TURUN
YLIOPISTO

Maximum Entropy Modeling of the Iron Age Settlement Distributions in River Valleys of Turku Region, Southwest Finland

Akseli Tolvi
Master’s Thesis
Degree Programme in History and Archaeology, Archaeology
School of History, Culture and Arts Studies
Faculty of Humanities
University of Turku
April 2023

Master's Thesis

Degree Programme in History and Archaeology, Archaeology
Akseli Tolvi
Maximum Entropy Modeling of the Iron Age Settlement Distributions in River Valleys of Turku Region, Southwest Finland
Number of pages: 71 pages, 4 appendices

Species distribution models (SDM) are predictive modeling tools widely used in analytical biology that have also found applications in archaeological research. They can be used to quickly produce predictive maps for a variety of use cases like conservation and to guide field surveys. Modern SDMs take advantage of advances in computing like machine learning and artificial intelligence to achieve better predictions.

In this study Maximum Entropy, or MaxEnt, machine learning SDM algorithm was used to create predictive models of the Iron Age settlement around Turku region in Southwest Finland, focusing on Aurajoki, Savijoki, and Vähäjoki river valleys. MaxEnt is the most popular SDM algorithm, largely due to its ability to create predictions based on presence-only data and consistently good performance. Only open access -data was used, and the selection of variables was based on availability and previous studies.

The results show that MaxEnt can create in some cases surprisingly accurate models based on archaeological information, but the results were limited by the quality of existing data. The most influential variable was distance to water, which was the majority contributor whenever present. Even without the variable, the predicted distributions followed the waterways closely due to the influence of other variables.

It was concluded that to improve the accuracy of the results the quality of the data should be a major focus. The results should also be tested through field surveys. Additionally, attention should be based on the model conception.

Key words: archaeology, the Iron Age, machine learning, MaxEnt, maximum entropy, predictive modeling, species distribution modeling, Southwest Finland

Table of contents

1 Introduction 5

1.1 Objectives and study area 6

1.2 Data and software 7

2 Background 8

2.1 The Iron Age 8

2.1.1 Sites and settlements 8

2.1.2 Environment and the people 10

2.1.3 Communities and subsistence 11

2.2 Species distribution modeling 12

2.2.1 Niche theory 13

2.2.2 Occurrence data and modeling methods 15

2.2.3 MaxEnt algorithm 16

2.2.4 Environmental data 17

2.2.5 Model evaluation methods 19

2.2.6 Applications in archaeology 20

2.3 Concept of scale 23

3 Modeling process 24

3.1 Data preparation 24

3.1.1 Occurrence data 25

3.1.2 Predictor variables 26

3.2 Predictor evaluation 28

3.2.1 Assessment of quality of environmental layers 29

3.2.2 Variable correlation 29

3.2.3 Environmental niches 32

3.2.4 Predictor responses 36

3.2.5 Variable significance 37

3.2.6 Summary 38

3.3 Model fitting and predictions 39

3.3.1 Simplicity versus complexity 40

3.3.2 Distance to water 42

3.3.3 Absolute or relative elevation 45

3.3.4 Location of suitable areas 48

3.3.5 Summary 51

4 Discussion 53

4.1 Human-environment relations 54

4.2 Data 55

4.2.1 Response variable 56

4.2.2 Predictor variables 58

4.3 On to the future 60

5 Conclusion 62

Abbreviations 63

References 64

Appendices 72

Appendix 1 Lyhennelmä (summary in Finnish) 72

Appendix 2 Training sites 79

Appendix 3 Predictor variables 80

Appendix 4 Models 85

1 Introduction

Locating remains of past human activity is a central objective of the study of archaeology. Especially with the rapidly changing world of the modern times the archaeological heritage is becoming increasingly threatened. New tools are needed to answer these challenges.

Predictive models are methods, which at their minimum aim to predict the locations of archaeological sites or material (Verhagen, 2007, p. 13). As such, they can also serve as a practical tool for selecting target areas for more intensive surveys. They are based on the expectation that locations of archaeological sites are not random but connected to certain environmental factors (Debenjak-Ijäs, 2018, p. 7). The origins of their use in archaeology lie in the new archaeology movement in the late 1960s (Verhagen, 2007, pp. 13–14), though there has been interest in the subject even before that (Yaworsky et al., 2020, p. 2). In Finland, predictive modeling has been practiced for a few decades. In the 1990s Kirkinen (1996) carried out an early pioneering work, while more recent examples include theses of Tiilikkala (2017) and Debenjak-Ijäs (2018).

Archaeological predictive models can be divided into two distinct approaches (Verhagen, 2007, pp. 13–14): data driven, and theory driven. The data driven approach is the most common of the two (Verhagen, 2007, p. 15) and it relies on statistical tests to find correlations between site locations and environmental features. A theory driven approach, meanwhile, starts from a hypothesis on location preferences and selects and weights landscape parameters based on that.

Recent developments in computing power have allowed for the creation of more powerful statistical models. Species distribution models (SDM) are methods of predictive modeling widely used in analytical biology (Phillips et al., 2006, p. 232) and have also seen some limited use in archaeological research (e.g., Yaworsky et al., 2020 or Rafuse, 2021). Their roots lie deep in ecological theory (Hutchinson, 1957), and they range from simple envelope models to more sophisticated methods utilizing machine learning and artificial intelligence.

In this thesis I create a predictive model using the most widely used SDM algorithm MaxEnt or Maximum Entropy (Feng et al., 2019, p. 10365). I begin by laying the framing of the study through a look in the study area and the objectives. Afterwards, I discuss the archaeological context and the theoretical background of the SDMs as well as their applications in archaeology, before moving onto more practical topics from data preparation and evaluation to model fitting. In the final chapters of the thesis, I discuss the results and their implications. A summary of this thesis in Finnish is included in appendix 1. Referencing was done following APA guidelines.

1.1 Objectives and study area

In this study I will test the use of species distribution modeling archaeological context in South-West Finland. The AOI (Figure 1) is a 24 km by 24 km square bounding box covering the river valleys of Aurajoki, Vähäjoki, and Savijoki and containing parts of municipalities of Aura, Kaarina, Lieto, Masku, Nousiainen, Paimio, Raisio, Rusko, and Turku. Compared to the provincial scale models created by Tiilikkala (2017) and Debenjak-Ijäs (2018) the scope of this study is more limited. The focus is on the methodology and the aim is to create multiple models to test a number of variables in several model configurations.

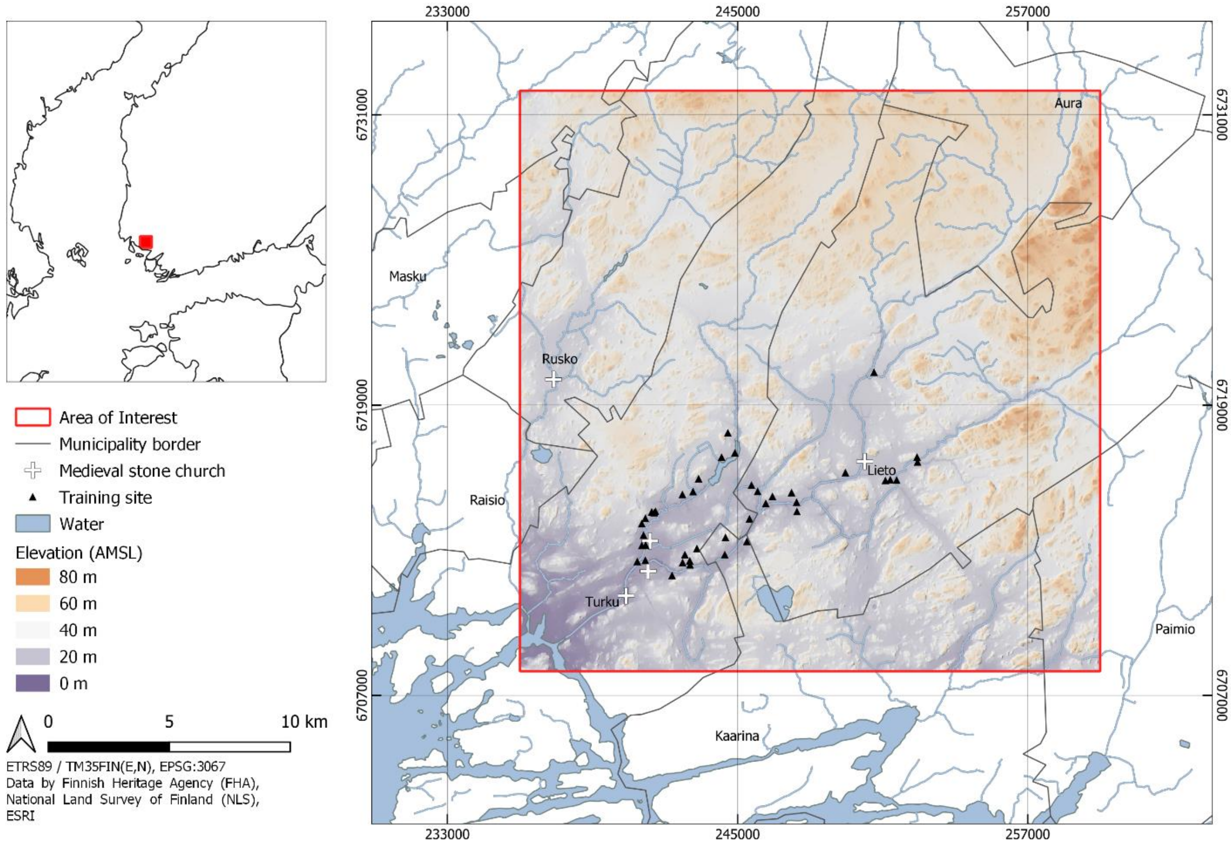


Figure 1. Map of the area of interest (colored) in the Turku region. Black triangles represent the Iron Age settlements used in model training.

This study is based on a hypothesis that known locations of Iron Age settlements can be used as a proxy to identify presence of the Iron Age people and thus as the occurrence data for an SDM. The main objective of this study is not to create accurate maps of suitability to locate new sites, but to test whether species distribution modeling algorithms like MaxEnt can, in general be used to produce predictive maps from archaeological records. Therefore, the main research questions are as follows:

- Can SDM algorithms like MaxEnt identify favorable areas for human settlement and can they find connections between archaeological site locations and certain environmental factors?
- Which combination of environmental variables yields the best results?
- Which variables explain the distribution pattern of the Iron Age settlements the best?

1.2 Data and software

This thesis is based on several open-access datasets available from various sources. The species data is based on All the sites of the cultural environment registers of the Finnish Heritage Agency (for research purposes) information product –dataset (Finnish Heritage Agency [FHA], 2021, retrieved 19.12.2021). The environmental layers are based on Elevation model 2 m (National Land Survey [NLS], n.d. -a, retrieved 2.12.2021) and Topographic database (NLS, 2021, retrieved 2.12.2021), Superficial deposits 1:20000/1:50000 (Geological Survey of Finland [GTK], 2015, retrieved 25.11.2021) and Shoreline – River Network (Ranta10) (Finnish Environment Institute [SYKE], 2012, retrieved 13.1.2022) datasets. ETRS89 / TM35FIN(E,N), EPSG:3067 reference system was used during data preparation and modeling as well as in all visualizations.

QGIS 3.16.3 with GRASS 7.8.5 was used for most of the processing of geospatial data as well as the final visualizations. A few analyses were carried out in ArcGIS Pro 2.9.1. Additional data processing and visualization was done in MS Office Excel.

Predictive modeling was carried out in R environment, using MaxEnt 3.4.4 software (Phillips et al., n.d.) interfaced with R 4.2.0 through RStudio 2021.09.2 using an interfacing function in library dismo. Additionally, functions from libraries raster, rgdal, maptools and rJava were used. Finally, some data visualizations were made in R using functions in scatterplot3d – library.

2 Background

This chapter discusses the framing and the theoretical background of this study. It begins with an overview of the Iron Age, focused on the study area, and continues into SDMs starting from the theory and moving on towards the more practical subjects and finally to their applications in archaeological studies. Lastly, the final subchapter delves on the subject of scale and discusses its implications on the study.

2.1 The Iron Age

Due to the nature and limitations of the research material, the period of interest of this study is, rather vaguely, the Iron Age. The Iron Age is the third and youngest part of the three-age system formulated by C.J. Thomsen in 1836 to order the collections in the Royal Museum of Antiquities in Copenhagen (Heizer, 1962, p. 259). The three-age theory was developed at the time when scientific prehistoric archaeology was beginning to be practiced, which led to its widespread adoption and eventually becoming the standard classification in studies of Eurasian prehistory. However, unlike the classification with its clear distinction between the ages, development from Bronze Age to Iron Age was gradual. Neither does the beginning of the Iron Age mean that the use of bronze ceased. As such the three-age system is fundamentally an artificial, but also highly practical construct.

In this study I use the Finnish chronology. The Iron Age in Southwest Finland begins after the end of the Bronze Age at around 500 BCE and ends at around 1200 CE at the dawn of the Middle Ages (Raninen & Wessman, 2015, p. 216). The period is divided into the early Iron Age, consisting of pre-roman, early roman and late roman periods, the middle Iron Age, consisting of migration and Merovingian periods, and the late Iron Age, consisting of Viking and crusade periods. The length of the period introduces some challenges to modeling, which will be discussed later.

2.1.1 Sites and settlements

As the objective of this thesis is to model the distribution of Iron Age settlements, it is essential to consider how the term is defined. The definition of the term settlement is broad, and it is often considered to be a residual category including sites with disputable identification (Seppälä, 2000, pp. 193–194). According to Seppälä (referencing Taavitsainen, 1992), some actual settlements may have been misidentified as, for example, cremation cemeteries under level ground. According to Tokoi (2020, p. 3) site is often defined as a settlement by excluding other potential uses like burials. This method, however, also runs the risk of misidentification.

Tokoi (2020, p. 3) raises the question whether all sites showing signs of human activity should be considered settlements or if the settlement is only some sort of core area. Seppälä (2000, p. 194) states that if the artefacts, structures, topography, and soil composition support the definition, it should be used even if the site has previously been classified as a burial site. Further questions raised by Tokoi (2020, pp. 3–5) include, whether a site used for camping for one night should be considered settlement or if the definition should imply a semi-permanent activity. As it stands, the identification of the Iron Age settlements is a topic littered with uncertainties.

In addition to identification of type, the dating of Iron Age sites is also problematic (Seppälä, 2000, pp. 194–195). Artefacts of the period are often not unlike those of historic times as the changes in daily life have happened slowly. For example, ceramic vessels of the so-called Iron Age type were manufactured still in the Middle Ages. Most of the known dwelling sites are the result of a cursory field survey and are only dated imprecisely to the Iron Age. The Iron Age is already a long period and the problem with dating ceramics only escalates the issue further.

An additional problem is the definition of a site and its representation. In European archaeological tradition a site is an area in which archaeological data are found in contrast to the surrounding area where the data is absent (Barford, 2000, pp. 85–87). In other words, there is a difference between ‘a site’ and ‘not site’. Barford also discusses two different definitions of the term identified in archaeological literature. One is a place where archaeological remains have been found, usually not in the original deposition layers, for example during a fieldwalking survey. The second are the places where the original depositional layers have been excavated. The information from the first category of the sites is tremendously less than from the second and they are the ones more likely to be misinterpreted. Most of the settlement sites within the study area fall within the first category.

In the geospatial dataset used in this study (FHA, 2021) the sites are represented by point or polygon features, with the point features being used for the modeling. The use of point features contrasts with the site, by definition, being an areal feature, which raises some issues to attention. First is the question of the point location. Why is the point located where it is and what does it represent? The point may, depending on the site, be around the centroid of the polygon feature or at an arbitrary location, which may or may not represent the overall characteristics of the site. An entirely different problem is presented by the area features. An area boundary implies a difference between a site and its surroundings, while, in reality, there likely never has been a clear border (Barford 2000, p. 87).

2.1.2 Environment and the people

The connection between people and the environment is the driving force behind predictive modeling in archaeology. Verhagen (2007, p. 13) states that predictive modeling assumes that the location of the sites is not random but tied to certain environmental factors.

According to Salo (1995, p. 23) referencing Orrman (1991, pp. 4–7) Iron Age settlement in Southwest-Finland was located on lighter clay soils formed during the Litorina period rather than the heavier Ancylus- and Yoldia- clays. According to Saloranta (2000, p. 38), the most significant factors for selecting the sites sites seem to have been the composition of topsoil, slope of the terrain and angle, most likely referring to the aspect of the slope. Early and middle Iron Age sites were mostly located in southern or southwestern slopes on fine sand. The late Iron Age sites on the other hand were located on even ground on clayey soils (Leppänen & Mäkelä, 2022, p. 32).

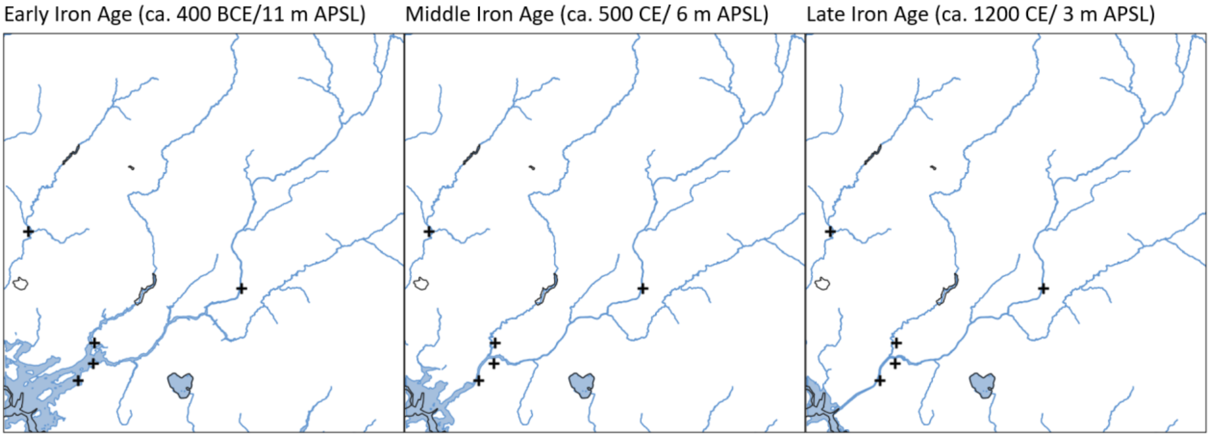

Early Iron Age (ca. 400 BCE/11 m APSL)
Middle Iron Age (ca. 500 CE/ 6 m APSL)
Late Iron Age (ca. 1200 CE/ 3 m APSL)

Figure 2. The change of the extent of water following land uplift from early, middle, and late Iron Age in the study area. Black line shows the modern coastline and black crosses locations of medieval stone churches.

A key factor driving natural change in the landscape has been land uplift (Figure 2), which in Turku area has been circa 3,8mm/a + 1,5 % / 100a (Kinnunen, 2019, pp. 127–130). While land uplift has played a key role in the formation of the landscape its effects during the Iron Age are mostly constrained to the proximity of the modern Turku urban area and with only limited influence on the rest of the study area. The people may have followed the water receding from the ascending land (Salo, 1995, p. 22), but the influence on the study area diminished during the middle and late Iron Ages. However, land uplift may give certain sites terminus post quem – datings (Seppälä, 2000, p. 195).

A commonly identified issue with studies researching relations between archaeological heritage and the environment is environmental determinism. The issue is strongly associated with the processual school of archaeology (Arponen et al., 2019, p. 2). The environmental determinism implies that human actions in pre-history were reactions to, or determined by, environmental factors, rather than to cultural processes. It has raised strong counterreactions over the past few decades and there are still antagonistic debates between the different schools of thought.

2.1.3 Communities and subsistence

The Iron Age was a very long period, and the communities and their subsistence strategies changed drastically over time. The Iron Age saw the emphasis on subsistence changing from hunter-gatherers towards agriculture and animal husbandry. The settlements changed from single farms early on to villages in the late Iron Age. All of this leads to the communal landscape of the late Iron Age being very different from that of the early.

The earliest iron age finds in the area date to the very beginning, in the turn from bronze age to pre-roman period (Saloranta, 2000, p. 15). However, the findings from that era have been scarce (Lehtonen, 2000, pp. 52–53). According to Lehtonen, the early settlement seems to have concentrated in Vähäjoki valley, where Saloranta identified two central settlement units, Kärsämäki and Saramäki during the early Iron Age. According to Lehtonen, the only reliably dated site from the period in Aurajoki valley is the Aittamäki burial ground at Vanhalinna, Lieto. During the early Iron Age, a single farm was the most common type of dwelling (Saloranta, 2000, p. 26).

During the Late Roman Period, the scarcity of finds in Aurajoki valley continues and the number of burials in Kärsämäki and Saramäki cemeteries decreases, which may suggest a decline in settlement of the river valleys (Lehtonen, 2000, p. 53). The decline continues until the migration period when the number of settlements drastically increases, and the focus seems to move from Vähäjoki valley to Aurajoki valley. During this period, the settlement of Aurajoki valley also almost reaches its late Iron Age extent (Salo, 1995, p. 19).

The earliest form of agriculture practiced in the area was based on mobile slash-and-burn cultivation (Saloranta, 2000, pp. 25–28). The early agriculture was small-scale, but around the third century CE sedentary form of slash-and-burn cultivation started being practiced, which lead to the landscape becoming more open due to the move towards extensive agriculture. Transition towards field cultivation began around the sixth to seventh century CE and was contemporary with the estimated inception of village settlement (Salo, 1995, pp. 28–31). From around the same period, there are also signs of settlement descending to the low-lying clay soils towards river (Saloranta, 2000, p. 38) as well as increasing population density. However, agriculture becoming more common does not necessarily mean every community based their living on it. For example, in the case of Kärsämäki, agriculture may not have originally been the primary means of living (Saloranta, 2000, pp. 38–39). For predictive modeling that relies on correlations between the sites and the environment multiple subsistence practices pose a challenge, as a broad group of sites may contain a number of subgroups, each with different needs.

2.2 Species distribution modeling

Species distribution models (SDM) are methods for predicting a species geographic distribution by estimating the relations between recorded species occurrence and environmental characteristics of their locations (Elith et al., 2011). They are commonly used methods in several disciplines such as analytical biology, ecology, evolution, and reserve planning (Phillips et al., 2006, pp. 231–232). Modern SDMs range from simple envelope models to advanced methods utilizing machine learning and artificial intelligence that combine ecological theory with advances in information technology and statistics (Elith & Leathwick, 2009).

SDMs have over the years been referred to by many terms, such as habitat suitability models (HSM), or environmental niche models (ENM). There have been arguments about the terms and situations they should be used in (e.g., Peterson & Soberón, 2012b). In this study, I will refer to these methods as SDM due to the relative neutrality of the term (Elith & Leathwick, 2009, p. 688).

An SDM study begins with model conception, where the basic idea of the model is set. The next step is collecting relevant occurrence and environmental data, which leads to the selection of modeling method or algorithm (Elith & Leathwick, 2009, p. 678). The model is then fitted to the training data and the results are evaluated after which the predictions are mapped. Finally, the process is iterated upon until the results are satisfactory.

2.2.1 Niche theory

The roots of species distribution modeling lie deep in the ecological niche theory, which means understanding the concept of niche is central to understanding the mechanics behind SDM. Ecological niche is a species position and match to specific environmental conditions in an ecosystem (Polechová & Storch, 2008, p. 1088). Since its inception, the niche has been a controversial concept and there have even been discussions whether the term should be discarded entirely (e.g., McInerny & Etienne, 2012).

The concept of ecological niche was first introduced by Grinnell (1917) as sets of environmental conditions under which a species can survive and reproduce. This description of the niche has become known as the fundamental niche. In 1927 Elton took a different approach to the concept by describing ecological niche as the role of a species in an environment. In Hutchinson later built upon this idea (1944, p. 20, footnote) by describing niche as "the sum of all environmental factors acting on the organism" and further defined it as "a region of an n-dimensional hyper-space, comparable to the phase-space of statistical mechanics."

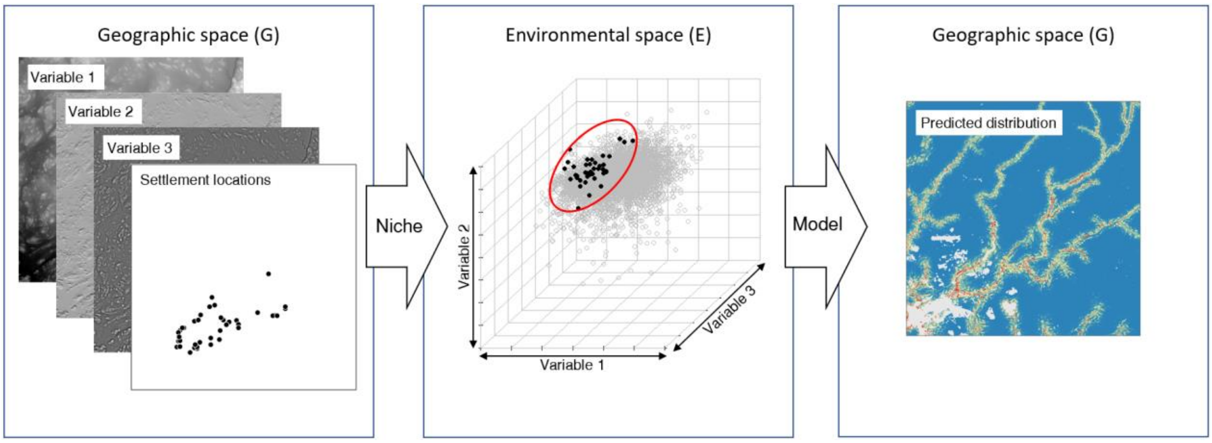


Figure 3. Visualization of the interplay between environmental space (E) and geographic space (G) in mapping species distributions.

Hutchinson later (1957) expanded upon this idea by presenting that the fundamental niche of a species can be represented in an n-dimensional hypervolume in which each dimension corresponds to an environmental factor, thus introducing two key concepts of SDM and modern niche theory: The abstract n-dimensional space defined by environmental parameters known as environmental space or E, and the two- or three-dimensional spatial space defined by map coordinates and elevation, known as the geographic space or G (Elith & Leathwick, 2009, pp. 681–683). Therefore, niche is an abstract construct that lies within E and distribution is niche projected into G (Figure 3).

SDMs fitted using only environmental predictors model the variance of occurrence in E (Elith & Leathwick, 2009, pp. 681–682). In effect, they are completely ignorant of the locations and distances within G. The results may look spatially informed, but this is a result of spatial autocorrelation in the environment, which means that locations close to each other have similar environmental properties. Spatial autocorrelation is crucial for the SDM predictions (Peterson & Soberon, 2012a, p. 794) and it also reduces the impact of locational uncertainty of occurrence data (Naimi et al., 2011, p. 1507).

Fundamental niche is not the appropriate term in all situations, as it covers the entire range of environmental conditions the species can inhabit on a permanent basis, while generally only a subset of it is actually inhabited. Thus, additional terms have been devised to cover other situations. Realized niche is the subset of fundamental niche the species actually inhabits (Polechová & Storch, 2008, pp. 1089–1090). Potential niche is the actual conditions of the fundamental niche existing within the relevant time and landscape (Jackson & Overpeck, 2000, p. 197).

To help describe impacts of combinations of factors on distribution Soberón and Peterson (2005) suggested a concept, which has become known as the BAM-diagram (Figure 4), standing for biotic, abiotic and movement (Soberón, 2010, pp. 160–161). BAM configuration of a model describes the types of predictor variables used in the prediction and thus the parts of the distribution modeled.

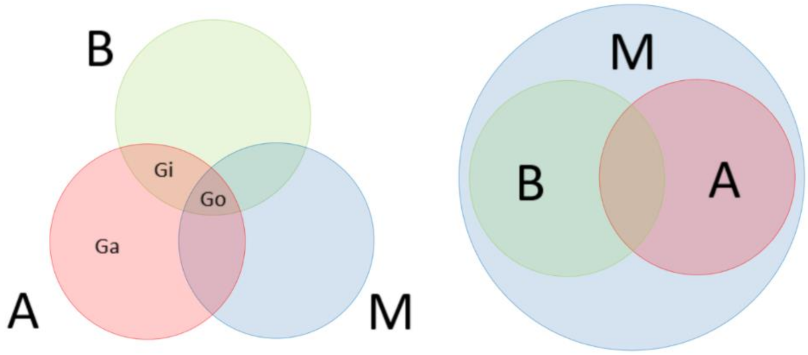


Figure 4. A BAM-diagram (left). The parts of the distribution modeled depend on the environmental variables used to describe the constraints within the landscape. The distribution is divided into abiotically suitable distribution (Ga), occupied distribution (Go), invadable distribution (Gi) and potential distribution (Gp = Go + Gi). In this study, the entire modeled area is considered accessible by humans and as such, movement is not a limiting factor (right).

2.2.2 Occurrence data and modeling methods

Occurrence data is a set of data consisting of recorded species presences and absences, as well as environmental background records. Presence data consists of sites where the presence of a species has been recorded and absence data where the species has been recorded absent. Background data, sometimes also referred to as pseudo-absence, are data that is used to capture the variability of environmental parameters in a landscape. The type of occurrence data is the single largest influence behind the selection of modeling method. The methods are divided by the type of data they use into three categories: presence-only, presence-background and presence-absence.

