Archaeological Predictive Modelling of Prehistoric Settlements in Lantau Island, Hong Kong

SRN: 23109181

Year of submission: 2024

Module code and title: INST0062 MSc Dissertation

Name of supervisor: Dr. Oliver Duke-Williams

This dissertation is submitted in partial fulfilment of the requirements for the Master's degree in MSc Digital Humanities, UCL.

Electronic word count: 12847

Referencing style: Harvard

Declaration

I have read and understood the College and Departmental statements and guidelines concerning plagiarism. I declare that:

- This submission is entirely my own original work.
- Wherever published, unpublished, printed, electronic or other information sources have been used as a contribution or component of this work, these are explicitly, clearly and individually acknowledged by appropriate use of quotation marks, citations, references and statements in the text. It is 12918 words in length.
- Any use of generative AI has been clearly acknowledged in my work and is permitted under the assignment's categorisation according to UCL's AI guidelines.

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Dr. Oliver Duke-Williams, for their guidance, insights and support throughout the research process. His expertise and mentorship have been instrumental in shaping this work. I would also like express my gratitude to my personal tutor, Dr Vasileios Routsis, for his care and support.

I would to appreciate the resources and facilities provided by the Hong Kong Antiquities and Monuments Office, which have helped me with research process.

I would also like to thank my family, friends and colleagues for their inspiration and encouragement during my Master's program.

Abstract

This study utilized machine learning methods to assess the archaeological potential of the prehistoric period on Lantau Island, Hong Kong. The dependent variable in the prediction study is the archaeological potential, while the independent variables consist of various environmental factors. The selected machine learning algorithms for the study are MaxEnt and SVM (One Class SVM), suitable for presence-only data. Two sets of archaeological data, including Sites of Archaeological Interests (SAIs) and archaeological survey reports, were gathered to represent archaeological potential. Additionally, diverse environmental data were collected as environmental variables.

The prediction results indicate that both MaxEnt and SVM (One Class SVM) are capable of generating reliable predictions. Both models suggest that areas with archaeological potential predominantly cluster along coastal regions characterized by lower elevation levels. The extent of archaeological potential areas varies depending on the archaeological dataset used in the model training process, resulting in either widespread or concentrated predictions. The study also highlights the significant roles of environmental variables such as elevation and proximity to the coast, while factors like distance to alluvial deposits and slope play minor roles in predicting archaeological potential locations. These findings align with previous archaeological studies, indicating that low-level coastal areas along sandbars were preferred landscapes for human settlement during the prehistoric era.

Table of Contents

Figures List

Tables List

1 Introduction

1.1 Introduction

Digital technologies have played a significant role and served different purposes in archaeology, notably the use of Geographical Information System (GIS), digital reconstruction, network analysis, computer simulation and statistical analysis. This study explores the statistical analysis aspect of digital archaeology and attempt to predict potential prehistoric settlements sites in Lantau Island, Hong Kong.

Predictive modelling in archaeology is utilised to identify the potential of archaeological sites within an area through the study of the relationship between the discovered sites and their environments. Such approach has often been used in archaeological research and cultural resource management, benefiting the conservation of discovered and undiscovered archaeological sites.

The method of archaeological predictive modelling can be dated as early as the 1950s and the methodology has been evolving with the advance of digital technologies, such as GIS. Studies related to archaeological predictive modelling over the years have suggested that the prediction is influenced by the geographic environment and other factors associated to the location of known archaeological sites. Environmental factors such as elevation, slope, distance to water bodies, etc. are commonly considered in prediction models.

The implement of statistical analysis and machine learning methods allow the quantitative study of archaeological sites data and make use of known archaeological sites to find unknown locations of archaeological potential.

1.2 Research Question

This research attempts to use machine learning algorithms and statistical analysis to predict the locations of potential prehistoric archaeological sites in Lantau Island, Hong Kong. The modelling makes use of machine learning algorithms and statistical analysis to predict the

potential based on known prehistoric archaeological sites within the region. Two machine learning algorithms are considered in this study, 1) Maximum Entropy (MaxEnt) and Support Vector Machine.

Besides of predicting the location of archaeological sites, the two machine learning models help measure the correlation between the archaeological sites and their associated environment. The study explores relationship of archaeological sites and the environment where past humans were likely settling in prehistoric Lantau Island, Hong Kong.

Research Extent

The Lantau Island is the largest island of Hong Kong. The island, sitting at the southwest of Hong Kong, is primarily composed of mountainous terrain. Past human activities can be dated as early as the Neolithic period such as the stone circle was found in Fan Lau, located on the southwest coast of the island.

This study focuses on the prehistoric period settlements, including the Neolithic period and Bronze Age (Figure 1-1).

Figure 1-1 Extent in red of the Lantau Island main island, Hong Kong

1.3 Aims and Objectives

Aims

A1. To establish a prediction of prehistoric archaeological potential of Lantau Island, Hong Kong.

A2. To understand the correlation between the environmental factors and archaeological potential areas in Lantau Island, Hong Kong through statistical analysis.

Objectives

O1. Gather and review datasets and machine learning algorithms suitable to predicting the prehistoric archaeological potential of Lantau Island, Hong Kong. (see A1)
    a. Conduct a comprehensive review of available datasets relevant to the archaeological potential in Lantau Island, Hong Kong from archaeological reports and journals.
    b. Evaluate the machine learning algorithms (MaxEnt and SVM) for archaeological predictive modelling.
    c. Compile appropriate and available environmental datasets of Lantau Island, Hong Kong.

O2. Develop a predictive model for the prehistoric archaeological potential of Lantau Island, Hong Kong. (see A1)
    a. Prepare datasets for model training and testing.
    b. Carry out and fine-tune the prediction process to obtain results.
    c. Evaluate the performance of the prediction model.

O3. Interpret the variable importance of the environmental variables in predicting the archaeological potential in MaxEnt and SVM. (see A1 and A2)

2 Literature Review

This literature review aims to explore existing research on archaeological prediction modelling studies and case studies of using machine learning algorithms and statistical analysis for predicting the likelihood of finding archaeological sites. It also explores the archaeological background of Lantau Island to link the prediction methodology to this research's extent.

2.1 Archaeological Background of Lantau Island, Hong Kong

2.1.1 Landscape Archaeology of Hong Kong

The sea level was 100-120m lower than present level during the last Ice Age and at around 15,000 years ago, South China was a plain where Taiwan is linked to the mainland China and Hong Kong was at the inner edge of the plain at around 200km from the coast (Fyfe et al. 2000). Between 15,000 to 4,000 BP, sea level raised rapidly as the results of melting glaciers, any human activities within the former plain would have been buried underwater by the raised sea-level. The sea level was +1m to +3m higher than present level at around 7,000 to 6,000 BP and regressed at about 4,000 BP and 2,000 BP, reaching the present sea level (ibid). The rising sea level played a role in human occupation of Hong Kong were suggested by Meacham (1984;2009), that humans were brought in to settle in Hong Kong by the rise, that all known inland sites were dated later than 4500 B.C.

The relationship between prehistoric archaeological sites in Hong Kong and their environment and landscape has been studied. So (1969) suggested that the landscape of archaeological sites can be categorized into four categories, a detail description of the landscapes is also discussed in Shang and Ng (2010).

1. Sand bar parallel to bay: Involved farming and fishing activities of small communities. Lagoon behind sand bar.

2. Tombolo: Sites found on the western part of tombolo aligning north to west to shelter from east and northeast wind. Formation of the sandy isthmus is associated to the rising of sea levels.
3. Coastal terrace: Terraces on the edge of hillslope of marine origin.
4. Alluvial terrace: Terraces on both sides of rivers. Usually rivers in the New Territories.

2.1.2 Archaeological Studies in Lantau Island

The Lantau Island, being the largest island in the western part of Hong Kong, saw significant archaeological discoveries of human settlements at different time periods. There are currently 54 Sites of Archaeological Interest (SAIs) in Lantau Island. Sites of Archaeological Interest are areas identified by the Antiquities and Monuments Office to be heritage sites and possess higher potential in finding archaeological context (Antiquities and Monuments Office 2024). Archaeological works are usually conducted to further investigate these sites or other parts of the Lantau Island. Out of the 54 SAIs, 33 are related to the prehistoric period.

Archaeological studies in the Lantau Island can be dated back to the 1920s where human activities were found along the coast. Further archaeological studies have found archaeological artefacts dated as early as the mid-Neolithic period (Islands District Board 1994).

A notable SAI Sha Lo Wan is located at the northern coast of the Lantau Island. Artefacts such as coarseware, geometric corded pottery, polished stone tool, quartz ring, bronzeware dated to the Neolithic and Bronze Age are found. In addition, are kilns dated to the Tang Dynasty and ceramics and coins dated to the Song and Yuan Dynasties (Drewett 1995; Guangzhou Antique Archaeology 1998). Many other SAIs had similar discoveries where prehistoric or historical artefacts were unearthed. Meanwhile, majority of the sites only found small artefacts such as pottery sherds which are insufficient to conclude any sites as settlements. However, since these findings indicate past human activities and archaeological potential, there locations should be used for the archaeological prediction. It is also noted that some of the sites have been disturbed by modern activities, such as agricultural and urbanization developments.

Some archaeological publications are inaccessible or untraceable to provide a comprehensive review, especially for the archaeological studies in the early and mid-20th century. Information extracted from these reports are limited to referencing by other archaeological study reports.

2.1.3 Background of Archaeological Prediction Model

Archaeological Predictive Model has become a common approach by research institutes and cultural resource managements in predicting the probability of archaeological sites presence (Wang et al. 2023; Verhagen and Whitley 2012; Yaworsky et al. 2020) and played a role in the conservation of archaeological resources. For instance, predictive modelling helps identify areas with archaeological potential in impact assessment, responsible consultancy can provide suggestions on what further actions to mitigate any disturbance to the potential archaeological resources.

Predictive modelling relies on either a representative sample from the region or a fundamental understanding of human behaviour to make the prediction (Kohler and Parker 1986). The underlying assumption is that the selection of settlement locations by humans is not random but rather influenced by their natural environment (Kvamme 2006). Furthermore, predictive modelling is based on the premise that similar types of archaeological sites tend to occur in the same or similar places (Renfrew and Bahn 2016). The concept allows researchers to quantify these locations into spatial data as a foundation for predictive modelling.

The assumption of quantitative predictive modelling is that the probability of any land unit to contain archaeological lies between 0 and 1 (where 0 represent 0% chance and 1 represent 100% chance likely), any location with known archaeological potential must be considered as 1 (Whitley 2003). The calculation to the site probability estimates can be explained using Bayesian probability terminology. Although Bayesian statistics is not implemented in all predictive models, the underlying structure is assumed in all of them (ibid). The probability of any land unit being an archaeological site is the sum of the probabilities of all exhaustive and mutually exclusive variables that cause a land unit to be chosen as a site or to intentionally be made one (Whitley 2003; Pearl 2000), implying the

relationship between archaeological sites and possible variables. It is further explained that the effect on the prediction models by variables vary that they each carry a weight influencing probability, suggesting the degree of correlation between archaeological sites and possible variables (ibid). However, this should be studied carefully since variables can be correlated with the variables themselves instead of the site presences.

Environmental data are commonly used in modelling while some included cultural data into their models. Willey, one of the pioneering works has played a pioneering role in predictive modelling where he discovered that locations of human settlements are closely related to the natural environment (Willey 1953). Subsequent studies by researchers have shown similar conclusion where environmental features influence human choices of settlements (Warren 1990; Kvamme 1992; Bauer et al. 2004; Wang et al. 2023). Environmental features can be categorised into different types (Yaworsky et al. 2020), including resource distribution (such as distance to water bodies), environmental productivity (agricultural productivity), climate (such as mean temperature) and landscape attributes (such as slope). Among these features, the topographic landscape attributes such as elevation and curvature and resource distribution such as distance to water bodies have been widely utilised in predictive modelling due to their strong correlation with settlement patterns (Wang et al. 2023; Luthfi et al. 2019).

However, the use of archaeological prediction models is not without controversy. Researchers have argued that the predictive models have been overly environmentally deterministic (Gaffney and van Leusen 1995; Wheatley 2004), suggesting that archaeology should not be studied through the lens of environmental factors alone. They suggested that apart from the environments, social and cultural factors are integral components of human behaviour and should be considered in archaeological research. Whitley's work is noted to study the cognitive aspects in predictive modelling, such as GIS modelling of cognition with spatial proxies (Whitley 2003). However, the incorporation of social or cultural and cognitive factors are complex and not without challenges, especially when they are more difficult to map (Kamermans 2010). It is suggested that to incorporate a wider spectrum of factors in predictive modelling, three issues should be tackled, statistical improvements, quality of archaeological dataset and the development of non-environmentally based models

(Verhagen 2010). Woodman (2002) in his study on the use and abuse concluded that “before researchers attempt to incorporate the more intangible social, cognitive, political and aesthetic factors, it would be wise to employ the appropriate statistical techniques required to deal with the complexities which already exist in even the most basic tangible and quantifiable environmental criteria”. This influences this research to study the machine learning algorithms in predictive modelling to study how well they perform and the result of the prediction.

While the criticism on environmental determinism holds merit, it is important to recognise that predictive models, despite their deterministic nature provide valuable insights into the preferred environments of past human settlements. Through quantitative analysis, these models allow us to observe patterns and gain deeper understanding of the relationship of how environmental factors influenced human behaviour and decision-making process.

2.2 Archaeological Prediction Modelling Case Study

2.2.1 Model Selection

Statistical approach has been implemented in archaeological prediction modelling in archaeological investigations of various regions, such as south-central Utah, USA (Yaworsky et al. 2020), northeast Romania (Nicu et al. 2019), Indonesia (Luthfi et al. 2019), Netherlands (Verhagen 2007), northeast Israel, China and Japan (Wachtel et al. 2018; Wang et al. 2023).

Concluding the previous archaeological prediction modelling, there are three types of statistical approach commonly used, including regression, frequency ratio and machine learning.

The regression approach performs logistic regression and fit models by maximising the likelihood of observing the given data. It obtains the prediction results through the relationship of a dependent variable (the archaeological site presence) and multiple independent variables (environment variables).

The frequency ratio approach observes the relationship between the dependent variables and independent variables to find their correlation. It is commonly used in landslide susceptibility, gully erosion susceptibility mapping, etc (Lee 2005).

The machine learning approach provides more flexibility and parameters influencing predictive power and avoid overfitting. The approach commonly involves iterations, permutation testing and regularisation to obtain an optimal setting for the model. The machine learning algorithms used in this study are MaxEnt and SVM.

The model being used to predict archaeology is traditionally logistic regression for its capability in of logistic regression classification. This method requires a present-absence dataset, which is a dataset of locations having and not having archaeological sites respectively (Wachtel 2018). However, archaeological data usually only consist of present data since present data indicate the presence of archaeological site / archaeological potential, while the absence of data does not necessarily indicate the absence of archaeological potential. This is usually addressed by creating random points to indicate absence of site, also called as pseudo-absence points or background points, which represent the available environment within the study area and are not considered as true absence.

On the contrary, the Maximum Entropy Modelling (MaxEnt) is suggested to perform better in predicting archaeological sites (ibid). Philips in his works on species distribution model suggested that one of the advantages of using MaxEnt model is the data requires presence data only accompanied by the environmental data. It is also capable of generating results when the amount of training data is limited (Philips 2006). Different comparative studies are comparing MaxEnt to linear regression and other statistical models, suggesting that MaxEnt provide better performance when predicting archaeological sites (Yaworsky et al. 2020; Noviello 2018; Wachtel 2018). Apart from the prediction results, the MaxEnt model generate results of response curves showing the relationship between predicted archaeological potential locations and environmental variables. It also generates result of variables importance indicating which environmental variables influence the prediction the most.

Besides of the MaxEnt, the Support Vector Machine (SVM), specifically OneClassSVM which is a SVM algorithm of unsupervised outlier detection (Schölkopf 2001), also only needs

presence data and environmental data for prediction. There are limited resources or previous studies discussing the SVM approach in archaeological prediction model. This study attempts to study its performance in archaeological studies.

See 2.3 Machine Learning Algorithm: MaxEnt and SVM for details on the machine learning models.

2.2.2 Archaeological and Environmental Data Selection

There are two studies of archaeological prediction model in the Fuxin of Liaoning Province (Wachtel et al. 2018) and Shaanxi Province (Wang et al. 2023) making use of the MaxEnt algorithms and provided reasonably good results. The environmental variables in both studies are selected base on domain knowledge and previous archaeological studies. The variables selected for the Fuxin region includes proximity to agricultural land, distance from modern villages, aspect, slopes distant from main river, land curvature, elevation ad distance from pasture land. The variables selected for the Shaanxi Province includes elevation, slope, roughness, relative degree of the land surface, plan curvature, profile curvature, cutting depth and distance from major rivers.

In the Environmental Impact Assessment (EIA) of the New Territories North in Hong Kong (Civil Engineering and Development Department 2023 Appendix 12.4), MaxEnt modelling is used in cultural resource management to establish the archaeological potential areas in north New Territories. The data being used in the study is slightly different than those generally used in other archaeological studies. For environmental data, distance to resources such as watercourse, coast or hill are derived from geology instead of the mapping of actual locations of the resources. The is due to limit of data and the environment data is affected by modified landscape by modern activities. For archaeological data, its uses of Sites of Archaeological Interests (SAIs) (Antiquities and Monuments Office 2024) and historical villages as archaeological sites data. The rationale of using these data is that archaeological finding data within its study area is very scarce. The study generates prediction for prehistoric and historical period in the study area. The dataset for this study is not accessible. The methodology of this study is inspired by the EIA's methodology and applied to a different region, the Lantau Island.

2.3 Machine Learning Algorithm: MaxEnt and SVM

2.3.1 Maximum Entropy (MaxEnt)

The MaxEnt algorithm begins by assuming a uniform distribution, where all locations are equally likely to have a presence point, (i.e. a distribution of maximum entropy) (Jaynes 1957; Yaworsky et al. 2020). This forms a probability density function across the range of each environmental variable, with the density proportional to the frequency of each variable value in the landscape.

MaxEnt uses a set of "features" to model the distributions described above, which are created from predictor variables (the environmental variables) and can include linear and non-linear transformations, as well as interactions between variables (Elith et al. 2011; Yaworsky et al. 2020). Initially, the model assumes a uniform distribution for each feature, based on the frequency of values in the landscape. Coefficients are randomly generated, and the fit to the presence sites is estimated using a log-likelihood function. These coefficients are then adjusted using a random walk in parameter space, and the log-likelihood is re-estimated. If the fit improves, the new coefficients are retained; otherwise, they are rejected. Regularization is employed to prevent overfitting, limiting the increase in log-likelihood and relaxing the constraints. This may lead to the rejection of certain features or variables that do not improve the model (Yaworsky et al. 2020).

The output of MaxEnt consists of raw probabilities, summing to one across the region and representing the relative rate of occurrence on the landscape. These probabilities are then converted to log-odds and further transformed into interpretable probabilities (ibid). The underlying model fitting of MaxEnt is based on generalized linear models, but with the inclusion of features that allow for more complex variable responses and interactions (ibid, Philips et al. 2006).

Wachtel et al. (2018) concluded the MaxEnt model is based on two principles, 1) the expectancies comparison principle and 2) the maximal entropy principle. The first step to a

probability map would be comparing the expectancies of all variables which is the mathematical concept of an average (ibid).

Wachtel et al. (2018) and Wang et al. (2023) in their studies of archaeological predictive modelling explained the mathematical steps of the MaxEnt model as follows.

