Rate (UDR) =	0.001
Detection Rate =	0.024
Positive Prediction Gain (PPG) =	2.786
Negative Prediction Gain (NPG) =	0.057
False Negative Rate (FNR) =	0.037
Detection Prevalence =	0.345

Region 4/5 East - Riverine Section 5

		Known Sites		
		Present	Absent	
Model Prediction	Present	13324	709460	722784
	Absent	130	1447783	1447913
		13454	2157243	2170697

Sensitivity / TPR =	0.990
Specificity / TNR =	0.671
Prevalence =	0.0062
Kvamme Gain (Kg) =	0.664
Accuracy =	0.673
Positive Prediction Value (PPV) =	0.018
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.006
Positive Prediction Gain (PPG) =	2.974
Negative Prediction Gain (NPG) =	0.014
False Negative Rate (FNR) =	0.010
Detection Prevalence =	0.333

Region 4/5 East - Riverine Section 6

		Known Sites		
		Present	Absent	
Model Prediction	Present	17653	553013	570666
	Absent	504	1170611	1171115
		18157	1723624	1741781

Sensitivity / TPR =	0.972
Specificity / TNR =	0.679
Prevalence =	0.0104
Kvamme Gain (Kg) =	0.663
Accuracy =	0.682
Positive Prediction Value (PPV) =	0.031
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.010
Positive Prediction Gain (PPG) =	2.967
Negative Prediction Gain (NPG) =	0.041
False Negative Rate (FNR) =	0.028
Detection Prevalence =	0.328

Region 4/5 East - Riverine Section 7

		Known Sites		
		Present	Absent	
Model Prediction	Present	54511	628216	682727
	Absent	6200	1319270	1325470
		60711	1947486	2008197

Sensitivity / TPR =	0.898
Specificity / TNR =	0.677
Prevalence =	0.0302
Kvamme Gain (Kg) =	0.621
Accuracy =	0.684
Positive Prediction Value (PPV) =	0.080
Negative Prediction Value (NPV) =	0.995
Unexpected Discovery Rate (UDR) =	0.005
Detection Rate =	0.027
Positive Prediction Gain (PPG) =	2.641
Negative Prediction Gain (NPG) =	0.155
False Negative Rate (FNR) =	0.102
Detection Prevalence =	0.340

Region 4/5 East - Upland Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	1069	3466435	3467504
	Absent	28	7142619	7142647
		1097	10609054	10610151

Sensitivity / TPR =	0.974
Specificity / TNR =	0.673
Prevalence =	0.0001
Kvamme Gain (Kg) =	0.665
Accuracy =	0.673
Positive Prediction Value (PPV) =	0.000
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	2.982
Negative Prediction Gain (NPG) =	0.038
False Negative Rate (FNR) =	0.026
Detection Prevalence =	0.327

Region 4/5 East - Upland Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	6350	3819859	3826209
	Absent	55	6991114	6991169
		6405	10810973	10817378

Sensitivity / TPR =	0.991
Specificity / TNR =	0.647
Prevalence =	0.0006
Kvamme Gain (Kg) =	0.643
Accuracy =	0.647
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) =	2.803
Negative Prediction Gain (NPG) =	0.013
False Negative Rate (FNR) =	0.009
Detection Prevalence =	0.354

Region 4/5 East - Upland Section 3

		Known Sites		
		Present	Absent	
Model Prediction	Present	3644	2517779	2521423
	Absent	16	7208095	7208111
		3660	9725874	9729534

Sensitivity / TPR =	0.996
Specificity / TNR =	0.741
Prevalence =	0.0004
Kvamme Gain (Kg) =	0.740
Accuracy =	0.741
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) =	3.842
Negative Prediction Gain (NPG) =	0.006
False Negative Rate (FNR) =	0.004
Detection Prevalence =	0.259

Region 4/5 East - Upland Section 4

		Known Sites		
		Present	Absent	
Model Prediction	Present	10111	4481978	4492089
	Absent	607	10256056	10256663
		10718	14738034	14748752

Sensitivity / TPR =	0.943
Specificity / TNR =	0.696
Prevalence =	0.0007
Kvamme Gain (Kg) =	0.677
Accuracy =	0.696
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) =	3.097
Negative Prediction Gain (NPG) =	0.081
False Negative Rate (FNR) =	0.057
Detection Prevalence =	0.305

Region 4/5 East - Upland Section 5

		Known Sites		
		Present	Absent	
Model Prediction	Present	1272	4700656	4701928
	Absent	0	11203425	11203425
		1272	15904081	15905353