Presence-only and presence-background methods are the most common due to the type of data being readily available. Presence-absence methods on the other hand are somewhat rare because absence data is difficult to collect. Attitudes towards presence-only modeling vary wildly. Early SDM methods were restricted to envelopes and distance measures, (Elith & Leathwick, 2009, p. 689), but desire to use the widely available species records collected into herbarium and museum databases has spawned a number of more advanced methods that compare the presence records with background values (Elith et al., 2011, pp. 43–44). Presence-absence models are generally considered the most advanced, and many acknowledge that use of absence data would increase the robustness of their model (Elith & Leathwick, 2009, p. 689). However, some argue that absence data brings a whole new set of potential problems into the model (Ortega-Huerta & Peterson, 2008, p. 206).

A frequent problem with presence-only datasets is sample selection bias, in which case some areas have been sampled more intensively than others (Elith et al., 2011, p. 45). Most commonly available datasets have been collected opportunistically over long periods of time as is the case with cultural environment registers used in this study, which, as is common with archaeological inventory data (Yaworsky et al., 2020, p. 17), has been gathered over decades for cultural heritage management. Even when the presence-only data has been collected specifically for a study the survey is usually limited to relatively small areas. Furthermore, for example past disturbances may have caused local extinctions, which results in species being absent from an otherwise suitable environment (Elith et al., 2011, p. 45).

Modeling with absence data may help alleviate some of the limitations, however, they may come with their own set of uncertainties because absence data is difficult to verify (Ortega-Huerta & Peterson, 2008, p. 206). Absence records could be a result of a species being present but undetectable, absent from a suitable environment, or the environment being unsuitable. Unreliability of absence data has been a topic of discussion, which has caused arguments towards presence-only modeling as it diminishes the risk of unreliable absences (Elith et al., 2011, pp. 44–45). On the other hand, it has been suggested that when modeling with both presences and absences the biases of each dataset may cancel each other out. However, the biggest factor limiting the use of presence-absence models remains that true-absence data is difficult to come by.

Nowadays presence-background models are common as they offer advantages over presence-only methods while not being limited by availability of true-absence data. These methods use background data to characterize the environment (Phillips et al., 2009, pp. 182–183). The environmental characteristics are then compared against the species preferences set by the presence data. The background is usually sampled evenly across the entire environment. However, other approaches designed to account for potential sample selection bias exist (Phillips et al., 2009, p. 196).

2.2.3 MaxEnt algorithm

MaxEnt, short for Maximum Entropy, is the most widely used SDM algorithm (Feng et al., 2019, p. 10365), originally developed to account for the limitations in presence-only data modeling (Elith et al., 2011, pp. 43–44). MaxEnt is a presence-background machine learning method. The presence and background data are used to give the algorithm an example of the environmental conditions the studied species inhabits, after which it works in iterations improving the results with each, until it achieves the optimal results by reaching the convergence threshold, or when the maximum iterations limit is reached (Phillips et al., 2004, p. 85). MaxEnt predicts the species geographic distribution by finding the distribution that is closest to geographically uniform, i.e., having maximum entropy probability distribution, constrained by features derived from environmental conditions (Phillips et al., 2017, pp. 887–888). Predicting performance of MaxEnt has been found to be consistently competitive with high performing methods (e.g., Elith et al., 2006; Ortega-Huerta & Peterson, 2008; Yaworsky et al., 2020).

Phillips et al. (2006, pp. 234–235) list several advantages of MaxEnt method. They include the ability to work with presence-only data and to use both continuous and categorical predictors. It uses regularization to address some of the issues with overfitting and sampling bias. It is also an efficient and flexible general purpose statistical method with algorithms guaranteed to converge to optimal maximum entropy probability distribution. Phillips et al. also list some drawbacks, most of which have since been addressed.

Species response to the environmental factors is rarely linear (Austin 2002, p. 106). MaxEnt has the ability to model complex interactions between species and the environment. To do this, the algorithm uses transformation functions on the predictors to produce features in six different classes: linear, quadratic, product, threshold, hinge, and category indicator (Phillips et al., 2006, pp. 237–238). Linear features are continuous variables without transformations while quadratic features are the square of linear features. Product features are a product of two continuous variables and are used to model interactions between environmental factors. Threshold and hinge features are used to model arbitrary responses to predictors (Phillips & Dudik, 2008, p. 163). Category indicator features are used for categorical variables.

By default, MaxEnt automates the feature selection (Elith et al., 2011, p. 46), however, using all available feature classes may not always be desirable. Merow et al. (2013, pp. 6–7) argue that depending on the species' responses to environmental factors, linear and quadratic features may be sufficient. Allowing too much flexibility may introduce noise, which may be difficult to differentiate from actual responses. Preselecting the features may be desirable as it can lead to a more interpretable model, however, other school of thought prefers letting the algorithm identify the meaningful feature classes.

2.2.4 Environmental data

Environmental data, called predictors or covariates define the environmental constraints the predictions of are based on. In SDM they are factors affecting the suitability of the environment for the target species (Elith et al., 2011, p. 46). The predictors can vary highly, with some enabling larger ranges, while others may limit the presence of the species. As such, they should be relevant and offer a complete picture of the factors affecting upon the target species. The selection process has spawned some discussion on the topic. The BAM-diagram, for example, touched on in chapter 2.2.1, was originally developed to guide the process.

The environment is an infinitely complex system that is impossible to perfectly represent using a limited number of variables, and as such, there is no single best approach to modeling occurrence-environment interactions. Simple models usually link model structure with hypotheses (Merow et al., 2014, p. 1273). Simple models can avoid problems like overfitting, where the model becomes too closely aligned with training data, but they may become misleading by missing relevant parameters. Complex models are usually non-parametric in that they do not make presumptions on predictor effects. They may be better at capturing the complexity of nature, but they may lose some interpretability as important patterns or processes are lost under noise. As such, both approaches should be explored.

There are several schools of thought when it comes to the predictor selection process. One of them (e.g., Huntley et al., 2008) prefers preselecting a limited set of variables especially important to the distribution of target species. However, according to Peterson and Soberon (2012, p. 794) this practice runs the risk of missing critical variables or giving insufficient information to the model during calibration process. Mac Nally (2000, pp. 668–669) criticized this approach calling it "statistical tinkering" that could never substitute for an intelligent preselection of the predictors built upon the existing theory and knowledge.

Austin (1980, pp. 18–19) identified three idealized types of environmental gradients to be considered in the context of plant species distribution: indirect, direct and resource. Indirect gradients, such as elevation or aspect, do not directly affect the species themselves, but affect and are correlated with variables that do. Direct gradients have a direct physiological influence on the species' growth but are not consumed by them (Austin, 2002, p. 105) and may be such as solar irradiance. Finally, the resource gradients are variables that are essential for species growth through consumption. Examples of these may include nutriments and water. The previously presented categories are not exclusive, and the same predictor may fall under any category depending on the case. Additionally, the predictors can be considered either distal or proximal depending on their positions in the chain of processes affecting species distribution, with factors closest to the target species, i.e., the most proximal, being causal (Austin, 2002, pp. 105–106). Proximal predictors are the ones that influence the target species most and thus should be prioritized in variable selection.

A common issue in ecological modeling is the collinearity of predictor variables. Collinearity in statistical modeling refers to linear relation of predictors (Dormann et al., 2013, p. 28). According to Yaworsky et al. (2020, p. 7) it is often overlooked in predictive modeling. Collinearity is, to some level, intrinsic to all real-world data (Dormann et al., 2013, p. 28). In many cases, collinear variables are expressions of the same underlying process. Feng et al. (2019, p. 10370) mention that effects of predictor collinearity in MaxEnt models have not been well understood despite frequent mentions in literature. Strong predictor collinearity can lead to overestimation of the predictive power and decrease in interpretability of the model

(Yaworsky et al., 2020, p. 7). Because of this, it is generally considered good practice to limit the number of correlated variables. Feng et al. (2019, p. 10372) found that, while not completely immune, MaxEnt regularization can limit the effects of collinearity to the point where removing highly correlated variables does not drastically improve the results.

2.2.5 Model evaluation methods

Results are generally evaluated using statistical tests or data resampling depending on the aims of the study (Elith & Leathwick 2009, p. 691). The need for robust suite of methods is generally recognized. However, opinions on the important properties in a model and how to test them vary.

Common methods for evaluating model performance are cross-validations with statistically independent test data (Merow et al., 2013, pp. 9–10). Cross-validation is sometimes, although rarely (Elith & Leathwick, 2009, p. 691), done against a completely independent dataset, which may run the risk of including datapoints different in nature to those in the model training data. A more common practice, especially in machine learning studies, is to split the research data into separate training and test subsets. The model parameters are set, and the predictions made based on the training data and the test dataset is used to validate the final fit of the model. There are some caveats though. Firstly, the test dataset needs to be large enough to provide statistically significant results and secondly, it needs to be representative of the whole dataset. Thus, splitting the data may not always be feasible, especially in the cases where the full dataset is already small. An alternative approach available in MaxEnt is k-fold cross-validation where the data is split into k subsets known as folds. The model is then trained with k-1 folds while reserving the final subset for testing. The model is run k times until each of the folds has been used as a test dataset.

Thresholding creates a binary output, and involves selecting a break value, above which the species is considered present and below which absent (Merow et al., 2013, pp. 9–10). Threshold value can be based on statistics, such as minimum predicted value at presence location, or on arbitrary values like user specified omission rate. When run, MaxEnt calculates several threshold values with a variety of conditions that can be applied to the models at will. Though many use cases, such as conservation, require binary output, Merow et al. recommend against thresholding whenever possible.

Originally developed to avoid selecting a single threshold, Receiver Operator Characteristic (ROC) and especially the Area Under (ROC) Curve statistic (AUC) derived from it, has become perhaps the most popular performance metric in SDM studies (Merow et al., 2013, p. 9). ROC describes the model's ability to separate presences from absences. It is derived by calculating specifity, or the portion of absences correctly predicted as absence, and the sensitivity, or the portion of presences correctly predicted as presence, from confusion matrices (Jiménez et al., 2020, p. 1572).

AUC is based on the ranking of occurrence locations and represents the probability that randomly chosen presence site will be ranked above randomly chosen absence or background site (Phillips & Dudik, 2008, p. 166). A completely random ranking would have an AUC of 0.5 while a perfect ranking would achieve the best possible AUC of 1.0. Generally, values from 0.5 to 0.7 are considered poor, from 0.7 to 0.9 moderate and above 0.9 excellent model performance (Biodiversity and Climate Change Virtual Laboratory [BCCVL], 2021).

There are some caveats to using AUC as a performance metric. For example, in presence-background modeling high AUC values indicate that the model can distinguish between presence and the background (Merow et al., 2013). This may not be desirable as background can include both presences and absences, which can lead to misleading measures (see i.e., Lobo et al., 2008).

MaxEnt features several measures that can be used to evaluate the fit and performance of the model (see Phillips, 2017). By default, MaxEnt gives a percent contribution and permutation importance values for each variable. Additionally, an optional jackknife test can be used to measure the significance of variables. MaxEnt can also create response curves that show the shape of the species' response to each environmental variable.

2.2.6 Applications in archaeology

While predictive modeling is common practice in archaeological research, SDMs have been implemented relatively rarely, though the results so far have been promising. Yaworsky et al. (2020) assessed four widely used SDM algorithms in archaeological predictive modeling and found that especially MaxEnt performed surprisingly well with archaeological data. Issues with the other algorithms mostly came down to the limitations of the type of data available.

Yaworsky et al. (2020, pp. 2–3) also identified several problems common in predictive modeling practice in archaeology. Firstly, there were theoretical problems like failure to take land use decisions changing over time into account or selection of predictor variables without appropriate theoretical backing. The second category were empirical problems such as limited spatial resolution of the environmental data and failure to identify functional and temporal subsets within the occurrence data. Finally, there were analytical problems like inappropriate or –adequate use of statistical methods and limited consideration for model evaluation.

Many of the theoretical and empirical problems mentioned can be traced back to the quality and availability of appropriate data. Though presence data is widely available in databases collected by museums and other institutions, these data are primarily collected for heritage management purposes, due to which they may not be fit for research purposes as is. Some of the problems may be within the data structure or the sampling. Often the data has been collected opportunistically over long periods of time, which can lead to inconsistencies, and while presence data is readily available, absence is rarely recorded in the databases. This limits the application of statistical methods with model assumptions of true absences (Yaworsky et al., 2020, p. 14). Another problem is the amount of information per record, which in some cases can be very limited. Additionally, sample selection bias is a common problem with datasets collected outside targeted surveys.

Temporal perspective limits the use of some predictors, as some environmental variables like vegetation are more prone to changes than others. In an ideal situation the predictors would depict the landscape how it was during the period of interest. This, however, would require rigorous landscape reconstructions and studies of the processes and changes that have taken place over time (e.g., Franklin et al., 2015, pp. 2–9), which are outside the scope of this study.

Archaeological predictive modeling has often faced the question of environmental determinism, which understands the past human actions as a result of environmental rather than cultural factors. This problem has been a subject of an ongoing debate between points of view of natural sciences and human sciences (Arponen et al., 2019, pp. 2–4). A common critique of determinism is it being limited to an incomplete account of all the potential causal connections. However, limitations in the availability of types of data are often the reality of archaeological research. There are also philosophical problems with determinism. Namely, whether the physical world can be explained by physical elements exposed to us by scientific theories or not. According to Arponen et al., the question, if human sciences ultimately study causal, as opposed to emergent, phenomena in the vein of natural sciences or if both branches would be better off not dealing with causal phenomena at all has been a subject of a long-standing debate. Perhaps the only thing that is certain is that any scientific study, be it from natural or human scientific view, can only grasp a fraction of the infinitely complex physical environment.

The previously discussed principles for selecting environmental data by classifying it into indirect, direct and resource gradients (Austin, 1980, pp. 18–19 and 2002, p. 105) can be utilized to assess the importance of environmental factors in the context of human species. However, they may not be directly applicable in all situations. Saloranta (2000, p. 38) identified the composition of topsoil, and slope and angle of the terrain as likely decisive influences in the selection of dwelling site. There are several reasons to suggest they may not be that important after all. Firstly, none of the factors listed are resources, which would be, for example, nutrition and water, but topsoil class may be linked to the former. Additionally, topsoil and slope can in some cases have direct influence, for example in the case of very steep slopes and hard to work barren soils, but in less extreme values humans may be able to overcome the challenges posed by them. Finally, while the angle or aspect of the terrain may be connected or correlated with other environmental factors like solar irradiance, the direction of the slope is hardly an intrinsic value. As such, the commonly maintained belief in preference towards southerly slopes is likely a result of other factors and potentially confirmation bias. Tiilikkala (2017, p. 66) studied the connection between aspect and the Iron Age sites in Kanta-Häme province and found that while some directions are emphasized in the material, the same slope directions are also the most common in the whole region.

The previously mentioned topsoil, slope and aspect have been common variables used in archaeological predictive modeling. Additionally, in their models based on overlay analyses, Debenjak-Ijäs (2018, pp. 42–43) used elevation and distance to water and studied the influence of solar irradiance, and Tiilikkala (2017, pp. 82–85) additionally separated distance to water into separate layers for streams and lakes. Yaworsky et al. (2020, p. 12) tested various machine learning approaches, including MaxEnt, and used a set of 34 variables, from which they selected 10 for the final models. The final 10 variables included east-west and north-south aspect, watershed size, net primary productivity, growing degree-days, mean temperature and separate cost distance layers to springs, streams and wetlands.

2.3 Concept of scale

Scale refers to the spatial and temporal dimensions of a process or an object and is characterized by grain and extent (Turner et al., 2001, p. 29). Grain describes the properties of the data or the analysis, i.e., the grid cell size of the predictors or spatial accuracy of the occurrence data (Elith & Leathwick 2009, p. 680), whereas extent refers to the size of the study area and the temporal range of the study (Turner et al., 2001, p. 29). The grain size should be relevant to the studied species or phenomenon, while the extent usually reflects the aim of the study (Elith & Leathwick 2009, p. 680).

The scale of the study is linked to the hierarchy, or level of organization, as the subject of the study (Turner et al. 2001, p. 34). In context of humans these levels may be, for example, individual, group, community, etc. The processes important to study change with the scale of the study, as what affects the actions of an individual is different to what affects the community. The same goes for the processes themselves as their scale domains vary drastically. As presented by Pearson and Dawson (2003, pp. 368–369) land use, for example, is relevant between site and landscape scales while climate becomes significant at regional and larger scales.

SDMs generally require uniform scale across the datasets used. This presents a problem when working with datasets gathered through different methods. Topographic variables in the modern days are usually collected through remote sensing methods like lidar or SAR as continuous raster surface at set resolution. Climate on the other hand is collected from individual measuring points varying distance from each other and the continuous surface is acquired through interpolation. Unifying the scale of various datasets can lead to the situation where some layers are presented with higher precision than the underlying data. A potential solution to this problem may be found in hierarchical multi-scale models (Elith & Leathwick, 2009, p. 681). However, they remain relatively untested and as such it is unclear whether they offer clear advantages over well-structured non-hierarchical models.

The study of SDMs often intersects between ecology and geography, which may lead to some confusion with terminology (Turner et al., 2001, p. 30). Geographers usually use cartographic scale, where large-scale refers to fine and small-scale to coarse resolutions. Ecologists, meanwhile, when talking about scale, usually refer to the extent, where large-scale means large and small-scale small size of a study area. To avoid confusion Turner et al. recommend the use of the terms fine and broad.

3 Modeling process

From model conception to fitting, species distribution modeling includes several steps. This chapter includes the walkthrough of the process, from preparation of occurrence and predictor data to their evaluation, and to fitting and running the model and the evaluation of results. Data preparation and modeling process was done using various software disclosed in their respective chapters.

MaxEnt was selected as the modeling method due to its generally good performance with a variety of, including archaeological (e.g., Yaworsky et al., 2020, pp. 17–18), presence-only data. The models made in this study are based on the locations of the Iron Age settlements within the study area as recorded in the cultural environment registers of the Finnish Heritage Agency (FHA, 2021). Guidelines for good practices in species distribution modeling presented by Araújo et al. (2019) were considered during the process. The original guidelines have been designed for biodiversity research purposes and were implemented in this study where applicable. The process was based on the workflow by Hijmans & Elith (n.d.).

3.1 Data preparation

Environmental data is based on Digital Elevation Model 2 m (NLS, n.d. -a) and Topographic database (NLS, 2021), Superficial deposits (GTK, 2015) and Stream network dataset (SYKE, 2012). A unified spatial grain of 20 m was chosen for all variables and models in this study, as it was deemed a good balance between to limit the effects of small inaccuracies of the training site locations and of noisiness caused by small terrain variations while still offering relatively small-scale detail. In total 14 environmental layers for nine unique variables were selected for this study and are listed in appendix 3. The list of variables was based previous archaeological predictive models made in Finland (e.g., Tiilikkala, 2017; Debenjak-Ijäs, 2018).

Most of the environmental layers were based on digital elevation model (DEM), which was also tested as a predictor. However, due to absolute elevations tendency to bias the results towards certain regions (e.g., Debenjak-Ijäs, 2018, p. 82), two topographic position index (TPI) layers were created to provide a relative elevation alternative. The TPI layers were created with neighborhood sizes of 200 m and 2000 m to compare against near and far landscape respectively. Additionally, commonly studied slope, aspect, topsoil, and solar irradiance layers were created to study their impact on distribution. The topographic wetness index (TWI) layer was produced to assess the impact of hydrological processes. Finally, six distance to water layers were created; three with cost distance and three with Euclidean distance algorithms. The three layers per method used water levels of 11 m, 6 m, and 3 m above present sea level (APSL) to account for the impact of land uplift in early, middle, and late Iron Age respectively.

3.1.1 Occurrence data

The occurrence data used in this study was derived from cultural environment registers for research purposes data product acquired from service of Finnish Heritage Agency in 19.12.2021. The datafile used was the Muinaisjaannospisteet_t_point shapefile-layer, which was first reduced to the size of the AOI using the clip –tool in QGIS.

The sites within the AOI were selected with an SQL query filtering by keywords 'rautakautinen' (Iron Age) or 'moniperiodinen' (multiperiod) from 'ajoitus' (dating) column, and 'asuinpaikka' (settlement) from 'tyyppi' (type) column. The results of the query were evaluated on a case-by-case basis using survey reports to remove sites with disputable identification and false positives caused by multiple sites of various types and datings being presented by a single record. The final sample of training sites includes 39 probable Iron Age settlement locations listed in appendix 2. Because of the small number of records, the data is not split into separate training and test samples.

Additionally, a set of 10000 background sample points was created using Random points in extent –tool in QGIS. The points were sampled equally across the entire landscape, thus including the expectation that settlements may exist anywhere within the study area (Merow et al. 2013: 6).

According to Araújo et al. (2019, pp. 3–4) there are a few critical questions to consider about the extent, sampling, spatial accuracy, and identification of the response variable. The final data has several issues regarding these topics. While the settlements exist over a large area, the sampling has been done opportunistically and cumulatively over a long period. Thus, even within the relatively small extent of the study area there are some areas that have been surveyed more intensely than others, while some may remain completely unmapped, all of which can lead to sample selection bias affecting the results. However, MaxEnt is generally regarded as an algorithm that performs well even with incomplete datasets, which might somewhat alleviate the problem (Phillips et al., 2006, p. 234). In the case of some sites, specifically the ones pre-dating GNSS measurements, there may be fluctuations in spatial accuracy. Additionally, even after being evaluated with the reports, the identification of settlement is uncertain in the case of many sites, as most of them have only been superficially surveyed and the conditions during observations have often been poor.

The surveys have yielded rather limited information about most sites. In many cases the dating and type identification is based on a few scattered artefacts like pottery sherds. Because of this some of the sites may not be contemporary with the others and, as discussed by Seppälä (2000, pp. 194–195), some Iron Age and historical artefacts can be very similar, which can further the risk of incorrect dating. The scarcity of the early Iron Age finds in the area (Lehtonen, 2000, pp. 52–53) would suggest that most of the resulting sites are from latter parts of the Iron Age, which is, however, impossible to tell without further study. The sites with uncertain identification could also be removed, but this can lead to new problems by limiting the pool of available training data which runs the risk of introducing entirely new biases in the data.

Finally, MaxEnt uses point feature type training data, which may not be able to capture all the characteristics of the site. The sites can cover relatively large areas, while the point is located entirely inside a single raster grid cell, which may or may not be representative of the site overall. Points could be sampled into each raster cell within the defined settlement boundary. However, often the true extent of the settlement is unknown and furthermore, the increased number of closely clustered presence points may overemphasize certain environmental features.

3.1.2 Predictor variables

Most of the preparation of the environmental layers was done in QGIS and GRASS, with additional tasks being done in ArcGIS Pro. The layers were created in 2-meter spatial resolution with the exceptions of topsoil class and solar irradiance. After the layers were finished, they were resampled to final 20-meter spatial resolution used in the predictions. All layers were clipped to the extent of the AOI.

As many of the variables selected for this study are based on elevation data, the creation of variables was started by creating a merged digital elevation (DEM) model covering the entire study area using the GDAL merge -tool. In total sixteen 6 km * 6 km tiles of Elevation model 2 m -data were used. Next, aspect and slope layers were created based on the merged DEM using aspect and slope tools. Solar irradiance layer based of DEM was calculated using Area Solar Radiation (Spatial Analyst) -tool in ArcGIS Pro. The date interval was set to one year while other settings were kept at default.

Topsoil class variable was derived from surface soil type -column Superficial deposits 1:20000/1:50000 dataset from GTK, which describes the of topmost 40–90 cm soil layer (GTK, 2018). Check validity -tool was used to check the data for errors after which the data was run through the Fix geometries -tool. Afterwards the data was clipped to the extent of the DEM and converted to raster using GDAL rasterize (vector to raster) -tool with grain size set to 20 meters. The number of soil classes was reduced into fewer broader classes using r.reclass -tool in GRASS. The final number of classes was six with class with 10 being solid rock, 20 glacial till, 30 coarse-grained inorganic soil types, 40 fine sands, 41 clay soils and 50 peat and sludge soils. The remaining classifications, filling, unmapped and water, were classified as NULL as they were not expected to contain information related to site location.

Iron Age water levels were derived from the merged dem -raster. The raster was reclassified using r.reclass -tool into three layers using 3 m, 6 m and 11 m water levels corresponding to the water levels of late, middle, and early Iron Ages respectively (Kinnunen, 2019, p. 126). Areas that would have been below and above water level were classified as 0 and 1 respectively. The below water level areas were then polygonised with Polygonize (raster to vector) -tool. Thereafter the polygons were simplified with simplify -tool and further cleaned manually.

Distance to water layers were created by combining data from lake and waterway area shapefiles from topographic database (NLS, 2021) and waterways data from Finnish Environment Institute (SYKE, 2012). Data was compared to old aerial photographs to identify modern features and edited where deemed necessary. Modern reservoirs were removed, and their underlying waterways reconstructed based on the oldest available aerial photographs. Line features were polygonised using buffer -tool with 2-meter distance value. Finally, the layers were merged into three layers accounting for each water level using merge -tool. The resulting layers included Littoistenjärvi -lake from lakes -file, edited waterways from SYKE, waterway areas (NLS, 2021) connected to them and the respective water area polygon for each period. Three water layers corresponding to previous water levels were calculated in ArcGIS Pro using Cost Distance under Spatial Analyst Tools using the DEM as cost raster. Three additional layers were calculated using Euclidean Distance -tool to compare the results between the different methods. The result was six distance to water -layers; two for each of the three water levels corresponding to the two distance measures used.

The Topographic Wetness Index was calculated as a potential alternative to distance to water layers and was created following the workflow by Kopecký et al. (2021, p. 2). The first step was to transform the previously created slope layer from degrees to a new layer in radians. This was done by first making sure that there were no cells with slope of 0 degrees using the following equation:

$$(\mathrm{slope20m\_dg@1} \leq 0) * 1 + (\mathrm{slope20m\_dg@1} > 0) * \mathrm{slope20m\_dg@1}$$

Next, the layer values were multiplied by 0.01745 to convert the degrees into radians. Afterwards, the upslope contributing area was calculated from the dem using Flow Accumulation (qm of esp) -tool in SAGA toolbox in QGIS. Finally, the TWI layer was calculated using the equation:

$$\ln\big((\mathrm{upslope@1} * 20 * 20) / \tan(\mathrm{slope20m\_rad@1})\big)$$

The resulting layer is used to estimate water accumulation and is a function of slope and contributing area upstream. It can be used to determine how wet a certain place should topographically be (Grabs et al., 2009, pp. 15–16).

Topographic position index is an analysis method that compares the values of a central cell to average values of a neighborhood of a set size. Generally, this is used to find out the elevation of a cell compared to its environment to get its relative elevation. For this model, two TPI layers were created to describe the relative elevation using the Topographic position index (tpi) -tool in SAGA –toolbox. Neighborhood sizes were set to 200 meters, to account for near environment, and 2000 meters, for distant environment, with no distance weighing used.

3.2 Predictor evaluation

Quality evaluation was done using a number of methods. The assessment, where applicable, was based on data quality parameters as discussed by Veregin (1999). It was done to identify issues potentially affecting the results of the modeling. Additionally, the correlations between variables were tested through an analysis in R and the variables themselves as well as the model composition with tests available in the MaxEnt software (Phillips et al., 2022). The next chapters go through the quality analyses.

3.2.1 Assessment of quality of environmental layers

The preliminary quality assessment was done through visual inspection of the environmental layers and led to the conclusion that there are several sources of error in the variables, which are likely to affect the predictions. Firstly, the layers represent the modern environment, where roads spread across the landscape, dams have been built to create reservoirs and asphalt and concrete blanket large areas. Additionally, the elevation data is based on laser scanning data (NLS, 2016, p. 9) and as such does not fully represent the environment during period of interest as land use and geomorphological processes have affected the landscape. The changes should be relatively minor, but will affect the results of the model, nonetheless. While some steps were taken to improve the representation of certain variables, reconstruction of Iron Age landscape is outside of the scope of this study and as such, most of the modern features remain.

Secondly, each variable is an abstraction, or a simplified representation of a single factor of the environment, while on the whole a landscape is an infinitely complex entity. This means that even with a large set of predictors, some aspects, which may or may not be important for the modelled species, will inevitably be missed. There are also the possibilities of human biases affecting the variable selection process, which may lead to important variables being missed. In the end, the variables are selected by the researcher and represent their idea of what may or may not be important.