The first step to a probability map would be comparing the expectancies of all variables which is the mathematical concept of an average (Wachtel et al.2018). Assume C as the set of models satisfying all constraints f_i:

$$\mathrm{C} = \{P | E_{\mathrm{P}}(f_i) = E_{\tilde{\mathrm{P}}}(f_i), i = 1, \cdots, n\} \tag{1}$$

where P refers to the distribution extracted from observation and $\tilde{\mathrm{P}}$ is the real distribution looking for, and f_i is the feature function corresponding to the input variables n on the presence of archaeological sites, defined as:

$$f_i = \begin{cases} 1, \\ 0, \end{cases} \tag{2}$$

The model is constrained by the variables with setting the expectation from the observed data $(E_{\mathrm{P}}(f_i))$ to be equal to the model's expectation $(E_{\tilde{\mathrm{P}}}(f_i))$ under the maximum entropy condition. There is a large collection of distribution $\tilde{\mathrm{P}}$ satisfying (1), which the maximum entropy principle is used to find the most appropriate one (Wachtel et al.2018). Entropy is defined as:

$$H(P) = -\sum_{x \in X} \tilde{\mathrm{P}}(x) \ln \tilde{\mathrm{P}}(x) \tag{3}$$

where x is a random location within the study area X, and $\widetilde{\mathrm{P}}$ is the expected probability distribution. The logarithm function here makes independent sources to be additive (e.g., x within X). This equation sets the sum of probabilities to be 1. The probability distribution of archaeological site presence P^* is as follows:

$$P^* = \arg\max_{P \in C} H\ (P) \tag{4}$$

2.3.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning model mainly used as binary classifier. Its objective is to find the optimal hyperplane to separate the classes (Sanchez-Hernandez et al. 2007). It requires n-dimensional training vectors x_i (environmental variables) and labelled class $y = \{-1, 1\}$. The machine learning aims to find the suitable of parameter α in the decision function $f(x, \alpha)$ for the classification. The vectors x lie on the hyperplane satisfy $w \cdot x_i + b = 0$, that w is normal to the hyperplane and $|b|/||w||$ is the perpendicular distance from the hyperplane to the origin. The margin of the hyperplane is defined as the shortest distance (d_+ and d_-) between the hyperplane to the closest positive or negative training example, and should follow the constraints (Yan and Zheng 2007),

$$w \cdot x_i + b \geq 1, \quad if\ y_i = 1 \tag{5}$$

$$w \cdot x_i + b \geq -1, \quad if\ y_i = -1 \tag{6}$$

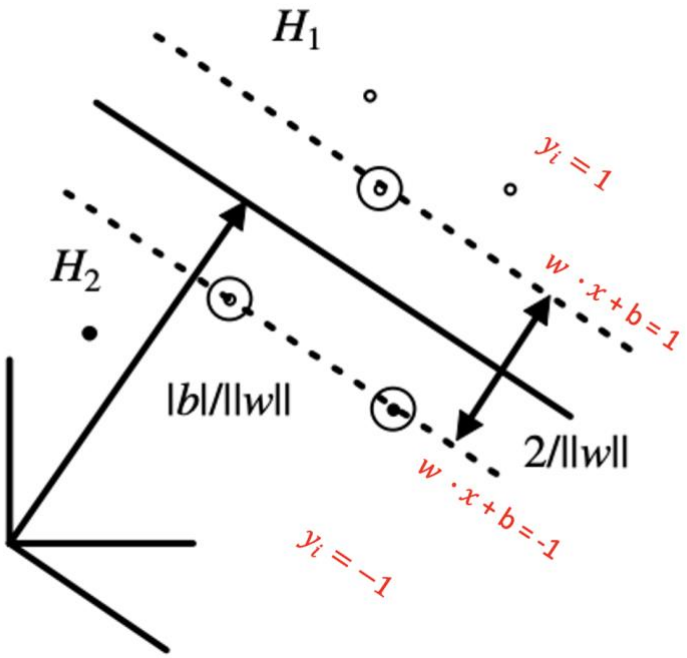


Figure 2-1 Linear hyperplane separating the two classes (ibid). The solid line being the hyperplane and dashed line being the margins. The circled points are the closest positive or negative training examples which are also called the support vectors.

Yan and Zheng (2007) further discussed the mathematical approach to optimisation approach and concluded with the decision function of nonlinear support vector machine where i are the iterations of the training vectors:

$$f(x) = \text{sgn}(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b) \tag{7}$$

SVM also support classification through the One-Class SVM algorithm, it is type of unsupervised machine learning useful for outlier detection and find the data belong to a class. Similar to binary classification in SVM, it finds the hyperplane and maximising its margin (Schölkopf et al. 2001).

Schölkopf et al (2001) and Yan and Zheng (2007) further discussed the mathematical process of the model and concluded with the decision function, notably y is no longer applied since the class is ignored in unsupervised machine learning:

$$f(x) = \text{sgn}(\sum_{i=1}^{l} \alpha_i K(x_i, x) - \rho)$$

(8)

SVM also allows parameters tweaking to adjust the training process to enhance performance and avoid performance, notably the kernel parameters which affect the transformation of input data to be separated by the hyperplane.

3 Methodology

This study involves three main stages, data collection, data mining and result interpretation. The following breaks down the stages in detail.

The data mining process involves processing data with machine learning algorithms to generate predictions. Two main algorithms are used, MaxEnt and SVM. MaxEnt is accessible through the software developed by Phillips et al. (2024). SVM requires Python scripting with the scikit-learn sklearn.svm package (Scikit-learn 2011).

The data and script are accessible in GitHub repository https://github.com/wei-hei-nip/LantauAPM.git.

The coordinate referencing system used in this study is the Hong Kong 1980 Grid System (ESPG: 2326).

3.1 Data Collection

Similar to the previous studies mentioned in the literature review, this research requires two sets of data for the prediction model, the dependent variable archaeological data and the independent variable environmental data.

3.1.1 Archaeological Data

The geographical locations of the archaeological site data are required as the dependent variable. There are mainly two sources of archaeological site data. The first is previous archaeological findings from archaeological journals and reports. previous archaeological finding is considered since it represents the geographic location where the archaeological context is found. The second is the Sites of Archaeological Interests (SAIs) which are areas possessing higher potential in finding archaeological context.

Two sets of data are stored for the two sources respectively.

Archaeological Data Set 1 (Previous Archaeological Findings)

This study records the geographical locations of test pit and borehole where prehistoric archaeological findings are found in archaeological studies in Lantau Island. The locations of findings recorded in reports are usually in three ways.

A. Without coordinate
Locations of archaeological findings are mentioned within the text only. However, the coordinates are not provided. It is impossible to record this data.

B. Appendix/Table with coordinate
The locations are recorded with coordinates and presented in tables or appendixes. The coordinates are usually presented in the Hong Kong 1980 Grid System (ESPG: 2326) or Hong Kong 1963 Grid System (ESPG: 3366).

C. Map
The locations are recorded graphically on maps while the coordinates are not provided. The maps are georeferenced in GIS and the points are plotted manually to record the data.

The data is stored as a point vector shapefile. The data includes the easting (X), northing (Y), corresponding archaeological reports, associated SAI and the value of environmental variable of the points.

Archaeological Data Set 2 (SAIs)

Ideally, previous archaeological finding is considered since it is the definitive location where the archaeological context is found. However, an insufficient amount of data may affect the accuracy of the machine-learning models. An extra set of archaeological data is considered inspired by the New Territories North EIA study (ibid). The areas of SAIs are not available online, which are only available in paper map format at the library in the Antiquities and Monuments Office. The maps are georeferenced in GIS and the areas are plotted manually to record the data.

The data is stored as a polygon vector shapefile. The areas are processed as points in the "Data Preparation" process. Two types of information are stored in the shapefile the name of SAI easting (X), northing (Y).

3.1.2 Environmental Data

Taking reference to previous species distribution models and archaeological prediction models. The prediction model considered various environmental datasets. Some data were excluded due to randomness, irrelevance, or disturbance from modern landscaping. Test runs with different variables during model development help identify environmental factors that are influencing the predictions (i.e., permutation importance and percent contribution generated by the MaxEnt model). The primary environment data collected are topography and geology data. These data are further processed during data mining.

3.2 Data Mining

Cross-Industry Standard Process for Data Mining (CRISP-DM)

This study makes use of the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Shearer 2000) to apply data mining methods. It is an industry-independent process model that explores and gains insight into the data using mathematics and analytic models. It is applied to this study since it aims to gain insight into archaeological sites and environmental data through machine learning methods. Figure 3-1 is a flowchart of the data mining process

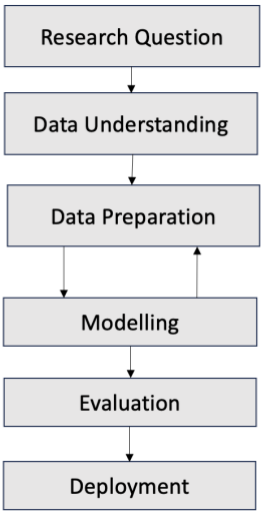


Figure 3-1 Flowchart of the data mining process of the study.

3.2.1 Research Question

The stage aims to establish the goal of the data mining process, which is equivalent to the aims and objectives of this study (see 1.3 Aims and Objectives).

3.2.2 Data Understanding

The stage aims to collect and explore the dataset, see above for data source and data collection methodology. In addition, it aims to establish an understanding of the collected data. This include employing descriptive statistics to summarise and describe the data and visualise the data to identify any patterns. It also examines the quality the data.

The datasets are processed to extract environmental variables values base on each archaeological finding locations through spatial join using GIS. The data understanding process utilises GIS and Python scripting for data visualisation.

Data Description

It provides an overview of the acquired dataset, including the number of archaeological findings samples and number of environmental variables. It discusses the archaeological data acquired and explain the archaeological context they represent, and any limitations of

the dataset. The meaning and purpose of each environmental variables are discussed, as well as any limitations and challenges encountered to acquiring the data.

Data Exploration

Key statistical properties such as mean, median, standard deviation, minimum and maximum of the continuous variables (i.e. elevation) are summarised. Histogram and box plots are useful in understanding the distributions of the archaeological finding data to the environmental variables. Scatter plots and correlation analysis are useful to explore the relationships between environmental variables and visualise any correlations.

Frequencies and proportions are calculated for categorical variables (i.e. geology). Bar charts are used to visualise the frequencies and understand if certain categories (i.e. geological deposit in geology) of the variables are represented more frequently in the data set.

Data Quality

An essential aspect of this stage is to ensure the data integrity. It evaluates the completeness of the dataset by identifying any missing values. If missing values are identified, it examines the impact of the missing data and what approach is appropriate to handle the missing data.

It also examines if the acquired data are accurate by referencing other source of data and ensure the data is in correct measurement unit and raster format.

Data Interpretation

The datasets are further analysed to summarise any insights or patterns. It summarises the initial findings and observations about the relationships between the archaeological settlements and the environmental variables. Highlighting the environmental settings that archaeological records are more commonly found. The interpretation should relate to the existing understanding of archaeological context on Lantau Island through existing archaeological studies.

3.2.3 Data Preparation

The data is then prepared to be readily input into the machine learning models. Several tasks are carried out in this stage using functions in GIS and Python scripting.

Prediction Extent

The prediction extent is further filtered by ancient coastline (Figure 3-2). The position of the coastline is dynamic which vary in different time periods. Archaeological studies in Hong Kong suggested that the coastline during the Neolithic Period and Bronze Age is generally lower and occasionally higher than the present sea level. Taking reference to the CEDD study (2023), the ancient coastline and submerged areas are identified by the superficial geological deposits of these areas were disturbed by the rising sea levels and tidal movements during or after the prehistoric period. It is achieved by filtering marine superficial geological deposits. The rationale to this approach is to locate the potential ancient coastline and establish its relationship with the archaeological finding locations, such as the distance and direction. Furthermore, it helps exclude regions that are less likely to yield significant archaeological findings where archaeological deposits were disturbed.

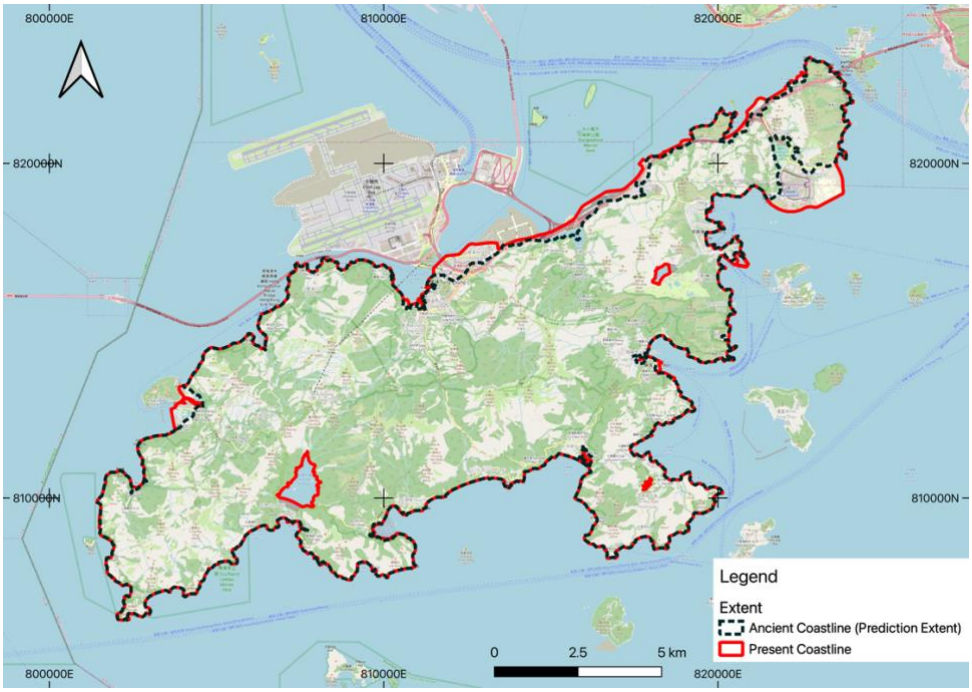


Figure 3-2 Present and ancient coastline of Lantau Island

Archaeological Data

Selecting Archaeological Data

Since this study focuses on the prehistoric archaeology on the Lantau Island, the archaeological data is filtered to only contain data with relevant archaeological context within the ancient coastline of the Lantau Island.

Format Archaeological Data

Archaeological Data Set 1 (Previous Archaeological Findings) are points data ready to be input to the machine learning models. Archaeological Data Set 2 (SAIs) are transformed into points based on the environmental variable raster data cells.

Environmental Data

Clipping Environmental Data

Environmental Data are clipped to prepare the data in this study desired extent, the main island of Lantau Island. Furthermore, to the data are clipped with the ancient coastline to reflect the past landscape.

The environmental data are clipped by the obtained ancient coastline using the clipping tool in GIS.

Construct and Integrate Environmental Data

New environmental variables are created or derived from collected environmental data. Such as creating new data based on the distance and direction of alluvial deposit, solid geology and ancient coastline.

Format Environmental Data

Finally, the environmental variables in raster format is clipped using GIS tools to extract the area of interest on Lantau Island. The environment raster data is processed into ASCII raster

format which is required in the MaxEnt model. SVM model is more flexible on the raster format.

The raster cell size is 30m.

The environmental and archaeological data collection and preparation for the finalised models are further discussed in Section 4.1- 4.2.

3.2.4 Modelling

Two machine learning models are used to generate predictions, the results are compared between the models to decide which perform better. These models include, MaxEnt and SVM.

Repeating Process

The modelling stage is a repetitive process where the same model is tested with different environmental variables and parameters. Testing with different environmental variables helps identify any redundant variables making the model simpler and prevent overfitting, that the model is only accurate when predicting training data but not for new data. Testing with different parameters helps find the settings for the model for optimal model performance.

Preparing new data or tweaking existing input data are occasionally needed for each repeated modelling process. The AUC score value and sensitivity (true positive rate) can be used to evaluate the model performance and determine the optimal settings (3.2.5 Evaluation).

The modelling process for is further discussed in Section 4.3 - 4.5.

3.2.5 Evaluation

AUC Value

The performance of the MaxEnt and SVM models are assessed by calculating the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. This method involves determining how well the model can differentiate presence and absence locations, presence being locations with archaeological potential, absence being without (Philips 2006). AUC values range from 0 to 1, with higher values indicating better model performance. An AUC value of 0.5 or lower suggests the model is unable to differentiate locations with or without archaeological potential, that the prediction is equal or worse than random chance. A lower AUC value suggests that there is no correlation between the environmental variables and the archaeological potential predictions by the models (Philips 2006). The AUC is calculated by comparing the archaeological data point (presence data) and background points randomly sampled within the study area (pseudo-absence data). The AUC values of MaxEnt and SVM prediction models were calculated by the MaxEnt software, and within the python script.

Train-test Split

It is crucial to prevent prediction models from overfitting. The train-test split is a common approach to test if a model is overfitting. The dependent variables (archaeological data) are split into training and testing dataset, usually 70%/30% or 80%/20%. The training data are used as input for the prediction model to generate predictions. The testing data act as unseen data for the model to predict.

The sensitivity or true positive rate is the calculation of the portion of archaeological data being predicted as positive (in this case predicted probability > 0.6 for MaxEnt and > 0 for SVM). It is calculated for both training and testing data to assess how well the high predicted probability areas match the testing data's locations.

$$\text{Recall} = \text{True Positive} / (\text{True Positive} + \text{False Negative})$$

3.2.6 Deployment

The finalised prediction of the MaxEnt and SVM models are generated once optimal settings are found. The prediction results are then interpreted to discuss the research questions (see below).

3.3 Result Interpretation

The prediction performance, result and variable importance are considered in this study.

3.3.1 Prediction Performance

It first interprets the prediction performance of the MaxEnt and SVM respectively as part of the modelling and evaluation stage of data mining. The models are compared by AUC scores and the test-train split method to determine how well they predict known archaeological locations.

3.3.2 Prediction Result

Then, it discusses the areas that are identified as archaeological potential areas that are not represented in the known archaeological data. The prediction result of each model is a probability map of the predicted region. The higher probability indicates the area is more likely to have archaeological findings. The results are also generated in raster data format to be visualised graphically.

The areas are identified mainly based on the prediction of the model with the best performance. However, the predictions of all three models are compared to evaluate the similarities or differences of the potential areas.

3.3.3 Response Curve and Variable Importance

Lastly, the variable correlation and variable importance between the environmental variables and the prediction will be studied to understand what environmental settings are favouring prehistoric archaeological settlements on Lantau Island. This is mainly studied using the MaxEnt model since it generates 1) response curves to illustrate how the value of a variable affects the predicted probability while keeping other variables at their average sample value and 2) the percentage contribution, permutation importance and jackknife analysis to determine which variables are more important when conducting the prediction (see 2 Literature).

Weights assigned to features (environmental variables) can be extracted for linear kernel SVM. The weight represents the importance of each variable in the classification. While the feature weights in SVM (non-linear) are less interpretable since only dual coefficients can be extracted which do not directly represent the classification ability of variables. SVM models also do not generate variable response curves. Hence, the MaxEnt model is used to study the effect of environmental variables on the prediction outcome.

The result and discussion should conclude which model performs better, what archaeological potential areas are identified and which variables are more important in the prediction.

4 Data and Model

The following reports the condition of the data collected and the details of the modelling process.

4.1 Selecting Archaeological Data

According to 3.1 Archaeological Data, there are two sources of archaeological data. The first is archaeological findings from previous archaeological studies, and the second is the SAIs.

4.1.1 Previous Archaeological Studies

Most of the archaeological reports in Hong Kong are stored as hard copies at the Reference Library of the Antiquities and Monuments Office. Requests are required to view the copies. The remaining proportion of the reports are usually environment impact assessment reports available online.