Sensitivity / TPR =	1.000
Specificity / TNR =	0.704
Prevalence =	0.0001
Kvamme Gain (Kg) =	0.704
Accuracy =	0.704
Positive Prediction Value (PPV) =	0.000
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	3.383
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.296

Region 4/5 East - Upland Section 6

		Known Sites		
		Present	Absent	
Model Prediction	Present	8442	3611624	3620066
	Absent	1346	7608807	7610153
		9788	11220431	11230219

Sensitivity / TPR =	0.862
Specificity / TNR =	0.678
Prevalence =	0.0009
Kvamme Gain (Kg) =	0.626
Accuracy =	0.678
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) =	2.676
Negative Prediction Gain (NPG) =	0.203
False Negative Rate (FNR) =	0.138
Detection Prevalence =	0.322

Region 4/5 East - Upland Section 7

		Known Sites		
		Present	Absent	
Model Prediction	Present	16754	3358076	3374830
	Absent	420	7294861	7295281
		17174	10652937	10670111

Sensitivity / TPR =	0.976
Specificity / TNR =	0.685
Prevalence =	0.0016
Kvamme Gain (Kg) =	0.676
Accuracy =	0.685
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) =	3.084
Negative Prediction Gain (NPG) =	0.036
False Negative Rate (FNR) =	0.024
Detection Prevalence =	0.316

Region 4/5 West - Riverine Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	5431	352580	358011
	Absent	0	810179	810179
		5431	1162759	1168190

Sensitivity / TPR =	1.000
Specificity / TNR =	0.697
Prevalence =	0.0046
Kvamme Gain (Kg) =	0.694
Accuracy =	0.698
Positive Prediction Value (PPV) =	0.015
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.005
Positive Prediction Gain (PPG) =	3.263
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.306

Region 4/5 West - Riverine Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	6566	454220	460786
	Absent	517	958635	959152
		7083	1412855	1419938

Sensitivity / TPR =	0.927
Specificity / TNR =	0.679
Prevalence =	0.0050
Kvamme Gain (Kg) =	0.650
Accuracy =	0.680
Positive Prediction Value (PPV) =	0.014
Negative Prediction Value (NPV) =	0.999
Unexpected Discovery Rate (UDR) =	0.001
Detection Rate =	0.005
Positive Prediction Gain (PPG) =	2.857
Negative Prediction Gain (NPG) =	0.108
False Negative Rate (FNR) =	0.073
Detection Prevalence =	0.325

Region 4/5 West - Riverine Section 3

		Known Sites		
		Present	Absent	
Model Prediction	Present	10155	629261	639416
	Absent	0	1311779	1311779
		10155	1941040	1951195

Sensitivity / TPR =	1.000
Specificity / TNR =	0.676
Prevalence =	0.0052
Kvamme Gain (Kg) =	0.672
Accuracy =	0.677
Positive Prediction Value (PPV) =	0.016
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 4/5 West - Riverine Section 4

		Known Sites		
		Present	Absent	
Model Prediction	Present	7712	512206	519918
	Absent	0	1192276	1192276
		7712	1704482	1712194

Sensitivity / TPR =	1.000
Specificity / TNR =	0.699
Prevalence =	0.0045
Kvamme Gain (Kg) =	0.696
Accuracy =	0.701
Positive Prediction Value (PPV) =	0.015
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.005
Positive Prediction Gain (PPG) =	3.293
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.304

Region 4/5 West - Riverine Section 5

		Known Sites		
		Present	Absent	
Model Prediction	Present	14111	389592	403703
	Absent	2464	797958	800422
		16575	1187550	1204125

Sensitivity / TPR =	0.851
Specificity / TNR =	0.672
Prevalence =	0.0138
Kvamme Gain (Kg) =	0.606
Accuracy =	0.674
Positive Prediction Value (PPV) =	0.035
Negative Prediction Value (NPV) =	0.997
Unexpected Discovery Rate (UDR) =	0.003
Detection Rate =	0.012
Positive Prediction Gain (PPG) =	2.539
Negative Prediction Gain (NPG) =	0.224
False Negative Rate (FNR) =	0.149
Detection Prevalence =	0.335

Region 4/5 West - Riverine Section 6

		Known Sites		
		Present	Absent	
Model Prediction	Present	1710	147099	148809
	Absent	62	299594	299656
		1772	446693	448465

Sensitivity / TPR =	0.965
Specificity / TNR =	0.671
Prevalence =	0.0040
Kvamme Gain (Kg) =	0.656
Accuracy =	0.672
Positive Prediction Value (PPV) =	0.011
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.004
Positive Prediction Gain (PPG) =	2.908
Negative Prediction Gain (NPG) =	0.052
False Negative Rate (FNR) =	0.035
Detection Prevalence =	0.332