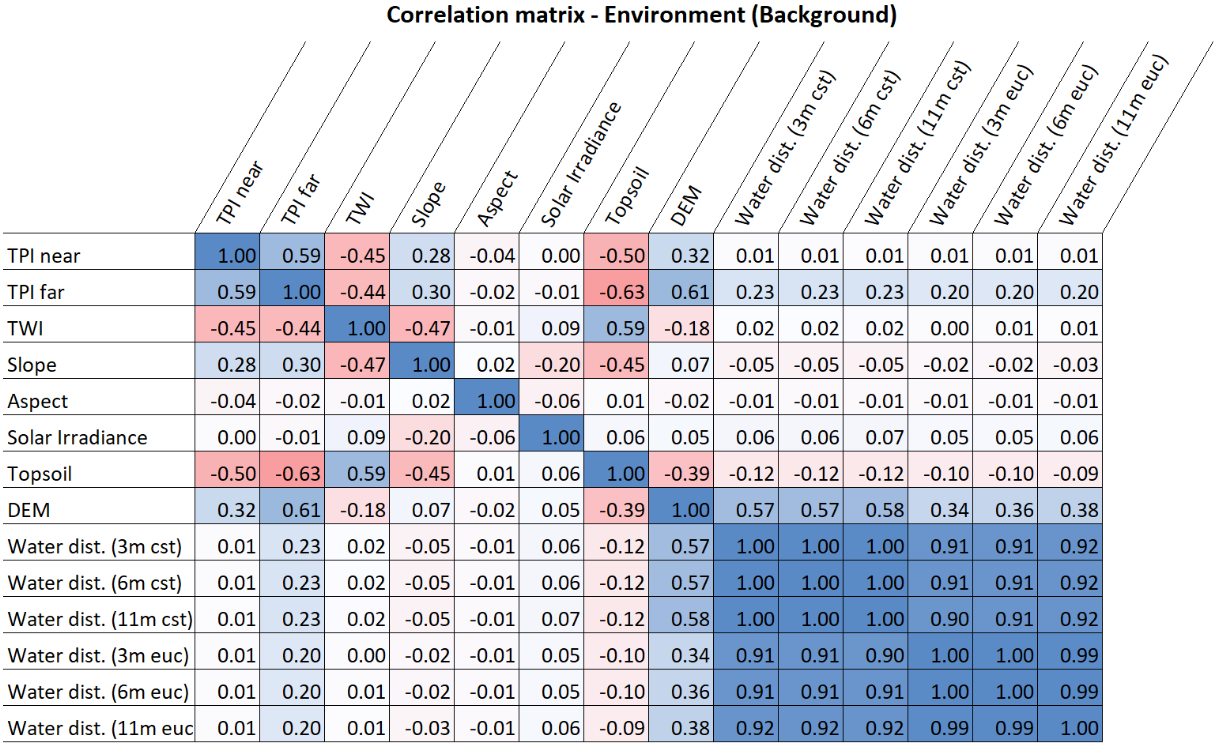
Additionally, while the TPI is consistent within its context, due to the way it is calculated, areas close to steep hills or cliff faces receive low values even if they otherwise would be suitable. For example, areas close to hillforts could be preferable to surrounding areas, but the TPI does not reflect this well. In the case of Vanhalinna hillfort finds that signal a presence of Iron Age settlement have been made below the western slopes (Lähdesmäki, 2000, p. 204). However, as this area receives low TPI values, it may be classified unsuitable in the predictions, depending on the significance of the variable.

3.2.2 Variable correlation

Because of the adverse effects strongly correlated variables can have on the models, it is deemed a good practice to review and limit the amount of covariation. For this purpose, a correlation matrix based on the values of the environmental layers at background locations was created (Table 1). An additional matrix was made using the values at presence locations

(Table 2) to see if connections between variables could be identified. When correlation value between two layers is positive their values are directly proportional and when negative, they are inversely proportional. In this chapter I will consider value ±0.5 a threshold of high correlation.

Table 1. Correlation matrix of all the predictors created. The table is based on the raster values at the background locations. Blue cells denote positive correlation and red cells negative. Darker color denotes stronger correlation.



Correlation matrix - Environment (Background)

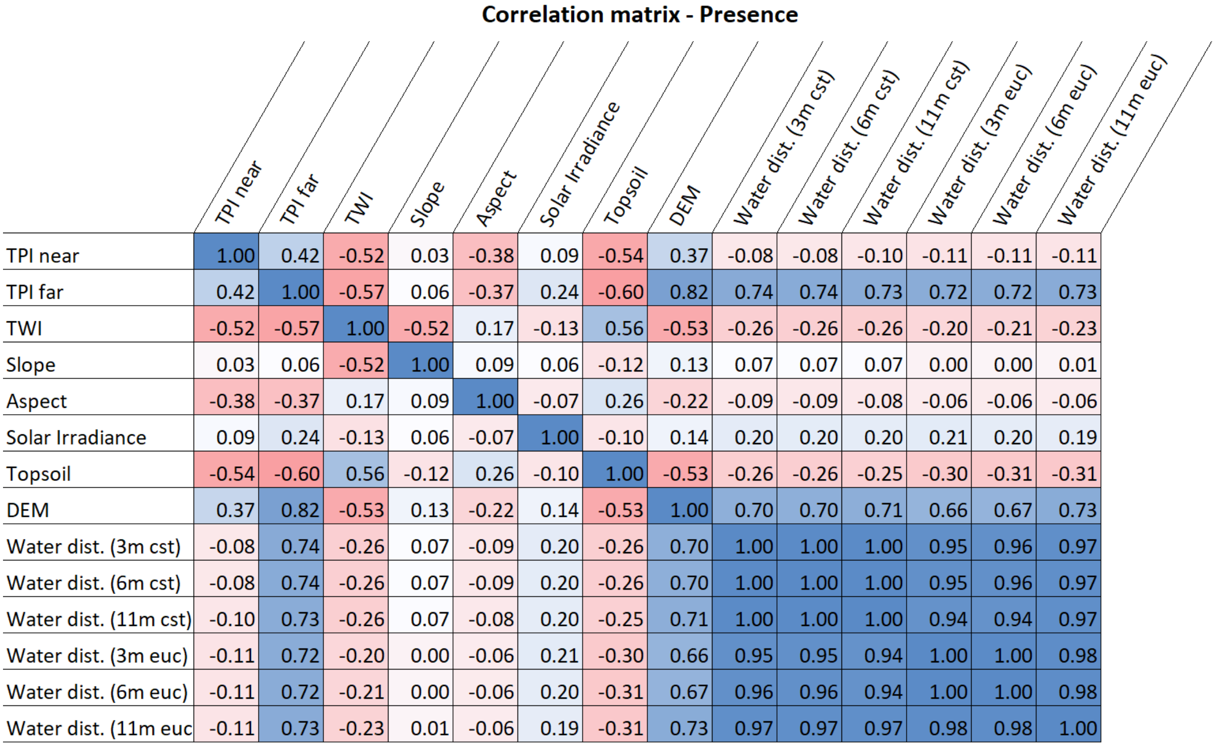
	TPI near	TPI far	TWI	Slope	Aspect	Solar Irradiance	Topsoil	DEM	Water dist. (3m cst)	Water dist. (6m cst)	Water dist. (11m cst)	Water dist. (3m euc)	Water dist. (6m euc)	Water dist. (11m euc)
TPI near	1.00	0.59	-0.45	0.28	-0.04	0.00	-0.50	0.32	0.01	0.01	0.01	0.01	0.01	0.01
TPI far	0.59	1.00	-0.44	0.30	-0.02	-0.01	-0.63	0.61	0.23	0.23	0.23	0.20	0.20	0.20
TWI	-0.45	-0.44	1.00	-0.47	-0.01	0.09	0.59	-0.18	0.02	0.02	0.02	0.00	0.01	0.01
Slope	0.28	0.30	-0.47	1.00	0.02	-0.20	-0.45	0.07	-0.05	-0.05	-0.05	-0.02	-0.02	-0.03
Aspect	-0.04	-0.02	-0.01	0.02	1.00	-0.06	0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Solar Irradiance	0.00	-0.01	0.09	-0.20	-0.06	1.00	0.06	0.05	0.06	0.06	0.07	0.05	0.05	0.06
Topsoil	-0.50	-0.63	0.59	-0.45	0.01	0.06	1.00	-0.39	-0.12	-0.12	-0.12	-0.10	-0.10	-0.09
DEM	0.32	0.61	-0.18	0.07	-0.02	0.05	-0.39	1.00	0.57	0.57	0.58	0.34	0.36	0.38
Water dist. (3m cst)	0.01	0.23	0.02	-0.05	-0.01	0.06	-0.12	0.57	1.00	1.00	1.00	0.91	0.91	0.92
Water dist. (6m cst)	0.01	0.23	0.02	-0.05	-0.01	0.06	-0.12	0.57	1.00	1.00	1.00	0.91	0.91	0.92
Water dist. (11m cst)	0.01	0.23	0.02	-0.05	-0.01	0.07	-0.12	0.58	1.00	1.00	1.00	0.90	0.91	0.92
Water dist. (3m euc)	0.01	0.20	0.00	-0.02	-0.01	0.05	-0.10	0.34	0.91	0.91	0.90	1.00	1.00	0.99
Water dist. (6m euc)	0.01	0.20	0.01	-0.02	-0.01	0.05	-0.10	0.36	0.91	0.91	0.91	1.00	1.00	0.99
Water dist. (11m euc	0.01	0.20	0.01	-0.03	-0.01	0.06	-0.09	0.38	0.92	0.92	0.92	0.99	0.99	1.00

Overall, the correlations between the variables, discounting the distance to water layers, are medium to low as seen in Table 1, and are either positive or negative based on the direction of the trend of values changing. Many of the variables include trends changing on the elevation axis. TPI layers, DEM and distance to water layer values tend to grow towards higher elevation, while in the case of TWI and topsoil class the opposite is true. When the trend is in the same direction, the layers receive positive correlation value and when in the opposite, negative. Aspect and solar irradiance do not have strong trends towards vertical directions and as such have generally low correlation values. Slope likely has some trends with steeper slopes in certain elevation ranges, which would explain the correlations with TPI layers.

Both TPI far and topsoil class have at or over ±0.5 correlation with three other layers: TPI far with TPI near, topsoil class and DEM, and topsoil class with both TPI layers and TWI. However, topsoil class is a categorical variable, which are treated differently by the model and as such, the correlation values are likely not comparable. TWI layer almost reaches the threshold of high correlation with the two TPI layers and slope layer. Slope, aspect, and solar irradiance do not have any correlations reaching the threshold. Finally, the DEM has over ±0.5 correlations with the TPI far and cost distance to water – layers.

Unsurprisingly, distance to water – layers correlate very strongly with each other. The values in the layers made using the same method change in proportion, due to which almost all of them get the strongest possible positive correlation values of 1 with each other. Out of the other layers, the distance to water –layers have strongest correlation values with DEM –layer. The layers calculated using the cost distance function have in this case stronger correlation as their value is based on the DEM used as cost raster while Euclidean distance only considers linear distance.

Table 2. Correlation matrix of predictor values at presence locations. The table is based on the raster values at presence locations. Blue cells denote positive correlation and red cells negative. Darker color denotes stronger correlation.



Correlation matrix - Presence

	TPI near	TPI far	TWI	Slope	Aspect	Solar Irradiance	Topsoil	DEM	Water dist. (3m cst)	Water dist. (6m cst)	Water dist. (11m cst)	Water dist. (3m euc)	Water dist. (6m euc)	Water dist. (11m euc)
TPI near	1.00	0.42	-0.52	0.03	-0.38	0.09	-0.54	0.37	-0.08	-0.08	-0.10	-0.11	-0.11	-0.11
TPI far	0.42	1.00	-0.57	0.06	-0.37	0.24	-0.60	0.82	0.74	0.74	0.73	0.72	0.72	0.73
TWI	-0.52	-0.57	1.00	-0.52	0.17	-0.13	0.56	-0.53	-0.26	-0.26	-0.26	-0.20	-0.21	-0.23
Slope	0.03	0.06	-0.52	1.00	0.09	0.06	-0.12	0.13	0.07	0.07	0.07	0.00	0.00	0.01
Aspect	-0.38	-0.37	0.17	0.09	1.00	-0.07	0.26	-0.22	-0.09	-0.09	-0.08	-0.06	-0.06	-0.06
Solar Irradiance	0.09	0.24	-0.13	0.06	-0.07	1.00	-0.10	0.14	0.20	0.20	0.20	0.21	0.20	0.19
Topsoil	-0.54	-0.60	0.56	-0.12	0.26	-0.10	1.00	-0.53	-0.26	-0.26	-0.25	-0.30	-0.31	-0.31
DEM	0.37	0.82	-0.53	0.13	-0.22	0.14	-0.53	1.00	0.70	0.70	0.71	0.66	0.67	0.73
Water dist. (3m cst)	-0.08	0.74	-0.26	0.07	-0.09	0.20	-0.26	0.70	1.00	1.00	1.00	0.95	0.96	0.97
Water dist. (6m cst)	-0.08	0.74	-0.26	0.07	-0.09	0.20	-0.26	0.70	1.00	1.00	1.00	0.95	0.96	0.97
Water dist. (11m cst)	-0.10	0.73	-0.26	0.07	-0.08	0.20	-0.25	0.71	1.00	1.00	1.00	0.94	0.94	0.97
Water dist. (3m euc)	-0.11	0.72	-0.20	0.00	-0.06	0.21	-0.30	0.66	0.95	0.95	0.94	1.00	1.00	0.98
Water dist. (6m euc)	-0.11	0.72	-0.21	0.00	-0.06	0.20	-0.31	0.67	0.96	0.96	0.94	1.00	1.00	0.98
Water dist. (11m euc	-0.11	0.73	-0.23	0.01	-0.06	0.19	-0.31	0.73	0.97	0.97	0.97	0.98	0.98	1.00

On the Iron Age settlements, the correlation values are overall stronger than in the background, though some of the differences can likely be explained as statistical anomalies caused by small sample size (Table 2). Some notable differences include slope having positive correlations with TPI layers and aspect having negative, TWI having strong negative correlation with DEM and distance to water layers having strong correlations with TPI far and stronger than previously correlations with each other.

In conclusion, there are some correlations between variables at around or higher than ±0.5 value. As such, they may have some effect on the models. Perhaps the most problematic correlations are between DEM and cost distance to water –layers, as the distance to water is likely a very important variable and having a variable highly correlated with it could emphasize it further and distort the results. As such, using the TPI as an alternative to DEM is preferred.

3.2.3 Environmental niches

Visualizing the occurrence locations in the environmental space can be useful in understanding the environmental conditions the species inhabits. The visualizations shown in this chapter were done by writing the predictor values at each presence and background location into a table, which was then visualized into three dimensional scatterplots. The data preparation was done in QGIS with point sampling tool –plugin and refined with Microsoft Office Excel. The scatterplot visualizations were done in R environment using functions in scatterplot3d –library. For purposes of clarity, the data was visualized in three separate scatterplots, each showing a space described by three variables. Only one of the six distance to water –layers was used, because the layers are very similar and visualizing all of them was not expected to carry additional value.

The first scatterplot (Figure 5) visualizes a space defined by topsoil class, cost distance to water and topographic wetness index. Due to the presence of categorical values in the plot, the data is formed into clearly defined clusters, each corresponding to a different soil class. Training sites appear in classes 10, 20, 30 and 41, corresponding to solid rock, glacial till, coarse-grained soils, and clay soils. Classes 40 and 50, corresponding to fine sands and peat and sludge soils do not contain any sites. The appearance of the sites is, overall, roughly proportional to the prevalence of the soil class within the landscape. Most of the sites are located on clay soils with solid rock being the second most common class. Coarse-grained soils class contains five presences and glacial till only one.

The preferences are more clearly visible on the TWI axis than the topsoil class. Most of the sites are located at places with TWI values from around 14 to 19 with median value being 15.5 and average 15.9. Raster cells with TWI values of over 20 are mostly confined to river channels or areas with natural streams. Additionally, some of the cells are on what now are open fields, which, judging by the characteristics of the rest of the areas with high TWI values, would likely be too wet for long term settlement. Areas with sub 15 values on the other hand, are generally high in the environment and often further away from water sources, which would likely make them somewhat undesirable.

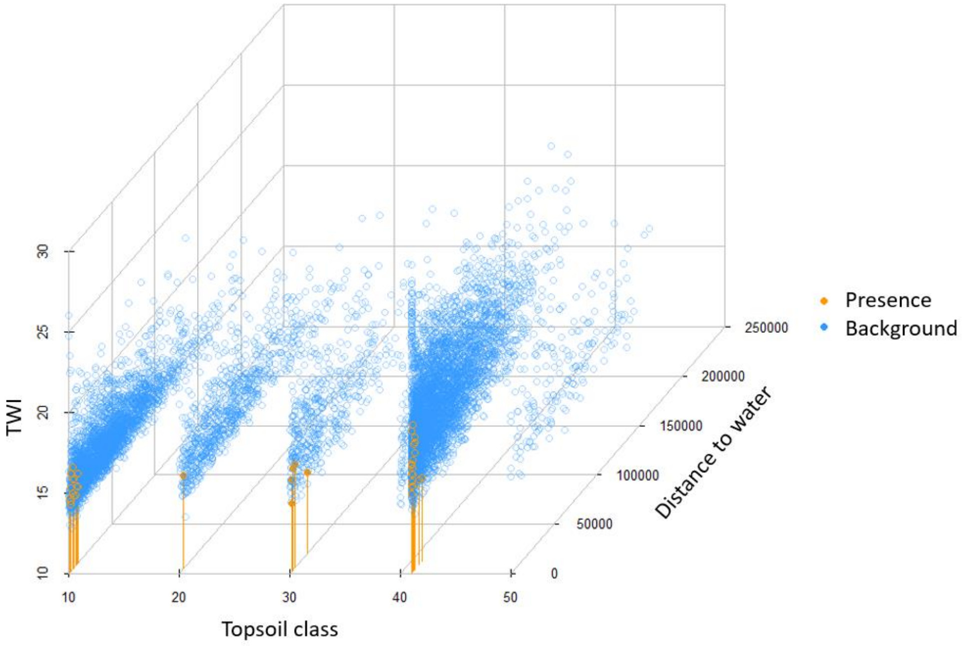


Figure 5. Locations of known archaeological sites in an environmental space defined by topsoil class (x), cost distance to water 3 m APSL (y) and TWI (z).

Distance to water shows by far the strongest biases out of the three, with settlements tending to be close-by to a water source. The cost units used in the visualized layer do not translate linearly to Euclidean distance and thus the relation varies on a case-by-case basis. An average settlement is at 4146 cost units or 220 meters, median values being 2425 cost units and 170 meters, from water, while the settlement furthest away is at around 20500 cost units distance, which translates to 694 meters Euclidean distance. The most distant point from water in the landscape is at around 215000 cost units distance, which is roughly 3300 meters. By Euclidean distance, the furthest away location in the entire landscape is 4769 meters from water. Based on this, only a very narrow band close to water sources is inhabited by the training sites, which, more than likely, will also be reflected in the predicted distribution.

The second scatterplot (Figure 6) is defined by solar irradiance, slope and aspect, and shows both the background and the presence locations scattered across the axes in a crescent shape, formed by the connection between solar irradiance and aspect. However, some potential preferences can be established. Most of the settlements are located on gentle, sub –10° slopes.

There are a few outliers on 10° –16° and one site on a 22° slope. Compared to all the values existing within the landscape, the preferences seem to be towards gentle slopes over flat land or steep slopes. These areas are likely less prone to water stagnation while still being practical for settlement. The median slope at a settlement location is 4.2° compared to 2.3° in the entire landscape. The average values are closer to each other at 5.3° and 4.3° respectively.

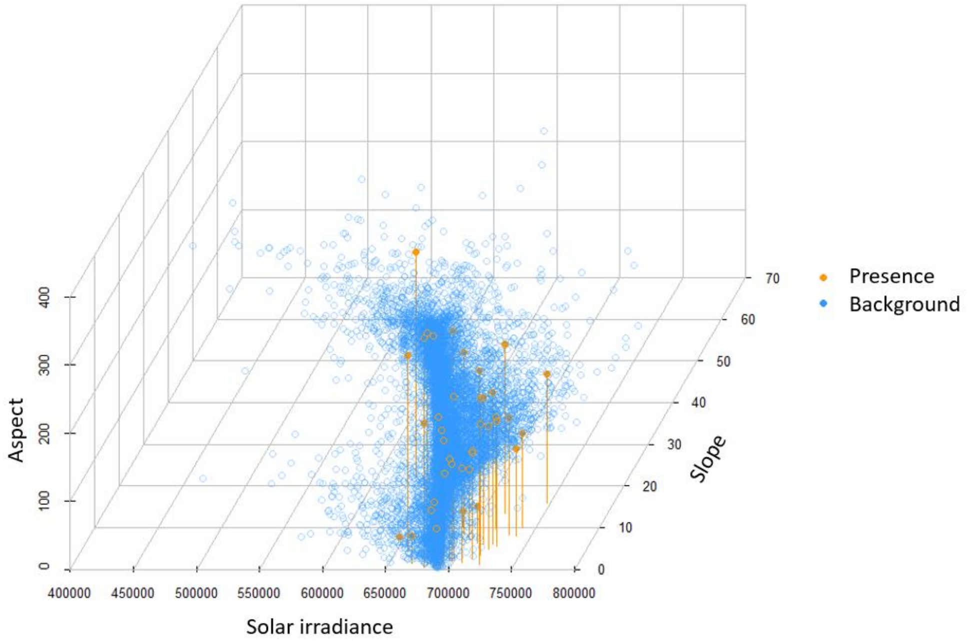


Figure 6. Scatterplot describing the settlement locations in an environmental space defined by Solar irradiance (x), slope (y) and aspect (z).

Settlements are scattered across all slope directions fairly evenly with slight inclination towards southerly slopes, which is also the most common slope direction in the landscape on whole. Compared to background values, the presences appear to have slightly stronger bias towards south, appearing to indicate slight preference. However, this potential preference is so small that it is likely to not be of much significance in the final predictions.

Settlement locations tend to receive slightly more solar irradiance than the environment on average; around 700000 WH/m^2 compared to around 689000 WH/m^2. The amount of solar radiation received by settlements ranges from 631000 WH/m^2 to 747000 WH/m^2, and the environment from 448000 WH/m^2 to 798000 WH/m^2. This shows modest preference towards areas receiving more than average solar radiation.

The third scatterplot (Figure 7) visualizes the site locations within a space defined by the two TPI layers and the DEM layer, and shows the background points forming an oblong shape, which is a result of covariation of the environmental factors. The settlements appear in the top half of the shape, mainly towards the end closer to the origin of the plot. The cluster is fairly well defined, which would point towards all the variables including some useful information.

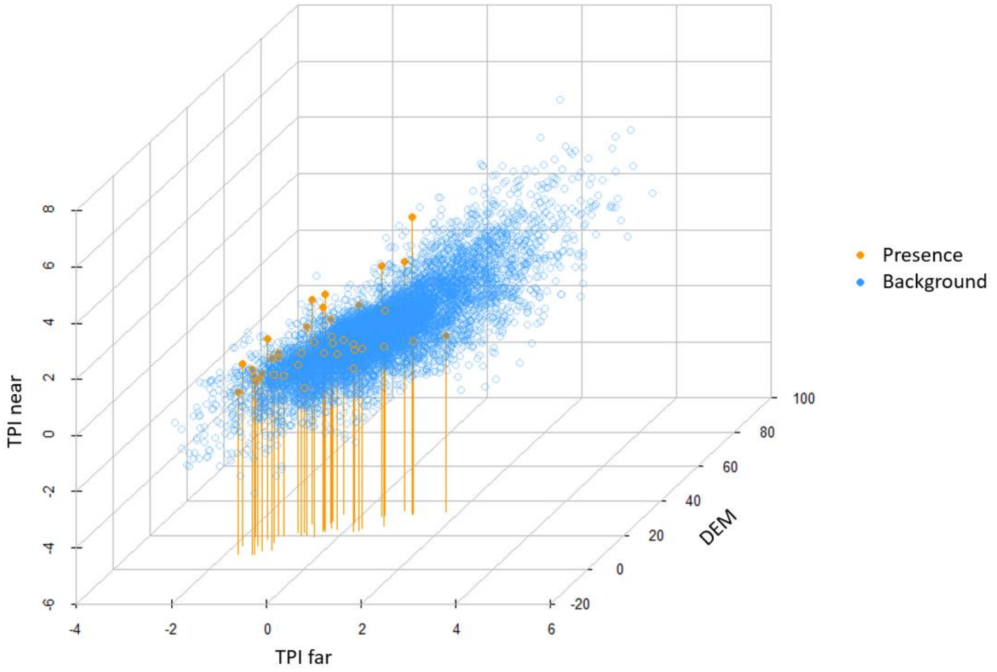


Figure 7. Iron Age settlements in an environmental space defined by relative and absolute elevation; TPI far (x), DEM (y) and TPI near (z).

The absolute elevation of the landscape, described by DEM, ranges from 0 m to 87 m APSL and averages to 40 m APSL. The training sites are contained within a range from 8 m to 35 m APSL with both an average and median elevation of 22 m APSL. Similar mean and median values indicate that the elevation values at presence locations are symmetrically distributed around 22 m APSL.

TPI near values in landscape range from -4.33 to 6.68 with an average of 0 and median of -0.12, compared to from -0.85 to 4.61 with an average of 0.78 and median of 0.40 on settlement locations. The difference between the mean and median indicates that the distribution of TPI near values on sites is skewed to the left, with most of the sites being in locations receiving values between 0.2 and 1.3 and a few being in locations receiving lower or higher values. Two of the sites are outliers on the TPI near axis with values of 4.61 and 2.92 appear. The values of TPI far on the other hand range from -3.56 to 5.33 with an average of 0 and median of -0.16 across the entire landscape and from -1.82 to 1.70 with an average and median of -0.49 on the settlements. The average and median of settlements once again indicate close to symmetrical distribution of values with most of the sites receiving values between -1.3 and 0.2.

3.2.4 Predictor responses

Response curves were created by running a model with all 14 environmental layers and all available features. The response curves presented in Figure 8, were created by modeling with each variable in isolation. Thus, effects of covariation are not visible in them. Only one of the six distance to water –layers is presented, because the rest of the layers show a similar response pattern on dry land. The layers using 11m APSL as the base water level show an increased response at sub-zero distances, but this is an anomaly resulting from several of the sites being underwater during the early Iron Age.

Topsoil class shows increased response of 0.87 at class 30, or coarse-grained soils and almost flat response of around 0.60 across classes 10, 40, 41 and 50. Class 20, corresponding to glacial till shows decreased response of 0.30. The topographic wetness index gets its highest response of 0.8 at the value of 14. At values under 14 the response is constantly high at almost 0.8, while from around 15 to 21 the response decreases to 0.1 before reaching 0 at around the value of 25. The response of distance to water is at its highest value of almost 1.0 at 0 distance and rapidly decreases to 0.1 at around 1500 cost units distance before reaching zero at 4000 cost units.

The response to solar irradiance starts at around 0.42 at the value of 450000 WH/m2 and increases slowly to 0.50 at the environmental average of 689000 WH/m2. After that it increases rapidly to 0.87 at 720000 WH/m2, turning into a curve that reaches the response of 1.0 at around 800 000 WH/m2. The response curve of the aspect is relatively flat with a slight increase from 0.55 to 0.67 between 10° and 140°, flat response of 0.67 from 140° to 220° and decrease to 0.62 before 330°. The response at 0° is 0.75 and at 360° it is 0.26, even though they are both north. This is likely a result of the model lacking the preceding information and considering the variable as linear instead of circular. The highest response to slope variable, 0.75, is at values between 5° and 10°. From 5° to flat land the response decreases sharply to 0.32 and from 10° onwards slowly, until reaching 0.04 at around 60°.

TPI far starts at a high response of around 0.9 until the value -1.7. At the higher values, the response decreases to 0.5 by the value of 0 and finally 0.1 at the value of 4.7. TPI near on the other hand starts low at 0.0 before slowly starting to increase at values above -3 and once again reaching 0.5 by the value of 0. The curve flattens at the response of 0.78 and TPI of 0.4, after which it smoothly increases to 0.93 at around TPI 7. DEM variable shows a high response of 1.0 at values below 10 m APSL, after which the response gradually decreases to 0.0 at around 55 m APSL.

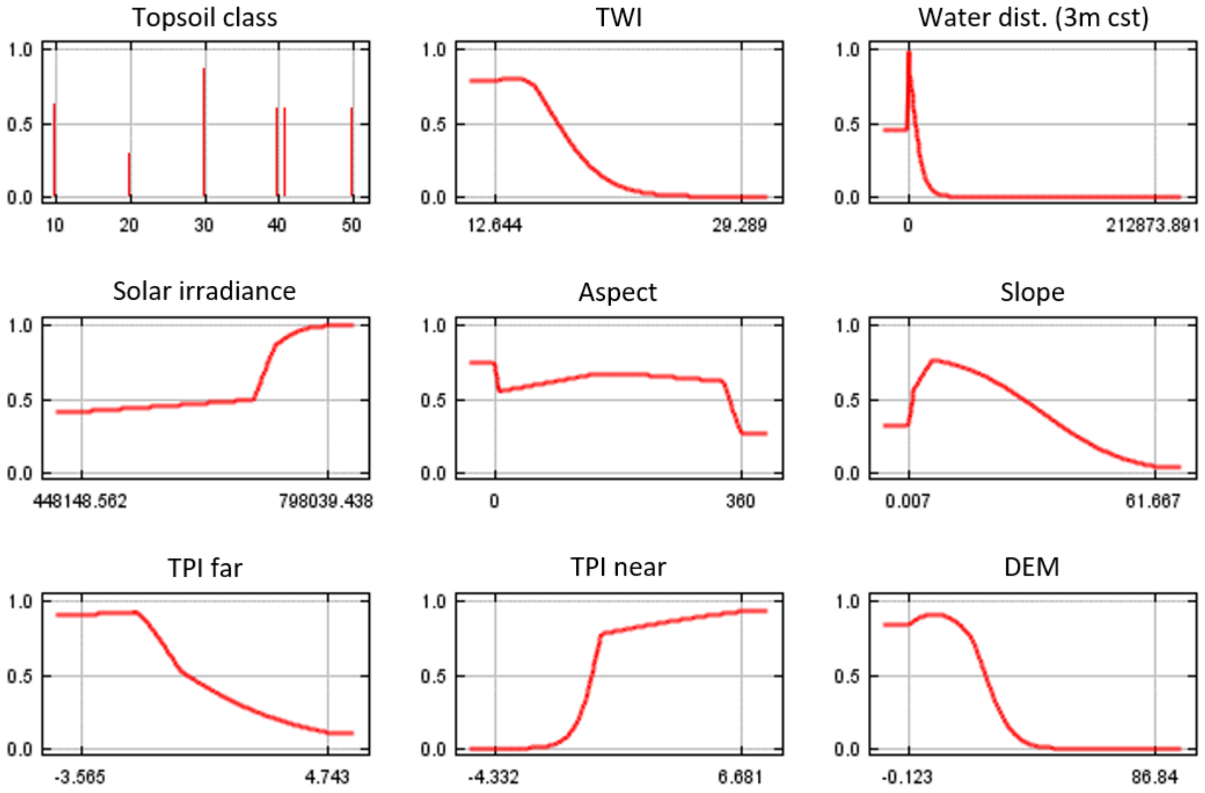


Figure 8.Response curves of a test model generated by MaxEnt.