There were difficulties in listing all the studies that are associated with Lantau Island comprehensively since access to the reports is limited and requires examining the reports' content which consumes a certain amount of time. Conveniently, the library catalogued the reference materials according to the SAIs, which include physical and digital materials accessible in the library and online. During data collection, the SAIs' catalogue and associated reports are requested and viewed. Hence, the data collected are mainly from locations within the SAIs.

There are 32 SAIs associated with the prehistoric period on Lantau Island, due to the limit of time only 17 were requested and viewed, and 11 of the 17 yielded locations useful to the predictions. The reports where the locations were yielded are included within the shapefile data. 32 location points were collected from the reference materials in total.

4.1.2 Sites of Archaeological Interest (SAIs)

The extent of the areas is also only accessible through requesting the reference materials from the library. Points are created within the extent of archaeological input data. Points are extracted with grid points of 30m, 60m, 90m and 120 distance matching the 30m cell size environmental data. The grid points are tested in the modelling process to test which set of grid points is optimal for prediction. However, after multiple iterations of prediction testing, the four sets of points yielded similar results, while the 30m grid points had higher AUC values. Hence, this study only focuses on the 30m grid points data. The 32 SAIs yielded 2669 points.

4.2 Selecting Environmental Data

Environmental data reflecting the historical environment are limited and assume present days environment are useful as predictor for archaeology. Similar archaeological studies noted environmental variables can be categorized as 1) resource distribution, 2) environmental productivity, 3) climate and 4) landscape attributes (Yaworsky et al. 2020).

4.2.1 Resource Distribution

Resource distribution data commonly seen in archaeological studies are water bodies and certain land types. Water bodies can include streams, rivers and coasts. Modern days rivers and streams are available online, but some of them were altered or defunct due to modern development. These data may not represent water resources accurately. Hence, taking reference to the CEDD (2023) EIA study, the locations of alluvial superficial geological deposits are extracted to represent the water resources.

Coastal resources are important to settlement patterns in Lantau Island since most of the archaeological findings are located near the coast. As suggested in 3.2.3 Data Preparation, the locations of marine superficial deposits are referenced to represent the coastal resources.

Another resource is the distance to hilly landscape which is represented by the extracted solid geological deposits.

The geology data is available online and published by the Civil Engineering and Development Department (2022). The Euclidean Distance is calculated with the functions in QGIS.

4.2.2 Environmental Productivity

Environmental Productivity refers to production activities such as agriculture and mining. However, the socio-economic activities are more challenging to map spatially and accurately. There were no environmental variables selected for this category for the finalised model.

4.2.3 Climate

Climate conditions such as temperature could influence human activities such as agriculture (Ramankutty et al. 2002). Morgan and Guénard et al. (2019) published a useful list of environmental data in 30m resolution. The list of data includes climate data, such as mean temperature, mean air pressure, mean rainfall, etc. These data were compiled by interpolating historical observatory station data between 1998 and 2007 throughout Hong Kong. However, preliminary prediction test suggests the climate data were less useful in prediction. Furthermore, it raised concern whether the climate data is reflective of the prehistoric period. There were no environmental variables selected for this category for the finalised model.

4.2.4 Landscape attributes

Landscape Attributes refer to the physical features of the area. Topography data such as elevation, aspect, and slope are selected for the prediction. Furthermore, geological data is also selected since geology is representative to the land type.

4.2.5 Environmental Variable Refinement

All of the environmental data are continuous except the geological data which is categorical. The MaxEnt model allows the input of continuous and categorical environmental data. SVM models require one hot encoding of categorical variables. It involves creating rasters of

binary value 0 or 1 for each category of the variable. 1 being the category is present. However, there are over a hundred superficial deposits and solid geology to encode which would increase the dimension of a model which is not ideal and potentially causes overfitting. Therefore, geology is selected associated with the known landscape in prehistoric archaeological sites (see 2.1.1 Landscape Archaeology of Hong Kong), being sandbar parallel to the bay, tombolo (not considered in this study as the extent only focuses on the main island), coastal and alluvial terrace. A total of 7 sets of data are extracted from the geological data (Table 4-1).

Table 4-1 Landscape Type of common prehistoric archaeological sites and associated superficial geology in Hong Kong

Landscape Type	Superficial Geology	Explanation
Sandbar	Qb	Quaternary, beach deposits: mainly sand
Sandbar	Qbs	Quaternary, back shore deposits: mainly sand or gravel
Sandbar	Qrb	Quaternary, raised beach deposits: mainly sand
Coastal / Alluvial Terrace	Qa	Quaternary, alluvium (undifferentiated)
Coastal / Alluvial Terrace	Qpa	Quaternary, Pleistocene, terraced alluvium
Coastal / Alluvial Terrace	Qd	Quaternary, debris flow deposits (undifferentiated)
Coastal / Alluvial Terrace	Qpd	Quaternary, Pleistocene, debris flow deposits

Sandbar	Qbb	Quaternary, beach deposits: mainly cobbles and boulders

The 14 environmental variables for the finalised models are elevation, slope, aspect, distance coast, distance to alluvial deposit, distance to solid geology and the encoded geology (Qb, Qbs, Qrb, Qa, Qpa, Qd, Qpd and Qbb).

4.3 Principle Component Analysis: Collinearity

Collinearity refers to the linear relationship of two or more predictor variables (the environmental variables). High collinearity in ecological modelling (Dormann 2013), and in this case archaeological modelling. A common approach to detect collinearity is the pairwise correlation coefficient (r), which $|r| > 0.7$ is suggested to be high collinearity. All 13 variables had $|r|$ lower than 0.7. The collinearity analysis as part of the exploratory data analysis is visualised in Appendix A.

4.4 MaxEnt Modelling

Modelling of MaxEnt algorithm was made available with the open-source software developed by Phillips et al. (2024) for species distribution model. It allows easy configuration of data input and parameters.

Since software modelling allows environmental data of continuous and categorical type, the initial test used a geological deposit raster which includes over hundred types of superficial deposits and solid geology. However, the raster was not considered afterwards, as too many categorical types of geology may cause complications to the geological selection, and to align with SVM to encode the different geological types as binary values. Instead, a list of one hot encoded geology (see 4.2.5 Environmental Variable Refinement) is selected as environmental variables input.

The maxent model can be replicated in multiple runs to further validate its performance. Cross-validation introduces randomness in the train-test split to further ensure the model is not overfitting. It is conducted by running the model with the same parameter multiple times. The dataset is divided into *k* number of sets where each iteration uses a set of the divided data as testing data and the rest as training data. The cross-validation replication method is selected for the 30m grid data since the data set is large. It is replicated 10 times. Meanwhile, the survey data points were not replicated, instead the validation was done by importing a test sample file which would be the 30m grid data. Such method could study how well the survey data could predict the SAI areas.

Other parameters to note during the MaxEnt modelling was the selection of features type which affects how the influence environmental variables towards the prediction and depends on the type of environmental variables. Different combinations of linear, quadratic, product, threshold and hinge were tested to see how they performed. It is also suggested that linear was always used; quadratic with at least 10 samples; hinge with at least 15; threshold and product with at least 80 (Elith et al. 2010). The 30m grid points selected linear, quadratic, product and hinge as feature types, while the survey data points selected linear, quadratic and hinge as feature types.

Response Curves and jackknife analysis are also generated for environmental variables interpretation. Table 4-2 shows the parameters of MaxEnt model. Two models are used to train the two data sets, 30m grid points and survey data points.

Table 4-2 Parameters specific of MaxEnt modelling

Parameter	Specific
Samples	30m grid points / survey data points
Environmental layers	Environmental variables folder
Create response curve	Checked
Make picture of predictors	Checked

Jackknife	Checked
Output format	Cloglog
Output file forat	asc
Features type	30m grid points: Linear, quadratic, product, hinge Survey data points: Linear, quadratic, hinge
Regularization multiplier	1
Max number of background points	10000
Replicates	30m grid points: 10 Survey data points: 1
Replicated run type	30m grid points: Crossvalidate Survey data points: Not replicated
Test sample file	30m grid points: No test sample file

	Survey data points: 30m grid points as test sample

4.5 SVM Modelling

The SVM model uses the same archaeological and environmental data input as the MaxEnt model.

The major concern of the SVM modelling is the kernel parameter. Linear kernels in SVM are useful when the data and be separated with a linear decision boundary, whereas non-linear kernels handle data that are not separable linearly. As suggested in Section 4.3.3, the linear kernel has weight assigned to each feature, while feature weights are less interpretable for non-linear kernel. Ideally, this study prefers the linear kernel which allows the study of the effect of environmental variables. However, test runs on the model are unable to generate meaning prediction, suggesting the data could not be linearly separated. The radial basis function (rbf) kernel is used in the finalised prediction.

Two models are used to train the two data sets, the parameters for 30m grid points are nu=0.5; kernel="rbf"; gamma= 1000, while the parameters for survey points are nu=0.5; kernel="rbf"; gamma= 'scale'. There are differences in gamma since the model for 30m grid points needs to be adjusted to find environment more similar to the environment of known archaeological locations. The Python script takes reference to the example code on scikit-learn by Prettenhofer and Vanderplas (2024). A version of the SVM script for this study is available on the GitHub repository and Appendix B.

5 Findings

5.1 Performance Comparison

5.1.1 AUC Score

Table 5-1 AUC Score of Prediction Models

Prediction Model	Testing Data	AUC Score
MaxEnt (30m Grid Point as Training Data)	Subset of training data from Crossvalidate	0.875
MaxEnt (Survey Point as Training Data)	30m grid point	0.924
SVM (30m Grid Point as Training Data)	Subset of training data	0.971
SVM (Survey Point as Training Data)	30m grid point	0.942

Four finalised models are generated, MaxEnt and SVM models trained using 30m grid points and survey points. The testing data used differs slightly in each model due to different modelling approaches and sample sizes. Judging from the AUC score alone the SVM appears to perform better than MaxEnt.

5.1.2 Sensitivity

The sensitivity provided a different perspective on the models' performance. Sensitivity is a better representation of how many positive samples are predicted as true positive. The first sensitivity is calculated with the points in the dataset as a unit.

Since SAIs are larger extent to represent potential areas, they usually include a wider variety of environments, and the prediction results tend to predict a portion of these areas instead of the entire area depending on the environmental variables. Therefore, a second sensitivity is calculated with the SAI as a unit where SAI is considered positive if there are respective positive points. The probability threshold is >=0.6 for MaxEnt prediction to be considered positive, and >0 for SVM.

Table 5-2 Sensitivity of prediction models in respect to training and testing dataset.

Prediction Model	Dataset	Sensitivity
MaxEnt (30m Grid Point as Training Data)	30m Grid Point*	2057/2669 = 0.771
		31/32 = 0.969
	Survey Point	30/32 = 0.938
		9/9 = 1
MaxEnt (Survey Point as Training Data)	30m Grid Point	331/2669 = 0.124
		23/32 = 0.719
	Survey Point*	22/32 = 0.688
		6/9 = 0.667
SVM (30m Grid Point as Training Data)	30m Grid Point*	1312/2669 = 0.492
		30/32 = 0.938
	Survey Point	18/32 = 0.563
		7/9 = 0.778
	30m Grid Point	260/2669 = 0.111

SVM (Survey Point as Training Data)		23/32 = 0.719
	Survey Point*	13/32 = 0.406
		4/9 = 0.444

* Training dataset to respective model.

The results show that the MaxEnt model trained with a 30m grid performed the best, exhibiting high sensitivity for both training and testing data. Both points and SAIs as a unit showed high sensitivity.

The MaxEnt model trained with survey data exhibited worse sensitivity, with significantly fewer points predicted - only 12.4% of the 30m grid points and approximately 68.8% of the survey data points were predicted. Additionally, around 70% of the SAIs as a unit were predicted in the training and testing dataset.

The SVM model trained with a 30m grid performed slightly worse than the MaxEnt version, predicting only around half of the training and testing dataset for points as a unit. However, the sensitivity was higher for SAIs as a unit, with 93.8% for training data and 77.8% for survey data.

The SVM model trained with survey data performed the worst, with the sensitivity of SAIs as a unit slightly better than that of points.

5.2 Prediction Map

The prediction maps for the four models are presented in Figure 5-1 - Figure 5-8. In Figure 5-1 - Figure 5-4, the prediction maps are portrayed with gradient colours, illustrating the relative potential of locations within the study area. Figure 5-5 - Figure 5-8 are binary-coloured maps representing designate locations deemed to possess archaeological potential, the threshold is >=0.6 for MaxEnt and >0 for SVM.