Region 4/5 West - Upland Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	1539	4765060	4766599
	Absent	21	10869502	10869523
		1560	15634562	15636122

Sensitivity / TPR =	0.987
Specificity / TNR =	0.695
Prevalence =	0.0001
Kvamme Gain (Kg) =	0.691
Accuracy =	0.695
Positive Prediction Value (PPV) =	0.000
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	3.236
Negative Prediction Gain (NPG) =	0.019
False Negative Rate (FNR) =	0.013
Detection Prevalence =	0.305

Region 4/5 West - Upland Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	1946	6800470	6802416
	Absent	0	13895871	13895871
		1946	20696341	20698287

Sensitivity / TPR =	1.000
Specificity / TNR =	0.671
Prevalence =	0.0001
Kvamme Gain (Kg) =	0.671
Accuracy =	0.671
Positive Prediction Value (PPV) =	0.000
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	3.043
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.329

Region 4/5 West - Upland Section 3

		Known Sites		
		Present	Absent	
Model Prediction	Present	3604	6915979	6919583
	Absent	0	17711952	17711952
		3604	24627931	24631535

Sensitivity / TPR =	1.000
Specificity / TNR =	0.719
Prevalence =	0.0001
Kvamme Gain (Kg) =	0.719
Accuracy =	0.719
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) =	3.560
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.281

Region 4/5 West - Upland Section 4

		Known Sites		
		Present	Absent	
Model Prediction	Present	1648	5198958	5200606
	Absent	0	10657122	10657122
		1648	15856080	15857728

Sensitivity / TPR =	1.000
Specificity / TNR =	0.672
Prevalence =	0.0001
Kvamme Gain (Kg) =	0.672
Accuracy =	0.672
Positive Prediction Value (PPV) =	0.000
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	3.049
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.328

Region 4/5 West - Upland Section 5

		Known Sites		
		Present	Absent	
Model Prediction	Present	11047	4268681	4279728
	Absent	0	10682426	10682426
		11047	14951107	14962154

Sensitivity / TPR =	1.000
Specificity / TNR =	0.714
Prevalence =	0.0007
Kvamme Gain (Kg) =	0.714
Accuracy =	0.715
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) =	3.496
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.286

Region 4/5 West - Upland Section 6

		Known Sites		
		Present	Absent	
Model Prediction	Present	2099	2484844	2486943
	Absent	0	6005877	6005877
		2099	8490721	8492820

Sensitivity / TPR =	1.000
Specificity / TNR =	0.707
Prevalence =	0.0002
Kvamme Gain (Kg) =	0.707
Accuracy =	0.707
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) =	3.415
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.293

Region 6 All - Riverine Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	5972	426892	432864
	Absent	172	960596	960768
		6144	1387488	1393632

Sensitivity / TPR =	0.972
Specificity / TNR =	0.692
Prevalence =	0.0044
Kvamme Gain (Kg) =	0.680
Accuracy =	0.694
Positive Prediction Value (PPV) =	0.014
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.004
Positive Prediction Gain (PPG) =	3.129
Negative Prediction Gain (NPG) =	0.041
False Negative Rate (FNR) =	0.028
Detection Prevalence =	0.311

Region 6 All - Riverine Section 2

		Known Sites		
		Present	Absent	
Model Prediction	Present	1830	876788	878618
	Absent	21	1801228	1801249
		1851	2678016	2679867

Sensitivity / TPR =	0.989
Specificity / TNR =	0.673
Prevalence =	0.0007
Kvamme Gain (Kg) =	0.668
Accuracy =	0.673
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) =	3.015
Negative Prediction Gain (NPG) =	0.017
False Negative Rate (FNR) =	0.011
Detection Prevalence =	0.328

Region 6 All - Riverine Section 3

		Known Sites		
		Present	Absent	
Model Prediction	Present	22850	716144	738994
	Absent	487	1473778	1474265
		23337	2189922	2213259

Sensitivity / TPR =	0.979
Specificity / TNR =	0.673
Prevalence =	0.0105
Kvamme Gain (Kg) =	0.659
Accuracy =	0.676
Positive Prediction Value (PPV) =	0.031
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.010
Positive Prediction Gain (PPG) =	2.932
Negative Prediction Gain (NPG) =	0.031
False Negative Rate (FNR) =	0.021
Detection Prevalence =	0.334

Region 6 All - Riverine Section 4

		Known Sites		
		Present	Absent	
Model Prediction	Present	5278	569327	574605
	Absent	53	1160204	1160257
		5331	1729531	1734862