3.2.5 Variable significance

Variable significance was estimated based on jackknife test available in the MaxEnt software. In jackknife test, models are created by excluding each variable in turn, after which each variable is modeled separately (Phillips, 2005, pp. 243–246). Finally, a model is created using all variables. The first set of models, represented by light blue bars in Figure 9, give us the loss of training gain when variable is not used and thus an estimate of how much information, not present in other variables, a layer contains. The second set, represented by the darker blue bars, gives us the training gain when a variable is used in isolation, and therefore an estimate of how much useful information it contains.

The environmental variable that achieves the highest gain when used in isolation is the cost distance to water. The training gain is barely reduced when the variable is excluded, which is most likely a result of the Euclidean distance to water variable being present in the test as well. The second highest gain is achieved by DEM, which also results in the largest loss of gain when excluded. The previously mentioned Euclidean distance to water has the third highest, and interestingly, its exclusion reduces the gain slightly more than the exclusion of cost distance layer. The fourth and fifth in gain are TPI near and TWI layers, which also show a fairly sizable reduction in gain when excluded, indicating they contain information not present in the other layers.

The rest of the layers achieve relatively low gain when modeled in isolation. The highest among them is TPI far, followed by solar irradiance, slope, topsoil class and aspect, respectively. The aspect achieves hardly any gain on its own and excluding TPI far or aspect results in barely any reduction in gain. Removing slope results in some loss of gain, and topsoil class or solar irradiance in moderate amount.

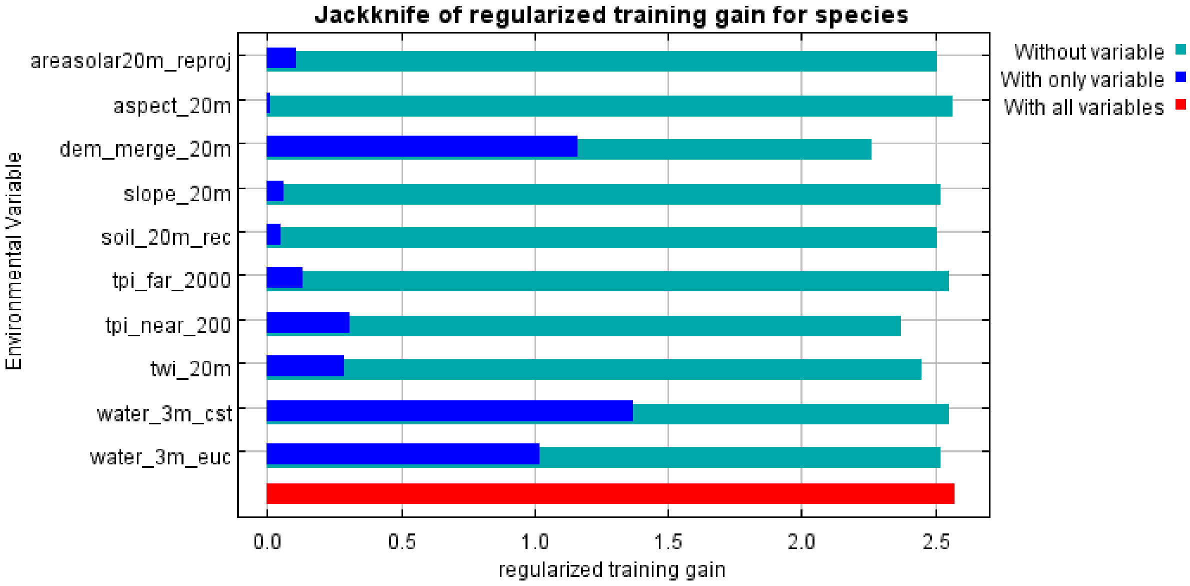


Figure 9. Results of the jackknife test of variable importance. Dark blue bars represent the gain when a variable is used in isolation, while the light blue bars show the gain when the variable is excluded. The red bar shows the gain when modeling with all listed variables.

3.2.6 Summary

Overall, the distance to water, which is also the only predictor present that could be considered a resource gradient, appears to be the most influential factor, and is followed by absolute elevation. Unlike distance to water, absolute elevation is most likely not a direct influence, but a factor affecting and correlating with a number of other, more functional influences. Additionally, the use of absolute elevation limits the predictions to a certain elevation range, which can leave otherwise suitable areas outside the predicted suitability (Debenjak-Ijäs, 2018, p. 82). Because of this, the variable is avoided in the final models, unless a model is created to specifically study its effects.

TPI near and TWI are moderately important. TPI near indicates that the settlements are in locally prominent places, such as above riverbanks. TPI far has more limited influence, as it characterizes the site compared to the distant environment, with not as immediate influence on the site. TWI characterizes the environment based on the wetness each cell should topographically have and the test data shows the settlements favoring the drier cells, which may be a question of living comfort.

According to Saloranta (2000, p. 38), topsoil composition, slope and aspect were likely the decisive factors when selecting a settlement site. However, based on the jackknife test (Figure 9) they appear to be the three least influential factors out of all the factors used in this study, with especially aspect bearing next to no significance. Saloranta further specifies that the sandy southern or southwesterly slopes were favored during the early and middle Iron Ages, while the late Iron Age settlements favored the heavy, fertile clayey soils. Due to the limitations of the research data, studying the differences between the early, middle, and late Iron Age sites is outside the scope of this study. The test data indicates some preference towards coarse-grained soils and 5°–10° slopes, but their influence needs to be studied further before conclusions can be drawn. Solar irradiance may have some influence, as the settlements are shown to have higher values than the global average and it may also work as a more functional replacement for aspect.

3.3 Model fitting and predictions

The open source MaxEnt software (Phillips et al., n.d.) was used to create the models. The software was run in R environment using an interfacing function in library dismo. The occurrence data was set as the response variable and the environmental layers as predictor variables. The final MaxEnt settings were determined with a series of test runs. The list of models discussed in the text is available in appendix 4.

To allow complex interactions between occurrence and environmental variables to be modeled, all features, linear, quadratic, product, threshold, and hinge as well as category indicator were allowed. Where applicable, the topsoil class variable was added to the predictor stack as a categorical variable, or as R calls them, a factor. Maximum iterations parameter was increased from default 500 to 5000 to always allow the algorithm to reach the convergence threshold, which was kept at its default value of 0.001 %.

Due to the limited number of sites, it was decided not to split the data into training and test sets, and instead use the entire set for training. This limits the validation options, but also removes the variation in the results caused by the training sample selection. Instead, k-fold cross-validation with five folds was run before each model to test the robustness of model configurations. The cross-validated run was reviewed before the final model run.

Additionally, response curves were generated for each model. Each of the five replicates, as well as an averaged one, was written on the disk in GEOTIFF-format in addition to the final model. Based on the test runs and data evaluation the following configurations were selected for testing:

1. Two models were created to compare simple and complex predictor configurations. The complex model used the full complement of variables, while the simple model reduced complexity by removing three variables found least significant in previous testing.

2. The differences in the distance measures and base water levels were tested by creating six models using a different distance to water layers. Additionally, a model using no distance to water information was created.

3. Six models were created to test the difference in results when different configurations of absolute and relative elevation information.

3.3.1 Simplicity versus complexity

The differences between complex and simple predictor configurations were tested by creating two models. Both models used solar irradiance, TPI far, TPI near and TWI cost distance to water with 3-meter base water level as predictors. Additionally, the complex model included aspect, slope, and topsoil class variables. The AUC score of the complex model was slightly higher, 0.958, than the 0.955 of the simple model. Thus, both models performed excellently based on statistical testing, with a small, inconsequential difference.

The response curves of the three variables not present in the simple model were reviewed. The response of the aspect variable is almost completely flat except for values close to 0° and 360°, both corresponding to north. Similarly, the responses to the various topsoil classes were almost flat, with the only exception being class 30 corresponding to coarse-grained soils.

Finally, the response of slope formed an s-curve beginning high on flat land and reaching close to zero gain at around 30° slope.

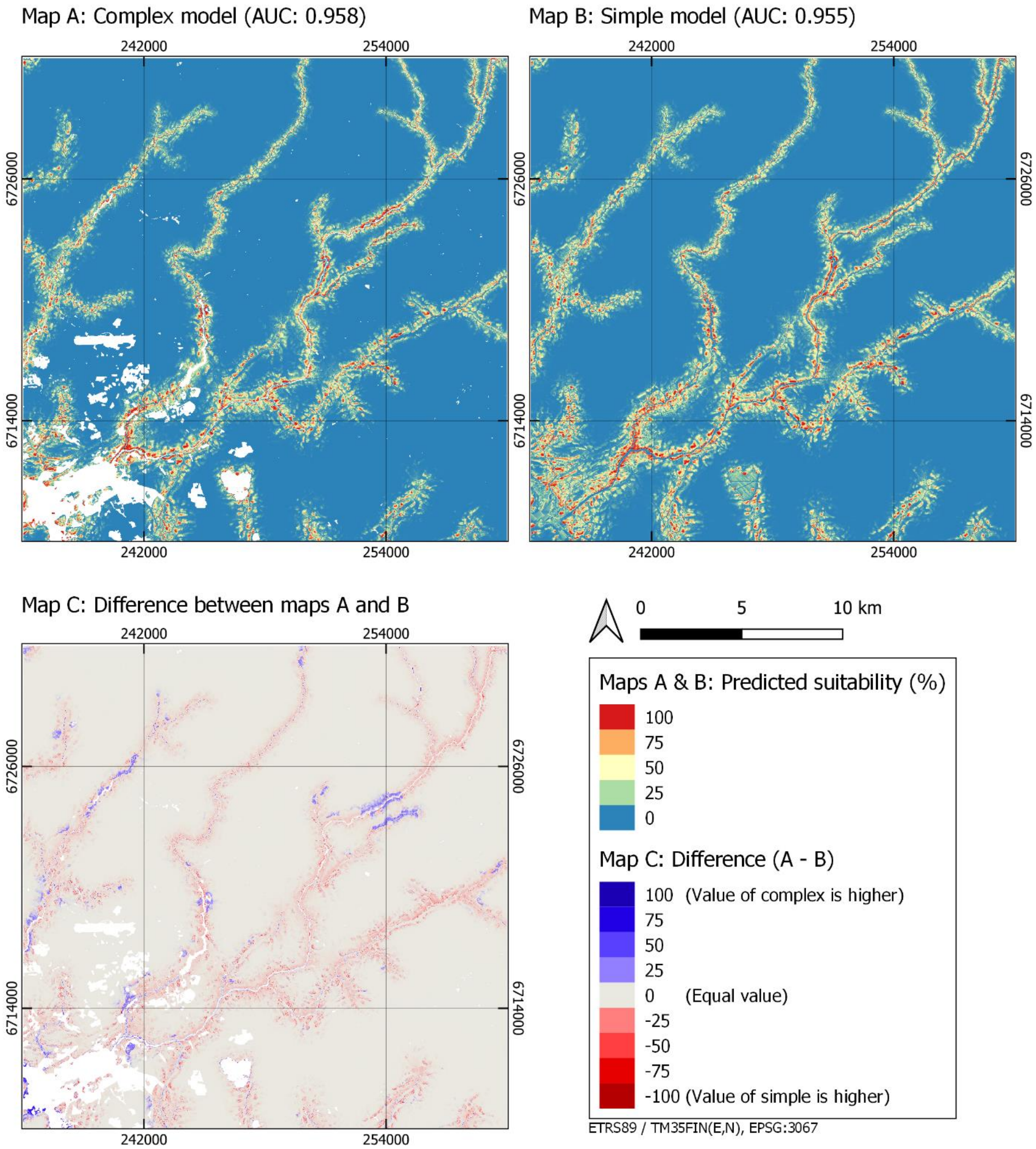

Map A: Complex model (AUC: 0.958)
Map B: Simple model (AUC: 0.955)
Map C: Difference between maps A and B
242000
254000
6726000
6714000
0
5
10 km
Maps A & B: Predicted suitability (%)
100
75
50
25
0
Map C: Difference (A - B)
100 (Value of complex is higher)
75
50
25
0 (Equal value)
-25
-50
-75
-100 (Value of simple is higher)
ETRS89 / TM35FIN(E,N), EPSG:3067

Figure 10. Comparison of complex and simple model configurations.

Overall, the distributions of both models follow the river network tightly. The most suitable areas, characterized by wide predicted distribution area around the rivers, are located upriver from the confluence of Aurajoki and Vähäjoki. On the banks of Aurajoki, they continue until around Lieto central conurbation, where the river turns towards north. On Vähäjoki the most suitable areas continue until around Jäkärlä. Finally, on the banks of Savijoki the distribution pattern is widest from the confluence to the Ankka and Raukkala area. Further upriver in all river valleys, the modeled distribution follows the rivers more tightly to the edge of the study area.

The differences between the two models are largely superficial. The complex model (Figure 10, map A) is noticeably noisy, while the simple model (map B) is comparatively smooth. At the edge of the river channel, the simple model shows slightly stronger predicted suitability. Additionally, the simple model includes distributions for built up areas, which are absent from the complex model, due to them being classified as NULL in the topsoil class layer.

For the most part the distribution patterns are very similar, but there are a few places with larger differences where, with one exception, the complex model shows a wider or stronger distribution area (Figure 10, map C). On the banks of Aurajoki there are two places: one in the middle of the Nummi district in Turku, and the other at Leinakkalankoski on the border of municipalities of Lieto and Aura. Additionally, on the north bank of Vuohenoja, a tributary of Aurajoki, around 800 meters south of the previously mentioned site, is an area with similar distribution pattern. In Vähäjoki valley, the situation is similar, with two major areas where the complex model shows wider distribution. The first one is located in Kärsämäki, between the Turku-Toijala railway and Kärsämäentie road. The second is further upriver, around the Jäkärlä brick factory area. Finally, the exception, where the simple model shows a wider distribution pattern, is on the northern bank of Savijoki between Pränikkälä and Pokkola.

Removing the three least significant variables has only a minor impact on the predictions. The largest visible difference between the models is that the simple model includes predictions for the areas where the topsoil is classified as NULL. Statistically the predictions of both models are almost equally good, as the simplified model saw only 0.003 decrease in AUC score.

3.3.2 Distance to water

To compare, how different base water levels and distance measures change the model predictions, seven models were created: six using different distance to water layers (Figure 11. maps A–F) and one not using the variable (map G). The other variables used in the models were solar irradiance, topsoil class, TPI far, TPI near and TWI variables.

AUC values for models using cost distance layers are very high and showed very little variance between different layers. They ranged from 0.955 to 0.957 AUC from the highest base water level (11 m) to the lowest (3 m). Similarly, although slightly lower, the AUC scores of models using Euclidean distance layers vary very little, as they range from 0.941, in the case of the 6 m base water level, to 0.942, in the case of both 11 m and 3 m base water levels. Finally, the AUC score of the model without the variable is 0.898, which can still be considered high level of performance.

The layers based on Euclidean distance (Figure 11, maps B, D and E) have a wider distribution pattern around rivers than the layers based on cost distance (maps A, C and F). This is more apparent further inland, potentially due to steeper banks and deeper river channels, which would increase the cost associated with the raster grid cells close to river. As expected, the distribution patterns follow the rivers closely, which interestingly, for the most part, is the case even when no distance to water information was provided (map G). Therefore, some of the other variables also have their optimum values located close to the rivers. Solar irradiance and topsoil class with their low significance are unlikely candidates for steering the predictions towards these areas, making the combination of the TPI layers and TWI more likely. TWI has a trend of higher values closer to rivers and its response peaks at around the value of 15. TPI near often has slightly elevated values close to the river channels due to the elevation difference between the channel bed and the top of the bank. TPI far values, on the other hand, change relatively linearly between the highest and the lowest points of the landscape due to the large neighborhood size.

The differences between the base water levels in cost distance layers are minor, mostly localized downriver, where the effects of land uplift are most readily observable. Between 3 meters and 6 meters base levels there is hardly any difference. Between 11 meters and other levels the difference is slightly more apparent, but still mostly limited to downriver to the Turku urban area. Upriver, the differences are predominantly confined to the immediate vicinity of the river due to the depth of the channel. In the layers based on Euclidean distance the situation is similar, with the notable exception that the distribution patterns in the layer using 11 meters base water level are thinner compared to the other layers.

A potential problem with distance to water variable is the extremely high percent contribution and permutation importance scores of the layers. MaxEnt estimates the percent contribution of the cost distance layers between 70.8 % and 71.4 %, while the Euclidean distance layers get scores between 61.6 % and 63.7 %. The permutation importance scores are equally high ranging from 77.7 % to 80.8 % in the case of cost distance layers and from 51.5 % to 55.5 % for Euclidean distance layers. Meanwhile, the scores of the rest of the layers range from 1 %

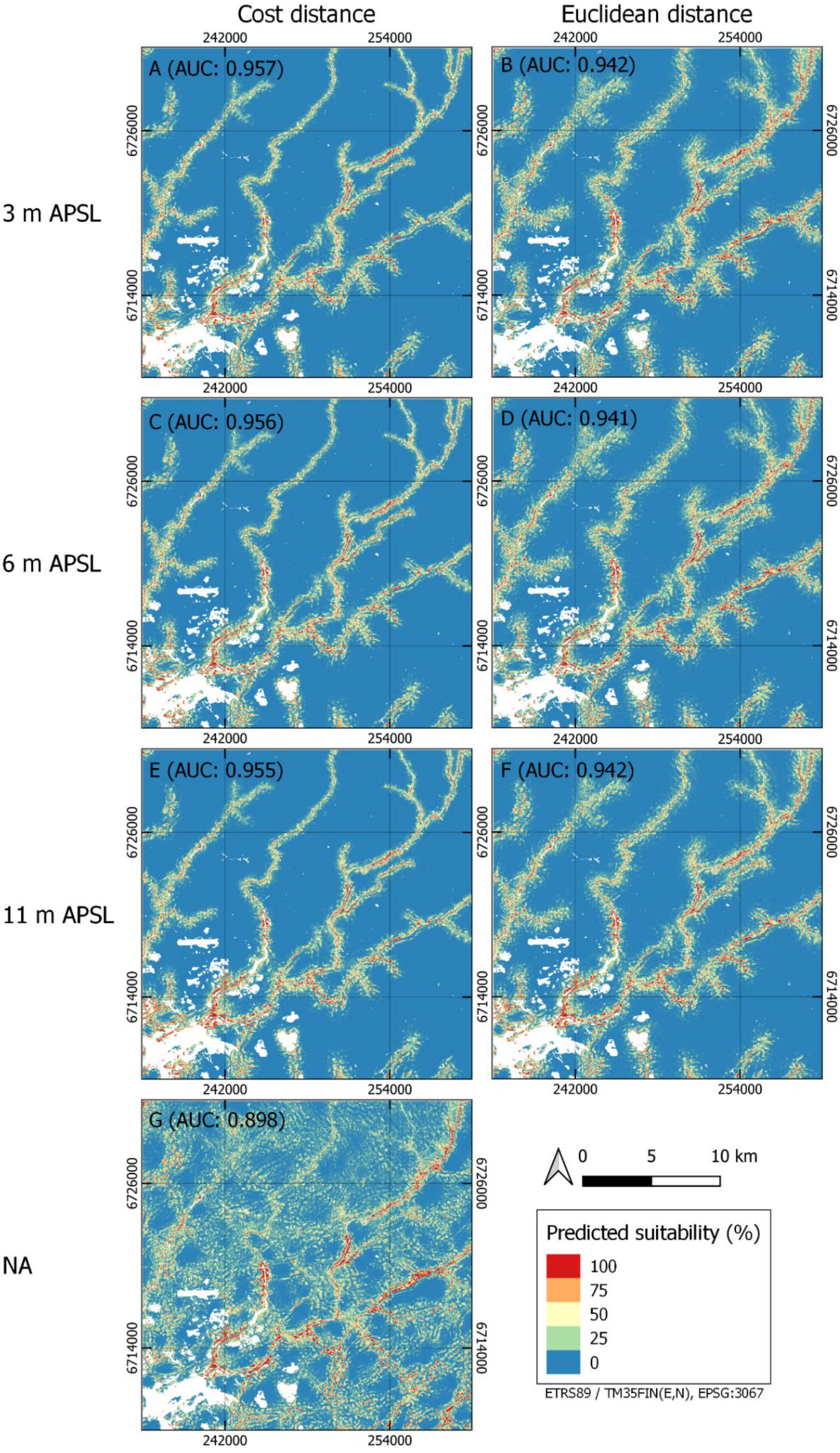


Figure 11. Comparison of models using distance to water –layers with different distance measures and base water levels.

to 15.6 % percent contribution and 0.5 % to 17 % permutation importance depending on the model. In the model not using distance to water variable the percent contribution estimates of the top three layers range from 25.4 % to 31.6 % and permutation importance scores from 22.6 % to 40.1 % making them relatively equal. When present, the distance to water –variables are by far the most significant variable, which may be a problem due to the predicted distribution overreliant on it.

Overall, every model, even when using no distance to water layer, performed well. The distribution patterns of the models using Euclidean distance are slightly wider and more even across the landscape than the patterns in the models using cost distance. This is a result of river channels in the study area being deeper further from the coast, which is only reflected by the cost distance layers. While the differences in the landscape are better shown in the models using cost distance, it may not be the best distance measure in all situations. Cost distance combines information from multiple sources, making it the more complex variable, while Euclidean distance may, as a simpler variable, be preferable in some cases.

3.3.3 Absolute or relative elevation

Six models with different configurations of elevation layers were created for the third set of comparisons. The first two models (Figure 12, maps A and B) represent the two main configurations under consideration with the first one using DEM for absolute elevation and the second one using both TPI layers for relative elevation. The second pair (maps C and D) test both TPI layers separately. Finally, the third pair tests configurations with all (map E) and no (map F) elevation layers. The rest of the predictors used in the predictions are Euclidean distance to water, solar irradiance, topsoil class and TWI. Euclidean distance was selected, as the cost distance layers are partially based on DEM layer and as such include some elevation information.

Based on the AUC score the best model is the one using all elevation layers (Figure 12, map E) at 0.975 and the worst is, expectedly, the one using no elevation information (map F) with AUC score of 0.930, which confirms that none of the layers worsens the results by being present. The AUC scores of the rest of the models in order of performance from best to worst were 0.969 for the model using DEM (map A), 0.942 for both TPI layers (map B), 0.937 for only TPI near (map C) and 0.933 for only TPI far (map D). Out of all the models the ones using DEM as a predictor outperformed those that did not, with TPI near coming in second. TPI far showed the least improvement over the model using no elevation data.

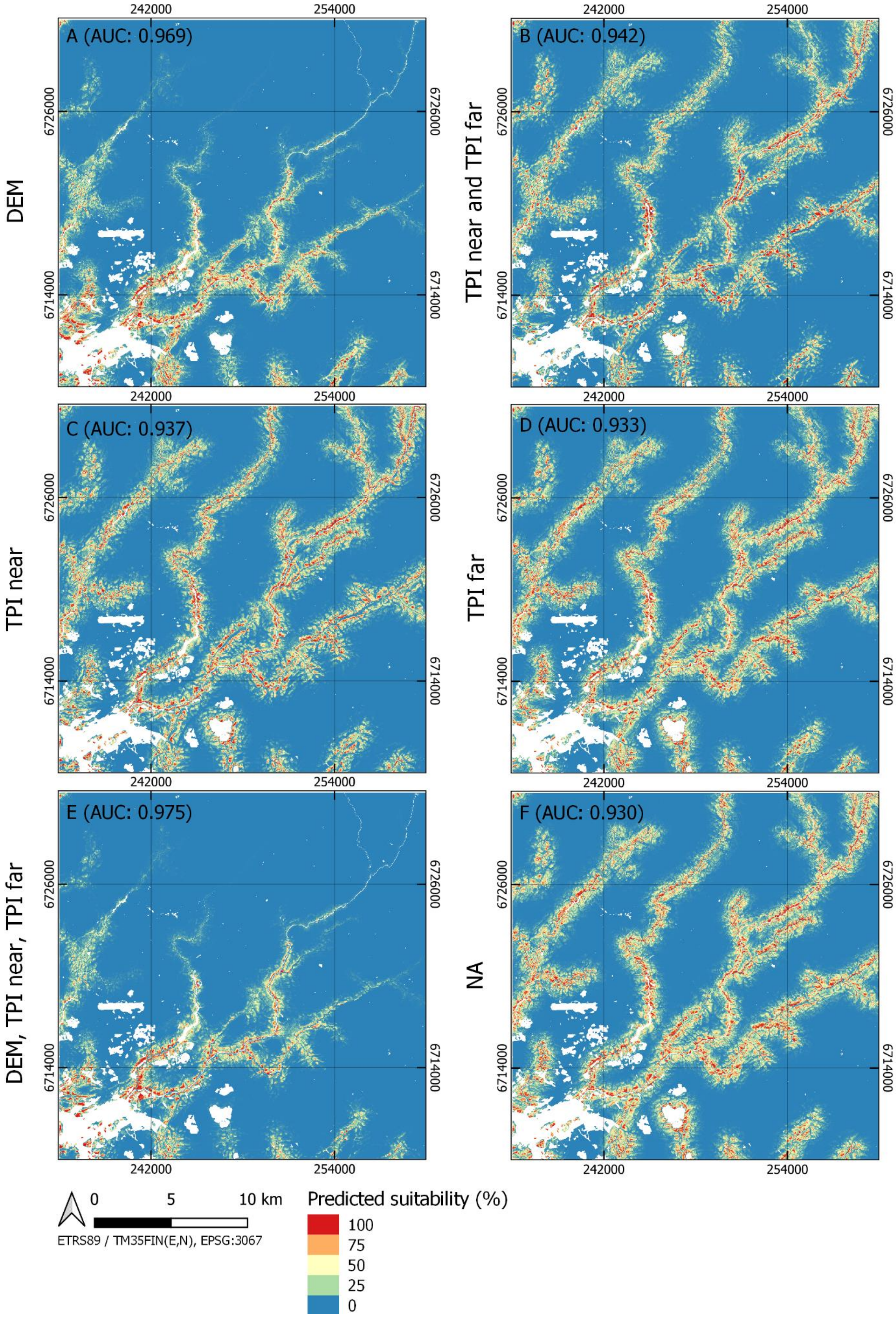


Figure 12. Comparison of models using various combinations of absolute and relative elevation layers.

Models partially based on DEM (Figure 12, maps A and E) have a distinct distribution pattern compared to the rest. Their predicted area is focused downriver, towards the coastline, while the rest of the models have predictions spread evenly across the entire landscape. In the model using all the layers the predicted area is narrower than in the model using only DEM. The distribution pattern of the former is also remarkably similar to the model using both TPI layers (map B) with the difference that the latter does not have the bias towards the coast the former has.

At glance, the distributions between the models using TPI near (Figure 12, map C) and TPI far (map D) look similar, but at a closer inspection there are clear differences. The distributions of the latter follow the rivers closely, while the most suitable locations in the former are slightly further away, above the banks. However, the TPI near model also includes highly suitable areas on some hilltops, likely resulting from the small neighborhood size failing to account for large-scale characteristics of the landscape. TPI far model fares better in that regard but does not appear to show as much differentiation in the low value regime. Finally, as every layer adds more limitations, the model using all the elevation layers (map E) shows the most confined distribution. In contrast, the distribution is most spread out in the model using no elevation data (map F).

Once again, every model showed high performance. The patterns of the predicted distributions, however, have major differences between absolute and relative elevation. The overall pattern confined by absolute elevation shows clear bias towards the coast, while in the case of relative elevation the distribution is spread evenly across the landscape. Both approaches have their strengths and weaknesses. Relative elevation may have more functional relevance in local environment, but on the other hand absolute elevation might be able to describe the actual distribution better as it better matches the current understanding of the extent of the Iron Age settlement in the area of interest (Juha Ruohonen, personal communication, February 23, 2023). Thus, it appears that the DEM layer includes information important for describing the locations of settlements, perhaps the distance to the mouth of the river, which is not present in either of the TPI layers.