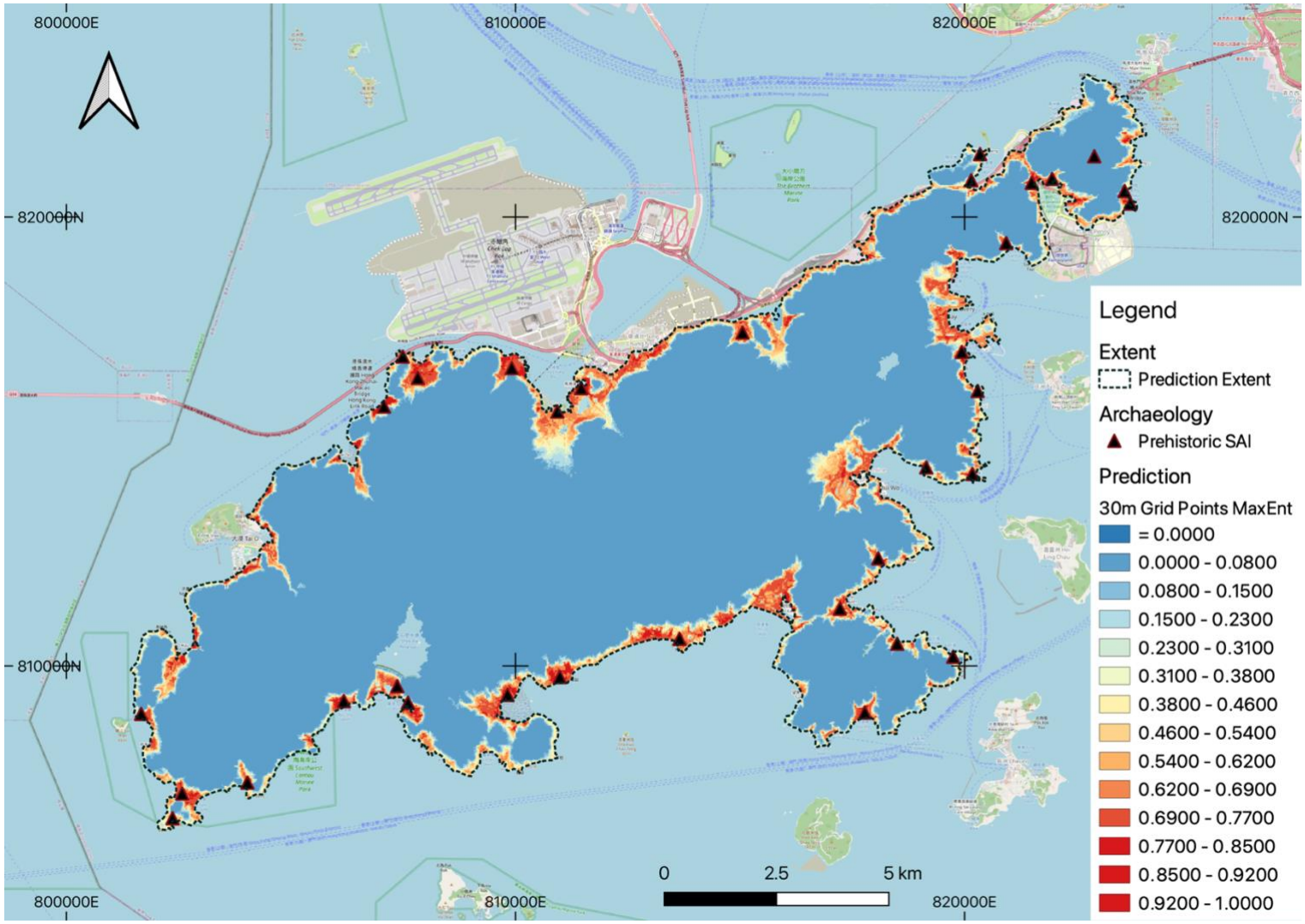


Figure 5-1 Continues prediction map of 30m grid point MaxEnt model

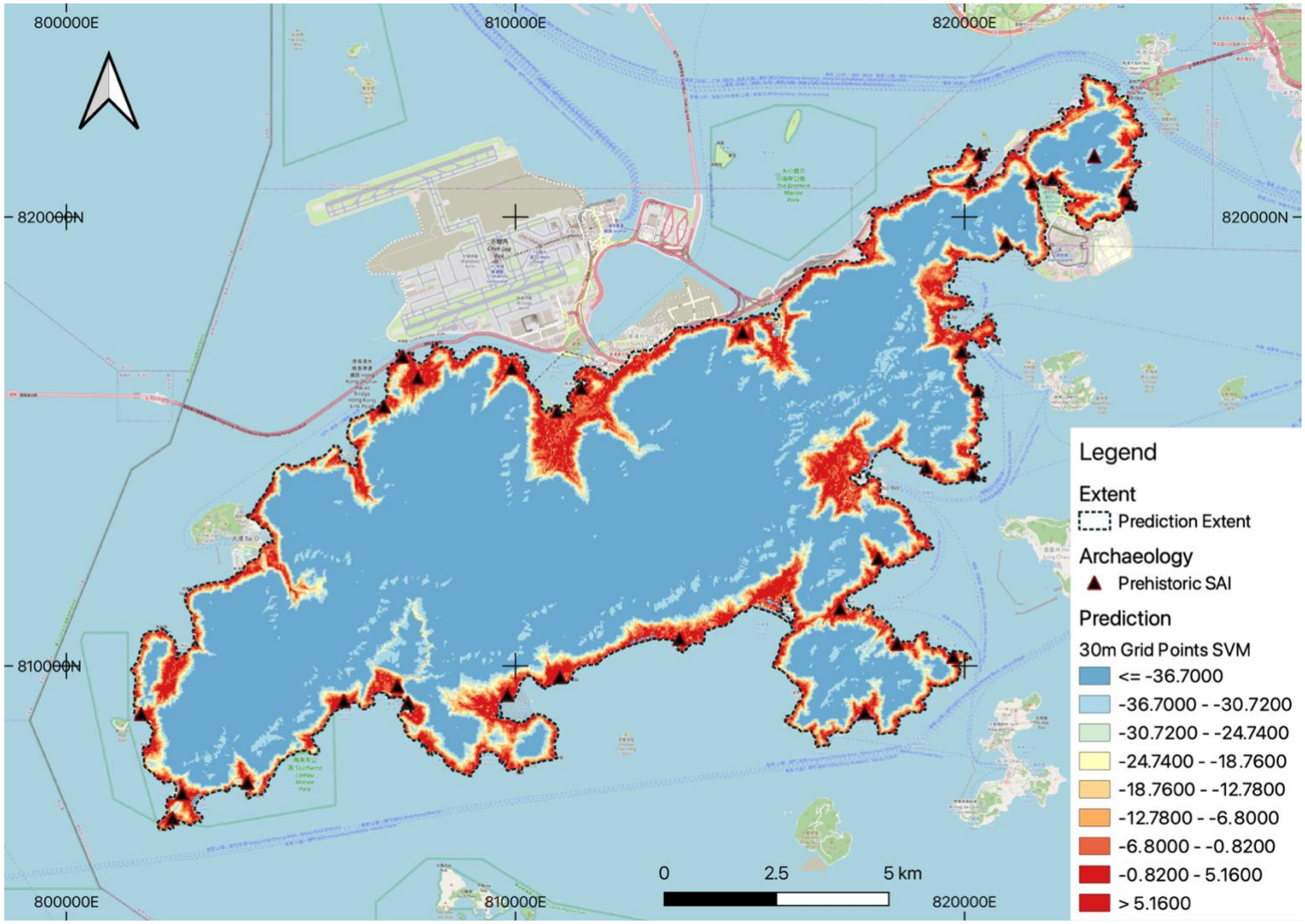


Figure 5-2 Continues prediction map of 30m grid point SVM model

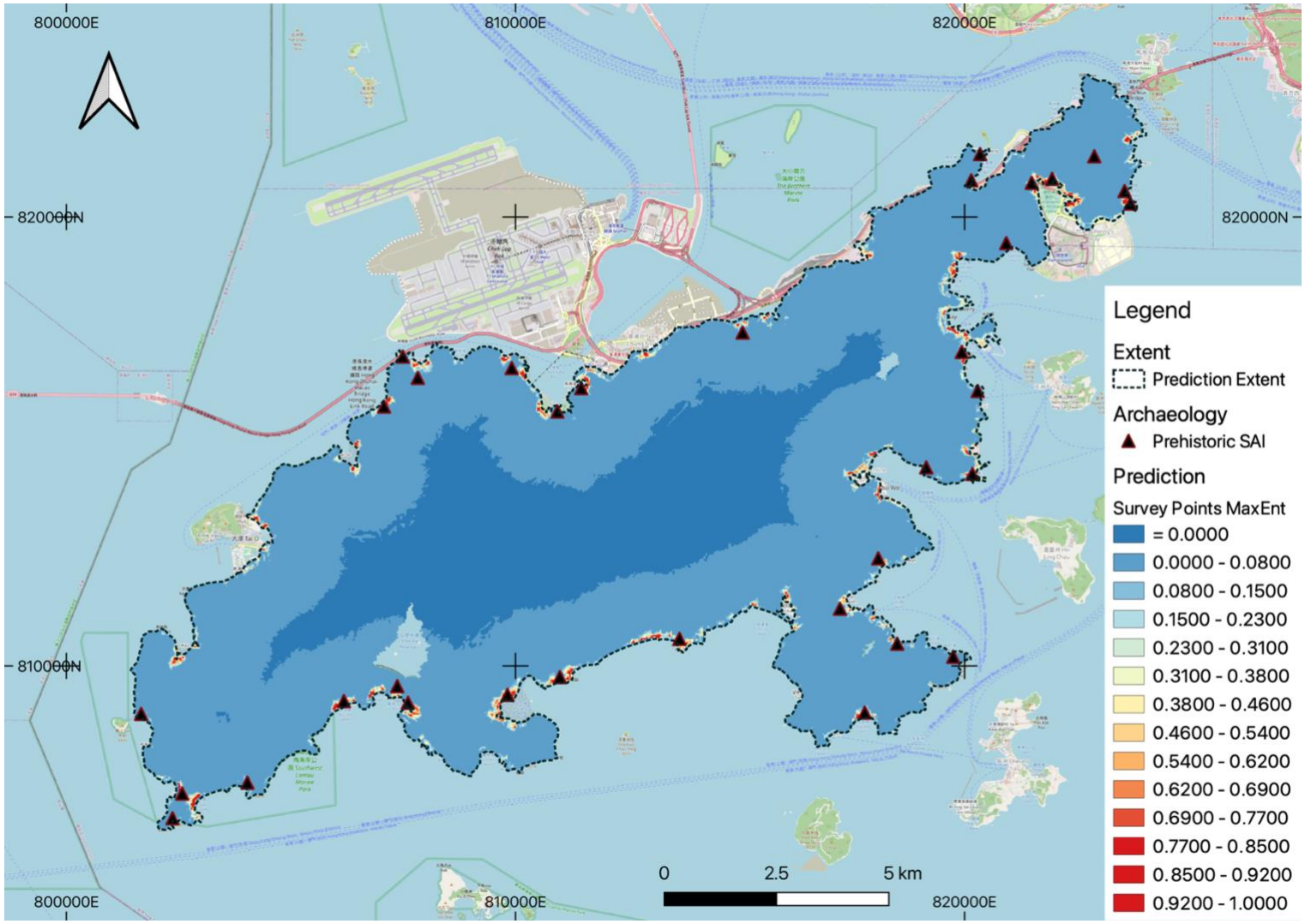


Figure 5-3 Continues prediction map of survey point MaxEnt model

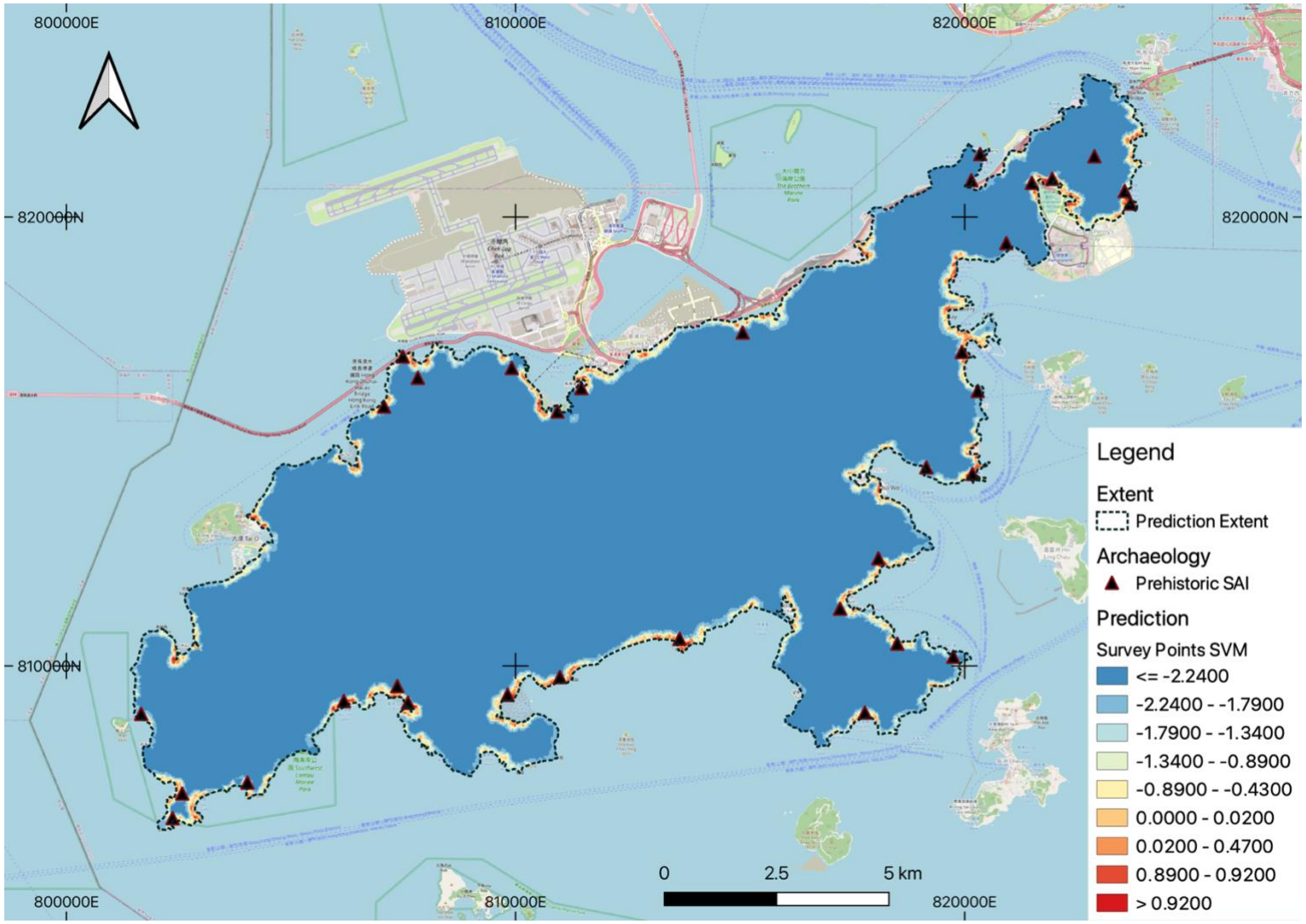


Figure 5-4 Continues prediction map of survey point SVM model

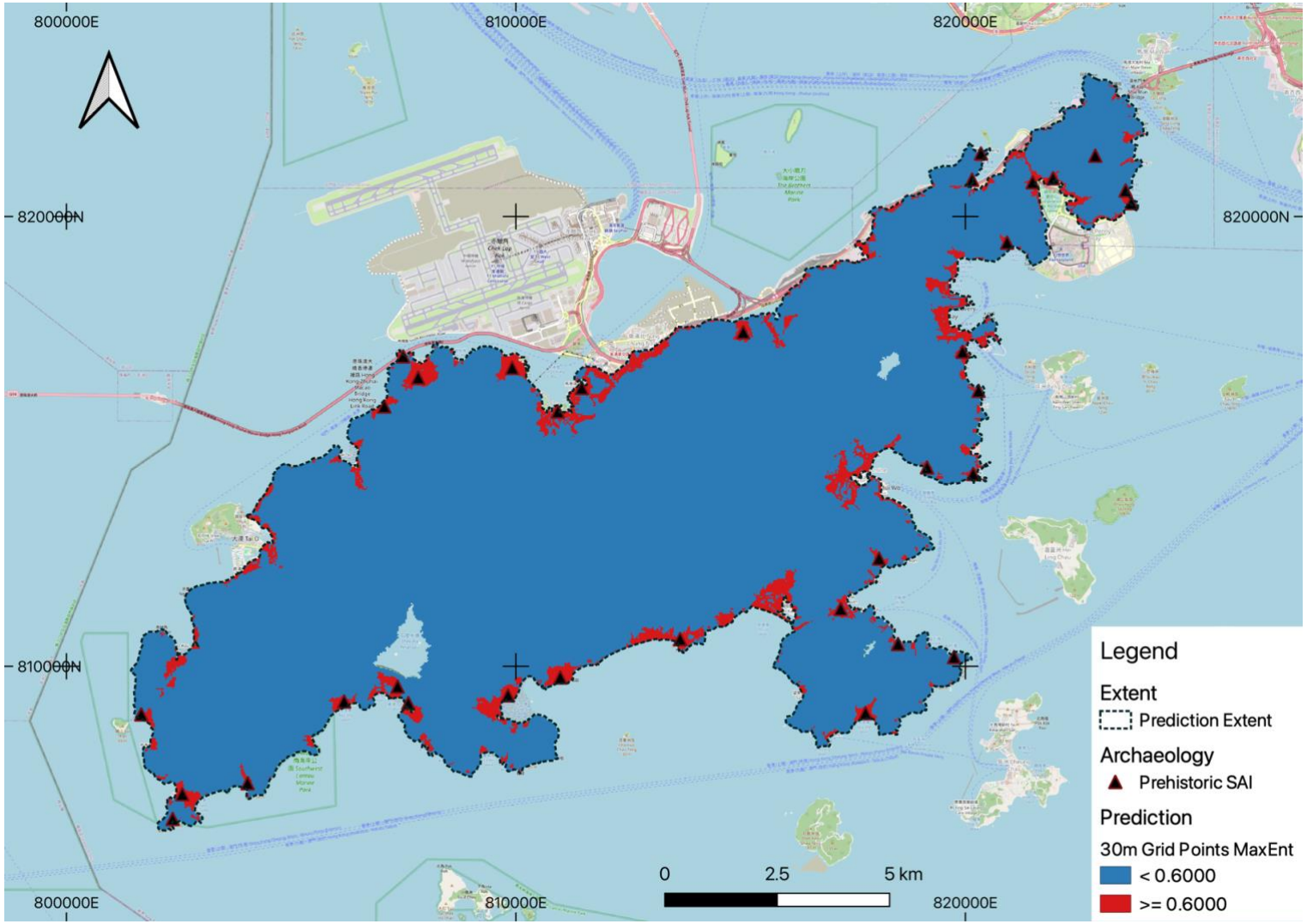


Figure 5-5 Binary prediction map of 30m grid point MaxEnt model

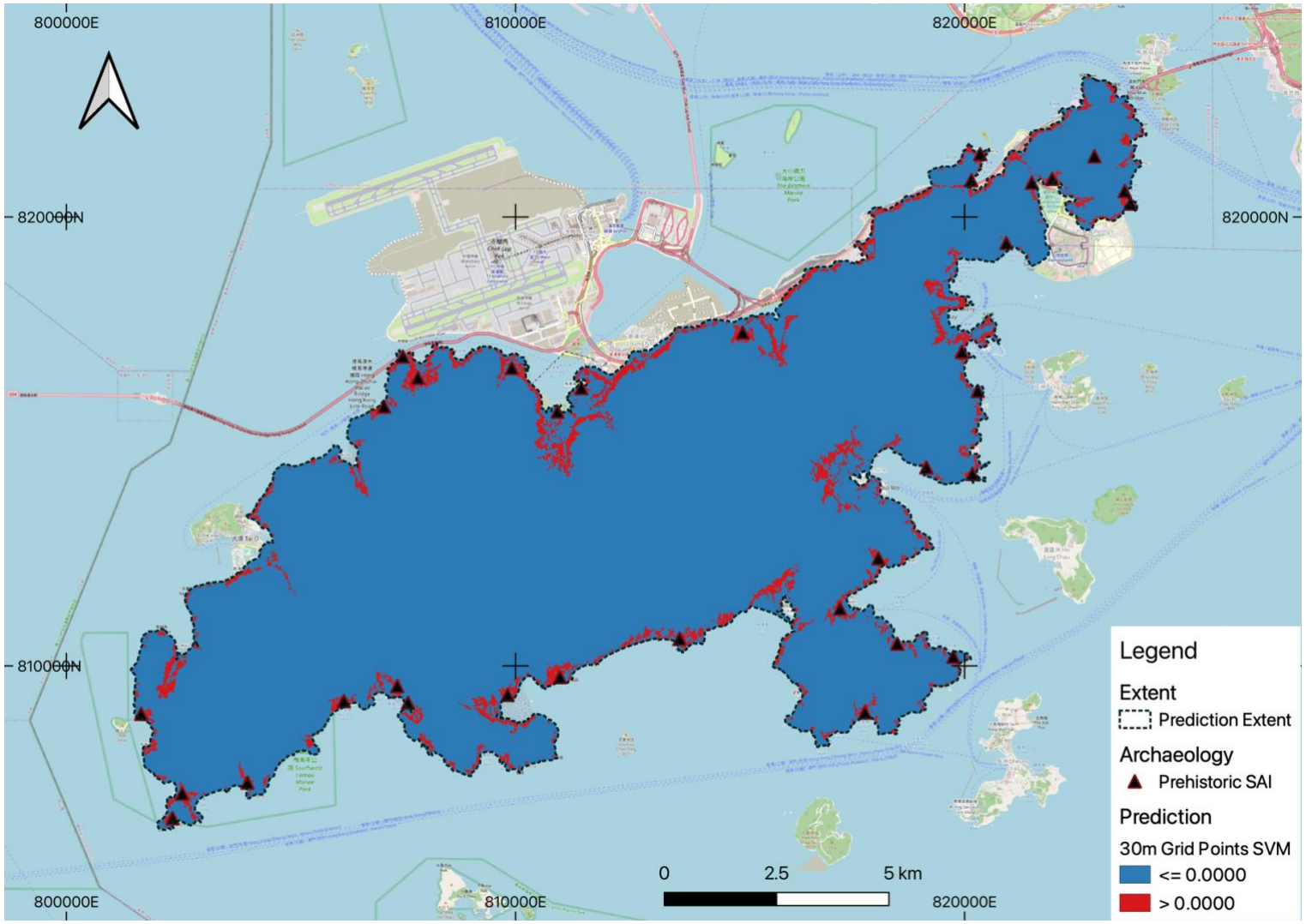


Figure 5-6 Binary prediction map of 30m grid point SVM model

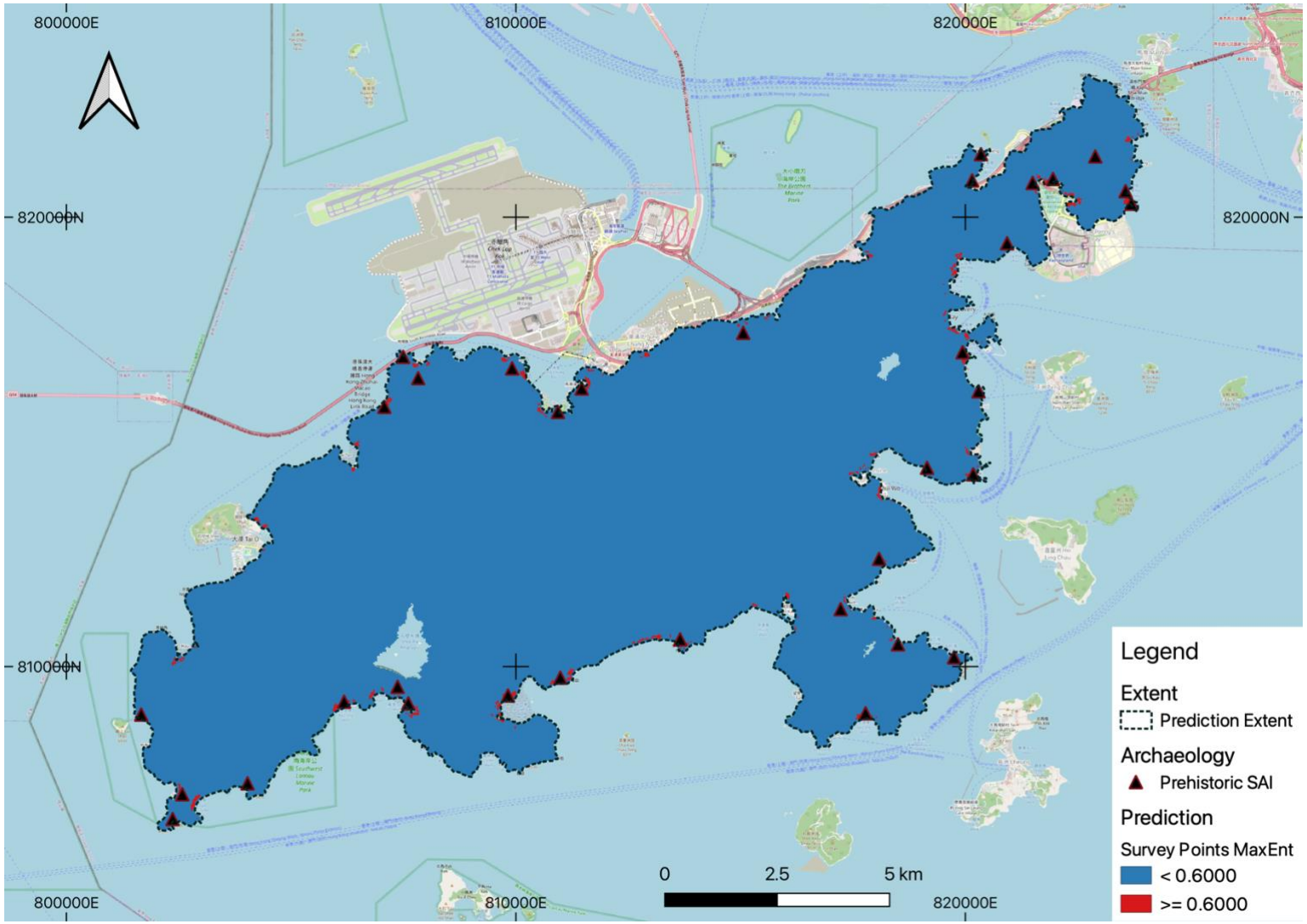


Figure 5-7 Binary prediction map of survey point MaxEnt model

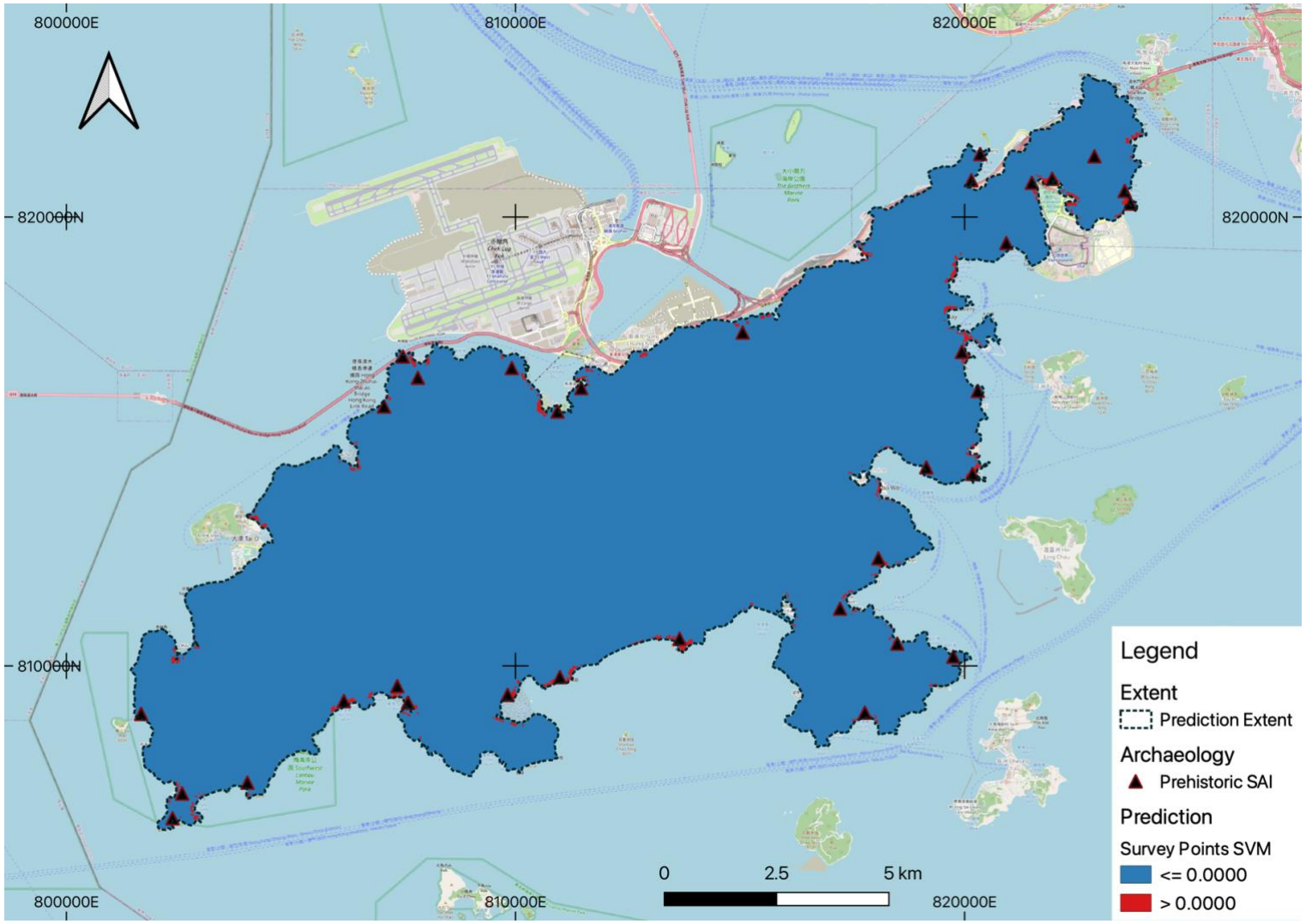


Figure 5-8 Binary prediction map of survey point SVM model

5.3 Effects of Environmental Variables on Prediction

MaxEnt generates three sets of information to study the effects of environmental variables on prediction, including percent contribution and permutation importance, response curve and jackknife analysis.

Percentage contribution and permutation importance measurements assessing the importance of environmental variables. Percentage contribution quantifies the relative importance of each variable influencing the model outputs, similar to variable weights in the model prediction process. Permutation importance focuses on the impact of each variable on the prediction performance, AUC. The first measure is obtained during the prediction process, while the latter depends on the final model (Phillips 2017).

Response curves graphically show the relationships between the environmental variables and prediction outcomes. Two sets of curves are generated, the first set of curves shows the effect on the prediction of the changing value of a variable while keeping all other environmental variables at their average sample value. The second set of curves shows the effect on the prediction of the changing value of the variable alone. The first set is studied since it shows the relationships while taking into account the variable's contribution to the prediction.

Jackknife analysis shows the effect of environmental variables on the prediction outcome by removing and only using the variable in the prediction. The analysis is reflective of the permutation importance.

5.3.1 Effects of Environmental Variables MaxEnt (30m Grid Points)

Table 5-3 Percent Contribution and Permutation Importance of environmental variables of 30m grid points MaxEnt model

Variable	Percent Contribution (%)	Permutation Importance (%)
Aspect	0.5	0.1
Elevation	86.7	67.6
Slope	2.5	6.6
Distance to alluvial deposits	0.9	1.2
Distance to coast	5.9	23.2
Distance to solid geology	2.1	0.4
Geology Qa	0.1	0
Geology Qpa	0.2	0.1
Geology Qd	1.2	0.7
Geology Qpd	0	0
Geology Qb	0	0
Geology Qrb	0	0
Geology Qbs	0	0
Geology Qbb	0	0

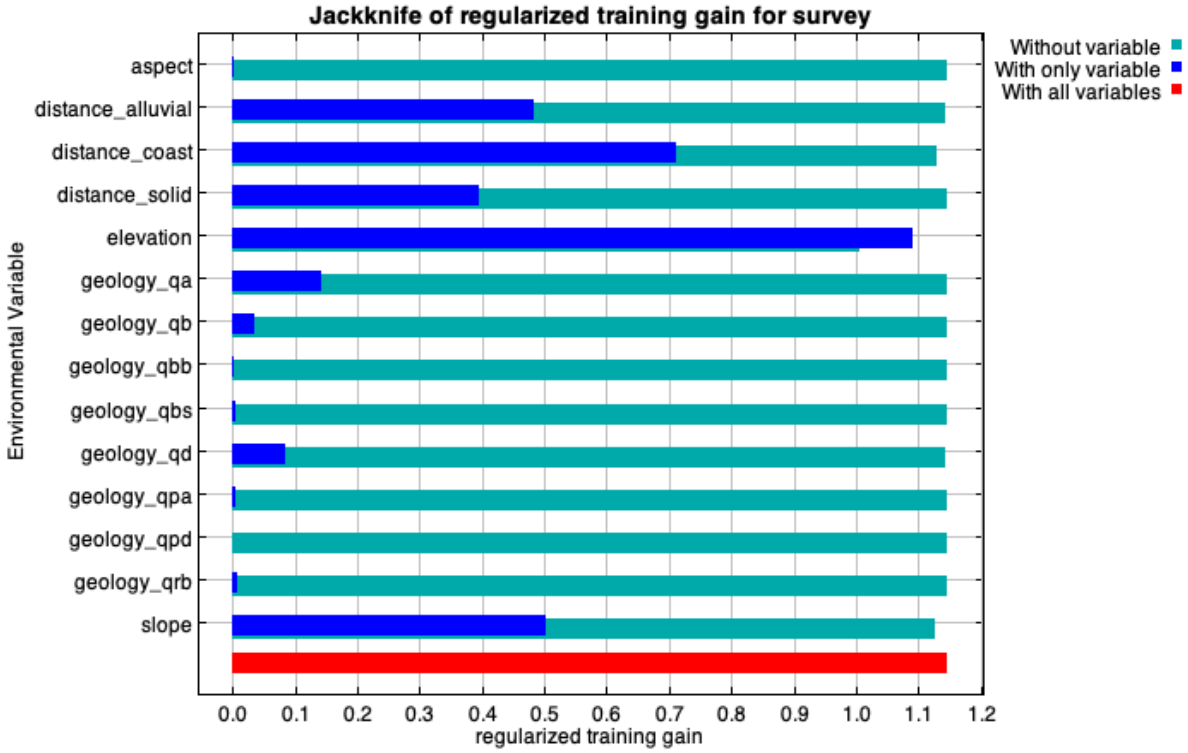


Figure 5-9 Jackknife analysis of 30m grid points MaxEnt model

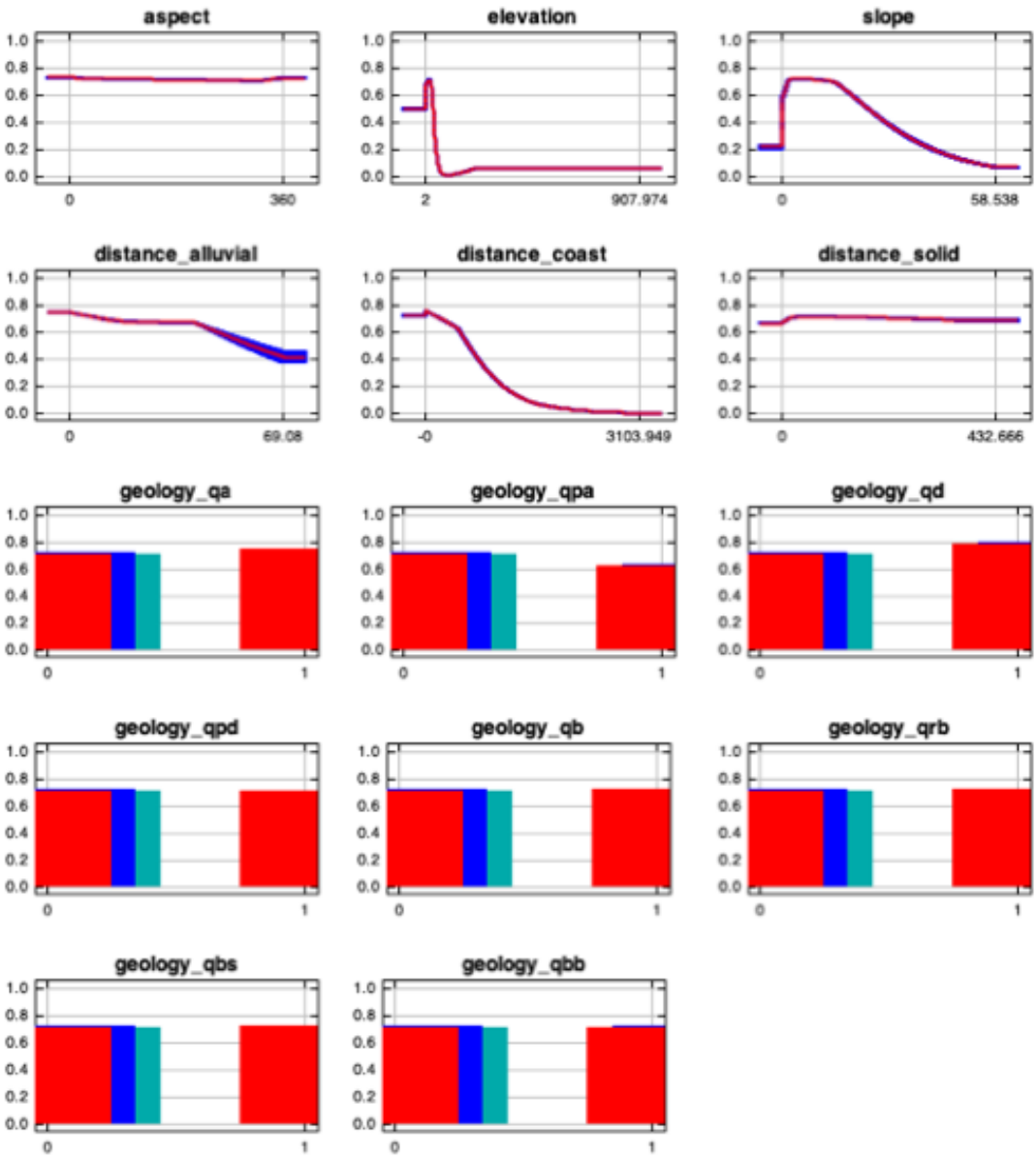


Figure 5-10 Response curves of environmental variables of 30m grid points MaxEnt model. The curves show the mean response of the 10 replicate Maxent runs (red) and and the mean +/- one standard deviation (blue, two shades for categorical variables).

Aspect

The percentage contribution and permutation importance of the aspect variable is 0.5% and 0.1% respectively, suggesting it plays an insignificant role in the model prediction. As shown in the jackknife analysis, the gain/performance did not increase when only using aspect for prediction and did not decrease when the model predicted without it. The flat response curve shows the archaeological potential did not increase as the aspect value varied, suggesting aspect does not affect the prediction.

Elevation