Sensitivity / TPR =	0.990
Specificity / TNR =	0.671
Prevalence =	0.0031
Kvamme Gain (Kg) =	0.665
Accuracy =	0.672
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) =	2.989
Negative Prediction Gain (NPG) =	0.015
False Negative Rate (FNR) =	0.010
Detection Prevalence =	0.331

Region 6 All - Riverine Section 5

		Known Sites		
		Present	Absent	
Model Prediction	Present	3818	148672	152490
	Absent	26	339902	339928
		3844	488574	492418

Sensitivity / TPR =	0.993
Specificity / TNR =	0.696
Prevalence =	0.0078
Kvamme Gain (Kg) =	0.688
Accuracy =	0.698
Positive Prediction Value (PPV) =	0.025
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.008
Positive Prediction Gain (PPG) =	3.207
Negative Prediction Gain (NPG) =	0.010
False Negative Rate (FNR) =	0.007
Detection Prevalence =	0.310

Region 6 All - Upland Section 1

		Known Sites		
		Present	Absent	
Model Prediction	Present	671	5128113	5128784
	Absent	35	15719647	15719682
		706	20847760	20848466

Sensitivity / TPR =	0.950
Specificity / TNR =	0.754
Prevalence =	0.0000
Kvamme Gain (Kg) =	0.741
Accuracy =	0.754
Positive Prediction Value (PPV) =	0.000
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	3.863
Negative Prediction Gain (NPG) =	0.066
False Negative Rate (FNR) =	0.050
Detection Prevalence =	0.246

Region 6 All - Upland Section 2*

		Known Sites		
		Present	Absent	
Model Prediction	Present	1321	8038467	8039788
	Absent	5	17463765	17463770
		1326	25502232	25503558

Sensitivity / TPR =	0.996
Specificity / TNR =	0.685
Prevalence =	0.0001
Kvamme Gain (Kg) =	0.684
Accuracy =	0.685
Positive Prediction Value (PPV) =	0.000
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	3.160
Negative Prediction Gain (NPG) =	0.006
False Negative Rate (FNR) =	0.004
Detection Prevalence =	0.315

* combined rock shelter and non-rock shelter specific models

Region 6 All - Upland Section 3

		Known Sites		
		Present	Absent	
Model Prediction	Present	1191	7766786	7767977
	Absent	163	28278254	28278417
		1354	36045040	36046394

Sensitivity / TPR =	0.880
Specificity / TNR =	0.785
Prevalence =	0.0000
Kvamme Gain (Kg) =	0.755
Accuracy =	0.785
Positive Prediction Value (PPV) =	0.000
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	4.082
Negative Prediction Gain (NPG) =	0.153
False Negative Rate (FNR) =	0.120
Detection Prevalence =	0.215

Region 6 All - Upland Section 4

		Known Sites		
		Present	Absent	
Model Prediction	Present	980	8633282	8634262
	Absent	23	15717017	15717040
		1003	24350299	24351302

Sensitivity / TPR =	0.977
Specificity / TNR =	0.645
Prevalence =	0.0000
Kvamme Gain (Kg) =	0.637
Accuracy =	0.645
Positive Prediction Value (PPV) =	0.000
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	2.756
Negative Prediction Gain (NPG) =	0.036
False Negative Rate (FNR) =	0.023
Detection Prevalence =	0.355

Region 6 All - Upland Section 5

		Known Sites		
		Present	Absent	
Model Prediction	Present	81	2657241	2657322
	Absent	0	5290570	5290570
		81	7947811	7947892

Sensitivity / TPR =	1.000
Specificity / TNR =	0.666
Prevalence =	0.0000
Kvamme Gain (Kg) =	0.666
Accuracy =	0.666
Positive Prediction Value (PPV) =	0.000
Negative Prediction Value (NPV) =	1.000
Unexpected Discovery Rate (UDR) =	0.000
Detection Rate =	0.000
Positive Prediction Gain (PPG) =	2.991
Negative Prediction Gain (NPG) =	0.000
False Negative Rate (FNR) =	0.000
Detection Prevalence =	0.334

Complete Model

		Known Sites		
		Present	Absent	
Model Prediction	Present	343773	97632926	97976699
	Absent	17149	228997869	229015018
		360922	326630795	326991717

Sensitivity / TPR =	0.952
Specificity / TNR =	0.701
Prevalence =	0.0011
Kvamme Gain (Kg) =	0.685
Accuracy =	0.701
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) =	3.179
Negative Prediction Gain (NPG) =	0.068
False Negative Rate (FNR) =	0.048
Detection Prevalence =	0.300