3.3.4 Location of suitable areas

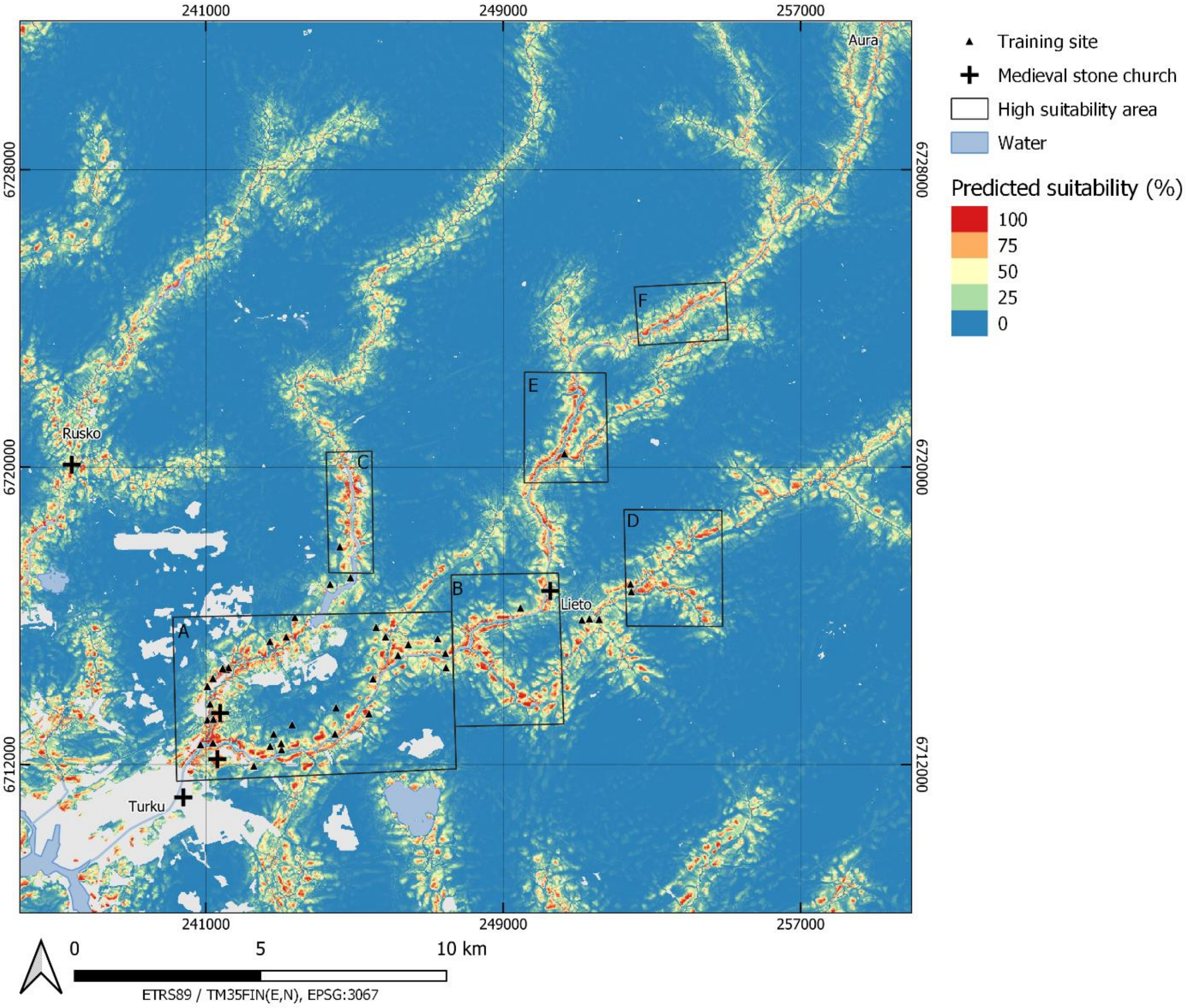


Figure 13.A map of the summary model.

An average of the previously discussed models was created for the discussion on the distribution of the most suitable areas. The summary model (Figure 13) was created by calculating an average for each of the previous three chapters. Then, an average of the resulting three models was then calculated, giving each of the issues equal weight. The resulting summary model is purely artificial and thus does not represent any particular issue.

The characteristics of a typical area with high predicted suitability were investigated through visual inspection. Most suitable areas are found right above the riverbanks, which is corresponding with the current view of the late Iron Age settlement locations (Leppänen & Mäkelä, 2022, p. 32). Almost all highly suitable areas are located within 400 meters of the river, after which the predicted suitability drops rapidly. Most of these areas are located on gentle slopes inclined towards river.

Six continuous areas of high predicted suitability were identified within the AOI and labeled A-F (Figure 13). Area A is located along Aurajoki and Vähäjoki rivers and continues from Turku urban area to Maaria reservoir along Vähäjoki and to the confluence of Savijoki along Aurajoki. It contains a number of known Iron Age settlements, including most of the training sites used. Area B (Figure 14, Map B) continues from around the confluence of Aurajoki and Savijoki upriver to Lieto conurbation on Aurajoki and to Kärpijoki on Savijoki. Compared to area A, area B contains very few previously known sites. There are, however, several sites located further upriver, on the banks of Savijoki, which may mean the riverside between the confluence and Kärpijoki, although relatively steep banked, has potential for containing previously unknown sites.

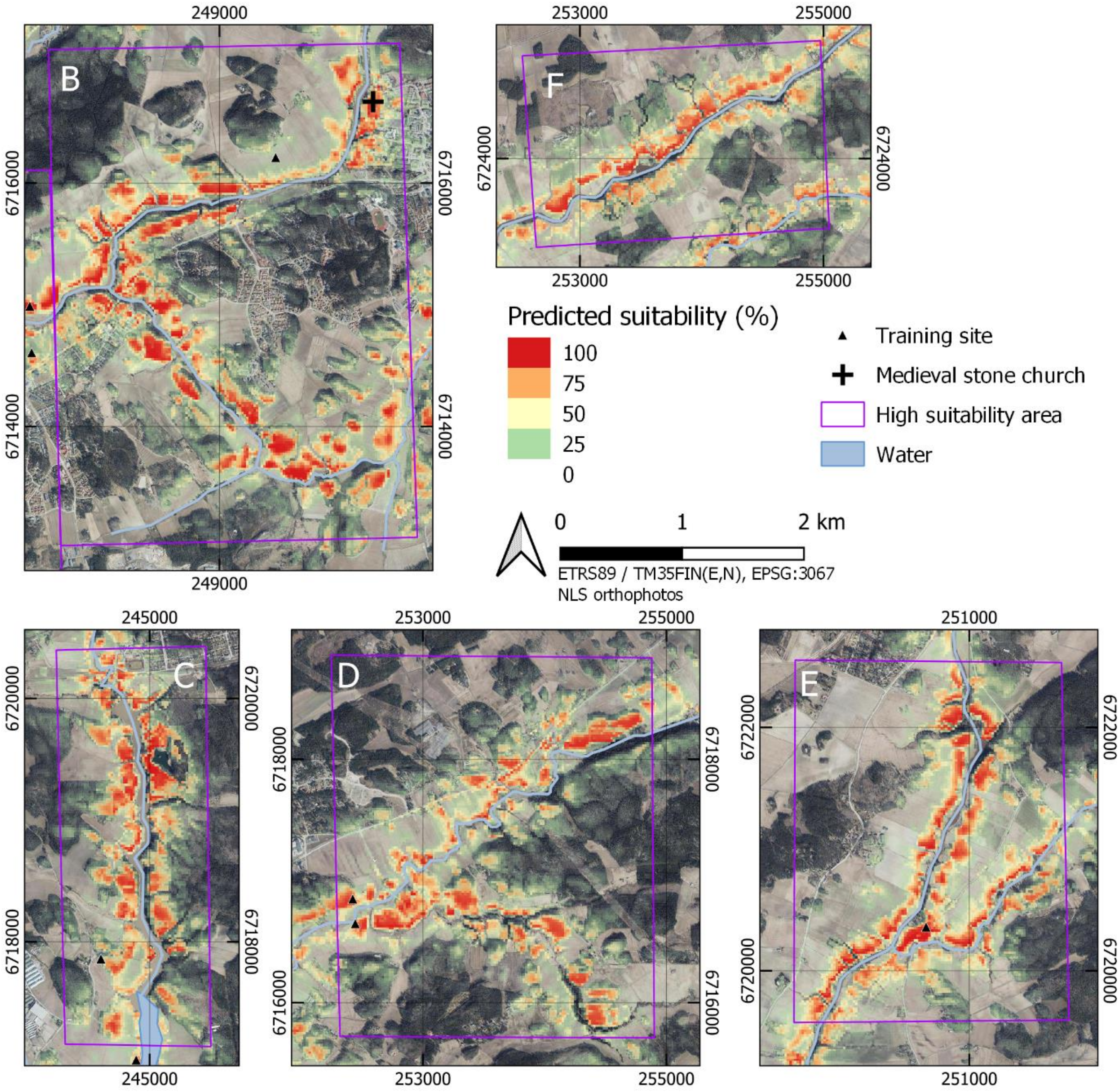


Figure 14. Areas B-F from Figure 13, overlaid with NLS orthophotos.

Area C (Figure 13; Figure 14, Map C) is located on the waterside of the northern end of Maaria reservoir. The area is just outside the areas with the highest density of known Iron Age settlements. The terrain is, however, somewhat difficult with steep slopes and the original river channel and the adjoining banks are underwater in the reservoir. Additionally, the area contains a large gravel pit, which has more than likely disturbed the results. Comparatively, area D (Figure 14, Map D) on the banks of Savijoki, between Raukkala and Pränikkälä, is interesting as it contains several Iron Age sites and additionally, a number of locations where scattered artefacts have been found (FHA, 2023).

Areas E and F are further upriver on Aurajoki river (Figure 13; Figure 14, Maps E & F). The former is located on the area surrounding Mäkkylä and Nautela. As was the case with area D, there are several known sites, in addition to which, a number of scattered artefacts have been found (FHA, 2023), making it likely that there are also unidentified sites. The latter on the other hand is further from known Iron Age sites on the area of Leinakkala and Leppäkoski. As there is little to no Iron Age found this far upriver, I interpret it as an unlikely candidate for a survey.

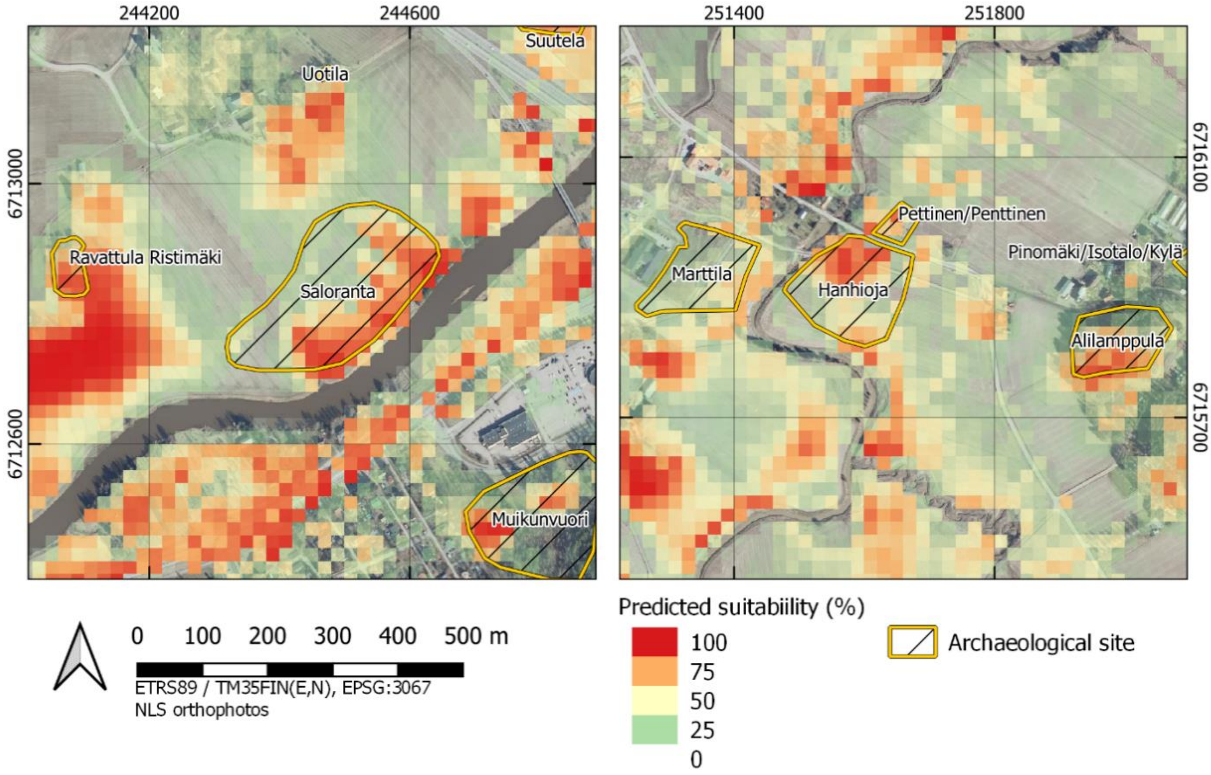


Figure 15. A detailed look showing the Iron Age sites of Saloranta (left) and Hanhioja (right) on the summary model overlaid on aerial image.

Finally, the summary model was inspected on a finer scale in comparison to two known settlements of Kaarina Ravattula Saloranta and Lieto Pettinen Hanhioja (Figure 15), both of which were also included in the training sample. On the Saloranta site, the artefacts have been found on a circa 300 meters long, 100 meters wide band along Aurajoki river (Mäkelä, 2021, p. 52). On Hanhioja, the artefact scatter matches relatively well the currently defined site boundary (Leppänen & Mäkelä, 2022, p. 31, Figure 4) seen in Figure 14. On the former, the predicted distribution matches the scatter almost exactly, while on the latter the prediction is more restricted than the actual scatter. Nevertheless, the predictions on both sites are surprisingly accurate making a case for the capabilities of MaxEnt.

3.3.5 Summary

In total, 47 models were created. The first 32 models were used as test runs for various settings and configurations. The final 15 of the models, as presented in chapters 3.3.1.–3.3.3., were created based on these observations two answer different questions. The first set, consisting of two models, tested the effects of using simple and complex sets of predictors and resulted in somewhat limited differences. The second set of seven models tested differences while using different distance measures for calculating the distance to water. The final set of six models tested the configurations of elevation layers. Finally, a summary model was created for chapter 3.3.4 to discuss the distribution of suitable areas.

The models were evaluated using various methods, such as visual inspection, AUC, and comparison of variable contribution. The highest AUC score of 0.975 was achieved by a model using all elevation variables (Figure 12, map E), which, however, may reinforce some variables through correlation, as the layers contain repeat information. The lowest score was achieved by a model without distance to water –variable (Figure 11, map G), with the AUC of 0.898, which is still generally considered high.

The most significant variables by contribution were absolute elevation and distance to water, and cost distance variables showed slightly higher contribution compared to Euclidean distance variables. Each additional variable further delimited the model area, and no variable was found to impair the results, though the effects of some variables like aspect were negligible. The results overall showed surprising similarity even with different selection of predictors.

Rather than continuous swaths of landscape, the predictions were largely fragmentary. This is in part due to modern construction, such as buildings, roads, and ditches, cutting through the landscape, creating disconnections, and in part due to the relatively fine grain of the predictors. Natural features, for example waterways, also play a part in creating similarly fragmentary patterns.

While in a few cases, known sites were systematically in all models left outside the suitable area, in some cases MaxEnt proved surprisingly accurate. In total, out of 39 training sites eight always fell outside the area with predicted suitability above 50 %. On ten sites the results varied. On the other side of the spectrum, three training sites, Turku Kärsämäki Radanvierusta, Lieto Vääntelä Uotila and Turku Koroisten tila were archetypical in the sense that their locations received suitability above 95 % in every model and in a few cases even the highest possible value of 100 %. These numbers should not be treated as proof of capability or incapability of the modeling method as all the sites discussed played part in model creation. However, based in the scattered finds mentioned in chapter 3.3.4, I would argue that the use of MaxEnt in archaeological predictive modeling is worth further study.

4 Discussion

While the models created for this study are simplistic in concept, based on the observations made in chapter 3.3.4, the predictions were in some cases surprisingly accurate. A conclusive evaluation, however, would require more robust evaluation methods, such as use of test dataset or field surveys. The models followed the known distribution of the Iron Age sites in the area of interest reasonably well, but a closer inspection reveals several issues. For example, the algorithm consistently failed to include in the model several areas where known sites used in model training are located. Most of the time the sites falling outside the predicted suitable area were the same, which indicates a subset of the training sample differs from the rest, as MaxEnt appears to evaluate them as outliers. Whilst MaxEnt evidently is capable of recognizing distribution patterns using the given data, ultimately, the models are only as good as the data the algorithm is provided.

At the beginning of this study, one of the goals was to find a combination of predictors yielding the best results, or in other words, best at representing the human niche. Ultimately, the question has proven rather arduous to get a fulfilling answer for, especially as the answer heavily depends on the definition of the best in this context. If we define best as simply having the highest AUC score, the answer would be simple. On the other hand, if we want to know which model best represents the Iron Age settlement patterns, the answer will depend on knowing the actual distribution of them. Based on the current understanding, the models using absolute elevation as a variable appear to be the closest (Juha Ruohonen, personal communication, February 23, 2023). This, however, does not mean that absolute elevation is the best or even a good variable at describing the settlement patterns.

A great advantage of topographic variables is their availability globally (e.g., SRTM), thus making their use low threshold for most use cases. Their use in predictive modeling, however, comes with a caveat that they only describe an attribute of the environment and not how the species of interest interacts with it. While topography influences, informs, and correlates with many processes that directly affect the habitability of the environment, one could argue that the influence is indirect at best. A viable alternative approach may be dividing the variables into their more functional components, all of which would be independently evaluated by the algorithm. This would, however, depend on identifying these components, which in archaeological context would require robust theories about human-environment interactions.

4.1 Human-environment relations

A common approach to studies on human-environment relations is the idea that the environment sets the boundaries within which the humans function (e.g., Coombes & Barber, 2005, p. 305). The boundaries can be limitations, outside which long term survival is not possible, or constraints, which direct the action towards a certain outcome. According to Arponen et al. (2019, p. 4), in palaeo-environmental studies a common way to look for these boundaries is looking for correlations between changes in the environment and a culture. Within the AOI of this study these correlations have previously been studied by, for example, Saloranta (2000, pp. 25–33). However, correlations only infer connections, but are not enough to identify causal links, which would need to be demonstrated to confirm, as the parallelism can be purely coincidental (Coombes & Barber, 2005, p. 305). This also applies to a situation where a modeling algorithm finds connections between predictor and response variables.

The debate whether human actions are a result of free will or determined by outside factors is one of the oldest philosophical discussions recorded (Stanton, 2004, p. 32). In the case of this thesis the debate is interesting as the implementation of the model already takes a stance for a deterministic approach. In fact, completely detaching the choice from the external factors would undermine the entire concept of this study and even the whole data driven approach on large. According to Arponen et al. (2019, p. 4) the modern understanding of human life includes a level of biologism, which views human life fundamentally as a biological phenomenon, which must abide by certain biological needs. The changes in the environmental conditions affect how the needs can be satisfied, thus affecting the biological creatures living in them. Arponen et al. conclude by arguing that the uncertainty regarding the degree of environmental influence is the driving force behind the whole debate on environmental determinism.

Data driven models have been criticized, as they represent an inherently environmentally deterministic approach. As stated by Verhagen (2007, p. 17), common critiques include the use of incomplete archaeological data, biased selection of environmental variables and neglecting of cultural variables and the changing nature of environments. Additionally, Verhagen notes that all the issues mentioned stem from failing to obtain data appropriate for capturing all the influences affecting site location and that shortcomings of data driven modeling become more prevalent with later, more developed societies as they become less dependent on the natural environment for their subsistence.

Representation of all the complexities of the studied culture appears a hopeless task (Coombes & Barber, 2005, p. 305) and I would argue that a model should not even strive to achieve it, because increased complexity also increases the difficulty of interpretation. Ultimately a model is an abstraction: a simplified representation of an exceptionally complex system. As Coombes and Barber continue, representation of the critical components should be feasible. In that case the issue is identifying these components.

We could approach the issue by using biological needs, like fresh water and food supply, as the basis and would likely be able to find linkages. In their article, Jones et al. (1999) were able to observe connections between palaeoclimatic data and archaeological record, supporting a theory that droughts affecting, for example the available food supply, had an effect on human subsistence and population. Bettinger (in Jones et al., 1999, 159), commenting on the article, however, pointed out that the evidence is circumstantial, and we have no idea what was actually limiting the population. Bettinger further highlights that while the connection between food supply and population certainly exists, it is rarely direct and thus cannot be expected to be.

Identifying causal influences, thus, is far from straightforward. Boserupian theory (Boserup, 1965, pp. 116–121) presents an idea that increasing food supply does not increase the population but enables higher population growth and that increasing population drives adopting new technologies to increase the food supply. In other words, increasing population puts pressure on the supply chain, which drives innovation from within the community. Thus, it appears whether resources are present or not is a more important factor than the amount of them. Lack of resources introduces boundaries, but they can be pushed further through innovation. Humans are extremely good at adapting to a wide range of conditions and modifying the environments to suit their needs. And even in inhospitable environments they are often able to find ways to survive, but they still need some sources of subsistence. Therefore, it may be reasoned with fair certainty that the truth lies somewhere in between complete determinism and completely free will.

4.2 Data

When conducting a predictive modeling study, an important step is to establish a concept or the objective of the model, as that dictates what data the model should be based on. As the main purpose of this study was to test viability of SDMs as a modeling method, the data selection was based on previously established approach in Tiilikkala (2016) and Debenjak-

Ijäs' (2017) theses. In hindsight, the results may have been improved with different, more thorough data selection process, but nonetheless, the results bring out interesting issues with this particular modeling approach. Additionally, even a different approach may not have improved the results significantly.

An important question in any study involving any sort of data usage is whether the used data is fit for purpose (Veregin, 1999, p. 178). The answer to the question regarding this study is multi-faceted. Considering the question within the framing of the research questions posed in chapter 1.1, I would argue that the data was fit, as they did not hinder answering the questions. This is in part due to the way the questions were formulated, as they allowed for an interpretive margin. The main objective was getting results; the quality was secondary. However, considering the question in more depth may be useful in the future.

Veregin (1999, pp. 178–183) discusses three components, spatial, temporal, and thematic, that together make geospatial data. They argue that usually the spatial dimension gets the most attention even though data is arguably all about a theme. Theme drives the collection and application of data and may be viewed as such an inherent part of it that the issues with it go unnoticed. The type of the issue and the way it manifests varies by its root cause.

The previous studies on archaeological predictive modeling have failed to reach a consensus on how the thematic composition of a model should be tackled. Because of this, the issue remains outside the scope of this study. However, there are several issues with the response and explanatory variables worth discussion. In the next chapters I ponder the issues in these data and consider recommendations based on the experiences from this study.

4.2.1 Response variable

The response variable in this study was based on cultural heritage registers. The training sites were manually picked from the full dataset using the information available within the data and the associated reports. During the process it became evident that even in an area with relatively active field of research like Southwest Finland, the information on many of the sites was lacking due to various circumstances such as inopportune weather conditions during the observation. This is a symptom of the data being collected primarily for management purposes over a long period, and the lack of resources for fundamental research. The result is, as is often the case with archaeological data, an assortment information of inconsistent quality. Therefore, most of the issues can be tracked down to the lineage of the data.

Due to the opportunistic and cumulative collection method, the data is lacking especially in completeness and consistency. The most important driver for data collection is land use development and the areas with most intensive land use receive the most attention, while more sparsely populated areas have been largely ignored. Because of this, the AOI being focused on the immediate vicinity of Turku is, in a way, a best-case scenario for the completeness of the data. Additionally, the situation overall has been somewhat improved in the recent years due to the growth of metal detecting, but large areas still remain uncharted.

The inconsistencies are a result of multiple factors including changes of methods and standards due to the long history of the data, the number of data collectors and the varying conditions during observation. Because of this, each record contained in the data is a product of its time for better or worse and can contain inconsistencies or even inaccuracies. Based on the reports, it can be difficult, even impossible, to ascertain dating and type of a site as many of the identifications are guesstimates based on scattered pottery sherds. Even with its issues, the dataset does, however, have distinct advantages over every other archaeological dataset in Finland: It is the most comprehensive overall dataset in existence and its open access availability.

Thus, I would argue that the use of cultural heritage registers as the source for occurrence data is justified in low level predictive modeling studies. However, if the aim of the study is to create representative models, alternative data should be sought. Optimally the data should only consist of verified occurrence locations, but in most cases the intensive surveys required are not attainable. Another, less intensive option could be using metal artefacts as the proxy (Juha Ruohonen, personal communication, February 23, 2023; Leppänen, 2022, pp. 30–32). Their number has increased dramatically due to the increase in metal detecting, and they are much easier to date typologically than pottery sherds. Because metal detecting is often practiced by amateurs, their locations may also be less skewed by expectations based on prevailing theories. Due to the active detectorist scene in Southwest Finland, the region would be a prime candidate for testing the hypothesis.

The expectation, when predicting a broad group of sites, such as settlements, is that they form a cohesive enough group that similarities can be identified. This may become a problem if a distinct subgroup that does not express those similarities exists. In the case of this study, this could happen if settlements practice different subsistence strategies. Hypothetically, a settlement practicing agriculture may have entirely different environmental requirements than a settlement based on fishing or trade. If subsistence strategies were to be included, their implementation and whether settlements following different practices can fit under the same umbrella needs to be considered. A way to test this would be to focus survey some of the areas with high predicted suitability.

The expectation of similarity also imposes limitations on the extent of the area a model can be generalized to. The predictions based on the training data used in this study may be generalizable to neighboring river valleys of Southwest Finland, but likely not further inland or to the neighboring regions. As the predictions are largely dependent on spatial autocorrelation (see, e.g., Naimi et al., 2011), the uncertainties grow further from the known data.

4.2.2 Predictor variables

Discussing the potential issues with the predictor variables requires diving into murkier waters, as the data driven approach used in this study is not well suited to answering such questions. The set of predictors used was based on previous studies and guesstimates of the environmental features that may be important for the target species represented by the response variable. Data availability was also a contributing factor. The final list included a number of primarily topographical variables, because they are relatively stable across time in the changing landscape. The selection put an emphasis on the abiotic environment instead of biotic, meaning that in effect the representation of biotic resources such as sources of sustenance is lacking. The results thus show the predicted distribution of the Iron Age settlements based on the form of the landscape instead of its function.

The environmental layers, except for soil type and river channels, were based on remote sensing data. Remote sensing has an advantage over other collection methods in that it produces spatially consistent and continuous data over large areas. A disadvantage is that only information that can be viewed from afar can be measured. The data produced is a snapshot of a moment of time, somewhat limiting their application in studies of the past. The environmental features, buildings, infrastructure, reservoirs and so on, that existed during the time of acquisition are depicted.

Clipping the modern features out and filling the gaps through interpolation is an option, which, however, could produce misinformation making the results more difficult to interpret. Because of this, I decided against removing the modern features apart from reservoirs because the prior river channels could easily be reconstructed based on old aerial photographs, which, around Turku urban area, have been extensively digitized and widely available. The coverage of the photographs varies highly and as such they may not be available in more sparsely populated areas.

Absolute values, like elevation above mean sea level are measured in relation to a standard making their values in different parts of the landscape comparable. On the other hand, relative values like topographic position index ditch the common standard and consist of values relative to their neighborhood. Because of the standardization, absolute values can appear artificial and irrelevant to a local perceivable environment. As a result, relative variables can appear superior at glance. They do, however, come with their own assortment of issues. Firstly, their values are not always comparable from one cell to another, because the analyses can reach the same result in multiple ways. Secondly, prominent features such as steep cliffs can cast large shadows across the landscape either lowering or elevating the surrounding values. Thirdly, by default the entire landscape is processed using unified neighborhood size while the relevant size may vary depending on the terrain. In future, integrating, for example, viewshed size, view distance or other attributes as part of the processing workflow could improve the variables by making them better adapted to the local variations of landscape.

A 20-meter grain was selected for all predictors and models created for this study. For the scope it was adequate, but the question whether a single grain is enough should be paid attention to. Arguably, the meaningful grain would depend on the scale of the activity or function; Pottery makers work at a finer scale compared to hunters who work at a much broader scale, for example. The same principle can be extended to the environmental variables as well. Location of a water source may affect the broad, kilometer scale location of a site, while the influence of the slope of the terrain may be a few dozen meters. The meaningful scale can also vary from area to area making a relatively flat area like Southwest Finland incomparable with an area with rougher terrain in, for example, eastern Finland. However, when considering scale and working with data, especially that acquired through methods other than remote sensing, it should also be noted that the sampling density limits the achievable precision limiting the scales that can feasibly be studied.

Finally, while humans and the environment were represented, their interactions were not. Modeling them can be attempted, but while humans use senses to perceive their environment and decide the course of action, computers base their predictions on statistics calculated from the values of provided data. Thus, even with all the recent developments in artificial intelligence applications, all computer-generated models are fundamentally based on mathematics. The models will doubtless drastically improve in the coming years, but as long as computing is not completely revolutionized the previous statement will hold true.

4.3 On to the future

SDMs commonly represent an inherently data-driven approach to predictive modeling. The approach used in this study was suited to answering the question of what kind of abiotic environments the archaeological sites fall into. The purpose of the model should be the main deciding factor on the approach. In archaeological context data-driven approach has some, sometimes major shortcomings, but they are comparatively quick to produce and iterate on and as such can be of great use as an assist in surveys. Theory driven approach may be able to produce more robust results, but the process is more involved and difficult to execute. Verhagen (2007, p. 17) draws the conclusion that the way forward is somewhere in-between, but I would argue that data-driven models still have their use cases where rapidity of development is more important than prediction accuracy.