Elevation has the highest percentage contribution and permutation importance, 86.7% and 67.6% respectively, suggesting elevation played the most significant role in the prediction. As shown in the jackknife analysis, it has the highest gain/performance when only using it for prediction and decreases sharply when the model is predicted without it. It also shows that the prediction relies heavily on elevation since both percentage contribution and permutation importance are > 50%. The response curves show that there was a sharp increase in archaeological potential between 0mPD to < ~50mPD and decrease as the value of slope increases, suggesting a lower elevation has higher archaeological potential.

Slope

The percentage contribution and permutation importance of the slope variable are 2.5% and 6.6% respectively, suggesting it plays a lesser role and has a small impact on the model prediction. The jackknife analysis suggests gain/performance increased when only slope is used in the prediction, while there was a minor decrease when the model predicted without it. The response curves show there was a sharp increase in archaeological potential between 0 ° to < ~5° and decreases as the value of slope increases, suggesting a gentler slope has higher archaeological potential.

Distance to alluvial deposit

Similar to slope, the distance to alluvial deposit variable plays an insignificant role. The percentage contribution and permutation importance of the distance to alluvial deposit

variable is 0.9% and 1.2% respectively. The jackknife analysis suggests gain/performance increased when only slope is used in the prediction, while there was an insignificant decrease when the model predicted without it. The response curves show that the archaeological potential decreases subtly as the distance to alluvial deposits increases, suggesting the potential is higher when the locations are closer to alluvial deposits.

Distance to coast

Distance to coast has the second highest percentage contribution and permutation, 5.9% and 23.2% respectively. The percent contribution remains small since the prediction relied heavily on elevation at 86.7%. While the 23.2 % permutation importance suggests it is an important variable for the performance of the model. The jackknife analysis suggests gain/performance increased when only the variable is used in the prediction, while there was minor decrease when the model predicted without it. The response curve shows the archaeological potential decreases as the distance to coast increases, suggesting the potential is higher when the locations are close to the coast.

Distance to solid geology

Similar to slope, the percentage contribution and permutation importance of the distance to solid geology variable are 2.1% and 0.4% respectively, suggesting it plays insignificant role. The jackknife analysis suggests gain/performance increased when only slope is used in the prediction, while there was insignificant decrease when the model predicted without it. The response curve is flat similar to aspect, suggesting it has minimum effect on the prediction.

Geology

Geology generally played insignificant role in the model prediction, deposits Qd, Qpa and Qa have 1.2%, 0.2% and 0.1% percent contribution respectively; 0.7%, 0.1% and 0% permutation importance respectively. Qpd, Qb, Qrb, Qbs and Qbb deposits have 0% percent contribution and permutation. The jackknife analysis and response curves also show the deposits have insignificant effect on the prediction.

5.3.2 Effects of Environmental Variables MaxEnt (Survey Points)

Table 5-4 Percent Contribution and Permutation Importance of environmental variables of survey points MaxEnt model

Variable	Percent Contribution (%)	Permutation Importance (%)
Aspect	2.7	0.5
Elevation	32.9	0.7
Slope	4.8	4.4
Distance to alluvial deposits	19.7	2.5
Distance to coast	31.5	90.2
Distance to solid geology	2.8	1.4
Geology Qa	0	0
Geology Qpa	0	0
Geology Qd	1.5	0.2
Geology Qpd	0	0
Geology Qb	1.1	0.1
Geology Qrb	0.9	0
Geology Qbs	2.2	0.1
Geology Qbb	0	0