As for modeling methods, generally, MaxEnt has been found to perform well with archaeological data (Yaworsky, 2020, p. 17). Although in the future more advanced models may be developed to replace it, for now there is no reason not to use it as it is seeing more widespread adoption through integration into major GIS-software such as ArcGIS Pro (ESRI, n.d.).

A natural development, and perhaps where the largest strides could be made, would be improving the quality of both the occurrence and predictor data. An obvious improvement would be ensuring concurrency of the occurrence data and the depiction of the environment and to narrow down the date range. They would, however, require extensive reconstructions of past landscape and as such would be prone to inaccuracies. A look at the legacy effects of past land use in the landscape (see, e.g., Foster et al., 2003) and establishing change trajectories would help push us in the right direction. The idea is based on the fact that the present is built upon the past. The continuities have been studied in the past through historical maps (e.g., Kuusela & Tiilikkala, 2008), however, often the studies have been limited to sites. In the Turku region, continuities where a historical settlement remained at the same place as a prehistoric one have been observed (e.g., Saloranta, 2000).

More refined conceptualization of the model could help in identifying the critical factors. Ecosystem services is a concept for describing and analyzing the benefits a species gets from the environment. Katz (2022, p. 1456) argues that the framework would make archaeologists the environmental settings and subsistence of past societies as well as provide perspectives into human-environment interactions. As an existing, well established, concept, ecosystem services thinking, integrated into predictive modeling, would be beneficial to finding the critical interactions and bridging the gap between archaeology and ecology.

The concept of ecosystem services is rooted in natural sciences, making archaeological studies utilizing them inherently transdisciplinary, which I would argue, is the way forward. There are numerous methods and concepts used by other disciplines that could be beneficial to archaeological predictive modeling. Multidisciplinary studies combining archaeologists' expertise on humans with the methodical expertise of other specialists would help reap these benefits.

5 Conclusion

In this thesis I set out to explore the use of SDM methods in the context of archaeological predictive modeling. In concept this study was not particularly ambitious, being based on previous studies, but the results were still encouraging, suggesting there exists a lot of untapped potential. Even if the conceptualization is kept relatively simple, MaxEnt is able to generate predictive models useful for practical applications like assisting in identifying potential areas for a more intensive survey.

With deeper, more theory driven conceptualization the models could be taken to a new level by finding the critical human-environment interactions. Additionally, especially trans- and multidisciplinary research would definitely benefit predictive modeling. The widely available geospatial datasets are restricted in scope and larger projects may have the resources needed to carry out environmental reconstructions needed to create the required predictor variables.

It is important to remember that the models are not infallible. Fundamentally, no matter how advanced, a predictive model is only that: A model, which is to say it is a simplification based on existing knowledge. The way to improve the model accuracy is to increase the background knowledge, which happens through fundamental research. In the best case, the models themselves, used in focusing field surveys, can become a part of a self-feeding loop creating the data they need. However, in this case, attention needs to be paid to the possibility of skewness in the knowledge gained with their assistance. As the models are based on existing information, their indiscriminate use has a danger of leading to a feedback loop that only strengthens the existing biases. Because of this, it would be important to test each iteration of the models on the field by surveying the areas predicted with both high suitability and those with low suitability. A natural continuation for this study could, for example, be conducting focus surveys on the areas discussed in chapter 3.3.4.

In conclusion, MaxEnt can be used to create predictive models from archaeological data, and the subject is worth further study. In the future the primary focus should be data quality, as that is where the biggest strides can arguably be made. If predictive modeling is integrated as part of the fundamental research process, the models themselves could serve as means for improvement by helping to provide new information based on the existing knowledge.

Abbreviations

AOI – Area of interest

AMSL – Above Mean Sea Level

APSL – Above Present Sea Level

AUC – Area Under (ROC) Curve

BAM – Biotic, Abiotic, Movement

DEM – Digital Elevation Model

E – Environmental space

ENM – Ecological Niche Model

FHA – Finnish Heritage Agency (Fi. Museovirasto)

G – Geographic space

GNSS – Global Navigation Satellite System

GTK – Geological Survey of Finland (Fi. Geologian Tutkimuskeskus)

HSM – Habitat Suitability Model

SDM – Species Distribution Model

SQL – Structured Query Language

SRTM – Shuttle Radar Topography Mission

SYKE – Finnish Environment Institute (Fi. Suomen ympäristökeskus)

NLS – National Land Survey of Finland (Fi. Maanmittauslaitos)

ROC – Receiver Operator Characteristics

TPI – Topographic Position Index

TWI – Topographic Wetness Index

References

Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R. I., García, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, *5*(1), eaat4858–eaat4858. https://doi.org/10.1126/sciadv.aat4858

Arponen, V. P. J., Dörfler, W., Feeser, I., Grimm, S., Groß, D., Hinz, M., Knitter, D., Müller-Scheeßel, N., Ott, K., & Ribeiro, A. (2019). Environmental determinism and archaeology. Understanding and evaluating determinism in research design. *Archaeological Dialogues*, *26*(1), 1–9. https://doi.org/10.1017/S1380203819000059

Austin, M. P. (1980). Searching for a Model for Use in Vegetation Analysis. *Vegetatio*, *42*, 11–21.

Austin, M. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, *157*(2), 101–118. https://doi.org/10.1016/S0304-3800(02)00205-3

Banks, W. E. (2017). The application of ecological niche modeling methods to archaeological data in order to examine culture-environment relationships. *Quaternaire (Paris)*, *28*(2), 271–276. https://journals.openedition.org/quaternaire/7966

Barford, P. M. (2000). Seen and unseen: sites in a landscape. In A. Nissinaho (Ed.), *Sites and Settlements* (pp. 85–99). Åbo Akademi Tryckeri.

Biodiversity and Climate Change Virtual Laboratory. (2021). *SDM – Interpretation of Model Outputs*. https://support.bccvl.org.au/support/solutions/articles/6000127046-sdm-interpretation-of-model-outputs

Boserup, E. (1965). *Conditions of Agricultural Growth* (First edition.). Boca Raton, FL: Routledge.

Coombes, P., & Barber, K. (2005). Environmental determinism in Holocene research: causality or coincidence? *Area (London 1969)*, *37*(3), 303–311. https://doi.org/10.1111/j.1475-4762.2005.00634.x

Debenjak-Ijäs, A. (2018). *Asutusta etsimässä: Menetelmiä myöhäisrautakautisen asutuksen paikantamiseksi* [Master's thesis. University of Helsinki]. HELDA – Digital Repository of the University of Helsinki. http://urn.fi/URN:NBN:fi:hulib-201803161482

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography (Copenhagen)*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. McC. M., Townsend Peterson, A., … & Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography (Copenhagen)*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J. & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity & Distributions*, *17*(1), 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Elton, C. (1927). *Animal Ecology*. London: Sidgwick and Jackson.

ESRI. (n.d.). *Presence-only Prediction (MaxEnt) (Spatial Statistics)*. Environmental System Research Institute. Retrieved April 4, 2023, from https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/presence-only-prediction.htm

Feng, X., Park, D. S., Liang, Y., Pandey, R., & Papeş, M. (2019). Collinearity in ecological niche modeling: Confusions and challenges. *Ecology and Evolution*, *9*(18), 10365–10376. https://doi.org/10.1002/ece3.5555

Finnish Environment Institute. (2012). *Shoreline – River Network (Ranta10)* [Data set]. Retrieved January 13, 2022, from https://ckan.ymparisto.fi/dataset/uomaverkosto

Finnish Heritage Agency. (2021). *All the sites of the cultural environment registers of the Finnish Heritage Agency (for research purposes) information product* [Data set]. Retrieved December 19, 2021, from http://paikkatieto.nba.fi/aineistot/tutkija.html

Finnish Heritage Agency. (2023). *All the sites of the cultural environment registers of the Finnish Heritage Agency (for research purposes) information product* [Data set]. Retrieved April 2, 2023, from http://paikkatieto.nba.fi/aineistot/tutkija.html

Foster, D., Swanson, F., Aber, J., Burke, I., Brokaw, N., Tilman, D., & Knapp, A. (2003). The Importance of Land-Use Legacies to Ecology and Conservation. *Bioscience*, *53*(1), 77–88. https://doi.org/10.1641/0006-3568(2003)053[0077:TIOLUL]2.0.CO;2

Franklin, J., Potts, A. J., Fisher, E. C., Cowling, R. M., & Marean, C. W. (2015). Paleodistribution modeling in archaeology and paleoanthropology. *Quaternary Science Reviews*, *110*, 1–14. https://doi.org/10.1016/j.quascirev.2014.12.015

Geological Survey of Finland. (2015). *Superficial deposits 1:20 000/1:50 000* [Data set]. Retrieved November 25, 2021, from https://hakku.gtk.fi/fi/locations/search

Geological Survey of Finland. (2018). *Maaperä 1:20 000/1:50 000 – Superficial deposits 1:20 000/1:50 000*. Retrieved March 31, 2023, from https://tupa.gtk.fi/paikkatieto/meta/maapera_20_50k.html#laatutiedot

Grabs, T., Seibert, J., Bishop, K., & Laudon, H. (2009). Modeling spatial patterns of saturated areas: A comparison of the topographic wetness index and a dynamic distributed model. *Journal of Hydrology (Amsterdam)*, *373*(1), 15–23. https://doi.org/10.1016/j.jhydrol.2009.03.031

Grinnell, J. (1917). Field tests of theories concerning distributional control. *The American Naturalist*, *51*, 115–128. https://doi.org/10.1086/279591

Heizer, R. F. (1962). The Background of Thomsen's Three-Age System. *Technology and Culture*, *3*(3), 259–266. https://doi.org/10.2307/3100819

Hijmans, R. J., & Elith, J. (n.d.). *Species distribution modeling*. Spatial Data Science with R. Retrieved April 4, 2023, from https://rspatial.org/raster/sdm/index.html

Huntley, B., Collingham, Y. C., Willis, S. G., & Green, R. E. (2008). Potential impacts of climatic change on European breeding birds. *PloS One*, *3*(1), e1439–e1439. https://doi.org/10.1371/journal.pone.0001439

Hutchinson. G. E. (1944). Limnological Studies in Connecticut. VII. A Critical Examination of the Supposed Relationship between Phytoplakton Periodicity and Chemical Changes in Lake Waters. *Ecology (Durham)*, *25*(1), 3–26. https://doi.org/10.2307/1930759

Hutchinson, G. E. (1957). Cold spring harbor symposium on quantitative biology. *Concluding remarks*, *22*, 415-427.

Jackson, S. T., & Overpeck, J. T. (2000). Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology*, *26*(sp4), 194–220. https://doi.org/10.1666/0094-8373(2000)26[194:ROPPAC]2.0.CO;2

Jiménez, L., Soberón, J., & McPherson, J. (2020). Leaving the area under the receiving operating characteristic curve behind: An evaluation method for species distribution modelling applications based on presence-only data. *Methods in Ecology and Evolution*, *11*(12), 1571–1586. https://doi.org/10.1111/2041-210X.13479

Jones, T. L., Brown, G. M., Raab, L. M., McVickar, J. L., Spaulding, W. G., Kennett, D. J., … Walker, P. L. (1999). Environmental Imperatives Reconsidered: Demographic Crises in Western North America during the Medieval Climatic Anomaly. *Current Anthropology*, *40*(2), 137–170. https://doi.org/10.1086/200002

Katz, O. (2022). The ecosystem services framework in archaeology (and vice versa). *People and Nature (Hoboken, N.J.)*, *4*(6), 1450–1460. https://doi.org/10.1002/pan3.10395

Kinnunen, J. (2019). Turun rannansiirtymisen uudelleenarviointi ja vertailu arkeologisten kaupunkikaivausten dendrokronologiseen ajoitusaineistoon. In R. Mustonen & T. Ratilainen (Eds.). *Pitkin Poikin Aurajokea: arkeologisia tutkimuksia. Turun museokeskuksen raportteja*, *23*, 121–133. https://www.turku.fi/sites/default/files/atoms/files/pitkin_poikin_e_kirja.pdf

Kirkinen, T. (1996). Use of a Geograhical Information System (GIS) in modeling the Late Iron Age Settlement in Eastern Finland. In T. Kirkinen (Ed.). Environmental Studies in Eastern Finland: Reports of the Ancient Lake Saimaa Project. *Helsinki Papers in Archaeology*, *8*, 19–61.

Kopecký, M., Macek, M., & Wild, J. (2021). Topographic Wetness Index calculation guidelines based on measured soil moisture and plant species composition. *The Science of the Total Environment*, *757*, 143785–143785. https://doi.org/10.1016/j.scitotenv.2020.143785

Kuusela, J.-M. & Tiilikkala, L. (2008). Malli myöhäisrautakautisesta asutuksesta lounaisessa Sisä-Suomessa. *Muinaistutkija*, *2008*(1), 14–27.

Lehtonen, K. 2000. Iron age settlement in river Aurajoki valley: Its pattern and relation to settlement of historic times. In A. Nissinaho (Ed.). *Sites and Settlements* (pp. 45–83). Åbo Akademi Tryckeri.

Leppänen, J. (2022). *Metallinilmaisinlöydöt rautakautisen toiminnan lähdeaineistona. Paikkatiedon ja löytökoostumuksen analyysi läntisessä Varsinais-Suomessa* [Master's thesis, University of Turku]. UTUPub. https://urn.fi/URN:NBN:fi-fe2022053139923

Leppänen, J. & Mäkelä, S. (2022). Rautakautisten peltokohteiden inventointia yhteistyössä metallinilmaisinharrastajien kanssa. *Ponsi 1 – Arkeologisia tutkimuksia. Raision museo Harkon julkaisuja*, *2*, 28–37.

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Lähdesmäki, U. (2000). Esimerkki suppean alueen kokonaisinventoinnista: Liedon Vanhalinnan alueen asutuskuvan tutkimus. In P. Maaranen & T. Kirkinen (Eds.). *Arkeologinen inventointi* (pp. 198–207). Gummerus Kirjapaino Oy.

Mac Nally, R. (2000). Regression and model-building in conservation biology, biogeography and ecology: The distinction between – and reconciliation of – “predictive” and “explanatory” models. *Biodiversity and Conservation*, *9*(5), 655–671. https://doi.org/10.1023/A:1008985925162

McInerny, G. J., & Etienne, R. S. (2012). Ditch the niche - is the niche a useful concept in ecology or species distribution modelling? *Journal of Biogeography*, *39*(12), 2096–2102. https://doi.org/10.1111/jbi.12033

Merow, C., Smith, M. J., Edwards Jr, T. C., Guisan, A., McMahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., & Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography (Copenhagen)*, *37*(12), 1267–1281. https://doi.org/10.1111/ecog.00845

Merow, C., Smith, M. J., & Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling species’ distributions: what it does, and why inputs and settings matter. *Ecography (Copenhagen)*, *36*(10), 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Mäkelä, S. (2021). *Löytökeskittymiä savimailla. Rautakautisten asuiinpaikkojen tutkimus pelloilta* [Master’s thesis, University of Turku]. UTUPub. https://urn.fi/URN:NBN:fi-fe2021121460507

Naimi, B., Skidmore, A. K., Groen, T. A., & Hamm, N. A. S. (2011). Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling: Spatial autocorrelation and positional uncertainty. *Journal of Biogeography*, *38*(8), 1497–1509. https://doi.org/10.1111/j.1365-2699.2011.02523.x

National Land Survey of Finland. (n.d. -a). *Elevation model 2008-2020, 2 m x 2 m* [Data set]. Retrieved December 2, 2021, from https://paituli.csc.fi/download.html

National Land Survey of Finland. (n.d. -b). *Elevation model 2 m*. Retrieved April 3, 2023, from https://www.maanmittauslaitos.fi/en/maps-and-spatial-data/expert-users/product-descriptions/elevation-model-2-m

National Land Survey of Finland. (2016). *Kansallisen maastotietokannan laatumalli – Korkeusmallit*. Retrieved March 31, 2023, from https://www.maanmittauslaitos.fi/sites/maanmittauslaitos.fi/files/attachments/2017/05/KMTK_korkeusmallit_laatukasikirja_2017-01-02.pdf

National Land Survey of Finland. (2021). *Topographic Database* [Data set]. Retrieved December 2, 2021, from https://paituli.csc.fi/download.html

Ortega-Huerta, M. A., & Peterson, A. T. (2008). Modeling ecological niches and predicting geographic distributions: a test of six presence-only methods. *Revista mexicana de biodiversidad*, *79*(1).

Pearson, R. G., & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, *12*(5), 361–371. https://doi.org/10.1046/j.1466-822X.2003.00042.x

Peterson, A.T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, *213*(1), 63–72. https://doi.org/10.1016/j.ecolmodel.2007.11.008

Peterson, A. T. & Soberón, J. (2012a). Integrating fundamental concepts of ecology, biogeography, and sampling into effective ecological niche modeling and species distribution modeling. *Plant Biosystems*, *146*(4), 789–796. https://doi.org/10.1080/11263504.2012.740083

Peterson, A. T. & Soberón, J. (2012b). Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. *Natureza & Conservação*, *10*(2), 102–107.

Phillips, S. J. (2005). A brief tutorial on Maxent. *AT&T Research*, *190*(4), 231-259.

Phillips, S. J. (2017). *A Brief Tutorial on Maxent*. Retrieved April 29, 2022, from http://biodiversityinformatics.amnh.org/open_source/maxent/

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: an open-source release of Maxent. *Ecography (Copenhagen)*, *40*(7), 887–893. https://doi.org/10.1111/ecog.03049

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample Selection Bias and Presence-Only Distribution Models: Implications for Background and Pseudo-Absence Data. *Ecological Applications*, *19*(1), 181–197. https://doi.org/10.1890/07-2153.1

Phillips, S. J., Dudik, M., & Schapire, R. E. (n.d.). *Maxent software for modeling species niches and distributions (Version 3.4.4)* [Computer software]. Retrieved February 2, 2022, from http://biodiversityinformatics.amnh.org/open_source/maxent/

Phillips, S. J., Dudík, M., & Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning* (p. 83).

Phillips, S. & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography (Copenhagen)*, *31*(2), 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Polechová, J. & Storch, D. (2008). Ecological niche. In S. E. Jørgensen, & B. D. Fath, (Eds.). *Encyclopedia of ecology* (pp. 1088–1097). Elsevier.

Rafuse, D. J. (2021). A maxent predictive model for hunter-gatherer sites in the Southern Pampas, Argentina. *Open Quaternary*, *7*(1). https://doi.org/10.5334/OQ.97

Raninen, S. & Wessman, A. (2015). Rautakausi. In G. Haggrén, P. Halinen, M. Lavento, S. Raninen, & A. Wessman, (Eds.). *Muinaisuutemme jäljet: Suomen esi- ja varhaishistoria kivikaudelta keskiajalle* (pp. 215–365). Gaudeamus.

Salo, U. (1995). Aurajokilaakson pronssikautinen ja rautakautinen asutus. Tietoja, tulkintoja ja kysymyksiä. In A. Nissinaho (Ed.). *Ihmisen maisema – Kirjoituksia yhteisön ja ympäristön muutoksesta Lounais-Suomen rannikolla* (pp. 1–45). Åbo Akademi Tryckeri.

Saloranta, E. (2000). Iron Age colonization and land use in the river Vähäjoki valley of Turku (Maaria). In A. Nissinaho (Ed.). *Sites and Settlements* (pp. 15–43). Åbo Akademi Tryckeri.

Seppälä, S. (2000). Rautakautiset kohteet – funktion, ajoituksen ja sijainnin problematiikkaa. In P. Maaranen & T. Kirkinen (Eds.). *Arkeologinen inventointi* (pp. 192–197). Gummerus Kirjapaino Oy.

Soberón, J. (2010). Niche and area of distribution modeling: a population ecology perspective. *Ecography (Copenhagen)*, *33*(1), 159–167. https://doi.org/10.1111/j.1600-0587.2009.06074.x

Soberón, J., & Peterson, T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1444), 689–698.

Stanton, T. W. (2004). Concepts of determinism and free will in archaeology. *Anales de Antropología*, *38*, 29–83.

Tiilikkala, J. (2017). *Rautakauden muinaisjäännökset Kanta-Hämeessä. Paikkatietoaineistojen analyysi* [Unpublished Master's thesis, University of Turku]. TYA.

Tokoi, A. (2020). *Asutuksen jälkiä Turun Kärsämäessä : Asuinpaikan luonne ja sen muutokset kivikauden lopulta rautakaudelle vuosien 2016 ja 2017 kaivaustutkimusten pohjalta* [Master's thesis, University of Turku]. UTUPub. https://urn.fi/URN:NBN:fi-fe202002125248

Turner, M. G., Gardner, R. H., & O'Neill, R. V. (2001). *Landscape ecology in theory and practice: pattern and process*. New York: Springer.

Veregin, H. (1999). 12. Data quality parameters. In P. Longley, M. Goodchild, & D. Maguire, (Eds.). *Geographical Information Systems, 2nd edition, vol 1: Principles and technical issues* (pp. 177–189). John Wiley & Sons. https://www.geos.ed.ac.uk/~gisteac/gis_book_abridged/

Verhagen, P. (2007). *Case studies in archaeological predictive modelling*. Leiden: Leiden University Press.

Yaworsky, Vernon, K. B., Spangler, J. D., Brewer, S. C., & Codding, B. F. (2020). Advancing predictive modeling in archaeology: An evaluation of regression and machine learning methods on the Grand Staircase-Escalante National Monument. *PloS One*, *15*(10), e0239424–e0239424. https://doi.org/10.1371/journal.pone.0239424

Appendices

Appendix 1 Lyhennelmä (summary in Finnish)

Pro Gradu -tutkielma

Historian ja arkeologian tutkinto-ohjelma, Arkeologia
Akseli Tolvi
Maximum Entropy Modeling of the Iron Age Settlement Distributions in River Valleys of Turku Region, Southwest Finland
71 sivua, 4 liitettä

Johdanto

Ennustavat mallinnukset ovat menetelmiä, joiden tavoitteena on ennustaa arkeologisen kulttuuriperinnön sijainteja. Ne ovat käytännöllisiä työvälineitä, joiden avulla voidaan muun muassa rajata inventoitavia alueita. Ennustavaa mallinnusta on tehty arkeologisessa tutkimuksessa Suomessa joitakin vuosikymmeniä. 1990-luvulla Tuija Kirkinen teki aiheesta varhaisia tutkimuksia ja viime vuosina aihetta ovat tutkineet muun muassa Jasse Tiilikkala ja Annukka Debenjak-Ijäs pro gradu –tutkielmissaan.

Tämän tutkielman konsepti rakentui mainittujen aikaisempien suomalaisten mallinnusten pohjalle. Tavoitteena oli testata luonnontieteellisessä tutkimuksessa käytettyjen menetelmien käyttöä arkeologisessa kontekstissa. Lajien levinneisyysmallit (tästä eteenpäin SDM) ovat ennustavien mallinnusten laji, jota käytetään runsaasti muun muassa ekologisessa tutkimuksessa. Tutkimuskohteeksi valittiin rautakautiset asuinpaikat.

Tutkimusalue sijaitsi Varsinais-Suomessa ja oli rajaukseltaan 24 km * 24 km neliö, joka kattoi kiinnostuksen kohteina olleet Aurajoki-, Vähäjoki- ja Savijokilaaksot ja sisälsi osia Auran, Kaarinan, Liedon, Maskun, Nousiaisten, Paimion, Raision, Ruskon ja Turun kunnista. Tutkimuskysymyksinä olivat, pystyvätkö SDM-algoritmit tunnistamaan ihmisasutukselle suotuisia ympäristöjä ja löytämään yhteyksiä ympäristötekijöiden ja muinaisjäännösten sijainnin välillä, mikä muuttujien yhdistelmä tuottaa parhaan lopputuloksen ja mitkä muuttujat selittävät rautakautisten asuinpaikkojen sijaintia parhaiten.

Aineistona käytettiin avointa dataa. Asuinpaikkojen sijainti johdettiin Museoviraston Muinaisjäännösrekisteristä ja ympäristömuuttujat Maanmittauslaitoksen, GTK:n ja SYKE:n avoimista paikkatietoaineistoista. Tutkielmassa käytettiin pääasiassa avoimeen lähdekoodiin perustuvia ohjelmistoja. Aineiston valmistelu tehtiin pääosin QGIS:llä ja Excelillä, minkä lisäksi joitakin analyysejä tehtiin ArcGIS Pro:lla. Mallinnus tehtiin R-ympäristössä, käyttäen MaxEnt 3.4.4-ohjelmaa dismo-kirjaston rajapintafunktion kautta.

Tausta

Tutkielman aikarajauksena oli tarkemmin määrittelemätön, suomalaisen kronologian mukainen rautakausi. Rajaus on lavea, sillä kausi kattaa lähes 2000 vuoden ajanjakson, alkaen n. 500 eaa. ja päättyen n. 1200 jaa. Aikajaksolle mahtuu lukuisia ympäristön ja elinkeinojen muutoksia. Ympäristön kehitykseen on vaikuttanut muun muassa maan kohoaminen, jonka vaikutus tutkimusalueella on tosin melko vähäinen. Myös yhteiskuntarakenne on muuttunut muun muassa asutuksen siirtymisen ja kyläasutuksen synnyn vuoksi. Lavea aikarajaus valittiin, koska muinaisjäännösrekisterin kohteiden iän määrittäminen on usein tehty pelkästään muutamien pintapoimintalöytöjen perusteella. Tarkempaa ajoitusta ei tämän vuoksi ole saatavilla ilman lisätutkimuksia, jotka eivät olleet toteutettavissa tämän tutkielman puitteissa.

Tutkielmassa käytetyt SDM-menetelmät perustuvat hutchinsonilaiseen ekolokeroteoriaan, jonka mukaan lajin peruslokero voidaan esittää joukkona ympäristötekijöitä tai -muuttujia näiden määrittämässä avaruudessa. Sen mukaan laji ja sitä määrittävät ympäristötekijät sijaitsevat kaksi- tai kolmiulotteisessa, koordinaattien määrittämässä maantieteellisessä avaruudessa G ja siirtämällä ne ympäristötekijöiden määrittämään avaruuteen E saadaan muodostettua lajin peruslokero. Kun avaruudessa E mallinnettu peruslokero siirretään takaisin avaruuteen G, saadaan mallinnettua lajin maantieteellinen levinneisyys eli luotua ennustava malli.

Tässä tutkielmassa käytettiin koneoppimista hyödyntävää Maximum Entropy eli MaxEnt -algoritmia. MaxEnt on presence-background algoritmi, eli se vertaa lajin esiintymissijaintien olosuhteita taustan, eli koko tutkimusalueen, olosuhteisiin. Se on nykyisin käytetyistä algoritmeista suosituin, koska se kykenee tuottamaan edustavia malleja tarvitsematta tietoa paikoista, joissa tutkittavaa lajia ei ole tavattu. Siinä on myös sääntelytoimintoja, joilla pyritään rajoittamaan muusta aineistosta poikkeavien havaintojen vaikutusta lopputulokseen. Algoritmin on lisäksi havaittu toimivaksi myös arkeologisen aineiston mallintamisessa.

Aineistojen valmistelu

Mallinnusprosessi aloitettiin käytettävien aineiston valmistelulla. Vastemuuttujaksi valmisteltiin luettelo tutkimusalueen rautakautisista asuinpaikoista. Pohja-aineistona käytettiin Museoviraston kulttuuriympäristörekisterien tietoja (haettu 19.12.2021). Tutkimusalueella sijaitsevat rautakautiset asuinpaikat poimittiin koko aineistosta SQL-kyselyllä, minkä tuloksena syntynyt luettelo käytiin läpi tutkimusraporttien perusteella. Lopullinen lista koostui 39 rautakautisesta asuinpaikkakohteesta. Lisäksi luotiin 10000 satunnaista havaintopistettä kuvaamaan ympäristön tausta-arvoja.

Selittäviksi muuttujiksi valmisteltiin neljätoista rasteritasoa kuvaamaan yhdeksää eri muuttujaa. Muuttujat perustuivat Maanmittauslaitoksen Korkeusmalli 2 m ja Maastotietokanta, GTK:n Maaperä 1:20000/1:50000 sekä SYKE:n Ranta10 – Uomaverkosto –aineistoihin. Valitut muuttujat olivat absoluuttinen korkeus (DEM), suhteellinen korkeus (TPI), rinnekaltevuus (slope), rinteen suunta (aspect), pintamaalaji, auringonsäteilykertymä, topografinen kosteusindeksi (TWI), sekä kustannusetäisyys ja euklidinen etäisyys vesistöön. Näistä suhteellista korkeutta kuvaavia tasoja valmisteltiin kaksi, toinen kuvaamaan lähimaastoa 200 metrin säteellä ja toinen etäistä maastoa 2000 metrin säteellä kohteesta. Lisäksi sekä euklidista että kustannusetäisyyttä vesistöön kuvaavia tasoja kolme vastaamaan rautakauden alku-, keski- ja loppuvaiheiden vedenpinnan tasoja.

Valmistelun jälkeen aineistojen laatua arvioitiin virhelähteiden ennakoimiseksi. Vastemuuttujassa merkittävä virhelähde oli ajoituksen epävarmuuden lisäksi tutkimusoletusten, sekä aineiston keruutapojen mahdollisesti aiheuttamat vääristymät. Lisäksi kohteiden kuvaaminen pisteenä voi aiheuttaa vääristymää, mikäli piste ei sijaitse kohteen kannalta edustavimmassa paikassa. Selittävissä muuttujissa virhelähteitä olivat puutteet niiden temaattisessa kattavuudessa, sekä se, että ne kuvaavat ympäristöä aineiston keruuhetkellä. Tasoja valmisteltaessa aineistoista poistettiin muun muassa nykyiset varastoaltaat, mutta muutoin rautakauden ympäristöä ei rekonstruoitu.

Muuttujia tarkasteltiin lisäksi analysoimalla niiden merkitsevyyttä, vasteita sekä tarkastelemalla niiden välisiä korrelaatioita. Merkitsevyyttä testattiin MaxEntin taittoveitsi-testillä (jackknife). Taittoveitsi-testissä algoritmi tekee jokaista muuttujaa kohden kaksi mallia: muuttujaa käytetään vuoroin ainoana muuttujana, ja vuoroin jätetään ainoana pois mallinnettavien muuttujien joukosta. Testissä kustannusetäisyys vesistöön tuotti parhaan mallin yksinään käytettynä, kun taas absoluuttisen korkeuden pois jättäminen heikensi tulosta eniten.

Vasteita tarkasteltiin MaxEntin luomien vastekäyrien pohjalta, sekä visualisoimalla asuinpaikat ja muuttujat kolmiulotteisiin sirontakuvioihin. Lukuun ottamatta rinteen suuntaa, kaikista käyristä pystyttiin havaitsemaan selkeitä preferenssejä asuinpaikkojen sijoittumisen suhteen. Selkeimmin mieltymykset näkyivät etäisyyttä vesistöön kuvaavalla käyrällä, jossa lähellä nollaa olevat arvot olivat suotuisimpia. Sirontakuvioista tehdyt havainnot tukivat vastekäyrien havaintoja.

Muuttujien välisiä korrelaatioita tarkasteltiin korrelaatiomatriiseista. Analyysissä havaittiin sekä positiivisia että negatiivisia korrelaatioita, jotka olivat pääosin matalia tai keskitasoisia. Korrelaatiot kahden merkitsevimmän muuttujan, absoluuttisen korkeuden ja etäisyyden vesistöön välillä olivat korkeahkot, minkä vuoksi näiden käyttämistä samassa mallissa pyrittiin välttämään.

Mallinnusprosessi

Mallit ajettiin sallien MaxEntin käyttää kaikkia sisältämiään toimintoja, eli lineaarinen, neliöllinen, tulo, kynnys, taite sekä luokkaindikaattori, vasteiden luomiseen monimutkaisten vuorovaikutusten mahdollistamiseksi. Iteraatioiden enimmäismäärä nostettiin oletuksesta viiteentuhanteen, jotta algoritmi saavuttaisi aina parhaan mahdollisen tuloksen. Vastemuuttujaksi asetettiin luettelo rautakautisista asuinpaikoista ja taustaksi luodut 10000 satunnaispistettä. Muuttujat asetettiin mallin kokoonpanon mukaan. Havaintopisteiden pienehkön määrän vuoksi päätettiin mallien koulutuksessa käyttää niitä kaikkia. Mallien kokoonpano testattiin kuitenkin ennen varsinaista ajoa k-kertaisella ristiinvalidoinnilla, minkä jälkeen lopullinen malli ajettiin käyttäen kaikkia vastemuuttujan havaintopisteitä.

Malleja arvioitiin tilastollisesti AUC-arvon avulla ja visuaalisesti. AUC-arvo vaihtelee 0 ja 1 välillä ja kertoo, kuinka todennäköisesti malli arvottaa satunnaisesti valitun positiivisen havainnon satunnaisesti valitun negatiivisen havainnon yläpuolelle. Käytännössä AUC-arvo 0 kertoisi mallin ennustavan aina väärin ja 1 mallin ennustavan aina oikein, 0,5 ollessa täysin satunnainen. Yleensä arvoja 0,5–0,7 pidetään heikkona, 0,7–0,9 keskivertona ja yli 0,9 erinomaisena mallin suorituskykynä.

Mallien ajossa päätettiin tarkastella kolmea kysymystä. Ensimmäisessä testattiin monimutkaista kaikkia muuttujia käyttävää kokoonpanoa ja yksinkertaista kokoonpanoa, josta oli poistettu kolme vähiten merkitsevää muuttujaa. Toisessa testattiin eroja eri merenpinnan tasoja ja eri menetelmillä laskettujen, etäisyyttä vesistöön kuvaavien tasojen välillä. Kolmannessa taas testattiin absoluuttista ja suhteellista korkeutta, sekä erilaisia näitä kuvaavien tasojen yhdistelmiä käyttävien mallien välisiä eroja.

Tulokset

Ensimmäisessä testissä tehtiin kaksi mallia, joista molemmissa käytettiin kumpaakin TPI tasoa kuvaamaan suhteellista korkeutta, auringonsäteilykertymää, topografista kosteusindeksiä, auringonsäteilykertymää ja kustannusetäisyyttä vesistöön 3 m mpy vedenpinnan tasolla. Monimutkaista kokoonpanoa testaavassa mallissa käytettiin lisäksi muuttujina pintamaalajia, rinteen kaltevuutta ja rinteen suuntaa. Mallien erot olivat pienet, minkä vuoksi niistä laskettiin erotus tarkastelun helpottamiseksi.

Yksinkertaistettu malli sai tilastollisessa testissä AUC-arvon 0,955, ja monimutkainen arvon 0,958. Visuaalisesti suurimmat erot mallien välillä olivat taajama-alueiden puuttuminen monimutkaisesta mallista, koska ne oli luokiteltu pintamaalaji-muuttujalla puuttuvaksi tiedoksi (NULL). Muutoin suurimmat erot olivat monimutkaisen mallin kohinaisuus, mikä näkyi mallien erotuksessa yksinkertaisen mallin korkeampina arvoina, sekä muutama laajempi yhtenäinen alue jokilaaksoista, joilla monimutkainen malli sai korkeampia arvoja. Pääosin jälkimmäiset sijoittuivat alueille, joilla moderni maankäyttö aiheuttaa häiriöitä, minkä vuoksi eroavaisuudet tulkittiin käytännössä merkityksettömiksi. Tämän vuoksi seuraavista malleista päätettiin jättää rinnekaltevuutta ja rinteiden suuntaa kuvaavat muuttujat pois. Pintamaalaji-muuttujaa päätettiin kuitenkin käyttää.

Toisessa testissä tehtiin yhteensä seitsemän mallia: kolme mallia käyttäen euklidista etäisyyttä vesistöön, yksi kutakin rautakauden vedenpinnan tasoa kohden, sekä kolme käyttäen kustannusetäisyyttä. Lisäksi tehtiin yksi malli, jossa etäisyys vesistöön -muuttujaa ei käytetty. Samalla menetelmällä laskettuja etäisyyksiä käyttävät mallit olivat keskenään hyvin samanlaisia. Kustannusetäisyyttä käyttävien mallien AUC-arvot vaihtelivat 0,955:n ja 0,957:n välillä, euklidista etäisyyttä käyttävien hieman matalampien 0,941:n ja 0,942:n välillä. Malli, jossa muuttujaa ei käytetty sai AUC-arvon 0,898, eli sekin jäi ainoastaan 0,002 pisteen päähän erinomaisena pidetyn suorituskyvyn rajasta.

Suotuisimmat alueet sijoittuivat kaikissa malleissa pääosin jokien varsille ja erot olivat jälleen pienehköjä. Eri vedenpinnan tasoista, samalla menetelmällä laskettuja etäisyyksiä käyttävien mallien väliset erot olivat suurimmillaan Turun kaupunkialueella, jota ei mallinnettu pintamaalaji-tiedon puuttumisen vuoksi. Eri menetelmien vertailussa havaittiin, että euklidista etäisyyttä käyttävissä malleissa suotuisa alue ulottui hiukan kauemmas, joista erityisesti tutkimusalueen itäosissa, kuin kustannusetäisyyttä käyttävissä malleissa, todennäköisesti jokiuoman rinteiden korkeuden ja jyrkkyyden vaihteluiden vuoksi. Myös mallissa, jossa etäisyys vesistöön -muuttujaa ei käytetty, suotuisimmat alueet sijoittuivat jokien varsille, todennäköisesti TPI ja TWI muuttujien vaikutuksesta.

Kolmannessa testissä tehtiin kuusi mallia käyttäen eri absoluuttisen ja suhteellisen korkeuden kokoonpanoja. Ensimmäiset neljä mallia käyttivät korkeuden kuvaamiseen DEM-tasoa, molempia TPI tasoja yhtaikaa, lähimaastoa ja etäistä maastoa kuvaavia TPI tasoja erikseen. Viimeisistä kahdesta mallista toinen käytti kaikkia korkeutta kuvaavia tasoja yhtaikaa, ja toinen ei käyttänyt niistä mitään. Korkeimman AUC-arvon, 0,975, saavutti malli, joka käytti kaikkia tasoja yhtaikaa. Pelkästään absoluuttista korkeutta käyttävä malli saavutti arvon 0,969, molempia TPI tasoja käyttävä 0,942, lähimaaston TPI:tä käyttävä 0,937 ja etämaaston TPI:tä 0,933. Heikoimmin tilastollisessa testissä menestynyt malli, jossa korkeutta kuvaavia tasoja ei käytetty, sai AUC-arvon 0,930.

Visuaalisesti suurimmat erot mallien välillä olivat absoluuttista korkeutta käyttävien ja sitä käyttämättömien mallien välillä. Ensimmäisissä suotuisa alue on leveämpi jokien alajuoksuilla, ja kapenee yläjuoksua kohti, kun taas jälkimmäisissä suotuisa alue on suurin piirtein tasalevyinen koko tutkimusalueella. Absoluuttista korkeutta käyttävät mallit seurailivat paremmin nykyistä käsitystä rautakautisesta asutuksesta, mikä kertoo muuttujan todennäköisesti sisältävän epäsuorasti tietoa, jota suhteellista korkeutta kuvaavissa muuttujissa ei ole. Yksi mahdollisuus on, että etäisyys joen suuhun on vaikuttanut asutuksen sijoittumiseen.

Lopuksi kaikista käsitellyistä malleista laskettiin yhteenvedoksi keskiarvoistettu malli. Mallissa jokaiselle edellä käsitellyistä kysymyksistä annettiin sama painoarvo, minkä vuoksi jokaisen testin malleista luotiin ensin erikseen keskiarvoistetut mallit, minkä tuloksena syntyneille kolmelle mallille laskettiin keskiarvo, jota käytettiin keskustelussa suotuisien alueiden sijoittumisesta tutkimusalueelle.

Mallista tunnistettiin kuusi suhteellisen yhtenäistä korkean suotuisuuden aluetta. Näistä erityisesti kaksi, Aurajoen varrella Mäkkylän ja Nautelan alueella ja Savijoen varrella Raukkalan ja Pränikkälän alueella, vaikutti lupaavalta, sillä kulttuuriympäristörekisterin tietojen mukaan niiltä on löydetty rautakautisia irtolöytöjä. Lisäksi tarkasteltiin kahta viime vuosina tarkemmin tutkittua, mallien koulutukseenkin käytettyä rautakautista asuinpaikkakohdetta ja tarkasteltiin miten niiden löytöalueet vastaavat ennustettua suotuisaa aluetta. Kaarinan Ravattulan Salorannan kohteella ennustus vastasi hyvin tarkasti pintapoiminnassa havaitun asuinpaikan laajuutta. Liedon Pettisten Hanhiojan kohteella ennustettu alue oli jonkin verran pienempi kuin todellinen löytöalue.

Keskustelu

Erityisesti yhteenvetomallista tehtyjen havaintojen perusteella tulokset vaikuttivat lupaavilta, mutta niiden tarkempi arvioiminen vaatisi lisätutkimuksia ja erityisesti havaintojen todentamista kentällä. Myös joitakin ongelmia havaittiin. Esimerkiksi alueita, joilla sijaitsee mallien koulutuksessakin käytettyjä rautakautisia asuinpaikkoja, oli säännönmukaisesti jäänyt ennustetun suotuisan alueen ulkopuolelle. On mahdollista, että ne syystä tai toisesta erosivat muista kohteista niin paljon, että MaxEnt arvioi poikkeamiksi. Poikkeavuus voisi olla seurausta esimerkiksi kohteiden ajoittumisesta eri tavoin kuin muut tai erilaisista elinkeinoista.

Tutkielmassa tehdyissä malleissa etäisyys vesistöön ja absoluuttinen korkeus olivat merkitsevimmät muuttujat, mikä ei kuitenkaan tarkoita niiden olevan hyviä muuttujia. Käytetyt muuttujat perustuivat avoimesti saatavilla olleisiin topografisiin aineistoihin. Niiden etuna on käytännössä globaali saatavuus, mutta ennustavassa mallinnuksessa on huomioitava, että ne kuvaavat ainoastaan yhtä ympäristön piirrettä. Topografiset aineistot eivät esimerkiksi kuvaa, miten tutkittava laji käyttää ympäristöään tai on vuorovaikutuksessa sen kanssa. Nykyisin laajalle levinnyt, muun muassa paleoekologisessa tutkimuksessa käytetty ymmärrys on, että ympäristö asettaa ihmistoiminnalle rajoja, joiden sisällä ihminen voi toimia vapaan tahdon mukaan. Ihminen on kuitenkin samalla hyvin sopeutumiskykyinen laji, joka pystyy tarvittaessa levittämään näitä rajoja. Kaikki tämä vaikeuttaa vuorovaikutusten tunnistamista.

Tutkielmassa tehdyt mallit edustavat datalähtöistä lähestymistapaa, joissa tutkimusaineiston valinta perustuu usein sen saatavuuteen. Tämän vuoksi datalähtöisissä tutkimuksissa aineisto ei aina ole täysin tarkoituksen mukainen, minkä lisäksi se voi olla temaattisesti puutteellinen. Mahdollisissa jatkotutkimuksissa aineiston laatuun ja temaattiseen kattavuuteen panostaminen olisikin varmin keino parantaa ennustusten tarkkuutta. Inventoinnin suunnittelun apuna käytettynä mallit itsessään voivat toimia keinona aineiston laadun parantamiseen ja aktiivisen tutkimuksen ja harrastustoiminnan vuoksi Varsinais-Suomi tarjoaa oivalliset puitteet testata tätä käytännössä.

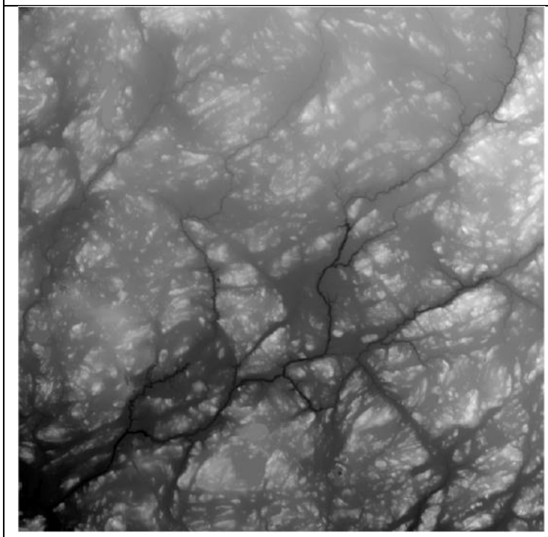
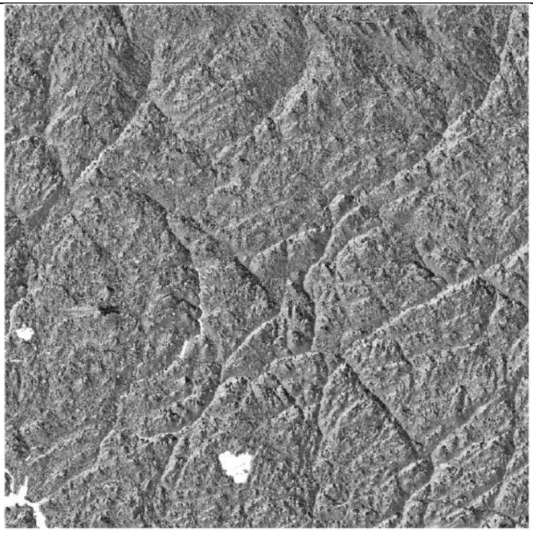
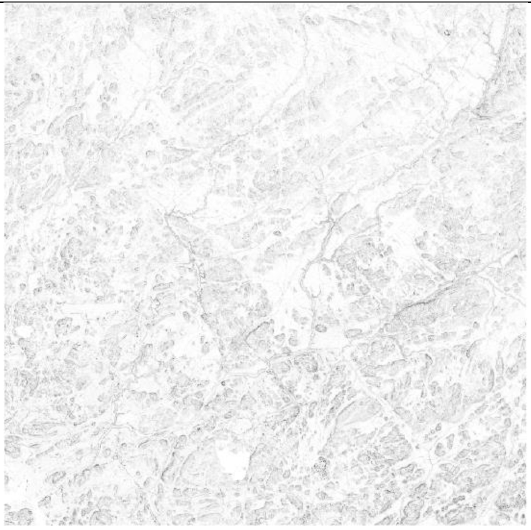
Appendix 2 Training sites

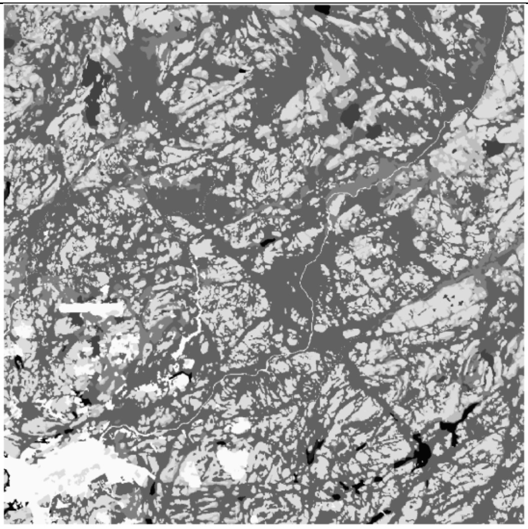
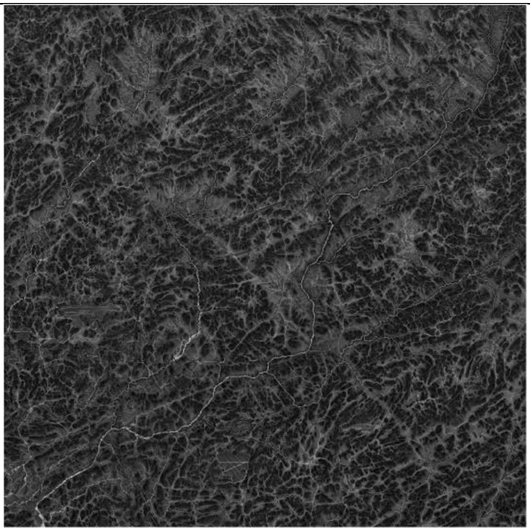
Table 3. All sites used in model training. Listed are site ID (fi. Mjtunnus) in the cultural environment registers, site name and site coordinates in ETRS89 / TM35FIN(E,N), EPSG:3067 reference system.

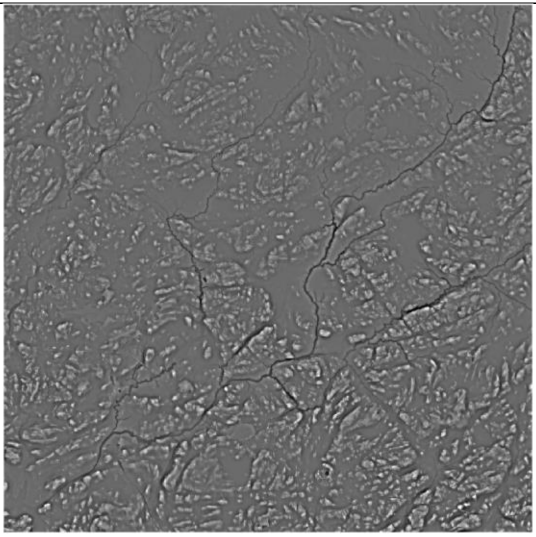
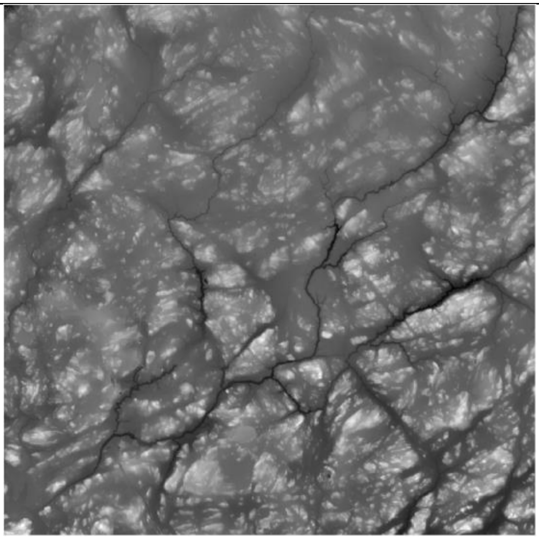
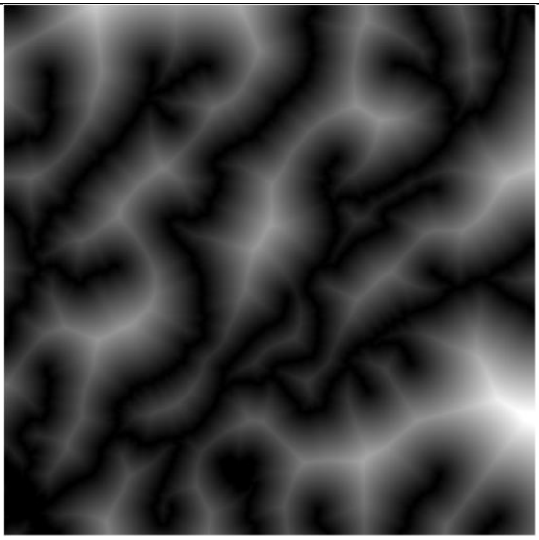
Site ID	Site name	Easting	Northing
853010066	Alfa II	240853	6712539
423010050	Alitalo	249462	6716210.602
1000021179	Ammatillinen kurssikeskus	241039	6714106
853010063	Haaga	245494	6714305
423010053	Hanhioja	251579.769	6715907.579
853010035	Hankkismäki	244340	6716850
1000019440	Hiidenkartano	241555	6714560
853010038	Hillamäki	244600	6717860
423010052	Karvala	251115.34	6715889.736
853010024	Kaupunkitätilä	243318.489	6713075.824
853010023	Komonen 2	242822	6712823
853010007	Koroisten tila	241186	6712586
1000011418	Kukonharja	245830.469	6715434.888
423000004	Kyläkallio	252444.802	6716650.441
1000019444	Kärsämäen Marttila	241450	6714590
853010015	Kärsämäki Radanvierusta	241606	6714610
1000011421	Laurila 2	246444.223	6715227.977
1000032908	Linnavuori 2	244501	6713535
1000011111	Maarian kirkon lounaispuoleinen pelto	241200	6713220
1000039446	Marttila	251317	6715922
853010032	Marttilan vasikkahaka	243390	6715960
423010009	Mikola	247456.816	6714606.23
1000001405	Paaskunnan asuinpaikka	242290	6711960
1000021082	Pappilanpelto	241110	6713630
1000012301	Rantapelto	246167.174	6714938.75
423010033	Ryökäs	250644.505	6720360.941
1000041861	Saloranta	244472	6712824
853010025	Sipipelto	243023	6712568
1000011424	Säteri 2	247232.904	6715386.924
853010067	Taskulan ranta	241040	6713200
202010023	Tähkäpää 2	245384.658	6713368.723
423010058	Uotila	247441.818	6714989.083
1000003008	Virnamäenpuiston ranta	243033.605	6712401.099
853010011	Virnamäki 1	242727	6712494
853010064	Vähätalon mäki/Ihamuotilan kylätontti	242719.713	6715309.925
1000025745	Yli-Junnila	244890	6717030
1000028108	Ylirihko	245577	6715695
853010009	Yrjönmäki	243159	6715430
423010055	Äyräs	252423.81	6716853.356

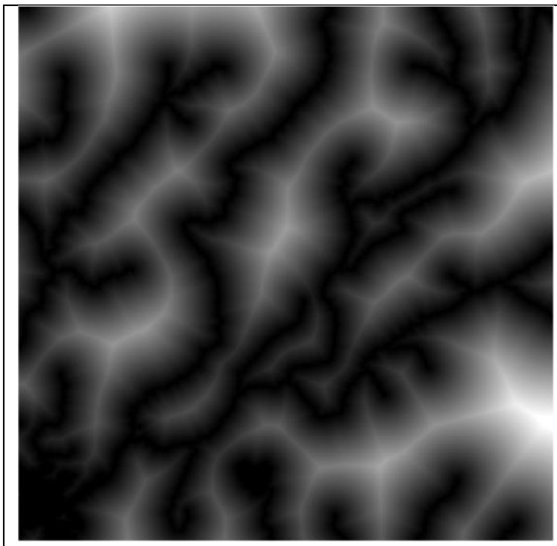
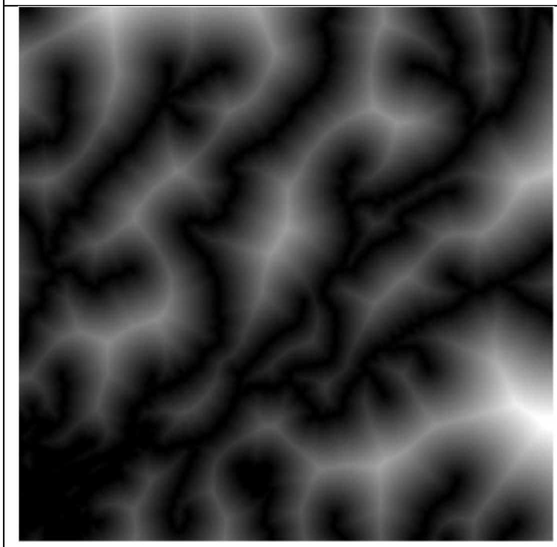
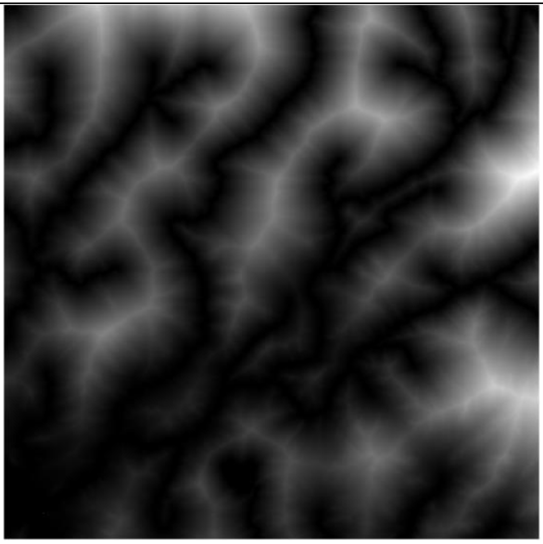
Appendix 3 Predictor variables

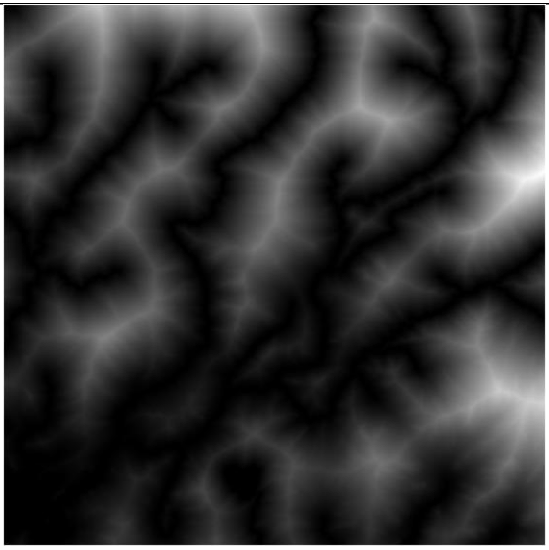
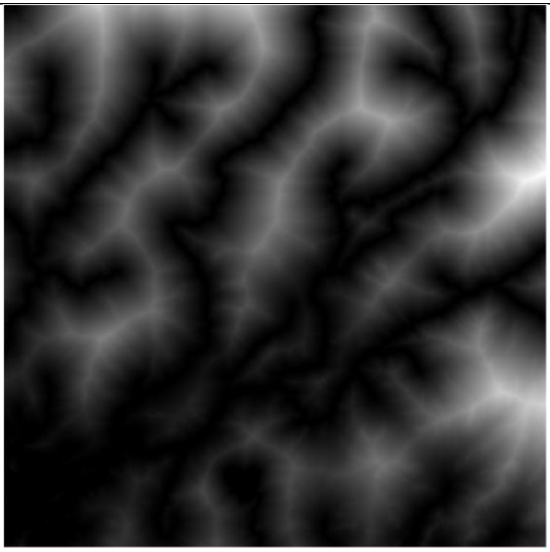
Table 4. Predictor variables used in this study.