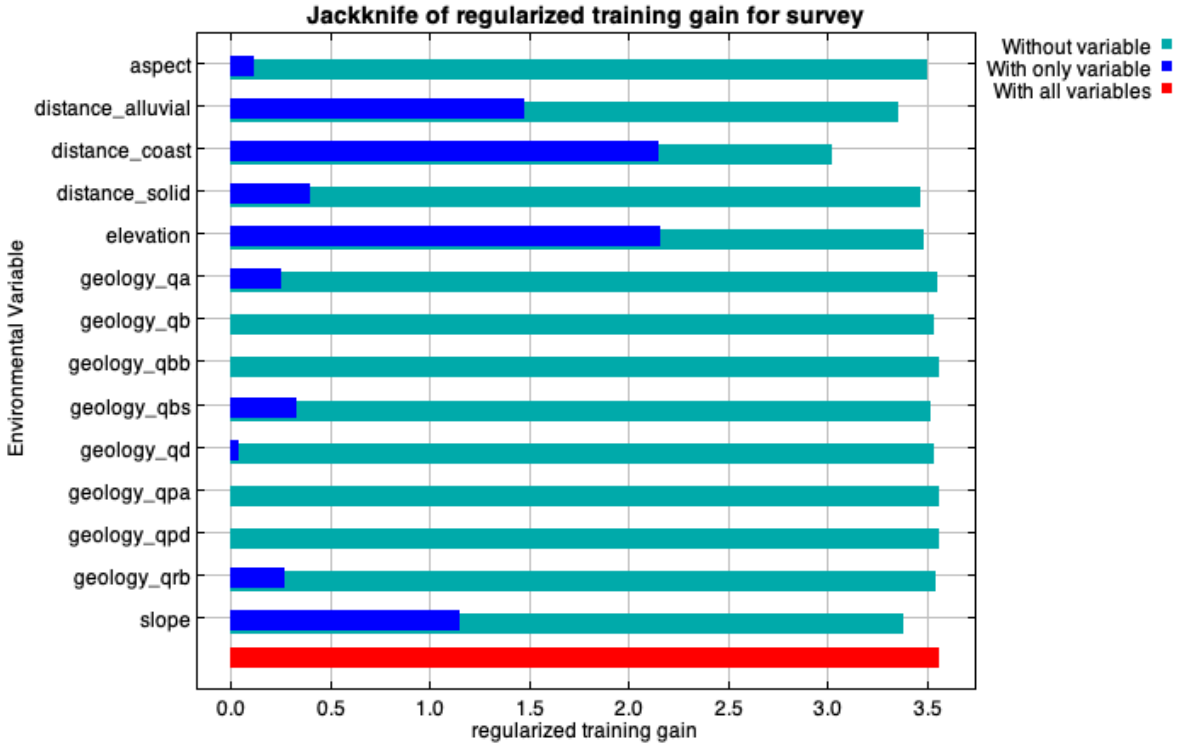


Figure 5-11 Jackknife analysis of 30m grid points MaxEnt model

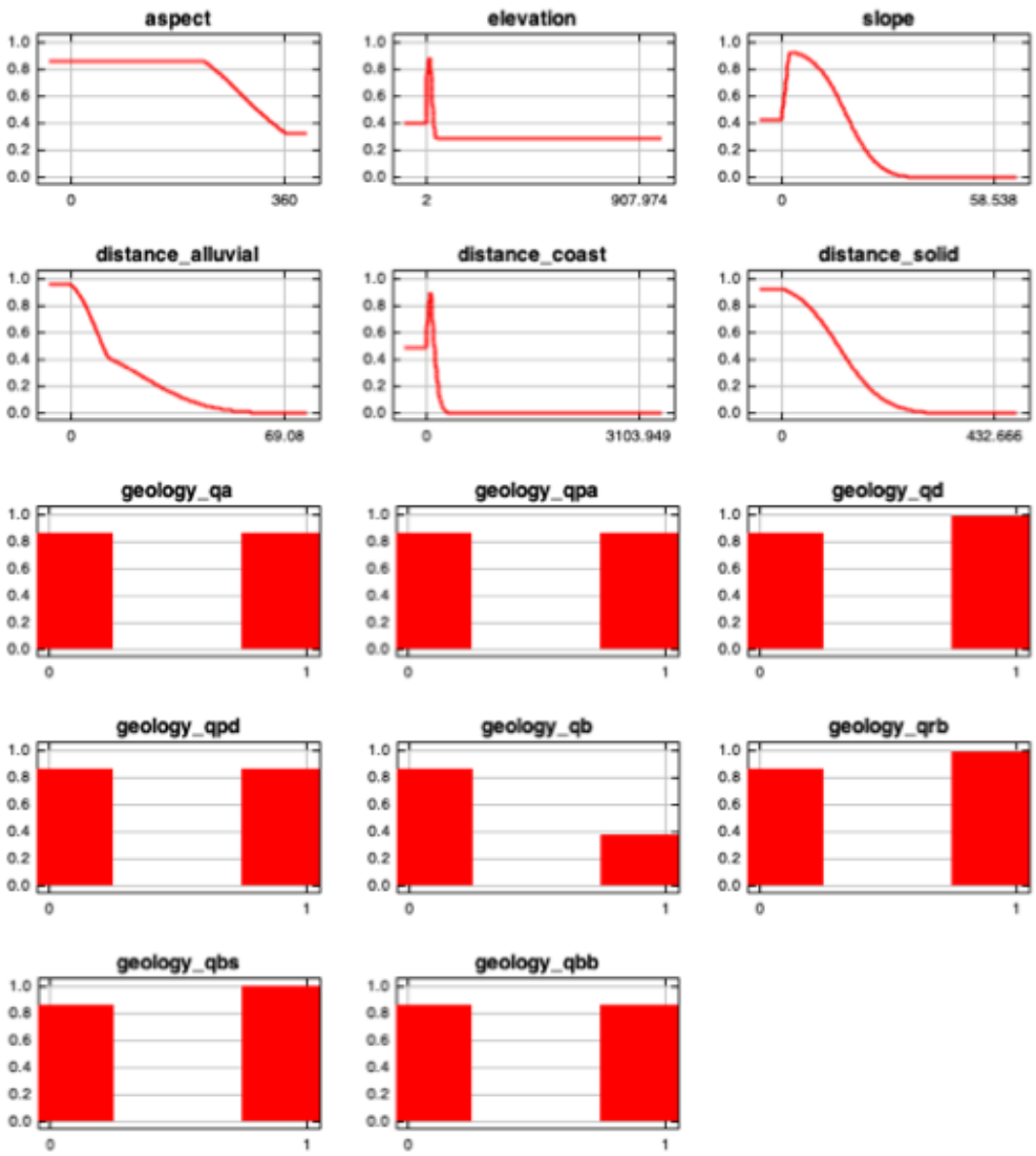


Figure 5-12 Response curves of environmental variables of survey points MaxEnt model.

Aspect

The percentage contribution and permutation importance of the aspect variable are 2.7% and 0.5% respectively, suggesting it plays a minor role in the model prediction it has unnotable influence on the model performance. As shown in the jackknife analysis, the gain/performance increased slightly when only using aspect for prediction and did not decrease when the model predicted without it. The response curve decreases at constant rate between 225° and 350°, suggesting lower archaeological potential when aspect is facing west, northwest and north direction. However, the interpretation may not be reliable as the sample size and variable importance are low.

Elevation

Elevation has the highest percentage contribution and permutation importance, 32.9% and 0.7% respectively, suggesting elevation played the most significant role in the prediction. Its percent contribution is also not as dominant as the 30m grid points model. The high percent contribution and low permutation importance imply that the variable weighted heavily during the prediction, but played a less significant role in the model performance. This is reflected in the jackknife analysis that it has high gain/performance when used alone, suggesting it has useful information for the prediction, but has little effect on the performance when the model is predicted without it. The response curve shows higher archaeological potential at < 50mPD and decreases as elevation increases.

Slope

The effect of slope in the survey points model is similar to the 30m grid points model. The percentage contribution and permutation importance of the slope variable are 4.8% and 4.4% respectively, suggesting it plays a lesser role and has a small impact on the model prediction. The response curve also suggests a gentler slope has higher archaeological potential.

Distance to alluvial deposit

The percentage contribution and permutation importance of the distance to alluvial deposit variable are 19.7% and 2.5% respectively. Distance to alluvial deposits plays a notable role in the prediction but has less influence on the performance. The jackknife analysis suggests gain/performance increased when only the variable is used in the prediction, while there is a slight decrease when the model is predicted without it. The response curves show that the archaeological potential decreases as the distance to alluvial deposits increases, suggesting the potential is higher when the locations are closer to alluvial deposits.

Distance to coast

Distance to coast has the second highest percentage contribution and the highest permutation, 31.5% and 90.2% respectively, suggesting that the variable plays a significant role in the prediction. The dominantly high value of permutation importance suggests it has huge impact towards the model's performance. It is reflected in the jackknife analysis that gain/performance largely increased when only the variable was used in the prediction, while there was a significant decrease when the model predicted without it. The archaeological potential increases between 0m and ~50m, then decreases as the distance increases.

Distance to solid geology

The effect of distance to solid geology in the survey points model is similar to the 30m grid points model. The percentage contribution and permutation importance of the variable are 2.8% and 1.4% respectively, suggesting it plays a less significant role. The jackknife analysis suggests gain/performance slightly increased when only the variable is used in the prediction, while there was unnotable decrease when the model predicted without it. The response curve is flat similar to aspect, suggesting it has minimum effect on the prediction. The response curve shows that the archaeological potential decreases when the locations are further from solid geology, which could be explained by the small distance of the samples (mean = 37m). The response curve using only the variable (the second set of response curves), may better represent the relationship of distance to solid geology and archaeological potential that the potential remains low when the locations are too close to the solid geology (Figure 5-13**Error! Reference source not found.**). Nonetheless, the percent contribution and permutation importance remain low for the variable to be representative.

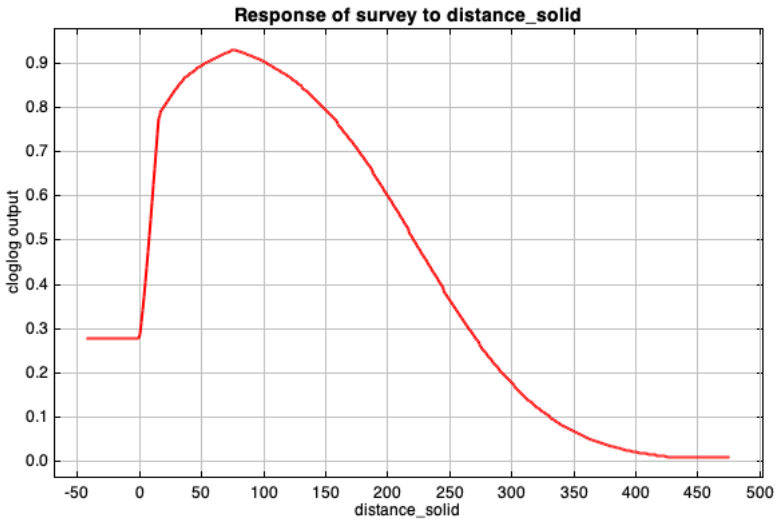


Figure 5-13 Response curve of Distance to Solid when only the variable is used in the prediction

Geology

Geology also generally played an insignificant role in the model prediction, deposits Qbs, Qd, Qb and Qrb have 2.2%, 1.5%, 1.1% and 0.9% percent contribution respectively; 0.1%, 0.2%, 0.1 and 0% permutation importance respectively. Qa, Qpa, Qpd and Qbb deposits have 0% percent contribution and permutation. The jackknife analysis and response curves also show the deposits have insignificant effect on the prediction. The response curves generally show the potential is higher when the locations are located in this geology. Except for Qb that the potential is higher if the locations is not the deposit, but the response may be unreliable as there are simply no samples with Qb deposits for it to calculate the potential.

6 Discussion

6.1 Discussion on MaxEnt and SVM

The AUC scores of all four models are >= 0.8, suggesting that the models are capable of making reliable predictions, as an AUC score of 0.5 indicates random prediction. Judging from the AUC score alone, the SVM appears to perform better than MaxEnt.

Differing from the AUC values, the MaxEnt prediction model outperformed the SVM models, with higher sensitivity observed for both points and SAIs as a unit. The differences in conclusions could be attributed to variations in the evaluation approach.

Furthermore, two concerns arise when examining sensitivity:

- The sensitivity of SAIs as a unit is higher than points as units. This approach may not be optimal, as SAIs are considered predicted with potential if a single raster cell (30m * 30m) is predicted with potential. However, it may suggest that certain areas are more likely to have archaeological potential within each SAI's extent.

- Models trained with 30m grid points performed better than survey points. The 30m grid points data encompassed a wider range of environments, with all survey points located within the SAIs. Therefore, it is expected that models trained with 30m grid points can capture the environmental characteristics of the survey points. Conversely, since survey data points represent a limited range of environments, only a small portion of SAIs resembling the survey points' environment are likely to be predicted.

In addition, the sensitivity of SAI reflects the survey points are able to train models to identify areas in at least half of the SAIs, suggesting the small sample size survey data is able to perform prediction.

6.2 Discussion on the Prediction

The predicted areas with archaeological potential predominantly cluster along the coastal regions characterized by lower elevation levels. Notable differences emerge between the 30m grid point models and the survey points models in terms of spatial distribution:

- 30m Grid Point Models: The prediction outcomes of these models reveal widespread areas identified as having archaeological potential. Particularly, a significant portion of the coastal regions on the island are considered as having archaeological potential.
- Survey Points Models: In contrast, the models based on survey data exhibit more concentrated areas of predicted potential, where predicted potential areas are focused on coastal bays.

While the prediction results demonstrate consistency in identifying coastal and bay areas with archaeological potential across models trained on similar datasets. There are variations in the extent and locality of these areas. The nuanced differences in the predicted potential areas are likely the influence of the modelling techniques on the spatial distribution of archaeological potential within the island.

The prediction maps reflect the environmental preferences of prehistoric people on Lantau Island, emphasizing low-level coastal areas. This supports the findings from previous archaeological investigation suggesting prehistoric human preference of locations close to bay. In addition to predicting the known archaeological locations, the prediction map identifies various coastal areas to having archaeological potential that are not investigated archaeologically.

The influence of environmental variables further supports the coastal preference of prehistoric humans. Analysis from MaxEnt models of 30m grid points and survey points indicates that elevation and coast proximity play significant roles, while distance to alluvial deposits and slope have minor roles in predicting archaeological potential locations. The response curve suggests that locations at lower elevations and closer to the coast are more likely to have higher archaeological potential. This aligned with archaeological studies that prehistoric humans prefers locations close to water bodies for resources and sustainability. Low-level and flat areas allow easier accessibility to these resources.

Geology does not significantly impact the prediction, especially in models trained with 30m grid points, which encompass a broader environmental variety. However, models trained with survey data reflect a slight preference for raised sand beach environments. The response curves suggest higher potential in locations on raised beach deposits (Qrb) and

back shore deposits (Qbs), with slightly lower potential on beach deposits (Qb). Two possibilities for these results are: raised beach deposits may offer more favourable sandbar environments than beach deposits, or there may simply be an absence of samples with beach deposits for potential calculation.

6.3 Research Limitation

6.3.1 Data Collection

The collection of archaeological data was incomplete due to time limitations when requesting archaeological reports from the AMO. An estimation suggests that one-third of the archaeological reports were not viewed. Furthermore, since the report data was not digitally available, they were collected and stored in a database manually through georeferencing, potentially leading to inaccuracies.

Most environmental data used in this study is similar to that used in archaeological prediction models, such as elevation, slope, aspect, and geology. However, some of the data varied slightly due to limitations in data availability. Alternatives were sought, such as distance to alluvial deposits, distance to marine deposits (coast), and distance to solid deposits. Furthermore, the historical environment is not comprehensively reflected due to challenges in reconstructing the past environment and data limitations.

6.3.2 Spatial and Temporal Limitation

The resolution of the predictive model could pose challenge since the scale is smaller than the models from other archaeological prediction model studies. The suitability of predictive modeling at a smaller scale and finer resolution remains uncertain, since the impact of larger and smaller environment may vary towards human settlement pattern. For instance, larger rivers, smaller rivers or even streams were taken into account as water resources at smaller scale and finer resolution, while the correlation between larger rivers and human settlements is well-documented, the relationships between these smaller rivers or streams and settlement could be less obvious

In the discussion above, it is noted that the environmental data does not comprehensively represent the historical environment, especially considering the dynamic nature of the environment throughout history, with factors such as climate, landscape, and sea-level changing at different points in time. This study encountered challenges in reconstructing the past environment for modelling purposes.

Additionally, it is observed that some landscapes within Lantau Island have been disturbed or developed, leading to disruptions such as urban construction and reclamation. Issues related to reclamation were addressed by excluding areas confirmed to have no archaeological potential when establishing the ancient coastline. However, disturbances from urban development remain a challenge in reconstructing the past landscape.

6.3.3 Assumption and Simplification

Since the scale of other similar studies is larger, they are able to utilize known archaeological settlements for prediction purposes. However, the number of known settlements on the island is limited, with only the settlement at Pa Tau Kwu being well-documented and discovered during data collection. Instead of settlements, this study relies on archaeological findings, such as pottery sherds, as indicators of archaeological potential. It is acknowledged that archaeological findings, especially small finds, do not definitively represent human settlements.

To address the scarcity of archaeological data from reports, SAIs were employed to signify areas with archaeological potential. The study assumed that the entire extent of SAIs holds archaeological potential, even in areas without documented findings. The widespread coverage of predicted potential areas by model predictions indicates that a diverse range of environments is considered as potential areas. However, uncertainty remains regarding whether the extent includes disturbances as potential areas.

As discussed in the Literature Review (ref), this predictive modelling approach has been criticized as environmentally deterministic, a critique that this study cannot evade. A major limitation lies in the challenges of spatially mapping non-environmental factors, such as

cultural and economic influences. It is also noted that the complex environment has been simplified for the prediction purposes, with only a limited selection of environmental factors taken into account.

Furthermore, the machine learning approach is dependent on previously known archaeological studies and would only locate the areas of environment resembling known archaeological sites. Hence, potential archaeological sites of different environmental settings are neglected.

6.4 Further Study

6.4.1 Refining Modelling Approach

This study considered the MaxEnt and SVM models, with the former likely being the most suitable algorithm in archaeological prediction modeling. However, as technology is advancing rapidly and new machine learning approaches are being introduced, it is worth delving into the nuances of predictive modeling approaches and exploring new or advanced techniques that could be employed in the study. It is also worth exploring how to refine the MaxEnt and SVM models to perform better in identifying archaeological sites, including optimizing the parameters. Furthermore, it should be noted that machine learning is not the only method for establishing archaeological potential; traditional geospatial analysis, overlaying environmental and cultural data to identify potential areas, is also valuable with established knowledge in archaeology.

The computational archaeological method is a theoretical approach since it does not replace but complements traditional archaeological survey and fieldwork. The archaeological potential of the areas remains unconfirmed until proper archaeological surveys and fieldwork are conducted to verify their presence. Any future archaeological investigation could contribute new archaeological data to further refine and improve the model.

6.4.2 Historical Environment Reconstruction

One of the limitations faced in this study is the reconstruction of the historical environment. Although challenging, attempts to reconstruct the past environment can be made. Not only can this potentially enhance the performance of predictive modelling, but it is also useful in other archaeological studies, such as agent-based modelling. Approaches such as remote sensing, photogrammetry, and 3D modelling have been used to identify hidden features or topographic changes that may prove helpful. Another example is Historic Landscape Characterization (HLC), a method used to recognize the historic character of the landscape. Such approaches have been employed in Hong Kong (Atha and Turner 2022), where a list of historic character types was drafted, including HLC types associated with modern days and historical periods. It is worth exploring how such an approach could aid in the environmental-based modelling approach.

6.4.3 Different Prediction Extent and Location

The same predictive modelling approach can be applied to various regions and extents not only in Hong Kong but also in south-eastern China. Further studies could be undertaken by collecting new archaeological and environmental data to generate new predictions. This would allow the exploration of new relationships between settlements and their environments in different extent and regions, and the examination of potential differences in settlement patterns on Lantau Island compared to other locations. By broadening the scope of analysis to encompass diverse regions and extents, researchers can gain a more comprehensive understanding of the different relationship between human settlements and the surrounding landscape different geographical contexts.

6.5 Implication

This predictive modelling study provides a quantitative method to explore the relationships between human settlements and the environment. It indicates that potential archaeological

sites on Lantau Island are concentrated in sandy bay areas, aligning with findings from previous archaeological studies in Hong Kong.

The study predicts several areas that have not been archaeologically investigated. These predictions could serve as a reference for potential areas of interest for future archaeological surveys. They could also inform development planning by highlighting areas that should be avoided due to their likelihood of containing archaeological remains. However, it is important to recognize the limitations and assumptions inherent in this study. The extent of the predicted potential areas is not definitive, and further archaeological investigations are recommended to validate the archaeological potential of these areas.

7 Conclusion

Previous archaeological studies have suggested the use of machine learning algorithms for predicting potential archaeological locations. MaxEnt has emerged as a popular supervised machine learning model in archaeological predictive modelling. One of the significances of the model is the ability of requiring presence only data and spatial data, which are known archaeological locations in this study and its surrounding environment. In addition, is its ability to present the relationships between the known archaeological potential and their environment. The model's application has been demonstrated through various archaeological studies in different regions, such as south-central Utah, USA, northeast Romania, Indonesia, Netherlands, northeast Israel, China and Japan. Another model OneClassSVM, although rarely seen in archaeological studies, is an unsupervised learning version of the SVM model, also requiring presence only data. This study attempts to employ archaeological predictive modelling to locate archaeological locations, understand the relationships between the archaeological locations and their environment and compare the results of MaxEnt and OneClassSVM.

This study employed the Cross-Industry Standard Process for Data Mining (CRISP-DM) process to collect and prepare data and build and validate model. Two-sets of archaeological data were used in this study, 1. extent of sites of archaeological interest transformed into grid points of 30m distance and 2. archaeological findings location gathered from previous archaeological studies. 14 environmental variables are used in the finalised models, including elevation, slope, aspect, distance coast, distance to alluvial deposit, distance to solid geology and the encoded geology (Qb, Qbs, Qrb, Qa, Qpa, Qd, Qpd and Qbb).

The predictive modelling generated the following results.

1. Performance
   Both MaxEnt and SVM are able to perform prediction to establish the archaeological potential on Lantau Island, Hong Kong.
2. Prediction Map

   Broadly, the predicted areas with archaeological potential predominantly cluster along the coastal regions characterized by lower elevation levels. While the

prediction results demonstrate consistency in identifying coastal and bay areas with archaeological potential across models trained on similar datasets. There are variations in the extent and locality of these areas. Noteworthy distinctions emerge between the 30m grid point models and the survey points models in terms of spatial distribution:

- 30m Grid Point Models: The prediction outcomes of these models reveal widespread areas identified as having archaeological potential. Particularly, a significant portion of the coastal regions on the island are considered as having archaeological potential.

- Survey Points Models: In contrast, the models based on survey data exhibit more concentrated areas of predicted potential, where predicted potential areas are focused on coastal bays.