Image	Information
	Variable: Absolute elevation Name: DEM Lineage: Elevation model 2m (NLS, n.d. -a) Grain: Original 2 m, resampled 20 m
	Variable/Name: Aspect Lineage: Elevation model 2m (NLS, n.d. -a) Grain: Original 2 m, resampled 20 m
	Variable/Name: Slope Lineage: Elevation model 2m (NLS, n.d. -a) Grain: Original 2 m, resampled 20 m

	Variable/Name: Topsoil class Lineage: Superficial deposits 1:20 000/1:50 000 (GTK, 2015) Grain: 20 m
	Variable/name: Solar irradiance Lineage: Elevation model 2m (NLS, n.d. -a) Grain: 20 m
	Variable: Topographic Wetness Index Name: TWI Lineage: Elevation model 2m (NLS, n.d. -a) Grain: 20 m

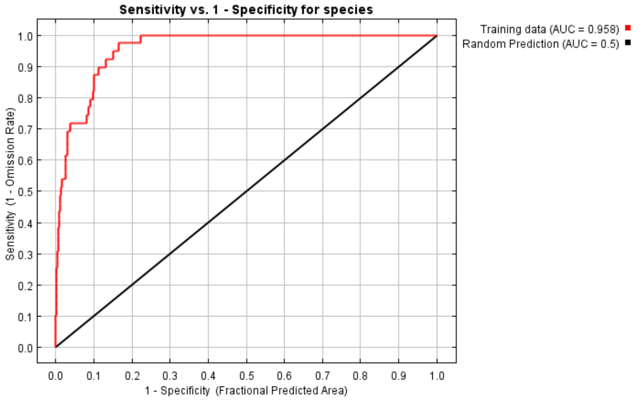
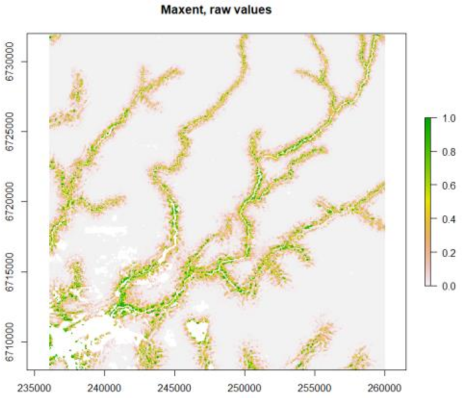
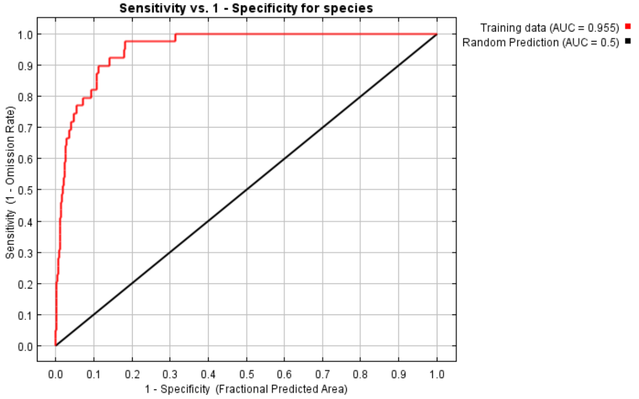
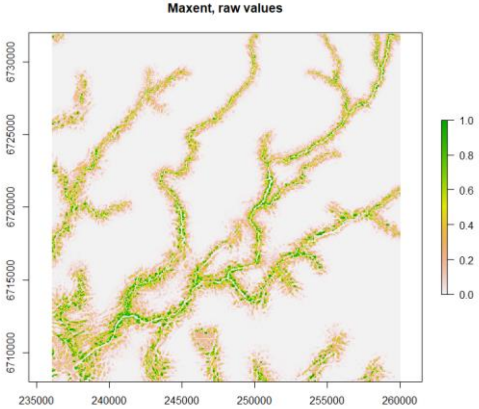
	Variable: Topographic Position Index (near) Name: TPI near Lineage: Elevation model 2m (NLS, n.d. -a) Grain: 20 m
	Variable: Topographic Position Index (far) Name: TPI near Lineage: Elevation model 2m (NLS, n.d. -a) Grain: 20 m
	Variable: Distance to water (Euclidean, 3 m) Name: Dist. to water (3m Euc) Lineage: Topographic database (NLS, 2021); Shoreline – River Network (Ranta10) (SYKE, 2012) Grain: Original 2 m, resampled 20 m

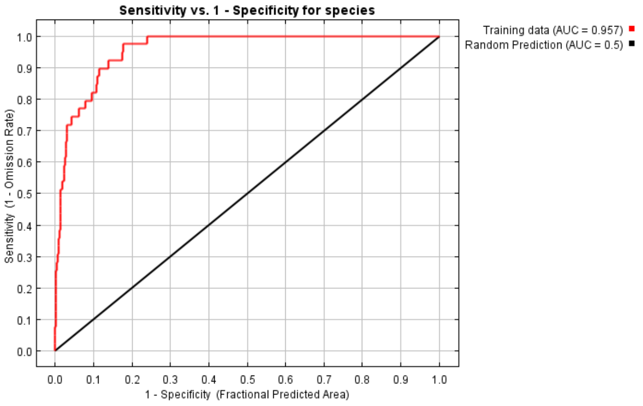
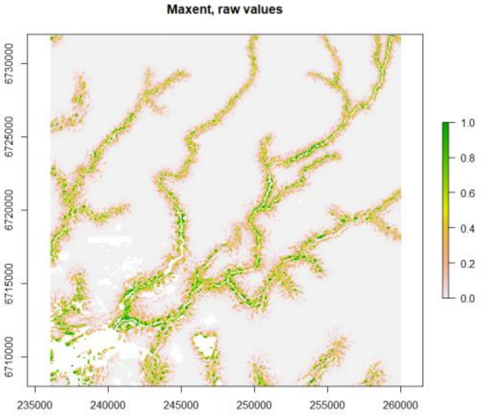
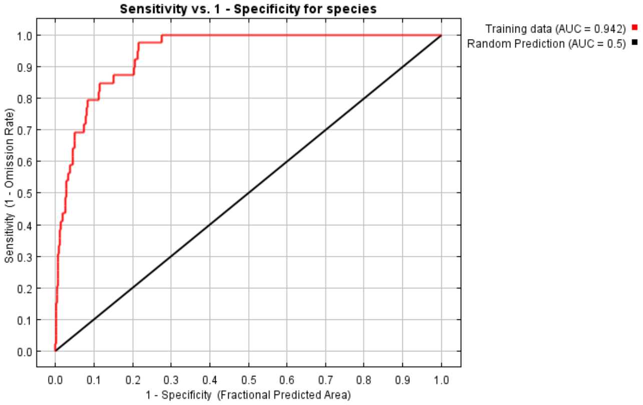
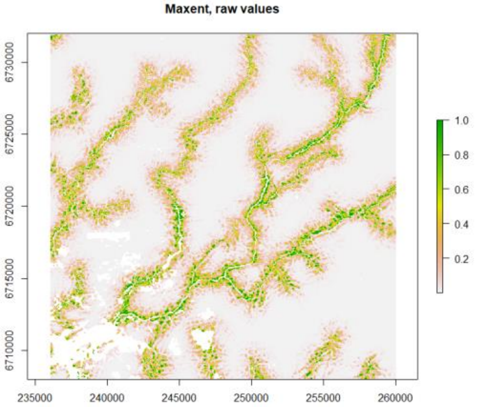
	Variable: Distance to water (Euclidean, 6 m) Name: Dist. to water (6m Euc) Lineage: Topographic database (NLS, 2021); Shoreline – River Network (Ranta10) (SYKE, 2012) Grain: Original 2 m, resampled 20 m
	Variable: Distance to water (Euclidean, 11 m) Name: Dist. to water (11m Euc) Lineage: Topographic database (NLS, 2021); Shoreline – River Network (Ranta10) (SYKE, 2012) Grain: Original 2 m, resampled 20 m
	Variable: Distance to water (Cost distance, 3 m) Name: Dist. to water (3m cst) Lineage: Elevation model 2m (NLS, n.d. -a); Topographic database (NLS, 2021); Shoreline – River Network (Ranta10) (SYKE, 2012) Grain: Original 2 m, resampled 20 m

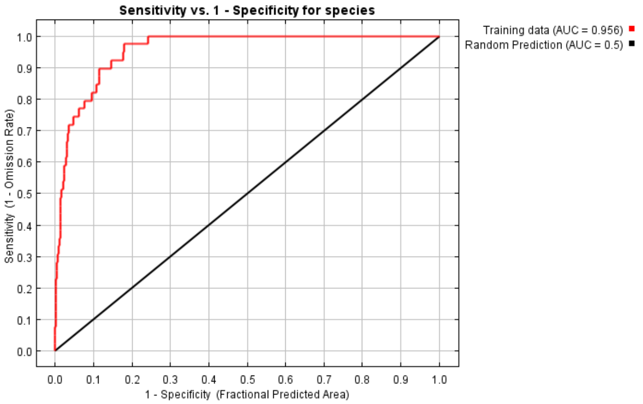
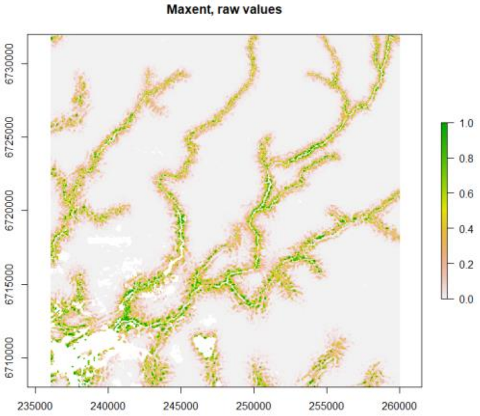
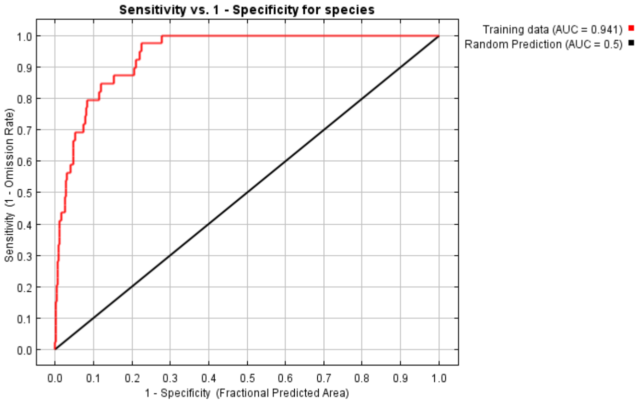
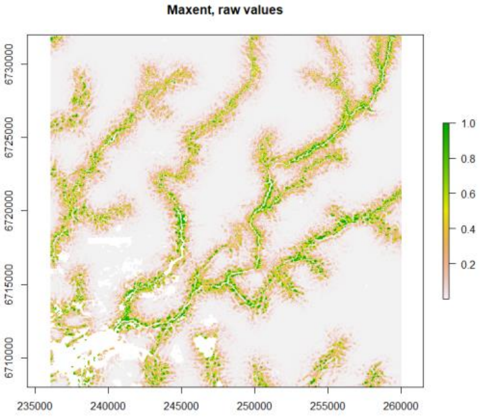
	Variable: Distance to water (Cost distance, 6 m) Name: Dist. to water (6m cst) Lineage: Elevation model 2m (NLS, n.d. -a); Topographic database (NLS, 2021); Shoreline – River Network (Ranta10) (SYKE, 2012) Grain: Original 2 m, resampled 20 m
	Variable: Distance to water (Cost distance, 11 m) Name: Dist. to water (11m cst) Lineage: Elevation model 2m (NLS, n.d. -a); Topographic database (NLS, 2021); Shoreline – River Network (Ranta10) (SYKE, 2012) Grain: Original 2 m, resampled 20 m

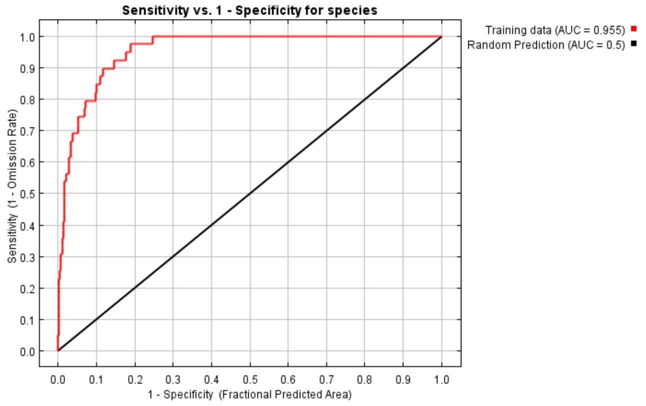
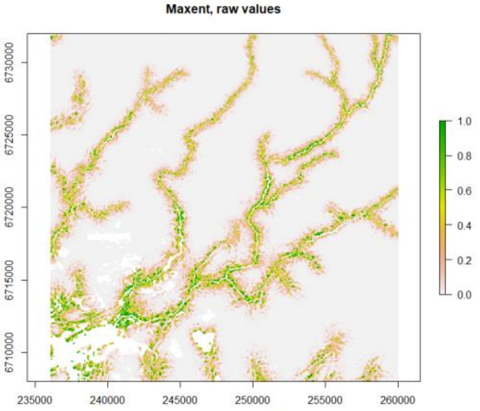
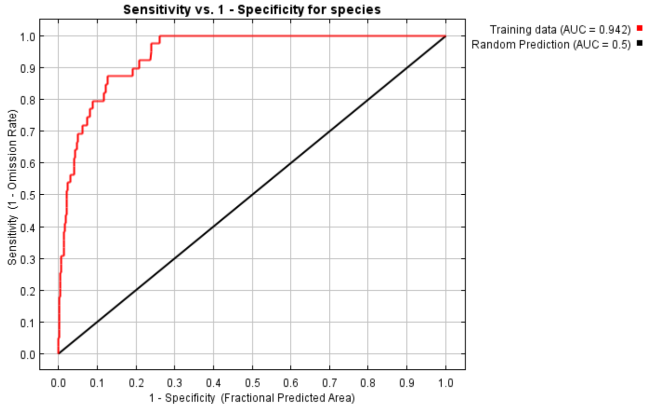
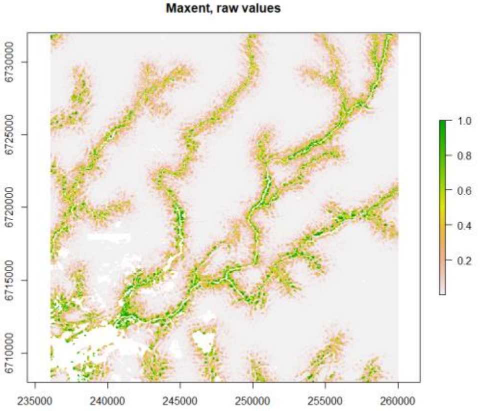
Appendix 4 Models

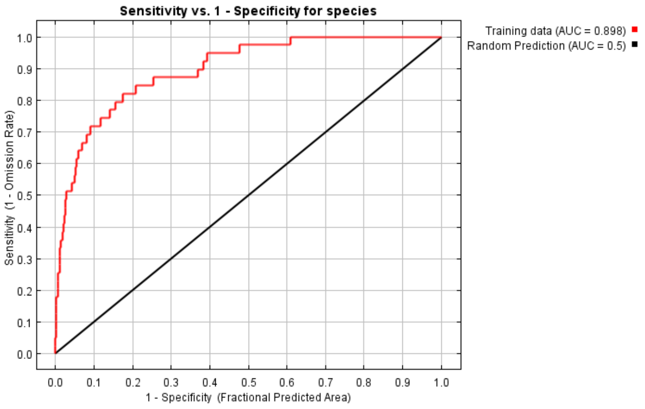
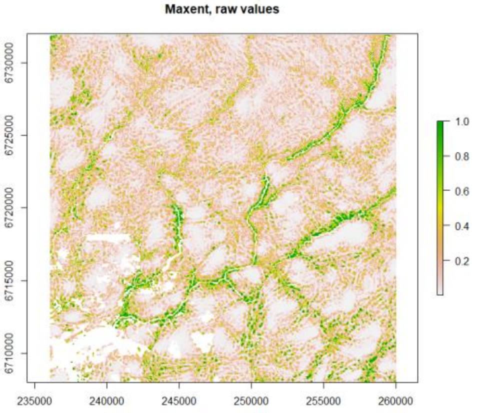
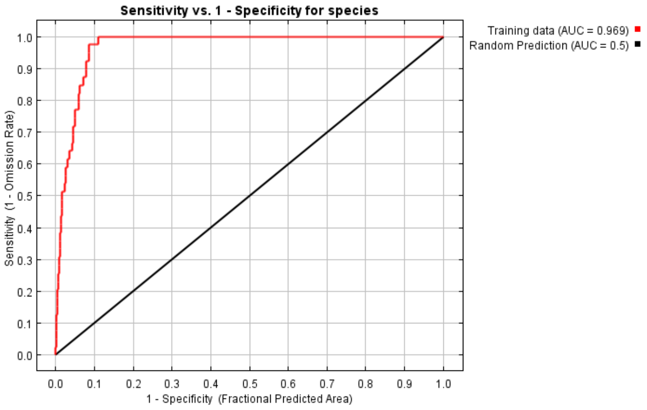
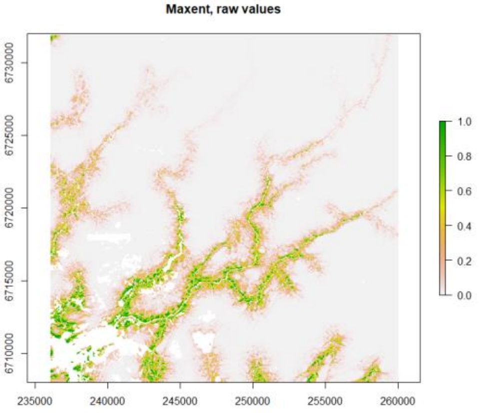
Table 5. Models discussed in the text.

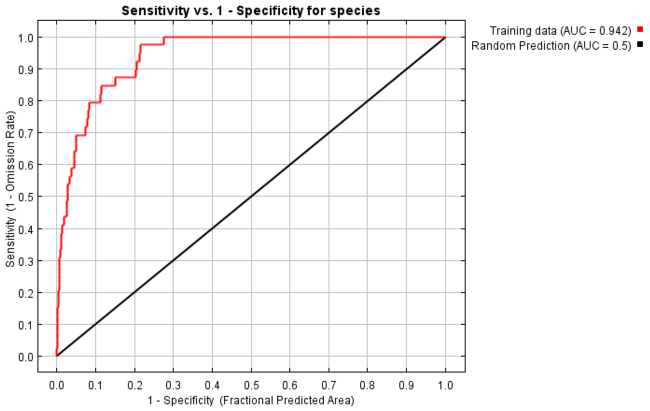
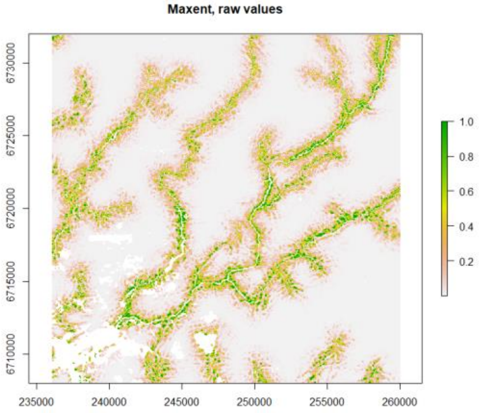
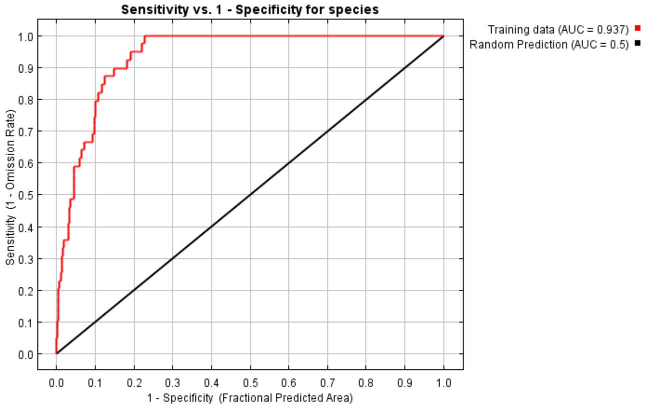
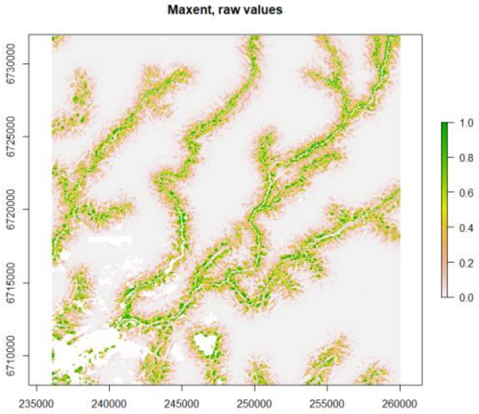
Name	Variables	Additional information	ROC-curve	Image (R plot)
01_complex_A	Topsoil class Aspect Slope Solar irradiance TWI TPI far TPI near DEM Dist. to water (3m cst)	AUC: 0.958 Features: L, Q, P, T, H		 Maxent, raw values 6730000 6725000 6720000 6715000 6710000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2 0.0
02_simple_B	Solar irradiance TWI TPI far TPI near DEM Dist. to water (3m cst)	AUC: 0.955 Features: L, Q, P, T, H		 Maxent, raw values 6730000 6725000 6720000 6715000 6710000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2 0.0

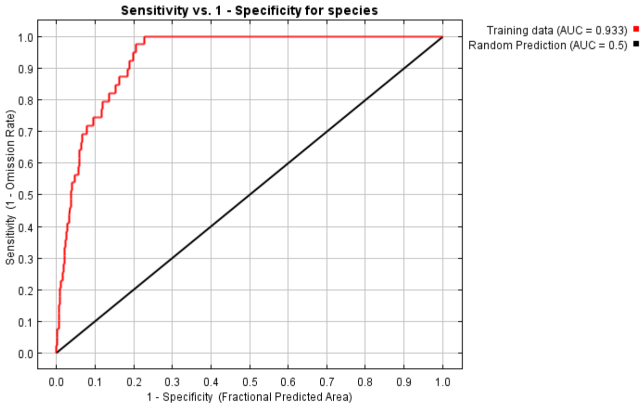
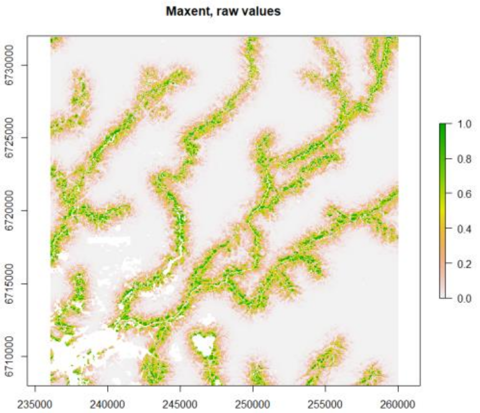
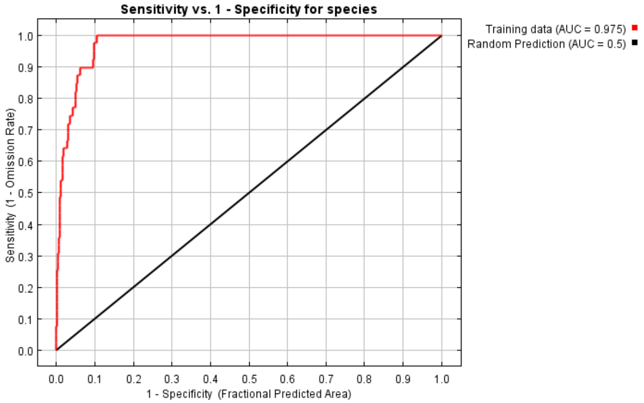
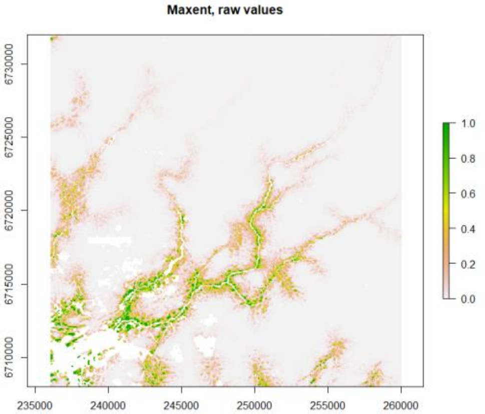
03_watertest_A	Topsoil class Solar irradiance TWI TPI far TPI near Dist. to water (3m cst)	AUC: 0.957 Features: L, Q, P, T, H		 Maxent, raw values 6730000 6725000 6720000 6715000 6710000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2 0.0
04_watertest_B	Topsoil class Solar irradiance TWI TPI far TPI near Dist. to water (3m Euc)	AUC: 0.942 Features: L, Q, P, T, H		 Maxent, raw values 6730000 6725000 6720000 6715000 6710000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2

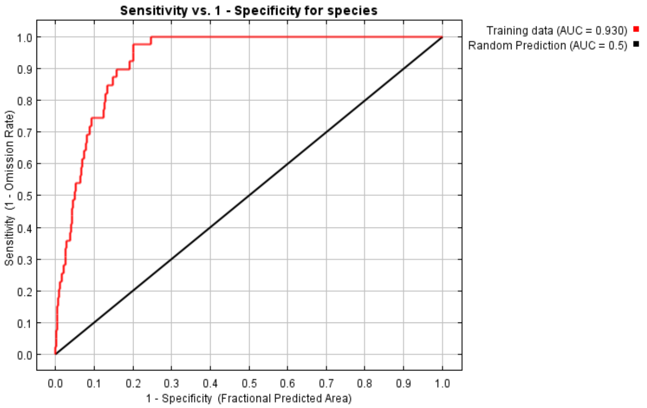
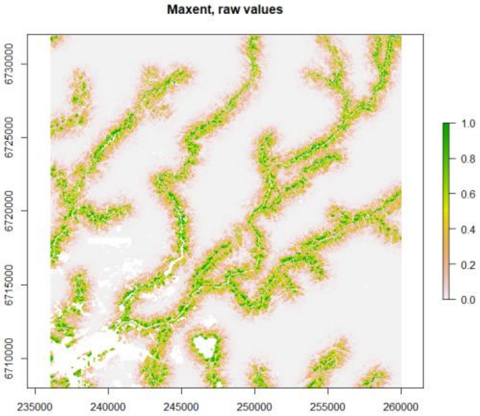
05_watertest_C	Topsoil class Solar irradiance TWI TPI far TPI near Dist. to water (6m cst)	AUC: 0.956 Features: L, Q, P, T, H		 Maxent, raw values 6730000 6725000 6720000 6715000 6710000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2 0.0
06_watertest_D	Topsoil class Solar irradiance TWI TPI far TPI near Dist. to water (6m Euc)	AUC: 0.941 Features: L, Q, P, T, H		 Maxent, raw values 6730000 6725000 6720000 6715000 6710000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2

07_watertest_E	Topsoil class Solar irradiance TWI TPI far TPI near Dist. to water (11m cst)	AUC: 0.955 Features: L, Q, P, T, H	 	 Maxent, raw values 6710000 6715000 6720000 6725000 6730000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2 0.0
08_watertest_F	Topsoil class Solar irradiance TWI TPI far TPI near Dist. to water (11m Euc)	AUC: 0.942 Features: L, Q, P, T, H	 	 Maxent, raw values 6710000 6715000 6720000 6725000 6730000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2

09_watertest_G	Topsoil class Solar irradiance TWI TPI far TPI near	AUC: 0.898 Features: L, Q, P, T, H		 Maxent, raw values 6730000 6725000 6720000 6715000 6710000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2
10_elevtest_A	Topsoil class Solar irradiance TWI DEM Dist. to water (3m Euc)	AUC: 0.969 Features: L, Q, P, T, H		 Maxent, raw values 6730000 6725000 6720000 6715000 6710000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2 0.0

11_elevtest_B	Topsoil class Solar irradiance TWI TPI far TPI near Dist. to water (3m Euc)	AUC: 0.942 Features: L, Q, P, T, H		 Maxent, raw values 6710000 6715000 6720000 6725000 6730000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2
12_elevtest_C	Topsoil class Solar irradiance TWI TPI near Dist. to water (3m Euc)	AUC: 0.937 Features: L, Q, P, T, H		 Maxent, raw values 6710000 6715000 6720000 6725000 6730000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2 0.0

13_elevtest_D	Topsoil class Solar irradiance TWI TPI far Dist. to water (3m Euc)	AUC: 0.933 Features: L, Q, P, T, H		 Maxent, raw values 6710000 6715000 6720000 6725000 6730000 235000 240000 245000 250000 255000 260000 0.0 0.2 0.4 0.6 0.8 1.0
14_elevtest_E	Topsoil class Solar irradiance TWI TPI far TPI near DEM Dist. to water (3m Euc)	AUC: 0.975 Features: L, Q, P, T, H		 Maxent, raw values 6710000 6715000 6720000 6725000 6730000 235000 240000 245000 250000 255000 260000 0.0 0.2 0.4 0.6 0.8 1.0

15_elevtest_F	Topsoil class Solar irradiance TWI Dist. to water (3m Euc)	AUC: 0.930 Features: L, Q, P, T, H		 Maxent, raw values 6710000 6715000 6720000 6725000 6730000 235000 240000 245000 250000 255000 260000 1.0 0.8 0.6 0.4 0.2 0.0