3. Effects of Environmental Variables on Prediction

   The percentage contribution, permutation importance and jackknife analysis generated from MaxEnt models of 30m grid points and survey points suggested elevation and coast plays a major role, while distance to alluvial deposits and slope play a minor role in predicting archaeological potential locations. The response curve also reflected that locations at lower level of elevation and closer to coast are likely to have higher archaeological potential.

This study encountered several limitations, particularly in data collection, spatial and temporal considerations, and assumptions made in the process. These challenges underscore the complexity of predictive modeling in archaeological research. This highlighting the need for further refinement and exploration in future studies.

The implications of this study extend to providing a quantitative method for investigating human-environment relationships and identifying potential archaeological sites. The predicted areas could guide future archaeological surveys and urban development planning by highlighting areas of archaeological significance. However, it is crucial to acknowledge the study's limitations and the need for validation through archaeological fieldwork to confirm the presence of archaeological remains in the predicted areas.

Bibliography

Antiquities and Monuments Office (2024) *Sites of Archaeological Interest in Hong Kong*. Available from url: https://www.amo.gov.hk/en/historic-buildings/archaeological-interest/index.html. Accessed on 3 June 2024.

Civil Engineering and Development Department (2023) *Agreement No. CE 20/2021 (CE) First Phase Development of the New Territories North – San Tin / Lok Ma Chau Development Node – Investigation*. https://www.epd.gov.hk/eia/register/report/eiareport/eia_3022023/Index.htm

Dormann, C.F., Elith, J., Bacher, S. et al. (2013) 'Collinearity: a review of methods to deal with it and a simulation study evaluating their performance', Ecography, 36, pp. 27-46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Drewett, P.L. (1995) *Neolithic Sha Lo Wan: A Late Neolithic Settlement at Sha Lo Wan, Lantau Island, Hong Kong*. AMO Occasional Paper No. 2.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Yung, E. C. and Yates, C.J. (2011) 'A statistical explanation of MaxEnt for ecologists', *Diversity and Distributions*, 17, pp. 43-57.

Fyfe, J.A., Shaw, R., Campbell, S.D.G., Lai, K.W. and Kirk, P.A. (2000) *The Quaternary Geology of Hong Kong*. Hong Kong: Geotechnical Engineering Office.

Gaffney, V. and van Leusen, M. (1995) 'Postscript – GIS, environmental determinism and archaeology: a parallel text', in Lock, G. and Stančič, Z. (eds) *Archaeology and Geographical Information Systems: A European Perspective*. United Kingdom: Taylor & Francis, pp. 367–382.

Guangzhou Antique Archaeology Institute 廣州市文物考古研究所 (1998) *Xiang Gang Wen Wu Pu Cha Da Yu Shan Bei Qu Gong Zuo Bao Gao 香港文物普查大嶼山北區工作報告*. Unpublished.

Islands District Board (1994) *Heritage of the Islands District*. 2nd Edn. Hong Kong: Island District Board.

Kohler, T. A. and Sandra C. Parker. (1986) 'Predictive Models for Archaeological Resource Location', *Advances in Archaeological Method and Theory*, 9, pp. 397–452. http://www.jstor.org/stable/20210081.

Kvamme, K. L. (1992) 'A Predictive Site Location Model on the High Plains: An Example with an Independent Test', *Plains Anthropologist*, 37, pp. 19-40.

Kamermans, H. (2010) 'The Application of Predictive Modelling in Archaeology: Problems and Possibilities', in Nicolucci, F. and S. Hermon (eds.) *Beyond the Artifact. Digital Interpretation of the Past.* Proceedings of CAA2004, Prato 13–17 April 2004. Budapest: Archaeolingua, pp. 271-277.

Kvamme, K. L. (2006) 'There and Back Again: Revisiting Archaeological Locational Modeling', in Mehrer, M.W. and Wescot, K.L. (eds.) *GIS and Archaeological Site Location Modeling*. United Kingdom: Routledge, pp. 3-38.

Lee, S. (2005) 'Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data', *International Journal of Remote Sensing*, 26(7), pp. 1477–1491. https://doi.org/10.1080/01431160412331331012

Luthfi A. M., Sigit H. M., and Bowo S. (2019) 'MaxEnt (Maximum Entropy) model for predicting prehistoric cave sites in Karst area of Gunung Sewu, Gunung Kidul, Yogyakarta', *Proc. SPIE 11311,* Sixth Geoinformation Science Symposium, 113110B. https://doi.org/10.1117/12.2543522

Meacham, W. (1984) 'Coastal Landforms and Archaeology in the Hong Kong Archipelago', *World Archaeology*, 16(1), pp. 128-135. https://www.jstor.org/stable/124693

Meacham, W. (2009) *The Archaeology of Hong Kong*. Hong Kong: Hong Kong University Press.

Morgan, B. and Guénard, B. (2019), 'New 30 m resolution Hong Kong climate, vegetation, and topography rasters indicate greater spatial variation than global grids within an urban mosaic', *Earth System Science Data*. 11(3), pp.1083-1098. https://doi.org/10.5194/essd-11-1083-2019

Nicu, I.C., Mihu-Pintilie, A., Williamson, J. (2019) 'GIS-Based and Statistical Approaches in Archaeological Predictive Modelling (NE Romania)', *Sustainability*, *11*, pp. 5969. https://doi.org/10.3390/su11215969

Noviello, M., Carfarelli, B., Calculli, C., Sarris, A. and Mairota, P. (2018) 'Investigating the Distribution of Archaeological Sites: Multiparametric vs Probability Models and Potential for Remote Sensing Data', *Applied Geography*, 95, pp. 34-44. https://doi.org/10.1016/j.apgeog.2018.04.005

Pearl, J. (2000) *Causality: Models, Reasoning and Inference.* United Kingdom: Cambridge University Press.

Phillips, S. J., Anderson, R. P. and Schapire, R. E. (2006) 'Maximum Entropy Modeling of Species Geographic Distributions', *Ecological Modelling*, 190, pp. 231-259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S.J. (2017) *A Brief Tutorial on Maxent*. Available from url: https://biodiversityinformatics.amnh.org/open_source/maxent/Maxent_tutorial2017.pdf. Accessed on 2024-07-31.

Philips, S.J., Dudík, M. and Schapire, R.E. (2024) *Maxent Software for Modeling species Niches and Distribution (Version 3.4.1)*. Available from url: http://biodiversityinformatics.amnh.org/open_source/maxent/. Accessed on 2024-06-04.

Ramankutty, N., Foley, J.A., Norman, J. et al. (2002) 'The Global Distribution of Cultivable Lands: Current Patterns and Sensitivity to Possible Climate Change', *Global Ecology and Biogeography*, 11, pp.377-392. https://doi.org/10.1046/j.1466-822x.2002.00294.x

Refrew, C. and Bahn, P. (2016) *Archaeology: Theories, Methods and Practice (6th ed.)*. United Kingdom: Thames & Hudson.

Sanchez-Hernandez, C., Boyd, D.S., Foody, G.M. (2007) 'Mapping specific habitats from remotely sensed imagery: Support vector machine and support vector data description based classification of coastal saltmarsh habitats'. *Ecological* informatics, 2, pp. 83–88. https://doi.org/10.1016/j.ecoinf.2007.04.003

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A. and Williamson, R. (2001) 'Estimating Support of a High-Dimensional Distribution'. *Neural Computation*. 13(7). pp. 1443-1471. https://doi.org/10.1162/089976601750264965

Shang, Z and Ng, W.H. (2010) *The Research and Discussion of Hong Kong Archaeology*. Beijing: Cultural Relics Press.

So, C.L. (1969) '*Land Forms and Archaeology*', Journal of Hong Kong Archaeological Society, 1.

Pedregosa et al. (2011) 'Scikit-learn: Machine Learning in Python', *JMLR 12*, pp. 2825-2830.

Prettenhofer, P. and Vanderplas, J. (2024) *Species distribution modelling*. Avaiable at https://scikit-learn.org/stable/auto_examples/applications/plot_species_distribution_modeling.html#references (Accessed: 29 July 2024)

Wang, Y., Shi, X. and Oguchi, T. (2023) 'Archaeological Predictive Modeling Using Machine Learning and Statistical Methods for Japan and China', *International Journal of Geo-Information*, 12, pp. 238. https://doi.org/10.3390/ijgi12060238

Warren, R. E. (1990) 'Predictive Modelling in Archaeology: A Primer', in Allen, K.M.S., Green, S.W., and Zubrow, E.B.W. (eds.) *Interpreting Space: GIS and Archaeology*. United Kingdom: Taylor & Francis, pp. 90-111.

Wachtel, I., Zidon, R., Garti, S. and Shelach-Lavi, G. (2018) 'Predictive modeling for archaeological site locations: Comparing logistic regression and maximal entropy in north Israel and north-east China', *Journal of Archaeological Science*, 92, pp. 28-36.

Wheatley, D. (2004) 'Making Space for an Archaeology of Place', *Internet Archaeology*. 15. http://intarch.ac.uk/journal/issue15/wheatley_index.html

Willey, G. (1953) '*Prehistoric Settlement Patterns in The Viru Valley, Peru*', *Bureau of American Ethnology Bulletin*, 155, pp. 1-453.

Whitley, T.G. (2003) 'Causality and Cross-purpose in Archaeological Predictive Modeling', in Magistrat der Stadt Wien (eds.), *Computer Applications in Archaeology Conference Vienna,* Austria.

Woodman, P.E. and Woodward, M. (2002) 'The Use and Abuse of Statistical Methods in Archaeological Site Location Modelling', in Wheatley, D., Earl, G. and Poppy, S. (eds.), Contemporary Themes in Archaeological Computing, Oxford: Oxbow Books.

Verhagen, P. (2007) *Case Studies in Archaeological Predictive Modelling*. Netherlands: Leiden University Press.

Verhagen, P., Kamermans, H., van Leusen, M., Deeben, J., Hallewas, D. and Zoetbrood, P. (2010) 'First Thoughts on the Incorporation of Cultural Variables into Predictive Modelling', in Nicolucci, F. and Hermon, S. (eds.), *Beyond the Artifact. Digital Interpretation of the Past*. Proceedings of CAA2004, Prato 13–17 April 2004. Budapest: Archaeolingua, pp. 307-311. https://proceedings.caaconference.org/paper/58_verhagen_et_al_caa_2004/

Verhagen, P. and Whitley, T.G. (2012) 'Integrating Archaeological Theory and Predictive Modeling: A Live Report from the Scene', *Journal of Archaeological Method and Theory*, 19, pp.49–100. https://doi.org/10.1007/s10816-011-9102-7

Yaworsky, P. M., Vernon, K. B., Spangler, J. D., Brewer, S. C. and Codding, B. F. (2020) 'Advancing Predictive Modeling in Archaeology: An Evulation of Regression and Machine Learning Methods on the Grand Stair-case Learning Methods on the Grand Staircase-Escalante National Monument', *PLOS ONE*. https://doi.org/10.1371/journal.pone.0239424

Yan, J. and Zheng, J. (2007) 'One-Class SVM Based Segmentation for SAR Image'. In: Liu, D., Fei, S., Hou, Z., Zhang, H. and Sun, C. (eds.) *Advances in Neural Networks – ISNN 2007*. ISNN 2007. Lecture Notes in

Computer Science, vol 4493. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72395-0_117

Appendix A – Exploratory Data Analysis

Set Up

In [ ]:
```
# import libraries
import geopandas as gpd
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

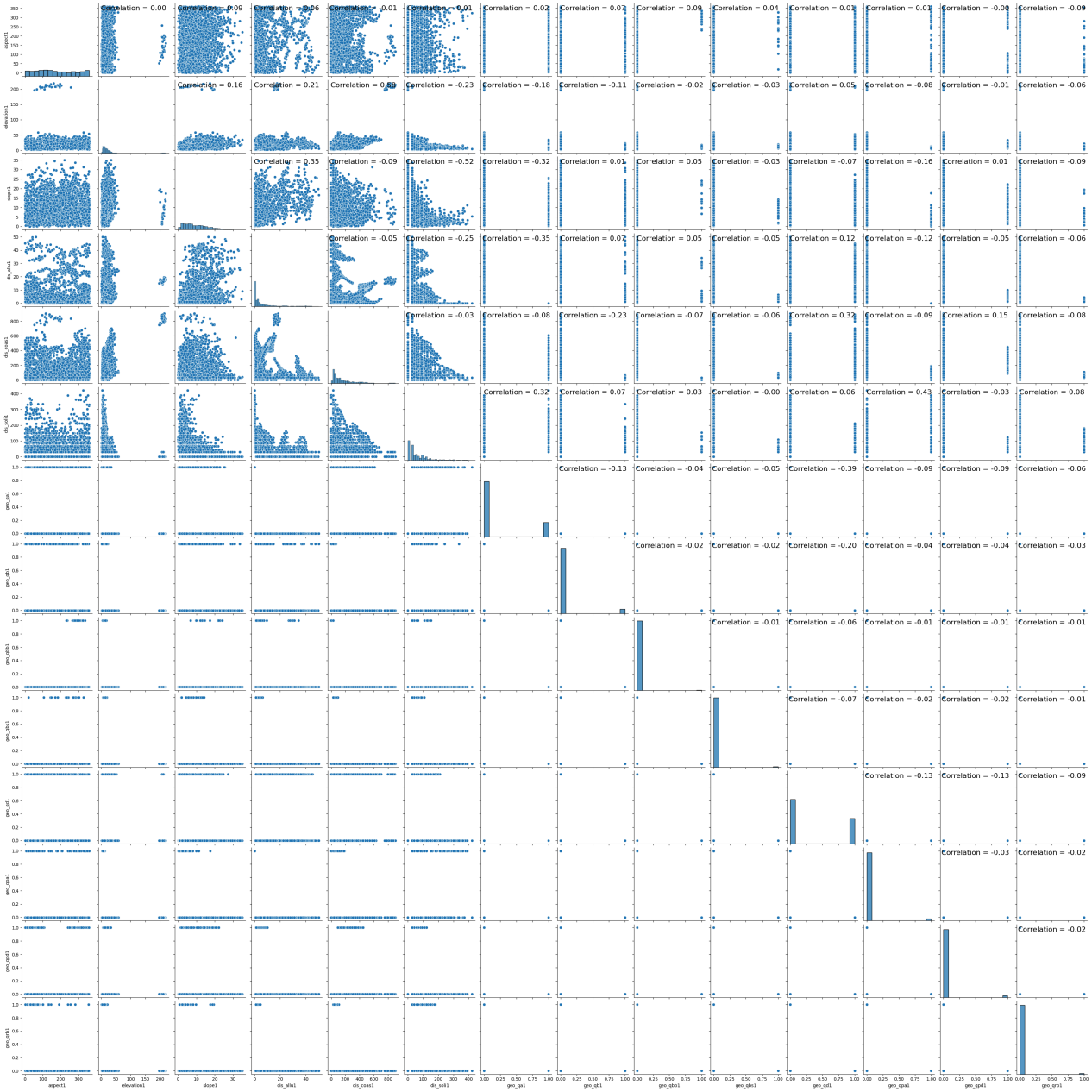
In [ ]:
```
# 2 sets of input data point

# 30m grid points
gdf_30m = gpd.read_file("  ") # input 30m grid points file location

# survey points
gdf_survey = gpd.read_file("  ") # input survey points file location
```

Plotting Pairwise Relationship of Environmental Variables

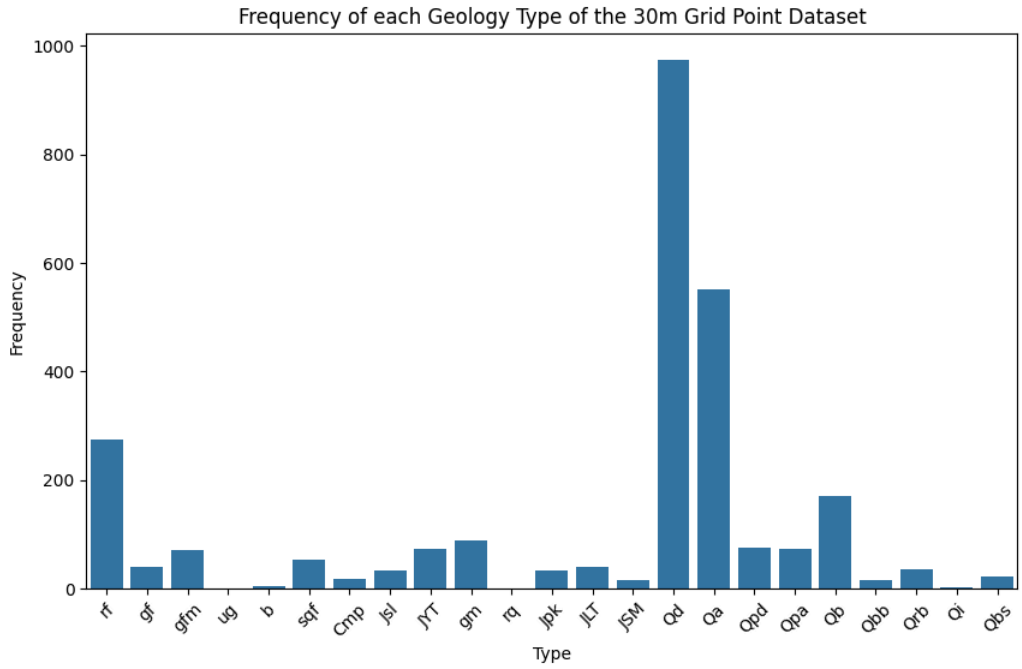
In [ ]:
```
# Environmental Variables Column Names
# aspect1 as Aspect
# elevation1 as Elevation
# slope1 as Slope
# dis_allu1 as Distance to Alluvial Deposit
# dis_coas1 as Distance to Coastline
# dis_soli1 as Distance to Solid Geology
# geo_qa1 as Qa geology
# geo_qb1 as Qb geology
# geo_qbb1 as Qbb geology
# geo_qbs1 as Qbs geology
# geo_qd1 as Qd geology
# geo_qpa1 as Qpa geology
# geo_qpd1 as Qpd geology
# geo_qrb1 as Qrb geology

# columns to visualise for 30m grid points
columns_30m = ['aspect1', 'elevation1', 'slope1', 'dis_allu1', 'dis_coas1','dis_soli1', 'geo_qa1', 'geo_qb1', 'geo_qbb1', 'geo_qbs1'
       'geo_qd1', 'geo_qpa1', 'geo_qpd1', 'geo_qrb1']

# columns to visualise for survey points
columns_survey = ['aspect1', 'elevation1', 'slope1', 'dis_allu1', 'dis_coas1','dis_soli1', 'geo_qa1', 'geo_qbs1',
       'geo_qd1', 'geo_qrb1']
```

30m Grid Point Dataset

In [ ]:
```
# Create a new dataframe for gdf_30m
df = gdf_30m[columns_30m]

# Create the pair plot for gdf_30m
pairplot = sns.pairplot(df)

# Calculate and display the correlation coefficient for each pair of variables
for i, col1 in enumerate(pairplot.axes):
    for j, col2 in enumerate(col1):
        if j > i:
            correlation = np.corrcoef(df[columns_30m[j]], df[columns_30m[i]])[0, 1]
            col2.text(0.5, 0.9, f"Correlation = {correlation:.2f}", transform=col2.transAxes, ha='center', fontsize=16)


plt.show()
```



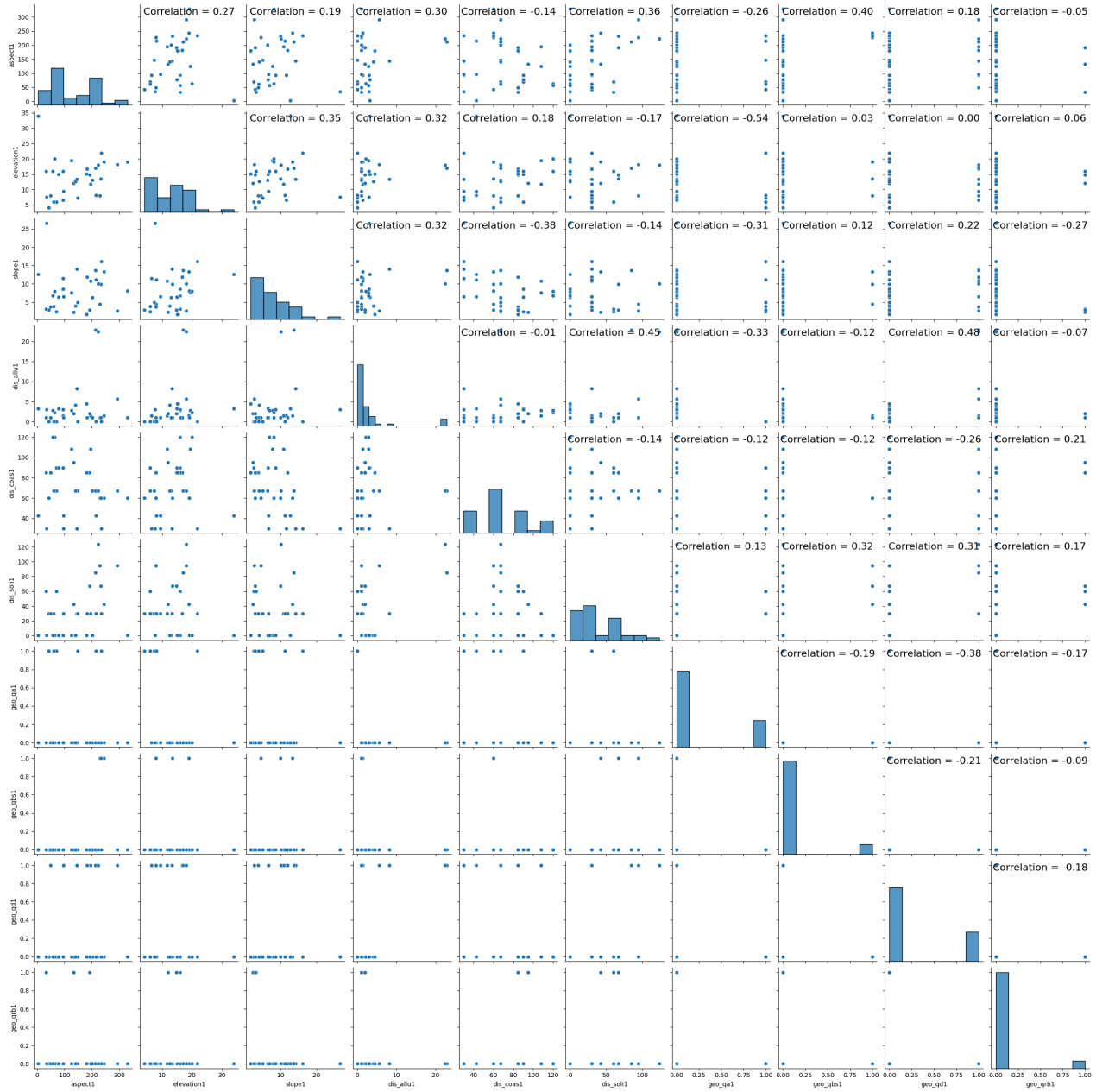
In [ ]:
```
# Visualise geology type of the 30m Grid Point

# Set the figure size
plt.figure(figsize=(10, 6))  # Adjust the width and height as desired

# Create a bar plot of the frequency
sns.countplot(data=gdf_30m, x='Type')

# Rotate the x-labels for better readability
plt.xticks(rotation=45)  # Adjust the rotation angle as needed

# Set the x and y axis labels
plt.title('Frequency of each Geology Type of the 30m Grid Point Dataset')
plt.xlabel('Type')
plt.ylabel('Frequency')

# Show the plot
plt.show()
```



Survey Report Dataset

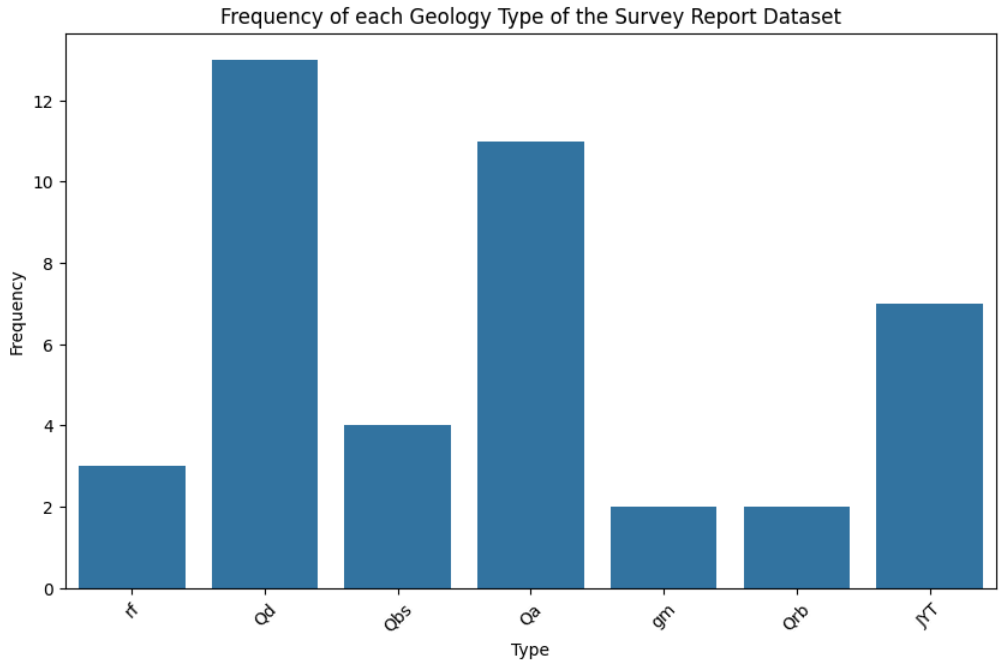
In [ ]:

```
# Create a new dataframe for gdf_survey
df = gdf_survey[columns_survey]

# Create the pair plot for gdf_survey
pairplot = sns.pairplot(df)

# Calculate and display the correlation coefficient for each pair of variables
for i, col1 in enumerate(pairplot.axes):
    for j, col2 in enumerate(col1):
        if j > i:
            correlation = np.corrcoef(df[columns_survey[j]], df[columns_survey[i]])[0, 1]
            col2.text(0.5, 0.9, f"Correlation = {correlation:.2f}", transform=col2.transAxes, ha='center', fontsize=16)

# Show the plot
plt.show()
```



In [ ]:

```
# Visualise geology type of the 30m Grid Point

# Set the figure size
plt.figure(figsize=(10, 6))  # Adjust the width and height as desired

# Create a bar plot of the frequency
sns.countplot(data=gdf_survey, x='Type')

# Rotate the x-labels for better readability
plt.xticks(rotation=45)  # Adjust the rotation angle as needed

# Set the x and y axis labels
plt.title('Frequency of each Geology Type of the Survey Report Dataset')
plt.xlabel('Type')
plt.ylabel('Frequency')

# Show the plot
plt.show()
```


Frequency of each Geology Type of the Survey Report Dataset
Frequency
0
2
4
6
8
10
12
rf
Qd
Qbs
Qa
gm
Qrb
JYT
Type

Appendix B – SVM Script

In [ ]:
```
# setup
import glob
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

import geopandas
import rasterio

from sklearn import metrics, svm
from sklearn.model_selection import train_test_split
from sklearn.utils import Bunch

class_name = 'prehistoric'
asc_folder_path = 'asc' + '/*.asc' # 'folder path of environmental variables folder' + '/*.asc' for iteration of the variables, change accordingly

pts_path_30m = 'shp/30m.shp' # 30m points path, change accordingly
pts_path_survey = 'shp/survey.shp' # survey points path, change accordingly
```

In [ ]:
```
# Extract env rasters values such as,
#    grid size  (grid_size)
#    number of rows (Nrow_y)
#    number of columns  (Ncol_x)
#    left(west) most corner in longtitude   (y_left_lower_corner)
#    bottom(south) corner in latitude   (x_left_lower_corner = [])
#    coverages  (coverages)

def prepare_env_asc_data(asc_folder_path):
    grid_size=[]
    Nrow_y=[]
    Ncol_x = []
    y_left_lower_corner = []
    x_left_lower_corner = []
    coverages = []

    for asc in glob.glob(asc_folder_path):
        raster = rasterio.open(asc)

        grid_size.append(raster.meta['transform'][0])
        Nrow_y.append(raster.meta['height'])
        Ncol_x.append(raster.meta['width'])
        y_left_lower_corner.append(raster.bounds[1])
        x_left_lower_corner.append(raster.bounds[0])
        coverages.append(raster.read()[0])

    # check if the values except coverages are matching
    # if not may require reviewing outside this script
    if len(set(grid_size))!=1 or len(set(Nrow_y))!=1 or len(set(Ncol_x))!=1 or len(set(y_left_lower_corner))!=1 or len(set(x_left_lower_corner))!=1:
        print('!! Warning Env rasters are not matching !!')
        return None
    else:
        print('- env_data prepared successfully')
        data = dict(grid_size= grid_size[0],
                    Nrow_y= Nrow_y[0],
                    Ncol_x = Ncol_x[0],
                    y_left_lower_corner = y_left_lower_corner[0],
                    x_left_lower_corner = x_left_lower_corner[0],
                    coverages = np.array(coverages))
        return data

# Preview env_data
# env_data = prepare_env_asc_data(asc_folder_path=asc_folder_path)
# for i in env_data:
#     print(env_data[i])
```

In [ ]:
```
# Prepare archaeological presence data points for train test split
def prepare_presence_point_data(pts_path, test_pts_path = None, x_long_colname = "X", y_lat_colname = "Y", define_class_value = class_name):
    points = geopandas.read_file(pts_path)

    if test_pts_path: # for survey point model, training data = survey points, testing data = 30m points
        points_test = geopandas.read_file(test_pts_path)

        points_train = points.loc[:,[x_long_colname, y_lat_colname]]
        points_test = points_test.loc[:,[x_long_colname, y_lat_colname]]

    else: # for 30m point model, train-test split 30m points
        points_train, points_test = train_test_split(points, train_size=0.7)

    points_train["class"] = define_class_value
    points_test["class"] = define_class_value

    points_train = points_train.to_records(index=False, column_dtypes={'X': '<f4', 'Y': '<f4', 'class': 'S22'})
    points_test = points_test.to_records(index=False, column_dtypes={'X': '<f4', 'Y': '<f4', 'class': 'S22'})

    pts_data = dict(train=points_train, test=points_test)

    print('- pts_data prepared successfully')
    return pts_data

# Preview pts_data
#pts_data = prepare_presence_point_data(pts_path=pts_path_30m, test_pts_path = None, x_long_colname="X", y_lat_colname="Y", define_class_value=class_name)
#print(pts_data['train'])
#print(len(pts_data['train']), len(pts_data['test']))
```

In [ ]:
```
# Construct the map grid from the batch object
def construct_grids(data):
    # x,y coordinates for corner cells
    xmin = data['x_left_lower_corner'] + data['grid_size']
    xmax = xmin + (data['Ncol_x'] * data['grid_size'])
    ymin = data['y_left_lower_corner'] + data['grid_size']
    ymax = ymin + (data['Nrow_y'] * data['grid_size'])

    # x coordinates of the grid cells
    xgrid = np.arange(xmin, xmax, data['grid_size'])
    # y coordinates of the grid cells
    ygrid = np.arange(ymin, ymax, data['grid_size'])

    print('- x, y grid constructed successfully')
    return xgrid, ygrid

# Preview xgrid, ygrid
#xgrid, ygrid = construct_grids(data=env_data)
#print(xgrid.shape)
#print(ygrid.shape)
```

In [ ]:
```
def create_archaeology_bunch(class_name, train, test, coverages, xgrid, ygrid):
    """Create a bunch with information

    This will use the test/train record arrays to extract the
    data specific to the given archaeology class.
    """
    bunch = Bunch(name=" ".join(class_name.split("_")[:2]))
    class_name = class_name.encode("ascii")
    points = dict(test=test, train=train)

    for label, pts in points.items():
        # choose points associated with the desired species
        pts = pts[pts["class"] == class_name]
        bunch["pts_%s" % label] = pts
        #print(bunch["pts_%s" % label])

        # determine coverage values for each of the training & testing points
        ix = np.searchsorted(xgrid, pts["X"])
        iy = np.searchsorted(ygrid, pts["Y"])
        bunch["cov_%s" % label] = coverages[:, -iy, ix].T

    print('- bunch created successfully')
```

```
    return bunch

# Preview prehistoric_bunch
# prehistoric_bunch = create_archaeology_bunch(class_name=class_name, train=pts_data['train'], test=pts_data['test'], coverages=env_data['coverages'], xgrid=xgrid, ygrid=ygrid)
# for i in prehistoric_bunch:
#     print(prehistoric_bunch)
```

In [ ]:

```
def plot_prediction(env_data, pts_data, plot_name, result_folder, nu=0.5, kernel="rbf", gamma= 'scale'): # nu, kernel, gamma OneClassSVM parameters

    # Set up the data grid
    xgrid, ygrid = construct_grids(data=env_data)

    # The grid in x,y coordinates
    X, Y = np.meshgrid(xgrid, ygrid[::-1])

    # create bunch
    prehistoric_bunch = create_archaeology_bunch(class_name=class_name,
                                                 train=pts_data['train'],
                                                 test=pts_data['test'],
                                                 coverages=env_data['coverages'],
                                                 xgrid=xgrid,
                                                 ygrid=ygrid)

    # take reference of coverages[1] to decide land and water. coverages[1] is the elevation raster.
    land_reference = env_data['coverages'][0]

    # Fit, predict, and plot.
    print("_" * 80)
    print("Modeling distribution of '%s'" % class_name)

    # Standardize features
    mean = prehistoric_bunch['cov_train'].mean(axis=0)
    std = prehistoric_bunch['cov_train'].std(axis=0)
    train_cover_std = (prehistoric_bunch['cov_train'] - mean) / std

    # Fit OneClassSVM
    print(" - fit OneClassSVM ... ", end="")
    clf = svm.OneClassSVM(nu = nu, kernel = kernel, gamma = gamma)
    clf.fit(train_cover_std)
    print("done.")

    # Plot map
    print(" - plot coastlines from coverage... ", end="")
    plt.contour(
    X, Y, land_reference, levels=[-9998], colors="k", linestyles="solid"
    )
    print("done.")

    # Predict species distribution using the training data
    print(" - predict... ", end="")
    Z = np.ones((env_data['Nrow_y'], env_data['Ncol_x'],), dtype=np.float64)

    # Predict only for the land points
    idx = np.where(land_reference > -9999)
    coverages_land = env_data['coverages'][:, idx[0], idx[1]].T

    pred = clf.decision_function((coverages_land - mean) / std)
    Z *= pred.min()
    Z[idx[0], idx[1]] = pred

    levels = np.linspace(Z.min(), Z.max(), 25)
    Z[land_reference == -9999] = -9999
    print("done.")

    # plot contours of the prediction
    print(" - plot prediction... ", end="")
    plt.contourf(X, Y, Z, levels=levels, cmap='Greens')
    plt.colorbar(format="%.2f")

    # scatter training/testing points
    plt.scatter(
        pts_data['train']['X'],
        pts_data['train']['Y'],
        s=2**2,
        c="red",
        marker="^",
        label="train",
    )
    plt.scatter(
        pts_data['test']['X'],
        pts_data['test']['Y'],
        s=2**2,
        c="black",
        marker="x",
        label="test",
    )
    plt.legend()
    plt.title(plot_name)
    plt.axis("equal")
    print("done.")

    # export prediction as raster
    # this method take reference of the input environemental raster
    print(" - create raster... ", end="")
    raster = rasterio.open(glob.glob(asc_folder_path)[0])
    new_dataset = rasterio.open(
        result_folder,
        'w',
        driver=raster.driver,
        height=raster.height,
        width=raster.width,
        count=raster.count,
        dtype=raster.meta['dtype'],
        crs=raster.crs,
        transform=raster.transform,
        nodata=raster.nodata
    )
    new_dataset.write(Z, 1)
    new_dataset.close()
    print("done.")

    # background points (grid coordinates) for evaluation
    print(" - compute AUC... ", end="")
    np.random.seed(13)
    background_points = np.c_[
        np.random.randint(low=0, high=env_data['Nrow_y'], size=10000),
        np.random.randint(low=0, high=env_data['Ncol_x'], size=10000),
    ].T

    # Compute AUC with regards to background points
    pred_background = Z[background_points[0], background_points[1]]
    pred_test = clf.decision_function((prehistoric_bunch['cov_test'] - mean) / std)
    scores = np.r_[pred_test, pred_background]
    y = np.r_[np.ones(pred_test.shape), np.zeros(pred_background.shape)]
    fpr, tpr, thresholds = metrics.roc_curve(y, scores)
    roc_auc = metrics.auc(fpr, tpr)
    print("done.")

    print("_" * 80)
    print("\n Area under the ROC curve : %f" % roc_auc)
```

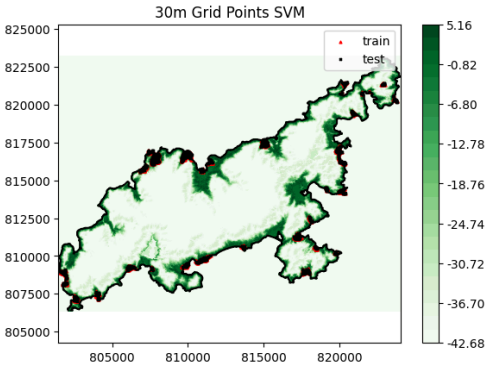
In [ ]:

```
# 30m grid points SVM
# Load data
env_data = prepare_env_asc_data(asc_folder_path=asc_folder_path)
pts_data = prepare_presence_point_data(pts_path=pts_path_30m, test_pts_path = None, x_long_colname="X", y_lat_colname="Y", define_class_value=class_name)

# Prediction
Z = plot_prediction(env_data, pts_data, plot_name = '30m Grid Points SVM', result_folder = 'SVM_30m/svm.asc', nu=0.5, kernel="rbf", gamma= 1000) # result folder path change accordingly
```

```
- env_data prepared successfully
- pts_data prepared successfully
- x, y grid constructed successfully
- bunch created successfully
```

```
Modeling distribution of 'prehistoric'
 - fit OneClassSVM ... done.
 - plot coastlines from coverage... done.
 - predict... done.
 - plot prediction... done.
 - create raster... done.
 - compute AUC... done.
```

```
 Area under the ROC curve : 0.970572
```



In [ ]:
```
# Survey points SVM
# Load data
env_data = prepare_env_asc_data(asc_folder_path=asc_folder_path)
pts_data = prepare_presence_point_data(pts_path=pts_path_survey, test_pts_path = pts_path_30m, x_long_colname="X", y_lat_colname="Y", define_class_value=class_name)

# Prediction
Z = plot_prediction(env_data, pts_data, plot_name = 'Survey Points SVM', result_folder = 'SVM_survey/svm.asc', nu=0.5, kernel="rbf", gamma= 'scale') # result folder path change accordingly
```

```
- env_data prepared successfully
- pts_data prepared successfully
- x, y grid constructed successfully
- bunch created successfully
```

```
Modeling distribution of 'prehistoric'
 - fit OneClassSVM ... done.
 - plot coastlines from coverage... done.
 - predict... done.
 - plot prediction... done.
 - create raster... done.
 - compute AUC... done.
```

```
 Area under the ROC curve : 0.942072
```